Testing English Collocations : Developing Receptive Tests for Use with Advanced Swedish Learners

Gyllstad, Henrik

2007

Testing English Collocations

Developing Receptive Tests for Use with Advanced Swedish Learners

Henrik Gyllstad

LUND
UNIVERSITY

Testing English Collocations

Developing Receptive Tests for Use with Advanced Swedish Learners

ISBN: 978-91-628-7296-0
Printed by Media-Tryck in Lund 2007

Details about this doctoral dissertation

This version of Henrik Gyllstad's doctoral dissertation is a digital PDF-version of the hard copy submitted for his viva on 27 October 2007, which, on that same day, was publicly defended and approved for the degree of Doctor of Philosophy, at Lund University, Sweden.

Viva opponent:

Reader Dr. Norbert Schmitt, University of Nottingham

Examining Committee:

Professor Moira Linnarud, Karlstad University
Dr. Britt Erman, Stockholm University
Dr. Jonas Granfeldt, Lund University

Supervisors:

Dr. Marie Källkvist, Lund University
Professor Beatrice Warren, Lund University

ABSTRACT

The research reported in this thesis has two main aims. The first aim is to develop tests capable of yielding reliable and valid scores of receptive knowledge of English collocations as a single construct, for use with advanced L2 learners of English. Collocations are seen as conventionalized, recurring combinations of words, and the targeted types are adjective + NP and verb + NP. The second aim is to chart the levels of receptive collocation knowledge in advanced Swedish learners of English, and investigate the relationship between receptive collocation knowledge, vocabulary size, and learning level. In a series of seven empirical studies, involving students of English in Sweden as well as native speakers of English, the two main aims of the thesis are addressed through three research questions. The informants in Sweden are L2 learners of English at upper-secondary school and university level, who have had 8 and 11 years of classroom instruction in English.

The results show that the two tests developed – called COLLEX and COLLMATCH – yield reliable scores, and show evidence of different types of validity, such as construct validity, concurrent validity, and face validity. Further investigation is needed in terms of content validity, and certain lingering problems are identified with regard to ceiling effects. It is furthermore shown that a) scores on COLLEX and COLLMATCH increase as a function of learning level, b) the two tests discriminate well between learners of different proficiency levels, and between learners and native speakers of English, and c) scores on COLLEX and COLLMATCH correlate highly with scores on a receptive vocabulary size test. The results suggest that there is a close relationship between advanced learners' vocabulary size and receptive collocation knowledge. The difference in receptive collocation knowledge between higher and lower proficiency learners is argued to stem from a dominating conceptual processing mediation of L2 forms through L1 forms for the lower proficiency learners, coupled with less exposure to the target language. The results also suggest that 4-6 months of full-time university-level studies are not enough for a measurable increase in receptive collocation knowledge to emerge. There is furthermore evidence to suggest that there is a progression in receptive collocation knowledge concomitant of learning level, overall language proficiency, and vocabulary size. This arguably favours a great deal of language exposure as an important factor for implicit acquisition of collocations, in addition to explicit instruction. COLLEX and COLLMATCH are quick to administer, hold appeal with test-takers, and so long as their limitations are noted they may be used as tests of receptive collocation knowledge, both as proficiency tests and as research tools.

Acknowledgments

There are a number of people to whom I owe enormous thanks. First and foremost, I would like to thank my supervisor Marie Källkvist, who has given me sustained support and guidance throughout my thesis work. You knew when to nudge me and when to leave me alone. Thank you for believing in me, and for letting me pursue the direction which this thesis is a result of. I would also like to thank Beatrice Warren for supervising me during the early and mid stages of the project. Your thinking has inspired me greatly.

A very special and heartfelt thank you goes to Paul Meara, who took such good care of me during my extended stays in Swansea, and who acted as an informal co-supervisor during certain phases of the project. Your support, unprecedented generosity, and mix of academic rigour and sense of humour make you a very special person, to whom I owe a lot.

I would like to thank my doctoral student colleagues in English Studies at the Centre for Languages and Literature in Lund, and the members of the English linguistics seminar. In addition, I would especially like to thank Birgitta Berglund, for making it possible to use my tests as part of regular end-of-term vocabulary exams; Lennart Nyberg, Carita Paradis, and Lars Hermerén, for showing interest in my work; Eva Kirsebom, for keeping a check on all of us, and Julienne Sandgren-Stewart, for always being cheerful and supportive.

The following people have helped me in various ways during my thesis work, or in other ways been generally supportive and friendly, and for this I am very grateful: Andy Barfield, Gunnar Bjärnlid, Frank Boers, Christopher Butler, Karin Ekbom, June Eyckmans, Tess Fitzpatrick, Geoff Hall, Birgit Henriksen, Jan Hulstijn, Nuria Lorenzo-Dus, Horst Löfgren, Iain McGee, Jim Milton, Valéria Molnár, Pierre Palm, Robert Lee Revier, Lars Stenius Stæhr, Cornelia Tschichold, Joost van de Weijer, Brent Wolter, and Alison Wray. My upper-secondary school teacher Ove Jonasson deserves a special mention for his inspirational role in making me study English in the first place.

In general, I would like to thank: all the people at CALS in Swansea; the members and co-lurkers of the VARG network around the world; the members of FLaRN; all the informants who more or less voluntarily took my tests. A big thank you goes to Louise Callmer at Latinskolan in Malmö, for helping me out with access to upper-secondary school informants.

The following foundations have generously financed my studies abroad, and participation in conferences: STINT, the Swedish Foundation for International Cooperation in Research and Higher Education; Crafoordska Stiftelsen; Stiftelsen Hilma Borelius stipendiefond; Stiftelsen Lektor Ture Betzéns donation nr 1; Stiftelsen Syskonen Anna Cecilia och Otto Sigfrid Granmarks stipendiefond; Stiftelsen Fil dr Uno Otterstedts fond för främjande av vetenskaplig undervisning och forskning; Knut och Alice Wallenbergs stiftelse.

On a more personal note, I would like to thank all my friends in Sweden for much needed distractions from tiring thesis work; the Simons family in Austria for your hospitality, unflagging support and encouragement; my sister Ulrika and her family for being there, and last but certainly not least, Magdalena, for your love: “I hab’ di’ sooo lieb!” Finally, I would like to dedicate this thesis to my mother, Lena Gyllstad, and in loving memory of my father, Bertil Gyllstad.

TABLE OF CONTENTS

1 INTRODUCTION 1

1.1 FIELD OF RESEARCH 1
1.2 THESIS AIMS 2
1.3 MAIN RESEARCH QUESTIONS 3
1.4 THESIS OUTLINE 4

2 THEORETICAL BACKGROUND AND PREVIOUS WORK 6

2.1 INTRODUCTION 6
2.2 TRACING THE USE OF COLLOCATION IN THE LITERATURE 6
2.2.1 Introduction 6
2.2.2 The frequency-based tradition 7
2.2.3 The phraseological tradition 11
2.2.4 The best of two worlds? - Researchers combining frequency-based and phraseological approaches to collocation 15
2.3 CRITERIA RELEVANT TO THE OPERATIONALISATION OF 'COLLOCATION' IN THIS STUDY 17
2.3.1 The nature of collocation 20
2.3.2 The nature of elements 23
2.3.3 The number of elements 24
2.3.4 Grammatical relation and structure 24
2.3.5 Adjacency 26
2.3.6 Frequency of co-occurrence 26
2.3.7 Lexical Fossilisation 27
2.3.8 Semantic opaqueness 28
2.3.9 Uniqueness of meaning 31
2.3.10 Summary: the treatment of collocation in this thesis 32
2.4 COLLOCATION KNOWLEDGE – DEFINING THE CONSTRUCT 33
2.4.1 Fundamental considerations 33
2.4.2 Defining the knowledge construct theoretically 34
2.4.3 Towards an operational definition of the construct 42
2.4.4 Reviewing empirical studies of L2 collocation knowledge 50
2.5 TEST THEORY 62
2.5.1 Introduction 62
2.5.2 Construct 62
2.5.3 Reliability 63
2.5.4 Validity 67
2.5.5 The application of test theory in this thesis 69

3 OPERATIONALISING RECEPTIVE COLLOCATION KNOWLEDGE INTO TEST FORMATS: COLLEX 1 AND COLLEX 2 71

3.1 DEVELOPING AND PILOTING COLLEX 1 71
3.1.1 Introduction 71
3.1.2 Preliminary considerations 71
3.1.3 The COLLEX test format 73
3.1.4 Methods 76
3.1.5 Results 80
3.1.6 Discussion 85
3.1.7 Conclusion 89
3.2 DEVELOPING AND ADMINISTERING COLLEX 2 90
3.2.1 Introduction 90
3.2.2 Methods 90
3.2.3 Results 94
3.2.4 Discussion 98
3.2.5 Conclusions 102

4 INVESTIGATING THE RELIABILITY AND VALIDITY OF COLLEX 3 AND COLLEX 4, AND DEVELOPING THE COLLMATCH TEST FORMAT 103

4.1 INVESTIGATING THE VALIDITY OF COLLEX AND DEVELOPING COLLMATCH 103
4.1.1 Introduction 103

4.1.2 Background ... 103
4.1.3 Methods ... 105
4.1.4 Results ... 113
4.1.5 Discussion ... 123
4.1.6 Summary and conclusions ... 129
4.2 DEVELOPING A NEW COLLMATCH FORMAT, ADMINISTERING IT TOGETHER WITH COLLEX 4, AND INTRODUCING A MEASURE OF VOCABULARY SIZE ... 130
4.2.1 Introduction ... 130
4.2.2 Background ... 130
4.2.3 Methods ... 132
4.2.4 Results ... 136
4.2.5 Discussion ... 147
4.2.6 Summary and conclusions ... 151

5 ATTEMPTS AT COMING TO GRIPS WITH CEILING EFFECTS AND TEST GENERALISABILITY ... 153

5.1 DISCUSSING WEAKNESSES OF PREVIOUS VERSIONS AND PILOTING NEW COLLEX AND COLLMATCH VERSIONS ON SWEDISH TEACHER STUDENTS AT UNIVERSITY LEVEL ... 153
5.1.1 Introduction ... 153
5.1.2 Previous versions of COLLEX and COLLMATCH – merits, problems and possible remedies 154
5.1.3 Piloting new versions of COLLEX and COLLMATCH ... 163
5.2 ADMINISTERING COLLEX 5, COLLMATCH 3, AND VLT M VERSIONS TO ADVANCED SWEDISH STUDENTS AND ENGLISH NATIVE SPEAKERS ... 178
5.2.1 Introduction ... 178
5.2.2 Methods ... 178
5.2.3 Results ... 180
5.2.4 Discussion ... 193
5.2.5 Summary and conclusions ... 199

6 VALIDATING COLLEX 5 AND COLLMATCH 3 AGAINST OTHER VOCABULARY AND PROFICIENCY TESTS ... 201

6.1 INTRODUCTION ... 201
6.2 METHODS ... 202
6.2.1 Considerations for the study design ... 202
6.2.2 Material ... 204
6.2.3 Informants ... 206
6.2.4 Research questions ... 206
6.2.5 Test administration and scoring ... 207
6.3 RESULTS ... 208
6.4 DISCUSSION ... 211
6.4.1 Are COLLEX and COLLMATCH scores more closely related to results on a vocabulary size test or a vocabulary depth test? ... 211
6.4.2 What is the relation between reading comprehension and each of the following variables: vocabulary size; vocabulary depth; collocation (COLLEX); collocation (COLLMATCH)? ... 218
6.5 SUMMARY AND CONCLUSIONS ... 220

7 DISCUSSION ... 221

7.1 INTRODUCTION ... 221
7.2 SUMMARIZING THE MAIN FINDINGS FROM THE EMPIRICAL STUDIES ... 221
7.3 DISCUSSION OF MAIN FINDINGS ... 225
7.3.1 Introductory remarks ... 225
7.3.2 Research question 1 ... 225
7.3.3 Research question 2 ... 239
7.3.4 Research question 3 ... 242

8 CONCLUSIONS, IMPLICATIONS, AND SUGGESTIONS FOR FURTHER RESEARCH ... 251

8.1 INTRODUCTION ... 251
8.2 MAIN FINDINGS AND CONCLUSIONS ... 251
8.3 LIMITATIONS ... 252
8.4 IMPLICATIONS ... 252

8.4.1 Testing 252
8.4.2 Learning and teaching collocations in a foreign language 253
8.5 SUGGESTIONS FOR FURTHER RESEARCH 254
REFERENCES 256
APPENDICES 3A-5N 266

1 Introduction

1.1 Field of research

Vocabulary and grammar are both indispensable aspects of knowledge that second language (L2)[1] learners need to acquire. The importance of vocabulary in communication cannot be underestimated, as emphatically pointed out by Wilkins (1972:111): "Without grammar very little can be conveyed, without vocabulary nothing can be conveyed". It is perhaps insights like these which have led to an upsurge of interest in L2 vocabulary over the last two decades, after having been a somewhat "neglected" aspect in linguistic research (Meara 1980).

The primary concern of L2 vocabulary has largely been single, orthographic words, and Moon (1997) argues that it is natural to focus on the word as the primary unit when discussing vocabulary knowledge, and that dictionaries help to reinforce this focus. It is beyond doubt that knowing many words is an advantage for all language learners. However, certain learner categories need to attain native-like command of an L2. Consequently, especially for advanced learners, e.g. university-level students, teacher students, translators and other professionals, possessing a large vocabulary *per se* is simply not enough. This has been pointed out by Wray (2002:143):

> To know a language you must know not only its individual words, but also how they fit together.

Thus, in order to be able to communicate effectively, in addition to knowing many words and their more frequent, core meanings, learners must also acquire knowledge about the combinatory potential of those words in relation to other words in the language. Again, in the words of Moon: "Text studies and corpus studies have revealed the significance and intricacy of the links between words […] their strong clustering tendencies and the patterns that are associated with them" (1997:40). A problem here is that grammatical rules alone do not predict why certain patterns and combinations of words are preferred to others in a specific language (Pawley & Syder 1983). Furthermore, if vocabulary is predominately learnt and taught as single words, this potentially leads to lexical incompetence on the part of the L2 learners (Farghal & Obiedat 1995).

The purpose of the project reported in this thesis is to construct tests that measure Swedish learners' knowledge about the combinatory potential of words in the English language. More specifically, the type of word combination that is targeted is 'collocation'. A definition of 'collocation' suitable for the purposes of this thesis is presented in Chapter 2. Here it suffices to say that 'collocations' are seen as conventionalized, recurring word combinations. The following English sequences may serve as examples: *say a prayer, draw a conclusion, make a mistake, do justice,* and *lose count*. Certain combinations of words are simply preferred to others in a specific language, and restrictions apply that do not follow from the grammar system of the language. Interestingly, the following plausible word combinations are

[1] In the thesis, the term 'second language' (L2) will be used interchangeably with 'foreign language' (FL) to denote a language that a person acquires after the native tongue. I will predominately use the term 'L2' since it is frequently used in applied vocabulary research.

unidiomatic if used with same intended meaning as those given above: **tell a prayer*, **pull a conclusion*, **do a mistake*, **make justice* and **drop count*.

The fact that collocations like those above pose problems to L2 learners is well-attested (see e.g. Channel 1981; Linnarud 1986; Biskup 1992; Bahns & Eldaw 1993; Farghal & Obiedat 1995; Howarth 1996; Granger 1998; Schmitt 1999; Gitsaki 1999; Källkvist 1999; Bonk 2001; Mochizuki 2002; Barfield 2003; Nesselhauf 2005 and Barfield 2006). Even though we know that collocations are challenging to L2 learners, and that collocational knowledge is seen as something that normally distinguishes between L1 and L2 speakers of a language (Schmitt 2000), there is a lack of reliable and properly validated test instruments with which learners' knowledge of collocations may be measured. The present thesis is an attempt to fill this void.

1.2 Thesis aims

This thesis has two main aims. The first aim is to construct, use, and evaluate the effectiveness of tests of receptive knowledge of English collocations as a single construct. For this purpose, two tests, called COLLEX (collocating lexis) and COLLMATCH (collocate matching), were developed. The second aim, which hinges on the first, is to investigate the performance of advanced Swedish learners of English, at different learning levels[2] in the Swedish education system, in terms of their receptive knowledge of English collocations, and in relation to their performance on vocabulary size tests.

Through a series of experiments and test administrations, the behaviour of COLLEX and COLLMATCH will be scrutinized in the pursuit of acceptable levels of validity and reliability[3]. A substantial part of the thesis will be devoted to empirically evaluating the tests and the scores they yield in the light of Classical Test Theory (CTT). The tests are investigated with respect to item quality, focusing on the level of difficulty of the items as well as their power of discrimination between informants with different abilities. Further analyses address guessing behaviour and the way it affects the quality of the tests, informants' perception of the tests, how well the tests discriminate between native speakers of English and Swedish-speaking learners of English, and how scores on the tests relate to scores on other tests of collocation knowledge.

By targeting in my tests frequently occurring English collocations which in turn are combinations of high-frequency word elements, it will be possible to empirically show whether knowledge of these high-frequency, single orthographic words, is beneficial to discriminating between native-like collocations and infelicitous, unidiomatic combinations of these words.

An additional aim of this thesis, and an important motivation behind the creation of COLLEX and COLLMATCH, has to do with washback[4]. According to Bailey (1996: 259), "washback is generally defined as the influence of testing on teaching and learning". As was pointed out above, vocabulary knowledge has traditionally been seen as knowledge of single orthographic words, and also tested as such. It is only recently that a call for more focus on

[2] The expression 'learning level' is taken to reflect the overall progression in an education system, e.g. primary school > secondary school > upper-secondary school > university, and also the progression within a certain study phase, e.g. first-term university students > second-term university students > third-term university students.

[3] The terms 'validity' and 'reliability' will be explained and discussed in Chapter 2.

[4] In the literature, the terms 'washback' and 'backwash' are used interchangeably.

tests of multi-word items has been made in relation to L2 vocabulary (Read 2000). As will become apparent, COLLEX and COLLMATCH were used in exams at university level. If students are subjected to tests where also collocations are tested, then this is likely to raise their awareness of collocations and the problems they may pose. This may provide an incentive to consciously study collocations as well as lists of single words in preparations for exams.

1.3 Main research questions

The two main aims from above are operationalised into three primary research questions (RQs). The first research question (RQ1) relates to language testing:

RQ1 Is it possible to develop tests measuring receptive knowledge of English collocations as a single construct, capable of yielding reliable and valid scores, for use with advanced Swedish learners of English?

The first research question (RQ1) is primary to this study, and as will become clear, it serves as a prerequisite for questions 2 and 3. RQ1 consists of several elements that require a brief explanation. Firstly, it is widely agreed that collocational knowledge is a particular kind of lexical knowledge, and an important one to boot (Pawley & Syder 1983; McCarthy 1990; Lewis 1997; Melka 1997; Schmitt 2000; Nation 2001; Wray 2002). My hypothesis is therefore that it should be possible to measure it as a single knowledge construct[5]. This means that collocation knowledge is a separate skill which can be measured as a stand-alone trait, albeit potentially interdependent on other closely related lexical constructs. Nothing in the previous attempts at constructing test-like measures of collocation knowledge – notably Bonk (2001) and Barfield (2003, 2006) – seems to impose any restrictions in this regard.

In terms reliability, if we consider previous work in the field of L2 vocabulary testing, my hypothesis is that it should be possible to construct tests that yield reliable scores, since many successful attempts have been made (see e.g. Meara & Buxton1987; Vives Boix 1995; Read 1998; Schmitt *et al.* 2001). With regard to validity, although it is in theory possible to construct a valid test, validity is a more nuanced quality of a test. It is not uncommon for test experts to disagree as to the validity of a particular test (Alderson *et al.* 1995), and validation is a perpetual process. For this reason, it is more difficult to hypothesize about the feasibility of aiming for the creation of valid tests.

The second research question (RQ2) concerns aspects of learning:

RQ2: What is the relationship between Swedish L2 learners' vocabulary size and their receptive knowledge of collocations?

[5] The term 'construct' is primarily a psychological term, but is used extensively in language testing (see e.g. Chapelle 1998; Alderson *et al.* 1995, Bachman & Palmer 1996). According to Davies *et al.*, a construct is a trait that a test is intended to measure. More specifically, it is "an ability or set of abilities that will be reflected in test performance, and about which inferences can be made on the basis of test scores" (1999:31).

It has been suggested that learners with large vocabularies are more proficient in a wide range of language skills than learners with smaller vocabularies (Meara 1996). This makes it reasonable to assume that this is the case also for collocation knowledge. However, until empirical support is presented, assumptions like these must be treated with caution. It is not unlikely that a large vocabulary will have a positive effect on receptive knowledge of collocations, but is the relationship in that case linear, and will the relation be similar across groups of learners at different proficiency levels? Furthermore, is it possible to possess a large vocabulary without having a good command of collocations? The empirical work in this thesis is aimed at addressing these issues.

It is in comparisons with other variables, like vocabulary size, that the creation of reliable and valid test tools is particularly important. Very little can be said about a learner's knowledge until there is a tool capable of yielding scores that reflect that knowledge in a reliable and valid way. This is why RQ1 serves as a prerequisite for RQ2 and RQ3.

Research question 3 (RQ3) also relates to the learning of collocations:

RQ3: What is the relationship between the learning level of Swedish L2 learners' of English and their receptive knowledge of collocations?

With respect to most language skills, an increase is expected as a student progresses to a higher level in an education system. Thus, a university student of English is normally expected to outperform an upper-secondary school student in most language skills. However, when it comes to collocation knowledge, this has not been sufficiently investigated empirically. It is not self-evident that collocation knowledge develops this way. Schmitt (2000) has argued that collocational knowledge is relatively difficult to achieve, and Melka (1997) that knowledge of a word's frequent collocates, i.e. the other words with which it co-occurs, implies a "higher" degree of familiarity with that word. These suggestions could be taken to mean that only very advanced students have developed a stable and high level of knowledge. This could in turn mean that no or small differences are present between students at different learning levels below the most advanced levels. There could be learning plateaux, where no tangible development can be observed, and conversely learning 'spurts' where students' proficiency is enhanced rapidly over a short period of time. These potential scenarios make research question three (RQ3) interesting and warranted.

In sum, RQ1 addresses a more practical and concrete process, namely that of constructing and evaluating tests, whereas the issues addressed in RQ2 and RQ3 have more theoretical ramifications in advancing our understanding of how collocational knowledge may develop, and in what way it is related to other variables, such as vocabulary size.

1.4 Thesis outline

The next chapter (Chapter 2) starts with a review of how the term 'collocation' has been treated in previous, relevant research. This review is meant to show the complexity of the term, its usage and definitions. The account of the different ways of approaching and defining collocation is also an important linchpin based on which I will subsequently operationalise the concept of collocation in this study. This will constitute a prerequisite for item selection for COLLEX and COLLMATCH. This chapter also reviews previous research targeting L2 collocation knowledge. Finally, necessary considerations in language testing are discussed.

Chapters 3-6 report on a series of seven empirical studies in which gradually refined versions of COLLEX and COLLMATCH are developed and investigated with a focus on aspects of reliability and validity, using data from Swedish students of English at upper-secondary and university levels, as well as native speakers of English studying at university level. COLLEX and COLLMATCH are also positioned in relation to standardized tests of vocabulary size, vocabulary depth and reading comprehension.

In Chapter 7 the findings of the series of studies are discussed and the aims and research questions are revisited. Finally, in Chapter 8, conclusions are drawn, implications are discussed, and suggestions for further research are made.

2 Theoretical background and previous work

2.1 Introduction

'Collocation' is far from being a well-defined term, and it has been investigated through many different approaches. In this chapter, I initially trace how the term 'collocation' has been used in the research literature, in particular within two dominating traditions. Key work within each of the two traditions is reviewed. As a second step, a number of criteria deemed relevant when defining 'collocation' for the purposes of testing are introduced and discussed in the light of the literature. Thirdly, the process of testing collocation knowledge will be addressed through a definition of 'collocation' as a knowledge construct, both theoretically and operationally. Following this I review the small number of empirical studies in which L2 collocation knowledge has been investigated, with the emphasis on studies using some sort of test tool. All these steps are needed to show the complexity of the field and the heterogeneity of collocation as a concept. Finally, basic notions within test theory will be explained. The primary purpose of this section is to explain fundamental considerations from Classical Test Theory (CTT). Anyone familiar with language testing and CTT can skip this section.

2.2 Tracing the use of collocation in the literature

2.2.1 Introduction

It is not an exaggeration to say that the ways in which collocation has been defined in the literature are quite diverse (see e.g. Fontenelle 1998:191; Stubbs 2004:107). Different scholars have tackled the concept in many different ways. Nesselhauf (2004) attribute the divergent use of the term 'collocation' to the fact that it has been used by researchers working in many different fields, and that the aims and methods of their investigations have governed the various definitions given.

The word 'collocation' itself can be traced as far back as the 17th century, when it was used by Francis Bacon in his *Natural History* from 1627, but not as a linguistic term. Supposedly, the first time it was used as a linguistic term was more than a century later, in 1750, by Harris, who used it to refer to the linear constellation of words (Palmer 1933). It was not until the 1930s, however, that the term was used in a way that is reminiscent of the dominant present day use, when Palmer (1931:4) used it to denote "units of words that are more than single words". This denotation lies close to more recent uses, such as "a natural combination of words" (McCarthy & O'Dell 2005:4), and "the way words combine in a language to produce natural sounding speech and writing" (*Oxford Collocations Dictionary* 2002:vii).

It will be convenient to acknowledge the fact that collocation, despite its definitional heterogeneity has traditionally been approached from two different angles in the literature of the second half of the 20th century. In one of them, collocation is intrinsically connected to frequency and statistics, predominantly advocated by scholars working within the fields of Corpus Linguistics and Computational Linguistics. I will refer to this tradition as the frequency-based tradition. In the other, the view on collocation has been largely inspired by Russian phraseology, and is more tightly linked to the fields of Lexicography and Language Pedagogy. I will refer to this tradition as the phraseological tradition.

First, I will account for the frequency-based tradition. In the subsequent subsection, I will in turn review the work in the phraseological tradition. In addition, I will also discuss approaches to collocation which straddle the aforementioned two traditions. A guiding and delimiting principle when carrying out this review is the focus on work which is relevant to the subsequent operationalisation of collocation in this thesis, and the development of test tools.

2.2.2 The frequency-based tradition

2.2.2.1 Relevant work on 'collocation'

In this tradition collocation is approached from a frequency perspective. In general, collocations are seen as units consisting of co-occurring words at a certain distance from each other, and a distinction is often made between frequently and infrequently co-occurring words (Nesselhauf 2005). In the following review, I will concentrate on the work by Firth, Halliday, and Sinclair.

The frequency-based tradition and its proponents are sometimes referred to as Firthian and Firthians, owing to the pioneering work by Firth (1951, 1957, 1968). Firth was the scholar who made the term collocation more widely known linguistically. Firth essentially saw collocation as a means to get to a word's meaning. It was this view that made him majestically proclaim: "You shall know a word by the company it keeps!" (1957:179), thereby giving collocation a central position in the theories of word meaning. Firth's main contribution is his advancement of "collocation" as a technical term, accompanied by the application of a "test of collocability" (1951:194). Firth suggested that part of the meaning of a word could be established by collocation, and he saw collocation as an abstraction at the syntagmatic level, "not directly concerned with the conceptual or idea approach to the meaning of words" (1951:196). Firth seemingly envisioned several types of collocations, as can be seen in his uses of "habitual", "common", "general" and "usual" collocations as opposed to "more restricted technical", "unique", "personal" and "a-normal" ones (1951). He did not, however, state what separates these types from one another. Across Firth's work, it is not possible to find a clear and consistent definition of collocation. There is variation, for example, when it comes to how many words may make up a collocation (between 2 and 11 orthographic words, e.g. "tender love" and "Is all the world drowned in blood and sunk in cruelty" (1951:196)). Firth furthermore seems to assume that word forms are involved in collocation, not lexemes. Another interesting aspect is whether a word under study is part of the actual collocation or not. In a later article he sees a collocation as being "the mere word accompaniment, the <u>other</u> word material in which [words under study] are most commonly or most characteristically embedded" (1957:180, my underlining). Thus, according to Firth, the specific word studied does not belong to the entity called collocation.

A second researcher positioned in the frequency-based tradition is Halliday (1961, 1966). Together with Sinclair, Halliday took the collocation baton, as it were, from Firth, and they are therefore commonly referred to as "neo-Firthians" (Mitchell 1971:36). They developed Firth's ideas on collocation and, as we will see in the passages to follow, advanced the formalization around the concept. This formalization will prove highly relevant to the research carried out in the present thesis. As opposed to the writings of Firth, Halliday attempts to define collocation in more detail (1961:276):

> …the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur, at n removes (a distance of n lexical items) from an item x, the items a, b, c …

There are several parts to this definition attempt that are relevant to the present thesis, and which therefore deserve extensive attention.

Firstly, the use of the term 'lexical item' should be noted. A lexical item in Halliday's view "may be a morpheme, word, or group (at least)" (1961:274). The term 'group' can best be seen to correspond to 'phrase', but not consistently. Halliday generally sees lexical items to be lexemes including all their possible derivations. This is evident in statements like the following: "*Strong*, *strongly*, *strength* and *strengthened* can all be regarded for this present purpose as the same item; and *a strong argument*, *he argued strongly*, *the strength of his argument* and *his argument was strengthened* all as instances of one and the same syntactic relation." (1966:151). These relations are seen as discontinuous abstractions. This means that Halliday's view clearly contrasts with Firth's in that Halliday treats lexical items as the entities involved in collocation, not word forms, and in the fact that the word under study, or rather lexical item under study, is intrinsically part of the collocation per se.

Secondly, we may note in the definition above the attempt to deal with the proximity in which collocating items appear: "…a distance of n lexical items…". However, Halliday does not develop this thought further, though it is clear that the distance may range across sentence borders: "*I wasn't altogether convinced by his argument. He had some strong points but they could all be met*" (1966:151, my underlining). He further qualifies this by proposing that "…lexis seems to require the recognition merely of linear co-occurrence together with some measure of significant proximity, either as a scale or at least a cut-off point. It is this syntagmatic relation that is referred to as 'collocation'" (1966:152). From an evaluative perspective, then, Halliday does not give a specific delimitation for this proximity. As we shall see later in this subsection, though, this problem is dealt with by Sinclair.

Thirdly, he introduces collocation as a statistical concept by saying that it is quantifiable as a probability of co-occurrence. However, he seems to view co-occurrences of all probabilities as collocations: "Any given item […] enters into a range of collocations, the items with which it is collocated being ranged from more or less probable" (1961:276). In a later article, though, he claims that in a lexical analysis, account should be taken of the frequency of an item in a stated environment relative to its total frequency of occurrence. He even goes as far as to use the term "significantly different", and in a discussion using the lexical item *strong* he predicts that "…there will be environments such that *strong* occurs with a probability greater than chance." (1966:156). This clearly suggests that the analysis of collocation must be accompanied by a measure that can reveal if words and their collocates appear together by chance or not.

In his account of 'collocation', Halliday introduces the terms 'node', 'collocate' and 'span' to refer to the item under study, the co-occurring item, and the specified environment in which the node and the collocate may co-occur, respectively. In doing this, Halliday definitely explicates the concept of collocation to a point which Firth's sometimes rather indistinct style of writing could not reach (cf. Robins 1961).

Sinclair (1966, 1970, 1987, 1991) takes the groundwork laid out by Firth and Halliday even further, at least in terms of operationalising them into a very comprehensive and text-driven research programme. One of Sinclair's main contributions in the work on collocation is

the attempt to solve some of the practical problems concomitant with a Firthian view of collocation. Sinclair took Firth's original ideas with him in the undertaking of the OSTI (Office of Scientific and Technical Information) project (see Krishnamurthy 2004), and later also the COBUILD project, one of the largest and most ambitious lexical research projects ever carried out (Carter 1998:167).

To Sinclair, Lexis as a field of study was focused on describing "the tendencies of items to collocate with each other" (1966:411). As with Halliday, Sinclair saw the lexical item as the entity under study within lexis, at least during the early stages of his research. Later on, he abandoned the notion of lexical item in favour of the word as the unit which enters into collocations (1987, 1991). Since a lexical item could not exclusively be associated with an orthographic word, but also other structures like morphemes and multiverbal items, this change made Sinclair's view more operationalisable. He also later changes the word "environment" to "text" (1991; Sinclair *et al.* 2004), and it seems feasible to assume that Sinclair generally treats collocation as a predominantly textual phenomenon.

Since Sinclair presents the characteristics of collocation more clearly than did Firth and Halliday, it makes sense here to take a closer look at some of these characteristics. Firstly, when it comes to how many words can make up a collocation, Sinclair is not totally consistent across his publications. In the OSTI report of 1970, which was officially published only in 2004, Sinclair and his co-workers still talk about "items" (Krishnamurthy 2004:10), and delimits the number to two. This is also done in an article from 1974 (Jones & Sinclair 1974:19). In more recent articles, though, he defines collocation as "…the occurrence of two or more words within a short space of each other in a text" (1991:170). He also stresses that collocation patterns are normally restricted to pairs of words, but that "there is no theoretical restriction to the number of words involved" (1991:170). The last quote highlights a second characteristic, having to do with the inclusion or not of the word under examination. For Sinclair, the word under examination, called the 'node', is part of the collocation per se. Consequently, this is a point where he differs from his master Firth, as we saw earlier in this section. Furthermore, in Sinclair's view, words that collocate do not have to be adjacent (1987:325). As to the distance that collocating words may be separated from one another, Jones & Sinclair (1974:21) propose that empirical evidence suggests that a span size of ± 4, i.e. 4 locations (number of orthographic words) to the left and to the right, respectively, of the node, constitutes the optimal environment within which 95% of that node's collocational influence occur. It was furthermore found that significant collocations were mostly found in span positions immediately next to the node, i.e. ± 1. The span was said to operate without any consideration taken of syntax, punctuation, and change of speaker. However, he later uses an example of the word *back* for suggesting that "few intuitively interesting collocations cross a punctuation mark." (1987:327).

Just like Halliday, Sinclair takes a statistical view of collocation, but basically considers all co-occurrences of words to be collocations. He makes a distinction, though, between "casual" and "significant" collocation, reminiscent of Firth's earlier division between e.g. 'habitual' and 'unique' collocations. He also outlines in more detail how the significant collocations could be singled out, by suggesting a formula for its calculation:

$$\frac{n * s * f}{p}$$

Figure 2.1 A formula for calculating the probability of an item occurring in a span, adapted from Sinclair (1966:418)

In the formula presented as Figure 2.1, *n* represents the number of times a particular node, the item or word under investigation, occurs in a delimited text; *s* stands for the span, i.e. the number of lexical items or words on each side of a node that is considered relevant to that node; *f* stands for the total number of occurrences of a particular item; and *p* stands for the total number of occurrences of items in a text. The resulting statistic is the probability of a collocate to appear within the span of a particular node. This, Sinclair suggests, may then be compared with the observed, actual number of times that the collocate occurs with the node, and statistical tests may be used to assess the significance of the discrepancy between the two values (1966:418).

It is not fair to talk about Sinclair's work without mentioning his modelling of how meaning arises from language text. This model is relevant since it has strong links to the concept of collocation. Sinclair proposes two principles of interpretation: 'the open-choice principle' and 'the idiom principle' (Sinclair 1991). The former envisages language text as the result of a very large number of complex choices. This view is, Sinclair claims, often called "a slot-and-filler" model. Texts are then seen as a number of slots that are filled from a lexicon. The slots are filled from the lexicon storage of words, if various local constraints are satisfied. The latter principle is an important complement to the open-choice principle. One of its stronger claims holds that "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (1991:110). This claim has more recently been elaborated in research about formulaicity and formulaic language (see e.g. Erman & Warren 2000; Wray 2002; Wiktorsson 2003; Schmitt 2004).

2.2.2.2 Summary of key aspects from the frequency-based tradition

In an attempt to summarize the key aspects from the review of the work carried out in the frequency-based tradition, we have seen suggestions that part of the meaning of a word could be established by collocation, and that several types of collocations exist, although not clearly defined. We have furthermore seen a definition of 'collocation' as the syntagmatic association of lexical items, where lexical items are lexemes including all their possible derivations. Technical terms like 'node', 'collocate' and 'span' have been proposed, and the proximity in which collocating items appear has been discussed. Here, empirical evidence suggested span sizes of ± 4 as the optimal environment for a node's collocational influence. Also, collocation as a statistical concept, quantifiable as a probability of co-occurrence, was introduced.

Other key aspects that emerged in the review was the proposal of Lexis as a field of study focusing on the description of tendencies of items to collocate with each other, and a

distinction between ‘casual’ and ‘significant’ collocations. A formula for the discrimination between these two types was presented. A model involving two principles: the ‘open-choice’ and the ‘idiom’ principle, was also suggested.

Time has now come to look at the other major tradition and its treatment of collocation.

2.2.3 The phraseological tradition

2.2.3.1 Relevant work on ‘collocation’

The treatment of collocation within the phraseological tradition can be seen to have been heavily influenced by work carried out first and foremost in Russia in the 1940s (Cowie 1998b, 1998c). Russian phraseologists like Vinogradov (1947) and Amosova (1963) postulated descriptive linguistic categories that later on have been elaborated on by British phraseologists. The point that unites researchers in the phraseological tradition is the treatment of collocation as a word combination, displaying various degrees of fixedness (Nesselhauf 2005). In the following review, I will concentrate on key aspects in the work of Cowie (1981, 1988, 1991, 1994, 1998a), Howarth (1996, 1998a, 1998b), Mel'čuk (1998), and Benson *et al.* (1997).

As opposed to most Russian phraseologists, who to a large extent have focussed their efforts on the description and classification of more fixed word combinations, Cowie (1981, 1988, 1991, 1994, 1998a), having a keen interest in language learners and their problems, is also interested in less fixed word combinations. Cowie basically classifies word combinations into two major types: ‘formulae’ and ‘composites’ (1988), where the former are units of sentence-length which normally have pragmatic functions, whereas the latter are units from below the sentence level. Collocations, according to Cowie, are part of the composite type, and as such units "which permit the substitutability of items for at least one of its constituent elements (the sense of the other element, or elements, remaining constant)” (1981:224). He exemplifies this through *run a business* in which *a business* may be substituted by *a theatre* or *a bus company*.

Cowie sees collocations as associations of two or more lexemes (or roots) occurring in a specific range of grammatical constructions. The last part of the definition is a clear example of how the phraseological approach differs from the frequency-based (Neo-Firthian) approach. In the latter, any two words can form a collocation, irrespective of word class and syntactic relation. What is also interesting is that Cowie talks about collocations as “abstract composite[s]” (1994:3169) which can be realized in patterns, e.g. *heavy rain* and *rain heavily*. Thus, it seems as if Cowie sees collocations both as abstractions and as some sort of instantiations, or “patterns” as he words it. This is in fact reminiscent of Halliday’s view, where *a strong argument*, *he argued strongly*, and *his argument was strengthened* (1966:151) were argued to be instances of one and the same syntactic relation.

Some of the interesting features of Cowie’s view on collocations are that they are transparent and in most cases lexically variable, but that they are characterized by arbitrary limitations of choice at one or more points. Cowie exemplifies with combinations like *cut one’s throat, slash one’s wrist*, **slash one’s throat*, and ?*cut one’s wrist*. He also proposes a sub-class, which he calls ‘restricted collocation’. The term itself is believed to stem from Aisenstadt (1979), and is defined as “word-combinations in which one element (usually the verb) [has] a technical sense, or a long-established figurative sense which [has] lost most of its analogical force” (Cowie 1991:102). This is in turn based on Vinogradov’s and Amosova’s

classifications of phraseologically bound units. Cowie gives the following examples of restricted collocations:

(1) run a deficit
(2) abandon a principle
(3) deliver an address

In an attempt to define the term restricted collocation even further, Cowie discusses its salient characteristics. He notes, firstly, that in the case of transitive verb + object noun combinations, the verb has special semantic properties. Either it is of the delexical type: *have, take, put, give*, or it has a long-established figurative sense, as in *reach an agreement, enjoy support* and *champion causes*. Secondly, he proposes substitutability to be a criterion. In this respect he argues that from the standpoint of the noun, whereas sometimes only one verb may be used in the required sense, in other cases a small set of more or less related verbs are possible. For example, in the case of the authentic newspaper text sample *he possessed a powerful antipathy towards income policy*, Cowie notes that the noun *antipathy* limits the number of synonymous verbs considerably. He suggests that only *have* and *feel* are possible in the same sense. The reverse perspective is also possible. From the viewpoint of the verb, several or only one object noun may be possible with a retained sense. As we will see in the accounts of the work of Howarth and Mel′čuk, the aspect of substitutability (or commutability) is a very important one in the phraseological approach.

On the whole, Cowie argues for a scalar analysis of word combination categories. The proposed scale ranges from "transparent, freely recombinable collocations at one end to formally invariable, unmotivated idioms at the other" (1994:3168). In fact, four different types of referential word combinations are suggested: free combinations (*drink one's tea*), restricted collocations (*jog someone's memory*), figurative idioms (*close ranks*), and pure idioms (*spill the beans*). Cowie stresses the fact that it is sometimes difficult to draw a line between the four categories, and some collocations are said to lie close to idiom-like combinations. Especially, it is argued, in collocations with delexical verbs (e.g. *bring, have, make, take*), for example *take (good) care of*, a part-for-part substitution is impossible and the combination displays a high degree of frozenness.

A second important figure in the phraseological tradition is Howarth (1996, 1998a, 1998b). Howarth's work lies close to that of Cowie, in that he follows the Russian phraseological tradition in postulating a model that separates idioms from collocations from free combinations. In this regard, his work is based on Arnold (1986), Cowie (1988), and Gläser (1988). Howarth acknowledges the value of investigating language use through corpora, referring to work in the Firthian vein, but states that frequency-based approaches alone do not suffice: "…phraseological significance means something more than what any computer algorithm can reveal" (1998:27). As his starting point, following Cowie's notion of "composite units", he draws a further distinction between "grammatical composites" and "lexical composites". This distinction depends on the word class of the constituent words. For lexical composites, the constituent words are nouns, verbs, adjectives and adverbs in different combinations. For grammatical composites, combinations such as preposition + noun, and adjective + preposition are included. Howarth here largely follows Benson (1985) who made a similar distinction between grammatical collocations and lexical collocations. It should be

noted that this division is also comparable to Firth's classification of collocation and colligation.

Howarth's category of lexical composites is divisible into two coarse categories: non-idiomatic and idiomatic. This two-way classification is, however, in fact a continuum. According to Howarth, by applying such criteria as restricted collocability, semantic specialization, and idiomaticity, four groups can be discerned. The continuum is shown in Table 2.1.

Table 2.1 A collocational continuum, after Howarth (1996:47, 1998a:28).

Category	free collocations	restricted collocations	figurative idioms	pure idioms
Definition	Combinations of two o more words in which the elements are used i their literal sense. Each component may be substituted without affecting the meaning c the other	Combinations in which one component is used in its literal meaning, while the other is used in a specialised sense. The specialised meaning of one elemen can be figurative, delexical or in some way technical and is an important determinant of limited collocability at the other. These combinations are, however, fully motivated	Combinations which have figurative meanings in terms of the whole. They may permit arbitrary synonymous substitution of one or more elements. They have current literal interpretation and are clearly motivated.	Combinations that have a unitary meanin that cannot be derived from the meanings of the components. They permit almost no substitution, and are unmotivated.
Example	*blow a trumpet*	*blow a fuse*	*blow your own trumpet*	*blow the gaff*

Howarth stresses the fact that a model like the one suggested holds an inherent characteristic: fuzzy boundaries. There are items which are considered to be more central members of a category and those that lie between.

An important aspect of Howarth's work is his preoccupation with the less central role that "linguists and teachers" have given collocations compared to free combinations and idioms (1998:42). He proposes more work to be carried out analysing learners' potential problems in the middle ground, that of restricted collocations. In his published doctoral thesis from 1996, Howarth claims that collocations present a particular challenge for linguistic description because of three main features. Firstly, one element in a collocation generally has greater freedom of co-occurrence than the other in a given sense. Secondly, the relationship between elements in a collocation is mostly unidirectional, not bidirectional. Thirdly, a collocation can be seen to have internal grammatical structure that contributes to its meaning as a whole. These three features can be exemplified in a collocation like *adopt a policy*. The sense of the verb *adopt* in the above collocation can be seen to be limited to a finite group of semantically related nouns, such as *measure*, *scheme*, and *approach*. The noun *policy*, on the other hand, possesses a much larger range of combinatory verb partners, e.g. *discuss*, *present*, *vote on*, which furthermore may display a higher degree of semantic heterogeneity. In terms of directionality, the figurative sense of *adopt* is created by its co-occurrence with *policy*. Lastly,

the collocation *adopt a policy* is analysable as a syntactic structure consisting of a transitive verb followed by a direct object.

A third researcher in the phraseological tradition who deserves attention is Mel′čuk (1998). Mel′čuk's phraseological framework is just like those of Cowie and Howarth heavily inspired by the Russian lexicology tradition. His treatment of collocation is part of a theory called Meaning-Text Theory (Mel′čuk, 1998), and his aims are said to be both theoretical and practical, where the practical aim should be read as lexicographic description. On the whole, Mel′čuk's system represents a highly formalized and very ambitious undertaking in the typology of collocations. The main field of application of the system are so-called Explanatory Combinatorial Dictionaries, which are lexical databases containing semantic representations of set phrases. My account of this system will be based on Mel′čuk (1998).

True to the Russian legacy, Mel′čuk draws up a system where collocations are part of a larger class for which the cover terms 'set phrases' or 'phrasemes' are used. These phrasemes are in turn divided into 'pragmatic phrasemes' and 'semantic phrasemes'. The former correspond to Cowie's 'formulae' and the latter to his class called 'composites'. The extension of pragmatic phrasemes is so-called 'pragmatemes'. This group consists of ready-made expressions like greetings, proverbs, and sayings. The further subdivision of the semantic phrasemes gives us 'Idioms', 'Collocations', and 'Quasi-idioms'. In less formalised language, Mel′čuk sees collocations as combinations consisting of two elements. One of these elements is chosen based on its meaning, whereas the other element is chosen contingent on the other element. This means that one element is free and the other one is not. Mel′čuk 's (1998:30) formal definition of the group called collocations is as follows:

> A COLLOCATION AB of language L is a semantic phraseme of L such that its signified 'X' is constructed out of the signified of one of its two constituent lexemes—say, of A—and a signified 'C' ['X' = 'A + C'] such that the lexeme B expresses 'C' only contingent on A.

The formulation "B expresses 'C' only contingent on A" covers four different subtypes of collocation:

a) Collocations containing a delexical (or 'support', 'light') verb (e.g. *give a look, launch an appeal*);
b) Collocations containing a dependent lexeme meaning which only occurs with one or a few lexemes (e.g. *black coffee, French window*);
c) Collocations containing a dependent lexeme meaning (intensifiers) that can be used together with other lexemes in the same sense, but its meaning cannot be expressed by a possible synonym (e.g. *strong coffee*);
d) Collocations in which one lexeme is dependent on the other lexeme because the meaning of the latter is utterly specific (e.g. *the horse neighs, rancid butter*).

A central part in Mel′čuk's system is played by so-called Lexical Functions (LF). A lexical function is a general and abstract meaning. This general meaning is coupled with a deep syntactic role, which can be expressed by various lexemes. In a LF, a so-called 'keyword'

selects another element, called 'value'. In *lend support*, *support* is the keyword, and *lend* is the value. Mel′čuk states that around 60 so-called "Simple Standard LFs have been recognized so far in natural languages" (1998:32). Given examples of LFs are 'Magn', which means "intensely" and "very" and is an intensifier (*stark naked*). Another LF is 'Oper', which is normally a support verb with the meaning "do" or "perform" (*lend support*). An interesting notion related to support verbs is that Mel′čuk calls them "semi-auxiliaries" (1998:37). This is because they are said to play important semantic-syntactic roles. The LF '$Oper_i$' (short for Lat. *Operari* 'to do, carry out'), together with '$Func_i$' (short for Lat. *functionare* 'to function'), and '$Labor_{ij}$' (short for Lat. *laborare* 'to work, toil') are all support verbs which are considered semantically empty in relation to the keyword lexical unit (LU). The LU is by necessity a noun which corresponds to the name of an action, an activity, a state, a property, a relation, etc.

Finally, Benson *et al.* (1997) is a dictionary which identifies collocations as phrases which are "fixed, identifiable, non-idiomatic…" (p. xv). Even though they call their dictionary "a dictionary of English word combinations", their main object of investigation is collocation. They distinguish between grammatical and lexical collocation. In the former, a dominant word (noun, adjective, or verb) is combined with a preposition or grammatical structure, such as a clause. Benson *et al.* identify eight major types of grammatical collocations, with each consisting of a varying number of subtypes. For example, type G8 comprises no less than 19 different English verb patterns. The lexical collocation is a word combination consisting of nouns, adjectives, verbs, and adverbs, but no function words. Also for each of these seven main categories, a number of different structures are postulated. This places Benson *et al.* (1997) into the phraseological tradition.

2.2.3.2 Summary of key aspects from the phraseological tradition

A summary of the key aspects which have emerged in the above review shows that collocation in the phraseological tradition is some kind of word combination, displaying various degrees of fixedness. In many cases, elaborate classification systems have been drawn up to discriminate between the range of transparent, freely recombinable collocations and formally invariable, unmotivated idioms, sometimes in the form of a continuum, where criteria such as restricted collocability, semantic specialization, and idiomaticity are used. We have also seen that researchers within the phraseological tradition dismiss frequency as the only important criterion for the identification of collocations.

Definition attempts suggest collocations to be associations of two or more lexemes occurring in a specific range of grammatical constructions. A subclass, called 'restricted collocation', has also been proposed, which entails word-combinations in which one element evokes a delexical, technical, or figurative sense. Finally, we have seen that so-called support verbs are considered semantically empty in relation to a 'keyword' noun, and that the keyword selects the verb element, called 'value'.

2.2.4 The best of two worlds? - Researchers combining frequency-based and phraseological approaches to collocation

In addition to the frequency-based and phraseological traditions, there are a number of researchers that are not as easily labelled, or who apply criteria which can be found in both the frequency-based and the phraseological camps. There are also researchers who do not use the term collocation, but who clearly refer to structures that are widely seen as collocations. In

this section, I will briefly account for a number of such views, notably Mitchell (1966, 1971), Greenbaum (1970, 1974), Kjellmer (1984, 1987, 1991, 1994), Stubbs (1995), Altenberg (1993, 1998), and Nesselhauf (2003, 2005).

Mitchell (1966, 1971), although working in the Firthian tradition, differs from researchers like Halliday and Sinclair in that he does not acknowledge a separation between lexis and grammar (Mitchell 1971). Instead, he argues that the study of collocation must incorporate both grammar and semantics. He sees collocations as consisting of roots, which is an abstraction based on inflectional and derivational forms of a word. This abstraction can furthermore be realized in various syntactic patterns. For example, the collocation *heavy damage* can be realized as *heavy damage* (adjective + noun), *to damage heavily* (verb + adverb), *heavily damaged* (adverb + passive participle) (1966:337). His dependence on frequency is captured by the fact that he uses "habitualness" as a criterion (1971:54).

Greenbaum (1970, 1974) also argues for the necessity of taking syntactic relationships into account when analysing collocations. Just like Mitchell, he furthermore sees frequency as a factor of interest: "we may also wish to take account of the tendency for certain collocations rather than others to be likely in a language" (1970:1). However, Greenbaum criticizes The Sinclarian item-oriented approach which is seen as divorcing collocations from syntactic considerations. Also, he is critical of the fact that item-oriented approaches do not stipulate the maximum distance between collocating items. He calls his own approach an integrated approach and uses a number of elicitation tests to investigate collocational behaviour of certain intensifiers (e.g. *certainly, really, badly, greatly, entirely*).

Kjellmer's (1984, 1987, 1991, 1994) view on collocation is clearly frequency-based, but he presupposes certain syntactic structures in his analyses. His work on collocations has a relatively applied basis in being linked to the production of a collocation dictionary (1994), based on the one-million-word Brown Corpus. In Kjellmer's words, a collocation is "a sequence of words that occurs more than once in identical form [in a text corpus] and which is grammatically well-structured" (1987:133). Examples of retrieved collocations based on this definition are *to be, had been, one of* and *United States*. In his collocation dictionary from 1994, we are told that only adjacent items are regarded as collocations (1994:xiv). For Kjellmer, a word corresponds to an orthographic word. Furthermore, the elements called 'words' are in fact word forms, not lemmas (1994:xix). The 19 grammatical patterns which Kjellmer acknowledges, a classificatory scheme based on work in Swedish by Allén (1975), include for example noun phrase (*the big question*), adverb or preposition plus preposition (*out from, away to*), nominal head plus related structure word (*job as, question whether*), and co-ordinated elements (*openly and honestly, quiet but impressive*) (1994:xxii-xxix).

An interesting inclusion in Kjellmer's class of collocations is idioms. He defines the latter as "a collocation whose meaning cannot be deduced from the combined meanings of its constituents" (1994:xxxiii). This inclusion differs from other researchers working in a purely phraseological tradition. Kjellmer argues that the borderland between idioms and other collocations is a very fuzzy area, and is content with saying that an attempt to separate idioms from other collocations "would create many difficulties [in a work devoted to English collocations in general] and serve no useful purpose" (1994:xxxiv). Another point that separates Kjellmer from mainstream phraseologists is that he accepts *drink water* as a collocation, a sequence that would fall under the heading 'open combination' or 'free combination' in most phraseologically based approaches because of its unrestrictedness.

Stubbs (1995) sees collocation as a relationship of habitual co-occurrence of words, either lemmas or word-forms. This view positions Stubbs in the frequency-based tradition. However, he does in practise use grammatical relations as an identification criterion for collocates of a node word: "Collocates which occur as subject or object of the verb CAUSE or as prepositional object of the noun Cause…" (1995:27).

Altenberg (see e.g. 1993) seldom specifically uses the term collocation in his research. Instead, he uses terms like "recurrent verb-complement constructions" (1993:227), and "recurrent word-combination" (1998:101)[6]. By the latter, he means "any continuous string of words occurring more than once in identical form", in a corpus (p. 101). In his analysis, starting out with computerised searches in a corpus, he subsequently subdivides the material into grammatical structures. This view clearly positions him both in the frequency-based and the phraseological tradition, with a slight emphasis on the former.

Nesselhauf (2003, 2005) is primarily working in the phraseological tradition, but she uses frequency as a complementary method in analysing learner corpora. She sees collocation as "arbitrarily restricted lexeme combinations" (2005:1) and draws on work by Howarth (1996) in analysing verb + noun combinations in corpora. She proposes three categories of word combinations: free combinations, restricted collocations, and idioms, and she uses degree of restrictedness in either of the word elements to distinguish between the three categories.

2.3 Criteria relevant to the operationalisation of 'collocation' in this study

It is not an exaggeration to say that the approaches to collocation reviewed above form a rather motley crew. Although two major, influential traditions were discerned – a more frequency-based approach and a more phraseological approach – we have sometimes seen a fair degree of overlap between the two, and we have also seen researchers who more eclectically use criteria from both traditions. In sum, no conventionalized and widely agreed definition was found. Consequently, it is clear that collocation is a complex concept.

For the purposes of this thesis, the term 'collocation' as it will be used in the remainder of this thesis will be defined. In doing this, I will use the key aspects that emerged in the review above in order to select criteria which will help me define the term collocation in such a way that it enables me to single out certain word sequences for inclusion in my tests. In this respect it is important to distinguish between 'collocation' as a linguistic concept per se, and a definition or operationalisation of 'collocation' as a knowledge construct, necessary for the selection of items for language tests. Both of these perspectives need to be addressed. For this purpose, drawing on work by notably Nation (2001), Nesselhauf (2004), and Siepmann (2005), I will use nine criteria based on which the concept and use of collocation will be investigated. These nine criteria are argued to capture both the nature of 'collocation' as a more linguistic concept, and aspects necessary for the operationalisation of the term in the present thesis. Before the nine criteria are presented, a short presentation and discussion of the three works referred to above are called for.

Nesselhauf (2004) identifies ten variables relevant to the way collocation has been used in the literature: (i) frequency of occurrence, (ii) transparency, (iii) variability, (iv) grammatical relationship, (v) the nature of the elements, (vi) the types of elements, (vii) the number of

6 Altenberg does mention the term collocation in his 1998 publication (p. 103), but it is not his object of investigation.

elements, (viii) consecutive or separated elements, (ix) the nature of the phenomenon itself, and (x) the equality of the relationship between elements. Nesselhauf shows that considerable variation exists across these ten variables, not least in relation to the two earlier mentioned approaches. Despite the multifariousness of existing definitions, she proposes that a common denominator prevails across these definitions: collocation is "some kind of syntagmatic relation of words" (2004:1).

Siepmann (2005) assumes three different approaches to collocation: "frequency-based, semantic, and pragmatic approaches" (p. 409). The first approach, rooted in a continental European research tradition, is seen to assume a particular meaning relationship between the constituents of a collocation. The second approach, rooted in the British tradition, is occupied with statistically significant co-occurrences of words, whereas the third approach can be seen to make recourse to contextualisation theory. Siepmann poses five questions related to the definition of collocation. The five questions used to review the three approaches concerns (a) how many elements make a collocation?, (b) what elements make a collocation?, (c) if collocations are arbitrary, (d) if a distinction can be made between collocations and phraseology, and between collocations and free combinations, and (e) whether collocations are monosemous and monoreferential.

Nation (2001) suggests what he calls ten "scales" to be used for the identification and classification of collocation, a term which he says refers to "a group of words that belong together" (p. 317). He furthermore adds that collocation, from a learning perspective, should be seen as "items which frequently occur together and have some degree of unpredictability" (p. 317). Nation's ten scales are best presented in a table. For this purpose, consider Table 2.2 below. The ranges of the ten scales, shown in the right column in the table, are graded from most lexicalised to least lexicalised, sometimes with a midpoint added within parentheses. As can be seen from the table, there is considerable overlap between Nation's scales and the criteria and questions proposed by Nesselhauf and Siepmann. This is especially so between the scales proposed by Nation and the criteria proposed by Nesselhauf, and to a lesser extent between Siepmann's questions and the criteria and questions of the two former. For example, Nation's scales 1 and 2, concerning frequency of co-occurrence and adjacency, are directly related to Nesselhauf's criteria (i) and (viii). Furthermore, it is possible to collapse and incorporate Nation's scales 3-6, all having to do with grammatical aspects, into Nesselhauf's criterion (iv).

Table 2.2 Ten scales for classifying groups of words as collocations, proposed by Nation (2001:328ff)

Scale for classification	Scale range and description
1. Frequency of co-occurrence	frequently occurring together <-> infrequently occurring together
2. Adjacency	next to each other <-> separated by several items
3. Grammatically connected	grammatically connected <-> grammatically unconnected
4. Grammatically structured	well structured <-> loosely related
5. Grammatical uniqueness	grammatically unique <-> grammatically regular
6. Grammatical fossilisation	no grammatical variation <-> (inflectional change) <-> changes in par of speech
7. Collocational specialisation	always mutually co-occurring <-> (one bound item) <-> all occurring in a range of collocations
8. Lexical fossilisation	unchangeable <-> (allowing substitution in one part) <-> allowing substitution in all parts
9. Semantic opaqueness	Semantically opaque <-> semantically transparent
10. Uniqueness of meaning	Only one meaning <-> (related meanings) <-> several meanings

Nation's scale 3 concerns whether some kind of grammatical connection must apply, or if lexical cohesion can also make two words into collocates, for example, if the word *silk* and the word for a colour are collocates by virtue of appearing in the same text environment. Scale 4 focuses on the degree to which two items must be structured in order to enter into a collocational relationship. The question is whether structures like *although he* and *the very* are considered to be sufficiently structured to be passed for collocations. Scale 5 is related to scale 4 in that it deals with the extent to which a collocation is formed according to grammatical rules, or whether certain features of rules are violated. Scale 6 has to do with the extent to which a collocation may be manipulated through for example grammatical inflections, change of word order, or change of part of speech. Scale 7 from Nation's list refers to the mutual exclusiveness of collocating items. An example is *hocus pocus* where neither of the two parts normally appears without the other. This could be seen to relate roughly to Nesselhauf's criterion (iii) variability, as can Nation's scale 8. This scale has to do with the extent to which a word may be substituted by another word bearing a related meaning. Scale 9 covers the classic concept of compositionality (see e.g. Saeed 2003), whereby the meaning of a phrase is determined by the meaning of its components. A classic textbook example of a non-compositional phrase is *kick the bucket*, where the meaning of the whole (~die) cannot be deduced from the meaning of the words. It should be noted, though, that it is possible to interpret *kick the bucket* literally, in addition to the idiomatic meaning. Scale 9 is therefore in close correspondence with Nesselhauf's criterion (ii) transparency. Scale 10, finally, refers to the fact that a collocation may invoke one or several meaning interpretations. This was just exemplified through the polysemous word sequence *kick the bucket*. This scale is reflected in Siepmann's question (e), which addresses the question whether collocations are monosemous. Siepmann's questions (a) and (b) correspond to Nesselhauf's criteria (vii) and (v) and (vi), respectively.

The above comparison demonstrates the fact that many criteria overlap or are even the same across the three lists. At this point it must be made clear that even though all of these criteria may be relevant to a discussion of what a collocation is from a more general linguistic perspective, all criteria are not necessarily relevant for the purposes of the present project.

What is most important here is to select criteria which will help me define the term collocation in such a way that it enables me to single out certain word sequences for inclusion in my tests. For this reason, not all of the criteria from the lists above will be taken into consideration. By taking Nation's list of ten scales as my point of departure, and collapsing scales that are arguably close to each other into one criterion (3-6), leaving certain scales out (7), and adding three of Nesselhauf's criteria ((v), (vii), and (ix)), the following list of nine criteria, relevant to the present thesis, is created.

Table 2.3 Criteria relevant to operationally defining 'collocation' for the purposes of test item selection.

Criteria
1. The nature of collocation
2. The nature of the elements in a collocation
3. The number of elements in a collocation
4. Grammatical relation and structure in a collocation
5. Adjacency of elements in a collocation
6. Frequency of co-occurrence of elements in a collocation
7. Lexical Fossilisation
8. Semantic opaqueness
9. Uniqueness of meaning

The nine criteria in Table 2.3 will be discussed one by one below, and relevant research literature will be reviewed. At the end of the discussion of every criterion, the view adopted in the present thesis will be given. Subsequent to this discussion, a definition of collocation will be presented, as it will be used in the present thesis.

2.3.1 The nature of collocation

The first of our criteria concerns the nature of collocation. In the light of the research literature, it seems possible to view collocation in one of three ways, in terms of its nature: either as a textual phenomenon, i.e. physical instantiations in a text, or as some kind of abstraction, in terms of links between words in a language system, or as a combination of both of these. The clearly dominating view among researchers is one in which collocation is seen both as some kind of abstraction, and a more textual phenomenon. This view is adopted in, for example Nesselhauf (2005), Cowie (1998), and Howarth (1996).

The view taken in this thesis is that collocations are both textual instantiations and abstractions. A textual instantiation can be either a written text, or a spoken text which is in some way recorded and transcribed in writing. However, since texts are produced by language users, it seems reasonable to assume that any textual instantiations stem originally from associative connections between words present in these language users' minds. Hoey (2005) argues that the textual instantiation view is more of a method, and what is really important is the abstraction view, which tells us something interesting about the psycholinguistic aspects of collocation as a phenomenon. According to Hoey, "the first view gives no clues as to why collocation should exist in the first place" (2005:4). This is an important observation. What makes collocation interesting is the assumption that it stems from associative connections between words in the mental lexicons of language users.

It is beyond the scope of this thesis to fully account for the processes involved in the making of these associative connections, but it seems that these are reminiscent of something called 'chunks'. The term chunk was originally coined by Miller (1956), but no clear definition was given. More recently, the term is explained as "…a unit of memory organisation, formed by bringing together a set of already formed chunks in memory and welding them together into a larger unit" (Newell 1990:7). Ellis (1996:107) explains chunking as "the development of permanent sets of associative connections in long-term storage and […] the process that underlies the attainment of automaticity and fluency in language". In this view, and relevant to the present discussion, a word form can be seen as a unit of memory organisation, and a chunk can consist of several word forms that are associatively connected with one another. Even though it is difficult, if not impossible, to demonstrate the existence of these connections empirically (see, though, Wiktorsson 2003; Knutsson 2006), the conduct of word association studies is one attempt to find support for the psycholinguistic validity of these connections in the minds of speakers (see e.g. Meara 1982; Kruse *et al.* 1987). Native speakers are normally good at finding associations between words, whereas L2 learners often fall short in terms of this skill. Meara asserts that L2 learners fail to see "connections between words that are obvious to native speakers" (1996:48), while Partington (1998) claims that knowing what are normal collocations is part of a native speaker's communicative competence.

Interestingly, Hoey (2005) suggests that semantic priming[7] is the key factor behind the forming of collocations. In this view, collocation is only accountable for if we assume that every word is mentally primed for collocational use. This is taken to mean that words become loaded with contexts and co-texts in which they are encountered, and whenever a word is repeated in use, the load increases and is strengthened, as long as the same contexts and co-texts co-occur.

An important aspect to discuss is the relation between collocation and formulaic language. The question at hand is whether collocations are inherently formulaic or not. Bonk (2001:114-115) argues that there is a difference between collocation and formulaic speech, in that "collocation is best understood as connections between items in the mental lexicon based on lexical and semantic characteristics, and not as a chunked storage and production strategy *per se*, as formulaic speech may prove to be". In order to evaluate claims like these, however, we need to know what definition is adopted for formulaic speech as such. Bonk does not supply such a definition. In terms of the wider term formulaic language, a number of relevant definitions exist. Consider first Wray's definition of a "formulaic sequence" (2002:9):

> a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved as a whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.

Wray's definition, which she admits is inclusive as to the linguistic units that may be subsumed under the cover term, has points in common with Sinclair's (1991) distinction

[7] Priming is defined as "Mental activation of a concept by some means, or the spread of that activation from one concept to another; also, the activation of some target information by action of a previously presented prime; sometimes loosely synonymous with the notion of accessing information in memory" (Ashcraft 2006: 572).

between language produced according to the open choice principle and the idiom principle (see section 2.2.2 above). In Wray's definition, a sequence subject to generation or analysis by the language grammar lies very close to what Sinclair refers to the open choice principle, whereas the prefabricated notion in Wray's definition corresponds well with Sinclair's idiom principle. A seemingly related concept to that of formulaic language is provided by Erman & Warren:

> A prefab is a combination of at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization (2000:31).

Erman & warren argue that the conventionalisation of the word combination they choose to call 'prefab' suggests that it is memorized, even though they admit that no proof exists. Many different kinds of word combinations are included in their group of prefabs, notably idioms, compounds, habitual collocations, and prepositional and phrasal verbs. One criterion has to be met in order for collocations to be included: they have to be non-compositional. The primary criterion used for distinguishing prefabs from non-prefabs is "restricted exchangeability" (2000:32), by which is meant the restrictions on a least one member of the prefab in terms of its replacement by a synonymous item, without causing either a change of meaning, change of function, and/or idiomaticity. This is reminiscent of Howarth's term restricted collocability (see section 2.2.3). I will come back to this notion in 2.3.8 below.

Wray's use of formulaic sequence from above seems to be compatible with the terms associative connections and chunks discussed above, and Erman & Warren's notion of prefab also appears to be a kindred notion. However, Erman & Warren's definition of prefab does place restrictions on what combinations or words are included. Any sequence of words stored and retrieved as a whole, along the lines of Wray's definition, does not pass the requirement of being called a prefab, nor does any kind of associative connection between words. The keyword seems to be conventionalization. The conventionalization aspect of Erman & Warren (2000) is closely related to Pawley & Syder's (1983) notion of nativelike selection, whereby native speakers make use of thousands of lexical phrases which in turn make up only a small proportion of the total set of grammatical sentences available along the lines of the full creative potential condoned by the syntactic rules of a language. Nativelikeness appears to be that use of language forms which is readily acceptable to native informants as "ordinary" and "natural" (1983:193), and "idiomatic usages" (1983:196). A similar point is made by Stubbs:

> Corpus studies show that what typically occurs in language use is only a small percentage of what seems possible within the language system. A large amount of language use consists of words occurring in conventional combinations. Such collocations are not an idiosyncratic and peripheral phenomenon, but a central characteristic of language in use. Native speakers' unconscious knowledge of collocations is an essential component of their idiomatic and fluent language use and an important part of their communicative competence. (2001:73)

I will have more to say about conventionalization and nativelikeness later when I define the construct to be measured by my tests. However, the question addressed here is whether collocations are inherently formulaic. Wray (2002) suggests that collocations are formulaic sequences for native speakers, but they are essentially not so for non-native speakers. She argues, from the perspective of language acquisition, that native speakers start out with big units (formulaic sequences in the form of collocations) which they may or may not analyse into parts. Wray illustrates this with the sequence *major catastrophe*, which is stored as a sequence to be conventionally used when talking about a big disaster. If they do analyse this sequence, then this implies a process whereby the two words, *major* and *catastrophe*, become loosened or separated. For adult L2 learners, collocations are in contrast seen as separate items (words) which may become paired. Wray argues that learners, on encountering a sequence like *major catastrophe*, break it down into separate word meanings like 'big' and 'disaster', and claims that they store no information about the two words going together. When needing to express the notion of a big disaster after this occasion, they would have no memory of *major catastrophe* as the pairing of words originally encountered (2002:209). It is the process of pairing words up which causes difficulties for these learners, because there are simply too many options: *big*, *large*, *major*, *huge* etc.

It should be pointed out that the above account is hypothetical, and no empirical support is presented for these claims by Wray. There is therefore clearly a need for further research in the field of formulaic language and language learning. However, these processes are intuitively appealing as possible explanations for non-native speakers' problems with lexical restrictions. It is probable that some sort of continuum might have to be introduced, where learners of lower levels of proficiency are more relevant to the above descriptions, whereas learners of higher levels of proficiency may be observed to function more like native speakers.

Summing up the view taken in this thesis, about the nature of collocation, collocations are textual instantiations that stem originally from associative connections between words present in the minds of native speakers of a language.

2.3.2 The nature of elements

An intriguing question which has far-reaching consequences for our treatment of collocation, and test item selection, is what elements can be seen to collocate. The literature review above reveals no consensus across researchers in this regard. To Firth, the elements are word forms. To Halliday, they are so-called lexical items, which may consist of either morphemes, words or groups. A word group is normally interpreted as phrases. Furthermore, Kjellmer (1984) propagates that consideration should be taken of word forms, whereas Sinclair (1991) refers to lexemes. When it comes to the position taken in the present study, if we accept the assumption that text instantiations of frequently co-occurring words (higher than chance probability) serve as evidence of the existence of associational links between words in the minds of speakers, then it seems reasonable to see lemmas as the collocating elements. Thus, *brisk walk* and *brisk walks* are both instantiations of the collocation BRISK + WALK. Furthermore, textual instantiations like *says a prayer*, *said a prayer*, and *saying a prayer* are based on the abstraction SAY + PRAYER.

2.3.3 The number of elements

A question that must be addressed when defining 'collocation' is that of how many elements, i.e. orthographic words, it may consist of. Most researchers reviewed above take two items to make up a collocation, and many claim that two or more do. Siepmann (2005) argues that collocations are normally treated by researchers as binary units, and although trigrams exist, these are often reducible to binary structures (e.g. German *allgemeine Gültigkeit haben* (to be generally valid) -> (*allgemeine* + *Gültigkeit*) + *haben*). However, there are examples of phrasal verbs, such as *put out* in the combination *put out a fire*. Depending on its syntactic realization, the concept underlying this combination can be seen to consist of three elements, as in *putting out fire* and *put out fires*, or four elements, as in *put out a fire* or *put out the fire*. Normally, elements like determiners and prepositions are not included in analyses that treat combinations like *launch an appeal* as a binary unit. The article is a variable operating with a certain amount of optionality. It may for example be substituted by a determiner quantifier such as *several* followed by the plural noun *appeals*. The view adopted in this thesis is that collocations are essentially and predominantly binary structures, especially in the abstraction sense of the word: the psycholinguistic association in the minds of speakers. This means that the following examples of sequences, if found as instantiations in texts, are all assumed to boil down to the underlying schematic, binary sequence KEEP + SECRET: -> *keep secrets*, *keep a secret*, and *keep a mysterious secret*.

2.3.4 Grammatical relation and structure

For testing purposes, it is logical to restrict the number of patterns used in a test. If too many patterns are included, there is a risk that scores will be hard to interpret. Also, in order to be able to generalize from scores to some sort of underlying ability, a large number of items would be needed from each pattern. For these reasons, this thesis focuses on verb + NP and adjective + NP collocations, respectively. Siepmann (2005) argues that the verb + NP is the more common type, and Altenberg (1993) says that the verb and its complementation are of particular interest, since "they tend to form the communicative core of utterances where the most important information is placed" (p. 227).

We saw in the above review that researchers within the frequency-based research tradition vary when it comes to whether the collocating elements in a collocation must be part of some sort of grammatical relation or not. By grammatical relation is meant syntactic patterns such as verb + direct object and adjective + NP. Sometimes further elements are needed in these patterns to make the language structures well-formed and idiomatic, e.g. articles and pronouns. We saw that researchers like Firth (1951, 1957, 1968), Sinclair (1966) and Halliday (1966) did not postulate such relation, whereas Kjellmer (1984) and Greenbaum (1974) did. As for the researchers belonging to the phraseological approach (Mel′čuk 1998; Cowie 1991, 1998; Benson *et al.* 1997) some sort of grammatical relation is inherent in the methodology adopted. If grammatical relations are ignored, we run the risk of calling sequences like *and the* and *but for* collocations. Even though we may observe a large number of recurrent uses of the sequence *and the*, it does not make much sense using these sequences as target collocations in a test, primarily for pedagogic reasons. It therefore seems sensible to restrict ourselves to certain predefined syntactic patterns. Kjellmer (1994) makes use of no less than 19 pattern categories. A feasible list of prospective candidates could include such patterns as adjective + noun, noun + verb, noun + noun, adverb + adjective, verb + adverb and verb +

noun (see Nesselhauf 2005). All of these are what Benson *et al.* (1997) refer to as 'lexical collocations' (nouns, adjectives, verbs and adverbs). Their other subgroup of collocation is called 'grammatical collocation', referring to combinations in which "a dominant word noun, adjective, verb" occurs together with "a preposition or grammatical structure such as infinitive or clause." (1997:xv). Thus, examples of 'grammatical collocation', or, 'colligation' as it is sometimes called (Carter 1998:59), are *abide by, interested in*, and *admiration for*[8].

In terms of grammatical fossilisation, an aspect having to do with the extent to which a collocation may be subjected to different kinds of syntactic variation or manipulation, certain conditions can be argued to apply to collocation. As opposed to idioms (e.g. *kick the bucket, bite the dust*), collocations can generally be seen to allow considerable manipulation, such as passivisation, pronominalisation, fronting, clefting, insertion of material (Fontenelle 1998) and tense marking. This does not, however, mean that all collocations allow all these types of manipulation. Consider (4) and (5) below:

(4) a. he will say a prayer
b. he said a prayer
c. he is saying a prayer
d. ?a prayer was said by him

(5) a. he will make a mistake
b. he made a mistake
c. he is making a mistake
d. a mistake was made by him

In (4a-c) we see that for the collocation SAY + PRAYER, use in future time reference, the past tense as well as the present continuous, in addition to the simple present tense, come across as acceptable variants. As for passivisation (4d), however, it is doubtful whether this manipulation is acceptable. In (5a-d) we see that also passivisation seems perfectly acceptable for MAKE + MISTAKE, in addition to the tense variations. One way of looking at fossilisation is in the form of continuum. An example of an expression residing close to the fully fossilised end of that continuum is provided by Carter (1998). Consider (6a-c).

(6) a. *cats and dogs were rained
b. *it's raining dogs and cats
c. *it's raining heavy cats and dogs

Carter points out that *it's raining cats and dogs* a) cannot be passivised, b) does not lend itself to word order change, and c) does not allow insertion.

In the present thesis, inflectional changes are seen as a ubiquitous feature of collocations. This is in line with the view that lemmas are the underlying elements of collocations, and different word forms make up instantiations of the same collocation.

[8] Interestingly, on a slightly anecdotal note, colligations like *interested in* and *angry with* have been part of the study of English grammar in the Swedish education system for a long time. Grammatical collocations can therefore be seen to have had more prominence in English language teaching than lexical collocations; the former have been explicitly taught, whereas the latter have generally not been taught.

2.3.5 Adjacency

Adjacency concerns whether collocating words must occur immediately next to one another, or whether they may appear within a certain defined distance from each other. This distance is normally operationalised as number of orthographic words. The view in which only adjacent word forms can be seen to form a collocation (see e.g. Kjellmer 1994) ignores the fact that word forms that attract each other may be positioned a number of orthographic word slots apart, at least for certain types of constructions. The attraction in question is evidenced in the results from the application of statistical methods mentioned in the previous criterion. This is particularly true for constructions such as the verb + object NP construction. In these constructions, it is not uncommon for a premodifying element to intervene between the verb and the noun functioning as the object of that verb. Take, for example *he made a horrendous mistake*, in which the word *horrendous* together with the indefinite article *a* separate the core elements of the collocation MAKE + MISTAKE. Thus, for the purposes of the present study, collocates making up a collocation are either adjacent, *or* found within a specified distance from each other. The exact distance will need to be specified in more detail. In this thesis, because certain syntactic patterns will be focused on, the distance will be dependent on the characteristics of syntactic structures. As pointed out by Evert & Krenn (2003), if the span size is kept small, it is unlikely to cover non-adjacent collocates of node words, where the potential collocations are structurally flexible. Conversely, a big span size leads to an increase in candidate collocations which in turn increases the amount of "noisy data" which needs to be discarded for lack of relevance. For example, it seems reasonable to assume that for adjective + noun combinations, either no other elements will occur between these two elements, or elements such as other adjectives, and adverbs may be observed in that position.

2.3.6 Frequency of co-occurrence

In the present thesis, in order to control for the frequency criterion, a large, computerized corpus (the British National Corpus (BNC[9])) will be employed in the pursuit of frequently recurring word combinations. My aim is to use collocations that display a high frequency. The exact cut-off score will be specified in each study. In this regard, in addition to the absolute number of co-occurrences of words, we ideally need some sort of statistical measure of significance. Many measures exist in which the idea is to establish whether two words occur together in specified text spans more often than would be expected considering the absolute frequencies with which the two words occur in the corpus as a whole. Such measures will be considered for use in the empirical investigations presented in Chapters 3-6. Below, I will account for the rationale behind this standpoint.

Frequency is an essential criterion to take into consideration with regard to a definition of 'collocation'. In order for a word combination to become conventionalised in a language or a speech community, that sequence must occur repeatedly in use across a substantial number of usage events. Conversely, if a word combination is limited in use, both in terms of number of speakers, and in terms of the overall frequency of occurrence, then the result is that it will not become conventionalised. As was shown in the above review, some researchers treat all co-occurring words as collocations, whereas some reserve it for recurrences of relatively high

[9] The BNC will be described in Chapter 3.

frequency. One inherent problem is of course to decide the cut-off point for when the frequency of a word combination is high enough for it to be called a collocation.

Since the aim is to construct tests of L2 collocation knowledge, it seems sensible to include word combinations that are frequently used by native speakers, in this case, native speakers of English. In the absence of diachronic data, we can probably assume that mere convention make native speakers use the verb *say* with the object *prayer*, and the verb *tell* with the object *joke*, but not **say + joke* and **tell + prayer*. As argued by Ellis (2002), both the recognition and production of words is a function of their frequency of occurrence in the language. The same thing goes for sequences of words. It stands to reason that through repeated exposure to authentic sources, learners gradually figure out what sequences of words are normally used in certain situations. Again, in the words of Ellis: "Nativelike competence, fluency, and idiomaticity require and awful lot of figuring out which words go together" (2002:157). The advent of computerised corpora, and the growing opportunity of statistical investigations of the patterning in the texts of those corpora, has shown that certain word combinations display a high degree of recurrence across text genres and different speakers or authors (see e.g. Hunston 2002; Moon 1998).

2.3.7 Lexical Fossilisation

In the present thesis, emphasis will be put on collocations that display some degree of restrictedness, at least with regard to one of its constituents. This means that collocations whose parts can all be substituted with the same semantic meaning retained, will not be primarily considered. However, as pointed out earlier, one relevant factor for selecting collocations for the tests to be constructed is frequency, and especially frequency combined with some measure of significance. It is hypothesized here that the more restricted a collocation is, the greater is its potential for being a combination of words that co-occur repeatedly more often than chance, in terms of probability.

When it comes to lexical fossilisation, or lexical substitutability, it can actually be seen to have strong links to the semantic opacity criterion targeted below. More about these links will therefore be discussed in section 2.3.8, and it suffices to say here that word combinations that display a low degree of substitutability, i.e. a high degree of fossilisation, tend to also to lie close to the more opaque, rather than the transparent side, of a semantic opacity continuum (compare Table 2.1 in section 2.2.3.1 above).

Lexical fossilisation has to do with the degree to which lexical substitutions may be carried out in a so-called collocational frame (Nation 2001). A basic assumption when it comes to these potential substitutions is that the objects of substitution are semantically-related items, i.e. words of related meanings. Nation (2001:331) brings up the example of the verb *entertain*. This verb may be used in one of its extended senses to mean 'to nurture', as seen in *entertain a belief*, *entertain an idea* and *entertain a desire*. Howarth (1996:111) extends the list of possible and semantically related nouns through *entertain a view*, *entertain an opinion*, and *entertain a notion*. The nouns all belong to the same set of collocates. Similarly, Stubbs (2001) uses the verb lemma CAUSE to exemplify related collocational patterns, as in *cause problems*, *cause concern*, and *cause trouble*. In these examples, the noun is lexically substitutable and the verb CAUSE may occur in a small number of related collocations. Stubbs calls the relation between a lemma and a set of semantically related words "semantic preference" (p. 65). The examples given above are all collocations in which a certain degree of substitutability applies. An example of a wholly lexically fossilized collocation is *curry*

favour, in which neither of the two elements is substitutable. This is also where more idiomatic word combinations reside.

2.3.8 Semantic opaqueness

The semantic opaqueness criterion has to do with the degree to which the meaning of a word combination is deducible from the meanings of its constituent parts. The traditional way of making a distinction between a 'collocation' and an 'idiom' is that the meaning of the latter is not fully deducible from the individual meanings of the constituent parts, whereas the meaning of the former generally is. Put another way, with regard to idioms, Sweet argues that "the meaning of each idiom is an isolated fact which cannot be inferred from the meaning of the words of which the idiom is made up" (1899:139). The term used for this state of deducibility is 'compositionality' (a term generally accredited to Frege). Thus, an expression is non-compositional if its overall meaning cannot be seen to be the function of the meanings of its immediate constituents. A related but different term often found in the literature is 'opacity' or 'opaqueness' (see e.g. Ayto 2006). Compositionality is probably best seen as an either-or phenomenon. Either a word combination is compositional or it is not. Semantic opaqueness, on the other hand, is best seen as a continuum on which phrases and expressions are positioned according to their degree of opacity. The continuum ranges from fully opaque to fully transparent. Fully-fledged idioms reside at the opaque end of the spectrum, whereas 'collocations' are generally treated as structures that occupy the middle ground and the sphere at the more transparent pole. Howarth (1996) exemplifies the distinction between more idiomatic word combinations and more transparent collocation types of word combinations by *foot the bill* and *fill the bill*. In the former, the use of the noun *bill* is literal, referring to a bill of payment whereas the use of the verb *foot* is highly specialised, corresponding to 'pay'. In the latter, the noun *bill* cannot be seen to make any analysable individual contribution to the overall meaning. Neither of these are compositional word combinations, though. In many cases, though, the distinction is not easily made. As a case in point, Wiktorsson (2003), following Warren (2001), discusses compositionality and opaqueness in relation to idiomaticity, a term generally referable to native-like choices of expressions in language use. Wiktorsson argues that opaque expressions are necessarily non-compositional, whereas she raises some doubts about whether transparent expressions must necessarily be compositional. Using the sequence *answer the door* as an example, she claims that although it can be seen to reside more to the transparent pole of the continuum, it is not fully transparent since the object answered is not really the door, but rather a person (2003:17). The verb *answer* in *answer the door* is an example of a restricted use[10]. No other verb can be used together with the object noun *door* with the same retained sense. The sense can be argued to be 'to open a door, prompted by the door bell ringing, or a knock, to see who is there and inquire their purpose'. Consider also the example *make the bed*. There are two meaning interpretations of this phrase. Either, it could refer to the literal construction of a piece of furniture, or it could refer to a process by which the piece of furniture is covered by a quilt and a cover in a tidy way. The noun *bed* is used in a semantically transparent way in both readings. The use of the verb

[10] According to searches in the British National Corpus, the verb may be used as the predicate of noun objects like question (1987 hits), phone (158 hits), query (91 hits), and prayer (32 hits). The noun door occurs 100 times in the object position of the verb answer.

make, on the other hand, being a very frequent, delexicalised verb, is not fully transparent in the second reading, and the phrase is therefore arguably non-compositional. The process of 'arranging the quilt and the covers in a tidy way' is not inferable from the use of *make*.

In the above account of the phraseological approach to collocation, we saw that Howarth (1996, 1998a) proposes a word combination continuum ranging across four types: free collocations, restricted collocations, figurative idioms, and pure idioms, based on *inter alia* Cowie (1994). It is widely agreed that the type that poses the greatest challenge to learners is restricted collocations, much depending on the fact that restrictions often seem arbitrary. In this thesis, therefore, focus will be put on restricted collocations. Following Howarth (1996:47, 91), these are word combinations in which one element is used in its literal meaning, whereas the other is used in a specialised sense. Drawing on work by Aisenstadt (1979), and Cowie (1991), Howarth argues that the specialised sense may in turn be subdivided into either figurative, delexical, or technical uses. These three terms warrant further explication.

As to the figurative use, an example like *surf the Internet* serves to illustrate a collocation in which the verb SURF is used in a figurative sense. Figurative uses of language are non-literal in that they do not primarily purport their original more concrete everyday meaning. Croft and Cruse (2004:193) define figurative language as "language use where […] conventional constraints are deliberately infringed in the service of communication", and claims that the motivation for its use is a speaker's feeling that no literal use will produce the same effect. However, it is not always totally clear where to draw the line between literal and figurative uses of language. As pointed out by Saeed (2003), language change leads to shifts in meaning of words, for example through metaphorical extension. Metaphorical extension is seen as a process whereby a new idea is depicted by way of something more familiar (p. 15), such as the use of *mouse* for the cursor controlling device for a computer. Other examples of figurative uses of a verb are *catch a cold* where CATCH does not carry one of its more prototypical literal meaning of 'seizing an object with one's hands', and *draw a conclusion* where DRAW is similarly used in a sense that extends away from the prototypical literal senses having to do with either 'sketching' or 'pulling'. It should be noted here that it is fairly common for verbs to display both literal and figurative senses. Howarth (1996:99) provides an example in *assume* in the sense 'accepting something as true before there is proof' versus *assume* in the sense 'begin to act in or exercise; take on', with *assume the validity of something* as an example of a literal use, and *assume responsibility* as an example of a figurative use.

In terms of delexical uses, taking verb + NP combinations as examples, the verb PAY in *pay a visit* is used in a delexicalised way. Similarly, MAKE in *make an arrangement* also lacks distinct meaning. In the literature, terms used to denote these kinds of verbs are 'light verbs' (Jespersen 1965; Butt 2003) and 'support verbs' (Mel′čuk 1998; Fillmore *et al.* 2003). These are semantically neutral verbs that are not predicating fully, even though they follow the standard verb complement schema. These neutral verbs can turn an event noun or state noun into a verb-phrase like predicate (Fillmore *et al.* 2003). Often, with regard to light/support verbs and their complementation, the term Support Verb Constructions (SVCs) is used (see e.g. Nesselhauf 2005; Langer 2005; Storrer 2006). The following general characteristics are attributable to SVCs. Firstly, the verb is often delexicalised to some extent and semantically bleached. Secondly, the SVC often has a corresponding, stand-alone verb derivable from the noun component of the SVC, as seen in *say a prayer -> pray*. Another

approach to this is to view the noun component as typically being a nominalization of a verb or an adjective (Storrer 2006). An interesting claim made about SVCs is that it is the noun that selects the verb rather than the other way around (Fillmore *et al.* 2003). In the words of Langer (2005:172):

> the verb does not semantically subcategorize any of its syntactic complements. This means that the noun is the predicate of the construction, the verb has mainly syntactic relevancy.

This can be argued to have also psycholinguistic implications. If a speaker wants to express a proposition in which someone performs the ritual of a prayer, then this concept is expressed by the noun form *prayer*. Having accessed this concept, and having made the mapping between the concept and the word form, the speaker must then select a verb to go with that noun. In order to follow linguistic convention, the speaker then chooses a form of the verb lemma SAY to form the full construction *say a prayer*. In terms of compositionality, Langer argues that prototypical SVCs are semi-compositional (2005). The noun is normally fully transparent, whereas the verb is to some extent semantically reduced, or rather, lexicalized.

An example of a restricted collocation in which one element has a technical meaning is *shrug one's shoulders*. The verb SHRUG has a very narrow meaning, which cannot be retained in a combination with any other noun. In this use it is therefore monosemous. Other examples of verbs used in a technical sense is CAST, as in *cast a vote*, meaning largely 'to vote', and PRESS, as in *press charges*, with the meaning 'accusing somebody formally of a crime'. It is common for collocations consisting of verbs displaying a technical sense to occur in a special register. The use of CAST in the example above is found predominately in political and editorial discourse, whereas the use of PRESS is found mostly in legal texts and newspaper articles. It is not always easy to distinguish collocations in which a verb has a technical sense from collocations in which a verb is used figuratively (see Nesselhauf 2005:33). Howarth (1996) suggests that it is perhaps not so much the semantics of the verb, as it is the occurrence in a specific register that makes a collocation technical rather than figurative. Also, he suggests, "the verb needs to be selected by the noun" (p. 94). This assertion corresponds to the ones made above by Fillmore *et al.* (2003, and Langer (2005), seeing the noun as the core element in SVCs.

The type that Howarth refers to as pure idioms will not be included as an item type in my tests. Being non-compositional units, idioms do present problems to learners (Read 2000). However, it could also be argued that precisely because they are non-compositional, language users are more prone to notice them. If they are noticed, and not understood, it seems intuitive that language users will presumably put more effort into negotiating their meaning. This in turn will make them stand out, and increase the chances of them being memorized for later retrieval. This process has links to the 'depth of processing hypothesis' (see e.g. Schmitt 2000), whereby mental information will be remembered to a higher degree, the more this information is manipulated and thought about. Thus, the problem they are claimed to present to learners of a language could in fact be seen as an accommodating factor in the acquisition process. This could be put in contrast to restricted collocations. It is a moot point whether restricted collocations are compositional. What is clear, though, is that they are not always fully transparent. Collocations vary in terms of transparency, and one could claim that it is the transparency of collocations that makes them deceptive from the point of view of learning. A

point made by Read (2000) and Moon (1997), is that idioms are very infrequent in terms of their occurrence in corpora. This could of course have to do with the fact that corpora are poor representations of language use in general, but Moon sees the small number of idioms in corpora as general tendencies which reflect some kind of reality. A textbook example of a pure idiom can be seen in the word combination *kick the bucket* in its idiomatic sense of 'die'. In this combination, the meaning of the whole cannot be derived from the meanings of the components. Both the verb and the noun elements are semantically specialised.

The distinction made in this thesis between a pure idiom, a restricted collocation, and a free combination can be seen in 7 (a-c) below. In example 7, the same verb (KICK) is used in a free combination, (a) *kick the ball*, a restricted collocation, (b) *kick a habit*, and a pure idiom, (c) *kick the bucket*. I argue here that (a) is compositional, and fully transparent, whereas (b) and (c) are non-compositional, but with varying degrees of transparency.

(7) a. kick the ball
b. kick a habit
c. kick the bucket

In (7a) the meaning of the whole phrase is deducible from the meaning of the inherent constituents. The phrase itself is predictable and fully generative in the language system. By knowing the core meanings of KICK and BALL, the acceptable combination of the two in *kick the ball* can be predicted. This largely corresponds to the notions of selectional restrictions, subcategorizing features, and argument structures in the transformation-generative linguistics tradition (Warren 2003). No specialised sense exists in any of the elements. In (7b) the noun *habit* retains its literal meaning of 'something done often which is hard to stop doing', whereas the verb *kick* is used in a specialised sense, evoking a meaning interpretation along the lines of 'push away' or 'get rid of'. It is non-compositional in a strict sense, even though the word *habit* does contribute to the meaning of the whole. In (7c) the meaning of the whole phrase is not a function of the meanings of its constituents. The senses of *kick*, *the* and *bucket* are absent from the 'die' sense of *kick the bucket* (Pitt and Katz 2000). It is the combination in (7b) of a noun used in its literal sense, with a verb used in a specialised sense, that makes it into a restricted collocation. Furthermore, assuming the non-literal reading of *kick the bucket*, it is the lack of individual referentiality on the part of both the verb and the noun in (7c) that makes it a pure idiom. This shows that a verb's polysemy and the degree of restrictedness of its different senses are of great importance when it comes to classifying word combinations that the verb enters into as free combinations, restricted collocations, or idioms.

2.3.9 Uniqueness of meaning

In subsection 2.3.8 above, the polysemy of the word combination *kick the bucket* was discussed. Another example is *draw a line* (commonly used with a definite article), which in a similar way can evoke two readings. One can be seen as a free combination having a literal reading: the act of drawing a line, for example on a piece of paper. The other is an idiom, meaning 'to set a limit' or 'to distinguish between two related concepts'. These two uses show that word combinations can indeed be polysemous. However, the question at hand here is if

collocations along the lines of criteria discussed in 2.3.8 are polysemous. This question does not lend itself to the same straightforward answer. Siepmann (2005) gives the English word combination *avoid an accident* as an example of a polysemous collocation. He shows that it may be translated in two different ways into German and French, respectively. In French, the two alternative verb translations are *éviter* and *échapper*, and in German *vermeiden* and *entgehen*. In terms of English primary translation equivalents, the French verb *éviter* corresponds to 'avoid', whereas *échapper* corresponds to 'escape' (*Duden Oxford* 2005). In terms of the German verbs *vermeiden* and *entgehen*, these correspond to 'avoid' and 'escape', respectively (*Robert Collins* 1987). Even though these translations of the verb *avoid* are possible, I do not see *avoid an accident* as a collocation displaying some kind of restriction in the senses discussed in 2.3.8 above. I would consider it to be a free combination. The fact that I have chosen to use the criterion of specialised sense for distinguishing a free combination from a collocation might have the implication that collocations in my definition are normally not polysemous. However, a caveat is perhaps called for. It is conceivable that what I refer to as collocations could be seen to display at least two different senses in particular contexts. The distinction between two senses of the same form is, though, ubiquitous for word combinations where one is widely seen as belonging to the free combination pole, and another the more idiomatic pole of a semantic opaqueness continuum.

2.3.10 Summary: the treatment of collocation in this thesis

Having discussed the nine criteria above, and also provided principled statements about how collocation will be treated in this thesis, it will be convenient here to take stock of these discussions and summarize the view of collocation taken in this thesis. I see collocations as associative connections between words present in language users' minds, and these connections are manifested in language use in textual instantiations. The associative connections between words are abstractions which apply to lemmas. The different forms of lemmas make up collocations that are observable as textual instantiations. Collocations are minimally binary structures which to varying degrees require additional orthographic words, like determiners and modifiers, in their forms as textual instantiations. The words making up a collocation in a textual instantiation are either adjacent, or found within a specified distance from each other. Collocations generally allow considerable morpho-syntactic manipulation, and the orthographic words making up a collocation are grammatically related in their textual instantiations, following the grammar system of a language. Collocations are furthermore conventionalized units, and they are therefore frequently occurring in a speech community. This is observable in the high frequency of textual instantiations of collocations in corpora. Statistical methods can be applied to single out words that are observed together with a higher frequency than would be expected in relation to their individual frequencies in a corpus. Collocations are either compositional or non-compositional word combinations, displaying varying degrees of opaqueness, in which one element is used in its literal meaning, whereas the other is used in a specialised sense, the specialised sense being figurative, delexical, or technical. In the majority of cases, collocations are monosemous.

In conclusion, collocations are associative connections between word abstractions in the mental lexicons of language users, which in their textual instantiations are conventionalized word combinations consisting of:

two syntagmatically related and frequently co-occurring orthographic words, either adjacent or separated by a specified distance, where one of the words is used in a figurative, delexical, or technical sense, and where the meaning evoked by the combination as a whole, sometimes requiring additional lexical elements for grammatical well-formedness and usage convention, is either compositional or non-compositional, and varies in its degree of opaqueness.

Figure 2.2 A working definition of collocation as it will be used in the present thesis

Admittedly, the above definition consists of at least two unclear points. Firstly, the distance between two collocating orthographic words is unspecified. In the present thesis, this variable will be discussed and specified in relation to each specific study conducted. Secondly, the level of frequency required for two orthographic words to be considered a collocation will also be specified in relation to each specific study.

In the next section, collocation as a theoretical knowledge construct for testing will be defined.

2.4 Collocation knowledge – defining the construct

2.4.1 Fundamental considerations

In this section, I will discuss what is involved in knowing a collocation. This is the process of defining a construct, and it is a necessary first step in any creation of a language test. The term construct is primarily a psychological term, but is used extensively in language testing (see e.g. Chapelle 1998; Alderson *et al.* 1995; Bachman & Palmer 1996). According to Davies *et al.*, a construct is a trait that a test is intended to measure. More specifically, it is "an ability or set of abilities that will be reflected in test performance, and about which inferences can be made on the basis of test scores." (1999:31). Thus, if we are to construct a test of collocation knowledge, we must, in as detailed manner as possible, define what it is we intend to measure. We do this in the effort to try to link the underlying ability and the test performances of the potential test takers.

Bachman (1990) recognises the need for a three-stage analysis in this respect.

(8) a. the construct needs to be defined theoretically;
b. the construct needs to be defined operationally;
c. procedures must be established for the quantification of observations.

The theoretical definition (a), is a specification of the relevant characteristics of the ability we intend to measure, and its distinction from other similar constructs. If there are several subcomponents to a construct, then the interrelations between these must be specified.

When it comes to the operational definition of the construct (b), this process involves attempts to make the construct observable. To a great extent, the theoretical definition will govern what options will make themselves available. For example, the theoretical definition of the construct 'listening comprehension' suggests an operationalisation as a task in which information must be decoded aurally in some fashion.

With respect to the third stage (c), our measurement should be quantified on a scale. In general, four different types of scale are acknowledged in measurement theory: 'nominal',

'ordinal', 'interval', and 'ratio' scales. Depending on the nature of the ability being measured, one of these will prove more or less appropriate. Ideally, ratio scales provide the largest amount of information, but it is not always possible to apply them. For most purposes of language testing, interval scales are sufficient. For an accessible account of the four types of scales, see Heiman (2006).

In the effort to apply Bachman's three-step procedure, I will first discuss the process of defining the construct to be measured theoretically.

2.4.2 Defining the knowledge construct theoretically

In an attempt to try to define the construct of collocational knowledge theoretically, it will be necessary to first try to delimit the ability to be measured in the test in as a precise way as possible. It therefore seems wise to start with a discussion of the integral parts of collocations, i.e. words, and what can be known about them. In this thesis, words are defined as strings of consecutive letters surrounded by blanks as found in written texts (cf. Lyons 1977:18); thus, in principle, the term 'word' will denote an orthographic word. A collective term for words in a language is vocabulary. Vocabulary studies have over the last two decades seen a tangible increase in interest, noticeable in the numerous collections of papers and monographs published, solely devoted to lexis (Carter 1987; Meara 1987; Carter & McCarthy 1988; McCarthy 1990; Nation 1990; Arnaud & Bejoint 1992; Lewis 1993; Schreuder & Weltens 1993; Coady & Huckin 1997; Schmitt & McCarthy 1997; Carter 1998; Cowie 1998a; Haastrup & Viberg 1998; Singleton 1999; Read 2000; Schmitt 2000; Nation 2001; Bogaards & Laufer 2004). The fact that vocabulary has risen from the ranks, as it were, does not mean, however, that there now is a unified way to treat vocabulary. A central and enigmatic question within the field of vocabulary is what is involved in knowing a word. Several attempts have been made to capture the answer this question. Cronbach (1942) proposed a framework consisting of five aspects of word knowledge. Richards (1976) followed suit and proposed a more comprehensive set of eight descriptors. More recently, Nation (2001) has proposed a framework aimed at describing what is involved in knowing a word. This framework is more elaborate than the previously mentioned frameworks, but at the same time it is clear that in many ways, it draws on the work by Cronbach and Richards.

Nation's (2001) descriptive framework provides a logical starting point for the present attempt to define collocation knowledge. The word knowledge framework is shown in Table 2.4 below.

Table 2.4 Description of "what is involved in knowing a word" from Nation (2001:27).

Form	spoken	R	What does the word sound like?
		P	How is the word pronounced?
	written	R	What does the word look like?
		P	How is the word written and spelled?
	word parts	R	What parts are recognisable in this word?
		P	What word parts are needed to express the meaning?
Meaning	form and meaning	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	concepts and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	associations	R	What other words does this make us think of?
		P	What other words could we use instead of this one?
Use	grammatical functions	R	In what patterns does the word occur?
		P	In what patterns must we use this word?
	collocations	R	What words or types of words occur with this one?
		P	What words or types of words must we use with this one?
	constraints on use (register, frequency)	R	Where, when, and how often would we expect to meet this word?
		P	Where, when, and how often can we use this word?

R = receptive knowledge, P = productive knowledge

As can be seen in the table, the description consists of four columns. In the first column from the left, three primary fields of knowledge can be found: form, meaning and use. In turn, these three fields are each divided into three subfields. This is indicated by the second column. For example, for the primary field "Form", we find the subfields "spoken", "written", and "word parts". The third column consists of the letters "R" and "P", respectively. The letter R stands for receptive knowledge and the letter P stands for productive knowledge. In the fourth column, we find questions intended to capture one aspect of the word knowledge framework. All-in-all, the framework consists of no less than 18 sub-aspects of what it means 'to know' a word. In effect, based on the framework, it is possible to ask all of the questions in the rightmost column in relation to a language user and a specific word in a language. For the purpose of illustration, I may ask myself all of the 18 questions about a word like *capricious*.

I may start with the questions about the form of the word. For example, do I know what the word *capricious* sounds like, and do I know how it is pronounced? If I can demonstrate this in some way, then I know the spoken form of the word, both receptively and productively. I can then continue to ask myself questions about the written form of the word capricious. In the same manner, one question pertains to the receptive aspect and another to the productive aspect.

The distinction made between receptive and productive skills merits a discussion. It is customary for researchers to make use of a distinction between receptive and productive knowledge of vocabulary items. References to this distinction are traced back to the middle of the 19th century (Waring 1999). In relation to vocabulary, Nation (2001:24-25) defines receptive use as involving "perceiving the form of a word while listening or reading and retrieving its meaning", whereas productive use "involves wanting to express a meaning through speaking or writing and producing the appropriate spoken or written word form".

It is widely agreed that a language user, in general, can recognize and understand more words than she can use when speaking or writing. It stands to reason that cases where a learner uses a word in production, but is not able to recognize or understand it receptively, are exceptions. In general, there has to be an initial exposure to a word involving listening or reading that precedes the first productive instance of it. However, it is of course conceivable that a learner's first receptive encounter with the word merely involves recognition of the form, spoken or written, and that any subsequent attempts to use it may be infelicitous due to lack of understanding of the proper meaning of the word. Conversely, a learner may use a word frequently when speaking to connote a specific concept, but could in theory fail to recognize the conventionalized orthographic representation denoting the concept. This might be more common in cases where the learner's L1 is typologically different from the L2, i.e. belonging to a different language family with few cognate words and different orthography.

Research carried out on size differences between receptive and productive vocabulary of L2 learners (Waring 1997) has shown that learners scored better receptively than productively on a passive definition-matching test and a controlled active test. Also, the learners' receptive vocabulary became progressively larger than their productive vocabulary as their overall vocabulary size grew. In terms of test design, it is of course crucial to take the distinction of receptive versus productive into account. However, as with most other dichotomy-like phenomena, when put under the magnifying glass, it tends to lose its clear-cut nature. The distinction between receptive and productive vocabulary is no different. A popular metaphor to use in these contexts is the continuum, allowing for gradual differences. Melka (1997) discusses degrees of familiarity a learner might have with a word, stating that phonological, morphological, syntactical and lexical information about an item constitutes a very high degree of familiarity, whereas merely having visual recognition ability suggests a low degree of familiarity. On the whole, Melka admits to the existence of empirical evidence for a difference between receptive and productive vocabulary, but dismisses a proper dichotomy (1997:101), and suggests the use of a continuum with degrees of familiarity. Meara (1990) proposes a diverging view from that of Melka. Meara argues that active vocabulary may be seen as existing on a continuum, but that passive vocabulary may not. The reason for this is that passively known vocabulary may only be accessed by means of appropriate external stimulation. He claims that there are no internal links available between the 'passive' word and other words in the lexicon network. Furthermore, Read (2000:154-157) calls for more narrow definitions of the terms production and reception in relation to testing purposes, introducing 'recognition' and 'recall', and 'comprehension' and 'use'. Recognition is taken to involve tasks where a learner is supposed to show that she has understood the meaning of a target word presented to her. Recall involves the presentation of some sort of stimulus, based on which the learner is expected to recall the target word from memory. Comprehension and use are seen to involve more context-dependent and comprehensive measures. Comprehension involves a task where the learner must show whether she understands a word given in a context, whereas use is involved when the learner is asked to produce one or several words, for example in oral retellings, translations and picture description tasks.

Irrespective of Melka's and Read's elaborations, and despite Meara's proposal for a different analysis, the two-fold distinction is a widely used notion which to a great extent affects the thinking of test designers and L2 vocabulary researchers alike. Because of its wide-spread use, I will employ the term since it will make the definition of the construct I intend to

measure more readily understandable. However, I will modify them when necessary with Read's terms from above.

Going back to Nation's table of what is involved in knowing a word, what is of primary interest is the knowledge aspect called "collocation". Just like all the other aspects of word knowledge, collocation has a receptive and a productive side to it. For the receptive side, a language user is expected to know what words or types of words occur with a specified target word. For the productive side, knowing what words or types of words to use with a specified target word is expected in order to meet the criterion. A decision to focus on either the receptive or the productive side of collocation knowledge would have the benefit of making the construct to be tested more precise. Testing productive collocation knowledge could for example entail analysing learners' attempt to produce conventionalised word combinations, either in samples of written texts (see e.g. Nesselhauf 2005), or in more experimental word association designs (see e.g. Schmitt 1998b). Although learners' production of collocations is indeed an intriguing field of study, measuring this type of knowledge in a test will have a number of more practical consequences that need to be considered. A productive task à la Schmitt, in which prompt words are given to informants who are in turn expected to yield common collocates of those prompt words, would require more time per tested item for the informants. The procedure would in all probability mean heavy restrictions on the number of informants that could be tested. Also, and perhaps more importantly, the scoring procedure would be considerably more intricate since a system would have to be developed for quantifying the informants' responses in some way. In contrast, choosing to measure receptive collocation knowledge could be seen to bring a number of positive effects. Firstly, it would be possible to test a larger number of items in each test session. Secondly, an objective scoring key could with minimal effort be produced for the test. Thirdly, the testing of receptive skills would have the potential of being transferable to computerized test formats in a way that productive tests would not have to the same extent. For these reasons, I opted for testing receptive collocation knowledge.

Having decided to use receptive collocation knowledge as the construct to be measured, it will be necessary here to take a closer look at what this knowledge entails. In the previous paragraph I referred to Nation's description of what is involved in knowing a word in relation to the knowledge aspect of collocation (2001:27). In fact, Nation has adapted his framework to more specifically describe how the different aspects of word knowledge could be tested (2001:347). The word knowledge framework for testing is shown in Table 2.5 below. Starting on the left of the table, we recognize the first three columns from Table 2.4 above. The difference between the tables is column four. In this column in Table 2.5, we find questions pertaining to each aspect of word knowledge, aimed at guiding what is to be tested. With regard to receptive knowledge of collocations, the relevant question here is: "Can the learner recognise appropriate collocations?"

Table 2.5 Aspects of word knowledge for testing, from Nation (2001:347) [with correction].

Form	spoken	R	Can the learner recognise the spoken form of the word?
		P	Can the learner pronounce the word correctly?
	written	R	Can the learner recognise the written form of the word?
		P	Can the learner spell and write the word?
	word parts	R	Can the learner recognise known parts in the word?
		P	Can the learner produce appropriate inflected and derived forms of the word?
Meaning	form and meaning	R	Can the learner recall the appropriate meaning for this word form?
		P	Can the learner produce the appropriate word form to express this meaning?
	concepts and referents	R	Can the learner understand a range of uses of the word and its central concept?
		P	Can the learner use the word to refer to a range of items?
	associations	R	Can the learner recall this word when presented with related ideas?*
		P	Can the learner produce common associations for this word?
Use	grammatical functions	R	Can the learner recognize correct uses of the word in context?
		P	Can the learner use this word in the correct grammatical patterns?
	collocations	R	Can the learner recognize appropriate collocations?
		P	Can the learner produce the word with appropriate collocations?
	constraints on use (register, frequency…)	R	Can the learner tell if the word is common, formal, infrequent, etc.?
		P	Can the learner use the word at appropriate times?

R = receptive knowledge, P = productive knowledge

* = This wording is claimed to be referring to productive skills in Nation's original table but this must be a mistake. It has therefore been re-arranged to refer to receptive skills in the above table.

At first sight, it would seem possible to use this question to guide our construct definition. If we for the sake of argument ignore the receptive/productive distinction, there are a total of nine aspects of word knowledge that may be tested according to Nation's framework. This means that a test of collocation knowledge would only target one out of nine types of word knowledge. However, Nation's table, although helpful, obscures the rather complex cognitive processes assumed to be involved in receptive collocation knowledge. It will be argued here that a number of the word knowledge aspects laid out in the table can in fact be subsumed in the collocation word knowledge aspect.

In order to corroborate this claim, I will use a word combination which is argued to be a collocation as defined in this thesis. Consider the collocation *say a prayer*. It consists of three word class elements: a verb, a determiner, and a noun. In the process of knowing this collocation receptively, a process which is here taken to imply recognizing it upon presentation, a language user must arguably have a command of the following receptive aspects. Starting with the form field, assuming a written test, the spoken aspects are not relevant. The next aspect, the written form, however, is. Thus, an informant would initially have to recognize the written forms of all these three words. Moving on to the third aspect in the form field, the recognition of word parts, the relevance of this is uncertain. On the one hand, it could be argued that recognizing word parts is to a great extent a strategic competence (cf. Nagy 1997) which helps learners in the processing and acquisition of word forms. Nation asserts that there is value in seeing the knowledge of word parts as accommodating in the

process of recognizing words (2001). For example, even if a language user has never been exposed to the form *prayer*, knowing the verb form *pray*, and recognizing the suffix *–er*, might help him or her deciding if *say a prayer* is known. Thus, it is conceivable that also the recognition of word parts is involved in receptive collocation knowledge.

Moving up to the meaning field of the table, and the first aspect of form and meaning, it is possible, but not likely to any greater extent, that a language user would verify the recognition of *say a prayer* without an initial recourse to, firstly, the possible meanings of the three elements, and secondly, to the meaning of the whole phrase. If no meaning can be retrieved from the mental lexicon that may be linked to the form in question, then the language user is left with a situation in which the form has to be acknowledged only on the basis of it being an isolated form. From another perspective, a language user who can readily match the presented forms with meanings, and also the whole combination of those individual forms, will have no problem acknowledging the form as an occurring string of words in English, unless of course there is some kind of structural aspect that causes doubt. I will come back to what I mean by 'structural aspect' when I discuss the word knowledge aspect of grammatical functions. For a L2 user, the meaning of a target word is often related to a meaning in the L1, a translation equivalent. On the whole, it seems feasible to assume that form-meaning mapping is an auxiliary process, part of collocation recognition. The next word knowledge aspect is concepts and referents. It is a difficult task to answer the question of what a word's concept and referents are. Cruse defines 'concept' as "organized bundles of stored knowledge representing an articulation of events, entities, situations, and so on in our experience" (2000:127). Words in a language differ with respect to polysemy. Some words, especially high-frequency words, are highly polysemous. Miller even goes as far as saying that "it is a perverse feature of natural languages that the more frequently a word is used, the more polysemous it tends to be" (1999:12). Low-frequency words tend to display a lesser degree of polysemy (compare, though, Ruhl 1989). With highly polysemous words, some senses are clearly related and can be seen to share the same basic concept. This basic concept is best seen as some kind of abstraction of the sometimes very pragmatic specializations or modulations of senses. Whether language users need to pay heed to concepts when deciding whether a presented collocation is known or not is not all-together clear. However, if Cruse's definition from above is accepted, then the process of drawing on stored knowledge about words is very likely to be involved in the decoding of meaning of word forms. For this reason, aspects of concepts and referents are seen to be involved in receptive collocation knowledge. The last aspect in the meaning field is associations. The central issue concerns what other words are activated when a target word or a group of target words are presented. Using the example collocation *say a prayer*, this combination of words is believed to trigger and activate other words in the mental lexicon of a language user. Different categories of the types of word associations that are normally made exist, but commonly used subclassifications involve words that are syntagmatically linked to a target word, words that are paradigmatically linked to a target word, and phonologically linked words (also called clang associations) (see e.g. Wolter 2001). The syntagmatic links are often words that collocate with the target words. This implies that word association is inherently involved in receptive collocation knowledge.

The third and final field in Nation's framework consists of the following aspects of word knowledge: grammatical functions, collocations, and constraints on use. The first one, i.e. grammatical functions, can be seen to concern information like what part of speech and

grammatical patterns words enter into. It is hard to say whether knowing what parts of speech *say*, *a*, and *prayer* belong to facilitates the recognition of these words in a sequence as a collocation. It could be argued, though, that recognizing *say* as a verb, *a* as a determiner, and *prayer* as a noun, is necessary in order to accept the structure as a legitimate grammatical pattern of English. It should be noted that for a learner the linguistic terms used for these categories is not what is important, but rather that *say* is an action/a process, rather than an object. Acceptance would perhaps not be granted if, for example, the structure **said an pray* was presented.

When discussing the form-meaning mapping aspect above I promised to come back to what I called 'structural aspect'. What I mean by 'structural aspect' is in fact a deviation from the patterns of language that a language user has experienced. Such a structural aspect could for example have to do with the form of the verb not being met before in combination with the determiner and the noun. In theory, the language user might only have been exposed to the collocation where the past tense was used, e.g. *said a prayer*. In that case, a decision has to be made whether the collocation may also exist instantiated by the base form of the verb, as in *say a prayer*. The next aspect of word knowledge is collocation itself, and since it is the object under study here, no further comments will be made about it in relation to Nation's table. Instead, the final aspect of word knowledge is constraints on use. This aspect is essentially linked to sociolinguistic constraints, such as register, but also a factor like word frequency. Thus, having knowledge about the level of formality of words, and whether words are frequent in use or not in a language is the kind of knowledge addressed here. I would argue here that this aspect is probably not relevant for the recognition of collocations in English. It would certainly be relevant when studying production, where appropriateness of expressions in a certain context is of the essence.

The exemplification and discussion above go to show that there are several other aspects of knowledge involved in receptive collocation knowledge, than just the designated collocation aspect itself, and that receptive collocation knowledge can therefore be seen as a cognitively complex construct. Table 2.6 below indicates by the letter 'X' those additional knowledge aspects that have to some degree been identified as relevant to receptive collocation knowledge.

Table 2.6 Aspects of word knowledge for testing relevant to or subsumed in the construct 'receptive collocation knowledge' (Table based on Nation 2001:347).

Form	spoken	R	Can the learner recognise the spoken form of the word?	
	written	R	Can the learner recognise the written form of the word?	X
	word parts	R	Can the learner recognise known parts in the word?	X
Meaning	form and meaning	R	Can the learner recall the appropriate meaning for this wor form?	X
	concepts and referents	R	Can the learner understand a range of uses of the word and its central concept?	X
	associations	R	Can the learner recall this word when presented with relate ideas?	X
Use	grammatical functions	R	Can the learner recognize correct uses of the word in context?	X
	collocations	R	Can the learner recognize appropriate collocations?	X
	constraints on use	R	Can the learner tell if the word is common, formal, infrequent, etc.?	

R = receptive knowledge

I will come back to this table, and the claims made above in the next section. However, a number of comments are needed at this stage. Firstly, Nation's tables of word knowledge aspects provides a framework for what is involved in knowing a word (Table 2.4), and how these aspects could be tested (Table 2.5). As has been pointed out by Schmitt and Meara (1997), the framework is descriptive and does not have the power to explain the processes of acquisition for the different word knowledge aspects, or how they interrelate. However, they hypothesize that the different aspects must be interrelated. They also conclude, based on an empirical study, that two word knowledge aspects investigated – word affix knowledge and word association knowledge – were related with significant correlation coefficients in the range 0.3-0.5 (1997:30), a weak relationship. To varying degrees, both these word knowledge aspects, furthermore subdivided into productive and receptive skills, correlated also positively with scores on a vocabulary size test and scores on a general language proficiency test, but the correlations were not significant throughout.

The way Nation's framework will be used in this thesis is mainly as a descriptive tool based on which a theoretical definition of the construct receptive collocation knowledge can be understood. Based on the above discussion, where the types of word knowledge aspects assumed to be relevant to receptive collocation knowledge were highlighted, we are now in a position to propose a theoretical definition of receptive collocation knowledge as a construct, seen from a learning perspective:

> The knowledge necessary for appropriately recognizing that two or more words frequently occur together as conventionalized word combinations in a language, and accessing the meaning of these combinations to some degree.

Figure 2.3 A theoretical definition of the construct 'receptive collocation knowledge'.

The next step will be to define the same construct operationally. The basic considerations involved in that process will be discussed below.

2.4.3 Towards an operational definition of the construct

2.4.3.1 Introduction

Following Bachman's principle of fundamental steps in measurement, the knowledge construct of receptive collocation knowledge now needs to be defined operationally. This will make it possible to relate the knowledge investigated to an observed behaviour of some sort. Since the receptive knowledge of collocations is a property of the way words are associated with each other in the mental lexicon of a language user, it cannot be directly observed. We therefore need to make it observable in some way, and the method for making the knowledge observable is through testing it. A test can be carried out in many different ways, and operationalisations of knowledge constructs like receptive collocation knowledge may vary at different stages of a test development process. This is so because a certain operationalisation may not prove to be tenable in the light of obtained empirical data from a test administration, and may need to be changed. For this reason, it is not possible at this stage to operationally define receptive collocation knowledge. Instead, in the following sections I will draw up more general considerations relevant to the subsequent forming of operational definitions of the construct.

In Chapter 1, the lack of standardised tests of collocation knowledge was identified. In this thesis, since collocation knowledge is seen to be intimately related to general vocabulary knowledge, an expedient approach at this stage is to look at the field of vocabulary testing when it comes to finding suitable test formats and possible frameworks. As a basis for this endeavour, two influential distinctions within vocabulary testing will be discussed: vocabulary breadth and vocabulary depth. Furthermore, frequently used tests of vocabulary breadth and vocabulary depth will be reviewed, and their potential relation to the receptive collocation knowledge construct will be addressed. After that, basic considerations guiding my test construction will be addressed.

2.4.3.2 Testing L2 Vocabulary

2.4.3.2.1 Vocabulary breadth and vocabulary depth

At the present stage of research within vocabulary testing, two influential dimensions of lexical knowledge are assumed to exist: vocabulary breadth and vocabulary depth (see e.g. Wesche & Paribakht 1996; Greidanus *et al.* 2004; Read 2004). The terms are claimed by Read (2004:210) to have been used since the early twentieth century in various ways in the vocabulary literature. The more recent treatment of the two terms was however coined by Anderson & Freebody (1981:92-93) who asserted that:

> It is useful to distinguish between two aspects of an individual's vocabulary knowledge. The first may be called "breadth" of knowledge, by which we mean the number of words for which the person knows at least some of the significant aspects of meaning. … [There] is a second dimension of vocabulary knowledge, namely the quality or "depth" of understanding. We shall assume that, for most purposes, a person has a sufficiently deep understanding of a word if it conveys to

> him or her all of the distinctions that would be understood by an ordinary adult under normal circumstances.

Interpreting the defining parts of the quote above in relation to Nation's word knowledge framework presented above (see section 2.4.2) it is possible to assume that 'breadth' has to do with the form and meaning, and concepts and referents aspects of the framework, and that 'depth' is more closely linked to aspects such as word parts, associations, grammatical functions, collocations, and constraints on use.

2.4.3.2.2 Vocabulary breadth and its application in testing

Much work in vocabulary testing has been preoccupied with a dimension called 'vocabulary breadth'. Another term used for the same dimension is 'vocabulary size' (see e.g. Meara 1996). The two terms are used interchangeably in the literature to denote the same concept, and I will henceforth use vocabulary size in the present thesis to denote how many words a learner knows with regard to a basic meaning.

Several studies have been conducted with the aim of trying to estimate the size of a learner's vocabulary (Ellegård 1960; Goulden *et al.* 1990; D'Anna *et al.* 1991; Hazenberg & Hulstijn 1996). Basically, there are two conventionalized ways of going about this. One way is to take a sample from a dictionary and the other is to use a sample from a frequency list based on a corpus. The dictionary-based technique implies that a representative sample of words (every *n*-th word) is taken from the dictionary and that learners are tested on those words (see Nation 1993). The rationale behind this is that the score on the test may be generalised to the total number of words in the dictionary[11]. For example, if the sample consisted of one in every 10 words in the sample, then the test-taker's scores on the test would be multiplied by 10 in order to arrive at the overall vocabulary size. Examples of this approach can be found in Goulden *et al.* (1990) and D'Anna *et al.* (1991), who focused on native speakers. The technique used for the compilation of a frequency list is intrinsically based on some sort of corpus. The corpus may either be a general corpus or a specialised one. An example of a frequency list based on a specialised corpus is The Academic Word List (Coxhead 1998, 2000), and examples of well-known and commonly used frequency lists based on more general corpora are *The Teacher's Word Book*[12] (Thorndike & Lorge 1944), *The General Service List*[13] (West 1953) and a list based on the Brown corpus, provided by Francis & Kucera (1982). Normally, the words of frequency lists are arranged in different bands: the band containing the 1,000 most frequent words is called 1K; the band containing the second thousand most frequent words is called 2K, etc. Tests based on these types of bands are designed on the same assumption as the dictionary-based ones: if a test taker knows a proportion of the sample items from a particular band, then it is assumed that she will know a corresponding proportion of all the words in that band.

The basic relation between vocabulary size and receptive collocation knowledge is fairly obvious. Since receptive collocation knowledge has been defined as the knowledge necessary

[11] Employing a so-called spaced sampling for test purposes may lead to a sampling problem. If, for example, the first word on every fifth page is used, then due to the fact that high-frequency words have more entries per word and more spacious entries in the dictionary, the result will be that more high-frequency words will end up in the sample than there should be.

[12] Contains about 13,000 word families based on an 18,000,000 million word written corpus.

[13] Contains 2000 headwords based on a 5,000,000 word written corpus.

for appropriately recognizing that two or more words frequently occur together as conventionalized word combinations in a language, it stands to reason that this latter process requires knowledge of some kind of the single words making up the collocation. It is argued here that this knowledge implies minimally recognition knowledge of the word forms, but probably also a mapping of the forms to meanings. It is very unlikely that a language user could identify a collocation, in its sense of conventionalized word combination, without recognizing the inherent words of that collocation as precisely words. Furthermore, not knowing any of the possible senses of those words, would very likely make the process difficult. However, even though the vocabulary size dimension is intrinsically linked to receptive collocation knowledge as a construct, the more exact relation between the two is not clear. Does a large vocabulary automatically lead to a high degree of receptive collocation knowledge? Or, put another way, does knowing many single words also mean knowing how these single words may be combined into conventionalized sequences of words? No straightforward answers to these questions seem to be available in the literature. This fact makes them all the more interesting.

Since vocabulary size and receptive collocation knowledge are seen to be related, in a thesis devoted to test construction, widely used tests of vocabulary size deserve a closer look. Two of the more commonly used tests and test formats of vocabulary size are: The yes/no test, and the Vocabulary Levels Test, respectively. The tests are briefly described with reference to intended use and purpose, design and underlying assumptions, and advantages and possible drawbacks.

2.4.3.2.3 The yes/no test

The yes/no test, or the checklist test, as it is also called, is essentially a word recognition test in which the test-taker is asked to indicate whether the meaning of a substantial number of single words is known. The test as such measures vocabulary size, and it is most commonly used as a placement test (Nation 2001:348). The most well-known versions of the test were developed by Meara and Buxton (1987) and Meara and Jones (1990). For the original idea, Meara and Buxton give credit to Zimmerman *et al.* (1977), who used it with L1 speakers, as did Anderson and Freebody (1983). The latter introduced dummy words in the test in order to be able to see if a testee overstated her knowledge. The dummy words, imaginary made-up items, follow the word formation rules of the target language.

The test basically relies on self-report of knowledge of meaning on the part of the testee, but by measuring and taking into account the number of times a dummy word is said to be known, deductions can be made from the final score adjusting it downwards. The technique for controlling for false claims is taken from Signal Detection Theory (see e.g. Green & Swets 1966). For the mathematical formula used, see Anderson and Freebody (1983). The items for the test are taken from frequency lists. A random sample of words is chosen from each frequency band of a 1,000 words. Because of the simplicity of the test design, a large number of items can be tested in a very short time and the sampling rate can therefore be kept at a comparatively high level, even down to one word in 10 up to as many as 10,000 words (Meara and Buxton 1987:151). An excerpt from a standard pencil and paper test, taken from Meara (1996:43), is given below:

1	[ ]	regard	2	[ ]	invention	3	[ ]	calendar
4	[ ]	guest	5	[ ]	communist	6	[ ]	amagran
7	[ ]	galpin	8	[ ]	hudd	9	[ ]	construct

Figure 2.4 Test items from a Yes/No test (Meara 1996:43).

The test from which the excerpt is taken provides a one in 25 sample of the target vocabulary. It consists of a total of 60 items, of which 40 are proper words and 20 are dummy words. Test-takers are instructed to tick the boxes beside the words whose meaning they know, and to leave the words which are unknown unmarked. Guessing is not encouraged.

The merits of this test lie primarily in the fact that it is easily administered, quick to take and that it covers relatively many items. Computerized versions of the test have been developed, for example the EVST (Meara and Jones 1990) and the X_lex test (Meara and Milton 2003; Meara 2005). The tests do not measure total vocabulary size, but provides an estimate of size in relation to the 10,000 most frequent lemmas of English in the case of the EVST, and the 5,000 most frequent lemmas of English in the case of the X_lex. They take less than ten minutes to sit and heighten the ease of administration even further through automation and self-scoring. Like their predecessors, these more recent versions are used as placement instruments, but as pointed out by Milton (personal communication, 2004), in some cases, they seem to be used by education administrators for measuring student achievement within a specific curriculum.

In a more critical view, problems identifiable with the test include that it implies that words have just one meaning, and that a test-taker does not overtly show knowledge of what the tested words mean. Another problem has been identified concerning test-takers who have an L1 which harbours many cognate words to the tested L2. An example of this is French speakers taking the English test, where the close relationship of the two lexicons of the languages is suggested to be the reason why the test performances of this group of subjects did not correlate as well with other linguistic skills as was the case with other speaker groups (Meara 1996). Another problem that has been identified is the question of how best to adjust the scores observed for the correctly identified words (also called 'hits') in the test on the basis of observed 'false alarm' rates. False alarms refer to answers in the test, where a test-taker claims to know the meaning of a dummy word (a distractor), which is taken to indicate an overestimation of known proper words. Shillaw (1999), in a study involving L1 Japanese informants, observed that using only the scores based on hits produced more reliable results than did scores that were adjusted by subtracting false alarms. Similarly, Eyckmans (2004) observed high false alarm rates with French-speaking learners of Dutch.

2.4.3.2.4 The Vocabulary Levels Test

The Vocabulary Levels Test (VLT) was primarily designed to be used as a diagnostic tool, helping teachers to plan the vocabulary learning parts of language courses for students (Read 2000). It contains different levels which are each linked to specific levels of learning objectives. It was developed by Paul Nation and it has in its original format been published twice (Nation 1983, 1990), and recently in two updated versions (Nation 2001; Schmitt 2000). The test measures vocabulary size, i.e. estimates a learner's knowledge of common word meanings, and the task involves a matching of English words with English definitions. The test format consists of five parts, each relating to a particular frequency level of English. These levels are the first 2,000, 3,000, 5,000, and 10,000 words and a level called the

university word level, which is fitted in between the 5,000 and the 10,000 word levels. The university level was included in the test due to the fact that the test was initially aimed at testing international students coming to New Zealand for subsequent university studies. The frequency list for the university word level was based on Campion & Elley (1971), whereas the other four levels were based on Thorndike and Lorge (1944), where comparisons were also made with West (1953) and Kucera and Francis (1967)

The test items are provided in blocks. Below, examples of blocks from two of the test levels are shown:

The 2000 word level

1	arrange		
2	develop	____	grow
3	lean	____	put in order
4	owe	____	like more than something else
5	prefer		
6	seize		

The 5000 word level

1	decent		
2	frail	____	weak
3	harsh	____	concerning a city
4	incredible	____	difficult to believe
5	municipal		
6	specific		

Figure 2.5 Two blocks of test items from the Vocabulary Levels Test (version B, from Nation 2001: 416-420).

The VLT is a receptive, mono-lingual matching task. The words on the left are ordered alphabetically and the definitions on the right in order of increasing length. In the original test, there are six blocks like the ones above, six words and three definitions, in each level of the test. The words for each level were randomly selected. In each level, out of the six blocks, 3 ended up testing nouns, 2 testing verbs, and 1 testing adjectives (Schmitt *et al.* 2001).

The test allegedly works well as an informal diagnostic tool for teachers, and Meara has called it "the nearest thing we have to a standard test in vocabulary" (1996:38). New revised forms of the test have been produced (Schmitt 1993; Schmitt 2000; Schmitt *et al.* 2001), and also a productive version (Laufer & Nation 1999). Schmitt *et al.* (2001) found two new versions (versions 1 and 2) of the test format valid as measurements of general and academic vocabulary size of L2 learners through a range of analysis techniques.

The test format obviously benefits from being relatively quick to take and administer. It also gives a learner profile in relation to the five levels and not just a rough estimate of total vocabulary size. However, both the receptive and the productive versions of the test can be criticized for their poor sample ratio. The 18 test items in the original receptive version make up less than 1 per cent of the target words at the 2000 word level. The same goes for the productive version. Another potential problem concerns the fact that the individual items in a block are not straightforwardly independent of each other. The three tested target words share the same set of distractors, and the process of answering one of the target words may involve, to a varying extent, the other target words. If certain distractors can be eliminated, then guessing would have a considerable impact on the test results. This fact was acknowledged by Beglar & Hunt (1999), who called for further study into this issue. One such study was carried out by Kamimoto (2005), who used verbal protocols to analyse the test-taker behaviour of five Japanese low proficiency students. Kamimoto concluded that elimination of distractors together with blind guessing affected these students' overall scores on the test. Based on these results, caution was advised when interpreting scores on the VLT test.

2.4.3.2.5 Vocabulary depth and its application in testing

As opposed to vocabulary size, the concept of vocabulary depth refers to various more qualitative aspects of what is known about a word. As was seen above, Anderson and Freebody (1981) described it in relation to what would be understood by an ordinary adult under normal circumstances. The ordinary adult referred to by the authors is assumed to be a native speaker. Compared to the supply of studies on vocabulary size, the concept of vocabulary depth has been more sparsely explored, but a number of studies have been carried out more recently (see e.g. Wesche & Paribakht 1996; Qian 1999; Vermeer 2001).

Read (2004) provides a thoughtful account of how the term 'vocabulary depth' has been operationalised. He acknowledges three lines of development visible in the literature. The first one is called 'precision of meaning'. It refers to the degree to which a word's meaning is known, from having a vague idea to being able pin it down more specifically and elaborately. Read argues that one problem with this operationalisation of depth of word knowledge is that words vary in the extent to which they lend themselves to exact definition. For example, the meaning, or meanings rather, of high-frequency words are notoriously difficult to define precisely. It is easier, then, to define technical words more precisely, since they do not normally display the same degree of polysemy. A case in point when it comes to the polysemy of words is provided by Bogaards (2001:324), who uses the word *party* to show the difficulty of finding a unifying meaning. Consider (9a-e) below:

(9) a. Our neighbours are throwing a party tonight.
b. They were very grateful to the rescue party.
c. The Conservative Party has lost many votes.
d. The lawyer refuted the arguments of the other party.
e. Your party is on the line.

Bogaards argues that these senses are quite different for a learner of English as a foreign language, even though some of the senses might have the same root diachronically speaking. Thus, to have a more precise knowledge of the word *party*, one should ideally know all the above senses.

The second use of vocabulary depth according to Read (2004) is captured in the term 'comprehensive word knowledge'. It refers to a view in which several different word knowledge components are involved. This is in line with word knowledge frameworks proposed by Richards (1976), and the one discussed in section 2.4.2, by Nation (2001). Read points out that an attempt to test many different word knowledge aspects of the same target word, in the same test, complicates test design, since it takes a long time to tap into informants' knowledge of a handful target words (see e.g. Schmitt 1998a). However, the view taken in this thesis assumes that it is possible to subsume certain word knowledge aspects into the measurement of others. More specifically, it is argued that receptive collocation knowledge incorporates a number of other word knowledge aspects in addition to collocation knowledge itself (see section 2.4.2).

The third view of vocabulary depth presented by Read is 'network knowledge', which refers to "the incorporation of the word into a lexical network in the mental lexicon, together with the ability to link it to – and distinguish it from – related words" (2004:212). The assumption behind this view is that words in the mental lexicon of a language user are structured through links between these words, forming some sort of network (see e.g.

McCarthy 1990; Aitchison 2003; Meara & Wolter 2004). The standard way of mapping out a language user's lexical network is through word associations. As was briefly described in section 2.4.2, word associations are normally classified into paradigmatic, syntagmatic, and phonological associations.

As has been pointed out by Read (2004), the three approaches to vocabulary depth outlined above overlap with each other. More specifically, the comprehensive word knowledge approach is seen to subsume the other two. However, it is also possible to see the comprehensive word knowledge approach as an atomistic approach, whereas the network knowledge approach can be seen as a more holistic approach. This is because the former is focused on individual words in the mental lexicon, whereas the latter focuses on the mental lexicon as a whole.

Before tests of vocabulary depth are presented, it is relevant here to briefly discuss the relation between receptive collocation knowledge as a construct and the vocabulary depth dimension. Referring to the three interpretations of vocabulary depth presented by Read (2004) and accounted for above, if we assume that depth refers to precision of meaning, which in turn refers to a range between vague knowledge and more elaborated knowledge, then it can be argued that there is a link between this view and receptive collocation knowledge. Reconsidering the data in example (9a) from above, in order to recognise that the phrase *throw a party* implies a conventionalized use of the verb THROW in English, a language user must know not only a more basic or vague meaning of THROW, corresponding to something like 'quickly letting go of an object by moving one's hand or arm', but also the extended meaning of 'arranging'. Thus, recognizing collocations require more than a vague, basic meaning of words. If we adopt the approach of comprehensive word knowledge as our point of departure, and if vocabulary size is furthermore taken to mean the number of words in a language for which a language user knows a basic meaning, then vocabulary depth involves all other word knowledge aspects beyond a basic form-meaning mapping. Collocation knowledge would then constitute one aspect of depth of word knowledge. Finally, if we assume vocabulary depth to correspond to network knowledge, taken to mean the degree to which a language user has incorporated a word into a lexical network with appropriate links to other words, then there is also reason to see a relation between vocabulary depth and receptive collocation knowledge. It should be pointed out that network knowledge is more comprehensive than receptive collocation knowledge, since it assumes a number of different kinds of relation between words.

On the whole, the above discussion strongly suggests that there are points in common between the vocabulary depth dimension and the receptive collocation knowledge construct as defined in this thesis. For this reason, a closer look at a widely used receptive test of vocabulary depth test is warranted: The Word Associates Test (WAT).

2.4.3.2.6 The Word Associates Test

The Word Associates Test (WAT) was originally developed by Read (1993), largely inspired by Meara (Read 1993:359). The test was originally intended to measure knowledge of academic vocabulary, as represented by the words in the University Word List (UWL), an 800-word compilation based on various frequency counts of academic texts. The original objective of the test was to combine the measures of size and depth by covering a reasonable number of words and at the same time measure depth of word knowledge in some meaningful way (1993:358). Read devised a test in which subjects were presented with a prompt word

together with eight possible associates, some of which are related to the prompt word and some which are not. The task of the learner is to select the words that are conceived to be related to the prompt word. An example of the structure of the task item is given below:

denominator

common	develop	divide	eloquent
fraction	mathematics	species	western

Figure 2.6 An example task item from the Word Associates Test (from Read 1993:366)

The task is essentially a recognition task since test-takers are required to select answers from set alternatives. The concept of depth of word knowledge was represented through the associates' link to the prompt word in three ways: paradigmatic relationship, syntagmatic relationship and analytic relationship. Synonymy and hyponymy were used as cases of a paradigmatic relationship, whereas collocations were used for the syntagmatic relationship. The third relationship, analytic, was seen as involving an associate which represented one aspect or component of the target word, and which was part of the dictionary definition of that target word (Read 2000:181).

After initial testing, in which verbs, nouns and adjectives were used as prompt words, Read designed a revised version containing only adjective prompts (Read 1998). The reason for this was that learners with a good knowledge of vocabulary who did not know the prompt word could find the associates by looking for semantic links among the eight possible alternatives, and thus to a great extent guess their way to a correct answer (Read 2000:183). In the revised version, the words were chosen on the basis of their multiple meanings or range of uses. The revised depth test version was aimed at measuring "the extent to which learners were familiar with the meanings and uses of a target word" (Read 1998:43). The structure of a test item in the new version looks like this:

Sudden

beautiful	quick	surprising	thirsty

change	doctor	noise	school

Figure 2.7 An example task item from the Word Associates Test (new version) (from Read 2000: 184).

The words in the left-hand box are adjectival forms and the associates among them have paradigmatic relationships with the prompt word sudden. The words in the right-hand box are nouns and the associates among them are collocates of the prompt word; thus they have a syntagmatic relationship with the prompt word. About half of the items have two associates in the left-hand box and two in the right-hand box (2 + 2). The other half of the item set has either 1 + 3 or 3 + 1. This arrangement was adopted in order to reduce the factor of guessing, but still retaining a consistent number of associates in each item.

In its revised versions, the word associates test assesses word knowledge of high-frequency adjectives, presenting 40 items like the one shown above in Figure 2.6, focusing on synonymy

and collocations. The test is monolingual and the words, claimed to represent high-frequency academic vocabulary, are presented in isolation. In terms of its qualities, high concurrent validity has been demonstrated vis-à-vis a matching format test, in which the same target words were used (r = .85). Furthermore, reliable scores have been obtained for two revised versions (Rasch reliability: .90, and .93) (Read 1998:50).

The most obvious criticism directed against the test format is that guessing might play a greater role than is acceptable for a valid measure of word knowledge (Read 1998). The number of correct responses in each item is fixed (4). It is therefore possible for a test-taker, through guessing, to obtain a large number of correct responses without knowing the meaning of the target words.

2.4.4 Reviewing empirical studies of L2 collocation knowledge

2.4.4.1 A review of the research into L2 collocation knowledge

Although learners' problems with collocations are widely attested, the overall number of studies investigating learners' command of collocation is on the whole scarce. In this section, I will first make a general review of the field of studies of L2 collocation knowledge in order to outline what we know to date about how collocations are learnt. I will then in more detail review a number of studies that are particularly important to the present thesis. These will constitute studies in which more test-like and experimental instruments are used to tap learners' knowledge of collocations. All of the reviewed studies deal with English as a foreign language, and the review is restricted to studies published no later than 2005.

There is a great deal of variation in the studies conducted into L2 collocation knowledge in terms of methods, measures, the proficiency levels and L1s of the informants, as well as the number of informants. When it comes to methods, two main approaches have been adopted: on the one hand, studies analysing learner production in an essay corpus, and on the other hand, studies in which some sort of elicitation technique is used.

2.4.4.2 L2 collocation studies based on corpora of learner essays

A number of collocation studies involve analyses of corpora of L2 essays written in English, e.g. Howarth (1996), Granger (1998), Gitsaki (1999), and Nesselhauf (2003, 2005)[14]

Howarth (1996) investigated the English academic writing of 10 MA students of linguistics and English language teaching. The students, who were seen as advanced learners, represented eight different L1s (Cantonese, German, Greek, Japanese, Mandarin, More, Thai and Tswana), with an age range of 22-40. The essays of these learners, totalling almost 23,000 words, and each essay amounting to about 2,500 words, were analysed in terms of occurrence of free combinations, restricted collocations, and idioms, all verb + noun combinations. In a comparison with native speaker (NS) data, Howarth found that the 10 non-native speakers (NNS) used more free combinations (67% versus 60%), fewer restricted collocations, (25% versus 36%), and fewer idioms (1% versus 5%) than the NSs. Howarth concludes, based on the percentages, that idioms are "an insignificant phenomenon"

[14] Other studies also exist, e.g. Wiktorsson (2003) which investigates so-called 'prefabs', and Knutsson (2006) which investigates 'multi-word expressions'. Both are analyses of a large number of different types of word sequences, not only collocations, which make their scope too wide to be included in the present review.

compared to the large number of collocations, which shows the importance of collocations for effective communication. The NNS data showed that learners often use infelicitous verb + noun combinations which are blends of two acceptable native-like collocations. Another interesting result in this study is the very low correlation observed between the use (number) of restricted collocation and general English proficiency, at r = .15. From an evaluative perspective, the number of informants in this study is small, which places restrictions on its generalisability.

Granger (1998) analysed an English learner corpus subcomponent of the ICLE corpus[15]. The learner corpus material comprised a total of about 250,000 words, and consisted of argumentative essays and literature exam papers written by L1 French informants. The investigation focused on the use of intensifying adverbs in combinations such as *perfectly natural* and *closely linked*. By automatically retrieving all words ending in *–ly*, and subsequently sorting them according to pre-defined semantic and syntactic criteria, Granger found that the NNSs on the whole underused these amplifiers compared to NS baseline data, which were gathered from a local essay corpus[16], the ICE[17], and the LOB[18], and that they used atypical word combinations. In a few cases, the NNSs overused specific amplifiers – combinations with *completely* and *totally* – which were explained as "safe bets" (1998:148) in having direct translation equivalents and in displaying few collocational restrictions. Another interesting finding was Granger's claim that the NNSs seemed to use amplifiers more as general building bricks than parts of prefabricated patterns such as collocations. This is reminiscent of Wray's (2002) argument that collocations for L2 learners can be seen as separate words which become paired, and that collocations are broken down into separate word meanings, with no information stored as to the words going together.

Gitsaki (1999) analysed essays, approximately 200 words long, on different topics written by 275 L1 Greek learners of English. The learners ranged between 12 and 15 years of age, and they were divided into three groups, classified as 1) post-beginners, 2) intermediate, and 3) post-intermediate with regard to general English proficiency. Gitsaki based her investigation on the *BBI*[19] collocation dictionary (Benson *et al.* 1997), in that the learner essays were checked for occurrences of the 33 types of grammatical and lexical collocations described in the dictionary, together with four additional types which were added by herself. Each correctly provided collocation was marked as a particular token of one of the 37 adopted types. Gitsaki found that for two types of collocations: 'SV infinitive' (example: *we must work*) and 'SV(O) that-clause' (example: *they admitted that they were wrong*), the accurate use increased with increased proficiency level. In a further comparison, it was found that the three proficiency level groups differed significantly from each other in the use of the following collocation types, i.e. that they were used significantly more often in a certain group: group 1): 'SVc' and 'Adjective noun', group 2) 'Prep noun', 'SV to Inf', 'Prep Det Noun', 'Phrasal Verb', and group 3) 'Noun Prep', 'SV Inf', and 'SV(O) that'. The type 'Verb Noun (creation)' (e.g. *reject an appeal*) was infrequently produced in the essays. A problem with Gitsaki's study is that the proficiency measure is based on the same data as that from which collocation use was investigated. Also, with 37 types, and essays of only 200 words,

[15] The International Corpus of Learner English
[16] The Louvain Essay Corpus
[17] The International Corpus of English
[18] The London-Oslo-Bergen Corpus
[19] Benson, Benson and Ilson

the mean number of collocations used for each type is very low. Only 6 types of 37 had a mean use greater than 1.0 across the 275 informants.

Nesselhauf (2005) analysed the written production of 207 advanced German L1 learners of English at university level. The corpus consisted of 318 essays, totalling around 155,000 words. Nesselhauf analysed the use of verb-noun collocations and found 2,082 tokens. The verb-noun combinations that were considered were: 'verb + object' (*wage war*), 'verb + preposition + object' (*cope with a problem*), 'verb + adverbial' (*look out of the window*), 'verb + object + complement' (*call somebody a genius*), 'verb + object + preposition + object' (*take something into consideration*), and 'verb + object + *to* + infinitive' (*force teachers to* + inf.). The average number of collocations produced per learner was 10. Nesselhauf found that two thirds of the collocations produced were considered acceptable, and consequently that one third was unacceptable or deviant, and concludes that "verb-noun collocations frequently pose problems for learners, even at an advanced level" (2005:69). The most frequent deviant element in a collocation was the verb. In terms of factors correlating with collocation difficulty, it was found that congruence, i.e. a word-for-word equivalence of a collocation in the learners' L1, emerged as the most important factor. Degree of restriction of a collocation was also found to be an important factor. Nesselhauf observed less restricted collocations – based on verbs combinable with a sizeable group of nouns, but where exceptions apply, e.g. COMMIT + [something wrong or illegal] – to be more deviant than more restricted collocations – based on verbs combinable with a small set of nouns, e.g. *fell a tree* and *shrug shoulders*. Two other findings are important: length of classroom exposure was found to have no positive effect on collocation use, whereas length of exposure to the language (length of stays in English-speaking countries) was found to have a slightly positive effect. It is a pity that Nesselhauf did not subject her data to statistical analyses, but interpreted the data rather impressionistically, a shortcoming which unfortunately places restrictions on her findings.

2.4.4.3 L2 collocation studies using elicitation techniques

When it comes to studies in which some sort of elicitation technique is used, we find firstly two of the above reviewed studies, namely Granger (1998) and Gitsaki (1999). Other relevant studies include Biskup (1992), Bahns & Eldaw (1993), Farghal & Obiedat (1995), Bonk (2001), Mochizuki (2002), and Barfield (2003). The techniques used include translation from the L1 into English, cloze formats, and receptive multiple-choice tests. Since the present thesis is concerned with the development of test formats, studies in which more test-like instruments are used to tap learners' knowledge of collocations are of paramount interest. I will therefore review the studies in Bonk (2001), Mochizuki (2002), and Barfield (2003) in more detail than the other studies. I will first, however, account for the other enumerated studies.

In addition to her analysis of learners' written production, Granger (1998) also carried out a study in which 56 NSs and 56 NNSs (L1 French) of English were given a questionnaire consisting of 11 amplifiers (*highly, seriously, readily, blissfully, vitally, fully, perfectly, heavily, bitterly, absolutely,* and *utterly*), each followed by a list of 15 adjectives (*significant, reliable, ill, different, essential, aware, miserable, available, clear, happy, difficult, ignorant, impossible, cold,* and *important*) and were instructed to choose acceptable collocates among the adjectives. The informants were asked to mark particularly strong collocates with an asterisk. The NNSs marked considerably fewer combinations than the NSs (280 versus 384), with examples like *readily available* and *bitterly cold* marked by 43 NSs versus 8 NNSs, and

40 NSs versus 7 NNSs, respectively. Granger explains this by a "weak sense of salience" on the part of the learners (1998:152).

Gitsaki's (1999) study involved the same 275 learners as reported in the account of the essay-based study above, and her elicitation technique consisted of a) a cued translation task, with 10 sentences containing collocations (six types) to be translated from Greek into English, and b) a blank-filling task, with 50, 65, and 90 (for the three proficiency groups) English sentences containing specific collocations (eleven types) with one part missing. All the targeted English collocations were non-congruent with their Greek equivalents, and they were taken from textbook material used in all junior high schools in Greece. The main findings from the elicited data show that 'SVc' collocations (e.g. *he was a teacher*) are "core" collocations (1999:141) in that they were the most frequently used collocations by learners at all three proficiency levels. Furthermore, the 'SVc' type together with 'Adjective Noun' (*strong tea*) seem to be acquired early, whereas 'Noun that' (*he took an oath that he would do…*), 'SV Possessive V-ing' (*they love his clowning*), 'SVOO' (*she asked the pupil a question*), 'S(it)VO *to* Inf' (*it surprised me to learn of her decision*), 'Verb Noun (eradication)' (*reject an appeal*), and 'Adverb Adjective' (*deeply absorbed*) were avoided by all learners. Gitsaki argues that these types are structurally demanding, infrequent and/or fixed, and stresses the fact that the type 'Verb Noun (creation)' of lexical collocations (e.g. *make an impression*) was the most difficult to translate with accuracy, and to get right in the blank-filling test. Gitsaki also concludes that the results on the elicitation tests show that collocation knowledge develops as L2 learners' overall language proficiency develops. A shortcoming of this study is the fact that different sets of items were tested on the three different proficiency groups, and also different number of types of items. This makes comparisons between the groups less straightforward.

Biskup (1992) investigated how well a total of 62 Polish and German university students, considered to be very advanced students, translated verb + noun and adjective + noun (lexical) collocations from their respective L1s into English. Biskup chose to test production because she found in another ongoing study that "perception" meant no visible difficulty "since collocations are fully transparent" (1992:86). She found that the two groups produced the same mean number of correct responses, but with more restricted collocations produced by the Polish learner group than German group. Also, the Polish learners more often refrained from answering, whereas German learners supplied more paraphrases, results which Biskup takes as evidence of the German learners being more prone to risk-taking. From an evaluative perspective, it is not clear how many items were tested, or if the tested items were decontextualised items, sentences, or full texts with underlined items. By and large, the lack of clearly presented details about the items and the test instruments makes it difficult to fully evaluate Biskup's findings.

Bahns & Eldaw (1993) aimed at testing learners' productive knowledge of 15 verb + noun collocations. A total of 58 German university students of English, in years 1-3, participated in the study. Of these 58 subjects, 34 were given a translation task in which 15 German sentences were to be translated into English, and 24 subjects were given a cloze format in which the target collocations were inserted into English sentences with the verb collocate of a noun missing. The 15 verb + noun collocations were selected from various sources, such as learning materials and dictionaries, and were pre-tested on 2 native speakers as a validation measure. The subjects' answers were rated as acceptable or unacceptable by 3 native speakers. In terms of main findings, no significant differences were found between the two

groups as to the mean number of correctly answered items, 7.2 for the cloze group and 8.1 for the translation group, respectively. Bahns & Eldaw also concluded that collocation knowledge does not develop alongside general lexical knowledge. The conclusion was based on an analysis in which the measures of two assumed variables, general vocabulary knowledge and knowledge of collocations, were taken from the same data, and thus not independent. The analysis entailed taking the percentage of felicitously translated single lexical words in hypothetically ideal translations (83 lexical words x 34 students) and comparing this with the percentage of felicitously translated verbal collocates. On a critical note, the number of items tested in this study is fairly small. Also, since the measures of general vocabulary and collocation knowledge were taken from the same data, the conclusions drawn in this study cannot be seen as sufficiently robust.

Farghal & Obiedat (1995) conducted a study aimed at testing learners' knowledge of 22 common English collocations. A total of 57 L1 Arabic university students of English were tested, divided into two groups: A and B. The two groups were given separate tasks. Group A took an English fill-in-the-blank test with 11 items. In each item, one member of a collocation pair was given, and one was missing, and meant to be supplied. Group B took a test in which Arabic sentences were supposed to be translated into English. This test was based on the same target collocation material as the fill-in-the-blank test. The targeted collocate pairs were validated by two native speakers of English. In terms of results, the informants supplied a correct collocation in 18% (group A) and 5% (group B) of the cases. Farghal & Obiedat found that 4 lexical simplification strategies were used among the informants. The use of synonymy was the most frequently used strategy by both groups when a correct collocation was not produced (group A = 41%, and group B = 35%), followed by that of avoidance (27% and 21%). The two other strategies identified were L1 transfer (10% and 13%) and paraphrasing (4% and 25%). The main conclusion drawn in the study is that L2 learners cannot cope with collocations because "they are not being made aware of collocations as a fundamental genre of multi-word units" (p.326). Farghal & Obiedat claim that vocabulary is taught as single lexical items, something that leads to lexical incompetence on the part of the L2 learners. The number of items tested in this study is fairly small, and it is not clear how the test items were selected. Furthermore, the study seems to rest on the assumption that there is a self-evident relation of antonymy between the collocations used, an assumption that is scarcely tenable.

2.4.4.4 L2 collocation studies using test-like elicitation techniques

In this section, three studies will be reviewed in more detail since they are central to the development of test instruments in the present thesis.

Bonk (2001) reports a study whose main aim was to investigate the reliability and validity of a test instrument, and to correlate collocation knowledge with general English proficiency. A total of 98 university students, a majority of whom were L1 speakers of East-Asian languages, were subjected to a test battery consisting of 3 subtests of collocation knowledge and a general English proficiency measure. The subtests used were the following: a) a 17-item prompted recall verb+object collocations test of English sentences, each with a gap for a verb to be inserted, b) a 17-item prompted recall verb+preposition collocation test, also with English sentences, but each with a gap for preposition to be inserted, and c) a 16-item receptive test of figurative use of verb phrases, consisting of multiple-choice items with 4 sentences in each. The task for the testee was to judge which one of the four sentences did not contain a correct usage of the verb. Finally, d), a 49-item general language proficiency

measure was administered in the form of a 49-item condensed TOEFL test. Examples of items in the three collocation subtests are given in (10), (11) and (12) below:

(10) Punk rockers dye their hair red and green because they want other people to _____ attention to them.
(11) Many of the birds in the area were killed _____ by local hunters. (to exterminate)
(12) a. Are the Johnsons throwing another party?
b. She threw him the advertising concept to see if he liked it.
c. The team from New Jersey was accused of throwing the game.
d. The new information from the Singapore office threw the meeting into confusion.

The test battery was validated by administration to 10 native speakers. A total of 98 students participated in the main test administration. The students scored a mean of 25.3 (SD 7.3) out of 50 on the collocations test total, and their mean scores on the 3 subtests were close to 50% of the maximum score of the respective tests (8.7, 8.8 and 7.8) Their total mean score on the 49-item TOEFL test was 37.3 (SD 7.2). A Kuder-Richardson 20 analysis of internal consistency showed that the scores on the collocations test were reliably measured at .83. One of the subtests, however, the verb+preposition test, was found to yield a rather low and unacceptable reliability value at .47.

Bonk also carried out item analyses including item facility and item discrimination indices, and point-biserial coefficients[20]. These analyses showed that a majority of the items functioned as good, well-discriminating items. The mean item facility[21] for the three subtests was around .50, and the mean point-biserial correlation was .38, .27, and .34 respectively for the three collocation subtests. In terms of main findings, based on an Item Response Theory (IRT) Rasch analysis, and a Generalisability analysis, Bonk concluded that the 50-item collocations test worked well on the whole for the population, but that subtest 2, the verb+preposition test, was a somewhat weak link and that it could practically be discarded in favour of extending subtests 1 and 3.

Bonk found a moderately high level of correlation between general English proficiency and collocation proficiency (.73 after correction for attenuation). No instances of low proficiency and high collocation scores could be found, and no instances of high proficiency and low collocation scores either, although the middle range of scores displayed some variation.

One of the advantages of Bonk's study is the attempt to include a larger number of items (*k* = 50). He also subjected his data to rigorous statistical analyses through which he attempted to support his conclusions. If several variables are to be compared and correlated with each other, it is important to show that these variables were reliably measured.

On a more critical note, the task formats used by Bonk involve a fair bit of reading, and this raises the question of what is really measured. It could be the case that the subjects did not understand the sentence prompts and therefore did not answer an item correctly. If so, the test is more a measure of reading comprehension than collocation proficiency. Admittedly,

[20] Point Biserial methods correlate binary item scores (0, 1) with continuous total scores on a test. As with Discrimination Indices, Point Biserial correlation coefficients indicate how well an item discriminates between test-takers with high total scores and test-takers with low total scores on a test (see Henning 1987)

[21] Item facility denotes the degree of facility of a test item which is calculated on the basis of a group's test performance (Davies *et al.* 1999)

Bonk tried to control for this by qualitatively examining 25% of the answer sheets, finding that the subjects seem to have understood the prompts "the great majority of the time" (p.134). A further weakness is the unsystematic selection of test items, which seems to have been made on the basis of intuition only.

In Mochizuki (2002), 54 Japanese first-year university students, majors in German, Chinese, or Japanese, were tested on collocation knowledge, paradigmatic knowledge and overall vocabulary size. The aim of the study was to explore how Japanese learners of English develop two aspects of word knowledge, paradigmatic and collocational, and vocabulary size over one academic year. Over this period of time, the students received 75 hours of instruction (reading and conversation classes). The tests used were the following: a) a vocabulary size test, an adaptation of the Vocabulary Levels Test (Nation 1990, 2001), in which the task involved matching English words with Japanese translation equivalents, in Mochizuki's version 7 levels corresponding to 7 frequency bands, and b) a test of paradigmatic knowledge of 72 English words in a 4-choice format, and c) a collocation test of 72 words, the same words as in task b), also in a 4-choice format. Examples of subtests b) and c) are provided in (13) and (14), respectively, below:

(13)	job	(1) date	(2) sort	(3) star	(4) work
(14)	job	(1) answer	(2) find	(3) lay	(4) put

The task for the informant was to decide with which of the four alternatives there is a possible link – a paradigmatic one in the case of the paradigmatic knowledge test (13), and a syntagmatic one in the case of the collocation knowledge test (14). The target words in the tests were divided into four groups of 18, and each group consisted of six nouns, six verbs and six adjectives, all randomly selected, taken from one out of four word lists based on frequency counts. In terms of internal reliability of the test instruments, the values calculated (Cronbach's alpha) were .71 and .75 for the two administrations of the paradigmatic knowledge test, and .54 and .70 for the two administrations of the collocation knowledge test, which Mochizuki concludes to be moderately reliable.

When comparing the results obtained at the two administrations (April=T1 and January=T2), Mochizuki found that only in the case of the collocation test was a significant difference observable (41.7 (SD 5.4) at T1, and 42.8 (SD 6.4) at T2). The very modest lack of increase over the two administrations is explained by lack of motivation on the part of the learners. Following an argument advanced by Schmitt (1998), Mochizuki furthermore explains the fact that over time there was a significant increase in collocation knowledge, and not in vocabulary size and paradigmatic word knowledge, by the inherent inertia of knowledge of meaning. It is assumed that a learner's knowledge of word meanings does not change radically over time, whereas knowledge of syntagmatic relationships does.

As with Bonk's study described above, Mochizuki's study attempted to test a larger number of items (k = 72), which is positive. Also, values of internal reliability were reported, even though no reliability values were given for the vocabulary size measure. One administration of the collocation knowledge test showed a relatively low value of α .54. The value might be partially explained by the rather homogeneous group of learners taking the test. Homogeneous group scores generally result in low internal reliability values, since the calculation relies on a certain amount of variance (see Brown 1983:86). In contrast to Bonk's study, decontextualised items were used. An analysis missing in the study, I think, is a

correlation measure. It would be interesting to correlate the vocabulary size variable with the paradigmatic knowledge and collocation knowledge variables, respectively, to see whether and how these word knowledge aspects are interrelated.

Barfield (2003) reports a study aimed at testing a large number of decontextualised verb + noun collocations for recognition, and at comparing recognition patterns with those of the single verbs and nouns. A total of 93 Japanese university students participated in the study. They were undergraduates and post-graduates belonging to 4 different fields of study. A test instrument was created by taking 40 lexical verbs from a previous study. These verbs were taken from the Academic Word List (AWL) (Coxhead 2000), and the General Service List (GSL) (West 1953). As a second step, 3 noun collocates were chosen for each of the 40 verbs, based on data in the Cobuild Bank of English. Furthermore, 20 so-called 'mis-collocations' were created, intuitively, mainly based on other verbs' collocates. This was done as a means of checking the reliability of the test instrument. The result was a 120-item test consisting of 100 'real collocations' and 20 'mis-collocations'. The learners were presented with the test items and were asked to rate each collocation on a 4-state scale, as shown in Figure 2.8 below.

I	I don't know this combination at all.
II	I think this is not a frequent combination.
III	I think this is a frequent combination.
IV	This is definitely a frequent combination.

Figure 2.8 A 4-state scale of reported knowledge of verb+noun combinations, from Barfield (2003)

It is not clear exactly how the tested items were presented to the learners, but examples of the tested items are *adopt + approach*, *adopt + child*, **adopt + profit*, *break + ground*, *break + record*, and *break + rules* (asterisk indicates mis-collocation). Barfield first tested the learners' recognition knowledge of the 120 nouns and the 40 verbs, using a similar but slightly differently worded rating scale than that above. He found that the recognition of nouns was very high, with a mean score of 3.87 (SD .079) out of 4. The mean for verb recognition was also high, observed at 3.56. As for the verb+noun collocation test, the mean recognition for the total number of collocations was 2.56 (SD .39) out of 4.

Barfield argues that the results suggest that knowledge of individual verbs and nouns does not necessarily entail recognition of their combination in a verb + noun collocation. Looking at the recognition scores of the 100 real collocations, no significant differences were found between the group mean scores. Barfield found that these scores showed high reliability as measured by Cronbach's alpha (α = .97), and that reliability was high also for the mis-collocations (α = .93).

In terms of correlations with general proficiency, Barfield observed a relation between the recognition levels of the verbs and the nouns individually, but no correlation was established between general English proficiency and collocation recognition in Knowledge State 4. With one exception, all of the nouns and verbs of the top 20 most recognized collocations, e.g. *change mind*, *protect body*, *protect environment*, *explain reason* and *govern country*, were within the 3,000 most common words of English according to frequencies in the British National Corpus (BNC), which leads Barfield to conclude that the relative frequency of the

single words making up a collocation is a supporting factor in collocation recognition. Looking further at the 20 most recognized collocations, core sense in both the verb and the noun seemed to figure highly as the primary deciding factor (11 items). Another factor seemed to be the combination of an abstract noun + a verb in its core sense (8 items). The remaining collocation residing in the top 20 was a verb in specialised sense + concrete noun. Based on these findings, a 4-way division of semantic transparency for collocational recognition is suggested (2003:45), in which field 1 is suggested to be the easiest and field 4 the most difficult for learners. The 4-way division is shown in Figure 2.9 below.

		NOUN	
		CORE	NON-CORE
VERB	CORE	1) Semantic transparency in both components	2) semantic transparency driven by abstract noun
	NON-CORE	3) Verb in specialised sense with core noun	4) Semantic opacity in both components

Figure 2.9 A 4-way division of semantic transparency for recognition of collocations, taken from Barfield (2003:45).

Barfield's study is yet another example of efforts to use a large number of items. The selection of items is systematic, and the 4-state scale of knowledge used is interesting, since word knowledge is not an all-or-nothing type of knowledge. Another interesting feature is the fact that recognition of the constituent parts of the collocations, the single verbs and nouns, is tested. This is good since learners' claimed level of knowledge of a collocation may depend on their knowledge of the parts of the combination. On the minus side can be noted the fact that some of the mis-collocations are possible in certain contexts, a shortcoming admitted by the author. Examples of these are *explain address, approve opportunity* and *create temperature*, all of which could be rather feasible combinations, conditioned by the insertion of one or more lexical items in-between and around the verb and the noun: *to explain an address to someone*, *to approve of a job opportunity*, and *to create a temperature at which certain solid elements become liquid*. A final observation concerns the fact that no delexical verbs were used. It is noted in the literature that delexical verbs, such as *make, take, do, give* and *have,* occur frequently in English and that native-like, productive use in particular challenges learners, even at advanced levels (Källkvist 1999, Altenberg & Granger 2001, Nesselhauf 2005). For this reason, investigating learners' knowledge of collocations in which delexical verbs appear seems to be warranted.

2.4.4.5 Summarizing findings from the reviewed studies

The key characteristics of the above reviewed studies are summarized in Table 2.7 overleaf. From the review, a number of interesting trends have emerged that are relevant to the present thesis. Firstly, few studies have been carried out investigating learners' receptive knowledge of collocations. Most studies reviewed entailed analyses of learners' production. Biskup (1992) even argues that perception is unproblematic for learners, and that collocations are fully transparent. It is not clear that this is the case, and more empirical support is warranted for these claims.

Secondly, in the few studies that do exist, often a rather small number of items are tested, usually 10-20, with the exception of the last three reviewed above (Bonk 2001; Mochizuki 2002; Barfield 2003). The drawback of using few test items is that it is not possible to draw well-founded conclusions, especially so when item selection is made in an unsystematic way, or not described at all.

Thirdly, verb + noun (or verb + NP) collocations have been investigated to a fair extent, but it is quite clear that these word combinations are problematic to learners, even when the individual verbs and nouns are known.

Fourthly, reliability values of the test instruments *per se* are seldom reported. Again, the three studies by Bonk (2001), Mochizuki (2002), and Barfield (2003) are exceptions to this trend. Especially when different variables are compared, it is essential that the operationalised measures of the variables, i.e. the scores, show a decent degree of reliability. If too high a percentage of a score is marred by unsystematic variance, inconsistencies, not attributable to the underlying language ability of the test-taker, then less trust can be placed in any conclusions drawn from the score. As pointed out by Bachman: "in order for a test score to be valid, it must be reliable" (1990:160). Reliability is thus a necessary condition for validity.

Fifthly, the answer to the question whether collocation knowledge is closely related to general proficiency is inconclusive. In some studies, a clear relationship has been observed (Gitsaki 1999; Bonk 2001), whereas in other studies, no relationship was established (Howarth 1996; Barfield 2003).

Sixthly, and finally, with the exception of Gitsaki (1999), none of the studies reviewed compare learners at different learning levels when it comes to collocation knowledge. This means that we do not have clear picture of whether collocation knowledge increases as a function of higher level of study.

Table 2.7 Summarizing key characteristics of the reviewed studies investigating L2 collocation knowledge

Study	Method	Informants	Investigated collocation items/types	Findings/arguments/conclusions
Biskup (1992)	-L1 > L2 translation	34 L1 German university students of English 28 L1 Polish university students of English	-? Lexical collocations: V + N, Adj + N	-perception of collocations is unproblematic for learners -collocations are fully transparent -closeness between L1 and L2 important
Bahns & Eldaw (1993)	-L1 > L2 translation -L2 sentence cloze	58 L1 German university students of English	-15 Lexical collocations: V + N	-collocation knowledge does not develop alongside general lexical knowledge
Farghal & Obiedat (1995)	-L1 > L2 translation -L2 sentence cloze	34 L1 Arabic university students of English 23 L1 Arabic university teacher students of English	-22 Lexical collocations: Adj + N, N + N	- lexical simplification strategies were used extensively among the informants -L2 learners cannot cope with collocation and there is a lack awareness of collocations as a fundamental genre of multi-word units
Howarth (1996)	-Analysis of L2 essays	10 university students (different L1s) of linguistics and English language teaching	-Lexical collocations: V + N	-Learners use fewer restricted collocations than NSs -Learners' use of infelicitous V + N combinations are often blends of two acceptable native-like collocations. -No correlation between general proficiency and collocation use (r= .15)
Granger (1998)	-Analysis of essays (corpus) -L2 receptive recognition test	56 L1 French (university?) students of English (+56 NSs of English)	-Lexical collocations: Adv + Adj -165-item test: Adv + Adj	-Learners underused amplifier adverbs compared to NS baseline data -Learners seemed to use amplifier adverbs more as general building bricks than parts of prefabricated patterns such as collocations -Learners marked considerably fewer combinations than the NSs -Learners have a weak sense of salience
Gitsaki (1999)	-Analysis of L2 essays -L1 > L2 translation	275 L1 Greek high-school students (yrs 1, 2, and 3)	-37 types of grammatical and lexical collocations	-collocation knowledge develops as L2 learners' overall language proficiency

	-L2 sentence cloze			develops
Bonk (2001)	-L2 sentence cloze -L2 receptive recognition test	98 university students (different L1s) of different subjects	-50-item test lexical and grammatical collocations: V + N, V + prep, (fig. use of verb)	-correlation observed between general English proficiency and collocation proficiency (r = .73)
Mochizuki (2002)	-L2 receptive recognition test	54 L1 Japanese university students of different subjects	-72-item test lexical collocations: V + N, Adj + N, N + N	-a learner's knowledge of word meanings does not change radically over time, whereas knowledge of syntagmatic relationships does
Barfield (2003)	-L2 receptive recognition test	93 L1 Japanese university students of medicine, area studies, environmental studies, and humanities	-120-item test lexical collocations: V + N	- knowledge of individual verbs and noun does not necessarily entail recognition of their combination in a verb + noun collocation -No correlation between general proficiency and collocation knowledge
Nesselhauf (2005)	-Analysis of written L2 essays (corpus)	207 L1 German university students of English	2,082 (tokens) Lexical collocations: V + N	-Two thirds of the produced collocations were considered acceptable -verb-noun collocations frequently pose problems for learners, even at an advance level, and the most frequent deviant element in a collocation was the verb. -Factors correlating with collocation difficulty were L1-L2 congruence, and degree of restriction -Length of classroom exposure had no positive effect on collocation use -Length of exposure to the language (length of stays in English-speaking countries) had a slightly positive effect.

2.5 Test Theory

2.5.1 Introduction

This section gives an account of important considerations in test construction and test evaluation. Central aspects of testing are discussed, such as construct, reliability, and validity. Anyone familiar with the field of language testing, and the above central aspects may skip this section (2.5).

In the process of constructing any language test, there are a number of important steps to take along the way. McNamara (2000) draws an analogy between the test development process and that of the car company getting a new car on the road. The process of producing both products involves a design stage, a construction stage, and a try-out stage before the product is fully operational. McNamara notes, however, that the linearity that this suggests does not fit the nature of the test development process. Rather, a cyclic process characterizes it, in the sense that the use of the test produces evidence of its qualities. Before dealing with the intricacies of these stages, though, we need to first define what a test is. Carroll provides the following definition (1968:46):

> a psychological or educational test is a procedure designed to elicit certain behaviour from which one can make inferences about certain characteristics of an individual.

A measurement is the process of quantifying this behaviour or knowledge of test takers, and it involves the use of a test instrument calibrated on some kind of scale (Davies *et al.* 1999:118). Often in language tests, the ability being measured is done so indirectly. For this reason, it is essential that we define what it is we set out to measure. Only then is it possible to carry out various analyses in an attempt to show that our test is a functional and good test. Since it stands to reason that test takers' knowledge or command of English collocations is a mental ability, we need to pin-point this ability as a so-called construct.

2.5.2 Construct

The term 'construct' is primarily a psychological term, but is used extensively in language testing (see e.g. Chapelle 1998; Alderson *et al.* 1995; Bachman & Palmer 1996). According to Davies *et al.*, a construct is a trait that a test is intended to measure. More specifically, it is "an ability or set of abilities that will be reflected in test performance, and about which inferences can be made on the basis of test scores" (1999:31). Thus, if we are to construct a test of collocation knowledge, we must, in as detailed a manner as possible, define what it is we intend to measure.

As was pointed out in subsection 2.4, Bachman (1990) recognises the need for a three-stage analysis in this respect. Firstly, the construct needs to be defined theoretically. Secondly, the construct has to be defined operationally, and thirdly, procedures must be established for the quantification of observations. The theoretical definition is a specification of the relevant characteristics of the ability we want to measure, and its distinction from other similar constructs. If there are several subcomponents to a construct, then the interrelations between these must be specified. When it comes to the operational definition of the construct,

this process involves attempts to make the construct observable. To a great extent, the theoretical definition will govern what options will make themselves available. For example, the theoretical definition of the construct 'listening comprehension' suggests an operationalisation as a task in which information must be decoded aurally in some fashion. With respect to the third stage, our measurement should be quantified on a scale. In general, four different types of scale are acknowledged in measurement theory: nominal, ordinal, interval, and ratio scales (see Heiman 2006). Depending on the nature of the ability being measured, one of these will prove more or less appropriate. Ideally, ratio scales provide the largest amount of information, but it is not always possible to apply it.

We turn next to two most essential ingredients of any language test: reliability and validity, respectively. The discussion will be restricted to norm-referenced[22] tests due to the nature of the tests investigated in this thesis.

2.5.3 Reliability

As has been pointed out by Jones, the word 'reliability' evokes in its everyday sense powerful positive connotations (2001:1 [cited in Weir 2005:22]). In general, something that is reliable is good. A reliable device will behave in an expected way. The meaning of reliability in testing is clearly linked to its everyday meaning. However, there are more or less technical definitions of the term. Starting with one of the more straightforward definitions, Lado presents it in the following way (1961:330):

> Reliability has to do with the stability of scores for the same individuals. If the scores of students are stable the test is reliable; if the scores tend to fluctuate for no apparent reason, the test is unreliable.

Lado's view hints to a common characteristic of reliability: the fact that it is reflected in a test's power to rank-order test takers consistently according to their comparative true abilities across two test administrations. This means that the same test given twice (identical content) to the same individual should produce the same or a very similar score, provided that the ability measured in the test does not change in the time between the administrations. This is often referred to as a test's stability or 'test-retest reliability' (see Field 2005). A test that produces a great deal of variability in test scores or large distances between test takers' scores is less likely to have extensive exchanges of positions between test takers on an ability continuum (Henning 1987). A straightforward way to illustrate the concept of reliability (sometimes called 'consistency') is the following: If we ask a person to stand on a typical bathroom scale and note her weight, we expect her to weigh the same, under the same conditions, if repeating the procedure ten minutes later. If this does not happen, we might suspect that something is wrong with the measurement instrument: the scale. It would in the case of different results be a unreliable instrument.

In slightly more technical terms, reliability is the absence of measurement error. Davies *et al.* (1999:168) define reliability as "The actual level of agreement between the results of one

[22] In norm-referenced tests, a test-taker's scores are interpreted with reference to the performance of the other test-takers, in the light of the spreading of individuals along an ability continuum. Another approach is criterion-referenced tests, which are concerned with the nature of the task to be attained (Davies *et al*. 1999: 130).

test with itself or with another test. Such agreement, ideally, would be the same if there were no measurement error, […]". All measurements are more or less subject to inaccuracies. In any test, therefore, the goal is to minimize error and subsequently to maximize reliability. In a language test, the goal is for test-takers' underlying language abilities to be reflected in the test scores to as great an extent as possible. Conversely, factors other than those underlying abilities must have as little impact as possible on the test scores. Generally, two kinds of analysis are involved in the estimation of reliability: logical and statistical (empirical) analysis. Thorough logical analyses of a test can be supported through statistical analyses. Within the framework of Classical Test Theory (CTT), methods have been developed for the estimation of how reliable the test scores of a test are (see e.g. Bachman 2004). Since this thesis has a language testing focus, and the fact that I will be using these methods extensively in the subsequent chapters of the thesis, a presentation of the basic assumptions behind the methods is warranted.

The test scores of a group of test takers will display a certain amount of variance. Variance is a measure of variability, and as such it describes the extent to which scores in a distribution differ from each other. Variance is the average of the squared deviations of scores around the sample mean (Heiman 2006:93). Bachman (1990:350) proposes that the variance in the scores of a language test can be classified into four categories. As can be seen in Figure 2.10, in addition to a) the language ability we set out to measure, language test score variance may be due to b) 'personal characteristics', c) 'random factors', and d) 'test method', respectively.

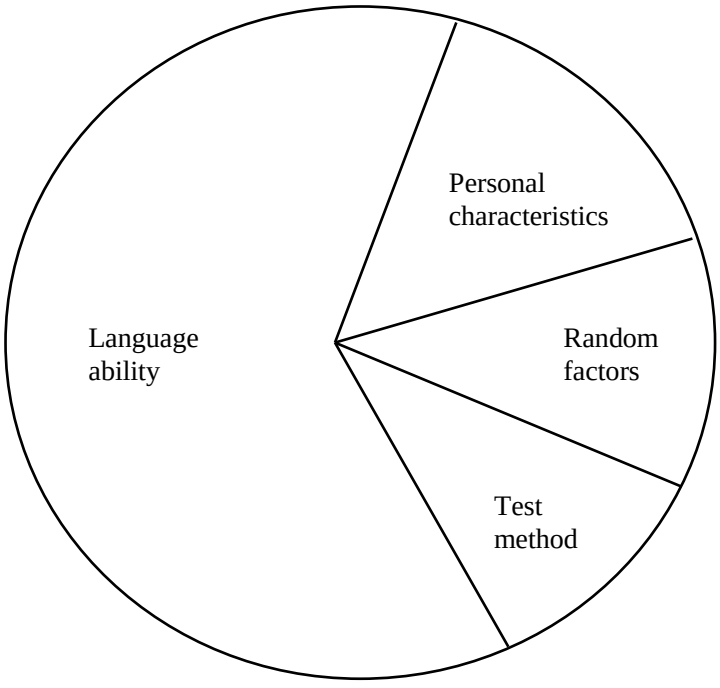

Figure 2.10 Sources of variance in test scores (from Bachman 1990:350)

Personal characteristics include attributes like age, gender, background knowledge, and cognitive abilities. These attributes are relatively stable and the variance stemming from them is 'systematic' since two individuals who differ in terms of these factors will perform differently on a test (Bachman 2004:156). Random factors refer to irregularities in test

administrations, e.g. technical problems with test tools or misprints in test questionnaires. They may also be conditions that affect test takers' performance, such as exhaustion, test fatigue, lapses in concentration, illnesses or emotional discomforts. These factors are 'unsystematic' since they may or may not affect the performances of different individuals. Finally, there are test method factors, such as the format of the test, e.g. multiple-choice or essay. Some individuals perform better on multiple-choice tests than essay-like tasks. The variance related to these factors is systematic.

In the CTT model, one basic assumption is that observed test scores consist of two components: a 'true' score component and an 'error' score component. The 'true' score reflects the underlying ability of an individual, and the 'error' score is due to factors other than the ability tested. In a similar vein, the variance of a set of test scores may be divided into observed score variance, true score variance, and error score variance. We should note here that this model collapses Bachman's three additional factors from above into one. These assumptions are illustrated in (15) and (16) below:

(15) $x = x_t + x_e$
(16) $s^2x = s^2_t + s^2_e$

In (15), x stands for the observed score, x_t the true score, and x_e the error score. In (16), s^2x stands for observed score variance, s^2_t true score variance, and s^2_e error score variance. Thus, reliability is seen as the proportion of the test score variance that is 'true score' variance (Bachman 2004:158). However, since there is no way of determining how big the true score variance is, this more theoretical definition has to be operationalised. We do this by postulating that reliability is the correlation between two sets of parallel scores. The logic behind this is that if we administer a test at least twice to the same group of test takers, we would expect them to score very similar results during the two test occasions. If they indeed do this, their respective pair of scores will display a high degree of correlation. This approach is called the test-retest approach and it provides a good way of establishing reliability of a test. However, it might not always be practicable to do so. For example, test takers might not be available for a second administration. Another problem is that a practise effect might distort the scores obtained in a retest.

The present thesis uses a reliability coefficient of internal consistency. This method allows us to compute a reliability estimate based on just one test administration. The specific type of coefficient to be used is Cronbach's alpha (Cronbach 1951). This computation, often designated Cronbach's α, is essentially a measure of scale reliability. It splits data (e.g., scores on a test) in two in every possible way and computes the correlation coefficient for each split, after which an average is computed of all the possible split values (Field 2005). Another way to see the computation is that the variance for each test item is related to the total variance for the test (all items). Because the coefficient is derived from item intercorrelations, it is the actual items in the test that are the primary source of error. The formula for Cronbach's coefficient α is given in Figure 2.11 below:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s^2_i}{s^2_x}\right)$$

Figure 2.11 Formula for the computation of Cronbach's coefficient α (Bachman 2004:163)

In the formula in Figure 2.11, k is the number of items in a test; $\sum s^2_i$ stands for the sum of the item variances; and s^2_x is the variance of the test scores (the scores on all k items). Coefficient values should in general be as high as possible. Those of .70 - .80 are often reported as acceptable (Field 2005), and those of .85 - .90 desirable and common (see Brown 1983).

Brown (1983) points to a number of factors that might influence the reliability coefficient of a test. Test length, firstly, will have an impact on reliability. Generally, longer tests are more reliable than shorter ones. This means that a vocabulary test of merely 10 items will most probably be less reliable than a vocabulary test of 50 items. The reason for this is that as the number of items increase, random measurement errors like lapses of concentration or blind guessing on the part of the test taker have a tendency to cancel each other out. Consequently, the observed scores will in a better way approximate true scores.

Secondly, the range of scores obtained in a test administration will have an effect on reliability. On the one hand, the scores of a very homogeneous group, homogeneous in terms of their underlying ability, will display a lower degree of variance. As a result, the reliability coefficient decreases. On the other hand, scores from a heterogeneous group of test takers will produce a greater degree of variance which will in turn increase the value of the reliability coefficient.

A third factor that affects reliability is the difficulty of the test. This factor can be seen to have links to the previously described factor. If a test is very easy for a group of test takers, then most of them will get the items in the test right. This will produce a so-called 'ceiling effect' (see Davies *et al.* 1999), and the result is that the test does not discriminate adequately among higher ability informants. Conversely, a 'floor effect' will be present if a test contains too many difficult items. The reason why these cases decrease reliability is that they in all probability narrow the range of the scores, which in turn results in low score variance. Ideally, then, norm-referenced tests should overall have a medium level of difficulty for the targeted subject group (cf. Klein-Braley 1991:81). Brown (1983:87) asserts that the largest variance occurs when the probability of obtaining a correct response on a test item is .50, i.e. when half of the test takers get the item right.

Fourthly, and finally, reliability figures for speeded tests are not appropriate. A speeded test is a test that provides too short a time limit for most test takers (Davies *et al.* 1999:183). If a majority of the test takers do not answer the items at the end of a test, then those items will display zero variance. Any correlation between these zero variance items and the other items of the test will be low.

Before we turn to the aspects of validity, an important point must be made pertaining to reliability. Very often, reliability values "of a test" are reported in the literature. There is a

misconception inherent in this language use. As is clarified by Bachman, citing the American Psychological Association,: "Reliability is a quality of test *scores*, and a perfectly reliable score, or measure, would be one which is free from errors of measurement (American Psychological Association 1985)" (1990:24). The reason for this is that a test may behave differently with different test taker groups. This is especially so when the test taker groups are very different with respect to their underlying ability. As a result of this, a language test must during its development phase be administrated to the type of learner group or groups for which it is eventually planned to be used. Only then will reported reliability values be relevant. It also follows that the use of an existing standardized test with a test taker group which is very different from the one specified in the test specifications will in all likelihood produce deviating reliability values.

In the present thesis, establishing that the scores on the investigated tests display a high level of reliability, within the above presented framework, is paramount since the intention is to investigate and compare test takers' performances on these tests with other tests and measures. An unreliable set of scores cannot be consistently related to other variables (Brown 1983:70).

2.5.4 Validity

The fact that I approached reliability before validity is not a coincidence. The reason is straightforward: there can be no validity without reliability. In Bachman's words, "When we increase the reliability of our measures, we are also satisfying a necessary condition for validity: in order for a test score to be valid, it must be reliable." (1990:160). Thus, if evidence of a test's reliability can be established, we have come a long way. However, there is no bi-directional relation between the two concepts: a reliable test is not automatically a valid test and vice versa. The validity associated with a test must therefore be investigated. As was seen to be the case for reliability, validity too involves both logical and empirical investigation. Furthermore, validity is not an all-or-nothing matter, but rather subject to degree (see Alderson *et al.* 1995; Messick 1989). This means that a test can for example be more or less valid for use with a certain test taker group under certain conditions.

Validity as a concept is generally treated in two ways in the literature. Either it is treated as a unitary concept, or, as is the more traditional way, it is seen to consist of several subcomponents. When treated as a unitary concept, the following definition by Messick (1989:13) is widely quoted:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment.

From this follows that it is not actually the test itself that should be validated, but rather the inferences drawn from test scores. Messick furthermore argues that test scores are a function not only of the test items, but also the persons responding and the context of the assessment (1995:741). We should therefore not state that "a test is valid". Instead, we should seek evidence for saying that the performances of test takers reflect the language ability which the test is designed to assess. Scores on a test are thus only valid with reference to the construct set out to be measured. Henning's definition captures this view: "Validity is the extent to which a test measures the ability or knowledge that it is purported to measure" (1987:198). As

we will see below, this overall definition can be strongly linked to one commonly used aspect of validity, namely 'construct validity'.

When treated as a concept consisting of several subcomponents or aspects, the following contenders are generally used in the literature (Henning 1987; Davies *et al.* 1999; Bachman 1990; Brown 1983; Messick 1989):

a) 'construct validity';
b) 'concurrent validity';
c) 'predictive validity';
d) 'content validity';
e) 'response validity';
f) 'face validity'.

Henning (1987) makes a distinction between empirical and non-empirical kinds of validity. In this respect, types a), b) and c) are empirical, whereas d), e), and f) are considered to be non-empirical. This distinction is made based on the need for data collection or not. Below, a brief description will be given of each of the respective types of validity from above. The descriptions are primarily based Henning (1987), Bachman (1990), and Alderson *et al.* (1995).

In general terms, 'construct validity' refers to the question whether a test measures what it purports to measure. The answer to this question is formed based on both logical analyses and empirical investigations. Of all the above enumerated types or facets of validity, 'construct validity' is seen as the most central one, since it can be seen to subsume all the other types (Messick 1989).This is so because all the other types contribute to score meaning. At the same time as it is the most central type of validity, it is also the most difficult type to establish since it cannot be measured directly. Empirical support for the existence of construct validity can be gathered through measures of internal consistency, which was treated above in the section on reliability, and differences between groups of language users as predicted by theory, through criterion-related measures. The fact that a reliability measure like internal consistency may be used as a means to gather evidence of validity shows that the two terms reliability and validity do not constitute a dichotomy, but rather that they are in many ways intertwined and complementary aspects.

'Concurrent validity', sometimes referred to 'criterion validity', refers to either the extent to which a test can be seen to correlate with another variable which is supposed to measure the same construct, or to the comparison between two or more groups of test takers differing in level of language ability. In this latter sense it is essentially a part of 'construct validity'. Concurrent validity is criterion-related in that a relationship is observed between the targeted test, and an additional criterion measure. The most common way of establishing concurrent validity is to administer a test purported to measure a specific construct with another test also claimed to measure the same construct. If a high correlation coefficient is observed between the two measures, then these is taken as support for concurrent validity in specific terms, and construct validity in general terms. In addition, a test that is not expected to correlate to any great extent can be administered. In this case, no relationship or a weak relationship is expected if separate constructs are to be claimed.

The focus of 'predictive validity' is the examination of whether scores on a test may predict future language behaviour. Empirically speaking, it is closely related to concurrent validity, since the future language behaviour of interest must in itself be tested in some way.

When it comes to 'content validity, this facet deals with establishing whether a test is relevant to a given area of language content or language ability. It normally involves a process whereby experts in a field scrutinize the content of a test in the effort to establish sufficient representativeness of the sample to the test construct.

Furthermore, 'response validity' can be seen as the extent to which test takers' responses reflect the underlying ability that the test purports to measure. Factors that may influence response validity are, for example, the clarity of test instructions, degree of familiarity with the test format, and motivation on the part of the test informants. As such, response validity is closely related to reliability (see Weir 2005).

'Face validity', finally, involves the extent to which a test measures what it is supposed to measure in the eyes of untrained observers, such as the test takers themselves. Data relevant to face validity can be gathered through verbal protocol analyses (see Jourdenais 2001), and through interviews with informants, or through the administration of a questionnaire in which questions can be asked about attitudes and reactions to, and feelings about, a test that has been taken (Alderson *et al.* 1995).

The overall validation methods employed in this thesis will be many-faceted in that I will try to show through argumentation and empirical testing that the scores on my tests can be used to infer a specific type of language knowledge. In effect, in a series of empirical studies, all of the traditional types of validity mentioned above will be covered. In terms of potential causes of invalidity, Henning (1987) suggests the misapplication of a test to be one of the most obvious ones. A test is only valid for the purpose for which it was developed, and any extension away from its specified use may result in invalid interpretations of test scores. Another cause of test invalidity is inappropriate selection of test content. In order to avoid this, test items must be selected in the light of the test construct. The informants taking a test may also cause test invalidity. Henning mentions insincerity, misinformation, and hostility on the part of the test informants as potential problems in this regard.

2.5.5 The application of test theory in this thesis

The fundamental test theory considerations discussed above will be applied in the empirical work presented in Chapters 3-6. A necessary first step has already been taken through the theoretical definition of the test construct. Secondly, an operational definition will be given, along with an outline of the scoring practise. Thirdly, I need to describe in detail the tasks involved in taking the tests, what cognitive processes may be involved on the part of the test takers, and also the process of item selection. fourthly, by administering the tests to learner groups differing in language ability, and to native speaker groups as control groups, and by carrying out various correlational analyses, e.g. the computation of internal consistency coefficients, I will empirically attempt to show that an acceptable level of validity is present in my interpretations of the test scores vis-à-vis the test construct.

3 Operationalising receptive collocation knowledge into test formats: COLLEX 1 and COLLEX 2

3.1 Developing and piloting COLLEX 1

3.1.1 Introduction

In this chapter, the rationale and the procedures behind the development of COLLEX will be described. This will in effect constitute procedures based on which the operational definition of the measured construct can be suggested. In addition, two initial studies set out to provide empirically based information about the quality and effectiveness of the test are reported.

3.1.2 Preliminary considerations

A good starting point in the effort to design a test measuring receptive collocation knowledge is to consider Read's (2000:7-13) 'three dimensions of vocabulary assessment'. Read's set of dimensions is intended to be used as a tool for deciding how to test vocabulary, and the underlying assumptions of the different approaches. Read's dimensions are shown in Figure 3.1. As can be seen in the figure, Read assumes a set of three dimensions that are relevant to the way vocabulary may be tested. The first dimension is focused on the construct tested. In a 'discrete' test, vocabulary is tested as an independent construct of its own, separated from other components of language competence. A vocabulary test can also address vocabulary as part of a larger construct. This approach is referred to as 'embedded. For example, knowledge of vocabulary could be measured as part of the assessment of academic writing ability. The second dimension relates to the range of vocabulary included in a test. A 'selective' vocabulary test is a test in which a set of target words have been selected, and test-takers are assessed in terms of how well they know these words.

Discrete A measure of vocabulary knowledge or use as an independent construct	←→	Embedded A measure of vocabulary which forms part of the assessment of some other, larger construct
Selective A measure in which specific vocabulary items are the focus of the assessment	←→	Comprehensive A measure which takes account of the whole vocabulary content of the input material (reading/listening tasks) or the test-taker's response (writing/speaking tasks)
Context-independent A vocabulary measure in which the test-taker can produce the expected response without referring to any context	←→	Context-dependent A vocabulary measure which assesses the test-taker's ability to take account of contextual information in order to produce the expected response

Figure 3.1 Dimensions of Vocabulary Assessment (from Read 2000:9)

Read points out that the words tested could be selected as individual words and then inserted into separate test items, or the words are picked out from a pre-selected text which is used as the basis for word selection. Another way of approaching the range of vocabulary tested is called 'comprehensive'. This is taken to mean a measure where all the vocabulary used by a test-taker, for example in a written or spoken test, is taken into account. In this way, the more comprehensive use of vocabulary is assessed, not particular words. The third dimension deals with the role of context in a test. More specifically, it has to do with the degree to which test-takers have to make use of a context provided in a test, in order to be able to answer a test item. The dimension is applicable either to a test as a whole, or the individual test items in the test. It will make sense here to use Read's dimensions in my effort to define the construct to be tested operationally.

In developing new tests of collocation knowledge, keeping the observations from the literature review in mind, I had a number of main concerns in addition to the overall aim of producing reliable scores and valid inferences. Firstly, I aspired to construct a test that made use of a large number of test items, but which at the same time would not take a long time to administer. The only way to make this practicable, using Read's set of dimensions from above, was to create a discrete test of receptive collocation knowledge. Secondly, concentrating on one, at most two, types of collocation would make test score interpretation easier. Thirdly, I intended to construct a test which would be easy to score and mark, and which would produce interval data, so that powerful quantitative analyses could be employed.

As to my first concern, ideally, when it comes to lexical knowledge, we would want to employ tests that create a trade-off between the number of items in the test and the degree of generalisability possible from these items to the underlying construct. In general, the more items in a test, the more reliably measured test scores can be achieved. However, since our potential informants are human beings we cannot expect them to concentrate for the time it would take them to sit a very long test. Lapses of concentration and general test fatigue would in all probability kick in, making measurement fraught with error. Thus, there is a clear trade-off between the aim to test many items and constructing a practicable test. Nation suggests that a good vocabulary test should contain at least 30 items in order to produce reliable scores (2001:345). I made a decision to initially use at least 50 items in my test[23].

A consideration concomitant of the desire to include many items was the choice of test task. I would have to come up with a task that was not too complex since this would lead to test takers having to spend more time on each item, a fact that would severely limit the number of items to be included in the test. At the same time, if the task was perceived as too simple and unchallenging, then test takers might not be motivated to do their best.

Having made the decision to test receptive collocation knowledge, yet another choice had to be made. Should the test task involve 'recognition' or 'recall' processes? This distinction refers to two different types of cognitive processes on the part of the language user. In a 'recall' process, the form or the meaning of a word is retrieved and supplied when triggered by some sort of prompt stimulus, whereas in a 'recognition' process the form or meaning of a word is recognized from a set of options (Laufer & Goldstein 2004).

A related question was whether I should make use of translation between L1 and L2. The test was intended to be used primarily with Swedish-speaking learners of English, and therefore, it would be conceivable to involve both Swedish and English in the test task.

[23] Meara (personal communication, 2006) sees a set of 50 items as more or less "ideal".

Nation (2001) addresses the fact that the use of first language translations in vocabulary tests is often frowned upon, whereas he himself fails to see any convincing argument why this should be avoided in general. However, for my test, at some stage of the test development process, I intended to use native speakers of English in terms of validation, as a control group. Using explicit translation as part of the task would make this impossible, or at least difficult. I therefore chose to construct a monolingual task in the sense that only words and structures in English would be used. Since explicit translation was excluded from the task, a receptive recall task was disfavoured. This left me with a receptive recognition task. In this task, test takers are presented with items in which they are instructed to choose an existing form from a set of options.

In order to be able to test a large number of items, I made a choice to use decontextualised items. This meant that my test would be a context-independent test, using Read's dimensions presented as Figure 3.1 above. Certainly, providing some sort of linguistic context around targeted test items makes any task more natural and authentic in that it is the way language appears to us as language users. However, as pointed out by Cameron (2002), it is reasonable to assume that learners presented with decontextualised test items do not make sense of the tested items in a decontextualised mental void. Rather, she claims, the recognition process may activate recall of previous encounters and their contexts. Also, it is arguable that the more context one adds to a test item, the more relevant is the question of what one is really measuring. More context means that reading comprehension and inferencing skills come into play, and this may in a way muddle the measure of the intended construct. This is what Messick refers to as "construct-irrelevant difficulty" (1995:742).

As to my second concern, dealing with score interpretation, I needed to concentrate on one or at most two types of collocation, since this would make score and test interpretation easier. The more types of structures are brought into a measurement, the more difficult it is to define what it is you are measuring. Consequently, I decided to primarily concentrate on verb + NP combinations. This type of combination was chosen first of all because of its frequent occurrence in language (Cowie 1991; Howarth 1996; Nesselhauf 2005; Siepmann 2005). Moreover, these combinations are reported to be notoriously difficult for learners (Biskup 1992; Bahns & Eldaw 1993). Altenberg claims that they "tend to form the communicative core of utterances where the most important information is placed" (1993:227). On the whole, this type of collocation has been researched to the extent that it makes sense to develop tests in which it is the main test item. This will add to the body of research and comparisons can be made between current claims and the results from the present study.

The third concern had to do with the scoring of the test. For the sake of simplicity, I decided against using a scale in which item responses are awarded anything from zero points to several points. Even though such scales may prove worthwhile in detecting partial knowledge (see e.g. Barfield's (2003) study reviewed in Chapter 2 above), I went for a straightforward system in which a correct answer was awarded 1 point and an incorrect answer 0 points.

3.1.3 The COLLEX test format

The above considerations resulted in a test format called COLLEX (collocating lexis). COLLEX is a binary, forced-choice format. It consists of a relatively large number of items (60). An item consists of two word sequences, juxtaposed horizontally. The word sequences are verb + NP combinations. In each item, there is a frequent and conventionalized English

lexical collocation together with a combination which is not frequent or conventionalized. Henceforth, the former will be referred to as a 'target collocation', whereas the latter, functioning as a distractor, will be referred to as a 'pseudo-collocation'. An example of COLLEX items together with the test instruction can be seen in Figure 3.2 below.

In each item, the noun is the same in the two sequences, whereas a different verb is presented in each sequence. Research has shown that verbs tend to be the more difficult elements to acquire and produce felicitously for L2 learners (Källkvist 1999, Nesselhauf 2005), and in many constructions, particularly support verbs constructions, verbs are considered semantically empty in relation to the noun, and that the noun selects the verb element (see Mel′čuk 1998).

The test format works by asking test takers to decide which one of the two word sequences they think is the most common one, and one that would be used by native speakers of English. The format with two juxtaposed choices was inspired by a vocabulary size test of single words, suggested by Eyckmans (2004). Eyckmans found that the binary format was easy to construct and that many items could be covered in a short period of time.

INSTRUCTION:

In the following test your task is to choose one out of two word combinations.

Choose the word combination that you think is the most common one, and the one you think native speakers of English would use in speech/writing, by putting a circle around it.

If you don't know, and have to guess, then tick the box to the right of the word combinations.

			tick the box if you are guessing
1	set the bed	make the bed	☐
2	drop count	lose count	☐
3	run a business	drive a business	☐

Figure 3.2 Instruction and sample items of COLLEX 1.

The underlying assumption of the test format is that one of the two choices is a frequently used, conventionalized word combination in English. Thus, this is a combination that the test takers might have encountered in their exposure to the English language. The other word combination – the so-called 'pseudo-collocation' – is not a frequently used or conventionalized word combination in English, and it is therefore unlikely that the test takers would have been exposed to it in their language input. They are therefore expected to choose the former over the latter, as long as they have some sort of knowledge guiding their choice. They might in this regard have a memorized "version" of the sequence stored in their mental lexicon. This "version" might be what Wray (2002) refers to as a formulaic sequence. It could also be seen as an abstracted construction in which insertions of variables are allowed and also inflections on the inherent elements. It may also be the case that several different and more fine-grained instantiations are stored, out of which one matches perfectly the presented

collocation in the test. In the event that they are not certain of which one figures in English use, they might resort to consulting their overall tacit knowledge of English to decide which one would be more likely.

The COLLEX format is at first glance a simple format, but maybe deceivingly so, for the knowledge needed to solve the task at hand on the part of the test taker is in all probability not insignificant. It will be argued here, firstly, that the test taker needs to possess some kind of knowledge of the meaning or meanings of the single words. For example, in an item like the following:

2 drop count lose count

the test taker needs to know the individual meanings of the words *drop*, *lose* and *count*, respectively. Furthermore, the test taker must make a judgement whether the combinations *drop count* and *lose count* are de facto combinations in English, i.e. if there is a meaningful and conventionalized relationship between the two words in each combination. Finally, the test taker must make a choice as to which one is a commonly used combination by native speakers of English. Thus, a type of knowledge that might be employed in this process is test takers' understanding of the polysemy of the single words making up a combination. Certain verbs, for example, have been seen to combine with certain types of objects (see Stubbs 2001:65). If the test taker 'knows' that *commit suicide* and *commit crimes* are acceptable combinations in English, then when presented with *commit a murder* he or she might decide that this is also an acceptable combination on the basis of analogy in terms of the semantic properties of the object noun. Stubbs argues that a semantic descriptor of the noun in this case could read "crimes and/or behaviour which is socially disapproved of" (2001:64).

The fact that real words are used, and not pseudo-words, as in some vocabulary size tests (e.g. Meara and Buxton 1987) is an advantage, since incidental learning of pseudo-words is avoided. It also means that a large number of real words are featured in the test.

Inherent in the format is also a simple control for guessing. When hesitating about which choice to make, test takers are instructed to indicate whether they resort to guessing (see Figure 3.2). There were two main reasons for this. First, the probability of answering an item correctly is .50. For this reason I felt I needed to get an indication of how frequently test takers needed to resort to guessing when answering the test items. The format theoretically allows someone to get all the items right by guessing, although in practise this would be improbable. As has been pointed out by Brown & Hudson, with a binary-choice format "examinees have a 50% chance of getting the answer correct even if they don't know the answer. However, if there are a large number of carefully designed true/false test items, the overall score should overcome much of the influence of guessing" (2002:66). They go on to conclude that on a 25-item test, a test taker has only 3 in 100,000,000 chances of getting a perfect score by guessing alone.

The second reason was that I was interested in analysing the total scores of the test takers in the light of their indicated rate of guessing. It could be the case that two test takers with the same score could be shown to have guessed in two very different ways. The possible impact of guessing on total test score was thus an important piece of information in the analysis of the behaviour of the test.

3.1.4 Methods

3.1.4.1 Item selection

3.1.4.1.1 Basic considerations

The following method was adopted for the item selection process. I decided to use a set of word lists developed by Paul Nation (Heatley *et al.* 2002) at Victoria University, Wellington, New Zealand. These lists cover the 1000 most frequent and the second 1000 most frequent word families of English. It also covers 570 word families that are frequent in upper-secondary school and university texts from a wide range of subjects. These 570 word families cannot be found among the first 2,000 words. A word family consists of a base form of a word together with inflected and derived forms. For example, the word family represented by the headword 'arrive' looks like this:

ARRIVE
- ARRIVAL
- ARRIVALS
- ARRIVED
- ARRIVES
- ARRIVING

The headword verb ARRIVE, the most common word class for this word family, has the following family members *arrival, arrivals, arrived, arrives* and *arriving*. The first 1,000 word list thus comprises around 4,000 forms or types in total. Out of the first 2,000 word families, about 165 are function word families and the rest are content word families. Henceforth, I will refer to Nation's frequency lists as follows:

1,000 most frequent word families	= 1K
The second 1,000 most frequent word families	= 2K
570 word families common in academic texts	= AW

The sources for Nation's 1K and 2K lists are West's (1953) General Service List of English Words, and for the AW list it is Coxhead's (2000) Academic Word List.

3.1.4.1.2 A corpus-based item selection

The target items for the test parts were selected in the following way. Based on Nation's lists, a database was created. In this database, 150 nouns (50 nouns from each of the three lists), were checked for frequent verb collocations. For this purpose, the *Oxford Collocations Dictionary for Students of English* (Crowther *et al.* 2002) was consulted. This dictionary is in turn based on searches in the British National Corpus (BNC). Before the nuts and bolts of the item selection are further explained, the use of corpora in general, and the BNC in particular, must be accounted for. Also, limitations of corpora use will be pointed out.

A corpus is, broadly speaking, a collection of texts in an electronic database (Kennedy 1998). The BNC is a multi-purpose corpus consisting of approximately 100 million words. One of the main aims of the construction of the corpus was to create a material that would reflect contemporary British English in its various social and generic uses (Kennedy 1998; Meyer 2002). The majority of the corpus consists of written British English material (about 90 per cent), and there is also a smaller part made up by spoken British English material (about 10 per cent). The material is effectively divided into 4124 so-called documents, where each document contains a sample of either written texts, or transcribed spoken discourse, and where a variety of different genres are represented. Most samples are of between 40,000 and 50,000 words (Aston & Burnard 1998:28). The written material was collected between 1960 and 1993, but no data are given as to when the spoken material was recorded. Table 3.1 below depicts the composition of the BNC in terms of genres and the percentage of the part (spoken or written) covered.

Table 3.1 Composition of the BNC with regard to text genres (based on Meyer 2002:31).

Part	Genre	Number of documents	Percentage of the written/ spoken pa of the corpus
Written	Imaginative	625	22%
	Natural science	144	4%
	Applied science	364	8%
	Social science	510	15%
	World affairs	453	18%
	Commerce	284	8%
	Arts	259	8%
	Belief & thought	146	3%
	Leisure	374	11%
	Unclassified	50	2%
	Total	3209	99%[1]
Spoken	Demographically sampled	153	41%
	Educational	144	12%
	Business	136	13%
	Institutional	241	13%
	Leisure	187	14%
	Unclassified	54	7%
	Total	915	100%

[1] Because of fractions being rounded up or down, the total does not add up to 100 per cent.

Computerized corpora allow researchers to investigate very large collections of data, to use their findings as sources of evidence for linguistic description and argumentation, and to do this beyond particular intuitions and preconceptions. Corpus searches furthermore offer techniques for counting and sorting linguistic material, and they come across as especially effective when it comes to collocation, in charting the tendency and probability of certain words to frequently co-occur in natural language. In this respect, it is arguably a more reliable guide to language use than, for example, speaker intuition. Hunston (2002:21) gives examples of collocations which learners of English tend not to use: (see Granger 1998): “acutely aware”, “painfully clear”, “readily available”, and “vitally important”. Hunston argues that it

is difficult for native speakers to have conscious access to these combinations, but that they are readily revealed through corpus search. However, there are limitations to corpus use. Firstly, through corpus analysis, it is possible to describe language use. How much faith we can put in our findings, though, hinges on the representativeness of the corpus we are using. We must therefore treat our findings with caution. Secondly, just because a certain pattern is not found in the corpus, it does not mean that the pattern is not used at all in a certain speech community. A limitation is thus that corpora merely tell us whether something is frequent or not. They do not tell us whether something exists or not. Despite these limitations, corpora constitute powerful tools in language research. In the words of Kemmer & Barlow (2000:xvi): "…corpus data provide a sampling of usage that can reflect general patterns very faithfully".

Returning to the BNC-based collocation dictionary, the main criterion for inclusion of words from the BNC in the dictionary was "typical use of language" (Crowther *et al.* 2002:viii). This resulted in the inclusion of 9,000 nouns, verbs and adjectives as headwords. These were included based on their frequent inclusion in typically used collocations. The collocations were chosen based on their frequency, their range (number and kinds of sources), and the contexts in which they appear in the BNC. For each of the 150 nouns in my own database, four to five possible verb collocates were recorded from the collocation dictionary. In addition, one or two 'pseudo-collocations' were created. This was done by keeping the noun constant, and combining it with a verb that does not normally collocate with it. From this list, nouns that combined with verbs, where the resulting combination was expected to present difficulty to Swedish learners of English, for "learning burden" (see Nation 2001) reasons, were selected. The notion of learning burden refers to the amount of effort required to learn a word. The general principle of the learning burden says that the more a word represents patterns and knowledge that are already familiar to a learner, the lighter is the learning burden. The patterns may come from the L1, from other languages, or from the learner's previous knowledge of the L2. Relevant to this, Ijaz talks about a 'semantic equivalence hypothesis' (1986:443):

> This hypothesis facilitates the acquisition of lexical meanings in the L2 in that it reduces it to the relabelling of concepts already learned in the L1. It confounds and complicates vocabulary acquisition in the L2 by ignoring crosslingual differences in conceptual classification and differences in the semantic boundaries of seemingly corresponding words in the L1 and L2.

Thus, even if a learner correctly interprets the reference of a new L2 word form, it seems unlikely that he or she will grasp the rather complex system of semantic and structural characteristics that that word displays, and an initial mapping onto a L1 equivalent is a common procedure (Nation 2001). Take as an example the English verb *keep*. In line with the above hypothesized processes, a Swedish learner might map this L2 form onto an L1 verb equivalent like *hålla* or *behålla*. As a consequence of this, it might be inconceivable to the learner that the Swedish V + NP collocation *föra dagbok* corresponds to 'keep a diary' in English. Possible non-standard suggestions, admittedly varied in probability, from Swedish L2 learners might be *lead a diary*, *conduct a diary* or *run a diary*. Thus, in COLLEX, the collocation *keep a diary* might be juxtaposed with the pseudo-collocation *lead a diary* since the latter is a possible 'bait' for learners who have not been exposed to the conventionalized collocation *keep a diary*. Using this principle, 60 items, each consisting of one target collocation and one pseudo-collocation, were selected for inclusion in the test.

In COLLEX, both elements of a collocation, i.e. the verb and the noun, were selected from the same or a higher frequency band. For example, a noun from 2K (the second 1,000 most frequent word families) was paired with a verb from either 2K or 1K. This was done to avoid the possibility of a learner knowing the noun but not the verb, based on a general assumption involving an expected correspondence between ascending order of difficulty and descending order of word frequency. Three test parts were created. Part one contained nouns and verbs from 1K. Part two contained nouns from 2K and verbs from either 2K or 1K. Finally, part three contained nouns from the AW list and verbs from the AW, 2K, or 1K lists. In practise, however, the great majority of the verbs in all the three parts came from the 1K band. Therefore, an assumption was made that for the relatively advanced learners that the test targets, these verbs would all be well-known in terms of their generalised basic meaning.

The test items were presented to a native speaker to minimize the risk of including pseudo-collocations that are in fact possible collocations. Furthermore, a number of Swedish colleagues, all near-native speakers of English, sat the test and were afterwards consulted on the feasibility of the pseudo-collocations in terms of their ability to attract answers from Swedish learners of English. Based on the findings from this process, a number of items were discarded or amended.

3.1.4.2 Material

In addition to the COLLEX test, a test of single word knowledge was also administered in the pilot test session. This test, called SINGLEX (single lexis), consisted of all the 60 nouns from the COLLEX test. SINGLEX was included to answer the question of whether learners knew the 60 nouns included in COLLEX. The rationale behind this was that I was interested in whether learners who knew the single word noun also knew a frequent collocation that this noun enters into. A multiple-choice format was adopted for testing learners' knowledge of the single word nouns. In the SINGLEX format, the test takers are exposed to an L2 word and must then select, from among three options, the L1 word whose meaning corresponds most closely to the meaning of the L2 word. The format was made sensitive by using L1 options which do not lie close to each other in meaning. The following criteria were followed (see Brown & Hudson 2002:68-71):

-The L1 options are grammatically consistent with the L2 stem (here: all nouns);
-The L1 options are of fairly similar word length;
-Wordiness is avoided by supplying only single words as options.

3.1.4.3 Informants

The informants were 19 Swedish teacher students of English, who at the time of testing were in their second year at university. Before university, they had studied English on average for 10 years in school[24].

[24] In Sweden, as an undergraduate student, you can pursue full-time studies in a subject like English for 4 terms (2 years). The first term of full-time studies is called the A level, the second term is called the B level, the third is called C level and the fourth D level. Being teacher students, the subjects taking my tests did not strictly follow the progression that students of general English do, but in terms of formal level of study, the subjects were judged to be on a proficiency level equivalent of somewhere in between the Swedish B and C levels. This judgement was made by an experienced university lecturer of English, who was teaching the group at the time of testing.

The reason why this group of students was chosen for the pilot was that their perceived proficiency in English matched the upper register of that of the target group for whom the test was eventually intended: from upper-secondary school students to university students. By piloting my test on this group, I believed that the results would provide a relevant indication of aspects like test difficulty, and also more practical matters, such as the time needed by this group of test takers to finish the tests.

3.1.4.4 Research questions

The chance of coming up with a good test at the first attempt is more or less microscopic. By testing our tests, as it were, we may elicit data that can guide our decisions on how to proceed with future testing sessions and aspects concerning test development, such as item selection, test formats and hypotheses.

The following questions were addressed:

1. Is the binary-choice format in COLLEX a viable one for testing verb + noun collocations?
2. Does knowing the meaning of high-frequency single nouns entail knowing common collocations that these nouns enter into?
3. Is guessing frequent in COLLEX and what effect does it have on test takers' scores?
4. Is the level of difficulty of COLLEX appropriate for the tested learner group?

The answer to the first question will be contingent on the answers to the subsequent three questions. With regard to question 2, it was hypothesized that the learners would produce very high scores on SINGLEX, whereas their scores on COLLEX would be lower. Prior to testing, there was no way of knowing to what extent guessing would be indicated on the test. Furthermore, it was not possible to hypothesize whether the level of difficulty would be appropriate for the tested informants.

3.1.4.5 Test administration

The test battery was administered in connection with a taught English course. I was kindly offered by the lecturer of the course to visit one of the classes to run the tests. The only information given to the students was that I was conducting research on English vocabulary. The single word test, SINGLEX, was administered first and was completed by a great majority of the students in about 5 minutes. After having collected the SINGLEX test sheets, the collocation test, COLLEX, was handed out. This test was completed by all the students in less than 10 minutes.

3.1.5 Results

The results of the study were analysed in two steps. First, descriptive results from the two tests were calculated. This included computing mean scores and standard deviations, both for the test as wholes, and also for each of the three parts of the tests. For COLLEX this analysis also included data on guessing frequency, and an estimate of test score reliability in terms of an internal consistency co-efficient (Cronbach's alpha). The guessing data were furthermore

subjected to inferential analyses. Second, an item analysis was carried out in order to discern individual item difficulties and item discrimination indices.

The research questions posed in the previous subsection will be addressed one by one. However, the answer to the first research question will be approached last, since it is largely contingent on the outcome of the other questions.

The results for SINGLEX are shown in Table 3.2 below. In terms of scoring methods, the learners were given 1 point for each correct answer and 0 points for each incorrect answer.

Table 3.2 Results on SINGLEX test of nouns, piloted in October 2004 (N = 18).

Value k	Total (60)	Part 1 (20)	Part 2 (20)	Part 3 (20)
Mean	59.7	19.9	20	19.8
S.d.	0.9	0.2	0.0	0.7

Out of the 19 subjects who took the test, 1 subject failed to give answers to items 51-60, probably by mistake, and was therefore excluded from the analysis of the SINGLEX data. Thus, the results of 18 subjects are reported here. The scores on SINGLEX were very high, resulting in a tangible ceiling effect. However, a high to very high set of scores on this test was more or less expected and perhaps not surprising given the high frequency of the test items and the sensitivity of the test format. The test was given, in the first place, to ensure that the students knew the nouns which were subsequently tested in terms of what verb collocations are acceptable with those nouns. 16 students scored the maximum point of 60; one scored 59, and one scored 56. Due to the very high scores, no reliability coefficient was computed for these scores. Based on these results, we may conclude that the subjects in the study knew the tested single nouns well.

As can be seen in Table 3.3 below, the scores on COLLEX were high but not as high as on the single word test. In terms of descriptive statistics, the overall mean was 51.7 (max. 60) with a standard deviation of 3.3. The means on the three respective parts were 18, 17.2 and 16.5 respectively. The subjects thus scored slightly better on part 1 than on part 2, and better on part 2 than on part 3, but with very small differences. The low values throughout for standard deviation—the variability of the data from the point of central tendency—show that the subjects can be seen as a homogenous group. The dispersion of scores is very small.

Table 3.3 Results on COLLEX test of verb + noun collocations, piloted in October 2004 (N = 19).

Value k	Total (60)	Part 1 (20)	Part 2 (20)	Part 3 (20)
Mean	51.7	18.0	17.2	16.5
S.d.	3.3	1.6	1.7	1.3
Reliability alpha	.54	.46	.42	-.34
Guesses (f)	91	24	29	38
Correct guesses (f)	47	11	13	23

The reliability of the test as measured by Cronbach's alpha is low, with a value of .54. The values for the respective parts are even lower, with part three having even a negative alpha. As was outlined in Chapter 2, a low reliability score implies an unacceptably high degree of measurement error. One probable explanation of the low value is the homogeneity of the group. We saw in Chapter 2 that this is one reason behind low reliability figures. A further explanation is related to the possibility of poor item quality. We will return to this eventuality at a later stage in this chapter. Yet another factor could be the relatively large number of guesses present. We can see that part three attracted more guesses than did parts 1 and 2, and it follows from this that guessing probably plays a part in the unacceptably low reliability value for this part. On the whole, the guessing behaviour reported merits further investigation. We will return to this issue presently.

When it comes to answering the second research question, it stands to reason that the subjects participating in the present study did not know all the common verb + NP collocations based on the nouns tested in the SINGLEX test. However, given the low reliability coefficient of the scores, we should be careful not to draw too far-reaching conclusions from this finding. Also, this result is not very controversial. It follows the wide agreement in the field of vocabulary research, that learners' knowledge of collocations and extended senses of frequent words is a constant obstacle on their way towards near-native speaker competence (Biskup 1992; Bahns and Eldaw 1993).

To explore the third research question, as to how frequent guessing is and how it may affect learners' scores, let us firstly refer back to Table 3.3. A total of 91 guesses were reported to have been made, and the subjects produced almost as many wrong guesses (44) as successful ones (47). Comparing the different test parts, more guessing was observed on part 3 than on part 2, and equally on part 2 than on part 1. This means that guessing increased as the frequency of the words decreased. In order to take a closer look at the guessing behaviour, the subjects were divided into three groups according to their total score on the test. These groups will be referred to as 'low group' (N = 6), 'mid group' (N = 7), and 'high group' (N = 6). Consider Table 3.4 below.

Table 3.4 Guessing behaviour with respect to three total score groups

Group	N	Score range	Mean no. of guesses	Standard deviation	Per cent correct guesses
High	6	53-58	2.17	2.64	62%
Mid	7	51-52	5.43	3.95	58%
Low	6	46-50	6.67	5.72	43%

In Table 3.4 we see that the six subjects with the highest COLLEX scores, called 'high', produced a mean of 2.17 guesses with a standard deviation of 2.64. The 'mid' group produced a slightly higher mean of 5.43 (s.d. 3.95), and, finally, the 'low' group produced the largest mean at 6.67 (s.d. 5.72). The large standard deviation for the 'low' group stems partly from one learner reporting as many as 17 guesses (7 correct and 10 incorrect). Incidentally, this subject was the lowest scorer on the test with a total of 47 points. Conversely, the subject with the highest total score on the test (58) produced the lowest number of guesses (0). An ANOVA comparing the guessing means of the three groups showed no significant effect of

group affiliation on guessing. This might be suspected considering the few subjects and the rather large variance observed in the data.

In order to further tease out the relation between the number of indicated guesses and the total scores of the subjects, a correlation analysis was carried out. Since the data set was rather small, and there were a number of tied ranks, a Kendall's tau (τ) test was used (Field 2005). The analysis showed that there was a significant negative relationship between the number of guesses indicated and the number of points on the test, $\tau = -.35$, $p = < .05$. This means that as the total score for a subject increases, the number of indicated guesses decreases. The conclusion we may draw from this points in a positive direction. Learners who are more skilled in the underlying ability which is intended to be measured by the test get a higher score than learners who are less skilled, and, most importantly, they do so without resorting to guessing to the same extent as the lower scoring learners. If our analysis of guessing behaviour had shown that high scorers and low scorers alike were guessing equally frequently, then it would have been likely that some of the high scorers reached high scores more by chance than by relying on an underlying language skill. This does not however seem to be the case.

When it comes to the outcome of the guessing behaviour, the high scoring group were the most successful guessers (62% correct guesses) followed by the mid scoring group (58%) and the low scoring group (43%). The results indicate that across informants participating in the study, those with higher scores guessed less often and were at the same time more successful when guessing compared to learners with lower scores.

In order to compare the guessing behaviour of informants with the same total score, five informants lying close to the mean score (51.7) were selected. This was done to see if very different guessing behaviour lay behind the same total score. The informants all reached a total score of 52. Their respective number of indicated guesses together with correct and incorrect guesses are shown in Table 3.5 below.

Table 3.5 A comparison of the guessing behaviour of five subjects with the same total score.

Learner ID	Total score	total no. of guesses	no. of correct guesses	no. of incorrect guesses
9	52	3	1	2
14	52	10	4	6
15	52	2	1	1
16	52	3	2	1
17	52	5	4	1

Informant 14 stands out with 10 indicated guesses, compared to learner 15 who only indicated having guessed twice. Among the five learners, learner 17 proved to be most successful in guessing with four out of five guesses being correct. The other four learners guessed correctly roughly half of the attempted times.

We are now in a better position to take stock of the guessing behaviour and its possible effects on the learners' scores. The mean number of guesses indicated by the 19 subjects is 4.7. From this we conclude that although guessing does occur, and a few subjects report to have guessed a large number of times, guessing does not occur to an alarming extent considering that the number of items is as high as 60. High scorers on COLLEX were found to guess less often than mid and low scorers, but with higher success rates. However, the

mean numbers of guesses of the three groups were not significantly different. Furthermore, guessing was on the whole found to correlate negatively with total scores on the test. A speculation was made whether the fact that a close to equal number of correct and incorrect number of guesses was made, could be taken as evidence of the reliability of the self-reported guessing. All in all, the results of the analyses seem to suggest that excessive guessing does not occur, that high scorers make few but often correct guesses, and that guesses seen across the learner group largely cancel each other out. However, the limited number of subjects in the study places restriction on the inferential statistical testing that can be carried out on the data.

As to our fourth research question, judging from the high mean scores on COLLEX, it is already clear that the test contained a large number of easy items. Thus, the test content was not totally appropriate for the tested learner group. The implications of this will be discussed below. An analysis of test items is an important step in further investigating test reliability since it will show which items are, for example, ambiguous and faulty. A test which contains many faulty items will tend to be unreliable (Bachman 1990:87). However, in order to arrive at a more informed picture of the test, a proper item analysis should be performed. In order to obtain information on how well each of the 60 test items in the first pilot of COLLEX worked, two values were computed: Item Facility (IF) and Item-total correlations (ITC). In the following paragraphs, I will briefly outline the way these two item indices work, and the way to interpret them. The computed values for each of the 60 COLLEX items can be found in Appendix 3A.

Item Facility expresses the proportion of the test takers who got an item right. A facility value ranges from 0 to 1. A very easy item which all test takers answer correctly means a value of 1.00 and a very difficult item which none of the test takers answer correctly means a value of 0.00. The ideal value is sometimes postulated to .5 (McNamara 2000:61). If a test constructor wants to get a wide spread of scores on a test, then he or she should select items with a facility value as close to .5 as possible (Alderson *et al.* 1995:81). Analysing the IF column of the table in Appendix 3A, we must conclude that COLLEX in its present version contains too many easy items. As many as 24 items have an IF value of 1, meaning that all subjects answered these items correctly. The mean IF of the 60 items is .86, which is at the high end.

Item-total correlation (ITC) is a common technique for computing item discriminability (Henning 1987:52). Item discriminability tells us how well an item discriminates between test takers of different levels of language ability. Ideally, we want more of the test takers with the highest total scores on a test to get an item right than test takers with low total scores. If this is not the case, then something is clearly wrong with the item in question. The computation involves correlating test-takers' scores for a given item and test-takers' scores for the test as a whole. In general terms, what is tested is if students with high total scores get the item right and if students with low total scores get the item wrong. This is what we expect from a well-functioning test. The value arrived at in an item-total correlation analysis ranges between +1 and -1. An item with negative values is clearly behaving badly as a test item. Following a proposal from Ebel (1979), items can be seen as functioning more or less well in reference to the following scale:

.40 and higher very good items
.30 to .39 reasonably good items possibly subject to improvement
.20 to .29 marginal items in need of improvement
below .19 poor items which need to be revised or eliminated

If applied to the item-total correlation values of the items in the table in Appendix 3A, we get the distribution presented in Table 3.6. If we follow Ebel's guidelines we are left with a rather pessimistic view about the quality of the items in the COLLEX pilot. One of the main reasons behind the obtained values is the fact that as many as 24 items display zero variance. This is due to the fact that all 19 subjects answered these correctly. Also, as many as 12 items display a negative item-total correlation. This points to the fact that for these items, one or several high scorers answered the items incorrectly, whereas low scorers answered the items correctly. As a consequence of this, the overall reliability of the test decreases significantly. Only seven items are considered to be very good items, if Ebel's guidelines are followed.

Table 3.6 Distribution of test items from COLLEX 1 into categories of discrimination following Ebel (1979).

Item-total correlation guidelines	.40 and higher very good items	.30 to .39 reasonably good items possibly subject to improvement	.20 to .29 marginal items in need of improvemen	below .19 poor items which need to be revised or eliminated
Items from COLLEX 1	3, 17, 21, 32, 44, 45, 55	4, 6, 29, 33	14, 20, 37, 38, 39, 5	7, 12, 16, 24, 25, 27, 30, 34, 41, 42, 47, 48, 49, 50 51, 52, 57, 59, 60 Value 0.00 (no variance): 1, 2, 5, 8, 9, 10, 11, 13, 15, 18, 19, 22, 23, 26, 28 31, 35, 36, 40, 43, 46, 54 56, 58,
Number of items	7	4	6	43 (19 + 24)
per cent of total number of items	12%	6%	10%	72%

3.1.6 Discussion

In this section, I will consider three main points: the item quality, the guessing behaviour, and the practicality of the test. Eventually, this discussion will guide the decision of whether COLLEX is a viable test worth developing further. Firstly, though, research question 2 will be addressed.

I asked whether knowing the meaning of high-frequency single nouns entails knowing common verb + NP collocations that these nouns enter into. In relation to this, I hypothesized that learners would produce very high scores on SINGLEX, whereas their scores on COLLEX would be lower. This is also what was found. Schmitt (2000:79) declares that "although it is not clear how collocational knowledge is acquired, it seems to be relatively difficult to achieve". As for the COLLEX items, all we can say is that the items in part 3 (AW) were slightly more difficult than those in parts 2 and 1, respectively. A possible explanation for this could be that the AW list words used in the items of part 3 were taken from a list of words that are common across academic texts. Consequently, not only were the words in part 3 less frequent, they were also words with a somewhat more restricted range than the words in parts 1 and 2. However, considering the low reliability of the test instrument, we cannot draw too far-reaching conclusions from this finding, and until a better test tool is developed, we should be very cautious about inferencing any strong claims from the study in this regard.

As was evident in the results reported in section 3.1.5, there is room for improvement of COLLEX as a test tool. One of the main problems has to with the seemingly poor item quality. It is obvious that far too many of the chosen test items are too unchallenging for the subjects taking the test. The mean facility value (Appendix 3A) amounted to .86. This is decidedly high and due to the fact that nearly half of the 60 items in COLLEX displayed individual facility values of 1.0. The test itself is eventually aimed to be targeted at intermediate and advanced learners of English, such as senior upper-secondary school students and university students. The subjects taking part in this pilot belong to the advanced register, and probably the upper part of that to boot. Therefore, we would expect them to do fairly well on the test. It is likely that upper-secondary school students would not perform equally well. All the same, the item analysis also showed that the items of the test had a poor discriminatory power overall. This all resulted in unreliably measured scores, evidenced by the disappointingly low coefficient for internal consistency (Table 3.3). A considerable improvement of the actual items of the test is therefore clearly called for. The question is what caused the seemingly poor item quality.

The problem may partly emerge from the way the items were selected. It will be argued here that the restriction on word frequency was unfortunate in this respect. The restriction meant that the collocating verb of a noun must not be taken from a lower word frequency band than this noun. This applied to both the verb of the target collocation as well as the verb in the pseudo-collocation. On the whole, this procedure can be seen to have impaired the content of the test since many interesting collocations could not be included in the test for the reason that the verb of the verb + noun collocation resides in a lower frequency band than the noun. Also, the process of finding suitable verbs for the construction of the pseudo-collocations suffered the same restriction. In the light of the many easy items in the present version, one remedy is to try to make the pseudo-collocations more plausible as choices for the informants. It might be the case that relaxing the restriction on word frequency will prove to be beneficiary in this respect. The result would then be that no strict separation of words and collocations into test parts would be enforced. This could lead to criticism saying that learners might not know a certain collocation because they do not know one of the component parts of the combination. This kind of criticism, though, would only be valid if the words used in the test would come from very low frequencies well beyond the proficiency range of the targeted student population. If relatively high frequency words are used, here tentatively

meant to refer to words from the 1K to 5K bands, then the intermediate and advanced students seen as the target group of the test would know these words.

Another unfortunate aspect of the current version of COLLEX was the inclusion of a number of faulty or at best ambiguous items. In hindsight, despite attempts to have a native speaker as well as Swedish-speaking colleagues check the items of the test prior to the administration of the pilot, some test items in the test were still infelicitous. Examples of such items are:

(17) item 11 keep a speech – give a speech
(18) item 22 pay attention – show attention
(19) item 34 turn a key – twist a key
(20) item 52 perform a task – solve a task

The main problem with these items is that both alternatives are to some extent possible. Even though one of them might be more frequent in use, the alternative might be conceivable in a certain context. Take for example item 11. The targeted collocation is *give a speech*, with the sense 'talking at a formal or semi-formal occasion in tribute to someone or something, often based on a rehearsed piece of text'. The constructed pseudo-collocation, *keep a speech*, meant to attract learners' attention as a viable choice. This is feasible since the item was meant to capture the concept of someone talking, and since the English verb *keep* and the Swedish verb *hålla* are translation equivalents. In Swedish, the form for the concept is *hålla tal*. Consequently, a learner who is not aware of the collocation *give a speech*, might due to transfer choose *keep a speech*. The problem, however, is that the form *keep a speech* exists in the sense 'to save (as a memento) a piece of written text once read out at a formal occasion'. The examples in (18-20) suffer from the similar kind of ambiguity.

In the present study, it was possible to investigate the guessing behaviour on COLLEX through a self-report method. The subjects of the study were instructed to indicate when they resorted to guessing on an item. The inherent problem with the method is that what is considered to be a guess may vary considerably from subject to subject. It is not possible to measure how sure an individual is of a choice in relation to whether a guess is indicated or not. This makes the method a rather crude one. In theory, we may illustrate the guessing aspect through either a discrete-state or a continuum model. In a discrete-state model, as a suggestion, a test-taker can be seen to be faced with the following 5 cognitive states:

1 = completely sure of choice A
2 = fairly sure of choice A
3 = choices A and B equally appealing
4 = fairly sure of choice B
5 = completely sure of choice B

An assumption in a model like this would be that test-takers are more likely to indicate a guess in states 2, 3 and 4, than in 1 and 5. We would also assume that state 3 would result in an indicated guess with a very high probability. However, in practise, subject A may indicate a guess at state 2 or 4, whereas subject B may not. Also, even though it may seem unlikely, learner C may indicate a guess at state 1 or 5. The difference between a discrete-state model and a continuum model is in my understanding the fact that a discrete-state model is likely to

be a simplification of the cognitive process it is supposed to illustrate. A continuum model would therefore represent more fine-grained differences in the cognitive process. Instead of positing discrete states, we would have a continuum ranging from 'completely sure of choice A' to 'completely sure of choice B' with the middle of the continuum corresponding to 'maximal doubt'.

Despite the inherent problems, the inclusion of a 'guessing gauge' in COLLEX provided several interesting insights. Firstly, it was possible to see how frequently the subjects of the study claimed to have guessed. For a 60-item test, a guessing mean of 4.7 does not seem to be very high. On the other hand, this relatively low guessing frequency is likely to be linked to the high mean facility value of the test, which was observed at .86. It is highly likely that a more difficult test will lead to a higher guessing mean. The fact that a significant negative correlation was found between number of guesses and total scores is revealing, but maybe not totally surprising. Assuming that an indicated guess means that a learner is experiencing something similar to cognitive state 3 from above, then the more often this state is experienced, the more instances in which an incorrect answer may be the result.

Secondly, we could see how many of these guesses were successful, i.e. resulted in a choice that was correct according to the test key. In this regard, the overall frequencies of correct and incorrect guesses were close to .5. The fact that the guesses are distributed this way is interesting since this could be taken as possible indirect evidence of the reliability of the self-reported guessing instrument. The logic behind this claim needs to be spelt out. Let us again consider the 5-state model from above. Statistically, pure, random guessing based on two options (state 3 in our model) has a theoretical probability of .5. This is similar to the distribution for coin tosses. In the long term, these situations will lead to a relative frequency of occurrence of .5 for either option. This is very close to the results gained from our guessing data. This might then suggest that learners in the study were truthful in their indication of guessing behaviour. If, instead, we had observed the ratio of correct guesses to incorrect guesses to be 3:1, then this could have meant that the learners were indicating guesses when they were in fact quite sure of the right answer.

Thirdly, by dividing the subjects into groups conditioned by their total score (high, mid, and low), it was possible to analyse whether there was a difference in guessing behaviour between them. In terms of absolute numbers, there was a difference between the groups, especially between the high group and the other two groups, but no statistical significance was found between the means of the three sub-groups in an ANOVA, possibly due to the limited sample. The high standard deviation in the scores, and the extensive overlap, suggest that the subjects of the three groups did not come from different populations.

On a positive note, anecdotal evidence tells us that the subjects taking COLLEX found the test to be an interesting and different kind of vocabulary test. Many informants said that they thought it tested vocabulary in a "new" and "fresh" way, compared to the tests that had been subjected to earlier. Even though we must be careful when it comes to anecdotal evidence, this is still encouraging.

Furthermore, it was possible to administer COLLEX in less than 10 minutes. Consequently, the format is a very practical one, covering many items in little time. Thus, the answer to research question 1 is positive. The COLLEX format is still seen as a viable format. Admittedly, no experimental procedure was employed to test the task format compared to other task formats, but the collective evidence in this study points to the actual items in the present version being the weak factor, not the test format itself.

3.1.7 Conclusion

In conclusion, provided that the item quality can be considerably improved, I think that the COLLEX format is worth pursuing. Therefore, I will develop a new version of COLLEX and put that to the test. This will be done in the following section (3.2), in which a study with a considerably larger group of informants will be carried out.

3.2 Developing and administering COLLEX 2

3.2.1 Introduction

In the previous section the results from an initial pilot administration of a collocation test called COLLEX were reported. On the plus side, the results showed that COLLEX is a practical test tool, which is quick to sit, contains many items, and is anecdotally perceived as an interesting test on the part of the test takers. On the minus side, many of the items used in the first pilot version proved to function poorly. The main reason for this was the fact that as many as 24 items out of 60 showed no variance due to the fact that all test takers answered these items correctly. If a well-functioning test is to be developed for a similar kind of target group, university-level learners of English who have had approximately 10 years of classroom exposure to English, then the test must clearly be made more difficult. As a part of this work, the distractors must be capable of attracting more answers. Another factor believed to have contributed to the negative outcome of the test administration was the fact that some items were ambiguous to an extent which made these items unreliable. As a consequence of the lack of variance in almost half of the items in the test, together with some faulty items, a low overall measure of reliability was observed. However, a further developed COLLEX test is believed to overcome these problems.

3.2.2 Methods

3.2.2.1 Item selection

For the second version of COLLEX, considerable improvements were aimed for in terms of item quality. Items that were proven to function poorly were either discarded or amended. In the pilot reported in section 3.1, the selection of target words was restricted in the sense that nouns were used as the basis of target pair selection. These nouns were selected from 3 frequency bands (1K, 2K and AW), 20 target nouns from each, in total 60 words, divided into three test parts. The collocating verbs for each test part were all taken from the same or higher frequency bands as the nouns. This restricted the choice of pseudo-collocations to an extent which might have affected the behaviour of the test negatively.

In the binary-choice format, some of the pseudo-collocations did not function as good distractors. The effect was also that the test failed to discriminate between more proficient and less proficient learners.

In the new version of the test, COLLEX 2, the criterion of using only the same or higher frequency of the component words in the items was abandoned. This meant that collocating words from lower frequencies than the noun prompt word were included in the test. In practice, however, a great majority of the two verbs in an item belonged to the same or adjoining frequency bands (the same thousand word band or, for example, one verb from band 1K and one from 2K). On the whole, even though care was taken to concentrate on high to moderately high frequency words for the items, sometimes obtaining distractor "credibility" took priority.

The frequency bands used in COLLEX 2 were those of Kilgarriff (1996), which is a BNC-based list available on the Internet. The main differences between the previously used frequency list (see 3.1.4.1) and the present one are that the former contains word families whereas the latter is a lemmatised list, and the fact that they are based on different source

material. The new frequency list was chosen on the grounds that it gives more precise frequency information. In the previously used list, words were classified as 1K words, 2K words, or AW list words. The AW list words are words commonly occurring in academic texts. The AW denotation only tells us that the words are less frequent than the 2,000 most frequent words (word families) of English. Thus, we do not know how infrequent these words are, only that they are common in a wide range of academic texts. The BNC-based list contains 6,318 words with more than 800 occurrences each in the whole 100-million-word corpus. This list can thus be divided into 1K, 2K, 3K, 4K, 5K, 6K, and 7K words. Kilgarriff's (1996) definition of a word approximates to headwords as used in EFL dictionaries. In this sense, nominal and verbal versions of a word are listed separately.

In terms of the items in COLLEX 2, the well-functioning verb + noun items from COLLEX 1 were retained. A few items showing zero variance in the first administration were also kept. This decision was based on the belief that they still held promise as decent items. The informants in the first pilot were undergraduate students, with an average of 10 years of English study behind them, and it was deemed likely that some items might still cause problems for less proficient students. These items were supplemented with newly created items. The number of items was increased from 60 to 65 in order to create a slightly larger pool of items to choose from in future testing sessions. A majority of the 65 items were verb + noun phrase items (52), but also adjective + noun items (13) were included. The aim was to try to keep the frequencies of the words making up the items as high as possible in order to minimize the impact of the learners' vocabulary size on the test performance. The aim of the COLLEX test was not to create a vocabulary size test (see section 2.4.3.2.2), but rather to test student's knowledge of collocations based on high-frequency words. The following words used in the second version of the COLLEX test came from a lower frequency than 1-4K:

Table 3.7 Words in COLLEX 2 with lower word frequency than 1- 4K.

5K	6K	7K+
crush (verb)	polish (verb)	dial (verb)
shed (verb)	sacrifices (noun)	fell (verb)
apologies (noun)		undo (verb)
conscience (noun)		visibility (noun)
slim (adj.)		smoker (noun)
		errand (noun)
		motorcycle (noun)
		fuse (noun)
		amends (noun)
		heed (noun)
		slender (adj.)
		fake (adj.)
		foul (adj.)

The words in the column labeled 7K+ in Table 3.7 above are words that cannot be found in Kilgarriff's (1996) word list. Consequently, these words do not have a high enough frequency to appear among the c. 6300 most common word lemmas in the BNC. Relying on my experience as a teacher of English to advanced EFL learners, my judgement tells me that some of the above words might possibly prove to be difficult for the learners, whereas others

despite their relatively low frequency will definitely not. For example, nouns like *apologies*, *smoker* and *motorcycle* are extremely unlikely to cause problems for advanced Swedish-speaking learners, but *errand* and *heed* might. For the purposes of cross-checking, all of the above 20 words were checked against a 15,000 word family list, also based on the BNC. The list was available at a website (Cobb 2006)[25].

As is evident in Table 3.8, some of the lower frequency words in the lemmatized list appear in a much higher frequency band in the word family list. The reason for this is that word families consist of a baseword, their inflections and the most common derivatives. For this reason, the verb *smoker* is not among the first seven thousand lemmas in the Kilgarriff list, whereas it is a 1K word in Nation's word family list, residing under the headword noun *smoke*. The tendency we can see in the comparison is that the word class based lemmatized list causes some words like *smoker* and *motorcycle* to become classified as low-frequency words, whereas the word family based list may inflate the frequency of some word forms, .e.g. the forms *shed* and *fell*.

Table 3.8 Comparison of word frequencies between a word lemma list and a word family list based on the BNC.

Form	Kilgarriff's (1996) lemmatized BNC list	Nation's (2006) word family BNC list
crush	5K (verb)	3K
shed	5K (verb)	2K
apologies	5K (noun)	2K
slim	5K (adj.)	3K
conscience	5K (noun)	4K
polish	6K (verb)	2K
sacrifices	6K (noun)	6K
dial	7K+ (verb)	3K
fake	7K+ (adj.)	5K
fell	7K+ (verb)	1K
foul	7K+ (adj.)	4K
undo	7K+ (verb)	3K
visibility	7K+ (noun)	3K
errand	7K+ (noun)	7K
smoker	7K+ (noun)	1K
motorcycle	7K+ (noun)	4K
fuse	7K+ (noun)	4K
amends	7K+ (noun)	4K
heed	7K+ (noun)	6K
slender	7K+ (adj.)	9K

The verb *shed* is not as frequent as the noun *shed*, and *fell* as the infinitive form verb is not as frequent as the past tense form verb, as in *fell trees* and *the tree fell*, respectively. If we take the findings based on the word family list into account, then a couple of words are expected to cause difficulties for the learners because of their low frequencies: *sacrifices, heed, errand* and *slender*. However, as was the case for pilot number one, the words in COLLEX 2 were tested also as single words, which means that this variable was controlled for.

[25] See Nation (2006) for details on the compilation of the list.

3.2.2.2 Material

In addition to the 65-item COLLEX 2, a 70-item single word test (SINGLEX) was administered. The items in this test were all the nouns figuring in COLLEX 2 together with those verbs or adjectives that were deemed difficult even for the informants of the present study. This judgement was made by myself together with two experienced university lecturers of English. Consequently, high frequency verbs like e.g. *set, make, run, drive, break, hit, put, do, draw* and *take* were not included in the single word test. The SINGLEX test was included to control for the possibility that the difficulty of a single word might influence learners' knowledge of a collocation which the word is included in. The format was the same as that used in the previous pilot, i.e. a multiple-choice test with three L1 words as choices (see section 3.1.4.2).

As opposed to COLLEX 1, COLLEX 2 was not divided into three parts, since this division in the previous version was governed by the frequency bands that the words were selected from.

3.2.2.3 Informants

The informants taking part in the study were 84 Swedish-speaking learners of English, out of which five indicated that they had other L1s than Swedish. They were all first term students of English at university level, with an average of eight years of English instruction behind them. At the time of testing, they were almost two thirds way through the first term of full-time English studies. In terms of perceived proficiency, these informants were on average on a slightly lower level than those taking part in the first pilot reported in subsection 3.1, and not as homogeneous. By using this learner group in the study, I intended to examine how first-term university learners of English might perform on my tests, and to administer the test to a slightly larger learner group than the one for COLLEX 1.

3.2.2.4 Research questions

On the basis of the results from the pilot of COLLEX 1, I concluded that the main problem was the relatively poor item quality. Therefore, one important aim of the COLLEX 2 administration was to test the new and hopefully improved items. In the light of the results of the pilot, the present test session was carried out with the following accompanying questions in mind:

1. Will a different selection method of test items in COLLEX 2 yield a test with better test reliability and item discriminatory values (internal consistency and item-total correlation values) than in COLLEX 1?
2. Is the level of difficulty of COLLEX 2 appropriate for first-term university students of English?
3. Is guessing frequent in COLLEX 2 and what effect does it have on test takers' scores?

Since test items were selected on a somewhat different basis it was hypothesized that the test would be more difficult and produce more reliable scores than the previous version, and therefore function better from a psychometric point of view. The principal reason behind this belief was the assumed improvement in terms of test item content and quality. In terms of an analysis of guessing, the guessing behaviour observed in COLLEX 1 was used as a benchmark for comparison.

3.2.2.5 Test administration

The test was administered to the 84 students at the beginning of a grammar revision lecture. Attendance at this lecture was not obligatory. The lecture was given as a help to the students approximately one week before they were due to sit a grammar exam. The test session was run in the following way. The two test parts were handed out to all students in an integrated test sheet at the beginning of the lecture. The students were told that the test was part of an on-going vocabulary research project and that their performance would in no way affect the grades in any of the courses they were taking. They were asked to do the test parts in the order they appeared on the test sheet, and not to go back to test part 1 once they had started test part 2. The order of the test parts were SINGLEX 2 followed by COLLEX 2. Most students finished the SINGLEX 2 part in 5 minutes, and the COLLEX 2 part in 10 minutes. Most students had finished both parts after 15 minutes and all students had finished the two parts after 20 minutes.

3.2.3 Results

The analysis of the collected data followed the procedures outlined in section 3.1.5 to a great extent. First, descriptive statistics were compiled, and the guessing behaviour was analysed. Second, an item analysis was carried out with the aim to investigate whether the new version of COLLEX was functioning better than the previous one in terms of item quality. In Table 3.9, the results on the single word test, SINGLEX, are presented.

Table 3.9 Results on SINGLEX 2 test of nouns, verbs and adjectives, run in November 2004 (N = 84).

Value	Total (70)
Mean	67.0
S.d.	2.58
Range	14 (57-70)

Following the trend from the previous pilot, the informants' scores on this test were very high with a mean score of 67.0. The Standard Deviation was strikingly low (2.58) which points to the fact that the sizable group of learners (N = 84) performed uniformly high on this test part of 70 items. Most words in the SINGLEX test had a facility value between .98 and 1.0. A closer examination of the facility values for the individual items revealed that certain words proved to be difficult for these informants. The words presented in Table 3.10 below proved particularly problematic. In the table, those words from the SINGLEX test with facility values of less than .9 are reported. It is evident that words like *heed, amends* and *foul* were not known to a high extent (IF values of .49, .56, and .62, respectively), and this is assumed to have an effect on the learners' knowledge of collocations that these words enter into. However, empirical evidence supporting this assumption will have to be presented.

Table 3.10 Words from SINGLEX 2 with the lower facility value than .9.

Item no.	Word	Item Facility
68	dial (verb)	.89
66	fell (verb)	.86
64	pursue (verb)	.82
65	shed (verb)	.74
9	slender (adj.)	.73
60	fuse (noun)	.70
70	foul (adj.)	.62
61	amends (noun)	.56
62	heed (noun)	.49

The least frequent word according to the word lists used, *slender*, was known by 73 per cent of the subjects in the study, which tells us that word frequency alone is not always the best predictor of word difficulty.

The scores on COLLEX are given in Table 3.11 below. One of the 84 subjects did not answer test items 41-65 and was therefore excluded from the analysis. The mean score was 52.0 and the standard deviation was 6.4. Comparing this mean score with the mean score produced by the informants in COLLEX 1, we may tentatively conclude that the present version of COLLEX was slightly more difficult. The mean score of the 19 subjects in pilot one corresponds to 86 per cent whereas the mean score of the 83 subjects in the present pilot corresponds to 80 per cent. However, since the content of the two versions of the test is slightly different it is difficult to say whether the test content or learner proficiency is the variable responsible for the difference in mean scores.

Table 3.11 Results on COLLEX 2 test of collocations, piloted in November 2004 (N = 83)

Value	Total (65)
Mean	52.0
S.d.	6.4
Reliability alpha	.82
Guesses (f)	838
Correct guesses (f)	449

In terms of overall reliability, COLLEX 2 produces acceptably reliable scores, as estimated through Cronbach's alpha; the internal consistency value of .82 means that the new version functions considerably better than the previous one. The main reason for this improvement is believed to be better item quality, but the fact that the present informant group is slightly more heterogeneous (s.d. 6.4) than the one in the COLLEX 1 pilot could also have played a part. The subsequent item analysis will show to what extent and in what way the items may have had a palpable impact on the improved, estimated reliability coefficient.

As to the guessing behaviour, the 83 subjects reported a total of 838 guesses. On average, this amounts to 10.1 guesses per learner. In comparison with the average number of guesses in the first pilot, the present subject group thus indicated that they guessed twice as many

times. We must however remember that the present test version consisted of 65 items compared to 60 in the previous version, which means 5 more items that could attract guessing behaviour on the part of the learners. As was the case for the first pilot, the proportion of correct guesses to total number of guesses was higher than for incorrect guesses. More correct indicated guesses (.54) occurred than incorrect ones (.46). The reason why the present learner group guessed to a higher extent is difficult to establish, but if we accept the assumption, supported by data in Table 3.3 above, that lower scoring learners will tend to guess more often than higher scoring learners, simply because their underlying ability to decide whether presented word combinations occur in English or not is not profound enough, then higher or lower ability in the construct measured could be the straightforward reason. Another reason could be that a more difficult test would attract more guessing than an easier one.

A correlation analysis was carried in which the informants' total scores were correlated with their number of guesses. In a similar fashion to that of the COLLEX 1 study, a Kendall's tau (τ) test[26] showed that there was a significant negative relationship between the number of guesses indicated and the number of points on the test, $\tau = -.33$, $p = < .01$. This means that there is a negative relation between the total scores of the learners and the number of guesses that they report.

In order to more closely examine the guessing behaviour, something which is warranted since guessing is likely to affect the reliability of the test scores, the subjects were divided into three groups according to their total score on the test. An effort was made to create as equally-sized groups as possible, and the groups are again referred to as 'low group' (N = 28), 'mid group' (N = 27), and 'high group' (N = 28). The data for the three groups are presented in Table 3.12 below, and the results show that the pattern observed in pilot 1 is apparent also in the present study.

Table 3.12 Reported guessing behaviour with respect to three total score groups.

Group	N	Mean score	S.d.	Score range	Mean no. of guesses	S.d.	Per cent correct guesses
High	28	58.89	2.35	55-62	6.25	5.40	67%
Mid	27	52.07	1.69	49-54	11.26	8.58	55%
Low	28	45.11	4.24	34-49	12.82	8.43	46%

The subjects in the mid and low scoring groups guessed to a much greater extent than the high scoring group. The mean number of guesses of the high group was 6.25 with a standard deviation of 5.40. The mid group produced a mean of 11.26 guesses, and a standard deviation of 8.58. The mean for the low group was 12.82, with a standard deviation of 8.43. As was the case in pilot 1, the high scorers were more accurate in their guesses (67%) than mid scorers (55%) and low scorers (46%). The guessing means of the three groups were subjected to an ANOVA. This analysis revealed a significant effect of group affiliation on number of guesses, $F(2, 80) = 5.70$, $p < .005$, $\eta^2 = .12$. A post hoc test (Gabriel) showed that the high group was

[26] This test was used since the data contained a number of tied ranks. A Spearman's rho (ρ) test gave an even higher significant, negative correlation of $\rho = -.40$, $p = < .01$, but it may produce inflated values when many ties exist.

significantly different from the mid and the low groups, respectively, whereas no significant difference existed between the mid and the low group in terms of mean number of guesses.

The items that attracted the highest amount of self-reported guessing are presented in Table 3.13 (>30 guesses).

Table 3.13 Items in COLLEX 2 that attracted a high degree of self-reported guessing and the observed IF values for these items.

No.	item	Total no. of guesses (correct/incorrect)	Item Facility
64	pay heed – show heed	49 (31/18)	.67
49	take root – make root	48 (23/25)	.54
42	kick a habit – undo a habit	45 (22/23)	.59
4	exercise one's rights – employ one's rights	36 (15/21)	.71
52	dress a wound – lay on a wound	36 (9/27)	.31
21	bring charges – run charges	34 (17/17)	.51
40	foul weather – poor weather	32 (9/23)	.22
62	blow a fuse – strike a fuse	31 (16/16)	.69

Interestingly, some of the single word items presented having low facility values in Table 3.13 above: *heed, fuse* and *foul*, occur in the items that attracted the highest number of guesses in COLLEX 2. From this we may conclude, on the one hand that poor knowledge of the component words of a collocation makes the recognition task in COLLEX 2 difficult under the assumption that many guesses on an item indicate that the learners found it difficult. On the other hand, some words that were not known to a great extent in the single word test were parts of collocations that proved to be answered correctly by most of the learners. For example, *make amends* was chosen by almost 90 per cent of the learners over the distractor *do amends* even though only around 50 per cent answered the word *amends* correctly in SINGLEX. Thus, the effect of unknown words on recognition of word combinations with these words as component parts is rather inconclusive. It needs pointing out, though, that the tasks in the two tests are slightly different. Both SINGLEX and COLLEX are receptive recognition tests, but SINGLEX requires informants to map an L2 word onto one out of three L1 words, whereas COLLEX requires informants to choose one out of two juxtaposed L2 word combinations. It is therefore conceivable that a learner may fail to find an L1 meaning for an L2 word, but manages to recognize a collocation consisting of this word together with another when presented with it. In this sense, the task in COLLEX can be seen to be slightly less demanding than the task in SINGLEX.

The results from the item analysis can be seen in Appendix 3B. Item Facility (IF) values together with Item-total correlation (ITC) values were computed for each of the 65 items of the test. When it comes to the facility values, COLLEX 2 seems to perform slightly better than COLLEX 1. Only three items display zero variance and the mean IF value is .80. This is a clear improvement compared to COLLEX 1, but still too high to be satisfactory for a norm-referenced test.

In terms of item-total correlation values, we can also see an improvement here compared to the previous test version. Using Ebel's (1979) guidelines for item quality, we get the distribution presented in Table 3.14 below. We can observe a clear overall increase in items above the .19 cut-off mark for 'poor items' (cf. Table 3.6). The 'very good items' category

increased from making up 12 per cent of the total number of items in COLLEX 1 to 18 per cent in COLLEX 2.

Table 3.14 Distribution of test items from COLLEX 2 into categories of discrimination power following Ebel (1979).

Item-total correlation guidelines	.40 and higher very good items	.30 to .39 reasonably good item possibly subject to improvement	.20 to .29 marginal items in need of improvement	below .19 poor items which need to be revised or eliminated
Items from COLLEX	3, 4, 16, 25, 26, 27 34, 36, 37, 48, 50, 55	5, 6, 7, 11, 12, 19, 22 23, 42, 46, 52, 62, 64	10, 15, 18, 21, 24, 28 39, 47, 51, 59, 60, 65	1, 8, 9, 13, 17, 20, 29, 3(31, 32, 33, 35, 40, 41, 43 44, 45, 49, 53, 54, 56, 57 58, 61, 63 Value 0.00 (no variance): 2, 14, 38
Number of items	12	13	12	28 (25+3)
per cent of total number of items	18%	20%	18%	42%

The 'reasonably good items' increased from 6 to 20 per cent, and the 'marginal items' group increased from 10 to 18 per cent. Furthermore, even though the large number of items (42 per cent) having item-total correlation values lower than .19 shows that yet further improvements must be made in terms of the discriminatory power of the test items, COLLEX 2 is clearly a step in the right direction towards a well-functioning test.

Summing up the results, in relation to my research questions, I have observed that both the overall test reliability, and the discriminatory power of the individual items are improved in COLLEX 2 compared to COLLEX 1, much along the lines of the hypothesis presented prior to the test administration. The level of difficulty is higher in COLLEX 2 than in COLLEX 1, but it is not possible to judge what effect the proficiency level of the informants had on the performance. In terms of guessing, the pattern from the COLLEX 1 pilot was similar also in this study, with significant negative correlations between total test score and number of indicated guesses. High scoring informants guessed significantly fewer times, and more accurately, than the mid and the low scorers.

3.2.4 Discussion

The overall aim of this second study was to develop a more reliable test. My first research question focused on the less restricted item selection method employed and whether this would help in making the test more reliable and give the test items better discriminatory power. The results were in this regard promising. The overall reliability of COLLEX 2, as measured through Cronbach's α, was observed at .82, which is satisfactory in comparison with the reliability of COLLEX 1 (Cronbach's α = .54). Thus, the construct aimed to be measured through COLLEX—receptive recognition knowledge of English collocations—was measured reliably. However, even though an α value of .82 is acceptable, it should be

possible to increase this value even further in subsequent testing sessions. The aim is to achieve a value of .9 or more. The question here, though, is what caused the increase in reliability.

As was discussed in Chapter 2, there are several factors that may affect a test's capacity to produce reliable scores. The length of a test is one of those factors. Compared to COLLEX 1, which consisted of 60 items, COLLEX 2 was slightly longer, containing 65 items. A longer test is generally more reliable than a shorter one as long as the added items are of similar or better quality than the original set of items. The item analysis in the present study showed that the overall item quality of COLLEX 2 was better than that of COLLEX 1, which seems to support the fact that I managed to add well-functioning items. However, a closer look at the items in COLLEX 2 is warranted. This will be done at a later stage in this discussion section.

Another factor that affects reliability is the level of homogeneity of the tested group of individuals. In section 3.2.2.3 above, it was estimated that the present informant group was on a slightly lower proficiency level than the group taking COLLEX 1. It follows from this that the present group might also be more heterogeneous than the COLLEX 1 group. A direct comparison is unfortunately not possible since the two groups sat different test versions, but the standard deviation of the scores produced by the informants in this study (S.D. 6.4 for N =83) is certainly greater than that of the informants in the previous study (S.D. 3.3 for N =19). Thus, the variability in the data is greater in COLLEX 2 than in COLLEX 1, with a reservation made for the unequal test lengths. If the assumption that the learner group in the present study is less homogeneous than the one in study 1 holds, then this may partly have caused the observed higher reliability values.

A third factor that may have caused the higher reliability level is better item quality. This brings us to the results of the item analysis, and it also touches upon research question two, which addressed the level of difficulty of the test for the student group used. In Table 3.15 below, ten items that were used in both studies are juxtaposed. Some of them are exactly the same, whereas some of them were slightly modified for the present study. Changes to items in COLLEX 2 compared to COLLEX 1 are shown through italics.

Table 3.15 A comparison of items used verbatim in COLLEX 1 and COLLEX 2, and items modified for COLLEX 2.

	COLLEX 1 (N = 19)			COLLEX 2 (N = 83)		
	item	IF	ITC	item	IF	ITC
A	run a business – drive a business	1.00	0.00	run a business – drive a business	1.00	0.00
B	keep a speech – give a speech	1.00	0.00	*hold* a speech – give a speech	.93	.16
C	pull a conclusion – draw a conclusion	1.00	0.00	*take* a conclusion – draw a conclusion	.95	.43
D	finish a fire – put out a fire	1.00	0.00	*turn out* a fire – put out a fire	.90	.40
E	speak a prayer – say a prayer	1.00	0.00	*tell one's prayers* – say *one's prayers*	.88	.55
F	drop count – lose count	.84	.62	drop count – lose count	.90	.18
G	catch a disease – receive a disease	.89	.29	catch a disease – receive a disease	.96	.33
H	set a deal – strike a deal	.47	.39	set a deal – strike a deal	.54	.45
I	do damage – make damage	.68	.27	do damage – make damage	.54	.29
J	lay a wound – dress a wound	.47	.50	lay *on* a wound – dress a wound	.31	.38

IF = Item Facility; ITC = Item-Total Correlation

Items A-E had an item facility of 1.00 in COLLEX 1. For COLLEX 2, item A, *run a business – drive a business*, was kept intact. The reason for this was that the item was still believed to be a good item, and that it would attract wrong answers with a slightly less proficient student group. As can be seen in the table, though, this did not happen in COLLEX 2. However, from a testing perspective, starting a test with a couple of easy items can have a positive effect on test-takers in that they feel confident. If a test starts with very difficult items, you run the risk of putting test-takers off, and they might lose interest in the test. As for items B-E, the modification made in the item resulted in lower facility values as well as good to very good item-total correlation values. Item F, however, displays a case where the same item, *drop count – lose count*, was used in both studies, and where better values were observed in COLLEX 1 than in COLLEX 2. The item facility value increased somewhat, from .84 to .90, in COLLEX 2, and the item-total correlation value decreased from .62 to .18. This means that the difficulty of the item stayed more or less the same, whereas its discriminatory power decreased. It is difficult to say why this happened other than some low scorers getting the item right and some high scorers getting the item wrong. Items G-H exemplify unmodified items that got a slightly higher item facility value in COLLEX 2, but an increase in discriminatory power. Item I is an example of an unmodified item in COLLEX 2 with a lower item facility value, and a slightly higher item-total correlation value. This is generally what we would aim to achieve with all items in COLLEX. The intention is to make the items slightly more difficult coupled with a higher discriminatory value. Item J, finally, exemplifies a modified item in which a lower facility value is observed, which is positive, but which at the same time displayed a lower item-total correlation value, which is negative.

The above analysis goes to show that it is possible to arrive at better items by using information from an item analysis, but it also shows that, from a norm-referenced testing perspective, already good items will sometimes function worse with a similar but not identical test group. In addition, there seems to be an intricate interplay between the item facility and the item-total correlation of an item. Making a too easy item more difficult may at the same time make it less discriminatory between high and low total scorers. Alderson *et al.* (1995) discusses this relation and points to the fact that only with items that have facility values

between .33 and .66 is it possible to get a maximum item-total correlation of 1.0. Furthermore, items with very high or very low facility values may still be good items if they have a relatively high discriminatory value. Subsequently, item 11 in COLLEX 2 (see Appendix 3B) *hit a number – dial a number*, which has a facility value as high as .99 but an item-total discrimination value of .30, can still be considered to be a good item. Similarly, items 12 *make an effort – commit an effort* (IF = .98, ITC = .30), 26 *pay a visit – do a visit* (IF = .95, ITC = .40), and 10 *strong competition – hard competition* (IF = .25, ITC = .29) are all good items in this respect.

The closer look at the items in COLLEX 2 has shown that it is essential to trial the items of a test with different learner groups, since an item might function well with one group but not as well with another. It must be stressed, though, that on the whole, COLLEX 2 is a clear improvement compared to COLLEX 1 in terms of item quality. The mean facility value (.86 > .80) is lower and the mean item-total correlation value is higher (.10 > .23.). However, the answer to research question two is closer to being negative than affirmative. The test is still judged to be slightly too easy from a norm-referenced testing point of view.

The third research question addressed the guessing behaviour in COLLEX 2. I asked whether guessing is frequent and what possible effects guessing might have on the test-takers' scores. It was found that guessing was more frequent among the 83 informants in this study than among the 19 informants of the previous study, if measured by mean number of guesses. COLLEX 2 attracted a mean of 10.1 guesses whereas the informants taking COLLEX 1 reported a mean of 4.7 guesses. The increase in number of guesses is believed to be due to COLLEX 2 being more difficult than COLLEX 1, and also that the students taking COLLEX 1 were in general slightly more proficient than the students taking COLLEX 2. This is assumed to have resulted in more guessing in COLLEX 1, since less proficient students arguably have a smaller knowledge base to rely on. The possible effect that the guessing might have on the scores was investigated partly through correlating the students' total scores with their reported number of guesses, and partly through creating three student groups (low, mid, and high) based on the total scores and comparing the means of these three groups. The result of the correlation analysis followed the result from the pilot of COLLEX 1 in that a significant negative correlation was found between scores and number of guesses. The result of the mean comparison by way of an ANOVA pointed to a significant difference between the means, and a post-hoc test showed that the high group was significantly different from both the mid and the low group. There were differences in absolute numbers in the first pilot but these were not significant. Thus, two studies have shown that high scorers report fewer guesses than low scorers and they are also more successful guessers than low scorers, and that these differences were statistically significant in the present study.

The conclusion we can draw from this is that lower proficiency learners will tend to guess more often and less successfully than higher proficiency learners on COLLEX. The guessing means of between 5 for a 50-item test and 10 for a 65-item test will prove useful information if and when we consider introducing some sort of correction for guessing formula in the test. Such a formula would deduct points according to a logarithm based on the number of incorrect answers and the number of choices available in an item. If a learner guesses about 10 times and is successful around 50 per cent of the cases, then this means that the score will hypothetically be inflated by around 5 points. Before, we introduce a correction formula, however, it is essential that we aim for an even higher reliability than the one obtained for COLLEX 2.

A final point to be addressed in this discussion section pertains to the nature of the test task in relation to test validity. In COLLEX 2, for each test item, an informant is asked to choose one out of two word combinations. The test instructions furthermore specify that the word combination that is deemed the most frequent, and also believed to be used by native speakers of English, should be chosen. From a validity perspective, this leaves us with an intriguing issue. There is no function present in the test that verifies that the informants actually understand the meanings of the word combinations presented to them. Consequently, in theory, informants may select an alternative based on other grounds than meaning knowledge. The kind of meaning knowledge I refer to here is first and foremost knowledge of appropriate L1 translation equivalents. Even though COLLEX is a receptive recognition test, which does not overtly ask informants to verify that they know the meaning of the English word combinations in their own L1, the test rests on the assumption that the word combinations that the informants correctly identify as frequent, native speaker-used word combinations, are also combinations for which the informants have some sort of deeper knowledge. This so-called deeper knowledge, I argue, goes beyond mere recognition of the L2 word combination, and is likely to involve to some extent semantic, grammatical and usage aspects (cf. Nation's (2001) descriptive model of word knowledge, accounted for in Chapter 2). This claim, however, needs to be corroborated and tested in some way. We will therefore set up an experiment in our next study that addresses this problem.

3.2.5 Conclusions

In conclusion, the results of the administration of COLLEX 2 reported here are encouraging. The overall aim was to try to improve the reliability of the test. Furthermore, I also wanted to see whether the level of difficulty of the test was appropriate for beginner university students of English. I would argue that the reliability observed is satisfactory, and that this can also be taken as a support for the construct validity of the test, but that the test may still be slightly too easy for the tested population. In the next chapter, in order to create a clearer view of the issue of difficulty, a further revised COLLEX test will be administered to around 100 informants studying English at different levels in the Swedish university system. In order to address test validity, an experiment will be conducted in which two versions of COLLEX will be given, one monolingual and one bilingual. The purpose is to see if the identification of target collocations is facilitated through the insertion of a Swedish translation of the target collocation in each item. Furthermore, a complementary receptive collocation test format called COLLMATCH will be piloted.

4 Investigating the reliability and validity of COLLEX 3 and COLLEX 4, and developing the COLLMATCH test format

4.1 Investigating the validity of COLLEX and developing COLLMATCH

4.1.1 Introduction

In the previous chapter, I described the administration and results of two studies involving a collocation test format called COLLEX. This chapter reports a study in which a further modified version of COLLEX is administered as an attempt to address issues emerging in the previous two studies. In addition, a second test format, named COLLMATCH, will be developed and trialled together with COLLEX.

4.1.2 Background

We are now in a position to take stock of the results from the two initial studies reported in Chapter 3. The first study was conducted on a small group of 19 third term teacher student learners, whereas it was possible to test a larger group of learners, 84 first term English learners, in the second study. The results of the first study showed that COLLEX 1 was a practical test, easy to administer, in which a fairly large number of items (60) could be tested in a short period of time. There was also anecdotal evidence that the format was appreciated by the participating students as a "new" and fresh way of testing vocabulary knowledge. However, there were a number of problems with the test. One of the main problems was the observed poor item quality. This resulted in an unacceptably low reliability value, and lack of discriminatory power of the individual test items. It was also clear that the first test version on the whole was too easy for targeted learner group. Far too many items displayed zero variance. The guessing behaviour of the subjects was elicited through a self-report function in the test itself, and it was found that the tested learners guessed on average 5 times in the 60-item test, a mean frequency which was largely deemed uncontroversial, but warranting further investigation. A relation was observed between total score on the test and indicated number of guesses, in that high scorers guessed less often than low scorers. Despite the apparent problems, considering the relatively small group of informants tested, COLLEX was still believed to hold promise as a test tool.

In the second study, a less restricted item selection method was used for COLLEX 2. Arguably, there was reason to believe that the item selection method in study 1 might have made a large number of items too unchallenging for advanced learners. The selection criterion specifying that the verbs used in each item must not come from a lower frequency band than the noun was thought to have resulted in too many weak distractors (pseudo-collocations). In addition to verb + NP collocations, a smaller number of other types of word combinations was used, such as adjective + noun collocations. However, the majority (>75%) were still verb + NP combinations. The results of this second test administration were encouraging. The reliability coefficient of the scores was considerably improved, from a disappointingly low α of .54 in study 1 to an acceptable α of .82. The number of items with zero variance (too easy

items) decreased from 40 per cent in study 1 to only 4 per cent in study 2, and consequently the discriminatory power of the test items as measured through item-total correlation estimations increased considerably. The analysis of the informant guessing behaviour largely showed the same pattern as that of study 1, with significant negative correlations between total scores and indicated number of guesses, although guessing was more frequent on average, and the mean number of guesses of the high scoring group was significantly lower than that of the low scoring group. This was believed to have been caused by more difficult items in conjunction with slightly less proficient learners.

Even though there is still room for improvement in COLLEX, we now have an acceptably reliable test tool with which we may both ask questions about learners and about the test tool itself. As was stated in section 2.5, reliability is a necessary but not sufficient condition for validity.

The present chapter tries to characterize the receptive collocation knowledge of groups of learners at different stages in the Swedish education system. The question is if a further modified version of the COLLEX test is sensitive enough to pick up any existing differences. As was pointed out in the summary of the literature review, there is a tangible lack of studies that compare learners at different proficiency levels in an education system. Biskup (1992) compared different L1 groups: German and Polish university students; Bahns and Eldaw (1993) indeed used university students in their first up to their third year of English study, but made no attempt to compare the learners in different years of study; Farghal and Obiedat (1995) did attempt a kind of comparison, but between teacher students of English and general English students; Bonk (2001) made no cross-sectional comparison in his study; Mochizuki (2002) used first year students only, and, finally, Barfield (2003), who used undergraduate and postgraduate university students, made a comparison based on field of study, such as medical students, fishery students, etc. Thus, a study that examines possible differences in connection to level of formal study is clearly warranted, and it would be interesting and worthwhile since we know little about how L2 collocation knowledge develops (Schmitt 2000).

In carrying out such a study, we are in fact collecting validation evidence through what Bachman calls ‘criterion validity’ (Bachman 1990:248). I will design a study that examines potential differences in terms of receptive recognition knowledge of English collocations among groups of individuals who are assumed to possess different levels of language ability. For this purpose, different Swedish university learner groups, as well as native speakers of English will be targeted. My aim in this regard is to investigate the discriminatory power of COLLEX. My assumption is that native speakers of English will be more proficient than Swedish learners of English in recognising English collocations. Furthermore, it is conversely assumed that Swedish second year students of English will be more proficient than Swedish first year students of English in recognising English collocations.

In addition to administering version 3 of COLLEX, a further test format of English collocation will be developed. Developing a second test format would make it possible to compare COLLEX with a similar but not identical test, preferably a test design that would entail a slightly different task for the test taker. This could then be administered together with COLLEX as part of a test battery. In general, having two different tests at our disposal for investigating learners’ receptive knowledge of English collocations was considered a worthwhile aim. Therefore, in this chapter, I will introduce COLLEX 3, and a first version of a new test format called COLLMATCH. First, I will present the methods used to construct the

two tests, and secondly, I will account for the results of an empirical study evaluating learners' performance on the two tests. Finally, I will evaluate the two formats. This evaluation will guide what steps to take next in the development of well-functioning tests of English collocations, and the mapping out of learner performance on these tests.

4.1.3 Methods

4.1.3.1 COLLEX 3

4.1.3.1.1 Item selection

For the third version of COLLEX, the best items from COLLEX 2 were selected for inclusion. This process entailed an analysis of the item facility as well as the item-total correlation value for each item. In general, the closer the item facility value is to .5, and the higher the item-total correlation is, the better (see section 3.2.4). Out of the 65 items from COLLEX 2, 47 items were chosen following these criteria. In addition to these 47 items, 2 new items were constructed, and 1 item from COLLEX 1 was re-used after due modification. This resulted in a 50-item test.

4.1.3.1.2 Introducing z-scores

In order to check the strength of the collocations, or put more accurately, the collocation 'significance', in COLLEX 3, so-called z-scores were computed. By computing z-scores I will be in a better position to decide how significant my test items are, and their relative importance. A z-score[27] is essentially a measure of how far a given value is from a mean. I will below account for the computation of z-scores, but first the rationale behind this approach needs to be unpacked. Thus far, I have relied on reported collocations from a collocation dictionary (Crowther *et al.* 2002), in turn based on analyses of the BNC, but I have neither known the specific strength of relationship between the component words of the collocations, nor the relative frequency of different collocations.

The description of the z-score calculation given here largely follows Oakes (1998:163-166). The notion of significance in relation to collocations is linked to the concept of statistical probability. A result is significant when its occurrence by chance is sufficiently low, as decided by a so-called alpha level. This alpha level, denoted *p*, varies across different research fields and research paradigms, but in linguistic research it is often set to .05. This value then serves as a cut-off point meaning that a there is only 5 chances in 100 that a result occurs by chance. Specifically, the z-score calculation for significant collocations rests on the probability of one lexical item (the node) co-occurring with another word within a specified distance or span being greater than chance expectancy. In order to be able to compute a z-score, the following data must be defined:

[27] Berry-Rogghe (1973) is credited with the z-score calculation.

Z: the total number of words in a text
A: a given node occurring in the text F_n times
B: a collocate of A occurring in the text F_c times
O: number of co-occurrences of B and A
S: Span size: the number of items on either side of the node considered as its environment

First, the probability of B co-occurring O times with A, if B were randomly distributed in the text, must be computed. Secondly, the difference between the expected number of co-occurrences and the observed number of co-occurrences is computed. The probability of B occurring at any place where A does not occur is expressed by:

$$p = F_c / (Z - F_n)$$

The expected number of co-occurrences is expressed by:

$$E = p\, F_n\, S$$

The formula for deciding whether the difference between the observed and the expected frequencies is given in Figure 4.1 below. In the formula, let $q = 1 - p$.

$$z = (O - E) / \sqrt{E\, q}$$

Figure 4.1 The z-score computation formula, adapted from Oakes (1998:163)

The z-scores were computed in the SARA software system developed for use with the BNC. The following passage gives an account of the way z-scores were retrieved. The search for collocations starts with a so-called 'word query', in which a particular word of interest – the node – is the starting point. For example, if we are interested in the verb *shed* and its collocates, a search for *shed* as a lemma can be made. This particular lemma yields 1,364 hits in the BNC. The concordance lines for these hits can subsequently be downloaded and analysed. If we are interested in compiling a list of the collocates of any of the verb forms of the lemma *shed*, then the collocation function is selected. This presents the collocation dialogue box. A screen shot of the dialogue box is shown in Figure 4.2 below.

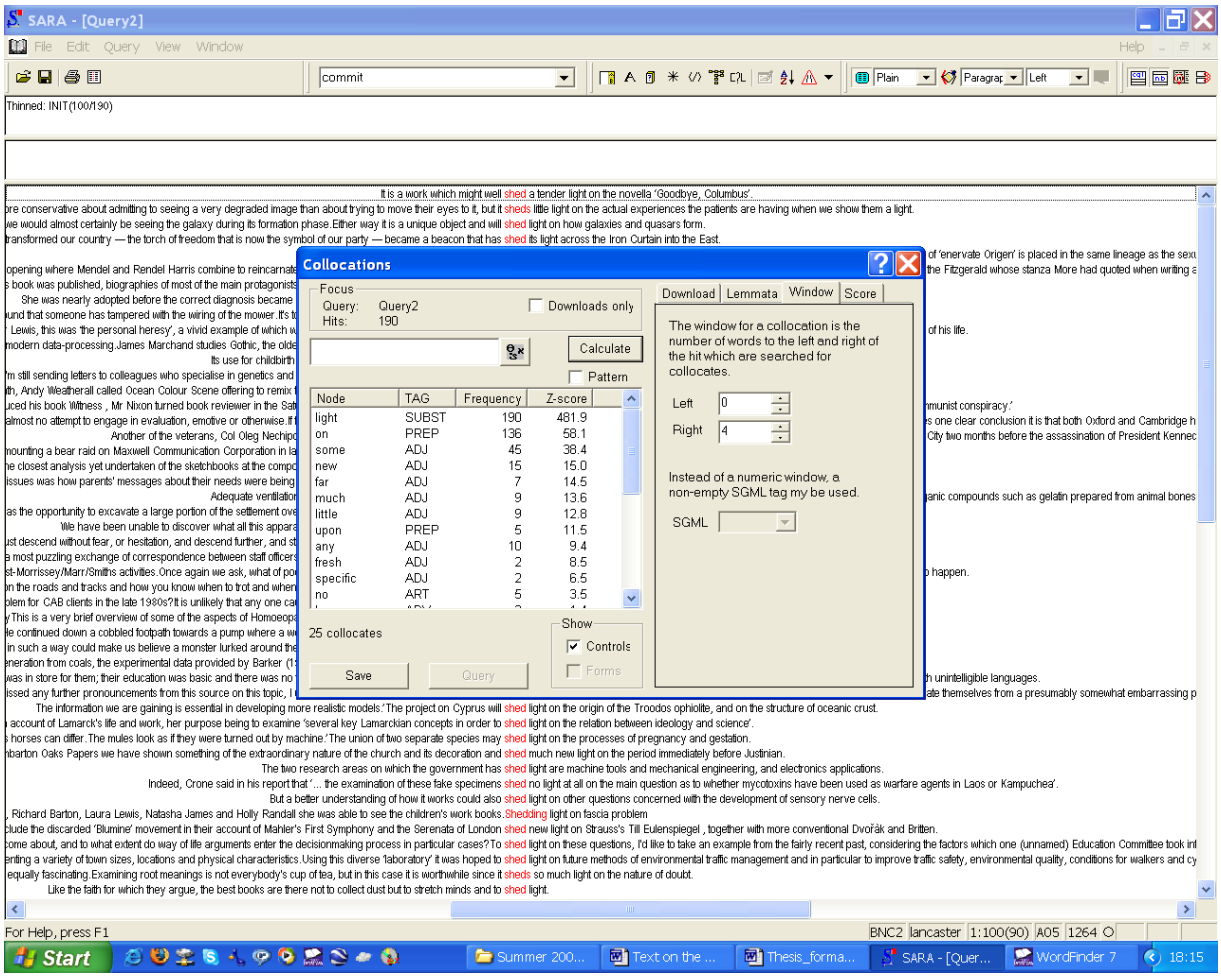

Figure 4.2 Screen shot from the SARA collocation dialogue box for BNC World Edition.

The screen shot shows the listed lemma collocates of *shed*, together with the absolute co-occurrence frequency and the z-score. The span is selected in the 'window' function, which in the screen shot was set at 'left' = 0, and 'right' = 4[28]. The result of the calculation is that there are 190 co-occurrences of the verb lemma *shed* and the noun lemma *light*. Furthermore, the z-score is 481.9, which is clearly significant. As a rule of thumb, z-scores of ≥ 3 are generally considered significant (see Barnbrook 1996:96). More specifically, in order to reach statistical significance based on a two-tailed test with a cut-off value of $p < .01$, we have to arrive at a z-score of at least 2.58 (see Oakes 1998:8-9 for a worked example). All combinations intended

[28] I discovered an error/bug in the way SARA calculates collocation frequency in the BNC World Edition. A discrepancy exists between the frequency value displayed for a collocation pair in the collocation dialogue box and the frequency value displayed in the concordance line mode. In correspondence with Oxford University Computing Service (Ylva Berglund Prytz, personal communication), the error was explained by the unfortunate fact that the collocation display seemingly miscalculates the collocation frequency by using a span which is one position smaller than the one that the user enters in the 'window' (span) function. Consequently, in order to arrive at the right collocation frequency number in the collocation dialogue box (seen in Figure 4.2), a span that is one position wider than the one you actually want must be entered. For example, a chosen span of 'right' = 4 actually calculates the frequency value of one position smaller, namely 'right' = 3.

to function as target collocations in COLLEX 3 were checked against this cut-off score, and values well over this minimum z-score were sought for. Conversely, those combinations intended to function as distractors (pseudo-collocations) were checked in order to avoid using significant collocations for this category. In some cases, z-scores were observed at a higher level than expected, based on my own intuition. In these cases, concordance lines were inspected in order to ascertain the circumstances behind the unexpected value. In some cases, other constructions than the one tested give rise to a high z-score. The items included in COLLEX 3 and their computed z-scores are presented in Appendix 4A.

As far as single word frequencies are concerned, compared to the account given of the words in COLLEX 2 (see Table 3.7), only 3 'new' words of lower frequency than 4K were used in COLLEX 3, based on Kilgarriff's (1996) list: *ripe* (7K+), *mature* (5K), and *awake* (6K). However, comparing these values to those found in Nation (2006), we get 4K, 3K, and 3K, respectively for these words. Thus, it is assumed that university students of English will have few problems with the single words that make up the collocations in COLLEX 3. However, in order to control for single word frequency, the words used in COLLEX 3 were tested separately as was done in the previous study.

4.1.3.1.3 Introducing a bilingual test format

A further feature of COLLEX 3 needs to be addressed. In the discussion section of the previous study I pointed at the fact that we cannot know whether informants sitting the COLLEX test format respond with the intended target item in mind. Each item in COLLEX is aimed at targeting a frequent English collocation. I cannot, however, be sure that the concept captured by the English collocation form is accessed by the informants as they process a test item. An alternative test task construction in COLLEX might therefore be called for. One way of addressing the issue is to introduce a bilingual test format. This format would consist of the same kind of basic item as in COLLEX 1 and COLLEX 2, but with a Swedish translation of the intended target collocation added. An example of a bilingual test item would look like this:

1 (be en bön) say a prayer tell a prayer

Figure 4.3 A test item in the bilingual version of COLLEX 3.

In the item above, the item number is followed by a Swedish translation of the target collocation intended to be captured. This is then followed by two proposed sequences. The advantage of introducing the Swedish sequence is that I can be sure that the informant is processing the same concept as intended by me as test constructor. This arguably increases the validity of the test. However, a disadvantage is that this effectively complicates the administration of COLLEX to informants whose L1 is not Swedish.

4.1.3.2 Developing an alternative test format: COLLMATCH

4.1.3.2.1 Test format

In general, the creation of an additional test format of receptive collocation knowledge would have two main positive effects. Firstly, one test could be used as potential concurrent validity support of the other. Secondly, in case of one test format not functioning well from a psychometric point of view, a second format would hopefully not suffer from the same shortcomings.

In creating a second test format aimed at tapping receptive collocation knowledge, my intention was to create a test that was easy to administer, that contained a large number of items, and that would produce meaningful and analysable interval data. A further aim was to develop a format and a task that differed to some extent from that of COLLEX, since it seemed futile to construct an identical test. I decided to use a grid format (see e.g. McCarthy 1990). In the format henceforth called COLLMATCH (collocate matching), the test taker is presented with a number of grids, each consisting of a 3 x 6 field design. I again decided to primarily focus on verb + NP combinations, but also to include a small number of adjective + noun combinations. An example of the COLLMATCH format can be seen in Figure 4.4 below. The grid consists of 3 verbs and 6 noun phrase objects. In each grid, an attempt was made to choose three verbs that shared some semantic feature. In the example below, the verbs *drop*, *lose* and *shed* can all be seen to share semantic properties that have to do with 'the release, volitional or not, of an object from another object or person'. The test taker is asked to indicate which of the 6 objects each verb felicitously may combine with. The number of possible combinations is not known to the test taker and in theory all or none, and every possible number in-between, is possible. The same object may be combined with more than one of the three verbs. Therefore, it is not possible to arrive at the right answer by a process of elimination. Each and every of the six alternatives above the grid must be tried for a potential match with all three of the words to the left of the grid. The instruction asks the informants to put a cross in the intersecting box of those words they think form combinations that exist in frequent use in English.

	charges	patience	weight	hints	anchor	blood
drop						
lose						
shed						

Figure 4.4 Example of a COLLMATCH 1 grid.

Just as in the COLLEX format, this format is a measure of receptive recognition knowledge of English collocations. However, the cognitive effort involved is believed to be somewhat more demanding than the COLLEX format, since the number of alternatives in COLLMATCH is much larger. In each grid, there are 18 items. Thus, each grid produces a fairly rich set of data. To a great extent, the format task can be seen to elicit answers to the question: '*What can be* V-ed?' Thus, based on the items in the grid above, the questions would be: *what can be dropped*?; *what can be lost*?; *what can be shed*? This should give an indication of learners' knowledge of the lexical restrictions, motivated or arbitrary, that must be abided by, if native-like sequences are the norm (see Howarth 1998a; Stubbs 2001). In some grids, the combinations are overlapping in the sense that two or even all three verbs may

share the same object. An example of this can be seen in the grid above, where both *shed + weight* and *lose + weight* are possible combinations. Since some of the verbs may enter into combinations in which the verb does not display its most common core meaning, the format can also been seen as measuring knowledge of word polysemy to some extent. Based on a logical *a priori* analysis of the test task, the COLLMATCH grid format can be argued to require the following from a test-taker:

a) recognition of the 9 words that make up each grid;
b) some degree of knowledge of the meanings of the 9 words that make up each grid;
c) a judgement about the potential relationship between the 9 words that make up each grid in terms of 18 possible combinations.

Thus, the receptive matching task in COLLMATCH is in fact fairly complex. In a similar attempt to create a vocabulary test of Catalan that measures both vocabulary size and quality of vocabulary knowledge, Vives Boix (1995) argues that the task in the test format called the Association Vocabulary Test (AVT) is not strictly a passive one, but that it rather forces the test taker to activate the two words of an item "in a deeper fashion that makes it QUASI-PRODUCTIVE." [upper-case letters from source retained] (p. 82). She furthermore claims that the test taker needs to know the specific meanings that make an association between the item words possible, in addition to knowing the meanings of the component words.

càrrec	:	alcalde	[a post : town mayor]
cua	:	gat	[tail : cat]

Figure 4.5 Example of test items from the Association Vocabulary Test (Vives Boix, 1995). Text within square brackets are my additions.

A claim like that made by Vives Boix above is difficult to test empirically, but other researchers give voice to supporting arguments. Melka argues, in a discussion about the complexity of what is involved in knowing a word, that certain degrees of knowledge, such as knowing the various meanings of polysemous words and also knowing collocations or idioms, could be labelled as "higher degrees of familiarity, close to productive knowledge" (1997:85).

4.1.3.2.2 Item selection

As with the COLLEX format, the items used in COLLMATCH are predominately words of high frequency. In version 1 of the format, the following verbs and adjectives were used as the word components to the left of the test format grid: *break, hold, keep, drop, lose, shed, say, tell, speak, beat, strike, perform, throw, cast, draw, take, make, pay, fair, blonde, light, hard, tough,* and *heavy*. Collocates of these verbs and adjectives were retrieved from the BNC, in the same fashion as for the COLLEX items. Z-scores were checked both for the intended target collocations as well as for the intended pseudo-collocations. Out of the 72 words in COLLMATCH 1, 9 have a lower frequency than 5K according to Kilgarriff's (1996) list. However, if we also retrieve values from Nation's (2006) list, we arrive at the comparison shown in Table 4.1 below. As can be seen in the table, a word like *drinker* ends up at the very high end of the frequency list when it comes to Nation's word family list, whereas *headway*

occurs as infrequently as in band 8K. It is likely that some of these words will pose problems to learners, and consequently, they will be incorporated in the test of single word knowledge that accompanies COLLEX and COLLMATCH.

Table 4.1 Comparison of word frequencies between a word lemma list and a word family list based on the BNC.

Form	Kilgarriff's (1996) lemmatized BNC list	Nation's (2006) word family BNC list
blonde	6K (verb)	5K
patience	6K (noun)	2K
anchor	7K+ (noun)	5K
sway	7K+ (noun)	6K
farewell	7K+ (noun)	6K
amends	7K+ (noun)	4K
headway	7K+ (noun)	8K
precaution	6K (noun)	4K
drinker	7K+ (noun)	1K

The items included in COLLMATCH 1 and their computed z-scores are presented in Appendix 4B, and the test version itself can be found in Appendix 4C.

4.1.3.3 Material

The material used in the study consisted of a 3-piece test battery. The two main parts of the battery were a 50-item COLLEX 3, and a 144-item COLLMATCH 1. The third part was a 40-item test of single word knowledge, called SINGLEX 3, and just like in the previous two studies, it served as a control for the informants' knowledge of the component words featured in COLLEX and COLLMATCH.

SINGLEX 3 contained single words, mostly nouns but also verbs and adjectives. Only words which were expected to present problems to the university level informants of the study were included. This selection process was carried out by myself together with two experienced university lecturers of English. The format was a multiple-choice format with 3 Swedish alternatives to choose from for each English word.

COLLEX 3 consisted of collocation pairs, of which one was a targeted real collocation, and one was a distractor (pseudo-collocation). The combinations were verb + NP but also some adjective + noun structures. COLLEX 3 was administered in two versions: one monolingual and one bilingual. In the bilingual version, Swedish translations of the targeted collocation in each item were supplied. This was done as a between-group experimental manipulation, with the monolingual COLLEX as a control, and the bilingual COLLEX as the experimental condition.

COLLMATCH 1 consisted of a total of 8 grids with 9 words in each grid. A total of 6 such grids featured verbs + NP, and two featured adjectives + nouns. In total, 144 word combinations were presented in the test, out of which 51 were intended as target collocations, and consequently 93 were distractors (pseudo-collocations).

4.1.3.4 Informants

The test battery was administered to a total of 119 informants, all of whom took the tests voluntarily. With only three exceptions, these were undergraduate students of English at Lund University, pursuing studies at different levels. Three informants were graduate students at the University of Wales, Swansea. The mean age of the informants was 24.1. Table 4.2 below gives an overview of the informants:

Table 4.2 Distribution of informants across study levels in test experiment 3.

Type of informant	Number of informants Total number	Number of L1 Swedes	Number of non-L1 Swedes
SWEuni1: First term students	46	39	7
SWEuni2: Second term students	42	37	5
SWEuni3: Third term students	22	21	1
SWEuni4: Fourth term students	6	3	3
ENGuniNS: Native speakers of English	3		
Total	119		

4.1.3.5 Research questions

In order to yield opportunities for further developments of COLLEX and COLLMATCH, the following main research questions were addressed:

1. Are COLLEX 3 and COLLMATCH 1 reliable tests in terms of internal consistency?
2. Do the test items in COLLEX 3 and COLLMATCH 1 have a satisfactory discriminatory power in terms of item facility and item-total correlations values?
3. Are there differences between different Swedish learner levels, and between different Swedish learner levels and native speakers, in terms of scores on COLLEX 3 and COLLMATCH 1?
4. In COLLEX 3, is there a difference between scores from informants who took the monolingual version and informants who took the bilingual version?
5. Is guessing frequent in COLLEX 3 and what effect does it have on learners' scores?

4.1.3.6 Test administration and scoring

The test administration was advertised prior to the test dates in a number of intact student groups, and volunteers were asked to stay on after class to take the test battery. A test battery consisting of 3 parts was administered to students of English at the Department of English, Lund University, in early February 2005. A great majority of the informants finished the test battery in 15-25 minutes. The three native speakers of English were sent the questionnaire via e-mail and they sent in their answers in the same manner.

In terms of scoring, all correct answers in the three tests were awarded with 1 point. Conversely, all incorrect answers received a score of 0 points.

4.1.4 Results

4.1.4.1 Reported results

In section 4.1.4.2, the overall results on the three tests, including data from all the 119 informants, will be given. Since the main interest here is how Swedish learners perform on COLLEX and COLLMATCH, in relation to native speakers of English, an analysis was carried out in which all informants with other L1s than Swedish or English were excluded. These results are reported in subsection 4.1.4.3. This leaves us with 103 informants.

4.1.4.2 Overall results

Descriptive statistics for the three tests were calculated. In Table 4.3 below, the score distributions on the respective tests are presented, and Figures 4.6, 4.7 and 4.8 show the frequency distributions.

Table 4.3 Score distributions and test characteristics of SINGLEX 3, COLLEX 3 and COLLMATCH 1 (N = 119)

Value	SINGLEX 3 N = 110*	COLLEX 3 N = 119	COLLMATCH 1 N = 119
k	40	50	144
MPS**	40	50	144
Mean	36.9	42.6	121.0
S.d.	2.8	5.5	8.3
Range	15	23	43
Minimum	25	27	97
Maximum	40	50	140
Skewness	-1.7	-.98	-.50
Kurtosis	3.8	.36	.24
Cronbach's α	.72	.84	.80
Guesses (f)	n.a.	606	n.a.
Correct guesses (f)	n.a.	383	n.a.

* = A total of 9 informants did not have sufficient knowledge of Swedish to take SINGLEX 3. As a result of this, only the scores of 110 informants are reported here.
** = Maximum Possible Score
n.a. = not applicable

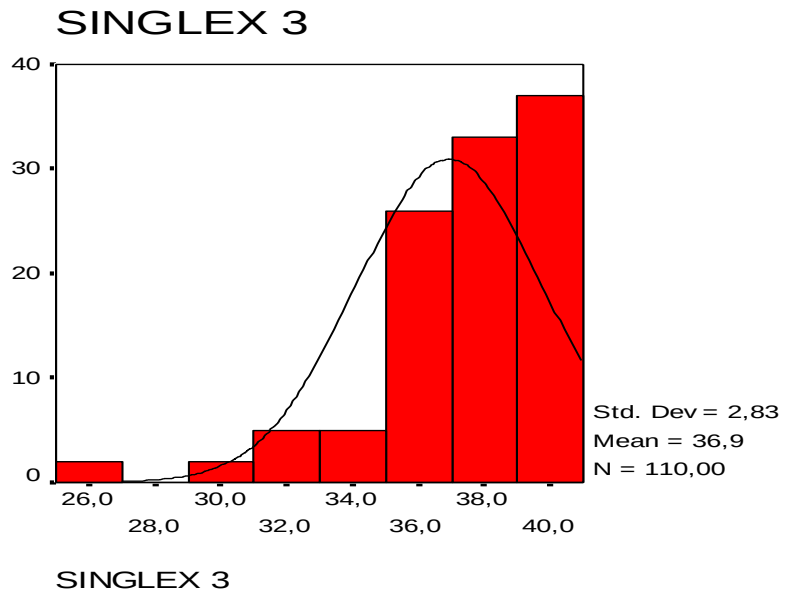

Figure 4.6 Frequency distribution of scores on SINGLEX 3.

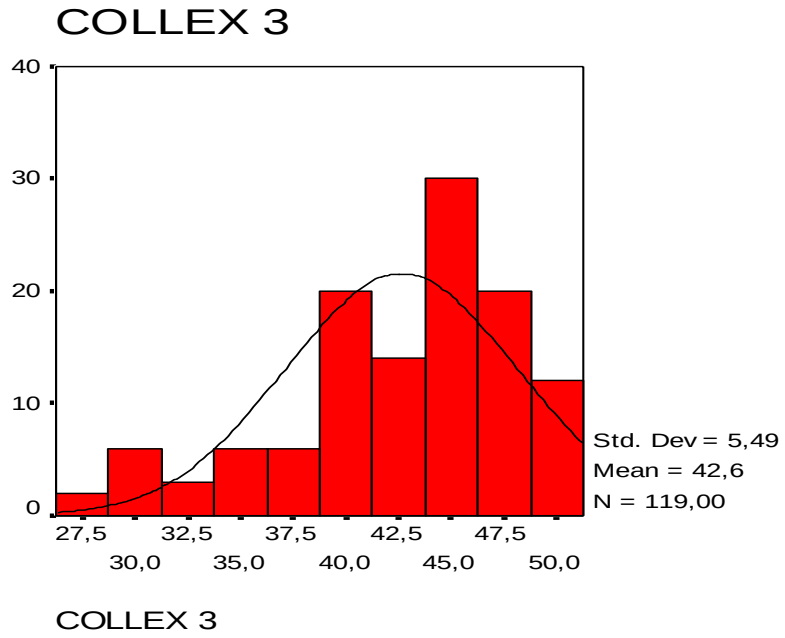

Figure 4.7 Frequency distribution of scores on COLLEX 3.

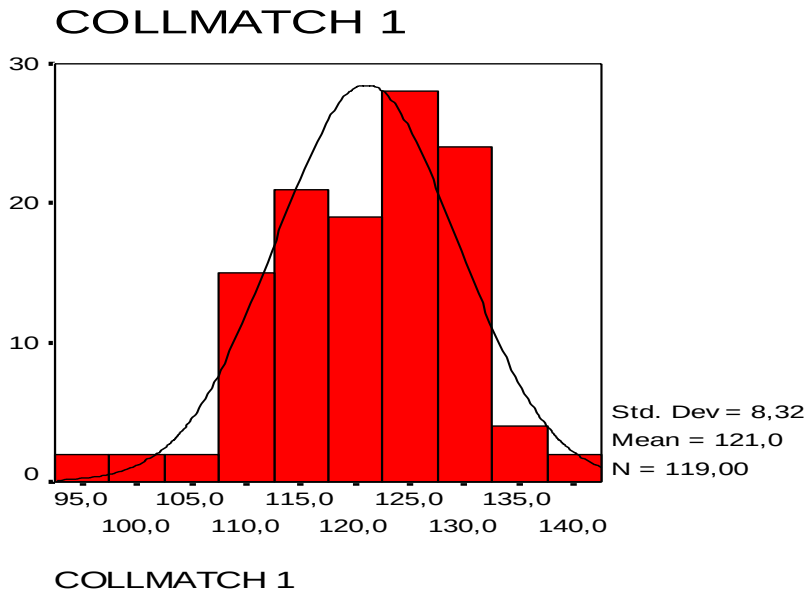

Figure 4.8 Frequency distribution of scores on COLLMATCH 1.

The SINGLEX 3 test displayed a non-normal distribution, whereas scores on both COLLEX 3 and COLLMATCH 1 stayed within the realms of normality. In terms of mean scores, the mean score on SINGLEX 3 was high at 36.9, which indicates that the tested population in general had few problems in terms of knowledge of the single words occurring in COLLEX 3 and COLLMATCH 1. The mean score on COLLEX 3 was relatively high at 42.6, indicating that a ceiling effect is present. The mean score on COLLMATCH 1 was observed at 121.0. All three tests displayed acceptable reliability coefficients, with .72 for SINGLEX 3, .84 for COLLEX 3, and .80 for COLLMATCH 1. Considering the many items in COLLMATCH 1, a test length of 144 items, a higher reliability value than .80 was expected. The fact that COLLEX 3, with about one third of the number of items in COLLMATCH 1 (50 vs. 144), produced a higher reliability coefficient goes to show that a longer test is not necessarily more reliable than a shorter test. An item analysis will shed light on the discriminability of the individual test items.

4.1.4.3 Cross-sectional data: Swedish learner groups and native speakers

4.1.4.3.1 Reported results

In the following subsections, the results from the test administration will based on cross-sectional comparisons. The following groups were used for this purpose (compare Table 4.2 above):

Table 4.4 Informant groups used in the cross-sectional analysis of the test data.

Informant group	Number
SWEuni1: Swedish first-term students of English	39
SWEuni2: Swedish second-term students of English	37
SWEuni3: Swedish third-term students of English	21
ENGuniNS: Native speakers of English	6
Total	103

These groups were formed by only using Swedish-speaking informants from the first, second and third term of study, and by discarding the 3 fourth-term students from any comparative statistical analyses due to the insufficient number of informants. A small group of native speakers of English was formed by taking 3 subjects from the first, second and third term of study, and pairing them up with the 3 native speakers from which data was gathered specifically. Furthermore, during the test administration of COLLEX 3, roughly half of the students were given a monolingual version of the test and half were given a bilingual version (see subsection 4.1.3.1). For this reason, each group (SWEuni1, SWEuni2, and SWEuni3) was divided into two sub-groups (M = monolingual, and B = bilingual).

When analysing test item quality, only data from informants from the Swedish learner groups were used. The rationale for this is that I am primarily developing a test for advanced Swedish learners of English, and it is important that we analyse test item quality based on the performance of these learners. It stands to reason that not least the observed item facility of the test items would be slightly inflated if also native speaker scores would be taken into account.

4.1.4.3.2 SINGLEX 3

Table 4.5 below presents the results for the test of single word knowledge: SINGLEX 3. The results show that the informants' knowledge of the single words was on a high level, as indicated by the high means of the respective groups (35.5, 38.2, and 38.0, respectively).

Table 4.5 Results on SINGLEX 3 (k = 40) by cross-sectional groups.

Group	N	M	S.d.
SWEuni1: Swedish first-term students of English	39	35.5	3.2
SWEuni2: Swedish second-term students of English	37	38.2	1.5
SWEuni3: Swedish third-term students of English	21	38.0	1.4

The low standard deviation scores tell us that the informants scored uniformly high, and that they display a high level of homogeneity in terms of their performance on SINGLEX. All but one informant (96 out of 97) scored a minimum of 32 points, which corresponds to 90% of the total score of SINGLEX 3. It may therefore be concluded that the learners did not have a problem with the single words.

No statistical comparison of means was made due to the fact that the SINGLEX test served as a control for single word knowledge.

4.1.4.3.3 COLLEX 3

When it comes to the informants' results on COLLEX 3, a number of analyses were carried out. Firstly, descriptive statistics for the four main groups were computed. As a second step, analyses aimed at establishing any effect of learner group affiliation on test scores were carried out. Thirdly, any experimental effect of the COLLEX test type (monolingual or bilingual) was tested for. Fourthly, an analysis of the self-reported guessing behaviour was conducted. The descriptive statistics are reported in Table 4.6 below. The notations used for the groups are explained under the table.

Table 4.6 Results on COLLEX 3 by cross-sectional groups.

Group	N	M	S.d.	Reliability	Mean number of guesses	Mean number of correct guesses
SWEuni1M	18	40.6	5.8	.82	5.6	3.8
SWEuni1B	21	40.2	6.0	.83	9.4[1]	5.7
SWEuni1 M+B	39	40.4	5.8	.82	7.6	4.8
SWEuni2 M	20	43.5	4.6	.79	5.6	2.9
SWEuni2 B	17	43.4	4.3	.75	5.8	3.6
SWEuni2 M+B	37	43.5	4.4	.76	5.7	3.3
SWEuni3 M	11	45.1	3.2	.67	4.6	3.6
SWEuni3 B	10	46.5	3.0	.68	4.2	3.1
SWEuni3 M+B	21	45.8	3.1	.67	4.4	3.3
ENGuniNS M	6	48.5	1.6	.50	.67	.50
ALL	103	43.1	5.2	.83	5.9	3.7

SWEuni1M = Swedish first-term students of English who took the monolingual COLLEX 3 test
SWEuni1 B = Swedish first-term students of English who took the bilingual COLLEX 3 test
SWEuni2 M = Swedish second-term students of English who took the monolingual COLLEX 3 test
SWEuni2 B = Swedish second-term students of English who took the bilingual COLLEX 3 test
SWEuni3 M = Swedish third-term students of English who took the monolingual COLLEX 3 test
SWEuni3 B = Swedish third-term students of English who took the bilingual COLLEX 3 test
ENGuniNS M = Native speakers of English who took the monolingual COLLEX 3 test

Notes: [1] One informant reported guesses on all 50 items (27 correct, and 23 incorrect guesses).
Maximum score = 50.

As can be seen in the table, the small group of six native speakers scored a mean of 48.5, in turn followed by the third term students (45.8), the second term students (43.5), and the first term students (40.4). We may also note that the higher mean score, the lower the standard deviation, which normally indicates that the better performing groups are more homogeneous in their performance. We should however note here that the groups differ quite considerably in size, and bear this in mind when we make comparisons. The mean score of the native speakers corresponds to 97%, which lends validation support to the test. The overall reliability of the test scores, as measured by Cronbach's alpha, was .83, which is satisfactory. However, as can be seen in the table, the reliability coefficients for the various groups were lower. It is likely that this effect is caused by group homogeneity. The low reliability for the native speaker group is above all believed to be due to the fact that as many as 43 out of 50 items had zero variance.

As to the potential presence of group effects on test scores, two factors violated the assumptions of a regular ANOVA: the unbalanced design (different group sizes), and unequal variance between the groups, as tested through Levene's test. For these reasons, appropriate alternative tests were used. A Welch test signalled a highly significant effect of learner group affiliation on scores on the test, Welch F (3, 31.7) = 18.24, p < .001. A post-hoc Games-Howell test showed that there was a significant difference between the group of first-term learners on the one hand, and third-term learners and native speakers on the other hand.

There were no significant differences observed between second term and third term learners. The difference between first-term and second-term learners was not significant, but it was very close to being so (p = .056), and is therefore interesting. The native speakers' scores were significantly different from all three Swedish learner groups. The significant differences are summarized in Table 4.7 below, where statistical significance is indicated through asterisks, and non-significance through the abbreviation n.s.

Table 4.7 Significant differences between group means on COLLEX 3.

	SWEuni1: Swedish first term learners of English	SWEuni2: Swedish second term learners of English	SWEuni3: Swedish third term learners of English	ENGuniNS: Native speakers of English
SWEuni1: Swedish first term learners of English		n.s.	*	*
SWEuni2: Swedish second term learners of English	n.s.		n.s.	*
SWEuni3: Swedish third term learners of English	*	n.s.		*
ENGuniNS: Native speakers of English	*	*	*	

* The mean difference is significant at p < .05

In order to test the hypothesis that the insertion of a Swedish translation in COLLEX 3 would affect test scores, the means of the informants taking the monolingual version of COLLEX were compared with the means produced by the informants taking the bilingual version. Under the null hypothesis, we assume that there is no difference between the means. An independent t-test was performed on the data which were normally distributed. In terms of group means, there was a minuscule difference between informants (N = 49) taking the monolingual version (M = 42.80, SE = .72) and informants (N = 48) taking the bilingual version (M = 42.67, SE = .78), but this difference was not significant t (95) = .122, p > .05. Thus, it was not possible to reject the null hypothesis, and consequently no effect of COLLEX 3 test version on test scores was observable among the tested population. In order to verify this result also between subgroups, pair-wise comparisons were carried out. The results from independent t-tests are shown in Table 4.8 below. No statistically significant differences existed. The analysis included all the Swedish learners in the study (total N = 97).

The guessing behaviour observed in COLLEX 3 amounted to a total of 606 self-reported guesses, out of which 383 were marked as correct guesses. This meant that more correct indicated guesses (.63) occurred than incorrect ones (.37). On average, the informants in the present study reported 5.9 guesses. The mean number of correctly made guesses was 3.7. The highest mean in terms of guessing was observed for the first term students (7.6), followed by second term students (5.7), and third term students (4.4).

Table 4.8 Pair-wise comparisons of subgroup means (monolingual and bilingual COLLEX 3 scores).

Group	Mean	S.d.	t-value	Sig.
SWEuni 1: Swedish first-term students of English				
Monolingual COLLEX 3 (N = 18)	40.6	5.8	0.20	p .845
Bilingual COLLEX 3 (N = 21)	40.2	6.0		
SWEuni2: Swedish second-term students of English				
Monolingual COLLEX 3 (N = 20)	43.5	4.6	0.60	p .953
Bilingual COLLEX 3 (N = 17)	43.4	4.3		
SWEuni3: Swedish third-term students of English				
Monolingual COLLEX 3 (N = 11)	45.1	3.2	-1.04	p .312
Bilingual COLLEX 3 (N = 10)	46.5	3.0		

The native speakers reported only 0.67 guesses per individual on average for the 50 items in COLLEX 3. It should be pointed out that one informant in the first-term student group reported guessing on all 50 items. An ANOVA showed no significant effects of group affiliation on levels of guessing, F (3, 99) = 2.69, p > .05.

A further analysis of the guessing behaviour was made in which the scores for all informants on COLLEX 3 were correlated with the number of guesses reported. A Kendall's tau (τ) test showed that there was a significant negative relationship between the number of guesses indicated and the number of points on the test, τ = -.33, p = < .01. This means that as scores on COLLEX 3 increase, the number of reported guesses decreases. Or, put another way, students with high scores on the test indicated fewer guesses than students with lower scores. A correlation analysis which took group affiliation into account resulted in the following coefficients:

Table 4.9 Correlations between scores on COLLEX 3 and reported guessing frequency by groups.

Group	N	Correlation (τ) between COLLEX 3 scores and guessing
SWEuni1: Swedish first-term students of English	39	-.22*
SWEuni2: Swedish second-term students of English	37	-.45**
SWEuni3: Swedish third-term students of English	21	-.10
ENGuniNS: Native speakers of English	6	.00

** Correlation is significant at the .01 level.
* Correlation is significant at the .05 level.

The correlation values tell us that no significant correlations were observable for neither the native speaker group, nor the third term student group. For the second and first term student groups, however, the negative correlations were significant.

The results from the item analysis of COLLEX 3 can be seen in Appendix 4D. This analysis was based on the 97 Swedish learners affiliated with groups SWEuni1 – SWEuni3. Item Facility (IF) values together with Item-total correlation (ITC) values were computed for each of the 50 items of the test. In terms of item facility values, COLLEX 3 displays a mean of .85. Only one item display zero variance. Compared to the administration of COLLEX 2 (.80), this mean is higher, which is not surprising, considering the fact that the informants'

general proficiency is assumed to be higher. The item-total correlation values, on the other hand, are higher in COLLEX 3, with a mean of .27 (.23 in COLLEX 2). Using Ebel's (1979) guidelines for item quality, we arrive at the distribution presented in Table 4.10 below.

Table 4.10 Distribution of test items from COLLEX 3 into categories of discrimination power following Ebel (1979).

Item-total correlation guidelines	.40 and higher very good items	.30 to .39 reasonably good items possibly subject to improvement	.20 to .29 marginal items in need of improvement	.19 and below poor items which need to be revised or eliminated
Items from COLLEX 3	2, 9, 19, 26, 27, 30, 35, 38	1, 4, 6, 7, 10, 13, 17, 18, 28, 29, 32, 41, 43, 45	3, 11, 16, 20, 21, 24, 25, 31, 39, 40, 42, 46, 48, 50	5, 8, 12, 14, 15, 23, 33, 34, 36, 37, 44, 47, 49 Value 0.00 (no variance): 22
Number of items	8	14	14	14 (13+1)
per cent of total number of items	16%	28%	28%	28%

Just as we observed a clear improvement in COLLEX 2 compared to COLLEX 1, we here observe a clear overall improvement in COLLEX 3 compared to COLLEX 2. Compared to COLLEX 2, although the 'very good items' in COLLEX 3 actually decreased by two percentage points, the 'reasonably good items' increased by 8 percentage points, and the 'marginal items' increased by 10 percentage points. Also, the 'poor items' decreased by 14 percentage units. Thus, even though the item facility mean is higher in the present study, the mean discriminatory power of the items in COLLEX 3, as measured through item-total correlation coefficients, is improved.

4.1.4.3.4 COLLMATCH 1

This section describes the results on COLLMATCH 1. A number of analyses were carried out. Descriptive statistics for the four main groups were computed, and, as with the COLLEX format, analyses aimed at establishing any effect of learner group affiliation on test scores were carried out. In addition, an item analysis was carried out in order to investigate the quality of the individual items of the test.

The descriptive results on COLLMATCH 1 are shown in Table 4.11.

Table 4.11 Results on COLLMATCH 1 (k = 144) by cross-sectional groups.

Group	N	M	S.d.	Reliability[1]
SWEuni1: Swedish first-term students of English	39	116.3	8.8	.79
SWEuni2: Swedish second-term students of English	37	122.5	5.8	.62
SWEuni3: Swedish third-term students of English	21	125.0	5.0	.56
ENGuniNS: Native speakers of English	6	133.8	5.4	.76
ALL	103	121.3	8.3	.80

[1] Cronbach's alpha

As can be seen in Table 4.11, the results of the COLLMATCH 1 test mirrored those of COLLEX 3 in that the highest mean score was obtained by the native speaker group (133.8), followed in turn by the Swedish third-term students (125.0), the Swedish second-term students (122.5), and the Swedish first-term students (116.3). The mean score of the native speakers corresponds to 93%, which lends validation support to the test.

The overall reliability of the test scores was satisfactory at .80. However, as with the COLLEX data, a low reliability coefficient was observed for the group of third-term students (.56).

In order to compare the seemingly different means from the four groups, a Levene's test was run to check the variance of these group scores. Since the result was significant, i.e. that the variance were significantly different, paired with the fact that unequal sample sizes were used, assumptions of a regular ANOVA were violated. Consequently, a Welch test was used instead. This test revealed a highly significant effect of learner group affiliation on scores on the test, Welch F (3, 22.7) = 15.94, $p < .001$. A post-hoc Games-Howell test showed that there was a significant statistical difference between the means of all the groups, except for the difference between second term and third term students. The significant differences are summarized in Table 4.12 below, where statistical significance is indicated through asterisks, and non-significance through the abbreviation n.s.:

Table 4.12 Significant differences between group means on COLLMATCH 1.

	SWEuni1: Swedish first term learners of English	SWEuni2: Swedish second term learners of English	SWEuni3: Swedish third term learners of English	ENGuniNS: Native speakers of English
SWEuni1: Swedish first term learners of English		*	*	*
SWEuni2: Swedish second term learners of English	*		n.s.	*
SWEuni3: Swedish third term learners of English	*	n.s.		*
ENGuniNS: Native speakers of English	*	*	*	

*The mean difference is significant at p < .05

These results are very similar to those on COLLEX 3. Again the native speaker group performed significantly better than all the three Swedish informant groups. No difference could be established between the mean scores of the second- and the third-term students.

An item analysis was carried out with the purpose of shedding light on the item quality of COLLMATCH 1. A table showing the item facility and item-total correlations of all the items is presented in Appendix 4E. The mean value for item facility was .84, whereas the item-total correlation mean was .14. Just as with COLLEX 3, the informants of the study thus scored very high results, and consequently, a large number of items in COLLMATCH 1 were too easy for the tested population, at least from a norm-referenced test perspective.

Table 4.13 Distribution of test items from COLLMATCH 1 into categories of discrimination power following Ebel (1979).

Item-total correlation guidelines	.40 and higher very good items	.30 to .39 reasonably good items possibly subject to improvement	.20 to .29 marginal items in need of improvement	.19 and below poor items which need to be revised or eliminated
Items from COLLMATCH 1	4, 16, 19, 38, 40, 45, 46, 50, 75, 76, 77, 82, 88, 136	2, 18, 21, 31, 37, 48, 49, 58, 60, 64, 72, 83, 95, 102, 108, 111, 120	8, 20, 29, 32, 39, 59, 61, 62, 63, 81, 97, 98, 99, 100, 103, 104, 105, 107 113, 127	3, 5, 6, 7, 9, 10,11, 12, 14, 15,17, 22, 24, 25, 26, 27, 28, 30, 33, 34, 35, 36, 43, 52, 54,55, 57, 65, 66, 67, 68, 70, 71, 73, 74,78, 79, 80, 84, 89,90, 91, 92, 93, 94,101, 109, 110, 112, 114 116, 117, 118, 119, 121,122, 123, 124, 125, 126, 128,129, 130, 131,132, 133, 135, 137, 138,139, 140, 141, 142,143, 144 Value 0.00 (no variance): 1, 13, 23, 41, 42, 44, 47, 51, 53, 56, 69, 85, 86, 87, 96, 106, 115, 134
Number of items	14	17	20	93 (75 + 18)
per cent of total number of items	10%	12%	14%	64%

The item-total correlation mean of .14 is rather low, in comparison with that of COLLEX 3, which was almost twice as high at .27. The overall poor quality of the items in COLLMATCH 1 is illustrated by the customary division shown in Table 4.13. As many as 93 items, or 64% of the total number of items, fall into the "poor item" category, which is a clearly disappointing result.

4.1.5 Discussion

In this section, I will discuss the results from the test administration of SINGLEX 3, COLLEX 3, and COLLMATCH 1. I will structure the discussion around the research questions that were presented in section 4.1.3.5. For sake of clarity, the specific research questions will be repeated here as the starting point of each discussion section. Questions 1-3 have bearings on both COLLEX 3 and COLLMATCH 1, whereas questions 4 and 5 relate only to COLLEX 3.

4.1.5.1 Are COLLEX 3 and COLLMATCH 1 reliable tests in terms of internal consistency?

The question of whether COLLEX 3 and COLLMATCH 1 are reliable tests in terms of internal consistency is of course clearly linked to the overall question of test reliability. As has been pointed out earlier in this thesis, it is widely agreed that reliability is a necessary but not sufficient condition for validity. Therefore, before we start comparing scores from groups of different language ability, we need to present clear evidence of the reliability of the measures we use. We should also recall that some researchers, *inter alia* Weir (2005), and Alderson (1991), see reliability as a type of validity evidence. From that perspective, a test's reliability is a valuable part of its overall validity. A measure of internal consistency is in a sense a measure of the homogeneity of the test items. As such, it has bearings on test content as well as test construct. A reliable test is a test whose scores consistently reflect the construct it is measuring. In a discussion about the use of correlational analyses for construct validation, Messick links item homogeneity, or internal-consistency reliability, with construct validity. This is so, he argues "because the degree of homogeneity in the test, as we have seen, should be commensurate with the degree of homogeneity theoretically expected for the construct in question" (1989:51).

The results presented in Tables 4.3, 4.6, and 4.11 give support to the claim that COLLEX 3 and COLLMATCH 1 are reliable tests as estimated through internal consistency (Cronbach's alpha). The reliability of COLLEX 3 was observed at .84 and .83, respectively, for the different group constellations used in the study. The values for COLLMATCH were .80 and .80, respectively, for the same constellations. We may therefore conclude, following Weir's view presented above, that we have now added one key aspect to the overall validity of the two tests.

One key issue with regard to test reliability is test length. For a test of 50 items like COLLEX 3, an overall reliability of .83 is acceptable, even though even higher values are naturally desirable from a general testing perspective. However, in the case of COLLMATCH 1, a test consisting of 144 items, a reliability of "only" .80 raises some questions. With a long test like that, a much higher reliability is in theory possible and perhaps even expected. It is difficult to say exactly what caused the absence of even higher values, but judging from the results presented in Table 4.13, it is highly likely that poor item quality is at play. This question will be addressed in more detail in 4.1.5.2 below.

4.1.5.2 Do the test items in COLLEX 3 and COLLMATCH 1 have a satisfactory discriminatory power in terms of item facility and item-total correlations values?

The question of test reliability discussed above is closely linked to item characteristics. In this regard, the aspects of item facility (or item difficulty) and item discrimination are essential pieces of information. If a test is too easy or too difficult for a specific population, then low reliability is often the result. We have already observed that the overall reliability of COLLEX 3 and COLLMATCH 1 is on an acceptable level. However, the analysis of the individual items puts us in a better position to elucidate what might be the cause of that level of reliability, and more importantly, how potential sources of measurement error could be eradicated.

Starting with COLLEX 3, we observed a mean item facility of .85 for the 50 items. This is a rather high mean, and it means that, on average, 85 per cent of the tested population (97 undergraduate Swedish learners of English) answered the items correctly. Compared to the earlier administrations of the COLLEX format, only one item displayed zero variance: item 22, *make a crime – commit a crime*. However, a large number of items were answered correctly by almost all of the informants tested (see Appendix 4D). In fact, more than half (27/50) of the 50 items in the test displayed an item facility of .90 or more.

The high level of item facility can be approached from two perspectives. From one perspective, it is clear that the learners from the tested population have a high degree of receptive knowledge of the collocations featured in the test. This can be seen as a positive result. Although we must be cautious about drawing too far-reaching conclusions based on the material, a tendency is discernable. Swedish university level learners of English seem to have a good receptive recognition knowledge of English verb + NP, and adjective + noun, collocations. From another perspective, this high level of performance is problematic. The perspective is that of norm-referenced test construction. Especially if COLLEX is to be used as a model-building tool, then the high means on COLLEX are not ideal. The reason is that too many learners produce close to maximum scores. Therefore, there is no room for improvement at the higher end of the scale. Consequently, we cannot make a finer ranking of these individuals in terms of their knowledge. My main aim has been to develop proficiency tests for upper-secondary school and university-level learners that can be used for diagnostic and placement purposes. However, if it was possible to come up with a reliable and valid model-building tool, then this would of course be an advantageous synergy effect. A potential remedy for the high means that could be used is the introduction of some sort of correction for guessing-formula. Such formula would deduct points from the overall score based on the number of wrong answers. A factor would be introduced that would take into account the number of choices in each item. However, before I introduce any correction formulae, the test should be administered to upper-secondary school students, in order to shed light on the performance of learners on a slightly lower level of general proficiency. Introducing a correction for guessing-formula is thus considered premature at this point.

The mean discriminatory power of the items in COLLEX 3, as estimated through item-total correlations, was observed at .27. Compared to its predecessor, COLLEX 2, this means that the discriminatory power is slightly higher in the present version (COLLEX 2: .23). This is an improvement. There are two likely reasons behind this improvement. Firstly, an attempt was made to include the best-performing items from COLLEX 2 in the present version. This means that poorly discriminating items were not included in COLLEX 3. Secondly, since the

tested sample of informants in COLLEX 3 arguably represent a wider range of general ability in English, this could have created a higher mean of item discrimination.

When it comes to COLLMATCH 1, the results in terms of item facility and item discrimination were discouraging. As a matter of fact, the item facility lies close to that of COLLEX 3, with a mean value of .84. In this sense, COLLMATCH 1 was difficult to the same extent as COLLEX 3. As to item discrimination, however, a much lower mean value was observed for COLLMATCH 1: .14. A quick glance at Table 4.13 tells us that a majority of the items in COLLMATCH 1 (93/144) did not function well in this regard. First of all, 18 items showed no discrimination at all, due to zero variance. Secondly, as many as 19 items displayed a negative item-total correlation. Clearly, these values are highly problematic.

As to the reasons behind the result, I would like to argue that the test format itself had a disadvantageous effect on the item material. My aim was to include words (verbs and adjectives) that shared some semantic component. For example, the items in block 3 were the verbs *say*, *tell*, and *speak*. Based on these words, I then selected six object NPs that could either collocate with these verbs or not. The phrases in block 3 were *a prayer, a language, a joke, farewell, a story*, and *lies*. For this total of 9 items, 18 combinations were possible. In practise, 6 of these combinations were intended to be target collocations, and consequently 12 were intended to function as pseudo-collocations. Thus, only one third of the combinations in that block were collocations, whereas two thirds were combinations that the learners were expected to reject. For the whole test, only 51 combinations were target combinations, and as many as 93 were pseudo-collocations. As a consequence of this, then, in hindsight, COLLMATCH 1 is more a test that taps into learners' ability to reject pseudo-collocations, than it is a test that taps into their ability to recognise real collocations. For this reason, changes to the format per se are deemed necessary. Even though COLLMATCH 1 displayed an acceptable overall reliability level, the prospect when it comes to improving the poor item discrimination level is bleak, due to the inherent restrictions that the format brings with it.

4.1.5.3 Are there differences between students from different Swedish university learner levels, and between different Swedish university learner levels and native speakers, in terms of scores on COLLEX 3 and COLLMATCH 1?

This question was based on aspects of validation. In a well-functioning test, we expect learners with different abilities in the construct measured to score differently from each other. More specifically, we expect learners with a good ability in the measured construct to produce higher scores than learners with a low ability. In the present study, data were collected from groups of Swedish students who were studying English at different levels at university. The levels are built on progression and there are several proficiency exams given at the end of each level. For example, there are tests of vocabulary (single word knowledge), grammar and translation, and oral fluency. Thus, to be allowed to continue to a higher level in the higher education system, a progressively higher level of proficiency is needed. The student groups in the study were thus assumed to be on different levels of general proficiency. However, there was no proof available saying that this meant that for example second term students were more advanced in terms of collocation knowledge than first term students. This leaves us with the question of whether receptive collocation knowledge of learners in general follows their general proficiency in a language. There is no unequivocal answer to this question, but Bonk

(2001), whose study was reviewed in Chapter 2, presents data relevant to the issue at hand. Let us recapture here some of the results arrived at in Bonk's study.

Bonk administered a 49-item general English proficiency test together with a 50-item collocation test to 98 learners of English. The proficiency test, which displayed a reliability of .85 (Kuder-Richardson 20), was in fact a shortened TOEFL test. Bonk found that the collocation test and the general English proficiency test correlated at .73 (r^2 = .53) after correction for attenuation. However, he entered a caveat about solely looking at proficiency as a predictor of collocational proficiency because of individual variation.

In another study, looking at the use of collocations, Gitsaki (1999) argues that it is possible to claim parallel development of collocation knowledge and language proficiency. However, no independent measure was used to establish proficiency. Instead, a number of measures, like lexical density, target-like use of articles, and words per T-unit in learners' 200-word essays were used as indicators, the same material in which collocation use was investigated. It is questionable if these findings are reliable when the variables were confounded like that.

Counterevidence can be found in a study by Howarth (1996), who manually investigated the use of verb-noun combinations in a corpus based on 10 essays. Howarth found no correlation (r = .15) between the general proficiency of a learner and the number and acceptability of the collocations used.

Thus, even though we are far from having a very clear and unified view of the relation between collocation knowledge and general proficiency, there is some evidence to suggest that there exists some sort of relation. In consequence, we would expect to find for example higher scores from third term learners than from second term learners on the collocation tests administered. Indeed, this is also what we find, even though the differences are relatively small, and they are not statistically significant throughout.

In COLLEX 3, we observed statistically significant differences between first term learners and third term learners, but not between either of these two groups and the group of second term learners, even if a difference between first-term learners and second-term learners was very close to being significant. This means that in comparison to the mean scores from learner groups only one term apart, no differences are observable. There could be two explanations for this. Either, the type of collocational knowledge tested in COLLEX 3, receptive recognition knowledge of verb +NP and adjective + noun combinations, does not develop to the extent that a difference is measurable. Or, it could be the case that COLLEX 3 as a test tool is not sensitive enough to pick up any existing differences. As to the first explanation, we should note that a university term at Swedish universities is 4.5 months long, and perhaps it is not realistic to expect a measurable growth of a learner's receptive inventory of collocations in this relatively short period of time, despite full-time studies (~40 hours/per week) containing a high degree of exposure to written texts, both fiction literature and technical texts. Unfortunately, we have no benchmark figures of collocation knowledge from other studies to draw on in this respect. An interesting comparison, however, can be made in terms of single word vocabulary size. Gyllstad (2004) analysed the results from a frequency-based, 120-item test of English single word knowledge used at many Swedish universities. Based on sampled groups of 30, he found that no difference was observable between first term and second term students (means: 69.5 and 69.2, respectively). Interestingly, this study found that the mean for third term students was considerably higher, with a mean of 80.8. Thus, it seems that there is a great deal of overlap between these learner groups, and that knowledge types

like receptive single word vocabulary, and receptive command of collocations does not seem to develop to a great extent over the period of just one term.

On the other hand, explanation number two proposed above is equally likely. COLLEX 3 consists of 50 items, and we have seen that the Swedish university informants in the present study produce high scores in general (> 80%). Thus, there is a tangible ceiling effect present and not much room for improvement. It could be the case that any existing difference cannot be picked up reliably by the test tool.

If we turn to the results for COLLMATCH 1, a slightly different picture emerges, however. In this test, significant differences were observed between first-term and second-term students, but not between second term and third term students. The conclusion I draw from this is that there are differences between the groups, but that the test tools used in this study are not always capable of picking up these differences, let alone statistically significant differences. This is a problem I will have to address if COLLEX and COLLMATCH are to be developed further.

A positive result from the test administration reported here is that native speakers of English scored highly on both COLLEX 3 and COLLMATCH 1. Admittedly, the group of native speakers was very small, and a fair amount of caution is needed when interpreting the results. The 6 native speakers scored 97 per cent of the maximum score on COLLEX 3, and 93 per cent of the maximum score on COLLMATCH 1, and their mean scores were statistically different from all the Swedish learner groups. This lends support to the validity of the tests, since the performance of native speakers can be seen to function as baseline data. Also, if we are not one hundred per cent sure of the true general proficiency of the Swedish learners in the study, it is a safe assumption that the native speakers have a higher ability than the Swedish learners in the construct measured.

4.1.5.4 In COLLEX 3, is there a difference between scores from informants who took the monolingual version and informants who took the bilingual version?

As was reported in the results section, no statistically significant differences were observed between the mean scores from learners who sat the monolingual version of COLLEX 3, and those learners who sat the bilingual version. Notably, minuscule differences occurred between the subgroups at each learner level, as evidenced by the means in Table 4.6 above. The biggest difference in this respect was observed among the third-term learners, where a mean of 46.5 was produced in the bilingual condition, compared to a slightly lower 45.1 in the monolingual condition. All the same, pair-wise comparisons showed no statistic differences in the subgroups (see Table 4.8).

A conclusion that I draw from this is that the insertion of a Swedish prompt, aimed at telling the informants taking the test what concept is targeted in each item, does not lead to either higher or lower scores. Even though we did not formally test whether the learners who took the monolingual format knew what concepts were intended to be targeted in each item, they performed neither better, nor worse than learners who were supplied with the targeted concept. My own initial hypothesis was that learners who take the bilingual version might benefit from knowing what concept is targeted. However, Britt Erman (personal communication) drew my attention to the fact that getting a Swedish prompt might in theory affect learners negatively, since the presence of erroneous L1 transfer might be bigger.

It should be noted that the monolingual format has one major advantage. It can be used with learners of English of various L1 backgrounds. This also allows for important validation administrations with native speakers of English. Thus, it seems worthwhile to keep the monolingual format of COLLEX.

4.1.5.5 Is guessing frequent in COLLEX 3 and what effect does it have on learners' scores?

In the earlier administrations of the COLLEX format—the 60-item COLLEX 1 and the 65-item COLLEX 2—a mean guessing frequency of 4.7 and 10.1, respectively, was observed. In the present study, the mean number of guesses amounted to 5.9. In that respect, guessing frequency is not considerably different from those previous administrations. As was done in the analysis of the scores on COLLEX 2, the scores on COLLEX 3 were correlated with the number of reported guesses. A similar result was arrived at, in that a significant negative correlation was observed. Thus, learners who report few guesses score higher than those learners who report several guesses. This must be seen as evidence of the fact that the distractors in COLLEX 3 are doing a reasonable job.

In terms of the estimated general proficiency of the learners of the three studies, the learners taking COLLEX 2 were considered to be on a slightly lower level than those taking COLLEX versions 1 and 3. Incidentally, these learners also reported the highest number of guesses. The learners of the present study consisted of a mix of learner levels, and although the mean number of guesses was in fact different in the groups, no statistical difference was reached. A mean guessing frequency of 5.9 could be interpreted to imply that the average informant taking COLLEX 3 guessed on around 6 items of the total 50. In a binary format like that of COLLEX, where the probability of getting an item right through blind guessing is a high as .5, this must been seen as a relatively unproblematic level. However, the fact that scores could in reality become somewhat inflated through guessing, together with the tendency for a ceiling effect, begs the question if it would not be wise to consider a scoring formula that corrects for guessing.

There are no doubt both pros and cons associated with the introduction of such a formula. On the negative side, there is always a risk that individuals who claim to guess, but who are in fact relying on partial knowledge to some extent, will be penalized in a an unfair way. Also, a fact that speaks against the introduction of a correction for guessing formula is the observed high reliability for the COLLEX 3 administration. If guessing were more excessive, then unsystematic error variance would most likely result in lower reliability coefficients. The reliability levels have hitherto been acceptably high, though, at least for the COLLEX 2 and COLLEX 3 administrations.

On the positive side, the tendencies towards ceiling effects could be remedied to some extent, and we would probably get rid of some of the negative skewness that the score distributions for the COLLEX format have produced so far. However, it is not all that clear that this move would outweigh the negative aspects discussed above. Therefore, on reflection, I am still reluctant to introduce a correction for guessing formula for the COLLEX test.

4.1.6 Summary and conclusions

In this chapter I have reported on a third study involving test tools aimed at tapping receptive collocation knowledge with advanced learners of English. At the outset, I presented five research questions pertaining to different aspects of the test tools, as well as the learners taking part in the study. I concluded that COLLEX 3 produced reliable scores, as well as acceptable item-total correlation values. Furthermore, I argued that close to maximum scores from native speaker performance gave validation support to the test. An experimental set-up involving a between-groups design revealed that no differences could be observed between mean scores from informants taking a bilingual version of COLLEX and informants taking the original monolingual version. However, I also concluded that there was a ceiling effect present, and that COLLEX 3 might not be sensitive enough to pick up subtle differences between Swedish learners of different abilities.

In terms of the new format that was introduced, COLLMATCH 1, I pointed at overall reliable scores, but noted the poor quality and discriminatory power of the items. This poor quality was argued to stem at least partly from restrictions that were imposed through the format itself. In effect, the test format was more a test of learners' ability to reject pseudo-collocation than their ability to recognise real collocations. Therefore, a continued development based on the COLLMATCH grid format was not considered a viable option.

In the next chapter, I will report on a study in which a further developed COLLEX, together with a new COLLMATCH format, are administered to upper-secondary school and university-level learners of English. In addition, a vocabulary size measure will be introduced in order to correlate this variable with receptive collocation knowledge.

4.2 Developing a new COLLMATCH format, administering it together with COLLEX 4, and introducing a measure of vocabulary size

4.2.1 Introduction

In this section, I will report on a study in which an attempt is made to improve the COLLMATCH test by changing the format and the method of item selection. Furthermore, a new version of COLLEX will be administered. I will also introduce a measure of vocabulary size in the test battery, as a way to control for general proficiency, and also to relate the construct intended to be measured—receptive recognition knowledge of collocations—to the construct of vocabulary size. In addition to administering the test battery to a large group of university students, the performance of a sizable group of upper-secondary school students will also contribute to our evaluation and understanding of the tests and the test constructs.

4.2.2 Background

In the previous study (section 4.1), I evaluated the third version of the binary test format called COLLEX in an administration comprising a total of 119 informants. The test version showed moderate promise with reliable scores, and decent discriminatory power, even though a tendency for a ceiling effect was present. The high performance of a small group of native speakers lent validation support to the test. An experiment furthermore showed that the insertion of a Swedish prompt did not affect test scores, neither positively nor negatively, and I therefore decided to continue pursuing a monolingual test format, and to try to further improve the test. In the study, I also introduced and trialled a second test format called COLLMATCH. The evaluation of the test results yielded evidence of an acceptable overall reliability, and the test discriminated fairly well between students of English at different levels of study. However, a large number of the test items displayed poor quality. Especially the item-total correlation values for the items in the test evidenced low values. Another negative feature of the test was that it in effect measured informants' abilities to reject pseudo-collocations rather than the targeted "real" collocations. This was seen to stem from the grid format in which three 'prompt' words shared the same six potential collocates. In the light of these findings, I concluded that the format itself virtually imposed restrictions on item selection which made the prospect for pursuing the development of the format rather bleak.

In this study, I aim to address two relevant issues. Firstly, there is a need to administer my collocation tests to a group of learners who possess a slightly lower general proficiency in English than the university learners tested in the previous studies. The main reason for this is my intention to develop test tools aimed at both university-level learners, and upper-secondary school level learners, of English. If a test is to be used with a particular group of learners, it must be trialled on individuals who can be argued to belong to that population. It is not uncommon for tests to be inadvertently used for purposes they were not originally intended for. We must also remember, that in terms of norm-referenced reliability, the reliability estimates arrived at for a particular test administration are strictly speaking valid only for the scores of that particular administration. Thus, even though COLLEX 3 was found to reliably measure the receptive collocation knowledge of university learners, it does not

mean that it will function reliably with upper-secondary school students. Evidence for this must be presented.

The second issue I aim to address is in fact two-fold. One aspect relates to assumed levels of general proficiency. In the previous study I compared university learners studying English at different levels. A comparison was made between the means of the respective groups on the assumption that they differed in terms of general proficiency. I found that both COLLEX 3 and COLLMATCH 1 discriminated between Swedish students one year apart in terms of level of study, and between Swedish students and native speakers of English. My assumption, though, was not verified by any external criterion. It was based on study level affiliation. The flaw with this principle is that it is perfectly possible that a student on a lower study level has a higher general proficiency than a student belonging to a higher study level. Therefore, I need a measure that can serve as an indicator of general proficiency, since this would allow me to group students according to that variable, rather than study level affiliation. This would make any comparisons between groups more interesting from a theoretical perspective, because learners at different levels of proficiency could potentially be shown to differ in receptive recognition knowledge of English collocations. For this reason, I decided to administer a test of general proficiency as part of my test battery.

One option in this regard is to use vocabulary size as an estimation of general proficiency. Meara and Jones (1988) found that a vocabulary size measure they developed correlated highly with students' scores on a general proficiency placement test, consisting of listening comprehension, grammar, and reading comprehension parts, supplemented by an oral interview. The authors observed positive correlations ranging between .66 and .72. Additional evidence that suggests the same can also be found in Laufer (1997) and Meara and Buxton (1987). Thus, there seems to be support for using scores on a vocabulary size measure as a rough indicator of general proficiency.

A second, related aspect has to do with the relation between the size of a learner's vocabulary and their receptive command of collocations. It stands to reason that vocabulary size, or vocabulary breadth as it is sometimes called, has a tangible effect on practically all language skills. As argued by Meara, "All other things being equal, learners with big vocabularies are more proficient in a wide range of language skills than learners with smaller vocabularies, and there is some evidence to support the view that vocabulary skills make a significant contribution to almost all aspects of L2 proficiency" (1996:37). I consequently expect that vocabulary size to some extent will correlate positively with scores on my tests. The question is how much. As was stated in Chapter 2, vocabulary size is generally seen as the number of words an individual knows. Thus, there is a clear focus on single words. The test formats I am developing—COLLEX and COLLMATCH—are in contrast focusing on the combinatorial potential of words. My test formats therefore go beyond the aspect of knowledge captured in vocabulary size measures. In fact, as we saw in Chapter 2, the collocation aspect of word knowledge is often targeted as one part of vocabulary depth measures (see e.g. Read 1993, 1998; Wolter 2005; Stæhr Jensen 2005). However, to me, it is not all-together clear that my tests are depth tests. This is an issue that I will come back to in Chapter 6 of this thesis.

In the previous three studies, I used a test called SINGLEX to test the more infrequent single words that were featured in the COLLEX and COLLMATCH test items. This was done to control for word difficulty. If learners do not recognize certain collocations, it might be because they do not know the single words that make up the collocations. By introducing a

proper measure of vocabulary size, it will be possible to analyse in more detail how smaller and larger vocabularies contribute to scores on my tests. The question is consequently in what way informants' vocabulary size affects their receptive command of collocation.

Summing up the considerations at hand, I will introduce a vocabulary size measure in my test battery, which can be used either as a rough indicator of general English proficiency, or as a variable in its own right. I will also collect data from upper-secondary-school level learners, since these are one of my target groups. This step is linked to one of the overall aims of the study: to construct and evaluate the effectiveness of tests of collocation knowledge, aimed for upper-secondary school and university-level learners of English.

4.2.3 Methods

4.2.3.1 Developing an alternative test format: COLLMATCH 2

4.2.3.1.1 Test format

As has been pointed out earlier, there were some obvious drawbacks with the COLLMATCH 1 test. One had to do with the test format per se. Despite the fact that the test consisted of as many as 144 items, only 51 of these were real collocations. This meant that the test primarily measured learners' ability to reject pseudo-collocations (65%), rather than their ability to recognize real collocations (35%). The large number of pseudo-collocations was to a considerable extent a function of the format per se, i.e. the grid with three verbs (or adjectives) and six shared potential collocates. For this reason, together with an unacceptably large number of poorly functioning items, a clear need for a new format presented itself.

One of the first orders of business was to decide on a modified format. There were a number of specifications that I intended the format to follow.

a) to tap learners' receptive recognition knowledge of collocations;
b) to be able to test a large number of items in a short time;
c) to include a fair portion of pseudo-collocations as a means to control the possibility of learners overstating their knowledge;
d) to test collocates of high-frequency verbs.

For COLLMATCH 2, I opted for a yes/no format. In fact, the grid format used in COLLMATCH 1 was also a type of yes/no format. In a typical yes/no format, a test-taker is asked to make a judgement about whether an item is or is not a word. It is also possible to ask whether a test-taker knows the meaning of the presented item. In the present format, I decided to ask my test-takers to indicate whether or not they think that a sequence of words presented constitutes a frequently occurring word combination in English. The reason for why I did not use terms like 'collocation' was that I didn't expect all of my informants to be familiar with this term. Furthermore, I did not ask if they knew the meaning of the presented items, since this is problematic when it comes to word combinations. The reason for why I think is problematic is that the task is subject to very different interpretations from different test-takers. Let us take a sequence like **pay patience* as an example. This sequence is intended to be a pseudo-collocation. However, if I ask someone if s/he knows what the sequence means, the answer might be yes, simply by virtue of decoding the meaning of the component parts, i.e. the two single words, and then inferring meaning that is plausible. This does not mean that

the same person sees the sequence as a frequently occurring word combination in English. Another person would perhaps answer 'no' to the question whether the meaning of the sequence is known, because the sequence could not be matched with any known concept in that person's mental inventory.

4.2.3.1.2 Item selection

In line with my earlier test versions, I aimed at keeping the frequency of the tested words fairly high. If word frequencies in a test are too low, then vocabulary size no doubt becomes a decisive factor. This will make it difficult to understand to what extent the test is actually measuring knowledge of collocations and not knowledge of single word meaning. I wanted to avoid this since the aim is to measure receptive recognition knowledge of collocations as an independent construct. Consequently, twenty high-frequency verbs, all taken from the first thousand most common words of English according to frequency counts based on the BNC (Kilgarriff 1996), were checked for frequent collocates. A large number of these verbs are de-lexical verbs, and they all display a high degree of polysemy. The 20 verbs were *have, do, make, take, give, keep, hold, run, set, lose, draw, say, break, raise, bear, serve, catch, pull, throw*, and *drop*. For each of the 20 verbs, five test items, consisting of the verb + NP, were constructed. This was done through creating lists of frequent collocates for each of the 20 verbs, and then selecting significant collocates based on z-scores. The NP was either a bare noun or an article plus a noun. A varying number of the 5 items for each verb was made up by a verb plus a pseudo-object NP, serving as distractors. In total, the 100-item COLLMATCH 2 consisted of 65 real collocations and 35 pseudo-collocations. As a result, the new format measures learners' recognition knowledge of real collocations to a greater extent than the old format. A row of five items is illustrated below in Figure 4.9:

a. draw the curtains	b. draw a sword	c. draw a favour	d. draw a breath	e. draw blood
☐	☐	☐	☐	☐

Figure 4.9 A row of five test items based on the verb *draw* in the modified COLLMATCH 2 format.

The task for the informant taking the test is to tick the word combinations they think occur frequently in the English language, and leave the boxes of the non-existing collocations blank. In many of the rows for each verb, the items capture different senses of the verb. In comparison to Swedish, for example, in the row for the verb *set*, the meaning potential of the verb in the collocations *set sail*, *set an example* and *set a trap* correspond to the Swedish verbs *sätta* (as in *sätta segel*), *statuera* (as in *statuera ett exempel*), and *gillra* (as in *gillra en fälla*), respectively, when it comes to conventionalized translation equivalents. The complete COLLMATCH 2 test is included as Appendix 4F.

4.2.3.2 COLLEX 4

For the fourth version of COLLEX, minor changes were made. As was concluded in section 4.1, COLLEX 3 behaved reliably and its items displayed acceptable item-total correlation coefficients. One problem, however, was the high facility values. Advanced learners scored very highly, which of course is very positive from a pedagogical point of view. From a testing perspective, though, it presents problems in terms of ceiling effects. An attempt was therefore

made to replace some of the items displaying high item facility values from COLLEX 3 with new items that were believed to be more difficult. The items included in COLLEX 4, together with their z-scores obtained from the BNC, are presented as Appendix 4G. The new items are marked in bold typeface. The COLLEX 4 test version is presented as Appendix 4H.

4.2.3.3 Introducing a measure of vocabulary size – the Vocabulary Levels Test

As was discussed in the introductory section, the introduction of a vocabulary size test in the test battery would allow for interesting analyses of the role vocabulary size may play in collocation recognition.

In Chapter 2, the most commonly used measures of vocabulary size were reviewed. I decided to use one of these for inclusion in the test battery: The Vocabulary Levels Test (see Nation 1990, 2001; Beglar & Hunt 1999; Schmitt *et al.* 2001). The version I decided to use was published in Schmitt (2000), and the same version was validated in a study published in Schmitt *et al.* (2001). This version of the test consists of five parts with ten 'blocks' in each part. Each 'block' consists of six words together with three definitions. An example of a test 'block' is shown below in Figure 4.10.

1. apply		
2. elect	a. __	choose by voting
3. jump	b. __	become like water
4. manufacture	c. __	make
5. melt		
6. threaten		

Figure 4.10 An item 'block' example from the Vocabulary Levels Test (VLT) (Nation 1990:265).

The five parts of the test correspond to five frequency levels, from which the inherent test item words were sampled. The frequency levels are 2,000, 3,000, 5,000, and 10,000. In addition, there is a level called ACADEMIC, which samples frequent words from academic texts across subjects and fields of study.

4.2.3.4 Material

The test material used in the present study comprised a test battery consisting of three parts. The three parts were:

a) Version 1 of the Vocabulary Levels Test (150 items) (Schmitt 2000).
b) COLLEX 4 (50 items)
c) COLLMATCH 2 (100 items; new format design)

It should be noted that one feature of earlier COLLEX versions could no longer be used. The boxes which informants could tick in each item, aimed at indicating guesses, had to be discontinued. The reason for this is given in 4.2.3.7 below. The COLLEX and COLLMATCH test parts as they appeared to the informants are shown in Appendices 4F and 4H.

4.2.3.5 Informants

The total number of students in the study was 188. In addition to university students, two intact classes of upper-secondary school students—10th graders and 11th graders—who have 7 and 8 years of classroom exposure to English, were subjected to the test battery. They were all students from a local upper-secondary school. One of the classes consisted of 10th grade students (N = 26), and the other consisted of 11th grade students (N = 28). All of these students had an obligatory school background of 9 years prior to entering upper-secondary school, which for most students meant having received English instruction for 6 to 7 years.

The university students were fulltime students of English at Lund University. They studied at different levels: either first term, second term, or third term. They had completed the mandatory nine school years, plus three years of upper-secondary school before entering university, which for most students meant having had received English instruction for nine to ten years.

4.2.3.6 Research questions

The following research questions are addressed in the study:

1. Are COLLEX 4 and COLLMATCH 2 reliable tests in terms of internal consistency, and do the test items have a satisfactory discriminatory power in terms of item facility and item-total correlations values?
2. Do COLLEX 4 and COLLMATCH 2 discriminate between upper-secondary school level students and university students?
3. What is the relation between vocabulary size and scores on COLLEX 4 and COLLMATCH 2, and does this relation vary according to study level affiliation?
4. Is there a relation between general proficiency in English and scores on COLLEX 4 and COLLMATCH 2?

4.2.3.7 Test administration and scoring

In terms of gathering data from university students, it was possible to administer the whole test battery as the obligatory departmental vocabulary exam, given at the end of each term. For policy reasons, it was not possible to administer the test battery to first term students, except for a very small group who followed an older curriculum. Therefore, in the exam, primarily second and third term university students of English participated. The university students taking the test had a maximum of 3 hours to complete the test battery, which for the overwhelming majority of the students was ample time. A majority of the students handed in after 60 to 90 minutes. Out of the total 134 university students who sat the test, 5 students used the full 3 hours of exam time to complete the test form.

The test battery used with the university students was administered to the upper-secondary school students a couple of days later. I visited a local school and administered the test battery myself in two consecutive sessions. The students were told that they were taking part in a research project, that the scores on the test would not affect their grades, but that they were expected to do their best. A majority of the total 54 upper-secondary school students who took the test completed the test battery in 40 minutes. The longest time was spent on the Vocabulary Levels Test with its 150 items. A few students handed in after 60 minutes.

The big difference in time spent on the test between the university students and the upper-secondary school students was primarily due to the fact that the test battery constituted an end of term exam—a high-stakes event—for the university students, a fact that meant that many students most likely took their time, and double-checked their answers several times before handing in. For the upper-secondary school students, the test session had no impact on their grades. The test was run in class at the end of term, after the final grades had been presented to the students.

The tests were scored in the following way. In VLT and COLLEX 4, correct answers were awarded 1 point, whereas an incorrect answer received 0 points. In COLLMATCH 2, a correctly identified real collocation was awarded 1 point, whereas a missed real collocation received 0 points. Conversely, a correctly rejected pseudo-collocation was awarded 1 point, whereas an incorrectly ticked pseudo-collocation received 0 points.

4.2.4 Results

4.2.4.1 Introduction

The results reported in this section will be structured as follows. In 4.2.4.2 I will present the overall descriptive statistics for the three tests in the test battery. In 4.2.4.3 I will present comparisons of the group means on the three tests (ANOVAs). In 4.2.4.3.4, I will carry out a number of different correlation analyses, and finally, in 4.2.4.4, I will form new groups based on scores from VLT, which will function as the criterion measure.

4.2.4.2 Results for all informants

Descriptive statistics for the 3 tests were calculated. Table 4.14 below shows the score distributions on the respective tests, and Figures 4.11, 4.12 and 4.13 show the frequency distributions. As can be seen in Table 4.14, the mean scores were relatively high on all three tests, with 125.0 for VLT 1, 39.4 for COLLEX 4, and 77.3 for COLLMATCH 2. Judging from the values of Kurtosis and Skewness, all three distributions fall within the landmarks of normality. The high means are clearly visible also in the frequency distribution tables shown below (Figures 4.11 – 4.13), where also the negative skewness of the tests is conspicuous. For the VLT 1 scores, there is a clear clustering of scores at the very high end of the distribution, and it is evident that a large group of informants were able to max out the test. The same tendency is visible for the COLLEX 2 scores, where close to 100 informants received scores between 35 and 50.

Table 4.14 Score distributions and test characteristics of VLT 1, COLLEX 4 and COLLMATCH 2 (N = 188)

Value	VLT 1 N = 188	COLLEX 4 N = 188	COLLMATCH 2 N = 188
k	150	50	100
MPS	150	50	100
Mean	125.0	39.4	77.3
S.d.	26.9	8.1	12.8
Range	113	29	57
Minimum	37	21	40
Maximum	150	50	97
Skewness	-1.2	-.63	-.69
Kurtosis	.40	-.86	-.36
Cronbach's α	.98	.91	.92

k =number of test items
* = Maximum Possible Score

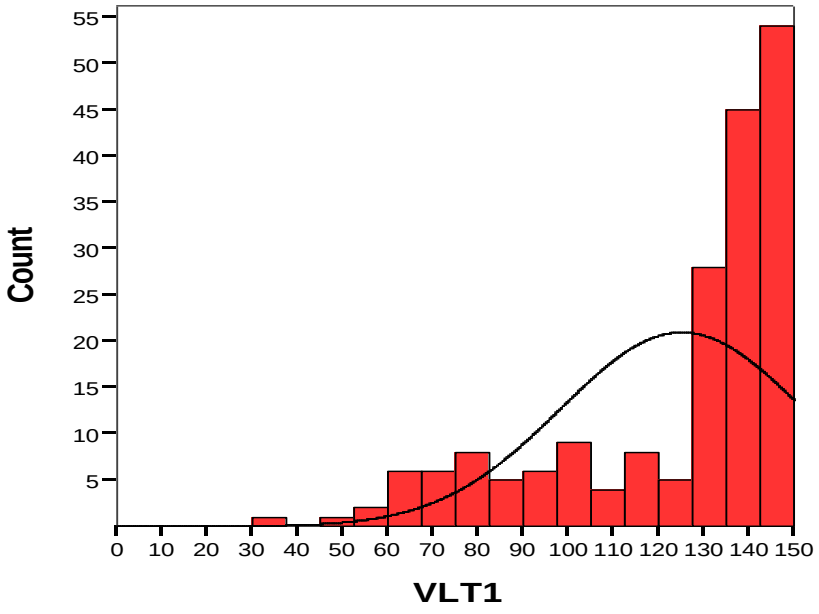

Figure 4.11 Frequency distribution of scores on Vocabulary Levels Test 1 (N = 188).

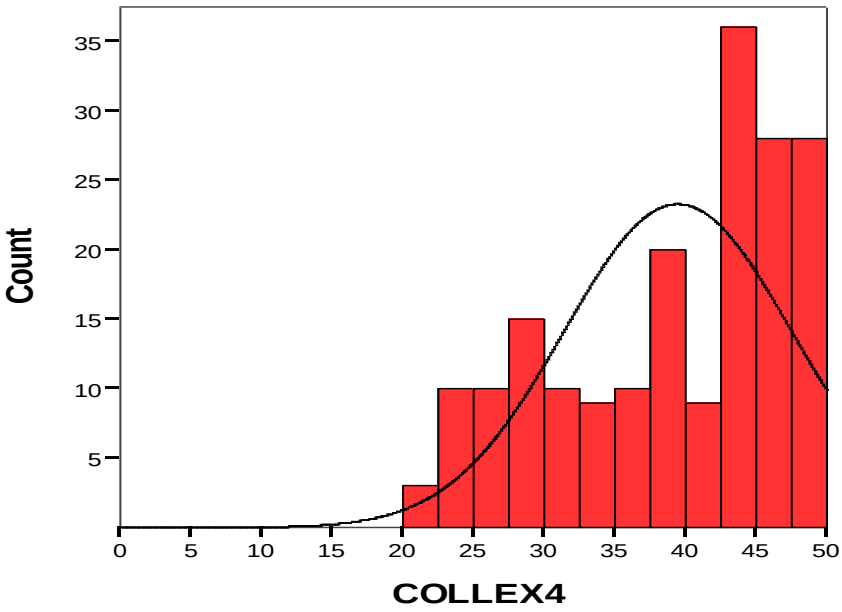

Figure 4.12 Frequency distribution of scores on COLLEX 4 (N = 188).

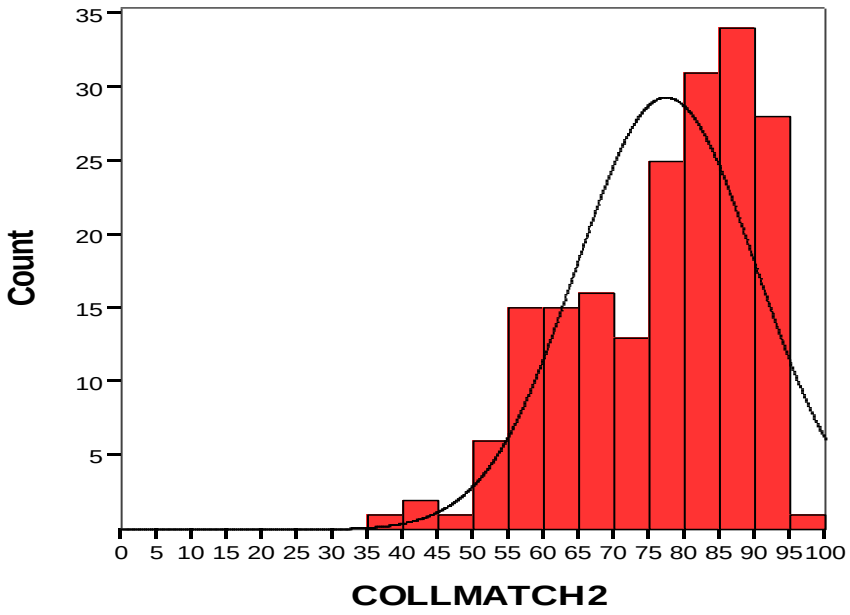

Figure 4.13 Frequency distribution of scores on COLLMATCH 2 (N = 188).

This tendency was not as clear on the COLLMATCH 2 test. It is also possible, although admittedly not very obvious, that the distributions verge on bi-modality, which suggests that there were two clearly different populations taking the tests. The further analyses reported in this section will shed more light on this issue.

The overall reliability coefficients, as estimated through Cronbach's α, were satisfactorily very high, at .98 for VLT 1, .91 for COLLEX 2, and .92 for COLLMATCH 2. This means that all three tests displayed a very high degree of internal consistency. It is also indicative and supportive of the fact that they were measures of a single, uni-dimensional construct, and that they seemingly functioned well in their ability to discriminate between test-takers.

4.2.4.3 Cross-sectional data: comparisons of Swedish learner groups

4.2.4.3.1 Learner groups used in this subsection

The cross-sectional data presented in this section is based on a number of groups of Swedish-speaking learners of English. The groups are shown in Table 4.15 below.

Table 4.15 Informant groups used in the cross-sectional analysis of the test data.

Informant group	Number
SWE10: Swedish upper-secondary school students (first year – 10th graders)	26
SWE11: Swedish upper-secondary school students (second year – 11th graders)	28
SWEuni1: Swedish first-term university students of English	7
SWEuni2: Swedish second-term university students of English	91
SWEuni3: Swedish third-term university students of English	36
Total	188

4.2.4.3.1.1 VLT

As can be seen in Table 4.16, scores on the Vocabulary Levels Test increase with higher level of study, with the exception of group SWE10 (10th graders) who scored better than group SWE11 (11th graders). Also, only a minuscule difference could be observed between mean scores of groups SWEuni2 (second term university students) and SWEuni3 (third term university students).

Table 4.16 Results on VLT 1 (k = 150) by cross-sectional groups.

Group	N	M	S.d.	Reliability[1]
SWE10 (10th graders)	26	95.3	17.1	.93
SWE11 (11th graders)	28	80.4	20.2	.95
SWEuni1 (1st term university)	7	129.0	10.6	.90
SWEuni2 (2nd term university)	91	140.5	7.6	.89
SWEuni3 (3rd term university)	36	140.8	5.5	.81
Total	188	125.2	26.6	.98

[1] Cronbach's alpha

Since Levene's test signalled unequal variances between the groups, and since the different group sizes violated the assumptions of a regular ANOVA, a Welch test was used. This test revealed a highly significant effect of learner group affiliation on scores on the test, Welch F (4, 32.6) = 97.07, $p < .001$. After having run a post-hoc Games-Howell test, I observed significant differences at $p < .05$ between the groups of 10th graders and 11th graders, and between these two and all three university student groups. No significant differences were found between the three university students groups.

The administration of the vocabulary size measure (VLT) provided excellent total reliability coefficients. Cronbach's alpha was estimated at α .98. The subgroups varied between α .81 and α .95. These coefficients are in line with earlier reported reliability values obtained for learner scores on the test (see Schmitt *et al.* 2001).

A closer look at the performance of the different groups on the five frequency levels in the Vocabulary Levels Test, revealed that scores on the whole decreased as a function of decreased word frequency. This analysis is shown in Table 4.17 below. The maximum score on each level is 30.

Table 4.17 Mean scores and standard deviations on VLT 1 word frequency levels by groups.

Group	Level 2000	Level 3000	Level academic	Level 5000	Level 10000
SWE10	26.7 (3.2)	24.0 (3.9)	21.1 (4.9)	16.9 (4.5)	6.7 (4.4)
SWE11	25.5 (3.8)	19.3 (5.9)	18.2 (4.7)	11.9 (5.6)	4.7 (4.0)
SWEuni1	29.9 (0.4)	29.0 (1.3)	27.6 (1.4)	26.7 (3.1)	15.9 (6.2)
SWEuni2	29.8 (0.5)	29.7 (1.0)	29.0 (1.3)	28.7 (1.8)	23.4 (4.3)
SWEuni3	29.9 (0.4)	29.9 (0.3)	29.4 (0.8)	28.6 (1.3)	22.9 (3.9)

The table shows that the test part consisting of academic words fit neatly between the 3K and the 5K levels, in terms of mean difficulty. The three university student groups performed well on the 2K, 3K, Academic, and 5K word levels, where they all had a mean above 26, which corresponds to 87 per cent of the total score for each level. The two groups of upper-secondary school students (SWE10 and SWE11) scored considerably lower, and these groups were also much less homogeneous as evidenced by the higher standard deviations, already at levels 2K, 3K and Academic.

4.2.4.3.1.2 COLLEX 4

The results on COLLEX 4 are presented in Table 4.18 below.

Table 4.18 Results on COLLEX 4 (k = 50) by cross-sectional groups.

Group	N	M	S.d.	Reliability[1]
SWE10 (10th graders)	26	29.9	5.1	.64
SWE11 (11th graders)	28	28.6	4.1	.45
SWEuni1 (1st term university)	7	34.5	6.7	.81
SWEuni2 (2nd term university)	91	43.8	4.7	.81
SWEuni3 (3rd term university)	36	44.2	3.3	.64
Total	188	39.4	8.1	.91

[1] Cronbach's alpha

As can be seen in the table, the mean scores of the respective groups mirror those observed on the Vocabulary Levels Test. Scores on COLLEX 4 increase as a function of higher level of study, with the exception of group SWE10 (10th graders) who again scored better than group SWE11 (11th graders), with means of 29.9 and 28.6 respectively. The results show that there was a clear difference in mean performance on COLLEX 4 between upper-secondary school students (groups SWE10 and SWE11) on the one hand, and university students (groups SWEuni1, SWEuni2, and SWEuni3) on the other.

The 3rd term students scored the highest mean (44.2), followed by the slightly lower mean score for 2nd term students (43.8). The small group of 1st term learners scored considerably lower, with a mean score of 34.5. A Welch test revealed a highly significant effect of learner group affiliation on scores on the test, Welch F (4, 33.7) = 101.75, $p < .001$. When analysed through a Games-Howell post hoc test, the observed differences were significant between 10th graders and 2nd and 3rd term university students, respectively. A significant difference was also observed between 11th graders and 2nd and 3rd term university students, respectively. Finally, a significant difference was also found between the scores of the 3rd term and the 1st term university students. Differences were minimally reached at $p < .05$.

Table 4.19 Significant differences between group means on COLLEX 4.

	SWE10: 10th graders	SWE11: 11th graders	SWEuni1: 1st term university students	SWEuni2: 2nd term university students	SWEuni3: 3rd term university students
SWE10: 10th graders		n.s.	n.s.	**	**
SWE11: 11th graders	n.s.		n.s.	**	**
SWEuni1: 1st term university students	n.s.	n.s.		n.s.	*
SWEuni2: 2nd term university students	**	**	n.s.		n.s.
SWEuni3: 3rd term university students	**	**	*	n.s.	

* The mean difference is significant at $p < .05$.

** The mean difference is significant at $p < .001$.

The notation n.s. indicates non-significance.

The overall scores were highly reliable with an internal consistency of α .91. As can be seen in the reliability column in Table 4.18, the coefficients for the 10th and 11th graders' scores, together with the university 3rd term students' scores, were low (.64, .45 and .64). The potential reasons behind this will be addressed in the discussion section.

The item quality in COLLEX 4, as based on the performance of the 188 learners of English, was satisfactory, with a mean item facility of .79 and a mean item-total correlation of .38. All the items and their values are shown in Appendix 4I. The mean item facility values for the different groups are shown in Table 4.20 below.

Table 4.20 Mean Item Facility values on COLLEX 4 by cross-sectional groups.

Group	SWE10	SWE11	SWEuni1	SWEuni2	SWEuni3
Mean Item Facility	.60	.57	.69	.88	.89

Table 4.20 shows that the mean facility values for groups SWEuni2 and SWEuni3 were very high, at .88 and .89, respectively. Clearly, even though the mean facility of the sample tested was lower than in previous test administrations reported on in this thesis, the means for the most advanced university student groups are slightly too high from a norm-referenced testing perspective.

4.2.4.3.1.3 COLLMATCH 2

The results on COLLMATCH 2 are shown in Table 4.21 below. The number of informants, the mean scores, the standard deviation, and the internal consistency of the scores are presented.

Table 4.21 Results on COLLMATCH 2 (k = 100) by cross-sectional groups.

Group	N	M	S.d.	Reliability[1]
SWE10 (10th graders)	26	62.1	8.6	.79
SWE11 (11th graders)	28	60.5	7.5	.71
SWEuni1 (1st term university)	7	71.5	11.4	.90
SWEuni2 (2nd term university)	91	84.3	7.3	.83
SWEuni3 (3rd term university)	36	84.5	5.7	.73
Total	188	77.2	12.7	.92

[1] Cronbach's alpha

The scores on the COLLMATCH 2 test mirrored those both on the Vocabulary Levels Test and COLLEX 4. Again, the 10th graders scored better than the 11th graders (62.1 compared to 60.5). The small group of 1st term university learners scored a mean of 71.5, and almost no difference was observed between the means of the 2nd term and 3rd term university students.

A Welch test was run for the group means. This test revealed a significant group effect on scores, Welch F (4, 33.4) = 83.29, p < .001. Table 4.22 below shows which group means were significantly different from each other, as evidenced through a Games-Howell test. The observed differences between the means of the groups were significant at $p < .001$ only for 10th graders and 11th graders on the one hand, and 2nd term and 3rd term university students on the other.

Table 4.22 Significant differences between group means on COLLMATCH 2.

	SWE10: 10th graders	SWE11: 11th graders	SWEuni1: 1st term university students	SWEuni2: 2nd term university students	SWEuni3: 3rd term university students
SWE10: 10th graders		n.s.	n.s.	**	**
SWE11: 11th graders	n.s.		n.s.	**	**
SWEuni1: 1st term university students	n.s.	n.s.		n.s.	n.s.
SWEuni2: 2nd term university students	**	**	n.s.		n.s.
SWEuni3: 3rd term university students	**	**	n.s.	n.s.	

** The mean difference is significant at $p < .001$.
The notation n.s. indicates non-significance.

The overall reliability of the new version of the test was found to be very high at α .92. The coefficient values for the different groups were lower, ranging between α .71 and α .90. These values are all acceptable, but they might still be somewhat low considering the large number of items in the test (k = 100). If analysing only the reliability of the scores on the 65 real collocations, the data are highly reliable at α .92. An analysis of the 35 pseudo-collocations yields a reliability coefficient of α .76. Thus, the students' ability to recognise real collocations was more reliably measured than their ability to reject pseudo-collocations.

The item quality in COLLMATCH 2 was on a lower level than that of COLLEX 4, but still satisfactory with a mean item facility of .77 and a mean item-total correlation of .32. All the items and their values are provided in Appendix 4J.

4.2.4.3.2 Correlation analyses

In order to investigate what role vocabulary size might have played for the informants' scores on COLLEX 4 and COLLMATCH 2 a number of correlation analyses were carried out. In the first analysis, the VLT scores of all 188 informants were correlated with the two collocation tests. As a first step in this analysis, scatterplots were created (Figures 4.14, 4.15, and 4.16), which clearly illustrate the negative skewness of all three tests, as evidenced through the clustering of scores in the upper right corner of the scatterplots.

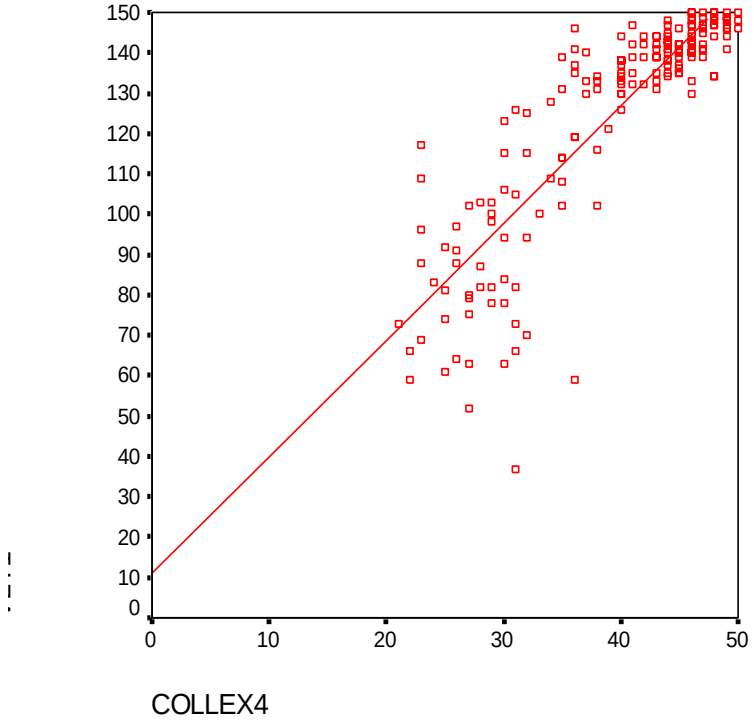

Figure 4.14 Scatterplot of VLT scores against COLLEX 4 scores (N = 188).

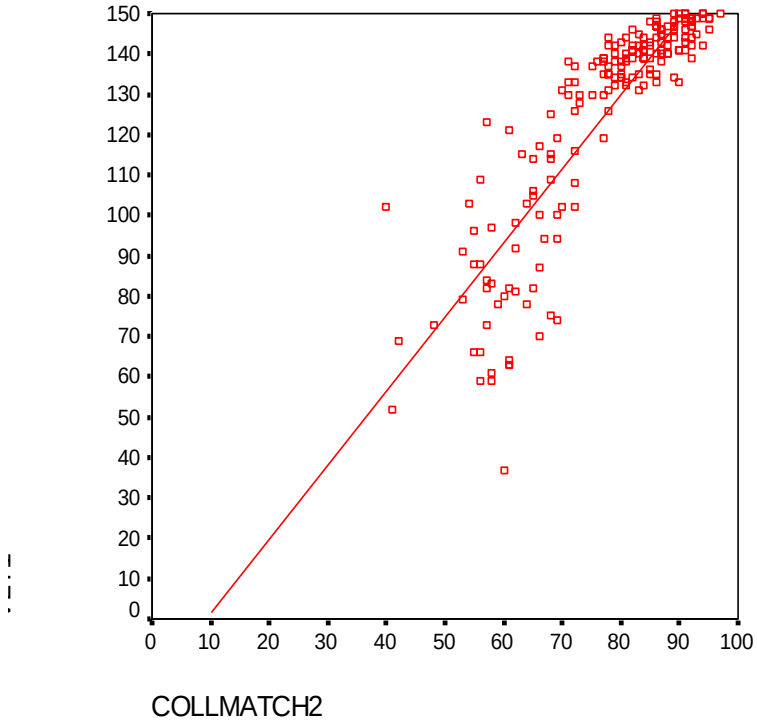

Figure 4.15 Scatterplot of VLT scores against COLLMATCH 2 scores (N = 188).

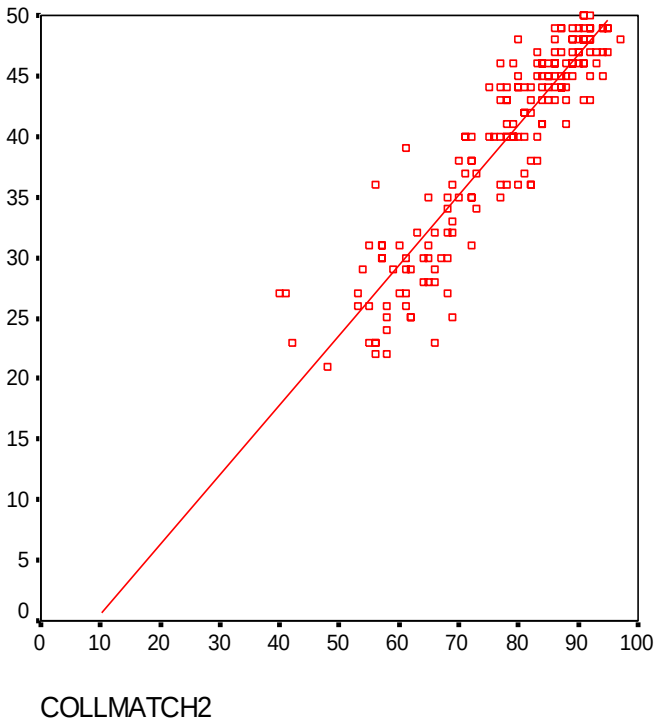

Figure 4.16 Scatterplot of COLLEX 4 scores against COLLMATCH 2 scores (N = 188).

One-tailed Pearson correlation coefficients were computed, which showed highly significant, positive correlations between the variables. The results are shown in Table 4.23.

Table 4.23 Correlations I between scores on VLT, COLLEX 4 and COLLMATCH 2 (N = 188)

Test	VLT	COLLEX 4	COLLMATCH 2
VLT	-	.87**	.87**
COLLEX 4		-	.92**

** Correlation is significant at $p < .01$, one-tailed.

As could be expected based on the clear tendencies in the scatterplots, there was a strong positive relationship between the vocabulary size measure and the two collocation tests. There was also a strong positive relationship between the two collocation tests.

Next, I ran a series of correlations to compare the respective groups in terms of their VLT scores, and scores on COLLEX and COLLMATCH. The results are shown in Table 4.24, for the VLT against COLLEX and COLLMATCH correlations, and Table 4.25, for COLLEX and COLLMATCH correlations. As can be seen in Table 4.24, significant positive correlations were observed for all the groups. This means that vocabulary size was a factor that can be associated with scores on COLLEX and COLLMATCH for both higher ability and lower ability students. The only value that sticks out is the correlation between VLT scores and COLLEX 4 for the 11th graders. It is difficult to say what caused this relatively lower correlation value, but considering the low reliability value observed for this group on COLLEX 4, only .45, a cautious approach in drawing conclusions must be adopted in general.

Table 4.24 Groupwise correlations (Pearson *r*) between scores on the VLT, and COLLEX 4 and COLLMATCH 2.

Group	N	VLT against	COLLEX 4	COLLMATCH 2
SWE10	26		.68**	.58**
SWE11	28		.40*	.51**
SWEuni1	7		.75*	.91**
SWEuni2	91		.74**	.83**
SWEuni3	36		.68**	.57**

** The correlation is significant at $p < .01$, one-tailed.
* The correlation is significant at $p < .05$, one-tailed

The group-wise correlations between the two collocation test scores show that these are inter-related for all groups, again with the 11th graders exhibiting a slightly lower value. It should also be noted that the correlation for the 1st term university group (SWEuni1) was not significant. However, the group only consists of seven informants, and this number is too small to yield significance.

Table 4.25 Groupwise correlations I between scores on COLLEX 4 and COLLMATCH 2.

Group	N	COLLEX 4 against	COLLMATCH 2
SWE10	26		.74**
SWE11	28		.50**
SWEuni1	7		.59
SWEuni2	91		.83**
SWEuni3	36		.74**

** The correlation is significant at $p < .01$, one-tailed.

4.2.4.4 New group divisions and comparisons based on scores from VLT

In the previous section, the mean scores of student groups were compared. The classification of these student groups were based on the formal level of study with which the learners were affiliated. The assumption behind the classification was that there is correspondence between level of study and general proficiency in a language. However, as was discussed in the background section (4.2.2), this classification may be slightly deceptive, since, for example, an upper-secondary school student could in theory possess a higher general proficiency level than a university level student. Thus, this assumption may not hold. For this reason, I will in this section classify the informants in the present study in a different way based on another assumption, namely that there is a correspondence between scores on a vocabulary size test, and general proficiency in a language. For this purpose, I ran a new analysis on my data, in which I tried a different classification criterion.

In my first analysis, I divided the group of informants into three groups of equal size. In order to form the groups, I eliminated the data from two informants from the analyses. I simply removed the informant with the lowest vocabulary size score (37), and, randomly, since there were many, one of the informants with the highest vocabulary size score (150). This gave me three groups of 62 informants in each. I called these groups LOW, MID, and

HIGH, based on what third the inherent scores belonged to, in the total distribution of scores. The mean and standard deviation VLT scores for each of the groups are displayed in Table 4.26.

Table 4.26 Means and standard deviations for VLT scores for three groups.

Group	N	M	S.d.
LOW	62	93.2	20.9
MID	62	136.6	3.4
HIGH	62	146.2	2.7

Based on the means of these three groups, I ran Welch *F* tests, which signalled a significant effect for group in both collocation test scores. The result for COLLEX 4 was F (2, 117.15) = 287.17, $p < .001$, and the result for COLLMATCH 2 was F (2, 116.74) = 257.11, $p < .001$. The means on the two tests for the three groups are shown in Table 4.27.

Table 4.27 Means, standard deviations, and statistical significance for COLLEX 4 and COLLMATCH 2 scores for three proficiency groups.

		COLLEX 4		COLLMATCH 2	
Group	N	Mean	S.d.	Mean	S.d.
LOW	62	29.6**	4.7	62.0**	8.0
MID	62	42.4**	3.5	81.4**	5.3
HIGH	62	46.2**	2.8	88.5**	4.4

** The mean is significantly different from other group means in the same test, at $p < .001$.

A post-hoc test (Games-Howell) signalled significant differences between all the three group means on both COLLEX 4 and COLLMATCH 2, as indicated by the asterisks in the table. Thus, based on the assumption of convergence between vocabulary size and general proficiency, students with higher proficiency in this study scored significantly better on COLLEX and COLLMATCH than did students with lower proficiency.

4.2.5 Discussion

The main goals of this study were to develop and investigate the effectiveness of a modified COLLMATCH test, and to administer this test together with COLLEX to a large group of students with different levels of language proficiency. I also wanted to investigate what role vocabulary size played in relation to scores on the two collocation tests. For these purposes, I collected data from a total of 188 students, with 134 being university students, and 54 being upper-secondary school students. The results of the study will be discussed with the research questions as points of departure.

4.2.5.1 Are COLLEX 4 and COLLMATCH 2 reliable tests in terms of internal consistency, and do the test items have a satisfactory discriminatory power in terms of item facility and item-total correlations values?

4.2.5.1.1 COLLMATCH 2

Starting with COLLMATCH 2, the changes introduced in the design of this version, compared to version 1, brought considerable improvements. A very high overall reliability coefficient was observed at .92, and the mean item facility of .77, coupled with a mean item-total correlation of .32, all lend positive support to the test and its items. All these values are positive improvements in comparison with those of the COLLMATCH 1 administration. The question is what can account for these improvements.

There is probably no easy answer to this question. One possible cause is the changed item format, and selection of test items. The fact that COLLMATCH 2 was more of a test of real collocation recognition, than pseudo-collocation rejection is believed to have had a positive effect. In COLLMATCH 2, about two thirds of the tested items were intended real collocations, whereas only about one third was intended real collocations in COLLMATCH 1. In terms of item selection, the focus on high-frequency verbs from the 1K band might have had a positive effect in that it stands to reason that these were all known by all the informants. Thus, verbs from a slightly lower frequency like *cast* and *shed* did not appear in the test. However, there were still a number of nouns of lower frequency present in the test, and the effect of single word frequency was not analysed.

Another possible cause has to do with the tested sample. In the COLLMATCH 1 administration, informants were exclusively university students. In the present administration of COLLMATCH 2, also upper-secondary school students were included. When it comes to a classical test theory reliability coefficient like Cronbach's alpha, a wider range of scores creates an increase in score variance, and this in turn creates higher reliability (see Brown 1983). The fact that the 100-item COLLMATCH 2 gave rise to a much higher reliability value than the 144-item COLLMATCH 1 illustrates the fact that having a longer test does not automatically lead to higher values as long as the overall item quality is not on an acceptable level.

The reliability of the different student groups was lower than the overall value, but still within the realm of acceptable levels (.70 - .90). Also in this regard COLLMATCH 2 was an improvement compared to its predecessor.

In conclusion, I believe the improvements were reached thanks to both an improved test *per se*, including the test format task and the item quality, and the inclusion of a wider range of scores.

4.2.5.1.2 COLLEX 4

The observed overall reliability for COLLEX 4 was on a par with that of COLLMATCH 2. With a Cronbach's alpha of .91, an item facility mean of .79, and a mean item-total correlation of .38, COLLEX 4 shows promise as a test tool. However, a word of caution is needed when it comes to the reliability of the subgroups. Especially the low reliability coefficient (.45) observed for group SWE11—the 11th graders—is a cause for concern. This level is clearly unacceptable. The reason for the high proportion of measurement error in the scores of these students is believed to come from a great deal of guessing. If there is much guessing, then this results in a great deal of variance that is unsystematic, and consequently

the measure will not reflect their true ability. Looking more closely at the item-total correlation values for the 50 tested items in the scores of the 11th graders group (N = 28), we see that as many as 17 out of the 50 items, almost 40%, have negative values. This means that on these items, many learners with low total scores on the test gave correct answers, whereas learners with high total scores gave wrong answers. Clearly the test does not discriminate well between learners of different proficiency levels in this group. All of these observations point to guessing as a highly probable cause.

In the scores of the 10th graders (N = 26, Cronbach's alpha .64), this negative trend is not so strong but we find 8 items with negative values. As for the scores of the 3rd term university students, we find 5 items with negative item-total correlations. In their case, the low overall reliability is at least partly believed to stem from high and homogeneous group scores.

4.2.5.2 Do COLLEX 4 and COLLMATCH 2 discriminate between upper-secondary school level learners and university learners?

On the whole, both COLLEX 4 and COLLMATCH 2 discriminate between upper-secondary school level learners on the one hand, and university learners on the other. The only violation of this pattern is the lack of significant differences between the tiny group of 1st term university students and the upper-secondary school level students. There are a couple of feasible explanations for this.

Looking at the standard deviations for the 1st term university student group, we see that they are relatively large, with 6.7 (Mean 34.5) for COLLEX 4, and 11.4 (Mean 71.5) for COLLMATCH 2. Clearly, the individuals vary quite considerably in terms of scores. This is furthermore corroborated by the wide range of VLT scores produced by the informants: 117, 121, 123, 126, 130, 139, and 147. These scores tell me that the small group is a rather heterogeneous group.

Another point is that the group consisted of only 7 informants. This is indeed a very small sample size, and it is in fact questionable if comparisons with other groups are meaningful. Also, the mean scores produced by first term students on COLLEX 3, in the previous study, were, relatively seen, much higher (Mean 40.4). All these findings point to problems having to do with the informant sample. This is not to say, however, that a larger sample size would result in a different mean score. It just means that the sample size used in the present study is too small to form a basis for any firm conclusions.

4.2.5.3 What is the relation between vocabulary size and scores on COLLEX 4 and COLLMATCH 2, and does this relation vary according to study level affiliation?

The results reported in section 4.2.4.3.2 above indicate that there is a strong positive relation between vocabulary size scores and scores on COLLEX 4 and COLLMATCH 2 for the informants in this study. Overall significant correlations were observed at .87 for both COLLEX 4 and COLLMATCH 2. That some sort of positive correlations would exist is perhaps not very surprising, since vocabulary knowledge has been shown to correlate positively with many other language skills (see Anderson & Freebody 1981). But the very high level is perhaps somewhat surprising. In comparison, Schmitt *et al.* (2004) observed lower and non-significant correlations between vocabulary size scores and formulaic sequence knowledge when tested on 94 students of English. Admittedly, the comparison between the formulaic sequence knowledge tested by Schmitt *et al.* and the collocation

knowledge tested by myself is not all-together straightforward. A better benchmark can be found in Stæhr Jensen (2005), who observed a correlation of .84 between a vocabulary size measure and a collocation subtest, part of a bigger test battery administered to 100 Danish university students of English. It seems, therefore, that high correlations can in fact be expected. Irrespective of correlation level, it is clear that vocabulary size seems to be a factor that influences scores on COLLEX and COLLMATCH.

When it comes to correlations for the different student groups of the study, these were slightly lower than the overall values. Table 4.24 is repeated here as Table 4.28.

Table 4.28 Groupwise correlations (Pearson *r*) between scores on VLT, and COLLEX 4 and COLLMATCH 2.

Group	N	VLT against	COLLEX 4	COLLMATCH 2
SWE10	26		.68**	.58**
SWE11	28		.40*	.51**
SWEuni1	7		.75*	.91**
SWEuni2	91		.74**	.83**
SWEuni3	36		.68**	.57**

** The correlation is significant at $p < .01$, one-tailed.
* The correlation is significant at $p < .05$, one-tailed

As can be seen in the table, the correlations, which were all significant, are of different strengths as an effect of student group affiliation. The highest correlations can be linked to the small first term university student group. The second highest pair of correlation values was observed for the second term university students (.74, and .83). It is difficult to interpret these results. Considering the performance on the different word frequency levels in the VLT, here repeated as Table 4.29, it is clear that the upper-secondary school students (groups SWE10 and SWE11) had problems with the words in the lower frequency bands. The maximum score on each level is 30.

Table 4.29 Mean scores and standard deviations on VLT 1 word frequency levels by groups.

Group	Level 2000	Level 3000	Level academic	Level 5000	Level 10000
SWE10	26.7 (3.2)	24.0 (3.9)	21.1 (4.9)	16.9 (4.5)	6.7 (4.4)
SWE11	25.5 (3.8)	19.3 (5.9)	18.2 (4.7)	11.9 (5.6)	4.7 (4.0)
SWEuni1	29.9 (0.4)	29.0 (1.3)	27.6 (1.4)	26.7 (3.1)	15.9 (6.2)
SWEuni2	29.8 (0.5)	29.7 (1.0)	29.0 (1.3)	28.7 (1.8)	23.4 (4.3)
SWEuni3	29.9 (0.4)	29.9 (0.3)	29.4 (0.8)	28.6 (1.3)	22.9 (3.9)

If assuming that mastery of a level presupposes scores of at least 25 out of 30, then these groups only reached mastery of the 2000-word level. The university student groups (SWEuni1, SWEuni2, and SWEuni3), though, all reached mastery of the 5000-word level. Based on these observations, it seems clear that the restricted vocabulary size of the upper-secondary school students was to some extent disadvantageous in terms of their ability to recognize collocations in COLLEX 4 and COLLMATCH 2. In fact, these two tests might have worked in part as tests of vocabulary size, not only tests of collocation knowledge. This

leads me to conclude that if I aim to minimize the influence of vocabulary size, especially if I want the tests to work well also with upper-secondary school students, I must further restrict the use of lower frequency words in future versions of the tests.

4.2.5.4 Is there a relation between general proficiency in English and scores on COLLEX 4 and COLLMATCH 2

In order to investigate whether there was an effect of general proficiency on COLLEX 4 and COLLMATCH 2 scores, I abandoned the initial division based on study levels, and instead divided the 188 informants into three groups (LOW, MID, and HIGH) according to their VLT scores. The data of two subjects were dropped in order to form equally sized groups of 62 informants in each. The comparison of the differences between the mean scores of the newly formed groups showed statistical significance for both collocation tests. From this I can conclude that general proficiency affects students' performance on COLLEX 4 and COLLMATCH 2. Pivotal to this interpretation is, of course, the acceptance of the assumed correspondence between scores on a vocabulary size test and general proficiency in a language. In section 4.2.2 I accounted for empirical evidence relevant to this claim. The advantage of using a vocabulary size measure for this purpose is the ease of administration. For example, most learners in my study finished the VLT in 10-15 minutes. This is considerably shorter than the time a full test of general proficiency normally takes. Thus, it is a quick measure with clear practical benefits. In a testing situation, practicality is not without importance. When administering as test battery, we must make sure that the overall length of the test instruments do not give rise to test fatigue. If this happens, we cannot use the collected data as intended, since the data would in all likelihood be partly infested with unsystematic variance. As a means to control for this undesirable effect, test batteries should be kept relatively short, without any loss of quality on the part of the test data collected. This is particularly true when we collect data in low-stake situations. The trade-off between practicality and quality of data renders, I argue, a vocabulary size measure as the VLT a sound estimate of general proficiency.

4.2.6 Summary and conclusions

In this study, I argued for a change in the design of COLLMATCH, and for the inclusion of a measure of vocabulary size in my test battery. I administered COLLEX 4 and COLLMATCH 2, together with the Vocabulary Levels Test to 188 students of English at different levels of study in the Swedish education system. The results were largely encouraging with high levels of reliability and overall good item quality, but with visible ceiling effects, particularly in COLLEX 4. The tests discriminated acceptably between upper-secondary school students and university students. Scores on COLLEX 4 and COLLMATCH 2 increased as a function of vocabulary size scores. In an attempt to elucidate the relation between general proficiency and scores on COLLEX 4 and COLLMATCH 2, the informants were divided into three proficiency groups. The results showed a clear effect of general proficiency on COLLEX 4 and COLLMATCH 2 scores.

On the whole, I maintain that there is good reason to continue developing both the COLLEX and the COLLMATCH test formats. In the next chapter, Chapter 5, I will attempt to decrease the observed ceiling effect in COLLEX, and through a different method of item selection I will try to make scores on COLLMATCH more generalisable to the underlying ability of receptive collocation knowledge which I argue is the measured construct.

5 Attempts at coming to grips with ceiling effects and test generalisability

5.1 Discussing weaknesses of previous versions and piloting new COLLEX and COLLMATCH versions on Swedish teacher students at university level

5.1.1 Introduction

In the previous two chapters, I described the development and continuous evaluation of COLLEX and COLLMATCH. As evidenced in Chapter 4, the tests can be argued to function reasonably well from a psychometric point of view: they produce high overall reliability coefficients; they discriminate between native speakers and non-native speakers of English; they discriminate between different levels of general proficiency, and they discriminate moderately well between Swedish students at different learning levels. However, some problems were nevertheless identified. There was a tangible ceiling effect present, especially with COLLEX. Also, more native speaker validation was called for, since, so far, the data from only a handful native speakers have been used. Furthermore, there was a tendency for low reliability for the scores produced by lower proficiency students. Furthermore, a general question that should be addressed is the level of generalisability of the test scores.

In this chapter, I will report on two studies aimed at tackling the above problems. The first study is a small-scale study in which an attempt is made to find a remedy for the ceiling effect problem in COLLEX, and to adopt a modified approach to item selection in COLLMATCH which is hoped to make generalisations from test scores to the overall construct of receptive recognition knowledge of collocations more straightforward. I will also administer a questionnaire in which questions will be asked about how informants perceive the test instruments and their qualities. This is hoped to give me valuable information pertaining to the validity of the tests. In the second study I will administer revised versions of COLLEX and COLLMATCH to a large group of Swedish students of English, and native speakers of English (in total c. 300 informants). The purpose is to establish acceptable levels of reliability and validity for these new versions in a large-scale administration. In particular, the data from a sizeable group of native speakers (> 30) will provide invaluable information based on which conclusions may be drawn about the validity of the tests.

The report of the small-scale study in 5.1.3 below will be preceded by a review of previous versions of COLLEX and COLLMATCH, and an indispensable discussion about the lingering problems associated with them, together with possible remedies. This will be done in section 5.1.2.

5.1.2 Previous versions of COLLEX and COLLMATCH – merits, problems and possible remedies

5.1.2.1 COLLEX

In my efforts to develop tests measuring receptive knowledge of English collocations, I have trialled four versions of the format called COLLEX. Table 5.1 shows the key features of those versions.

Table 5.1 Overview of key features of administered versions of the COLLEX format.

Test version	Items*	tested structures	Informants (students)	N	Mean	S.d.	Reliability
COLLEX 1	60	V + NP	2nd yr uni	19	51.7	3.3	.54
COLLEX 2	65	V + NP, Adj + NP	1st yr uni	83	52.0	6.4	.82
COLLEX 3	50	V + NP, Adj + NP	1st and 2nd yr uni + NSs	119	42.6	5.5	.84
COLLEX 4	50	V + NP, Adj + NP	1st and 2nd yr uni + 10th and 11th graders	188	39.4	8.1	.91

*Also indicates maximum point score

As can be seen in the table, the test versions have been administered to increasingly large groups of students, and to samples of increasingly heterogeneous abilities, judging from the standard deviations. We can also see that the reliability coefficients based on the scores elicited through the test tools have increased as well. The mean scores have in relation to the maximum scores of the different test versions fluctuated between 79 and 86 per cent[29].

In all versions, a test item has consisted of two juxtaposed word sequences. One of the sequences has been a targeted idiomatic collocation, whereas the other has been a distractor (also called pseudo-collocation). The informants have been asked to identify which of the two sequences is by them thought to be a frequent word combination, used by native speakers of English. In versions 1-3, a device for guessing indication was featured in the test. The informants were asked to self-report through a tick in a box if they were guessing. Thus, in versions 1-3, a COLLEX item looked like the example shown in Figure 5.1.

			tick the box if you are guessing
1	run a business	drive a business	☐

Figure 5.1 A sample item from the COLLEX format, versions 1-3.

[29] (COLLEX 1 = 86%, COLLEX 2 = 80%, COLLEX 3 = 85%, and COLLEX 4 = 79%).

In the most recent test version, COLLEX 4, the self-report guessing feature was not used. This was because the test administration was carried out as part of a regular exam, and the use of a self-reported guessing measure was therefore considered inappropriate. I simply foresaw that a large group of informants would be reluctant to admitting that they were guessing during a high-stake event like an end-of-term exam.

Through the trials of the different COLLEX versions, three problems have become evident:

a) From a norm-referenced testing perspective, the tests have been slightly too easy for the tested student samples, resulting in a ceiling effect;
b) The probability of successful guessing in an item is as high as .5;
c) Test administrations with lower ability students have sometimes produced somewhat unreliable scores.

A general principle that would remedy problems a) and b) is the creation of a more difficult test. This could in theory be achieved in a number of ways. I will in turn discuss the options of decreased word frequency, introducing a correction for guessing formula, and introducing more than one distractor in each test item.

One way of possibly making COLLEX more difficult is to use lower frequency words as part of the collocations. This would likely make the test items more difficult. However, a great disadvantage concomitant with this modification would be the increased influence of vocabulary size on test scores. This is important to avoid, since, firstly, my aim is to develop tests for both upper-secondary school and university students, and I have already observed that the vocabulary size means for the former group is somewhat problematic in this respect, and secondly, I am not intending to create a vocabulary size test. For these reasons, using lower frequency words would run counter to my aims.

A second way in which to make COLLEX more difficult, in the sense of lowering the mean score, would be the introduction of a correction for guessing procedure. Such a procedure involves reductions of measurement problems induced by informants' guessing the answers to test items, through a formula (Davies *et al.* 1999). In a correction for guessing formula referred to as 'correction for blind guessing' (see e.g. Eyckmans 2004), the raw scores from a multiple choice test are reduced based on the assumption that a person either knows the correct answer, or does not know the right answer, in which case blind guessing occurs. The formula is shown in Figure 5.2.

$$R - [W/(n-1)]$$

Figure 5.2 A correction for blind guessing formula

In the formula, let R be the number of correctly answered items in a test; let W be the number of wrongly answered items in the same test; let n be the number of alternatives in each item.

Applied to the COLLEX format, which consists of two alternatives in each item, one point for each wrongly answered item would be deducted from the sum of correct answers.

In order to evaluate the effect of the correction formula, I employed it on the data gathered in the COLLEX 4 administration. Based on the scores from the 188 informants, through the application of the above formula the overall mean went down from 39.4 to 28.7. Thus, the correction formula effectively reduced the mean score on the test. The overall reliability did not change, since the rank order of informant scores did not change. This is because the result is a simple linear transformation of the raw scores. An observed problem with the correction, however, was that nine students ended up with negative scores. This is no doubt an undesirable effect. For this reason, an alternative was tried out in which informants were not as heavily penalized. For each incorrect answer, 0.5 points was deducted. This brought the mean down to 34.0 and resulted in no negative individual scores. Naturally, no changes occur here either in terms of the observed reliability of the scores.

Even though the above correction formula may produce lower overall mean scores, they have little effect on the high scorers. For example, an informant who scores 48 out of 50 consequently got two items wrong. The score of this informant will be reduced by two points at most (or one point with the more lenient penalizing factor of 0.5), resulting in a corrected score of 46 (or 45). Since the real crux of the matter is how to tackle the ceiling effect, it seems that the introduction of a correction for blind guessing formula does not help. It should also be noted that the assumption behind the formula is an all-or-nothing kind. It is implied that informants either know the answer to an item, and answers correctly, or they do not know the answer and resort to blind guessing. This is clearly inconsistent with the way lexical knowledge can be argued to work. Firstly, it is probably not reasonable to assume that all guessing is blind guessing. It stands to reason that partial knowledge on a collocation test is psycholinguistically intuitive. There is support for these views in the literature. According to Burton, partial knowledge can be seen to imply the "possession of incomplete information that may improve the probability of a successful guess" (2002:807). Burton also argues that the concept of partial knowledge may include implicit (unconscious) memory. In a research review article, Schacter concludes, based on experiment data on implicit memory, that "subjects demonstrate that they possess a particular kind of knowledge by their performance on a task, yet they are not consciously aware that they possess the knowledge and cannot gain access to it explicitly" (1987:513). Nation (2001:349-350) reports research carried out on L1 learners concerning answer strategies during multiple-choice tests. The research, which compared high-ability (HIGH) and low-ability (LOW) readers, showed that 'knowing the answer' accounted for 8 (LOW) and 16 (HIGH) per cent of the items, whereas 'guessing the answer' accounted for 21 per cent of the items (LOW, with 35 per cent success rate) and 8 (HIGH, with 50 per cent success rate). The conclusions drawn were that guessing is not a major problem and that some sort of knowledge is the driving factor behind learners' responses. These guessing behaviours are corroborated by my own data on informants' guessing in Chapter 4, where high scorers on a collocation test reported fewer guesses than low scorers, and that they were also more successful guessers than low scorers.

A third modification that would in theory make the test more difficult is the introduction of a second distractor in the test item. With carefully constructed distractors, this would reduce the probability of successful guessing in an item from .5 to .33. We saw in Chapters 3 and 4 that the informants taking part in earlier test administrations indicated guessing means corresponding to 8% - 15% of the total number of test items. On the assumption that these

numbers include a proportion of so-called blind guesses, the scores of some informants could be slightly inflated. Thus, it seems obvious that the inclusion of a second distractor is an appealing measure with which to battle high mean scores and possibly also, but not to the same extent, ceiling effects. It is also feasible that reliability values could be positively affected by this step. I could thereby also address problem c) from above, relating to unreliable scores produced by lower ability students. To large extents, the occurrence of blind guessing in a test creates decreases in overall reliability, and also in terms of item-total correlation values, since low ability learners are likely to get some difficult items right and relatively easy items wrong. Making it more difficult for blind guessers to succeed in their guessing would consequently lead to a positive effect with regard to test reliability. As was mentioned, though, it would probably have a minimal effect on the high scorers, who have been shown to rarely resort to blind guessing anyway. However, in principle, I still think that it would be worth trying.

In conclusion, the above discussion has weighed the pros and cons of different potential remedies to high mean scores and ceiling effects. Neither item selection based on lower frequency words, nor the introduction of a correction for blind guessing formula were seen as appropriate steps, for different reasons. Creating a new COLLEX version with three alternatives in each test item, however, is believed to potentially damp down the high means observed in earlier versions, and also to increase test reliability, without any known negative side effects. Therefore, this modification will be carried out.

A final modification relates to the tested item structure types. So far, I have concentrated on verb + NP, and adjective + NP items. However, there are very few items of the latter kind in COLLEX, only 10 out of 50. It therefore seems wise, in the effort of developing a test capable of producing reliable and valid test scores, to stick to one kind of structure. This will make the test more homogeneous in terms of content, and score interpretation will be more straightforward.

5.1.2.2 COLLMATCH

In the previous chapter, I trialled two different versions of COLLMATCH. The first version, COLLMATCH 1, consisted of a grid format, in which verb and adjective prompts shared the same potential objects (head noun, in the case of adjectives). Figure 5.3 shows an example of a grid consisting of 18 items. The task for a test taker is to tick the intersecting box of words that may be felicitously combined in English. Just as for COLLEX, combinations that are believed to be used frequently by native speakers should be ticked. These are the intended target collocations.

	charges	patience	weight	hints	anchor	blood
drop						
lose						
shed						

Figure 5.3 Example of a COLLMATCH 1 grid.

The combinations that are not believed to be used in English are to be left unticked. For the second version, COLLMATCH 2, changes were made to the format. Instead of a grid format, a more traditional yes/no format was introduced. Based on 20 high-frequency verbs, a row of

five items was created for each verb. Figure 5.4 shows an example of a COLLMATCH 2 row of five items.

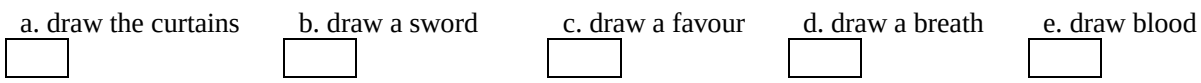

Figure 5.4 A row of five test items based on the verb draw in the modified COLLMATCH 2 format.

The test task instruction in COLLMATCH 2 asked informants to tick the boxes of those word combinations that they thought occurred frequently in the English language. The key features of the two COLLMATCH versions are shown in Table 5.2 below.

Table 5.2 Overview of key features of administered versions of the COLLMATCH format.

Test version	Items*	tested structures	Informants (students)	N	Mean	S.d.	Reliability
COLLMATCH 1	144	V + NP, Adj + NP	1st and 2nd yr uni + NSs	119	121.0	8.3	.80
COLLMATCH 2	100	V + NP	1st and 2nd yr uni + 10th and 11th graders	188	77.3	12.8	.92

*Also indicates maximum score

It is difficult to say to what extent the changes introduced in COLLMATCH 2 were the driving factors behind the improved characteristics: better reliability despite a shortened test length corresponding to 31 per cent, and a lower mean score in relation to the maximum score (77% for COLLMATCH 2 versus 84% for COLLMATCH 1). We know that greater variance in scores boosts reliability, but only if the observed increase in variance reflects differences in ability on the part of the informants vis-à-vis the intended underlying test construct. It is also clear that the lower mean score in COLLMATCH 2 stems from the lower ability of some of the informant subgroups. The fact remains, though, that the test characteristics of COLLMATCH 2 are more promising than those of COLLMATCH 1.

Apart from this, there are a couple of identifiable problems with COLLMATCH 2:

a) The lack of discrimination between a possibly omitted answer and an 'unticked' item;
b) The limited possibility of generalising results on the test to the underlying population of collocations.

Starting with problem a), In the COLLMATCH format, for each item, informants' responses can be classified in the following way. The informants are subjected to two different kinds of stimuli: collocations and pseudo-collocations. For each of these two types, either a 'yes' or a 'no' answer can be given. This amounts to four possible combinations, as depicted in Figure 5.5.

	"tick"	"no tick"
collocation	hit	miss
pseudo-collocation	false alarm	correct rejection

Figure 5.5 An item response matrix applicable to the COLLMATCH format.

The terms used in the matrix originate from Signal Detection Theory (SDT), a theory aimed at describing human sensory discrimination and decision-making behaviour in detection tasks (see e.g. Green & Swets, 1966). Hitherto in my test administrations, 1 point has been awarded for 'hits' and 'correct rejections' (grey areas in figure) and 0 points have been given for 'misses' and 'false alarms' (white areas in figure). Thus, students are rewarded not only for their ability to recognize collocations but also for rejecting pseudo-collocations. It should be noted that they are not given negative scores. One problem with this scoring method is that omitted answers cannot be separated from the answer category called 'no tick' above. In theory, an informant who did not tick an item box in the earlier COLLMATCH versions could either have left the box unticked volitionally, meaning that he or she judged the item word combination to be infelicitous (a pseudo-collocation), or the unticked box could have been a result of a lapse of concentration, meaning that no actual judgement was made about the item. This could be solved by introducing two small answer boxes under each item, e.g.:

catch a cold	draw a limitation
☐ yes	☐ yes
☐ no	☐ no

Figure 5.6 Introducing yes/no answer boxes in the COLLMATCH format items.

By using such answer boxes, omitted responses are controlled for. If at some stage a correction for guessing formula is to be applied, then a control must be introduced for omitted responses since we would for example want to penalize 'false alarms' more strictly than 'misses'. I have already indicated that troublesome tendencies of ceiling effects are more visible in COLLEX scores than in COLLMATCH scores. If, however, a correction formula is to be applied to the COLLMATCH scores, then one way of correcting scores would be to award points in the following way (see Figure 5.5 for reference):

1 point	hits, correct rejection
0 point	miss
-1 point	false alarm

Just as with the COLLEX format, an analysis was carried out in order to investigate how the correction scheme would affect the scores from the COLLMATCH 2 administration

(described in Chapter 4). Compared to the initial mean of 77.3, the application resulted in a new mean of 72.2. Thus, the penalty of -1 for false alarm responses created an overall mean decrease of around 5. Rather surprisingly, the application of the scoring method resulted in a slightly lower overall reliability coefficient, at .91 compared to the initial .92. This means that the changes in the rank order of scores, created by the scoring method, only affected reliability to a very small extent, and that it did so in a negative direction. However, the difference is so small that it is negligible.

Another approach would be to take the hits into account only when calculating raw scores. An informant's score would then equal the number of performed hits. The pseudo-collocations would then merely be used as a means for correcting the raw scores based on hits. This is perhaps of the essence when estimating vocabulary size. In vocabulary size tests, made-up non-words are used as a way to identify individuals who overstate their knowledge. Penalising an individual for ticking a non-word is somewhat justified since the informant cannot possibly have met that word in natural language exposure. The justification is not as straightforward in COLLMATCH since the words used are real words, they are not made-up. However, the combination of the words is not likely to occur in natural language by native speakers of English. Thus, informants are not overstating their knowledge, they are more making an infelicitous claim about the combinatory potential of a specific set of words. This makes it interesting not only to analyse how many collocations learners know (or recognise rather) but also to see how good they are at rejecting pseudo-collocations. This is part of the present task, and arguably also the construct, and the pseudo-collocations are not just there as a validation means.

In sum, it seems that I could arrive at lower mean scores by introducing a correction for guessing formula. However, a correction for blind guessing does not seem to be straightforwardly applicable to the COLLMATCH format. The pseudo-collocations are seen as an inherent part of the test construct. Also, the overall reliability of the scores was not improved by the application. Consequently, the disadvantages outweigh the advantages of introducing a correction for guessing formula.

Problem b) from above pertains to the scores on the test and their relation to the construct intended to be measured. In any test, the question of what construct is measured is of superordinate importance. My aim in this thesis is to measure receptive recognition knowledge of English collocations. In the COLLMATCH 2 test, twenty high-frequency verbs are presented together with some of their object collocates, and also distractors, in the form of objects that do not frequently occur together with the verbs. It stands to reason that all of the twenty verbs are well-known to the targeted informant sample: upper-secondary school and university-level students. The informants have arguably been exposed to these verbs on a large number of occasions, and they certainly know the core meaning in the sense of a Swedish translation equivalent. When it comes to the noun objects, I cannot with the same degree of certainty say whether they are all known or not. The results on a vocabulary size test (VLT) showed that upper-secondary school students did not reach an estimated mastery level of 5K words (mastery = at least 26 out of 30).

Irrespective of the recognition and knowledge levels, using high-frequency verbs in the fashion it is done in COLLMATCH 2 has its clear merits, but perhaps also some downsides. Arguably, although a 1K verb might be, and expectedly so, well known in terms of its more frequent core meaning, it does not mean that students automatically will know its collocates. If this is the case, since the same verb is featured together with five different objects, an

informant might lose up to five points for not knowing the collocates of that particular verb. Hypothetically, if this is the case for a couple of verbs in the test, the informant stands to lose quite a large number of points. In this way, a small number of words in the test have a large effect on the overall score.

As an alternative, if I were to introduce more word types in the test, the lack of knowledge of a smaller number of these words would not have the same big effect on the overall score. Instead of probing what knowledge students have about 20 verbs and a number of their collocates, I could introduce a unique verb in each test item. This would mean a move away from testing the combinatory potential of a smaller number of verbs (20) with possible NP objects, to testing a substantial number of verbs (100) with possible NP objects. The question that follows from this is whether this makes the test scores more generalisable.

Scores on a test can be investigated in two principal ways. Either the test scores are treated solely as test scores. A claim is made along the lines of an individual receiving score X on a specific test. Or, the test scores are used as a basis for inferences to some wider domain of language ability. In this way, the interpretation is not limited to the specific performance on the test, but is extended to some sort of general type of knowledge or skill of which the test performances are claimed to be examples. This second way of interpreting test scores is part of what Kane *et al.* (1999) call an interpretative argument. An interpretative argument can be seen to involve four interlinked aspects. Each link equals an interpretation inference. Figure 5.7 below models these aspects.

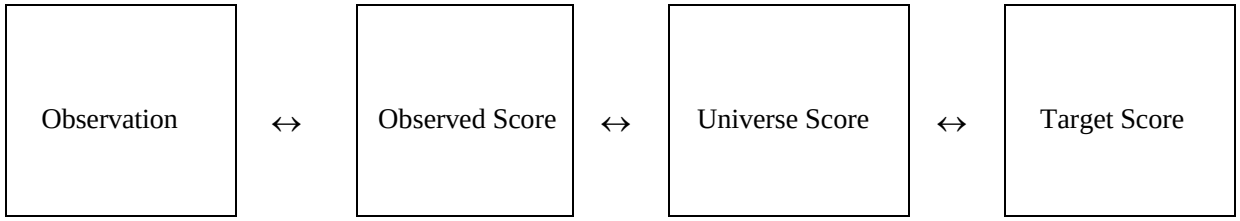

Figure 5.7 Interlinked aspects of an interpretative argument (after Kane *et al.* 1999:9).

The first aspect is the actual 'observation' of a performance, and this is in turn linked to aspect number two, an 'observed score'. The inference between the two largely rests on appropriate and clear scoring procedures. In objective tests like COLLEX and COLLMATCH, with prespecified sets of response options, scoring is argued to be a straightforward process, which clearly differentiates good performances from bad ones.

A second inference is made between the 'observed score' and a so-called 'universe score'. It involves a generalization from the actual performance on a test to a conclusion about expected performance on tasks similar to those in the test. Ideally, such a generalization is based on the assumption that 'observed scores' are based on random samples, or at least representative samples. Kane *et al.* (1999) argue that the evidence needed to support this kind of generalization may be collected in reliability studies. Reliability is at the heart of psychometric theory, and objective tests with up to hundreds of items tend to facilitate high levels of generalisability. This is so because as the size of the sample of observation for each

informant increases, generalisability increases too. Thus, I need to argue for the case that the sample of items for COLLMATCH is if not random, then at least representative. The observed performance must be extended beyond a narrow subdomain of the universe of generalization. I must also be able to demonstrate that the tests exhibit a large degree of consistency, for example through internal consistency coefficients like Cronbach's alpha. There is also reason to suppose that choosing collocations based on 100 rather than 20 high-frequency verbs should make the inferential link between an observed score and a universe score less tentative.

The third and last inference is basically an extrapolation from the 'universe score' to a 'target score'. A 'target score' is a potential performance in a target domain beyond a test. As pointed out by Bachman, "if our intended inference is a prediction about what test takers can do beyond the test, then we must assume that the tasks included in the test are representative of tasks in some target language use domain outside the test" (2004:263). The question is what kind of target language use domain COLLMATCH is capturing. I have argued on numerous occasions in this thesis that both COLLEX and COLLMATCH are tests of receptive recognition knowledge of English collocations. The interesting question is whether a receptive skill is in any way indicative of a corresponding productive skill. Are informants who score highly on the collocations in COLLEX and COLLMATCH also capable of using these collocations? This link is far from straightforward, and would have to be corroborated by empirical evidence. At the same time, it is not totally unreasonable to assume that there is some kind of relationship between a high degree of receptive knowledge of collocations, and the potential ability to use these collocations in writing and/or speaking, just as there is a relationship between receptive vocabulary size and productive vocabulary size. It is widely agreed that a person's receptive vocabulary skills are normally greater than his/her productive skills (see e.g. Melka 1997; Nation 2001), but a more exact relation is not definable, and there is evidence to suggest that large individual variation is at play.

Thus, it seems I am left with a situation where it is hard to evaluate the potential link between individuals' receptive collocation knowledge, and their ability to use collocations in language production. On the whole, though, I must tentatively and cautiously argue that it should be possible to predict that someone performing well on a test like COLLMATCH should perform better in a target domain, than someone performing badly on COLLMATCH, if the target domain is taken to be the ability to produce native-like written and spoken texts. This inference is the most problematic one since there is in fact no easily and straightforwardly definable target domain. As a comparison, if the target domain was the ability to drive a car, and our test was a close simulation of the situation of driving a car, the link would have been easier to make. An alternative and possibly more realistic candidate for the target domain could be the ability to judge whether word sequences used during natural language exposure were acceptable or infelicitous collocations. This could then be evidenced through criterion-related validity studies, in which test scores are compared to the performance of another measure of the same construct.

In sum, by employing the interpretative argument model presented in Kane *et al.* (1999), I will attempt to demonstrate that a different item selection method leads to improved generalisability in COLLMATCH. It will, however, be difficult to strongly argue for a clear-cut link to a lucid target domain.

Before I report on a small-scale study in which the new item selection will be used, I will account for another drawback of using the same verb repeatedly for a number of items. This

was pointed out to me when presenting the COLLMATCH 2 format to a number of experienced vocabulary testing researchers at an international conference[30]. My attention was drawn to the fact that the presentation of a "verb row" of five items could cause something that might be called a 'verb polysemy block' on the part of the test taker. The argument put forward was as follows. When processing the first item in a row, it is likely that the verb meaning inferred is retained in some way when processing the rest of the items in the row. For example, in a test item like *pull a trigger*, where the verb sense is literal, this sense might be retained in the mind of a test taker in such a way that a more metaphorical sense is subsequently blocked, for example in an item like *pull rank*. The effect would then be that a learner fails to recognize the latter test item as a perfectly acceptable collocation because the metaphorical reading of the verb *pull* is blocked by the preceding literal sense (Christopher Butler, and Paul Meara, personal communication). Whether this alleged phenomenon can be substantiated is up for debate. Some support for this idea could possibly be found in a study by Bobrow & Bell (1973), in which experiment participants were primed with either sentences having literal interpretations or sentences having idiomatic readings. They were then presented with ambiguous sentences with either reading possible. Those participants who had been primed with literal interpretations reported seeing literal meanings, and those who had been primed with the idiomatic set reported idiomatic interpretations. However, even though the priming effect might be present, more recent research has suggested that simultaneous computation of both literal and non-literal meanings take place (see e.g. Swinney & Cutler 1979). Consequently, it is a moot point whether the claimed existence of a 'polysemy block' rests on empirical support.

Having accounted for the previously administered COLLEX and COLLMATCH formats, and having discussed lingering weaknesses that the formats can be seen to be impaired by, together with possible remedies, I will in section 5.1.3 below report on a small-scale study aimed at finding out whether changes to the tests discussed above are potentially sound measures.

5.1.3 Piloting new versions of COLLEX and COLLMATCH

5.1.3.1 Methods

5.1.3.1.1 Item selection

5.1.3.1.1.4 COLLEX 5 – pilot version

The item selection for a new COLLEX test version was based on the version called COLLEX 4, described in Chapter 4. Firstly, only verb + NP items were used, which meant that adjective + NP items were discarded. Secondly, based on the best performing items from COLLEX 4, in terms of item-total correlation and item facility values, new items were created by adding a second distractor to each item. In this way, 40 test items were created. Figure 5.8 shows an example of what a modified COLLEX test item looks like:

[30] The 15th Vocabulary Acquisition Research Group Network Conference, Swansea, 9-11 September 2005.

1	a.	receive a cold	b.	achieve a cold	c.	catch a cold

Figure 5.8 An item example from the COLLEX 5 – pilot version format.

As in earlier versions of COLLEX, a test taker is asked to identify one word combination in each item which is believed to be a frequently occurring combination, used by native speakers of English.

Care was taken to choose high-frequency words making up the word combinations. In checking the frequencies of the individual words, the JACET 8000 (Ishikawa *et al.* 2003) word list, based on the BNC, was used. In the new 40-item version of COLLEX, a total of 112 different words (72 verbs and 40 nouns) were used, and 88 per cent of these words came from the 1-3K bands. The fact that some words used still belong to lower frequencies was governed by a need to make the distractors plausible, and this sometimes meant choosing a lower frequency word over a higher frequency word. Nouns from lower frequencies were *revenge* (6K), *fuse* (6K), and *apologies* (5K). Lower-frequency verbs were *lodge* (5K), *tidy* (5K), and *clench* (not in list). Furthermore, along the lines of procedures used in the previous test versions, z-scores were checked for both intended target collocations and distractors as a means to use conventionalized collocations as targets, and to ensure that distractors were not in frequent use as evidenced through the BNC. The COLLEX 5 – pilot version test, is shown in Appendix 5A, and the word frequencies for the words used in the test are shown in Appendix 5B.

5.1.3.1.1.5 COLLMATCH 3 – pilot version

Based on the discussion in 5.1.2.2 above, a different item selection method was employed for COLLMATCH 3. As a starting point, well-functioning items form earlier versions of COLLMATCH were selected for the creation of a shortlist. As in COLLEX, this meant picking items that displayed a combination of acceptable levels of item facility values and item-total correlation coefficients. In addition, verbs from the first four thousand words of English, according to the JACET 8000 list (Ishikawa *et al.* 2003), were analysed in terms of their noun collocates. Together with two experienced lecturers of English, candidate items were chosen. The aim was to use unique words in all items so that a word did not occur twice in the set of test items. In total, 200 words were selected for inclusion in the test. The main criterion followed was the choosing of verb + NP combinations in which the verb did not display its most typical core sense. Furthermore, combinations were chosen in which a certain degree of restriction is present in the verb use. Such a restriction stems primarily from technical, figurative or delexical uses of the verb. Examples of this can be seen in items like *run a bath*, *pay attention* and *throw a party*. In the first example, the verb *run* is used in a technical sense, which implies causing water to run from a tap. In *Collins COBUILD Advanced Learner's Dictionary* (Sinclair 2003), this sense of the verb is presented as sense number 23 out of 57 identified meanings, including phrases. The use of this technical, and slightly peripheral sense, makes *run a bath* into a collocation.

Example number two illustrates the verb *pay* in a non-monetary sense. The verb is used in a restricted sense, in which it can only be combined with a limited group of nouns, e.g. *heed, tribute,* and *visit*. In *COBUILD,* it is listed as a sense that occurs with some nouns to indicate

that something is given or done. This sense is ranked 11th out of 16 identified senses. The restricted combinability of the verb *pay* in this sense makes *pay attention* into a clear case of collocation.

The last example illustrates the verb *throw* in a clear, non-core sense, which corresponds to the process of organizing an event. This sense is given as number 15 out of 18 identified senses in *COBUILD*. In this case, the verb is heavily restricted as to its combinability with other nouns, under the condition of a retained sense. This use is informal, and arguably no other nouns are normally used together with *throw* in similar constructions. A combination like **throw a conference* is not acceptable, and the same thing goes for **throw a meeting*. However, free combinations like *throw a stone* are perfectly acceptable.

The process resulted in a list of target collocations together with pseudo-collocations. The proportion of target collocations to pseudo-collocations was 70/30. For all items, irrespective of category, z-scores were retrieved from the BNC to ensure significance for the target collocations and conversely lack of significance for the pseudo-collocations. The items were equipped with yes/no answer boxes. Examples of two items are shown in Figure 5.9 below.

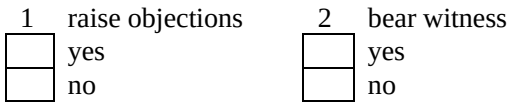

Figure 5.9 Item examples from the COLLMATCH 3 – pilot version format

As in earlier versions of COLLMATCH, a test taker is asked to identify word combinations which are believed to be frequently occurring combinations in English, whereas non-existing combinations are to be rejected. Identifying a word combination as existing is done by ticking the "yes" box, and a rejection is made through the ticking of the "no" box. The COLLMATCH 3 – pilot version test is shown in full in Appendix 5C, and the word frequencies for the words used in the test are shown in Appendix 5D.

5.1.3.1.2 Material

In addition to new versions of COLLEX and COLLMATCH, the test material used in the study consisted of a vocabulary size test, and two questionnaires, used for gathering information relevant to examining the validity of the two tests.

The vocabulary size test was a modified version of the Vocabulary Levels Test (VLT). The modification implied the removal of the 2K band of the test, and an augmentation of the 5K and 10K bands instead. The reason for this step was a purely practical one. I knew that a future large-scale test administration would be carried out in conjunction with the department's end of term vocabulary exam. This was arranged so that I could obtain data from a large group of students, with reasonable ease, and it would also have the positive effect of being data from informants assumed to do their best. This cannot always be taken for granted if volunteers are used as informants. Officials at the department feared that the relatively advanced university students at the department would score very high scores on the VLT if the test was given in its original version. This would lead to a pass cut-off score that would in turn be very high, in order to make the exam roughly equal in difficulty compared to exams given previously at the department. Therefore, the "easy" 2K band was taken out, and instead more items were added at the "more difficult" 5K and 10K bands. It seemed

worthwhile to get an indication of the difficulty level of this modified version of the VLT in the present small-scale study, before it was to be used in full-scale as part of an exam. The modified version contained items from version A, published in Schmitt (2000), and also items from version B, published in Nation (2001). The number of items in the modified test, here called VLT M (M stands for 'modified'), was 150, just as in versions A and B, and its structure is depicted below.

3K	30 items
ACADEMIC	30 items
5K	45 items
10K	45 items
Total:	150 items

The two questionnaires were incorporated in the test battery as a means to create a better understanding of how informants perceived COLLEX and COLLMATCH in terms of clarity of instruction, level of difficulty, level of appeal, and perceived tested ability. These are all aspects that can be argued to affect test validity.

The instructions of a test is normally the first thing that test-takers encounter, and as such, they play a major role in setting expectations and motivation in the test situation. According to Bachman and Palmer (1996:190), effective test instructions have three qualities: a) they are simple enough to understand; b) they are short enough not to take up too much of the test administration time, and c) they are sufficiently detailed for test takers to know exactly what they are expected to do. My question about the test instruction mainly concerned qualities a) and c), since I felt confident that they were short enough.

In terms of level of difficulty, I aimed to gather some kind of data which reflected the level of difficulty as perceived by the informants. Ideally, a test must not be felt to be too easy, since this might cause loss of motivation on the part of the test taker. Conversely, a test should not be too difficult, since this might also result in loss of motivation. I hoped that asking test takers to rate their perceived level of difficulty would give me at least a rough indication of whether the tests had a suitable level of difficulty.

The question appearing under the heading "level of appeal" was incorporated to roughly gauge the extent to which the tests appealed to the test takers. In addition to measuring the intended construct in a reliable way, ideally, a test should also be appealing and enjoyable. Reliable scores presuppose motivated test takers. It can be argued that enjoyable tasks are more likely to enhance test taker motivation than boring tasks (see Alderson *et al.* 1995:173). Therefore, I wanted to see if my collocation tests appealed to the test group.

The three questions above were constructed with a Likert scale of five points. A fourth question involved open-ended answers. This question was included in an effort to investigate what the test takers themselves thought the tests were measuring. This could provide interesting information for the overall validation process of the tests. The questionnaire used for both COLLEX and COLLMATCH is shown in Appendix 5E.

Thus, the following parts in the displayed order were administered in the study:

1. Vocabulary Levels Test (150 items, modified, version M)
2. COLLEX 5 (40 items)
3. Questionnaire on COLLEX 5
4. COLLMATCH 3 (100 items)
5. Questionnaire on COLLMATCH 3

5.1.3.1.3 Informants

A total of 25 informants participated in this study. They were teacher students at a university college in Sweden, and their mean age was 28.8 years (SD 6.6). They all studied English as one of their two major subjects. At the time of testing they had studied English for two and a half terms. This meant that they could be assumed to represent a fairly advanced group of students, with general proficiency roughly equivalent to that of the second-term students taking part in previous test administrations. This would give me a good indication of how the new versions of the collocation tests functioned with students at this level of proficiency.

Out of the 25 informants, six reported that they had other mother tongues than Swedish. Some of these claimed that they had more than one mother tongue, which I took to mean that they were in some sense bilingual. In the group there were two native speakers of English. Also, one of the native Swedish students reported that she had lived for 17 years in the UK, which meant that I considered her to possess near-native language skills. This assumption was confirmed by the students' lecturer.

All in all, I had a group of 20 Swedish-speaking students, and I also had a small group of 3 native/near-native speakers which could serve as a validation control group. In addition, I had 2 L1 speakers of other languages than Swedish.

5.1.3.1.4 Research questions

The following research questions were addressed in the study:

1. Does the 3-choice COLLEX show promise as a test format in terms of item facility and item-total correlations, and is there a ceiling effect present with the present group of informants?
2. Does the new COLLMATCH version show promise as a test format in terms of item facility and item-total correlations?
3. Do native speakers perform close to maximum scores on COLLEX and COLLMATCH?
4. What are the informants' opinions about COLLEX and COLLMATCH in terms of test instructions, perceived difficulty, and measured ability (face validity)?

5.1.3.1.5 Test administration and scoring

The test battery together with questionnaires were administered to the intact group of teacher students. A lecturer kindly offered me to gather the data at the end of a lecture in a course on sociolinguistics that the students were taking at the time. All students completed the test battery in 45 minutes.

The scoring was done as follows. In the VLT, one point was awarded for each successful match. In COLLEX, each correctly answered item was awarded one point, and each incorrect answer resulted in zero points. The scoring in COLLMATCH was performed in the following

way (see Figure 5.5 above). In each item, one point was awarded for 'hits' and 'correct rejections', and zero points was given for 'misses' and 'false alarms'.

As to the questionnaires, the answers in each scale were quantified by transforming them into numbers on a scale between 1 and 5, with 5 being the most positive response, and 1 being the least positive response. The answers to the open-ended questions were analysed qualitatively.

5.1.3.2 Results

5.1.3.2.1 VLT, COLLEX and COLLMATCH results

In Table 5.3 below, descriptive statistics are reported for the Vocabulary Levels Test, COLLEX 5, and COLLMATCH 3. In the table, values based on the scores from all 25 informants are reported.

Table 5.3 Score distributions and test characteristics of VLT M, COLLEX 5 and COLLMATCH 3 (N = 25)

Value	VLT M N = 25	COLLEX 5 N = 25	COLLMATCH 3 N = 25
k	150	40	100
MPS*	150	40	100
Mean	130.2	33.0	81.8
S.d.	12.6	3.2	7.9
Range	56	12	33
Minimum	93	27	65
Maximum	149	39	98
Skewness	-.98	-.28	.19
Kurtosis	1.9	-.77	-.35
Cronbach's α	.94	.58	.82

k =number of test items
* = Maximum Possible Score

In terms of score distributions, these were all normal as evidenced by the values of Kurtosis and Skewness, even though the scores on the VLT displayed a high level of Kurtosis at 1.9, bordering on non-normality.

The mean score on COLLEX was observed at 33.0 which corresponds to 82 per cent of the maximum score. The mean score on COLLMATCH was observed at 81.8 which in turn, if rounded up, also corresponds to 82 per cent of the maximum score. The small standard deviations of 3.2 and 7.9 respectively for COLLEX and COLLMATCH indicate that the tested informant group was relatively homogeneous.

An item analysis of the scores on COLLEX and COLLMATCH revealed the following item facility and item-total correlation values (see Appendices 5F and 5G for individual item values):

Table 5.4 Mean values for Item Facility and Item-total correlations for items in COLLEX 5 and COLLMATCH 3 (pilot versions) (N = 25).

Test	Item facility Mean	Item-total correlation Mean
COLLEX 5	.83	.13
COLLMATCH 3	.82	.17

Compared to earlier administered versions of COLLEX and COLLMATCH the Item Facility values shown in Table 5.4 are still fairly high, at .83 and .82, respectively. Since no direct comparison of mean item facility values was possible, in order to get at least a rough indication, I needed to compare the present data with those of a similar student group in terms of assumed general proficiency. I decided to use data from the previous test administration reported in Chapter 4. I took the mean score produced by 91 second term students, a group that I judged to be closest to the present informant group in terms of proficiency level. Since COLLEX 4 contained 50 items, and COLLEX 5 contained 40 items, I discarded 10 items from COLLEX 4 from the analysis. These 10 items were adjective + noun items, the type of items that were not used in COLLEX 5, so this procedure seemed logical[31]. In terms of the COLLMATCH test, both the previous version and the present one contained 100 items, so no truncation was needed. Furthermore, I decided to exclude three informants from the present data: two native speakers of English and one near-native speaker. This was done since I believed that the inclusion of data from these three informants would inflate the means from COLLEX 5. The comparison of means is shown in Table 5.5 below.

Table 5.5 A comparison of Item Facility Means from different versions of the COLLEX and COLLMATCH test.

Test version	Number of items	Number of informants	Item Facility Mean
COLLEX 4	40 (sampled)	91	.88
COLLEX 5 - pilot	40	22	.81
COLLMATCH 2	100	91	.84
COLLMATCH 3 - pilot	100	22	.80

The comparison is based on the assumption of similarity of proficiency levels between the two groups of informants. If this assumption is borne out, then the comparison shows that the new 3-choice COLLEX 5 seems to produce lower Item Facility means than the 2-choice COLLEX 4 (.81 compared to .88). In a similar way, but not as markedly, the new COLLMATCH 3 produced lower Item Facility means than the previous COLLMATCH 2 (.80 compared to .84). Admittedly, these comparisons can only provide approximate indications. Nevertheless, they point in a positive direction.

[31] It turned out that the initial mean Item Facility value based on 50 items in COLLEX 4 was .88, and the process of discarding 10 adjective + noun items did not alter that value.

In order to find validity support for the new versions of the tests, the responses from the small native speaker group were analysed. The respective scores on the three tests are shown in Table 5.6 below.

Table 5.6 Native speaker and near-native speaker performance on VLT M, COLLEX 5, and COLLMATCH 3.

	VLT M	COLLEX 5	COLLMATCH 3
Native speaker X (American English)	149	36	91
Native speaker Y (British English)	141	35	95
Near-native speaker Z (British English)	149	39	98

Firstly, when it comes to the VLT M scores, Native speaker Y surprisingly did not score the maximum or close to maximum point. A closer look at this person's responses showed that the errors were distributed as follows: -2 points in the ACADEMIC band, and -7 points in the 10K band. All the minus points in the 10K band came from omitted responses. It seems that even educated native speakers sometimes have problems with words from the 10K band. Alternatively, the omitted responses could have been caused by lack of motivation on the part of the informant.

Secondly, as to the COLLEX scores, the answers that were scored as wrong were scrutinized. In item 9, *grab an opportunity* was given as an answer by one of the native speakers (NSs). The targeted collocation was *seize an opportunity*, but the difference between *seize* and *grab* could possibly have to do with register, and *grab* is similar to the acceptable collocate *grasp*, and for these reasons the item should be modified. In item 10, *bring charges – run charges – push charges*, two of the native speakers answered *push charges* instead of the targeted *bring charges*. It is possible that the phonologically related *press charges* was targeted. I decided to change this item and use *press charges* as a target collocation in future versions. In item 11, the alternative *lend a complaint* was chosen by one of the NSs. It is viable that *lend* was misread as *land* which would perhaps be a possible but not conventionalized collocate with *complaint*. This item was kept intact. In item 15, *hold a speech* was chosen by two NSs. Based on data from the BNC, *give a speech* occurs 92 times whereas *hold a speech* does not occur. This item was also kept intact. An item that was discarded was *hold one's balance – keep one's balance – last one's balance*. Even though my analysis in Chapter 4 of concordance lines of *hold + balance* showed that in a large number of cases the phrase *hold the balance of power* was behind the obtained frequencies, I decided to replace this item. Three more items displayed discrepancies between the test key and the answers from the NSs, but these were seen as cases possibly stemming from lapses of concentration on the part of the NSs.

Thirdly, and finally, the COLLMATCH items displaying discrepancies between answers from the NSs and the test key were examined. In two cases, two or more of the NSs disagreed with the answer key. In item 64, *afford an opportunity* was said not to exist according to two NSs, and in item 89, *fill an aim* was ticked as an existing collocation by all three members of the NS group. As to the former, *afford an opportunity*, it was kept intact since there is corpus and dictionary evidence for its existence. In terms of register, it is a fairly formal phrase, and this could possibly have affected its level of rejection. As to the latter, *fill an aim*, my

judgement was that it was mixed up with *fulfil an aim*, which is a perfectly acceptable collocation. The truncated sequence *fill an aim*, however, is not, according to corpus and dictionary data. The difference in spelling is one cue that could have helped the watchful test taker. It is, though, a very subtle cue. On the whole, I decided to keep the item, despite the native speaker verdict. Out of the three NSs, the person with an American background gave the most answers that did not fit the test key. For example, phrases like *keep pets*, *pull a face*, and *realise a potential* were rejected. When it comes to the latter, the British spelling with an –s rather than a –z might have affected this informant. When it comes to the former two phrases, these might be more frequently used in British English than in American English. Since the British National Corpus was used to find collocates it cannot be ruled out that American English informants will sometimes not agree with the test key. I will have to return to this potential problem at a later stage in the research process.

In conclusion, based on the close examination of the NS answers, some items were modified, and some items were even discarded in the pursuit of developing better test versions. In addition, some items were kept intact as their existence was clearly supported by corpora and dictionaries, and feasible reasons to why the NSs responses did not match with the key could be presented.

5.1.3.2.2 Analysis of answers in questionnaires

Table 5.7 below shows the means of the tallied responses to the Likert scale questions from the COLLEX and COLLMATCH questionnaires.

Table 5.7 Mean scores and standard deviation scores for answers to COLLEX and COLLMATCH questionnaires.

Question	Scale			COLLEX mean (SD)	COLLMATCH mean (SD)
1.Level of test instruction comprehensibility	very easy	<5 4 3 2 1>	very hard	4.67 (0.70)	4.67 (0.82)
2.Level of perceived test difficulty	very easy	<5 4 3 2 1>	very difficult	3.52 (0.81)	3.13 (1.06)
3.Level of test appeal	very appealing	<5 4 3 2 1>	very boring	3.74 (0.65)	3.61 (0.78)

As to question 1, asking about the level of test instruction comprehensibility, the 25 informants gave very high marks both for COLLEX and COLLMATCH. Thus, it stands to reason that they felt that the instructions were clear and easy to understand. One student remarked that the inclusion of an item example in the instruction would enhance clarity. This is a fair remark and I will in the next test administration include such an example. On the whole, the responses from the informants speak in favour of the validity of the test instruction.

As to their answers to question 2, which asked about the level of perceived test difficulty, the informants seem to have felt that COLLEX was slightly easier than COLLMATCH. The mean for COLLMATCH is close to the mid category ‘average’, whereas the mean for COLLEX lies between the categories of ‘easy’ and ‘average’. Two native speaker informants

stated that they felt that COLLEX was very easy, which is perhaps not surprising. Two informants stated that they thought that COLLEX was difficult. Furthermore, ten informants said that COLLEX was easy and another ten said it was average. Since COLLEX and COLLMATCH are aimed to be used with Swedish learners of English, the inclusion of judgements in the data from two native speakers and a near-native speaker was likely to have boosted the mean scores. Indeed, if these three respondents were removed the perceived difficulty mean for COLLEX decreased to 3.3, and the mean for COLLMATCH decreased to 2.8.

As to the answers to question 3, the informants seem to have found COLLEX slightly more appealing than COLLMATCH, but the difference is tiny. For both tests, the mean judgements lie between the categories 'OK' and 'appealing with a very light tilt towards the latter. Interestingly, one of the native speakers commented on question 3, for which he gave a mark of 4.5, by adding "It's fun to see just how much comes completely naturally, and how 'odd' the incorrect phrases seem".

Question four in the questionnaire was an open-ended question. For this reason, the answers were not straightforwardly quantifiable. The question read: "What kind of knowledge is in your opinion measured in the test?". It should be noted that the only influence that I as a researcher might have had on their responses pertained to the test instructions. In these instructions I mentioned "word combinations" and the fact that some are "natural" and "frequent" in English. All the responses are presented in Appendix 5H, where the 25 informants are designated through codes. These codes will be used below when referring to the answers given. A general point to be made is that several comments include many different views, and a neat division into categories was not possible. I will below attempt to account for the more common answers. The answers were in a majority of the cases given in Swedish, and they have therefore been translated into English by myself.

Starting with the COLLMATCH answers, the test part that appeared first in the test booklet, ten of the informants gave answers that had some sort of bearing on collocations, phrases and/or expressions. Examples of answers are "awareness of English collocations" (UL01), "the combinatory potential of words; phrase knowledge" (UL05), "Tacit and subconscious knowledge, word combinations and phrases that one has 'collected' over the years" (UL08), and "It is knowledge which is first and foremost acquired in an English-speaking country. The knowledge chiefly measures [sic] idiomatic expressions in everyday speech" (UL21).

Five answers seem to allude to general English proficiency. Examples of answers are "General language skills, and spoken English skills" (UL11), "It deals not only with the comprehension of words, but also language familiarity, even if you know what the words mean, they may not necessarily together create something coherent. Therefore one must probably possess language fluency, not merely word knowledge" (UL19), and "If you know what combinations fit together then you know quite a lot of English. This test part measures general English skills, not just vocabulary, I dare say" (UL24). Compositionality aspects are captured by two informants, e.g. "…when words together form a meaning which the component words cannot create…" (UL19), and the comment from (UL21) quoted in the preceding paragraph. One of the native speakers supplied an answer that sticks out: "If you like poetry or not". It is not all together clear what the informant meant by this, but one interpretation is that the usage of unconventional combinations of words is a feature of poetry. The informant may thus have thought that extensive reading of poetry could make a person

more prone to accepting unconventional combinations, and therefore in effect do worse on the test.

What is interesting is that three answers refer to usage skills. One informant says, for example, that the test measures "The ability to use expressions in different situations" (UL09), and another suggests that it tests "How much English you have read, and which of it you yourself would apply in writing" (UL13). A third alludes to spoken skills (UL11). A statement that also in a way links the test content to usage is that made by Informant (UL15) who stated that the test measures "Possibly a kind of 'native' or 'vernacular' English. Doesn't feel like school English but rather more like real English in real situations".

The comments made about the COLLEX test mirror those made about COLLMATCH to a great extent, and in some cases the informants referred to their answer provided on the questionnaire already filled out. To an ever greater extent than for COLLMATCH, the informants alluded to collocations, phrases and/or expressions as being measured in COLLEX. This was done in 13 cases. Some examples are "Idiomatic expressions/collocations" (UL01), "Standard expressions in English" (UL09), "Phrases and word combinations, vocabulary knowledge" (UL03), and "Phrases that are important to know, especially interaction/conversation" (UL21). At least seven informants refer to vocabulary or word knowledge aspects, for example: "Synonym knowledge and vocabulary knowledge" (UL16), "Word comprehension" (UL18), and "I think it measures the average level of vocabulary acquired. I think it is very useful…" (UL23).

5.1.3.3 Discussion

This study was aimed at finding out whether the modified versions of COLLEX and COLLMATCH showed promise as reliable and valid tests of receptive collocation knowledge. It was also aimed at investigating the face validity aspects of the tests, by asking informants questions about the two tests.

All in all, the new 3-choice COLLEX (version 5 – pilot), which in the present version contained 40 items, worked well. With a relatively advanced group of informants, the mean of 33.0 is certainly high but not alarmingly so. Although a ceiling effect is still perceptible, there is not a large group of informants at the very high end of the score range. The mean item facility was lower than that observed for COLLEX 4 (see Table 5.1.5). The very small standard deviation indicates that the informant group was homogeneous, and this probably affected the reliability of the test, Cronbach's alpha = .58, which is indeed on the low side, and the mean item-total correlation of .13. There were some poorly performing items which need to be replaced. Also, it would probably be wise to lengthen the test by including at least ten more items, since this theoretically increases the chances of obtaining a higher internal consistency value. With carefully constructed additional items, it would be possible to stretch out the score range further. Although the low reliability is worrying, I feel confident in reaching higher reliability values with more heterogeneous groups in future test administrations.

The modified COLLMATCH test also worked relatively well. The observed score reliability was perfectly acceptable at .82, and with a wider score range, i.e. a less homogeneous group, the reliability coefficient could be expected to get even higher, together with a higher mean item-total correlation value. The mean item facility was lower than for COLLMATCH 2, a result which points in a positive direction.

The new method for item selection resulted in a test that presents the test-takers with a large number of words for which they were asked to judge the acceptability. Instead of basing the selection of collocations on 20 verbs, the new method relied on a selection of 100 verbs. High frequency verbs were aimed for, and 90 of the 100 verbs were taken from 1-3K. The remaining 10 verbs all came from band 4K. A close to identical selection method was followed for the nouns. A desired effect of this change in item selection method was the arguable increase in generalisability from observed scores to universe scores. A validation argument pertaining to this kind of inference rests on both logical aspects and empirical evidence. The basic logical aspects were discussed in section 5.1.2.2, but further implications need to be addressed here. It could in fact be argued that the kind of collocation knowledge tested in the present version of COLLMATCH is reminiscent of what Read calls "network knowledge" (Read 2004). According to Read, network knowledge is one of three ways to conceive of depth of vocabulary knowledge, the other two being "precision of meaning" and "comprehensive word knowledge" (see Chapter 2). Read points out that the three approaches overlap, and that the comprehensive approach conceptually speaking subsumes the other two. He sees the network interpretation as "learners' developing ability to distinguish semantically related words and, more generally, their knowledge of the various ways in which individual words are linked to each other" (p. 219). It is true that I have so far in this thesis seen collocation as a property of individual words. By employing a selection method that starts with verbs, and which subsequently retrieves noun collocates of those verbs based on corpus data, I have, from one perspective put focus on how much informants know about these individual words. I used Nation's word knowledge framework (2001) as my point of departure. However, the network notion discussed by Read focuses on the mental lexicon as a whole, as opposed to individual words. The network approach is normally taken to include a number of different kinds of associative links between words, notably paradigmatic, syntagmatic, phonological and analytic (see e.g. Söderman 1993; Singleton 1999; Read 2000). It is clear that COLLMATCH is directly aimed at syntagmatic links, and it is in this sense more narrow in its scope than for example the Word Associates Test (Read 1993), which targets also paradigmatic and analytic links. Despite the more narrow scope, by presenting informants with combinations based on a total of 200 verbs and nouns from the five thousand most frequent words of English, some possible and some infelicitous collocations, I will arguably tap into the mental lexicons of the learners and the degree to which there are associative links between these words. An assumption underlying this argument is the view that learners who recognize word combinations in which the inherent words appear in less typical senses (e.g. *throw a party*) will also recognize more typical senses (e.g. *throw a ball*). Learners who do recognize the former as an acceptable word combination in English have in the view accommodated the words *throw* and *party* in their network by way of forming a syntagmatic link between them. Thus, even though COLLMATCH has never been argued to be a 'bona fide' depth of vocabulary knowledge test, in its present version it has features that are affiliated with what Read (2004) calls network knowledge. I will come back to this notion in subsequent chapters.

Going back to the more empirical aspects of the tests, the analysis of the performance of two native speakers of English and a near-native speaker was carried out as part of the validation process. In the present versions of COLLEX and COLLMATCH these three informants all scored between 90 and 98 per cent of the maximum score, which can be seen to provide some evidence of test validity. Admittedly, there were a couple of items in each test

that in hindsight did not work well, and these will have to be modified or replaced. I probably cannot expect all native speakers to reach the maximum score on tests like COLLEX and COLLMATCH, due to variation in familiarity with the register of some collocations, and possibly also presence of test fatigue and lapses of concentration, but it must be seen as a problem if native speakers do not reach levels of around 90 per cent or more (cf. Greidanus *et al.* 2004).

The informants' answers to the questionnaires gave valuable insight into test characteristics such as the instruction, perceived test difficulty, and whether they enjoyed doing the tests. The last aspect is not without importance in the trade of language testing, since a test that is perceived as fun to sit is more likely to minimize measurement error like test fatigue and lapses of concentration. It is likely that it will also mean that informants will do their best. The fact that as many as 16 out of 25 thought that it was 'appealing' or 'very appealing' to do COLLEX, and 13 out of 25 thought that it was 'appealing' or 'very appealing' to do COLLMATCH, is a very positive finding. Even though two informants stated that it was 'boring' to do COLLMATCH (none for COLLEX), the overall result is clearly satisfactory.

Judging from the quantified mean opinion, the test instructions of the two tests do not seem to lack clarity, and I probably do not need to change these, except for the addition of item examples that illustrate how the task should be performed. On the other hand, the test instruction must not be ignored since it can presumably have an impact on the test task interpretation. A possible change for the COLLMATCH test could be to ask informants to tick the word combinations they know the meaning of. The potential risk with this kind of instruction could be that intended pseudo-collocations could be ticked as 'yes' responses by virtue of giving rise to some sort of meaning in the heads of the test takers. An item like **swing a secret*, which was included in COLLMATCH as a pseudo-collocation, could be seen to have a 'meaning', but arguably it does not have a conventionalized meaning that native speakers would readily acknowledge. As was expressed by one of the native speakers, it is often striking how "odd the incorrect phrases seem". It is not too bold a claim to say that this clear feeling is probably not present to the same extent with an L2 learner (see Meara 1996:48). Thus, no changes in the test instructions are warranted.

The question about the perceived difficulty attracted mostly either 'average' or 'easy' responses. Some informants commented that it was difficult to decide since they did not know their score. This is of course a valid point. Nevertheless, there is still a value in informants' estimate of difficulty level. Out of the 25 informants, 12 (48%) found COLLEX either 'average', 'difficult' or 'very difficult', and 18 (72%) found COLLMATCH 'average', 'difficult' or 'very difficult'. In conclusion, even though the mean scores show that the group as a whole did well, several individuals seem to have found the test challenging.

The open-ended question asking what kind of knowledge was being measured by the tests attracted a fairly heterogeneous set of responses. It's clear that a large number of students referred to some aspect of phraseology, and also vocabulary knowledge. These can all be seen to have clear links to the intended test construct. Among the answers, however, some answers stick out. One interesting category that emerged contained answers that linked the receptive test task to supposed usage skills. The informants seem to have felt that even though they did not produce any answers of their own – they rather picked their choices from alternatives on the piece of paper in front of them – they claimed that there was some sort of link to either written or spoken language ability. This is indeed interesting. In section 5.1.2.2 above, I

conducted a line of argument about the possible indicative link between scores on COLLEX and COLLMATCH and potential productive collocational skills. Naturally, the fact that a small number of learners make this connection is not evidence of its existence. However, it raises questions about the role of collocations. In a discussion about receptive and productive aspects of vocabulary, Melka (1997:85) claims that the knowing of collocations, as she puts it, is a "higher" degree of knowledge which is close to being productive. Intuitively, this makes sense even though it lacks empirical support. In a similar vein, Hill (2000) attributes collocation to be an important key to fluency. This is so, he argues, because by knowing a large number of collocations we can name complex ideas quickly, and we do not have to resort to using new language all the time. An assumption in Hill's claim is that by knowing a large number of collocations receptively, a learner is able to use a majority of these also productively. Even though this assumption is feasible, words of warning are given in the literature that relate to this assumption. Nation (2001:371) argues that increases in vocabulary size as measured in decontextualised vocabulary tests do not necessarily reflect an increase in vocabulary in use. Thus, it remains clear that we must present empirical support for this assumption, and not only rely on anecdotal argumentation.

5.1.3.4 Concluding summary

In this section, I have summarized the main findings from the test administrations reported in Chapters 3 and 4. In doing so, a number of weaknesses were identified both in COLLEX and COLLMATCH. Measures that were thought to remedy these weaknesses were discussed and implemented in new versions of the two tests. In an attempt to trial these modifications a study was set up in which 25 informants sat COLLEX 5 and COLLMATCH 3, both pilot versions, together with a modified version of the Vocabulary Levels Test. In addition, a questionnaire was administered in the effort to further evaluate the qualities of the tests, from a test-taker perspective. Even though the results indicated some lingering problems, they were on the whole positive. I therefore decided to continue the development of COLLEX and COLLMATCH along the lines suggested in this section.

The next step will be to administer these improved versions in a large-scale test, ideally involving several hundreds of students of English in Sweden, at different levels of study, as well as a sizeable group of native speakers of English. This is an important step as the present study, essentially a pilot, incorporated only a small group of informants. It is only through large-scale studies that important aspects of reliability and validity can be properly addressed, and more well-grounded conclusions can be drawn. Such a study will be reported in section 5.2.

5.2 Administering COLLEX 5, COLLMATCH 3, and VLT M versions to advanced Swedish students and English native speakers

5.2.1 Introduction

In the previous section, modified versions of COLLEX and COLLMATCH were piloted on a group of 25 teacher students. The modifications in COLLEX consisted of the introduction of a second distractor in each item, and in COLLMATCH a new item selection method. A questionnaire and a vocabulary size measure were furthermore included in the test battery. The modifications turned out moderately well, and a subsequent large-scale study was deemed necessary in order to investigate score reliability and validity more fully for the new test versions. It was seen as particularly important to gather validation data from a sizeable group of native speakers of English, since the previous studies have only made use of very small numbers of such informants. Consequently, in this section, I will report on a study in which a total of 308 informants, both learners of English in Sweden at different levels of study, and native speakers of English, were subjected to a test battery consisting of a 50-item COLLEX 5, a 100-item COLLMATCH 3, and a 150-item VLT version M.

5.2.2 Methods

5.2.2.1 Material

5.2.2.1.1 COLLEX 5 – full version

The creation of a full version of COLLEX 5 was based on the pilot version used in the study accounted for in section 5.1. In addition to changing some of the distractors (10 changes in total) that were seen to function relatively poorly in the previous study, 10 more items were added to create a test of 50 items. This was done in order to maximize test reliability. By making the test longer, the theoretical possibility of observing a higher reliability coefficient was increased.

The new test version is shown in Appendix 5I, and the frequencies of the individual words are shown in Appendix 5J.

5.2.2.1.2 COLLMATCH 3 – full version

Compared to the pilot version of COLLMATCH 3, presented in section 5.1, only one item was changed for the creation of the full version. This change involved a change of the object noun in item 9, from *make progress* to *make a move*. The new COLLMATCH version is presented in Appendix 5K, and the frequencies of the individual words are shown in Appendix 5L.

5.2.2.1.3 Vocabulary Levels Test – version M

A vocabulary size measure was incorporated in the test battery. The measure used was The Vocabulary Levels Test featured in a modified version, here called version M. The creation of this version was accounted for in section 5.1.3.1.2 above. In brief, the version is a mix of versions A (Schmitt 2000) and B (Nation 2001).

5.2.2.2 Informants

The total number of informants taking the test battery was 307. Three main informant groups can be discerned. The largest group consisted of university students of English at Lund University. These students were studying English at different levels at the time of the test administration. Most of them were first-term students, but sizeable groups of second as well as third-term students also took the test. The second largest group consisted of native speakers of English. These informants were all students at the Centre for Applied Language Studies at the University of Wales, Swansea. The third group of students who took part in the study were upper-secondary school students who at the time of testing were in the eleventh grade at a local school in Malmö, Sweden.

The specific distribution of informants across the above described groups is shown in Table 5.8 below[32].

Table 5.8 Distribution of informants across groups.

Informant group	Number
Upper-secondary School students of English in Sweden, 11th grade	26
University students of English in Sweden, first term	163
University students of English in Sweden, second term	49
University students of English in Sweden, third term	35
Native speakers of English, university students in Wales	35
Total	308

The university students were full-time students of English at Lund University. Being university students, they would have had to completed compulsory school, plus three years of upper-secondary school before entering university. This means that they had had received English instruction for 9 to 10 years.

5.2.2.3 Research questions

The following four research questions are addressed in this study:

1. Do COLLEX 5 and COLLMATCH 3 produce reliable test scores in terms of internal consistency, and do the test items have a satisfactory discriminatory power in terms of item facility and item-total correlations values?
2. Can COLLEX 5 and COLLMATCH 3 be argued to produce valid scores as tests of receptive recognition knowledge of English collocations?
3. What is the relation between vocabulary size and scores on COLLEX 5 and COLLMATCH 3, and does this relation vary according to study level affiliation?
4. Is there a relation between general proficiency in English and scores on COLLEX 5 and COLLMATCH 3?

[32] The unequal sizes of the particular groups are a reflection of practical matters involved in research. The majority of the test data was gathered as part of an end of term exam.

5.2.2.4 Test administration and scoring

The gathering of data in the study was done in the following ways. For the university students in Sweden, it was possible to administer the whole test battery as the obligatory departmental vocabulary exam, given at the end of each term. The students taking the test had a maximum of 3 hours to complete the test battery. It should be mentioned that a further test part was included in the exam, namely a contrastive vocabulary measure of L2 English word translation into Swedish. The results on that test part are not included in my analysis.

The same test battery used with the university students was administered to the upper-secondary school students a week later. By kind permission from a teacher contact, it was possible for me to visit a local school and administer the test battery myself in an intact group of students during an English class. I told the students that their participation was an essential part of a vocabulary research project, and that their scores on the test would not affect their grades, but that they were expected to do their best. A majority of the 26 upper-secondary school students who sat the test battery completed it in 40-45 minutes. All students had handed in after 70 minutes.

The tests were scored in the following way. In the VLT M and COLLEX 5 tests, 1 point was awarded for each correct answer, whereas 0 point was awarded for each incorrect answer. In COLLMATCH 3, a correctly identified target collocation was awarded 1 point, whereas a missed target collocation received 0 point. Conversely, a correctly rejected pseudo-collocation was awarded 1 point, whereas an incorrectly ticked pseudo-collocation received 0 point.

5.2.3 Results

5.2.3.1 Reported results

The results will be reported as follows. In 5.2.3.2, overall descriptive results are shown for all three tests, based on data from all informants combined. In 5.2.3.3, in order to be able to address the research questions raised, I report results based on analyses of a modified data set from subgroups, and with certain data not included.

5.2.3.2 Overall descriptive results

Descriptive results for all three tests were computed. One of the native speakers did not fill in one of the parts of COLLMATCH, and was therefore excluded from the analysis. Consequently, data from 307 informants were used. Table 5.9 below shows the score distributions on the tests, and Figures 5.10, 5.11, and 5.12 display the frequency distributions. The mean scores were high on all three tests. This was more or less expected since a great majority of the informants were university students of English, and the fact that the data for 34 native speakers were included.

Table 5.9 Score distributions and test characteristics of VLT M, COLLEX 5 and COLLMATCH 3 for all informants combined (N = 307).

Value	VLT M N = 307	COLLEX 5 N = 307	COLLMATCH 3 N = 307
k	150	50	100
MPS*	150	50	100
Mean	127.1	41.4	78.0
S.d.	18.6	6.8	11.1
Range	90	28	48
Minimum	60	22	51
Maximum	150	50	99
Skewness	-.99	-.80	-.06
Kurtosis	.66	-.09	-1.0
Cronbach's α	.96	.89	.89
Mean Item Facility	.85	.83	.78
Mean Item-Total Correlation	.35	.34	.26

k =number of test items
* = Maximum Possible Score

On the whole, the values for Skewness and Kurtosis indicate normality in terms of score distribution, even though the distributions on all three tests are more or less negatively skewed, as indicated by the high bars to the right, near maximum score end of the histograms.

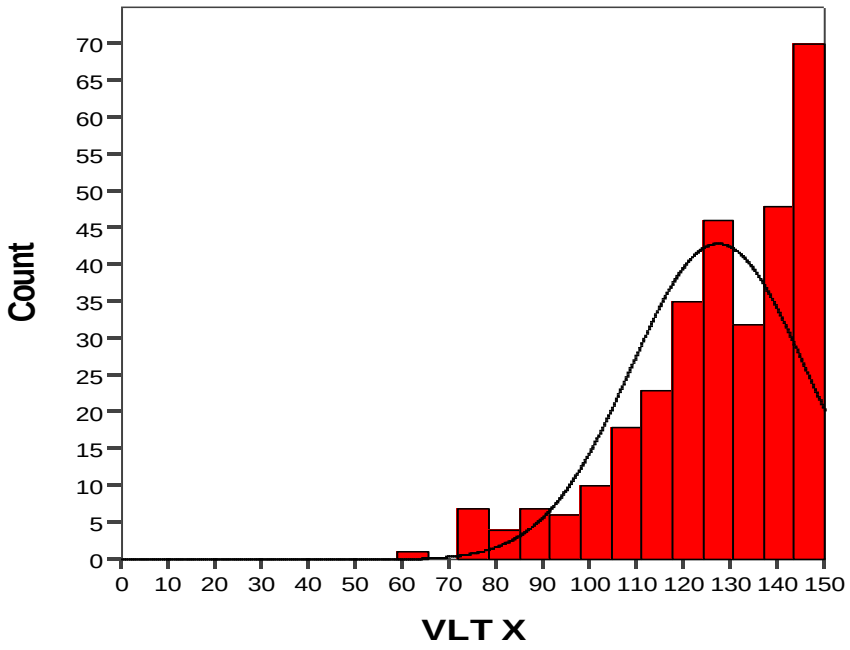

Figure 5.10 Frequency distribution of scores on VLT M (N = 307).

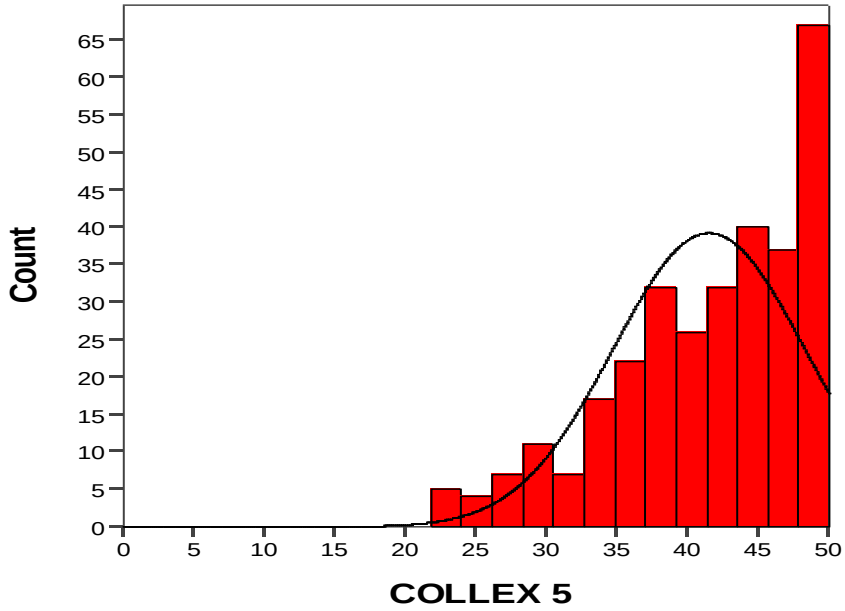

Figure 5.11 Frequency distribution of scores on COLLEX 5 (N = 307).

The scores on VLT M, COLLEX and COLLMATCH were reliable in terms of internal consistency, with Cronbach's alpha values between .89 and .96.

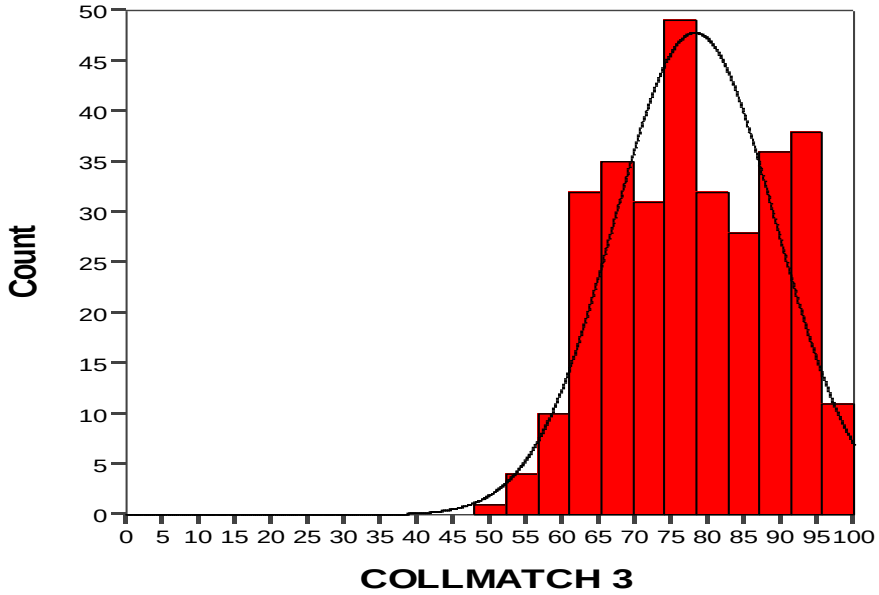

Figure 5.12 Frequency distribution of scores on COLLMATCH 3 (N = 307).

The mean Item Facility values for COLLEX and COLLMATCH were observed at .83, and .78, respectively. These values are fairly high, but the inclusion of the native speaker group

must be taken into account. A separate analysis of Swedish informant data only is made in 5.2.3.3 in an effort to try to estimate whether the changes made to the formats (see section 5.1) lead to the desired effects. Such an exclusion of native speaker data in the item analysis is warranted due to the fact that the tests are primarily aimed at Swedish students of English. The mean Item-total correlation values for COLLEX and COLLMATCH were observed at .34 and .26, respectively. The value for COLLEX is respectable, but the value for COLLMATCH is slightly lower than expected. A separate, follow-up analysis of intended target collocations and pseudo-collocations in COLLMATCH was made. This analysis showed that the former item category, consisting of 70 items, displayed a mean Item-total correlation value of .27, whereas the latter item category, consisting of 30 items, displayed a mean Item-total correlation value of .22. This means that the target collocations discriminated slightly more effectively between high-scoring and low-scoring informants, respectively, in terms of collocation recognition, than did the pseudo-collocation items.

5.2.3.3 Comparison between Swedish student groups and native speakers

5.2.3.3.1 Informant groups used in this subsection

In order to analyse the effectiveness of the new versions of COLLEX and COLLMATCH, and address the research questions, cross-sectional comparisons were carried out. Based on the original data presented in 5.2.3.2 above, data from a total of 269 informants were singled out for further analyses. The criterion for excluding data was the following. All informants who indicated that their L1 was not Swedish were removed. The rationale behind this was the wish to see how L1 Swedish students performed on the tests, in comparison with a designated group of native speakers. The excluded 38 informants had L1s like Finish, Polish, Bosnian, Bulgarian, Mandarin, Vietnamese, Arabic, Nepalese, Turkish and Italian. The remaining 269 informants were distributed as follows:

Table 5.10 Informant groups used in the cross-sectional analysis of the test data.

Informant group	Number of informants
SWE11: Swedish upper-secondary school students – 11th graders)	26
SWEuni1A: Swedish first-term university students of English – group 1	34
SWEuni1B: Swedish first-term university students of English – group 2	35
SWEuni1C: Swedish first-term university students of English – group 3	35
SWEuni1D: Swedish first-term university students of English – group 4	35
SWEuni2: Swedish second-term university students of English	39
SWEuni3: Swedish third-term university students of English	31
ENGuniNS: Native speakers of English at university level	34
Total	269

As can be seen in Table 5.10 above, the 139 Swedish first term students were divided into four subgroups, called SWEuni1A – SWEuni1D. This was done to facilitate subsequent inferential statistic analyses where equal or close to equal group sizes are preferable. The informants were randomly assigned to one of the four subgroups.

5.2.3.3.2 VLT M

The cross-sectional results on VLT M are shown in Table 5.11.

Table 5.11 Results on VLT M (k = 150) by groups.

Group		N	M	S.D.	Reliability α
SWE11:	(Swe 11th graders)	26	87.0	12.6	.86
SWEuni1A:	(Swe 1st term university)	34	127.7	14.4	.94
SWEuni1B:	(Swe 1st term university)	35	126.5	13.9	.93
SWEuni1C:	(Swe 1st term university)	35	128.0	14.4	.94
SWEuni1D:	(Swe 1st term university)	35	127.6	14.4	.94
SWEuni2:	(Swe 2nd term university)	39	132.0	10.9	.91
SWEuni3:	(Swe 3rd term university)	31	138.3	8.6	.89
ENGuniNS	(Eng Native speakers)	34	143.6	5.9	.86
Total		269	127.5	18.9	.96

The mean scores on VLT M increased with higher level of study. I terms of the Swedish informants, the 11th graders performed a mean score of 87.0, the four 1st term university student groups mean scores between 126.5 and 128, the 2nd term university group 132.0, and the 3rd term university student group 138.3. In comparison, the native speaker group scored a mean of 143.6. As can be seen in the table, these scores based on the modified data set also yielded reliable figures. The overall reliability for the groups combined was observed at .96, with values between .86 and .94 for the respective groups.

An analysis was carried out, in which the above reported mean scores were compared to one another. A Levene's test signalled significantly different variances in the data sets of the groups, wherefore a Welch's *F* test was used. The analysis yielded a significant group effect on the Vocabulary Levels Test (version M) scores, Welch F (7, 108.67) = 66.57, $p < .001$. A Games-Howell test showed the following statistically significant differences, shown in Table 5.12 below. As can be seen in Table 5.12, the mean score from the 11th graders' group (SWE11) was significantly different from all other group means. Furthermore, the native speaker group (ENGuniNS) mean differed from all groups except the third term university student group (SWEuni3). The mean of the latter group differed, in turn, from all Swedish student groups of a lower study level affiliation. What is particularly interesting with these results is the lack of statistical difference between the most advanced Swedish learners (SWEuni3) and the native speaker group (ENGuniNS). This begs the question whether these groups will differ in terms of scores on COLLEX and COLLMATCH.

Table 5.12 Differences between group means on VLT M.

	SWE11	SWEuni 1A	SWEuni 1B	SWEuni 1C	SWEuni 1D	SWEuni 2	SWEuni 3	ENGuni NS
SWE11		**	**	**	**	**	**	**
SWEuni1A	**		n.s.	n.s.	n.s.	n.s.	*	**
SWEuni1B	**	n.s.		n.s.	n.s.	n.s.	*	**
SWEuni1C	**	n.s.	n.s.		n.s.	n.s.	*	**
SWEuni1D	**	n.s.	n.s.	n.s.		n.s.	*	**
SWEuni2	**	n.s.	n.s.	n.s.	n.s.		n.s.	**
SWEuni3	**	*	*	*	*	n.s.		n.s.
ENGuniNS	**	**	**	**	**	**	n.s.	

* The mean difference is significant at $p < .05$.
** The mean difference is significant at $p < .001$.
n.s. = The mean difference is not significant.

One further analysis of the VLT M scores was carried out. The performance of the respective groups on the four frequency levels of the test was analysed. The results are shown in Table 5.13 below.

Table 5.13 Mean scores and standard deviations on VLT M word frequency levels by informant groups. The maximum score on each level is indicated.

Group	Level 3000 (k = 30)			Level Academic (k = 30)			Level 5000 (k = 45)			Level 10000 (k = 45)		
	M	S.d.	%	M	S.d.	%	M	S.d.	%	M	S.d.	%
SWE11	24.3	2.9	81	18.8	3.7	63	29.5	4.6	66	14.4	4.5	32
SWEuni1A	29.4	1.2	98	25.6	2.8	85	40.9	3.8	91	31.8	8.1	71
SWEuni1B	29.3	1.1	98	25.8	2.7	86	40.3	5.3	90	31.1	7.0	69
SWEuni1C	29.3	1.4	98	25.1	3.0	84	41.7	3.1	93	32.1	8.3	71
SWEuni1D	29.5	1.0	98	25.3	3.3	84	41.3	3.7	92	32.1	7.6	71
SWEuni2	29.6	0.7	99	26.5	2.7	88	42.6	2.2	95	33.4	7.0	74
SWEuni3	29.9	0.3	100	27.7	2.2	92	43.7	1.4	97	36.8	5.8	82
ENGuniNS	29.7	0.5	99	28.3	2.1	94	44.8	0.6	100	40.7	4.0	90

In the table, mean scores are presented for each level, together with the standard deviation (within parentheses) and the rounded percentage (within square brackets) of correctly answered items that the mean scores represent. Since the four levels consist of unequal numbers of items, the interesting figures in Table 5.13 are the percentages. As opposed to the analysis presented in Chapter 4 (Section 4.2.4.3.1.1, Table 4.17), the test part consisting of academic words was slightly more difficult than the 5000 word level, for all groups.

The performance of the most advanced Swedish student groups lies close to that of the native speakers up to and including the 5000 word level, but on the 10000 word level, there is a striking difference, with 74 and 82 per cent, respectively, for groups SWEuni2 and SWEuni3, and 90 per cent for ENGuniNS. It is thus clear that it is in this lower frequency band that the biggest difference can be found between advanced Swedish learners of English and native speakers of English in terms of vocabulary size. There is also a clear difference between the Swedish 11th grader group and the Swedish university learner groups. The gap between these groups becomes progressively bigger as a function of decreased word frequency.

5.2.3.3.3 COLLEX 5

The group-wise results on COLLEX 5 are shown in Table 5.14.

Table 5.14 Results on COLLEX 5 (k = 50) by groups.

Group		N	M	S.D.	Reliability α
SWE11:	(Swe 11th graders)	26	28.9	4.9	.65
SWEuni1A:	(Swe 1st term university)	34	41.3	6.0	.85
SWEuni1B:	(Swe 1st term university)	35	41.9	5.0	.82
SWEuni1C:	(Swe 1st term university)	35	41.3	5.7	.80
SWEuni1D:	(Swe 1st term university)	35	40.3	5.3	.80
SWEuni2:	(Swe 2nd term university)	39	42.5	4.3	.74
SWEuni3:	(Swe 3rd term university)	31	45.9	2.7	.58
ENGuniNS	(Eng Native speakers)	34	48.9	1.0	-.09
Total		269	41.7	6.8	.89

As can be seen in Table 5.14 there is a clear progression in scores. The lowest mean score was observed for the 11th graders, at 28.9. The four 1st term groups scored higher but slightly different means, ranging from 40.3 to 41.9. The 2nd term students scored a mean of 42.5, and the 3rd term students scored a mean of 45.9. The native speakers, finally, scored a mean of 48.9.

The overall reliability for the groups combined was observed at .89, as measured through Cronbach's alpha. The reliability coefficients for the different groups of Swedish informants varied between .58 and .85. For the native speaker group, a negative alpha value was observed.

In order to investigate the potential presence of a group effect, a Welch test was employed. The Welch test was used rather than an ANOVA since unequal group sizes existed. Also, unequal variance was observed across the groups. The Welch test signalled a significant effect of student group affiliation on test scores, Welch F (7, 102.38) = 96.64, $p < .001$. A Games-Howell post hoc test showed that differences between means were significant, except for the differences between any of the four 1st term university groups, and the 2nd term university group. The differences, in terms of levels of significance, are indicated in Table 5.15.

Table 5.15 Differences between group means on COLLEX 5.

	SWE11	SWEuni 1A	SWEuni 1B	SWEuni 1C	SWEuni 1D	SWEuni 2	SWEuni 3	ENGuni NS
SWE11		**	**	**	**	**	**	**
SWEuni1A	**		n.s.	n.s.	n.s.	n.s.	*	**
SWEuni1B	**	n.s.		n.s.	n.s.	n.s.	*	**
SWEuni1C	**	n.s.	n.s.		n.s.	n.s.	*	**
SWEuni1D	**	n.s.	n.s.	n.s.		n.s.	**	**
SWEuni2	**	n.s.	n.s.	n.s.	n.s.		*	**
SWEuni3	**	*	*	*	**	*		**
ENGuniNS	**	**	**	**	**	**	**	

* The mean difference is significant at $p < .05$.
** The mean difference is significant at $p < .001$.
n.s. = The mean difference is not significant.

In terms of mean Item Facility values for the 269 informants, this was observed at .83. Table 5.16 provides the mean Item Facility values for COLLEX 5. The values for the four first-term university students were collapsed into one in the table:

Table 5.16 Mean IF (Item Facility) values for items in COLLEX 5 by groups.

Group	SWE11 (N = 26)	SWEuni1 (N = 139)	SWEuni2 (N = 39)	SWEuni3 (N = 31)	ENGuniNS (N = 34)	All SWE groups combined (N = 235)
Mean IF	.58	.82	.85	.92	.98	.81

As could be expected, the item facility means increase by virtue of study level for the Swedish informants, and for the native speakers the value is close to maximum. The rightmost group in the figure consists of all the Swedish students combined. The very high IF value for the native speakers is positive from a validation point of view.

5.2.3.3.4 COLLMATCH 3

The results for COLLMATCH 3 by group are shown in Table 5.17. When it comes to group mean scores, the progression visible in the COLLEX scores (Table 5.14) is visible also in the COLLMATCH scores. In Table 5.17, mean scores increase across study levels, and native speakers score the highest mean:

native speakers of English > Swedish 3rd term university students > Swedish 2nd term university students > Swedish 1st term university students > Swedish 11th graders

Table 5.17 Results on COLLMATCH 3 (k = 100) by groups.

Group		N	M	S.D.	Reliability α
SWE11:	(Swe 11th graders)	26	63.0	6.4	.54
SWEuni1A:	(Swe 1st term university)	34	76.8	9.2	.81
SWEuni1B:	(Swe 1st term university)	35	77.9	8.5	.74
SWEuni1C:	(Swe 1st term university)	35	76.2	9.5	.86
SWEuni1D:	(Swe 1st term university)	35	75.1	9.6	.88
SWEuni2:	(Swe 2nd term university)	39	79.4	8.0	.81
SWEuni3:	(Swe 3rd term university)	31	85.2	6.9	.80
ENGuniNS	(Eng Native speakers)	34	92.9	3.3	.52
Total		269	78.7	10.9	.89

In terms of reliability, a respectable internal consistency coefficient of .89 was observed for the COLLMATCH 3 scores. Values on the lower end were observed for the upper-secondary school students, and the native speakers (.54 and .52, respectively). For the Swedish university student groups, Cronbach's alpha values ranged between .74 and .88.

A comparison of the eight group means revealed that a group effect existed. A Welch *F* test indicated significant differences between means, Welch F (7, 107.98) = 86.72, $p < .001$. In order to find out where these differences lay, a post hoc Games-Howell test was conducted. The exact same pattern as was found for the COLLEX means was found also for the COLLMATCH means. All means were different from each other except any of the four 1st term student means, and the 2nd term student mean. The results from the post hoc test are shown in Table 5.18 below.

Table 5.18 Differences between group means on COLLMATCH 3.

	SWE11	SWEuni 1A	SWEuni 1B	SWEuni 1C	SWEuni 1D	SWEuni 2	SWEuni 3	ENGuni NS
SWE11		**	**	**	**	**	**	**
SWEuni1A	**		n.s.	n.s.	n.s.	n.s.	*	**
SWEuni1B	**	n.s.		n.s.	n.s.	n.s.	*	**
SWEuni1C	**	n.s.	n.s.		n.s.	n.s.	*	**
SWEuni1D	**	n.s.	n.s.	n.s.		n.s.	**	**
SWEuni2	**	n.s.	n.s.	n.s.	n.s.		*	**
SWEuni3	**	*	*	**	**	*		**
ENGuniNS	**	**	**	**	**	**	**	

* The mean difference is significant at $p < .05$.
** The mean difference is significant at $p < .001$.
n.s. = The mean difference is not significant.

The only difference between Tables 5.15 and 5.18 is the significance level for the difference between groups SWEuni3 (third-term students) and SWEuni1C (first-term students, group C ($p < .001$. in Table 5.18). When it comes to Item Facility values, these were computed for the 269 informants and observed at .78. In an analysis of the different groups, the results arrived at are shown in Table 5.19. The values for the first-term university students were collapsed into one.

Table 5.19 Mean IF (Item Facility) values for items in COLLMATCH 3 by cross-sectional groups.

Group	SWE11 (N = 26)	SWEuni1 (N = 139)	SWEuni2 (N = 39)	SWEuni3 (N = 31)	ENGuniNS (N = 34)	All SWE groups combined (N = 235)
Mean IF	.63	.76	.79	.85	.93	.77

As could be predicted based on the mean scores presented in Table 5.17 above, the item facility means increase by virtue of study level for the Swedish students. The rightmost group in the figure consists of all the Swedish students combined. Again, the highest IF value was observed for the native speaker group (ENGuniNS). This lends validation support to COLLMATCH 3.

5.2.3.3.5 Correlation analyses

In order to address research question 4, whether there is a relation between vocabulary size and scores on COLLEX 5 and COLLMATCH 3, a number of correlation analyses were carried out.

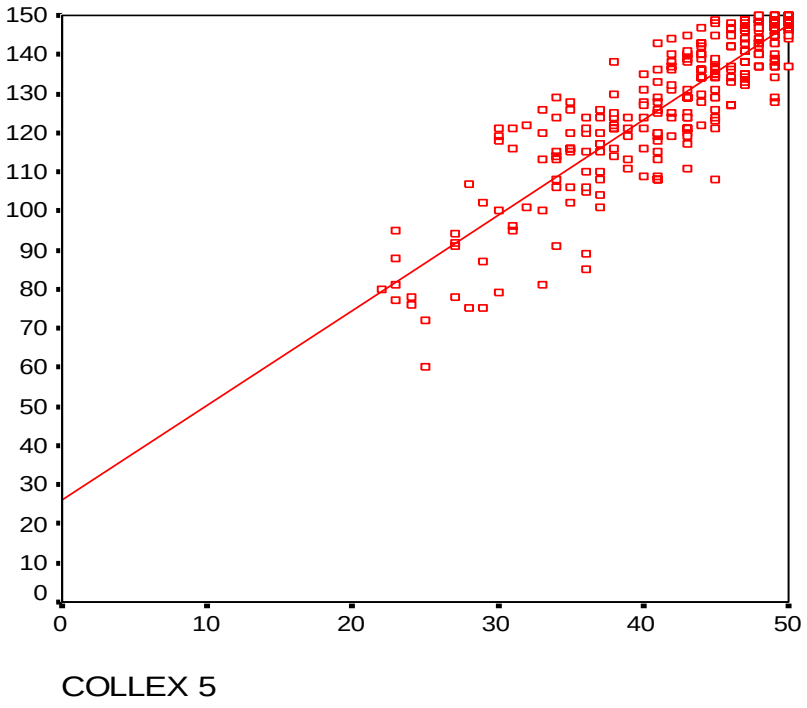

Figure 5.13 Scatterplot of VLT M scores against COLLEX 5 scores (N = 269).

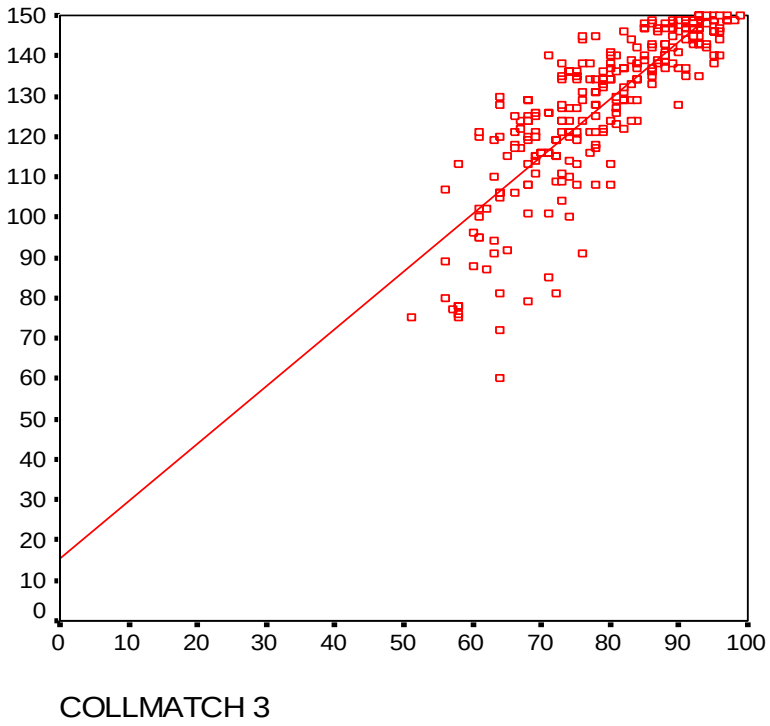

Figure 5.14 Scatterplot of VLT M scores against COLLMATCH 3 scores (N = 269).

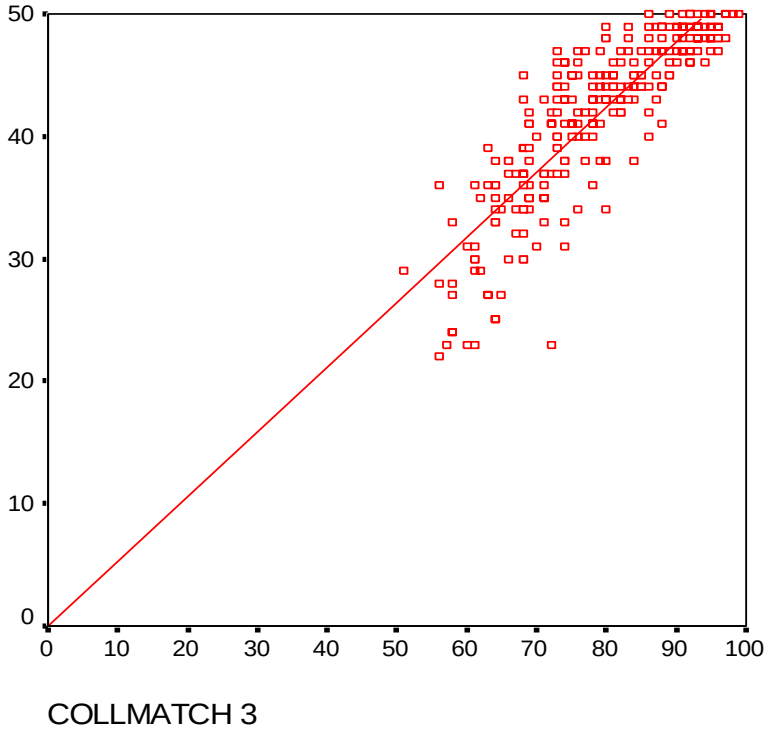

Figure 5.15 Scatterplot of COLLEX 5 scores against COLLMATCH 3 scores (N = 269).

As a first step, scatterplots were retrieved for the relations between the three variables: VLT M scores, COLLEX 5 scores, and COLLMATCH scores. The scatterplots are shown above as Figures 5.13, 5.14, and 5.15, and are based on all groups combined (N = 269). The three scatterplots foreshadow high positive correlations between the variables at hand. They also clearly show the negative skewness of the scores.

As a second step, a Pearson Product Moment test was used in order to arrive at correlation coefficients. The correlations are shown in Table 5.20 below.

Table 5.20 Correlations (Pearson *r*) between scores on VLT M, COLLEX 5 and COLLMATCH 3 (N = 269)

Test	VLT M	COLLEX 5	COLLMATCH 3
VLT M	-	.88**	.83**
COLLEX 5		-	.86**

** Correlation is significant at p < .01, one-tailed.

As was predicted through the scatterplot visualizations, the test signalled high, positive correlations between all three variables. In an attempt to ascertain whether correlations varied according to student group affiliation, separate correlation analyses were carried out for each group. In the first analysis, the scores on the vocabulary size test were used as the predictor value. The results of this analysis are shown in Table 5.21.

Table 5.21 Groupwise correlations (Pearson *r*) between scores on VLT M, COLLEX 5 and COLLMATCH 3.

Group	N	VLT M against	COLLEX 5	COLLMATCH 3
SWE11	26		.59**	.41*
SWEuni1A	34		.78**	.82**
SWEuni1B	35		.79**	.66**
SWEuni1C	35		.86**	.86**
SWEuni1D	35		.76**	.81**
SWEuni2	39		.67**	.75**
SWEuni3	31		.69**	.75**
ENGuniNS	34		.43**	.57**

** The correlation is significant at p < .01, one-tailed.
* The correlation is significant at p < .05, one-tailed

Correlations between VLT M and COLLEX 5 were higher for the first term university student groups (SWEuni1A-SWEuni1D), than for any of the other informant groups. The same trend was observed for COLLMATCH 3, with the exception of group SWEuni1B. The lowest correlations were observed for the upper-secondary school student group (SWE11), and the native speaker group (ENGuniNS). A possible reason for this could be the fact that the estimated reliability of the scores of these two groups was on the low end.

In the second analysis, COLLEX and COLLMATCH scores were correlated for each group. The results are shown in Table 5.22.

Table 5.22 Groupwise correlations (Pearson *r*) between scores on COLLEX 5 and COLLMATCH 3.

Group	N	COLLEX 5	against	COLLMATCH 3
SWE11	26			.52**
SWEuni1A	34			.87**
SWEuni1B	35			.64**
SWEuni1C	35			.82**
SWEuni1D	35			.84**
SWEuni2	39			.77**
SWEuni3	31			.67**
ENGuniNS	34			.22

** The correlation is significant at p < .01, one-tailed.

As can be seen in Table 5.22, all correlations were significant, except for the native speaker group. For three of the first term university student groups correlations were observed at .82 - .87. One of the first term university student groups reached a relatively lower correlation, at .64, moreover the same group for which a lower correlation was observed between VLT M scores and COLLMATCH scores. The correlation for the upper-secondary school student group (SWE11) was observed at .52, and for the 3rd term university student group .67. The correlation for the native speaker group (ENGuniNS) was .22.

5.2.3.3.6 New group divisions based on vocabulary size scores

In a similar vein to the analysis carried out in study 4 in Chapter 4, I created three new groups based on the vocabulary size scores (VLT M). This was done in order to address research question 5, which asked whether there is a relation between general proficiency in English and scores on COLLEX 5 and COLLMATCH 3. The rationale behind the method of analysis was the assumed correspondence between scores on a vocabulary size test and general proficiency in a language.

Based on the data from the 269 informants used in the previous analysis, I first removed the native speaker data (N = 34) since I was primarily interested in the performance of Swedish L2 learners of English. This gave me a data set of 235 Swedish informants. In order to be able to divide this group into three groups of equal size, I eliminated the data of one informant from the analyses. I randomly removed one of the informants with the highest vocabulary size score (150). This gave me three groups of 78 informants in each. I called these groups LOW, MID, and HIGH. The mean and standard deviation VLT M scores for each of the groups are displayed in Table 5.23 below. The next step was to check homogeneity of variance of the three groups, as a preparation for an ANOVA. A Levene's test showed that the variances for the groups were dissimilar, and therefore a Welch *F* test was used. This test signalled a significant effect for group in all three test means. The VLT M score means were significantly different from each other, Welch F (2, 140.06) = 398.87, p < .001, and post hoc Games-Howell.

Table 5.23 Means and standard deviations for VLT M scores for three groups.

Group	N	M	S.d.
LOW	78	103.7**	15.1
MID	78	127.9**	4.9
HIGH	78	143.5**	4.3

** The mean is significantly different from other means, at p < .001.

Similar analyses were carried on the COLLEX and COLLMATCH data. The result for COLLEX 5 was Welch F (2, 138.92) = 152.78, p < .001, and the result for COLLMATCH 3 was Welch F (2, 153.78) = 193.24, p < .001. The means on these two tests are shown in Table 5.24. A Games-Howell post-hoc test indicated significant differences between all the three group means on both COLLEX 5 and COLLMATCH 3, as seen in the table.

Table 5.24 Means, standard deviations, and statistical significance based on COLLEX 5 and COLLMATCH 3 scores for three groups.

		COLLEX 5		COLLMATCH 5	
Group	N	Mean	S.d.	Mean	S.d.
LOW	78	34.1**	5.8	67.2**	6.5
MID	78	41.5**	4.1	75.8**	5.9
HIGH	78	46.2**	2.5	86.8**	6.0

** The mean is significantly different from other group means in the same test, at p < .001.

In sum, based on the assumed convergence between vocabulary size scores and general proficiency in a language, students with higher proficiency in this study score significantly better on COLLEX and COLLMATCH than do students with lower proficiency.

5.2.4 Discussion

The study reported in this section (5.2) aimed at investigating whether the amended COLLEX 5 and COLLMATCH 3 formats that were piloted in a small-scale study in section 5.1 worked well psychometrically also in a large-scale study. In the following discussion section, I will structure my discussion around the research questions stated in section 5.2.2.5.

5.2.4.1 Do COLLEX 5 and COLLMATCH 3 produce reliable test scores in terms of internal consistency, and do the test items have a satisfactory discriminatory power in terms of item facility and item-total correlations values?

5.2.4.1.1 COLLEX 5

The question whether COLLEX 5 is capable of producing reliable test scores can be answered affirmatively based on the data gathered in the present study. Using Cronbach's alpha, the reliability coefficient was observed at .89 for the initial analysis including data from a total of 307 informants, both upper-secondary school students and university students in Sweden, as well as native speakers of English pursuing university level studies in Wales. In a subsequent analysis, in which informants with other L1s than Swedish were excluded, a reliability value based on 269 informants was observed at the same level: .89. A reliability coefficient of this magnitude is clearly satisfactory.

In retrieving coefficients for the scores by different student groups on COLLEX 5, however, lower values were observed (see Table 5.14). In particular, lower values were observed for Swedish third term university students (Cronbach's alpha = .58), and university level native speakers of English (Cronbach's alpha = -.09). A possible reason for arriving at lower values for subgroups in a test is restrictions in test score range. If a subgroup is homogeneous in terms of the ability measured in the test, the variance in the scores produced by that subgroup will be small, and as a consequence the reliability coefficient will be low. Support for viewing an informant group as homogeneous can be found in a low standard deviation. A closer examination of the Item Facility and Item-total correlation values will also render potential support. For the Swedish third-term university student group, the standard deviation was 2.7, and for the university level native speakers of English it was 1.0. These are both very small. The Item Facility for the Swedish third-term learner group was .92 and as many as 22 of the 50 items displayed an Item-total correlation of .00, due to zero variance in the group scores of those items. For the native speakers, the Item Facility was .98, and the number of items with zero variance was 31. Thus, these observations can be taken as causing the lower reliability values.

The negative reliability value for the native speaker group is an undesired result, but there is a feasible explanation. With a mean score of 48.9 and a standard deviation of 1.0, it is clear that the native speakers performed uniformly high scores. The negative alpha value is likely to stem from the striking lack of variance in the majority of items, together with the inability of the items producing variance to discriminate between informants with higher total scores and informants from lower total scores. In conclusion, considering the conditions accounted for above, the negative value is not as serious as it might appear at first sight. Moreover, the primary function of the native speaker data in the present study is validation. This aspect will be addressed in section 5.2.4.2 below.

The mean Item-total correlation value for COLLEX 5 was observed at .34, based on the 307 informants that took the tests. This is a wholly satisfactory level (see Ebel 1979), and it suggests that the items in COLLEX 5 discriminate well between students with high and low total scores. One item displayed the value of .00 (item 22: Target collocation: *keep a secret*). The value stems from zero variance, caused by the fact that all 307 informants answered this item correctly. Such an easy item gives no discriminatory information and should probably be discarded. However, the inclusion of an apparently easy item holds an advantage. It may

serve the purpose of providing confidence to lower ability informants. If these informants get the feeling that all items are very difficult, then they may lose interest, which would yield unreliable scores as a result. Therefore, there is a place for even very easy items in a test.

5.2.4.1.2 COLLMATCH 3

The internal consistency of COLLMATCH 3 scores, as measured through Cronbach's alpha, was also satisfactory. Based on 307 informants it was observed at .89, and when excluding students in Sweden with other L1s than Swedish, it remained at .89 (N = 269). Thus, the 100-item COLLMATCH test produced the same high reliability in scores as did the 50-item COLLEX. This raises the question of whether COLLMATCH is too long, or, more to the point, whether the same high reliability can be reached with fewer items. An interesting aspect relevant to this claim is the potential difference in reliability between the target collocations, and the pseudo-collocations. An analysis based on the results for the different groups (section 5.2.3.3.4) showed that the scores on the 70 target collocations generated a reliability of .88, whereas the scores on the 30 pseudo-collocations generated a reliability of .78. Thus, it seems to be possible to reach a reliability of around .90 based only on real collocations. The inclusion of pseudo-collocations affects the overall reliability marginally. Bringing matters to a head, an implication that follows from this observation is that the pseudo-collocations might in fact be a redundant feature of COLLMATCH 3. Interestingly, in an evaluation of the effectiveness of vocabulary yes/no tests, Shillaw (1999) even found that the reliability was higher for only real words (Cronbach's alpha .81) than for real words and non-words combined (Cronbach's alpha .61 and .73) in two versions of a test containing 80 real words and 20 non-words. This was not the case in the present study, however, where no tangible negative effect was found when it comes to the inclusion of pseudo-collocations. In fact, the removal of the pseudo-collocations could possibly lead to other undesired effects. One purpose of the inclusion of distractors in a yes/no test is to prevent informants from ticking all the items uncritically, and thereby receiving the maximum score. Dropping the distractors all together could therefore paradoxically lead to scores that do not reflect the true ability of the informants. Another reason for keeping the distractors in the COLLMATCH format is the fact that they are seen as part of the measured construct itself. The test measures both the ability to recognize conventionalized target collocations, and to reject pseudo-collocations. It therefore seems logical to keep the pseudo-collocations as part of the format as long as they contribute to test score information.

The question still remains, though, whether a shorter test could produce the same high level of reliability as the 100-item version. Why use a longer test, when a shorter one can do the trick? In order to test this, two analyses were carried out. In one of them, the original test was divided into two parts, with 50 items in each. In each 50-item part, there were 35 target collocations and 15 pseudo-collocations. In a second analysis, the pseudo-collocations were discarded, resulting in two test parts of 35 target collocations each. When checking the reliability coefficients for these four versions, the results shown in Table 5.25 were obtained. As can be seen in the table, the 50-item versions (1A and 2A) are slightly more reliable than the 35-item versions (1B and 2B), and the inclusion of 15 pseudo-collocations increases reliability by a couple of percentage points. The analysis shows that the pseudo-collocations do provide psychometric information, and that the 100-item COLLMATCH 3 could potentially be divided into two 50-item test versions, where a hypothetical overall reliability value of around .80 seems to be within reach for these two versions.

Table 5.25 Reliability coefficients for different versions of COLLMATCH 3 based on N = 269.

Test version	Reliability (Cronbach's alpha)
50-item COLLMATCH version 1A (35 real collocations + 15 pseudo-collocations)	.79
50-item COLLMATCH version 2A (35 real collocations + 15 pseudo-collocations)	.81
35-item COLLMATCH version 1B (35 real collocations)	.77
35-item COLLMATCH version 2B (35 real collocations)	.80

This level of reliability would be acceptable but perhaps somewhat lower than desired for a 50-item test of receptive collocation knowledge. It would of course be possible to experiment further with versions consisting of 90, 80, and 70 items, etc, but for the moment, going back to the question of test length, there is evidence to suggest that the 100-item version should be kept due to its capacity to produce reliability values of around .90.

In terms of reliability values for the different subgroups on COLLMATCH, these were better than those on COLLEX for the Swedish third-term university student group, as well as the native speaker group, with Cronbach's alpha observed at .80 and .52, respectively. Thus the negative value for the native speakers on COLLEX was not present in the COLLMATCH scores. The reason for the slightly lower values was again believed to be the homogeneity of the subgroups, indicated by the standard deviation of 3.3 for the native speakers. The reliability of .54 observed for the upper-secondary school students could possibly stem from blind guessing, but a specific study of guessing behaviour of these students is required to bear this speculation out, and such data are not available.

5.2.4.2 Can COLLEX 5 and COLLMATCH 3 be argued to produce valid scores as tests of receptive recognition knowledge of English collocations?

One piece of support for the validity (see section 2.5 for an account of different kinds of validity) of COLLEX 5 and COLLMATCH 3 can in fact be found in the reliability results discussed in the previous section. As was discussed in Chapter 3, Weir (2005) and Alderson (1991) see reliability as a type of validity evidence. Weir proposes reliability to be subsumed under the cover term "scoring validity" (p. 22). Following this view, by observing high values for reliability, a necessary but not sufficient condition for validity has been established.

Another piece of support for the validity of COLLEX 5 and COLLMATCH 3 can be found by seeing the groups in section 5.2.3.3 as differing in the ability being assessed in the tests. This is what Bachman (2004:290) calls "a non-equivalent groups design". It is based on the division of informants into different *a priori* ability groups. In the present study, I formed five such groups, with one group further subdivided into four subgroups. The formation of the groups was carried out with level of study as the criterion for Swedish students of English, and for the native speakers of English by virtue of being native speakers of English, and students at university level in Britain. The native speakers were hypothesized to have the highest ability in the measured construct, followed in turn by the highest level Swedish students. The overall hypothesized differences in mean scores can be summarized as follows (see Table 5.10 for explanations of group abbreviations):

$\bar{X}$ ENGuniNS > $\bar{X}$ SWEuni3 > $\bar{X}$ SWEuni2 > $\bar{X}$ SWEuni1 > $\bar{X}$ SWE11

The above differences in means were not observed to the letter, for no statistical differences were observed between the Swedish second term university students (SWEuni2) and the Swedish first term university students (SWEuni1) in neither COLLEX nor COLLMATCH. Except for this anomaly, the differences in means between the groups can be taken as evidence of test validity. I have been able to demonstrate clear differences between native speakers of English, Swedish university students of English, and Swedish upper-secondary school students when it comes to their receptive recognition knowledge of English verb + NP collocations.

The question is why no statistically significant difference was found between university students in their first term of study and university students in their second term of study. There are two competing explanations relevant to this observation. One of the explanations holds that no difference exists between these two groups. Since the groups were formed *a priori* based on formal level of study, it is quite possible that there is a mismatch between the ability that was the basis for the creation of the groups, and the ability that is intended to be measured in the two tests. A second-term university student is expected to be more skilled in English than a first term university student, but in reality that might not be the case. An indication of this can be found in the scores on the Vocabulary Levels Test (see Table 5.11). The means for the first term university student groups ranged between 126.5 and 128.0. The mean for the second term university student group was 132.0. Thus, there were differences in the mean scores, but these differences were not significant, and this is predictable if we look at the standard deviations. The first-term university student groups ranged between 13.9 and 14.4, whereas the value for the second-term university student group was 10.9. These scores clearly overlap and cannot be seen as scores coming from different populations.

A second explanation for the absence of statistical significance holds that neither COLLEX 5 nor COLLMATCH 3 are sensitive enough as test tools to pick up any existing difference. It is at the same time as difficult to reject this explanation as it is to confirm it. If we take the scores on the Vocabulary Levels Test as indicative of general proficiency in English, and if we also assume a correlation between general proficiency in English and receptive collocation knowledge, then no difference between the two groups should be expected. Therefore, the first explanation can be seen to be empirically supported, to some extent, if the premises are accepted, whereas the second is impossible to falsify on the basis of the existing data gathered for this study. I will therefore pursue the explanation that holds that it is not beyond reasonable doubt that Swedish university students in the first term and students in the second term in this study come from the same underlying population.

5.2.4.3 What is the relation between vocabulary size and scores on COLLEX 5 and COLLMATCH 3, and does this relation vary according to study level affiliation?

The analysis of the three test variables in the study showed that they were positively related. The scores on the Vocabulary Levels Test were observed to correlate positively both with scores on COLLEX 5 (r = .88) and COLLMATCH 3 (r = .83). Furthermore, the two

collocation tests correlated at the same high level with each other (r = .86). The results lie in the same region as those obtained in Chapter 4 (section 4.2) for correlations between vocabulary size scores and COLLEX 4 and COLLMATCH 2 (both r = .87). These present results thus corroborate earlier results, and they seem to suggest that scores on the collocation tests vary as a function of variation in vocabulary size scores. In other words, the larger vocabularies the informants had, as measured through the receptive Vocabulary Levels Test, the higher were their scores on the two receptive collocation tests. In the development of the COLLEX and COLLMATCH versions used in the present study, efforts were made to minimize the impact of vocabulary size on collocation test scores by using high frequency words for inclusion in the test items. As can be seen in Appendices 5J and 5L, a large majority of the words in COLLEX 5 and COLLMATCH 3 were taken from the first two thousand words of English according to JACET 8000 (Ishikawa *et al.* 2003). However, a small number of words were taken from lower frequency bands. It therefore cannot be explicitly ruled out that weaker students experienced problems with certain lower frequency words, and that this in turn affected their ability to recognize collocations in which these single words featured.

The answer to the follow-up question whether the relation between vocabulary size scores and COLLEX and COLLMATCH scores vary according to study level affiliation is affirmative, but the possible reasons behind the observed variation are less straightforwardly explained. Table 5.21 indicated that the highest correlations were found for the first term university student groups in terms of COLLEX 5, and for three of these four groups the highest correlations were also observed for COLLMATCH 3. The correlations were generally lower for the upper-secondary school student group, and for the second and third term students, as well as the native speakers. One possible explanation for the lower correlations is the lower reliability values observed for the same groups. Thus, where scores were not reliably measured, correlations based on those scores were low.

5.2.4.4 Is there a relation between general proficiency in English and scores on COLLEX 5 and COLLMATCH 3?

The answer to the fourth and final research question is conditioned by the acceptance of an assumption. The assumption holds that vocabulary size scores are indicative of general proficiency in a language. If the assumption is accepted, and there is empirical support for its existence (see references in section 4.2), then the results obtained in this study support the view that there is a relation between general proficiency in English and scores on COLLEX 5 and COLLMATCH 3. The really interesting question is if it is possible to make an inference from these observed test results to universe scores. It would be surprising if the results arrived at in this study are not indicative of abilities beyond the test contents, i.e. performance on tasks similar to those in the tests used here. One way to test this assumption is to carry out a concurrent validity study (see section 2.5) in which the same tests used here are administered together with other measures of the same abilities. By correlating, for example, a collocation test like COLLMATCH with another test also purported to be a test of collocation knowledge, we can find evidence either for a rejection of the inference, or an acceptance. In addition to this type of criterion validation, a logical analysis of the claimed test constructs of the compared measures must be carried out, and ideally, counterhypotheses should also be formed. This step implies incorporating in the test battery administered also a measure of an

ability which is not believed to correlate at a high level, or less strongly, with the measures under investigation.

A study such as that sketched above could potentially shed more light on the explanations for the high correlations between vocabulary size and receptive collocation knowledge, as measured through COLLEX and COLLMATCH. It seems there is a tangible relation between the vocabulary size construct and the collocation knowledge construct arguably measured in the two mentioned tests. An interesting comparison would include also a vocabulary depth measure. By comparing scores on vocabulary size and vocabulary depth measures with scores on COLLEX and COLLMATCH, it would be possible to see if the two collocation tests reside closer to size than depth scores, or vice versa.

5.2.5 Summary and conclusions

This study has provided further evidence for the interpretation of COLLEX 5 and COLLMATCH 3 test scores as reliable and valid indicators of receptive recognition knowledge of English collocations. The study also showed that scores on the two tests vary as a function of vocabulary size, and if vocabulary size is accepted as an indicator of general English proficiency, analyses carried out in the study corroborated a positive relation between receptive recognition knowledge of English collocations, and general proficiency. Finally, it was concluded that a study of concurrent validity should be carried out, in order to find out if test scores can be generalised beyond the actual tests, and whether the constructs arguably tested in COLLEX and COLLMATCH gravitate more towards the dimension of vocabulary size or the dimension of vocabulary depth. Such a study will be reported next, in Chapter 6.

6 Validating COLLEX 5 and COLLMATCH 3 against other vocabulary and proficiency tests

6.1 Introduction

The purpose of the present chapter is to report on a validation study, in which COLLEX 5 and COLLMATCH 3 were administered together with a vocabulary size test, a vocabulary depth test, and a reading comprehension test in order to observe potential concurrent validity with these tests. The results of the study are reported, and certain problems pertaining to a conceptual distinction between vocabulary size and depth are discussed, as is the role of the Word Associates Test as a proper depth test.

In section 5.2 of the previous chapter, the results of a large-scale study were reported. The results were on the whole promising. However, an observed high correlation between COLLEX and COLLMATCH on the one hand, and a measure of vocabulary size on the other, raised the question of what ability, or construct, rather, COLLEX and COLLMATCH are measuring. One interpretation of the high correlation would be that a measure of vocabulary size (such as the VLT), and COLLEX and COLLMATCH are measures of different aspects of the same underlying linguistic knowledge, which could be seen as 'lexical knowledge'. Another interpretation, however, would be that there is a causal relationship between the measures. The hypothesis would then be that vocabulary size determines the performance on COLLEX and COLLMATCH in that it takes a large vocabulary to recognize conventionalized collocations. The assumption behind this line of thinking is that a large vocabulary is built up incrementally through exposure to the target language. In this process, links between single words are believed to be forged in the mental lexicon, which in turn makes language users more collocationally skilled. The result, then, is that as more L2 words are learnt through exposure, knowledge about how words may be combined is also acquired.

If we accept this hypothesis, about the reason behind the high correlation between COLLEX and COLLMATCH, and vocabulary size, then a related question is how COLLEX and COLLMATCH relate to the concept of vocabulary depth. Will equally high correlations (r = ~.90) be observed, and if not, can we then claim that COLLEX and COLLMATCH are more size tests than depth tests? The question of what a test is measuring is at the heart of validity in general, and construct validity in particular. Consequently, a study was set up aimed at empirically investigating the relationships between COLLEX, COLLMATCH, a vocabulary size test, a vocabulary depth test, and also a fifth variable argued to measure something different. Concurrent validity support will be established if COLLEX and COLLMATCH correlate at a high level with another test also argued to measure the same, or similar, construct. Conversely, we expect there to be no, or a much lower correlation between COLLEX and COLLMATCH, and a variable which is argued to measure something different from these two tests, for example a grammar test or a reading comprehension test.

6.2 Methods

6.2.1 Considerations for the study design

In section 2.4 of Chapter 2, we saw that one assumption evident in recent developments in the field of L2 vocabulary testing is that lexical knowledge is made up of somewhat different, co-existing dimensions. The two most dominating and widely assumed dimensions are vocabulary size[33] and vocabulary depth (see e.g. Anderson & Freebody 1981; Wesche & Paribakht 1996; Qian 1999; Schmitt 2000; Greidanus *et al.* 2004). Vocabulary size denotes the number of words for which a basic meaning[34] is known by an individual, whereas vocabulary depth commonly denotes a more comprehensive knowledge beyond a basic meaning, entailing, for example, knowledge of multiple meanings of words (polysemy), grammatical functions, and common collocations. In fact, seeing collocation knowledge as part of vocabulary depth implies that COLLEX and COLLMATCH are something akin to depth tests. This is not a claim that I have pursued so far in this thesis, and no empirical support for this view has yet been gathered, but it certainly merits attention. I have pursued the argument, though, that receptive collocation knowledge as a single construct is likely to presuppose several other subcomponents of Nation's (2001) word knowledge framework (see section 2.4.2). However, as we have seen in the studies reported in Chapters 4 and 5, strong, or even very strong correlations were observed between vocabulary size scores, as measured through the Vocabulary Levels Test, and scores on COLLEX and COLLMATCH. This raises an interesting question. On the face of things, if we assume that vocabulary size and vocabulary depth are different dimensions of lexical knowledge, do COLLEX and COLLMATCH possibly gravitate more towards one of these dimensions than the other? By carrying out a validation study in which COLLEX and COLLMATCH are administered together with a vocabulary size measure and a vocabulary depth measure, it would be possible to find empirical support for the answer to this question. Correlations between the scores on the different variables would show whether COLLEX and COLLMATCH are more closely related to depth tests, to size tests, or if they measure something slightly different.

A number of studies have empirically compared vocabulary size with vocabulary depth. Vermeer (2001), testing 50 L1 and L2 Dutch kindergarten 5-year-olds, arrived at correlations ranging between .70 and .83 between a receptive vocabulary size measure, and an association task depth measure. Qian (1999) used the VLT as a size measure and found correlations between scores on that test with scores on the Word Associates Test (WAT), at .82, based on data from 74 L1 Korean and L1 Chinese ESL college and university students, predominately 18-27 year-olds. Nurweni and Read (1999), when administering a receptive vocabulary size measure and a WAT format depth measure to 350 L1 Indonesian ESL first-year university students, observed a correlation of .62 for the whole group, and in a subsequent analysis, in which the 350 students were subdivided according to scores on a general proficiency exam, .81 for high level students (10%); .43 for mid level students (42%); and .18 for low level students (48%). Thus, in previous studies, barring the low level student component of the

[33] Vocabulary size is often referred to as vocabulary breadth. These two terms are used interchangeably in the literature. I will in this chapter use the term vocabulary size.

[34] By basic meaning is meant the sense given in a dictionary as the most frequent and common one. An alternative to basic meaning is 'core meaning', which according to Sinclair is the one that first comes to mind of most people (1991:113).

latter study, high correlations between size and depth measures have been observed. Consequently, similarly high correlations can be expected in the present study.

From this follows also that based on previously observed high correlations between COLLEX and COLLMATCH scores and vocabulary size scores, high correlations can be expected between the two collocation test scores and a vocabulary depth test score. However, in addition to measures of size and depth and the two collocation measures, by incorporating a fifth measure, it would be possible to see if a certain variable contributes more to the variance in this additional measure than other variables. For example, if we hypothetically arrive at high inter-correlations between COLLEX, COLLMATCH, a vocabulary size measure, and a vocabulary depth measure, it could be that these four variables account to varying degrees for the variance in the additional variable. Their existence as separate constructs could then be justified. Bachman (2004:279) argues that "…if we want to support a claim that a particular test measures a particular area of knowledge or ability and not another, we need to administer many different tests that we claim measure different abilities". Such an approach is clearly linked to construct validity, i.e. the question of what skill or knowledge a test is testing.

Many alternatives as to what this fifth additional ability variable should be presented themselves. I ultimately decided to add reading comprehension as this additional ability for the following reasons. Firstly, from a practical point of view, it would be easy to administer a reading comprehension test as part of a test battery, as opposed to, for example, a speaking or writing test. With an objectively marked, standardized reading test, complexities involved in the administration and rating of a spoken test component could be avoided. Secondly, a test of writing skills would be considerably more difficult to assess in a straightforward way. Thus, we have five variables to administer in the validation study. The conceptual outline of such a study and the assumed relation between the inherent variables are illustrated in Figure 6.1 below. In the figure, the circles represent the five variables, with two assumed vocabulary dimensions (size and depth), the receptive collocation construct arguably measured in COLLEX and COLLMATCH, and lastly the reading comprehension construct. The unbroken arrows indicate an empirically established relationship through analyses in Chapters 4 and 5, whereas the broken arrows indicate a yet unknown relationship. As to the potential informants of the study, I needed data from the same level of students that were tested in the previous studies. Since the aim of the project reported in this thesis is to develop collocation tests for use with advanced students of English, at upper-secondary school and university levels, it seemed logical to use these types of students also for the present validation study. My aim was to gather data from a fairly heterogeneous informant group, in terms of general English proficiency. This was because COLLEX and COLLMATCH were constructed for use with both upper-secondary and university level students in mind. Furthermore, it would be interesting to see if students of differing general proficiencies would produce scores in certain ranges in the five variables. In order to get enough data for statistical tests, I estimated that I needed a minimum of 20-30 informants.

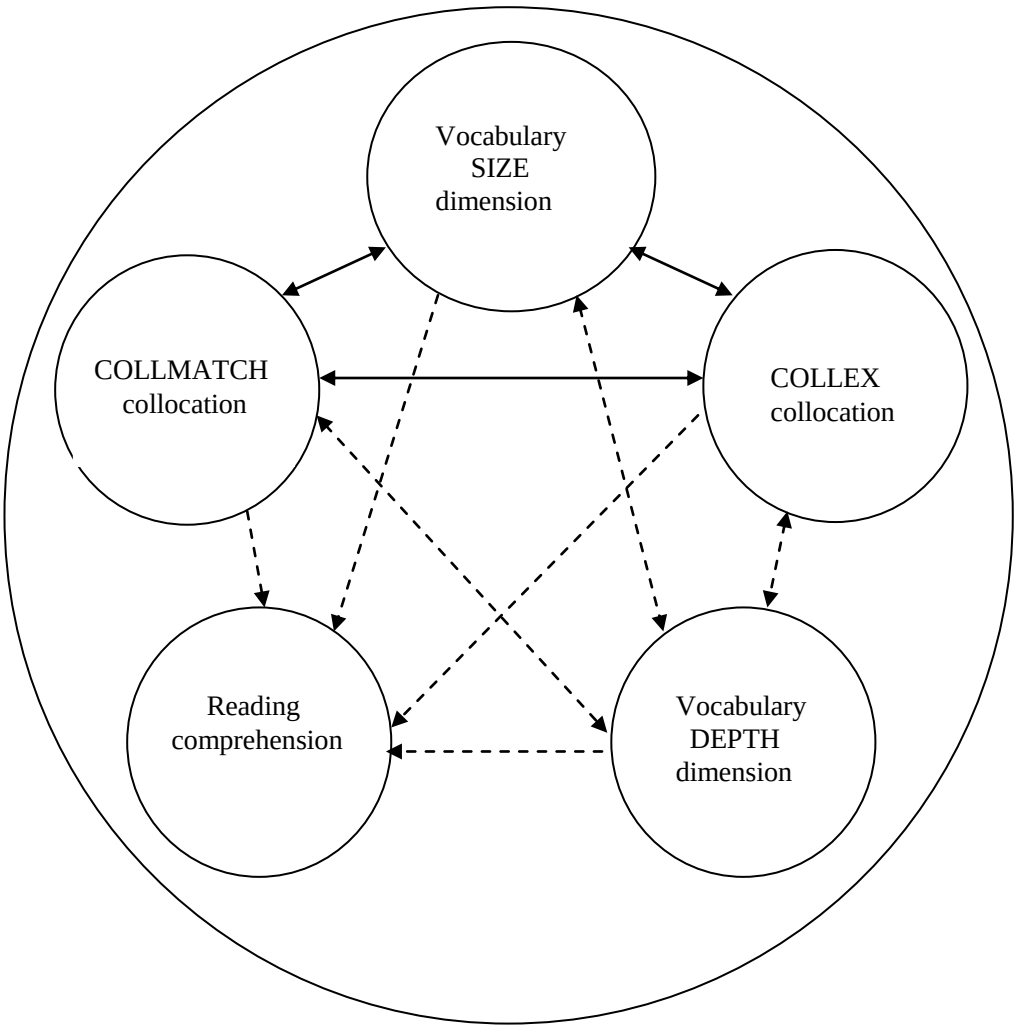

Figure 6.1 Empirically observed and hitherto unestablished relations between assumed vocabulary dimensions and test variables.

6.2.2 Material

Based on the rationale presented in the previous section, a test battery was created consisting of five parts, shown below.

a) COLLEX 5, 50 items (Appendix 5I);
b) COLLMATCH 3, 100 items (Appendix 5K);
c) Vocabulary Levels Test (VLT), version 1, 150 items (vocabulary size);
d) Word Associates Test (WAT), 320 items (vocabulary depth);
e) CAE Reading comprehension (RC) test (Cambridge ESOL Examination), 43 items.

The selection of tests needs commenting on. The COLLEX and COLLMATCH test versions were the same as those used in Study 6 in Section 5.2, and the reader is referred to a description of their design in that chapter. The Vocabulary Levels Test (VLT), version 1, was used in study 4 in Chapter 4 of this thesis. It was published in Schmitt (2000), and its design has been described in Chapter 2 of this thesis. In the previous two studies a modified version of the VLT was used, for technical reasons pertaining to the exam situations in which it was administered. Since none of these conditions applied in the present study, the version published by Schmitt (2000) was used.

As the vocabulary depth measure, the Word Associates Test (WAT) was used (Read 1993; 1998) as it is claimed to be one of the most well-known, and widely-used tests particularly targeting depth of word knowledge (Greidanus *et al.* 2004). Its design and characteristics were reviewed in Chapter 2. The version used in this study is intended primarily as a research tool (Read 1998:45), and it was retrieved in an electronic format from a webpage, where it has been made accessible as an on-line test by Tom Cobb (Cobb 2007). The electronic format was transformed into a paper and pencil test, and a test key was obtained from the provider of the webpage[35]. The test consists of 40 items, or blocks of items rather, each containing an adjective target word, and eight potential associate words. These eight words are in turn divided into two groups of four, with four adjectives in a box on the left, and four nouns in a box on the right. A sample test block is shown in Figure 6.2 below.

Sudden

☐beautiful ☐quick ☐surprising ☐thirsty	☐change ☐doctor ☐noise ☐school

Figure 6.2 An example block of items in the Word Associates Test (WAT).

The adjectives on the left are either potential synonyms of the target word *sudden*, or they represent one aspect of its meaning. As such, they are potentially paradigmatically linked to the target word. The nouns on the right are potential collocates of the target word, and are thus potentially syntagmatically linked to the target word. All in all, with 40 blocks comprising eight choices each, the test consists of as many as 320 choices. A test taker is instructed to select four of the eight words in the boxes as connected to the target word, and is furthermore told that there is no consistent number of correct answers on the left or on the right.

As to the reading comprehension test, I needed a test which would be challenging enough for university-level students of English with an advanced level of proficiency in English. At the same time, since I was planning on using data also from upper-secondary school students, I did not want to run the risk of using a test that was too difficult. I also wanted to use a standardized test, with acceptable levels of validity and reliability associated with its scores. In the light of this, a recent version of the University of Cambridge ESOL examination reading paper presented itself as a good choice. The paper is part of the Certificate in Advanced English (CAE), which is the second highest level Cambridge ESOL exam (after Certificate of Proficiency in English (CPE)), and it corresponds to level C1 of the Council of

[35] I would like to express my sincere gratitude to Tom Cobb (personal communication) for supplying the test key to the WAT version.

Europe's Common European Framework of Reference for Languages. According to information published on the Cambridge ESOL homepage, the reading paper "assesses your ability to read and understand a number of texts taken from books, newspapers and magazines. You are expected to be able to show understanding of gist, main points, detail, text structure or specific information, deduce meaning or recognise opinion and attitude." (Cambridge ESOL 2007). The reading test consists of four parts, two including multiple matching of a prompt to elements in a text, one gapped text from which paragraphs have been removed and placed in jumbled order after the text, and one text followed by four-option multiple-choice questions. The total number of questions/items is 43.

6.2.3 Informants

A total number of 24 informants took the five-part test battery. This number was slightly lower than hoped for, but the time required to take the complete test battery – between 2.5 and 3 hours – is believed to have affected the turnout. The informants were students from a local upper-secondary school, and students of English at Lund University. A breakdown of the informants and their study level affiliation is shown in Table 6.1.

Table 6.1 Informants in the present study and their level of study.

Informants	Number
Upper-secondary School students of English in Sweden, 11th grade	7
University students of English in Sweden, first term	3
University students of English in Sweden, second term	2
University students of English in Sweden, third term	12
Total	24

A total of seven volunteer upper-secondary school informants were recruited through contacts at a local school. A teacher was asked to try to recruit students with somewhat different abilities in English, so that not only the most proficient students would volunteer. The 17 university undergraduates were recruited during lectures at Lund University. Ideally, I would have liked to obtain an equal number of volunteers from the three university study levels, but this was not possible since I had to rely on the willingness of these students to participate. All 24 informants were rewarded for their participation in the form of a cinema ticket.

6.2.4 Research questions

The following research questions were addressed in the study:

1. Are COLLEX and COLLMATCH scores more closely related to results on a vocabulary size test or a vocabulary depth test, or equally related to both?
2. What is the relation between reading comprehension and each of the following variables: vocabulary size; vocabulary depth; collocation (COLLEX); collocation (COLLMATCH)?

6.2.5 Test administration and scoring

The test battery for the present study was comprehensive, containing five parts, and also demanding in terms of effort and time needed on the part of the informants. The time allotted for the reading comprehension part was 75 minutes, as recommended by specific CAE instructions for the test. For COLLEX and COLLMATCH together, approximately 25 to 30 minutes was deemed necessary. The time needed to take the VLT was estimated to range between 20 and 30 minutes, and for the WAT, around 20 to 25 minutes. All in all this meant that between 2.5 and 3 hours would be needed for each student taking the test battery. For the upper-secondary school students, the administration was divided into two sessions. The reason for this was twofold. Firstly, I did not have access to the students for more than 90 minutes at a time[36], which made two sessions necessary. Secondly, doing all five parts in a row was believed to be too cognitively demanding and tiring. For these reasons, session one contained the reading test and the WAT, and session two contained COLLEX, COLLMATCH and the VLT. For both sessions, taking place one week apart, I visited the school and administered and supervised the tests myself.

For the university students taking part in the study, an appointment was set up in each individual case, at a time of their choice, and they were sat in an adjoining office specially made available for this purpose. A majority of the students completed all five parts during one session. I specifically told them that it was essential that they take short breaks after having completed each part, and a long break after having done three out of five parts. Most university students handed in after approximately 3 hours, including breaks, but some needed slightly longer to finish. All students were asked to do their best, even though their performance had no bearing whatsoever on any course grades.

The test parts were marked in the following way. For the reading comprehension part, I produced a key in collaboration with an experienced lecturer of English as a key was not provided. The test was subsequently marked according to the instructions given in the test. Along the lines of the specified CAE marking instructions, the questions in two of the four test parts were given a mark of 2 points, and the other two a mark of 1 point. This resulted in a maximum score of 55 for the whole reading test.

The COLLEX and COLLMATCH test parts were marked in the same way as in previous studies: in COLLEX 5, 1 point was awarded for each correct answer, whereas 0 point was awarded for each incorrect answer; In COLLMATCH 3, a correctly identified real collocation was awarded 1 point, whereas a missed real collocation received 0 point; Conversely, a correctly rejected pseudo-collocation was awarded 1 point, whereas an incorrectly ticked pseudo-collocation received 0 point.

In the VLT test, correct answers were given 1 point, and 0 point was awarded for each incorrect answer.

Finally, for the WAT, much along the lines of the marking method adopted for COLLMATCH, correctly identified word associations were awarded 1 point, whereas missed associations received 0 point. Conversely, a correctly rejected distractor was awarded 1 point, whereas an incorrectly ticked distractor received 0 point.

[36] This was the length of their English lesson.

6.3 Results

Two main results will be reported in this section. Firstly, descriptive statistics for the scores from the 24 informants will be presented. Secondly, results from a number of different correlation analyses will be presented, aimed at providing answers to the two research questions.

The overall descriptive test results are presented in Table 6.2 below. In terms of scores, all tests except the Vocabulary Levels Test 1 (VLT 1) were normally distributed[37].

Table 6.2 Score distributions and test characteristics of VLT 1, WAT, CAE reading comprehension (READING), COLLEX 5, and COLLMATCH 3 (N = 24).

Value	VLT 1	WAT	READING	COLLEX 5	COLLMATCH 3
k	150	320	43	50	100
MPS*	150	320	55	50	100
Mean	133.5	263.0	41.0	42.2	80.2
S.d.	16.9	31.8	9.3	5.8	11.6
Range	67	129	30	19	43
Minimum	83	182	23	31	55
Maximum	150	311	53	50	98
Skewness	-1.4	-.87	-.50	-.65	-.67
Kurtosis	2.0	.89	-.92	-.94	-.27
Cronbach's α	.97	.96	.86	.86	.91

k =number of test items
* = Maximum Possible Score

With a kurtosis of 2.0, the score distribution of VLT 1 is markedly peaked, bordering on leptokurtosis. The distributions on all five tests were negatively skewed. The mean scores on all five tests are relatively high, corresponding to 89% of the maximum score for VLT 1, 84% for COLLEX 5, 82% for WAT, 80% for COLLMATCH 3, and 75% for the READING test. All five variables were reliably measured, as indicated by the high values for internal consistency (Cronbach's alpha ranging between .86 and .97). This is an important finding since subsequent meaningful correlation analyses presuppose reliably measured scores.

A closer look at the scores on the VLT 1, presented in Table 6.3, shows that the informants as a group performed well on the 2K and 3K levels, with means above 29.

Table 6.3 Mean scores and standard deviations on VLT 1 word frequency levels. The maximum score on each level is 30. (N = 24).

Level 2000	Level 3000	Level Academic	Level 5000	Level 10000
29.3 (1.1)	29.2 (1.4)	28.2 (2.7)	26.2 (5.2)	20.7 (7.8)

[37] Based on values for *Skewness* and *Kurtosis*. Values between -2 and +2 indicate a reasonably normal distribution (Bachman 2004:74).

The mean scores on the academic word level were also high, with a mean of 28.2. A small dip down to a mean of 26.2 is noticeable for the 5K level, and a clear decrease in mean scores is observable for the 10K level (20.7). Table 6.3 above shows that the informants scored progressively lower as the word level frequency decreased in the test. The results in the table also indicate that the academic word level fit well between the 3K and the 5K level in terms of mean difficulty. If we take a score of 26 as a cut-off score for mastery of a level in the test, following procedures in Schmitt *et al.* (2001), then the group as a whole mastered the 5K level in terms of the mean score reported in Table 6.3. However, an analysis of the scores of the 24 individuals showed that 8 of these (33%) did not reach the 5K mastery level, and 17 (71%) did not reach the 10K mastery level, as indicated in Table 6.4 below.

A conclusion that may be drawn from this is that around 33 per cent of the informants in the study may have had problems with certain words in COLLEX 5 and COLLMATCH 3 since they were taken from lower frequencies than the 3K band (see Appendix 5J and 5L).

Table 6.4 VLT 1 word frequency levels and the number of informants reaching the mastery cut-off score of > 26 out of 30 for each level (N = 24).

Level	Number of informants (per cent)
Level 2000	24 (100%)
Level 3000	24 (100%)
Level 5000	16 (67%)
Level 10000	7 (29%)

The next analysis was carried out in order to arrive at results that could be used to provide an answer to the first research question, i.e. whether COLLEX and COLLMATCH scores were more closely related to either vocabulary size scores or vocabulary depth scores. For this purpose a correlation analysis was conducted. The correlation values were computed using the Pearson Product Moment Correlation Coefficient. A one-tailed test was used since positive correlations were expected between all variables. The result is shown in Table 6.5 below.

Table 6.5 Correlations (Pearson *r*) between the five test battery variables (N = 24).

Value	VLT 1	WAT	READING	COLLEX 5	COLLMATCH 3
VLT 1	-	.93**	.69**	.90**	.90**
WAT	-	-	.80**	.85**	.89**
READING	-	-	-	.64**	.68**
COLLEX 5	-	-	-	-	.89**

** Correlation is significant at $p < .01$.

The results in the table require commenting. Firstly, moderate to strong significant, positive correlations exist throughout (.64 - .93) between the five main variables. It was indeed hypothesized that positive correlations would exist since four out of the five variables can be seen as some sort of vocabulary construct, and since it stands to reason that reading

comprehension skills are contingent on the knowledge of the inherent building stones of texts, viz. words.

Secondly, the scores on COLLEX 5 and COLLMATCH 3 correlate almost equally highly with both the vocabulary size measure (VLT 1, .90 for both), and the vocabulary depth measure (WAT, .85 and .89, respectively). This result implies that it is not possible to assert that COLLEX or COLLMATCH should gravitate more towards one of these assumed dimensions than the other. Further implications of this result will be discussed in section 6.4 below. The highest correlation was observed between the VLT 1 scores and the WAT scores, at .93. This is also an interesting finding which merits further discussion.

Having observed high, positive correlations between the COLLEX, COLLMATCH, VLT 1 and WAT variables, their relation to the reading comprehension variable is clearly of interest. As can be seen in Table 6.5, the correlations range between .64 and .80. Some of these levels of correlation perhaps come across as lower than expected, considering earlier findings of higher values, at around .80 (see e.g. Qian 1999). Irrespective of the various correlation levels, it is clear that the coefficients (.64 - .80) are lower than the levels between COLLEX, COLLMATCH, VLT and WAT (.85 - .93. This indicates that reading comprehension is indeed a different construct.

Out of the relationships with reading comprehension scores, the scores from the WAT were the ones that correlated most highly, at .80. In a basic correlation study, we cannot make any direct conclusions about causality, but we can take the correlation coefficient a step further by squaring it. The correlation coefficient squared (R^2) is a measure of the amount of variability in one variable that is explained by the other (Field 2005:128). As such, we can estimate the predictive value of a variable. Thus, it is possible to find out to what extent the reading comprehension scores can be explained by the different vocabulary-related variables. The proportion of variance in reading comprehension scores accounted for by the variance in the four vocabulary variables is shown in Table 6.6 below.

Table 6.6 Correlation coefficients squared (R^2): The variance in reading comprehension scores accounted for by four predictor variables (N = 24).

Predictor value	Reading comprehension (RC)
VLT 1	.48
WAT	.64
COLLEX 5	.41
COLLMATCH 3	.46

The proportions of variance accounted for are relatively modest on the whole, and the best predictor is the WAT, with VLT as the runner-up. With R^2 values between .41 and .64, we are still left with between .36 and .59 of the variance in the reading comprehension scores unaccounted for. As was pointed out earlier, the somewhat unexpected performance of some informants in the study may have distorted the picture.

In order to investigate the predictive capacity of the vocabulary size scores, the correlation coefficients obtained for this variable vis-à-vis the other four variables were also squared. The results are shown in Table 6.7 below. Out of the four variables, vocabulary size had the best prediction strength for scores on the vocabulary depth variable (WAT), at .87. According to Heiman (2006:188) values around and above .50 are "very large".

Table 6.7 Correlation coefficients squared (R^2): The variance in four variable scores accounted for by vocabulary size as the predictor variable (N = 24).

Predictor value	RC	WAT	COLLEX 5	COLLMATCH 3
VLT 1	.48	.87	.81	.81

This means that the squared coefficient value of .87 should be an indicator of a very important strong relationship. If we know someone's scores on the VLT 1, this score should prove valuable for identifying their depth of word knowledge, as measured through the WAT. Similarly, but not as strongly, VLT 1 scores are good predictors of COLLEX and COLLMATCH scores, both at R^2 = .81.

6.4 Discussion

Before the results of this study are discussed, a caveat is called for. The caveat bears upon the small sample of informants used in the study. Because of the small sample, caution must be observed when interpreting the results, and when discussing implications of the results. However, the results can certainly serve as tendencies which point in one direction or other. Also, despite the small sample of informants, all five measured variables were satisfactorily reliable, with values of internal consistency between .86 and .96. This fact serves as a prerequisite for subsequent correlation analyses.

In the following discussion section, I will structure my discussion around the research questions put forward in section 6.2.4.

6.4.1 Are COLLEX and COLLMATCH scores more closely related to results on a vocabulary size test or a vocabulary depth test?

An answer to this question was attempted through the design of a validation study consisting of five different variables. Two somewhat conflicting assumptions guided the study. One assumption held, based on empirical evidence from previous studies in this thesis, that there was a strong relation between scores on COLLEX and COLLMATCH, and vocabulary size. This raised the question whether COLLEX and COLLMATCH may be viewed more as size tests than depth tests. The second assumption held that collocation knowledge is strongly affiliated with vocabulary depth, evidenced through its mention in passages on depth or quality of word knowledge in the literature (see e.g. Read 2000; Schmitt 2000; Jiang 2004b).

The results showed that COLLEX and COLLMATCH scores correlated slightly more strongly with vocabulary size scores than with vocabulary depth scores, but the differences were tiny. This is an intriguing result and it leaves the door open to several possible explanations. If vocabulary size and vocabulary depth are assumed to be different dimensions of lexical competence, we would expect COLLEX and COLLMATCH to correlate more highly with one of them, than with the other. Formally, this did happen, but the difference was negligible. At the same time, the vocabulary size measure was observed to correlate very highly with the vocabulary depth measure, at .93. This has two interesting implications. One of them has to do with the influence of vocabulary size on COLLEX and COLLMATCH, and WAT scores, and the other with problems affiliated with the conceptual treatment of

vocabulary size and depth as independent 'dimensions'. I will below discuss these two, one at a time.

The first implication is that one could argue that COLLEX and COLLMATCH, and the WAT, are all influenced by vocabulary size to a great extent. In terms of the WAT, Wolter (2005) points to the fact that some of the words featuring in the version used in the present study are fairly low-frequency items, and that vocabulary size is therefore believed to have a considerable influence on test-takers' performance (p. 37). A closer look at some of the words featured in the WAT test version used in this study confirms this. For example, target words like *ample*, *synthetic* (both 6K), and *fertile* (7K), together with associate words like *cautious* (5K) and *plentiful* (8K) are clearly not high-frequency words. I will later on in this discussion come back to the qualities of the WAT as a vocabulary depth test. In the recent development process of COLLEX and COLLMATCH, attempts were made to minimize the influence of vocabulary size on scores by keeping the frequencies of the single words making up the word combinations in the tests as high as possible. However, a small number of low-frequency words (> 5K) are included in the tests (see Appendices 5J and 5L). The mean frequency of the words in COLLEX 5, with regard to frequency bands, is 1.6 for verbs and 1.8 for nouns. For COLLMATCH 3 the mean for verbs is 1.8 and for nouns 1.9. The mean for the 40 WAT words is 3.2. Thus, based on sheer frequencies of the single words, COLLEX and COLLMATCH should in theory be less dependent on vocabulary size then the WAT. The results obtained in the present study, however, did not quite support this assumption. By taking the data from Table 6.4 into account, it is clear that especially for upper-secondary school students, certain lower frequency words could be problematic, as the full range of words in bands 4K and 5K are probably not known. To do well on COLLEX and COLLMATCH, it is assumed that you have to know the meaning of individual words, and based on the frequencies of the single words of the two tests, you seemingly need a moderately sized lexicon of around 5000 words to do this.

In Table 6.7, I presented an analysis of how many informants passed a certain criterion mastery level on the vocabulary size test. It will be appropriate here to further analyse these data in relation to the COLLEX and COLLMATCH scores in order to corroborate this impression. If we for argument's sake form three groups based on the vocabulary size scores: group 3K, group 5K and group 10K, based on whether informants reached a criterion score of 26 out of 30 on the different word levels, we arrive at results provided in Table 6.8. The differences between the mean scores were all statistically significant: for VLT 1, Welch F (2, 10.28) = 31.70, $p < .001$; for COLLEX 5, Welch F (2, 12.47) = 34.50, $p < .001$; for COLLMATCH 3, Welch F (2, 12.93) = 19.00, $p < .001$. Based on the small data set, the result shows that learners with an estimated vocabulary size of at least 3000 words, but smaller than 5000 words, scored around 72 per cent on COLLEX and 67 per cent on COLLMATCH.

Table 6.8 Comparison of scores based on VLT levels criterion groups.

Groups	N	VLT 1			COLLEX 5			COLLMATCH 3		
		M	S.D.	Range	M	S.D.	Range	M	S.D.	Range
3K	8	114.9	15.5	83-136	35.8	4.0	31-42	67.4	8.9	55-76
5K	9	138.2	5.8	127-145	43.6	3.1	36-46	83.2	4.9	72-88
10K	7	148.9	0.9	148-150	47.9	1.4	47-50	90.9	5.3	85-98

Learners with an estimated vocabulary size of at least 5000 words, but smaller than 10000 words scored around 87 per cent on COLLEX and around 83 per cent on COLLMATCH, whereas learners with an estimated vocabulary size of at least 10000 words scored around 96 per cent on COLLEX and around 91 per cent on COLLMATCH. Thus, it seems to be possible to roughly predict COLLEX and COLLMATCH scores based on their vocabulary size scores.

However, as can be seen from the range scores, there is a fair degree of overlap between the scores on COLLEX and COLLMATCH produced by the individuals in the three criterion groups. What is interesting, for example, is that one learner (called X) from the 3K group scored as high as 42 (84%) on COLLEX. This individual was estimated to have a vocabulary of 3000 words according to the mastery criterion. With a score of 42 on COLLEX, X ended up at the same level as three informants who were estimated to have a vocabulary of 5000 words according to the mastery criterion. Their score on COLLEX was 43. However, if we compare these scores with the scores obtained on the COLLMATCH test, the pattern breaks. Our learner X scored 65, whereas our three 5K informants scored 84, 86, and 88, respectively. At a first glance, it seems possible that an individual with a relatively small vocabulary, in terms of single words, has a rather high level of receptive collocation knowledge. However, his performance on COLLMATCH did not quite match his high score on COLLEX. In fact, from an impressionistic point of view, it seems in general as if learners' scores on COLLMATCH can be roughly predicted by doubling their COLLEX score. This was not the case for our 3K vocabulary size learner (X). It should be pointed out though that VLT scores are rough estimations of vocabulary size, and a closer look at the performance of the 3K informant reveals that he was close to reaching the mastery level for the 5K band. He scored 23 out of 30. In fact, all of the informants from the 3K group except one were relatively close to the mastery criterion score for the 5K band. The one learner that was not close scored 8 out of 30 on the 5K band, and her scores on COLLEX and COLLMATCH were 31 and 59 respectively. We also find an example of a learner (Y) who has a fairly large vocabulary size (5K), but who does not perform the same high scores as the other 5K group members. Learner Y's scores on COLLEX and COLLMATCH were 36 and 72 respectively.

Thus, even though it seems that vocabulary size does explain COLLEX and COLLMATCH scores to a great extent, we also have evidence that suggest that some individuals do not conform to this pattern. How can these results be explained? One possibility is measurement error and the high probability of answering an item correctly by guessing. Another possibility is differences in learning strategies. There might for example be a difference between a learner who has had minimal exposure to natural L2 input, and possible reliance on list learning of vocabulary items in a classroom situation, and a learner

who has been exposed to natural L2 input to a large extent outside of typical classroom instruction. This is reminiscent of Meara & Wolter's (2004) hypothesis that learners with similar sized vocabularies might differ in respect of how organized their vocabularies are. It should be pointed out that these authors discuss 'vocabulary organisation' as a fundamental dimension of lexical knowledge, and although 'vocabulary organisation' has similarities to 'vocabulary depth', they are modelled on quite different assumptions. However, we saw in Section 2.4.3.2.3 that depth can be seen from three perspectives: precision of meaning, comprehensive word knowledge, and network knowledge (Read 2004), and Meara & Wolter's term organisation is closely associated with the network knowledge perspective. Meara & Wolter furthermore hypothesize that learners with large but weakly organized lexicons may behave differently from learners with similarly sized, but better organized, lexicons. As an example of a potential difference, they suggest text comprehension. Examining this hypothesis in the light of the data from the present study, I managed to find an example that could be indicative of this. Consider Table 6.9 below. The two learners (called A and B) had the same score on the vocabulary size test: 139. In terms of their profiles on the different levels, these are close to identical. Admittedly, there is a 1-point difference on the 5K and 10K levels, between the learners, but it is so small that it is negligible. If we thus treat them as having similar levels of vocabulary size, a striking difference emerges in terms of their scores on the other lexical measures. Learner A clearly scored better than B on COLLMATCH (88 vs. 80), on the WAT depth test (280 vs. 265), and on the reading comprehension test (53 vs. 43).

Table 6.9 Comparison of scores from two learners.

	Vocabulary size (VLT 1)						COLLEX	COLL-MATCH	READING	WAT
	Tot	2K	3K	AC	5K	10K				
Learner										
A	139	30	30	30	27	22	46	88	53	280
B	139	30	30	30	28	21	45	80	43	265

Following Meara & Wolter's hypothesis, learners A and B have similar-sized vocabularies, but the lexicon of learner B is more weakly organized (COLLMATCH and WAT scores), and her reading comprehension is weaker than that of learner A. The COLLEX scores are very similar, which points to a potential problem with COLLEX not having enough discriminatory power with very advanced learners.

Going back to the second implication of the results obtained in this study, an assumption saying that vocabulary size and vocabulary depth are different dimensions of lexical competence is perhaps faulty, or tenuous at best. As pointed out by Read (2004:221): "Although the tendency of authors since Anderson & Freebody (1981) has been to contrast the concepts of breadth and depth as if they are – if not polar opposites – at least quite distinct dimensions of vocabulary knowledge, the small amount of evidence that is available so far suggests that they are somewhat closely related". In a similar vein, Vermeer argues that "Breadth and depth are often considered opposites. It is a moot point whether this opposition is justified. Another assumption is that a deeper knowledge of words is the consequence of knowing more words, or that, conversely, the more words someone knows, the finer the

networks and the deeper the word knowledge." (2001:222). Vermeer explains the high correlations he observed (.70 and .83) in the following way: "The strong correlations between breadth and depth measures of vocabulary justify the position that there is no conceptual distinction between the two. The high correlations are a logical consequence of the fact that the lexical elements in the mental lexicon consists [sic] of interrelated nodes in a network, which specify the meaning of an element." (2001:231). In a similar vein, Qian asserts that "…the high correlation between the scores on the DVK [depth test] and scores on the VS [size test] strongly suggests that learners' scores on the depth and breadth dimensions of vocabulary knowledge are also closely, and positively, associated, which leads us to believe that development of the two dimensions is probably interconnected and interdependent." (1999). Vermeer's and Qian's arguments are intuitively appealing, and it seems logically sound to assume that there is a fair degree of developmental interaction going on. An individual must know (in the sense of having acquired a basic form-meaning mapping) a large number of words as a prerequisite for developing an extended and more detailed kind of knowledge of these words.

Irrespective of the apparent interdependedness of vocabulary breadth and depth, the conceptual treatment of the two does clearly harbour problems. Meara (personal communication) argues that it does not make sense to call depth a dimension, in the same way as size can be called a dimension. The reason for this is, he claims, that size is used to describe the whole lexicon, not as a property of a single word. You cannot have breadth or size of a single word, but as most researchers see depth, arguably as comprehensive word knowledge (see Read 2004), it is possible to have depth knowledge of a single word. The effect of this seems to be that the two "dimensions" are associated with different measurement characteristics. Meara's solution, presented in Meara & Wolter (2004), implies using a network metaphor, whereby depth is substituted for 'organisation'. Organisation refers to the degree of interconnectivity of the words in the lexicon. If a new word is added to the lexicon, then this is assumed to have implications for the rest of the network. Interestingly, Meara & Wolter found a modest level of correlation between scores on a test of overall vocabulary size and scores on a lexical organisation test tool called V_Links ($r < 0.3$), which was taken as support for the view that size and organisation are "more-or-less independent features of L2 lexicons" (2004:93). Wolter (2005), putting different versions of V_Links to the test, found similarly low, or even inverse (though not significant), correlations with vocabulary size. Wolter concludes that there is evidence to suggest that vocabulary organisation, as measured by V_Links (versions 2.0 and 4.0), and vocabulary size may develop orthogonally (2005:208). One of the true advantages of V_Links over tests like the WAT, and indeed COLLEX and COLLMATCH, seems to be the control for vocabulary size effects. Whereas the WAT contains a number of lower-frequency words, the words in V_Links are all 1K words. As long as a number of low frequency words, albeit few, are featured in tests like the WAT, COLLEX and COLLMATCH, it is perhaps not surprising that we observe high correlations with vocabulary size.

Before the results relevant to the second research question are discussed, the use of the current WAT version as a proper depth test merits further discussion. At the outset of the present study, I more-or-less accepted the WAT as a proper vocabulary depth test. On reflection, however, the use of the WAT for these purposes in the study is fraught with certain problems. In a paper reporting a validation study of the WAT, Read (1998) accounts for the target word selection process. The words were selectively chosen adjectives from a word list

(Nation 1986), claimed to give a comprehensive coverage of high frequency academic vocabulary (p.45). If we for argument's sake background the word academic for a moment, it is questionable to what extent the words are high-frequency words, and I have already pointed at the size-dependency effect that the inclusion of lower-frequency words has on the supposed depth of word knowledge scores of the WAT. However, the fact that Read states that the selection targets high-frequency academic [my underlining] vocabulary, means that it may be less well suited as a depth of vocabulary knowledge "research tool", a use which is specifically aimed at (p. 45). In a more critical view, it may serve well as a research tool of depth of academic vocabulary, specifically adjectives. Another point which may confound the results is the selective rather than random selection of target items. This method of item selection makes generalisations from test scores to a general underlying ability less straightforward, since it is difficult to know how the correct identification of associate words of certain target adjectives relates to a more general depth of knowledge[38].

Having identified certain shortcomings of the WAT as a general vocabulary depth test, its use for another purpose emerged. In lack of other measures of receptive collocation knowledge, a lack which moreover was one of the reasons behind the development of such measures in this thesis in the first place, one way to gather validation data for the scores on COLLEX 5 and COLLMATCH 3 would be to correlate these with parts of the full WAT administered in this study. The rationale behind this approach is the fact that half of the 320 items, i.e. 160 items, are words which are potentially linked to the target word on the basis of a syntagmatic relation, which to a great extent implies the same link that exists in what I have called collocation in this thesis. In the box to the right in the example given in Figure 6.2, syntagmatically related words like *sudden* (target word) + *change* (associated word) and *sudden* (target word) + *noise* (associated word) are shown. The example also contains two words intended to function as distractors: *doctor* and *school*. An informant who chooses to tick these two words as relevant to the target word is seen as making an idiosyncratic link between the target adjective word *sudden*, and the noun alternatives *doctor* and *school*. These two words would not enter into a sequence with the target word sudden seen as a collocation under the view taken in the present thesis. The task at hand in the WAT is to some extent reminiscent of the task in COLLMATCH 3. In fact the WAT task is even more reminiscent of the task used in COLLMATCH 1, where the same target word was tested for each combinatory potential with six other words (see section 4.1). The difference between the WAT task and the COLLMATCH 3 task is that the items in the latter are more independent of each other. In each item, a unique verb is combined with a unique noun. In the WAT, the same target adjective must be probed for its combinatory potential with as many as eight other words. Notwithstanding these differences, a test-taker is required to select frequently occurring word combinations (collocations) and resist the selection of word combinations that would not normally be used by native speakers (pseudo-collocations) in both tests.

There are similarities between the WAT and COLLEX too. Both tests are tests of receptive knowledge. However, in COLLEX, a test-taker is required to make a choice between three options in each test item. In each item, the same noun is presented together with three different verbs. The verb + NP combination that is deemed to be a frequently occurring combination in the English language (collocation) should be chosen over two other

[38] To be fair, the face value acceptance and use of the WAT (new version presented in Read 1998) as a more general vocabulary depth test and the accompanying problems naturally fall back on myself.

combinations that are not. This is in line with the syntagmatic "half" of the WAT, but the task is slightly different.

Irrespective of the identified differences, using the syntagmatic "half" of the WAT presented itself as practically the only viable choice for the purpose of a concurrent validity measure vis-à-vis COLLEX and COLLMATCH. As a first step, the scores on the WAT were therefore divided into two parts. One part consisted of the informants' scores on the paradigmatic half of each target word block, and the other consisted of scores on the syntagmatic half of each target word block. A reliability check of the scores on the two parts showed that they were highly reliable (Cronbach's alpha = .93 for the paradigmatic part, and .92 for the syntagmatic part). Since the syntagmatic links part of the WAT can be seen as assessing receptive collocation knowledge, a high correlation between scores on this part and the scores on COLLEX and COLLMATCH would be expected and would, if it was observed, be taken as concurrent validity support for the two latter as receptive collocation tests. The result of the correlation analysis is shown in Table 6.10 below. As can be seen in the table, significant high correlations were observed between COLLEX 5 and COLLMATCH 3 scores, respectively, and the scores on the syntagmatic part of the WAT, at .84 and .86.

Table 6.10 Correlations (Pearson *r*) between COLLEX and COLLMATCH, and parts of the WAT (N = 24).

Correlations	WAT paradigmatic half[1]	WAT syntagmatic half[2]
COLLEX 5	.81**	.84**
COLLMATCH 3	.88**	.86**

** Correlation is significant at p < .01., one-tailed.
[1] Only paradigmatic link items of the WAT test, k = 160
[2] Only syntagmatic link items of the WAT test, k = 160

This lends validation support to COLLEX and COLLMATCH, more specifically in terms of concurrent validity, as tests of receptive collocation knowledge. However, as is also evident in the table, high correlations were also observed for the paradigmatic half of the WAT, at .81 and .88. This clearly confounds the initial finding. In fact, for COLLMATCH 3 scores we observe a slightly higher correlation with the paradigmatic half of the WAT, than with the syntagmatic half. This is somewhat surprising and it is difficult to explain why this happened. With only 24 informants, though, caution must be taken not to draw any far-reaching conclusions based on these results, since chance factors could in theory cause construct irrelevant changes to ranked scores, which in turn affect correlation coefficients.

An analysis of the test key revealed that 73 out of the 160 items on the paradigmatic part were intended as associated words and consequently 87 were intended as distractors. For the syntagmatic part the numbers were reversed, in that 87 were intended as associated words, and 73 were intended as distractors. As many as 20 out of 24 informants scored better on the paradigmatic half on the WAT than on the syntagmatic half, and the scores on the respective halves correlated at r = .91. The mean scores (and Standard deviations) were 135.0 (16.0) on the 160-item paradigmatic part, and 128.0 (16.5) on the 160-item syntagmatic part, which corresponded to a mean Item Facility of .84 on the paradigmatic half, and .80 on the syntagmatic half. The difference between the raw score means was statistically significant,

$t(23) = 4.91$, $p < .05$, r = 51. This means that these informants were better at recognizing synonyms and rejecting non-associate words of the target words, then they were at recognizing collocates and rejecting pseudo-collocations. This could be taken as evidence of the fact that for informants in the present study, knowledge of collocations lagged behind knowledge of synonyms for the tested target words. This is indeed intriguing, and it supports the argument put forward by Schmitt (1998), that word knowledge is likely to be at least partially hierarchical, and that collocation knowledge is likely to occur at a relatively late stage, after other types of word knowledge have been acquired. Empirical support for this was found by Greidanus and Nienhuis (2001), and Greidanus *et al.* (2005), who observed on the part of the informants of these studies (L1 Dutch and L1 English university level learners of French, and French native speakers at university level) better performance at paradigmatic than syntagmatic items in a quality of word knowledge test of French, based on Read's (1993, 1998) WAT test. Similarly, Bahns & Eldaw (1993) have argued that collocation knowledge does not develop alongside general lexical knowledge. However, in my review of their study earlier in this thesis I questioned the method they used for arriving at this conclusion. Furthermore, the question is what is considered as general lexical knowledge.

6.4.2 What is the relation between reading comprehension and each of the following variables: vocabulary size; vocabulary depth; collocation (COLLEX); collocation (COLLMATCH)?

The results from the present study imply that the relationship between the four vocabulary-related variables contribute to reading comprehension in slightly different ways, but that the differences are indeed small, possibly due to the small number of informants participating in the study. In the initial correlation analysis, small differences were observed in terms of how the four different vocabulary-related measures related to reading comprehension. The significant correlation coefficients ranged between .64 and .80. In the subsequent analysis, where these correlations were squared, we saw that the vocabulary size measure (VLT) and COLLMATCH ended up explaining a similar amount of variance in the reading comprehension scores (.48 and .46), whereas COLLEX ended up lower than VLT and COLLMATCH, and the vocabulary depth measure (WAT). The amount of variance in reading comprehension accounted for by COLLEX was .41, and by WAT .64. This means that the vocabulary depth test (WAT) was the variable that accounted for most of the variance in the reading comprehension scores.

It is very difficult to explain the observed differences, and with the small sample of informants, there is a risk that the performance of one or two individuals will affect the correlations obtained to a large extent. Therefore, a larger sample of informants is needed for a possible replication study in order to find substantiated answers to the research question.

On an anecdotal note, I suspected that certain informants relied on guessing in the last part of the reading comprehension test. The reason for this was that their performance in the other three parts of the test was fairly good, but that many incorrect answers were given in the last part. It is possible that they suffered from test fatigue by this stage. In my own view, the reading test was difficult and more cognitively demanding, a view which I furthermore shared with the experienced lecturer of English who also took the test. This could have meant that it was difficult to the extent that the informants ran out of time, energy and motivation, and therefore resorted to guessing.

A closer look at the individual scores for the reading test revealed that some informants, who were expected to score highly on the reading test considering their vocabulary size scores and their level of study, received relatively low scores. Figure 6.3 below was produced in order to illustrate this. Informants (cases) whose scores lie above the regression line are those which performed relatively better on the vocabulary size test in relation to their performance on the reading comprehension test. For example, a learner who got a score of 83 on the VLT (case 1), received almost as high a score, 26, on the reading test as a learner with 136 on the VLT (case 11), scoring 28 on the reading test. Clearly, case 11 would be expected to perform better on the reading comprehension test, considering the high score in terms of vocabulary size.

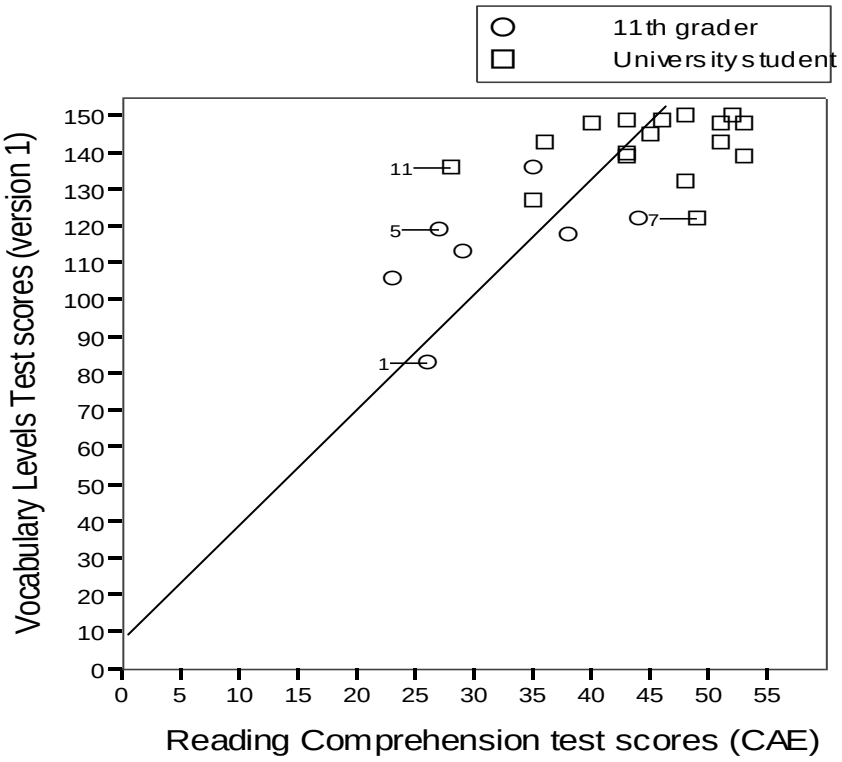

Figure 6.3 Correlation between scores on the Vocabulary Levels Test and CAE Reading Comprehension test (r = .69, N = 24).

There were also other anomalies in the lists of ranked scores, which are believed to have resulted in lower than expected correlations. Cases 5 and 7 scored 119 and 122 respectively on the VLT test, and are thus assumed to have a very similar vocabulary size. However, case 7 got an almost twice as high score on the reading comprehension test as case 5 (49 points vs. 27 points). A closer look at their respective performance on the different levels of the size test revealed that they performed very similarly up until the 10K level, where case 7 scored 15 out of 30, whereas case 5 scored 9 out of 30.

6.5 Summary and conclusions

Based on the results of the study, it was not possible to conclude that the COLLEX and COLLMATCH test scores gravitate more towards a vocabulary depth dimension (WAT) than a vocabulary size dimension (VLT), since strong correlations were observed between these four variables at r = .85 - .93. Two possible explanations for this were given. Either, language users need to possess both a large vocabulary and 'deep' word knowledge to do well on COLLEX and COLLMATCH, or the assumed and sometimes polarized distinction between vocabulary size and vocabulary depth must be questioned. Vocabulary size was concluded to be an important factor for the performance on the two receptive collocation tests. A high correlation with data from the syntagmatic part of the Word Associates Test was taken as support for concurrent validity of COLLEX and COLLMATCH as tests of receptive collocation knowledge, but equally high correlations with the paradigmatic part somewhat confounded this interpretation. The fact that the scores on the four tests of lexical knowledge (COLLEX, COLLMATCH, VLT, and WAT) correlated to a slightly lower extent (between .64 and .80, all significant), with the reading comprehension test was believed to stem from the fact that the reading comprehension test indeed measures a different construct. It was concluded that a study comprising a much larger group of informants, and a more careful selection of test tools would be needed to fully evaluate the relationships between reading comprehension and vocabulary size, vocabulary depth, and COLLEX and COLLMATCH scores.

7 Discussion

7.1 Introduction

This thesis is concerned with the development and evaluation of two tests of receptive collocation knowledge and the performance of advanced Swedish learners of English. The focus of investigation with regard to the tests per se has been the pursuit of evidence of valid and reliable scores, which would allow the tests to be used for educational as well as research purposes. In addition to the test development process, the focus has been to investigate the potential role of vocabulary size in determining learners' performance on the collocation tests, as well as the role that learning level, i.e. the number of years of classroom exposure to English, may have.

In this chapter, I will first summarize the main findings of the experimental work carried out in Chapters 3-6. I will then discuss these findings under three main headings, corresponding to the three research questions.

7.2 Summarizing the main findings from the empirical studies

As a suitable point of departure, consider again the main research questions proposed in Chapter 1:

RQ1: Is it possible to develop tests measuring receptive knowledge of English collocations as a single construct, capable of yielding reliable and valid scores, for use with advanced Swedish learners of English?

RQ2: What is the relationship between Swedish L2 learners' vocabulary size and their receptive knowledge of collocations?

RQ3: What is the relationship between the learning level of Swedish L2 learners' of English and their receptive knowledge of collocations?

Research question one addresses the qualities of COLLEX and COLLMATCH as test tools, whereas research questions two and three primarily concern aspects of learning. As was pointed out in Chapter 1, an affirmative answer to question one is more or less a prerequisite for the pursuit of answers to questions two and three. However, we must remember that validity is not an all-or-nothing quality, and different aspects of validity may be argued to exist to varying extents for a particular set of test scores.

The main findings of this research project are presented in Table 7.1 below (several pages). In the table, each study is briefly described by stating which version of COLLEX and COLLMATCH was used, any additional tests that were employed, the number and types of informants, the observed overall reliability, validity aspects and main findings.

Table 7.1 Summary of empirical studies reported in Chapters 3-6.

Study	Thesis section	Main test/s examined (number of items)	Additional test/s used	Informants (type)	Overall Reliability (main tests)	Validity and main findings (main tests)
1	3.1	COLLEX 1 (60 items)	SINGLEX 1	19 (Swedish university level)	Unacceptable (.54)	-Unreliable test scores -Ceiling effect -Poor item quality in terms of item-total correlation
2	3.2	COLLEX 2 (65 items)	SINGLEX 2	83 (Swedish university level)	Very good (.82)	-Reliable test scores -Ceiling effect tendencies -Improved but still somewhat poor item quality -High-scoring learners guessed less often and more successfully than low scoring learners
3	4.1	COLLEX 3 (50 items)	SINGLEX 3	103 (97 Swedish university level + 6 NSs of English)	Very good (.83)	-Reliable test scores -Decent item quality -Ceiling effect tendencies -High-scoring learners guessed less often and more successfully than low scoring learners -No effect of Swedish target prompt insertion on test scores -Native speaker scores provided validity support -Some discrimination between Swedish informants at different learning levels
		COLLMATCH 1 (144 items)			Very good (.80)	-Reliable test scores -Poor item quality in terms of item-total correlation -Undesirable outcome of COLLMATCH grid format design

4	4.2	COLLEX 4 (50 items)	VLT 1	188 (134 Swedish university level + 54 Swedish upper-secondary school level)	Excellent (.91)	-Highly reliable test scores -Acceptable item quality -Ceiling effect tendencies -High correlation with vocabulary size measure observed (r = .87) -High correlation with COLLMATCH observed (r = .92) -> concurrent validity -Some discrimination between Swedish informants at different learning levels -Evidence of relation between COLLEX 4 scores and general English proficiency
		COLLMATCH 2 (100 items)			Excellent (.92)	-Highly reliable test scores -Acceptable item quality -High correlation with vocabulary size measure observed (r = .87) -High correlation with COLLEX observed (r = .92) -> concurrent validity -Some discrimination between Swedish informants at different learning levels -Evidence of relation between COLLMATCH 2 scores and general English proficiency
5	5.1	COLLEX 5 – PILOT (40 items)	VLT M	25 (22 Swedish university level + 3 NSs of English)	Unacceptable (.58)	-Unreliable test scores -Evidence of face validity and to some extent response validity -Satisfactory outcome of new test design in terms of introduction of a 2nd distractor in each item, and lower mean scores relative to previous test versions.
		COLLMATCH 3 – PILOT (100 items)			Very good (.82)	-Reliable test scores -Evidence of face validity and to some extent response validity -Satisfactory outcome of new test design in terms of item facility and item response design

6	5.2	COLLEX 5 (50 items)	VLT M	269 (209 Swedish university level + 26 Swedish upper-secondary level + 34 NSs of English)	Very good (.89)	-Reliable test scores -Good item quality -High correlation with vocabulary size measure observed (r = .88) -High correlation with COLLMATCH observed (r = .86) -> concurrent validity -Evidence of relation between COLLEX 5 scores and general proficiency -Evidence of construct validity through NS comparison group -Evidence of construct validity through discrimination between Swedish learner groups of differing proficiency levels
		COLLMATCH 3 (100 items)			Very good (.89)	-Reliable test scores -Good item quality -High correlation with vocabulary size measure observed (r = .83) -High correlation with COLLEX observed (r = .86) -> concurrent validity -Evidence of relation between COLLMATCH 3 scores and general proficiency -Evidence of construct validity through NS comparison group -Evidence of construct validity through discrimination between Swedish learner groups of differing proficiency levels
7	6	COLLEX 5 (50 items)	VLT 1 WAT CAE READING	24 (17 Swedish university level + 7 Swedish upper-secondary level)	Very good (.86)	-Reliable test scores -Evidence of construct validity through correlation with WAT collocation part (concurrent validity) but somewhat confounded by correlation with WAT paradigmatic part. -High correlation with vocabulary size measure observed (r = .90) -High correlation with COLLMATCH 3 observed (r = .89) -> concurrent validity -High correlation with vocabulary depth measure observed (r = .85) -Moderate correlation with reading comprehension measure observed (r = .64)
		COLLMATCH 3 (100 items)			Excellent (.91)	-Highly reliable test scores -Evidence of construct validity through correlation with WAT collocation part (concurrent validity) but somewhat confounded by correlation with WAT paradigmatic part. -High correlation with vocabulary size measure observed (r = .90) -High correlation with COLLEX 5 observed (r = .89) -> concurrent validity -High correlation with vocabulary depth measure observed (r = .89) -Moderate correlation with reading comprehension measure (r = .68)

7.3 Discussion of main findings

7.3.1 Introductory remarks

Before I discuss the main findings in relation to the three research questions, it is worth emphasizing the merits involved in the test development process. Although attempts have previously been made at constructing discrete collocation tests, notably Bonk (2001), Mochizuki (2002), and Barfield (2003, 2006)[39], it seems that the present project is one of the more comprehensive endeavours yet undertaken[40]. There are several reasons for why this is the case. Firstly, previous studies have consisted of one-off attempts where an initial version of a test (barring pilot tests) has not undergone further development and validation. In contrast, the present thesis reports a series of seven studies in which various aspects of validity and reliability were investigated with regard to the COLLEX and COLLMATCH tests.

Secondly, the number of students tested in the present thesis is in most cases larger than in previous studies. For example, 98 informants participated in Bonk's (2001) study, 93 in Barfield's (2003, 2006), and 82 in Mochizuki's (2002). These numbers are no doubt respectable, but they are lower than those in studies 3 (103), 4 (188), and 6 (269) in this thesis. Even though my informants did not represent a true random sample, sample sizes are important in test development, especially when it comes to reliance on item analyses.

A third point has to do with the circumstances under which the test administrations were conducted. In the two major test administrations (studies 4 and 6), the test data from university informants were collected as part of an end-of-term vocabulary exam. In many cases, failing the vocabulary exam meant that they were not allowed to continue to study at the next level. The fact that the test battery was administered under such high-stake conditions means that I can be sure that the students were highly motivated to do their best. It stands to reason that a lack of such motivation is a highly problematic factor when doing empirical research.

Having highlighted some of the conditions under which the research in this thesis was undertaken, and possibly why it is unique, I will now continue by discussing my findings. This discussion will be structured around the three research questions.

7.3.2 Research question 1

7.3.2.1 Introduction

On the face of it, research question 1 (RQ1) can be answered either in the affirmative or the negative. There are however five 'subcomponents' to the question to consider. First, we need to take test format into account (receptive test), and also the construct (receptive collocation knowledge). Furthermore, we need to look at the potential evidence for reliability and validity that has emerged, and also the effectiveness of the test when used with the specified type of

[39] In addition, a number of studies have been reported in which some sort of elicitation tool of collocation knowledge was developed for experimental purposes, but not as a proper test, e.g. Channel (1981), Biskup (1992), Bahns & Eldaw (1993), Farghal & Obiedat (1995), Granger (1998), Schmitt (1998b), Gitsaki (1999), and Staehr Jensen (2005).

[40] Ambitious test development studies can also be found in Vives Boix (1995) and Wolter (2005), but these are not explicitly focusing on collocations.

informants. No doubt, this makes the question very complex and comprehensive. For this reason, it makes sense to try to synthesize the findings relevant to the subcomponents into an overall attempt to answer the question. In fact, it is possible to subsume the four subcomponents under two main ones: reliability and validity. A test format is intimately linked to both of these aspects, and so are the questions of construct and the targeted test-taker group. Consequently, I will in this section discuss all these points under two main headings: reliability and validity, respectively. As will become clear, though, reliability and validity are not opposite poles, and the discussion of one often touches upon aspects of the other. Furthermore, in terms of validity, Alderson *et al.* (1995) have stressed the fact that test validity is relative rather than absolute. This means that an interpretation must be made about the degree of relative validity that must be present for a particular test use. In the following subsections, I will discuss the findings of the empirical studies of this thesis in the light of the following types of validity: concurrent validity, face validity, and content validity. When it comes to construct validity, following Messick (1989), this is viewed as embracing all the other types of validity, and aspects of construct validity will consequently be addressed under these respective sub-headings.

7.3.2.2 Reliability

7.3.2.2.1 Reliability and its relevance to construct validity

The development of the COLLEX and COLLMATCH tests in this thesis has been guided by the assumption that it is possible to treat receptive collocation knowledge as a single, independent construct, and that it is possible to develop discrete tests of such knowledge. In general, if a test is aimed at measuring a single construct, empirical evidence supporting this fact should be collected. To this point, Messick (1989:51) argues that internal consistency reliability is relevant construct validity information, and the degree of homogeneity[41] should be commensurate with the level which is theoretically expected for the construct in question. Consequently, if internal consistency reliability can be used as one piece of evidence of construct validity, this is relevant to addressing the issue of receptive collocation knowledge as a single construct in RQ1.

As a first step, then, let us consider the reliability values observed for COLLEX and COLLMATCH in the different studies in this thesis. I used a reliability measure called Cronbach's alpha, which is a coefficient through which internal consistency is estimated (see section 2.5.3). In the reliability column of Table 7.1 above, the reliability levels observed for the different test administrations have been classified with nominal descriptors. These descriptors have been suggested by DeVellis (1991:85) to be interpreted in the scale presented in Figure 7.1 below. In the light of DeVellis's scale, it must be concluded that I have managed to create two tests of receptive collocation knowledge that are capable of producing highly reliable scores. For most test versions, the coefficients observed range between 'very good' and 'excellent'. The exceptions to the rule were COLLEX 1 (study 1) and COLLEX 5 – PILOT (study 5).

[41] A high level of internal consistency reliability is taken to imply a high degree of construct homogeneity.

Cronbach's alpha coefficient	Descriptor (interpretation of quality)
< .60	Unacceptable
.60-.65	Undesirable
.65-.70	Minimally acceptable
.70-.80	Respectable
.80-.90	Very good
> .90	Excellent

Figure 7.1 Scale descriptors for the interpretation of Cronbach's alpha, from DeVellis (1991:85)

These very high levels of internal consistency reliability can be used as empirical evidence of the capability of COLLEX and COLLMATCH to yield reliable test scores in more general terms. However, if the same evidence is to be used also for support of construct validity, then we must try to determine what levels of internal consistency can be expected for COLLEX and COLLMATCH, along the lines of Messick's argument above. To this point, Alderson *et al.* (1995:88) emphasize that the level of reliability expected for a specific test is contingent on the type and the length of the test, and the range of ability of the informants.

It stands to reason that with objective tests like COLLEX 5 and COLLMATCH 3, aimed at measuring a single construct, a very high level of reliability is warranted. This is so because they do not contain parts that are aimed at measuring different subcomponents of a construct, as opposed to, for example, a test of general proficiency, where parts like writing skills, oral fluency, listening comprehension and reading comprehension are more heterogeneous subcomponents of that construct.

Also, with tests between 50 and 100 items, tested on a large student sample (N = 269) consisting of upper-secondary school students, university undergraduates, and native speakers of English (as in study 6), it could be considered more or less expected to arrive at high reliability coefficients. In a comparison with reliability values observed for other receptive vocabulary tests in the literature, some having a more or less standardized test tool status, we find Meara and Buxton's (1987) yes/no test, with a with a KR-21[42] reliability of .91 for a 100-item test[43]; Read's (1993; 1998) WAT, with a KR-20[44] reliability of .92 for a 50-item test[45]; Vives Boix's (1995) Association Vocabulary Test, with a Cronbach's alpha of .85 and .88 for two versions of a 90-item test, and a Cronbach's alpha of .94 for a 120-item test; Nation's (1990; 2001) Vocabulary Levels Test, in a validation study by Schmitt *et al.* (2001), with a Cronbach's alpha ranging between .92 and .96 for two versions of a 150-item test; and Barfield's (2006) 99-item collocation recognition test where Cronbach's alpha was observed at .95 and .96. This pattern implies that values around and above .90 are aimed for. Thus, a lower value would raise questions about the quality of the tests. It should be emphasized, also, that the same test versions (COLLEX 5 and COLLMATCH 3) were observed to attract scores which were reliable at a similarly high level (.86, .91) when administered to a small number of informants (N = 24) assumed to be less heterogeneous in terms of proficiency. In the light of these observations, with reliability coefficients close to .9 the COLLEX 5 and

[42] Kuder-Richardson 21 formula.
[43] The 100 items consisted of 60 real words and 40 non-words.
[44] Kuder-Richardson 20 formula.
[45] In the reported version of the WAT, each of the 50 items contains 8 choices, all which require responses from informants. This means that it is in fact a 400-item test.

COLLMATCH 3 tests display a wholly acceptable level of quality as collocation tests with regard to aspects of reliability.

The high reliability values are probably best explained as a result of many factors, such as item homogeneity, item quality and test length. Item homogeneity was discussed above. In terms of item quality, the fewer ambiguous items, and the greater the discriminatory power, the better is the quality. In terms of test length, there is a trade-off between having a long test, which through its length alone increases the chances of a higher reliability, and having a shorter, more practicable test, since the risk of test fatigue and lapses of concentration is arguably smaller with a shorter test.

7.3.2.2.2 Ceiling effects and consequences for reliability

A finding which is relevant to discuss in connection with reliability is the tendency towards ceiling effects that was observed in the COLLEX scores. In several studies, I observed high mean scores for the university-level informants as a collective, and close to maximum scores for the most advanced university informant groups. Davies *et al.* (1999:19) call attention to the fact that a ceiling effect means that a test "does not discriminate adequately amongst higher ability learners". This means that, in terms of test scores, learners with a very high level of proficiency cannot be rank-ordered in a reliable way. Two learners may hypothetically differ in terms of receptive collocation proficiency, but the test is not sensitive enough to pick up those differences at the very high end of the test score scale: there is simply no headroom.

The reason behind the observed ceiling effect in COLLEX scores probably lies in the combination of very advanced informants and the test format itself. In terms of the former, we have seen in many studies that although groups of native speakers of English outperform Swedish university student groups, both on COLLEX and COLLMATCH, the differences between Swedish 3rd term university students of English and British university-level students are not very big (though statistically significant). This could be seen in scores on the Vocabulary Levels Test as well, where the Swedish 3rd term students' mean score corresponds to 92 per cent, and the British students' mean score corresponds to 96 per cent of the maximum score (see section 5.2.3.3.2). In terms of the test format, it seems that the receptive recognition, multiple-choice task in the 50-item COLLEX is slightly too easy for the most advanced students. Although attempts were made to remedy these effects, and some improvements were made, they did linger also in the most recent version (COLLEX 5). It is quite likely that a productive format would have been more difficult for these learners.

7.3.2.2.3 Inherent limitations of reliability estimates in Classical Test Theory

Despite the largely positive interpretations when it comes to test reliability, certain constraints exist when it comes to estimates within Classical Test Theory. One important qualification that needs to be made with regard to the observed reliability is that the values are intimately linked to the scores produced by the sample of informants on which the tests were trialled. This is an inherent drawback of reliability measures within the theory framework. As pointed out by Alderson *et al.* (1995:89): "The examinees' characteristics and the test characteristics cannot be separated…". It is consequently not possible to claim that COLLEX and COLLMATCH are reliable tests per se, but it is possible to claim that they are tests which have been empirically shown to be capable of producing reliable test scores with a certain

informant sample. This sample has been identified as consisting of the range between Swedish upper-secondary school and university-level students of English.

A model that does separate between characteristics pertaining to test takers and characteristics of a test per se is the item response model, found within the theory referred to as Item Response Theory (IRT) (Alderson *et al.* 1995; Bachman 2004). Roughly speaking, IRT models, like Rasch (see Henning 1987), make assumptions about the relationship between a test taker's ability and his or her performance on a specific test item. More specifically, the models assume that a test taker's response is determined by two factors: 1) the test taker's ability on an underlying trait, and 2) the characteristics of the items (Bachman 2004:141). Different models display different levels of complexity with regard to how many parameters they can handle: one, two, or three, corresponding to item difficulty, item discrimination, and guessing (Brown & Hudson 2002). Thus, with an IRT model, reliability can be estimated independently of the group of informants used, and this is a great advantage over CTT models.

Considering the discussion earlier about the potential guessing behaviour of certain learner groups being problematic, the application of the three-parameter Rasch model to the COLLEX and COLLMATCH data could no doubt provide useful information. Unfortunately, the application of an IRT approach to reliability has fallen beyond the scope of the present thesis, and it should be noted that a very large data set is recommended for Rasch analyses. Henning (1987:116) suggests samples of 100-200 informants for the one-parameter model, 200-400 informants for the two-parameter model, and as many as 1,000-2,000 informants for the three-parameter model. This must be considered if IRT models of reliability are to be used.

7.3.2.3 Validity

7.3.2.3.1 Concurrent validity

Concurrent validity denotes either the extent to which a test can be seen to correlate with another variable which is supposed to measure the same construct, or to the comparison of two or more groups of test takers differing in level of language proficiency. One type of analysis involved a correlation between COLLEX and COLLMATCH. Since both tests were constructed as tests of receptive collocation knowledge, a high correlation between them was expected. In three studies, very strong correlation values were indeed observed, as illustrated in Table 7.2 below.

Table 7.2 Correlation coefficients (Pearson *r*) observed between different versions of COLLEX and COLLMATCH.

Study	Test versions	Correlation I	N
4	COLLEX 4 and COLLMATCH 2	.92**	188
6	COLLEX 5 and COLLMATCH 3	.86**	269
7	COLLEX 5 and COLLMATCH 3	.89**	24

** Correlation is significant at $p < .01$, one-tailed.

Judging from these results, concurrent validity support for COLLEX and COLLMATCH as tests of receptive collocation knowledge has been empirically demonstrated. This conclusion would then be based on the assumption that the two tests measure the same construct. The fact that a perfect correlation was not observed could be seen as a positive outcome. Even though both tests are aimed at measuring receptive collocation knowledge, the tasks in the tests are slightly different.

In a further analysis, COLLEX and COLLMATCH scores were correlated with yet another measure of receptive collocation knowledge, namely the syntagmatic part of the WAT (Read 1993, 1998). The WAT is one of the more widely used depth tests, and Bachman claims that one of the first questions test users are likely to ask about a test is whether it is correlated with some standardized test (1990:249). As was described in Chapter 6, I decided to use the WAT as a more or less standardized concurrent criterion validity test. Since 160 out of 320 items tap knowledge of potential syntagmatic links between a target word and four associate words, this set of items could arguably be used as a concurrent validity measure. Significant correlations were observed between the two collocation tests (COLLEX 5 and COLLMATCH 3) and the syntagmatic part of the WAT at r = .84 and .86, respectively.

From a concurrent validity perspective, these correlations are no doubt positive. However, there is also an alternative interpretation. This has to do with the fact that all measures involved share the same test method: they are all multiple choice tests. Campbell and Fiske (1959:83) argue that:

> Reliability is the agreement between two efforts to measure the same trait through maximally similar methods. Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods.

The following figure adopted from Bachman (1990) illustrates Campbell and Fiske's argument in a clear way:

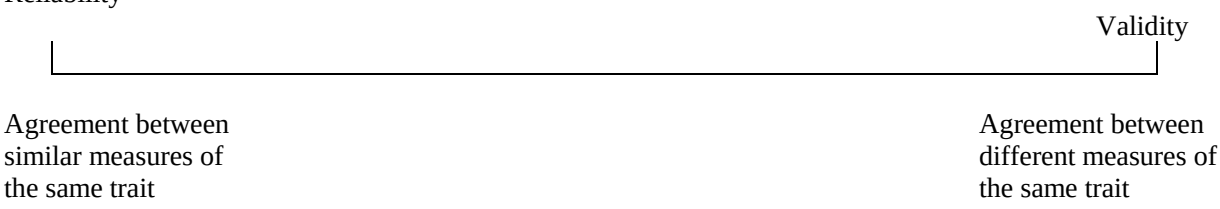

Figure 7.2 Relationship between reliability and validity (adopted from Bachman 1990:240).

If applied to the current discussion, using very similar test methods in COLLEX, COLLMATCH, and the WAT would mean that we are in fact dealing more with reliability than validity aspects. In terms of COLLEX and COLLMATCH, there are similarities in test content as well as test method. Furthermore, the tests were administered after each other under the same conditions. I will come back to the issue of test method later in this section. At this stage, though, we should consider the possibility that these factors played a role in the high correlations. A remedy to this would have been, for example, to administer COLLMATCH as an aural test, with the test items read out to the test takers, who then respond using an answer sheet, whereas COLLEX would be administered only as a paper-

and-pencil test, with all information to be processed in a written medium. However, this design would not be without problems, since listening comprehension skills would become a crucial factor. Thus, even though Campbell and Fiske's distinction between reliability and validity is thought-provoking, administering tests of the same trait (construct) through maximally different methods brings with it the possible interference of what Messick (1995) calls construct-irrelevant variance.

Before the discussion pertaining to concurrent validity aspects is rounded off, the correlations between COLLEX and COLLMATCH and WAT scores need to be revisited. I concluded above that the significant correlations between COLLEX 5 and COLLMATCH 3, and the syntagmatic part of the WAT were positive in terms of concurrent validation. However, in the same analysis correlations between COLLEX and COLLMATCH, and the paradigmatic part of the WAT were computed, and high correlations in the same region were observed between these measures (r = .81 and .88, respectively). This finding could at first sight be taken as a counterclaim to concurrent validity for COLLEX and COLLMATCH as tests of receptive collocation knowledge as a single construct. However, I think there are a number of explanations which should prevent us from sticking unconditionally to this initial conclusion. Firstly, the correlations between COLLEX 5 and COLLMATCH 3 and the whole 320-item WAT were .85 and .89. With such high correlations with the whole test, it is not surprising that we observe similarly high correlations with both the two halves.

Secondly, since the two halves in the WAT correlated highly with one another (r = .91), high correlations between both these and COLLEX and COLLMATCH would also be expected. Relevant to the high correlations with both the syntagmatic and the paradigmatic parts, Qian (1999:299) has suggested that knowledge of word meaning, here interpreted to denote the meaning of single orthographic words, has an impact on knowledge of collocation. This is intuitively appealing. For example, a learner who knows that the verb *run* means not only 'to move rapidly by using one's legs', but who has also discovered through exposure a meaning of *run* which amounts to 'to manage something', is perhaps more likely to recognize *run a business* as an English collocation, than a learner whose knowledge is restricted only to the first sense.

Thirdly, in the discussion of chapter 6 I pointed out the fact that some items in the WAT are relatively low-frequency words, and that the WAT therefore is prone to be vocabulary size-dependent. We observed a correlation between the vocabulary size scores (VLT 1) and the WAT scores at r = .93, which could be taken as evidence of this size-dependence. Possibly, the high correlation between WAT scores and VLT 1 scores could be partially explained by the fact that the tasks in the VLT test and the paradigmatic half of the WAT are very similar. In one (the WAT), an English target word is to be matched with up to four other English words which can be used to define the meaning of the target word, mostly near-synonyms. In the other (the VLT), in blocks of three, English target words are each to be matched, out of six choices, with a word, phrase or sentence that can be used to define the meaning of those target words. This fact, coupled with the aforementioned size influence, is a potential cause of the results we have observed. A consequential, and admittedly radical, view would then be that the paradigmatic half of the WAT functions as a vocabulary size test, but without the systematic sampling of words from certain frequency bands to certain word levels in the test. If we put this argument together with the arguments made earlier, i.e. that vocabulary size is an important factor in collocation recognition, then the strong correlations

between COLLEX, COLLMATCH, the VLT, and the WAT follow logically from construct interrelatedness and aspects of test design, such as test task and test content sampling.

From another perspective, it could be the case that a general underlying language ability causes all these variables to correlate. Also, as touched upon earlier in this section, test method might have played a role. This claim needs to be unpacked. Consider Figure 7.3 below, which is inspired by Bachman (1990). In the figure, the arrows beneath the four boxes are intended to illustrate that the four variables presented in the boxes all correlate highly with one another. Furthermore, an underlying ability together with a common test method are seen to affect the performance on and correlation between the four measures: the WAT (vocabulary depth), COLLEX (receptive collocation knowledge), COLLMATCH (receptive collocation knowledge), and VLT (vocabulary size), as indicated by the arrows going from the ellipses to the boxes.

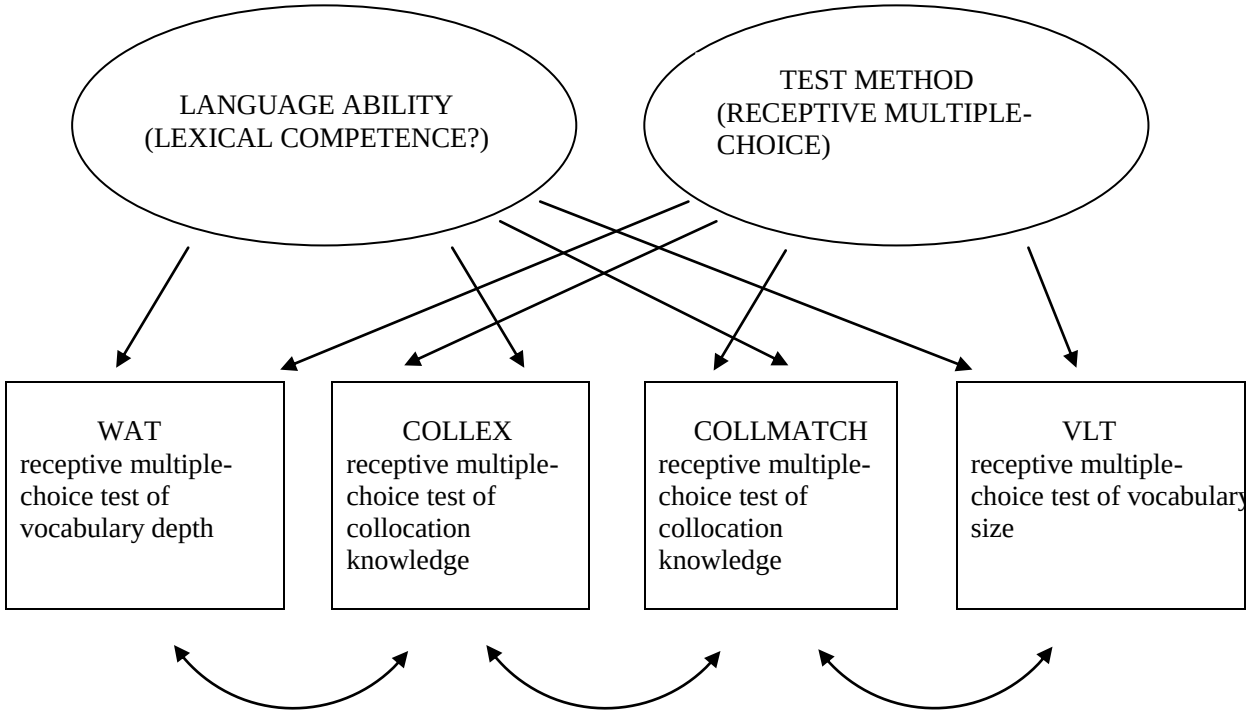

Figure 7.3 Correlations between four tests potentially caused by an underlying ability and test method.

A potential contender for the role as underlying ability is lexical knowledge. This can be seen as a comprehensive construct that comprises the enumerated subconstructs. The test method effect would lie in the fact that all four measures were multiple-choice, paper-and-pencil tests of receptive knowledge. Even though test method effects cannot explicitly be ruled out, it is unlikely that they played a major role. A more dominant factor was possibly an underlying trait, such as lexical knowledge.

An additional, intriguing fact is that a fifth variable was correlated with the four tests, namely reading comprehension. With this variable, however, as is evident from Table 6.5 in Chapter 6, the four measures correlated less strongly (between .64 and .80, all significant). There is one obvious explanation for this: that the reading comprehension test indeed measures a different construct, and also that it is a different test format. The levels of

correlation between, for example, the VLT scores and reading comprehension (.69) were in line with correlations that are commonly reported in the literature: between .66 and .75 reported in Thorndike (1973); .74 reported in Qian (2002); between .79 and .85 reported in Henriksen *et al.* (2004). On the whole, if we see the VLT, the WAT, COLLEX and COLLMATCH as all being closely related to some underlying lexical knowledge, then the observed high correlations between them, and the generally lower correlations observed between them and reading comprehension, could in fact be interpreted positively, since arriving at very similar correlations between all five variables would have left us with a result very difficult to interpret.

In sum, in the light of the above discussion, my interpretations of the results from the concurrent validity studies in the present thesis are predominately positive. High positive correlations between COLLEX and COLLMATCH, and other vocabulary constructs, such as size and depth, are not seen as overly problematic, but are argued to stem from construct interdependence and the inclusion of low frequency words in all the measures. The possibility of an underlying trait, such as lexical knowledge, causing the high correlations could not be ruled out, and some effect may have stemmed from test method aspects, even though the role of the latter was considered having minor importance.

7.3.2.3.2 Face validity

Face validity denotes the extent to which a test measures what it is supposed to measure, in the eyes of untrained observers, such as the test takers themselves. Aspects that may be judged by these types of observers are for example the test as a whole, specific test items, instructions, and time limits (Alderson *et al.* 1995).

The use of the term face validity and investigations thereof are not favoured in all camps of the language testing field. Bachman (1990:285-289) accounts for a large collection of critical voices against the notion. The gist of these criticisms seems to be that face validity is claimed to be unscientific and irrelevant. Cronbach (1984) even goes so far as to compare the unconditional acceptance of a test, based solely on its appearance as reasonable to the lay person (face valid), to the historical workings of phrenology, graphology, and tests of witchcraft. Indeed, it does seem wise not to rely only on face validity when investigating the quality of a test. However, as one piece of the validity puzzle, there is in my opinion a place for the investigation of face validity in test development projects. Consequently, face validity must be investigated, not in lieu of other aspects, but in conjunction with them. I thus concur in the opinion of Alderson *et al.* (1995:173) who argue that face validity is important in testing:

> For one thing, tests that do not appear to be valid to users may not be taken seriously for their given purpose. For another, if test takers consider a test to be face valid, we believe that they are more likely to perform to the best of their ability on that test and to respond appropriately to items. In other words, we believe that face validity will affect the response validity of the test.

The quote above emphasizes the link to reliability. If test scores do not reflect a test taker's true ability or knowledge, then these scores cannot be interpreted as valid indicators of that knowledge, and there is also a risk that low reliability values will follow. Consequently, face validity information is certainly relevant information. Data to this point were collected in

study 5 (section 5.1) from teacher students of English at university level, representing a fair range of general English proficiency levels. In terms of formal learning level, they had studied English for two and a half terms (almost 1.5 years), in addition to eight years at school. The data were collected in a structured way through a questionnaire, and also anecdotally throughout the series of studies, by taking informants' comments from different test administrations into account when deliberating changes to the tests.

On the whole, the results from the questionnaire data were positive in terms of face validity (see Table 5.7 in section 5.1.3.2.2). Based on descriptive means, the informants as a collective stated that the instructions of the tests were very easy to understand, that the levels of difficulty were average to easy, and that the tests appealed to them. Had the informants' mean judgements been considerably lower, for example if they collectively had stated that test instructions were unclear, that the tests themselves were boring, and that they were either very difficult or very easy, then this would have had serious consequences for the validity of the tests.

Firstly, unclear instructions are highly undesirable, since very different opinions about what the test task requires a test taker to do could potentially lead to unreliable scores. However, unclear instructions could probably be remedied in a fairly straightforward way.

Secondly, if the tests had been perceived as boring by the informants, then the cause of this would have had to be investigated. Among the possible reasons we could conjecture, for example, test length and task complexity. A too long test coupled with a test task that is not demanding enough would be negative in this regard.

Thirdly, a test that is either too easy or too difficult could not have been rectified without considerable changes to the test, and subsequent trialling of new versions would have had to follow. The fact that the students perceived the tests to be average in terms of difficulty meant that students at lower learning levels, such as university first term students, would potentially find the tests challenging, as would upper-secondary school students. At the same time, Swedish near-native speakers of English, as well as native speaker of English, would probably find them easy, and score close to the maximum score. The results obtained in terms of perceived difficulty of COLLEX and COLLMATCH on the part of the informants were therefore positive.

Whether the results from the open-ended question in the questionnaire were univocally positive is difficult to say. The informants were asked to state what kind of knowledge they thought was measured in the test. As was clear from the account of the results in section 5.1.3.2.2, many answers could be straightforwardly linked to the construct of collocation knowledge (10 and 13, respectively for COLLMATCH and COLLEX), which is positive. However, some answers also alluded to general proficiency, and language aptitude (5 and 6, respectively for COLLMATCH and COLLEX). This is not necessarily negative, since many target collocations tested in COLLEX and COLLMATCH involve seemingly arbitrary restrictions on lexical items. For example, in terms of delexical verbs, why do we say *do justice* but *make progress* and *take measures*, and not **take justice*, **do progress*, and **make measures*?

In hindsight, in terms of methodology, it is possible that a different question should have been asked. At the time when the study was carried out, I deliberately wanted to avoid using the term 'collocation', because I did not expect all informants to be sufficiently familiar with it. However, I could have given a definition of receptive collocation knowledge, such as the one suggested in Figure 2.3 (section 2.4.2), possibly accompanied by the working definition

of collocation given in Figure 2.2 (section 2.3.10), and then asked whether the informants thought that COLLEX and COLLMATCH constituted good tests of the knowledge of such linguistic items. This could possibly have yielded even better information relevant to the face validity of COLLEX and COLLMATCH.

7.3.2.3.3 Content validity

The summary in Table 7.1 above shows that evidence of many different kinds of validity has been observed: response, construct, concurrent, and face validity. In hindsight, however, one type of validity has been sparsely addressed thus far: content validity. For this reason, it needs to be discussed at some length here. Content validity denotes the extent to which a test is relevant to a given area of language content or language ability. Thus, the question at hand is whether the test content, in our case the collocation test items, is adequate and representative of the larger universe of items (target domain) of which the test is assumed to be a sample. The adequacy hinges on an *a priori* description and definition of the construct to be measured. Only then can judgements be made about whether the items of a test fit the specified construct. The representativeness is closely linked to whether potential aspects of a construct are covered in suitable proportions, for example if a construct consists of several subdomains which need to be tested. In Chapter 5 (section 5.1.2.2), based on an interpretative argument model proposed by Kane *et al.* (1999), I argued that an inference must ideally be possible from observed scores from my tests to a so-called universe score. An issue central to this possibility, and to content validity in general, is the method of item selection.

The item selection methods used in the different versions of COLLEX and COLLMATCH developed were all based on a word knowledge framework (Nation 2001). This made the selection approach taken into a word-centred approach, as opposed to a more holistic approach, in that single words were used as a point of departure. For example, the items for COLLMATCH 2 were all compiled by starting with twenty high-frequency verbs, all taken from the first thousand most frequent words of English, and subsequently selecting noun collocates of these verbs according to corpus data from the BNC. Another example can be seen in the selection of items for COLLMATCH 5, where 100 verbs functioned as a point of departure, and similarly noun collocates were then selected for each of these verbs based on corpus data. The underlying assumption was thus that collocation knowledge can be measured as a property of single words, i.e. that collocation knowledge is based on knowledge of single words in a language, and that these words in turn may be combined with certain other words in that language. For example, in natural language, the delexicalised, high-frequency verb *make* collocates with a large number of object nouns. By sampling some of these object nouns for a test, and asking informants if they recognize the combination of the verb + NP, I assumed that I could probe the knowledge that those informants have about the combinatory potential of the selected words, more specifically the knowledge of collocations. By trying to restrict item selection to higher-frequency words (an issue which will be discussed in the next section), which the informants were expected to know minimally in terms of a basic form-meaning mapping, I furthermore assumed that informants' collocational knowledge of these words could be mapped out.

However, on reflection, these assumptions, and the item selection methods used, are fraught with restrictions. The overall problem is that scores on COLLEX and COLLMATCH cannot be straightforwardly extrapolated into "scores" for the target domain, i.e. the universe of English collocations consisting of high-frequency verbs + nouns. Especially for COLLEX,

in which two or three combinations of verb + noun are juxtaposed, and an informant is asked to choose one of these over the others on the basis of it being a frequent, conventionalized collocation, it is difficult to say exactly how scores reflect overall collocation knowledge. The 50-item COLLEX 5 is essentially a measure of how capable learners are at identifying a conventionalized English verb + NP collocation when simultaneously presented with competing word combinations which may distract them. The distraction is linked to potential influences from the L1 (Swedish), or L2 forms which are in themselves intuitive alternatives to the intended target collocation, but which native speakers of English refrain from using due to mere convention. In comparison, in COLLMATCH 3, informants are presented with 100 word combinations out of which 70 are intended target collocations, and 30 so-called pseudo-collocations (distractors). The format is essentially a yes/no test, and as opposed to the COLLEX test, informants are required to make a judgement about each word combination in isolation, in the sense that no distractors exist in the item itself. The cognitive process can rather be seen as a matching between the test item, and the array of structures and meanings in the mental lexicon. Also, the cognitive process involved in responding to an item might involve either recall of stored whole combinations, or a word-for-word analysis.

In relation to the *a priori* specified construct, I am confident in claiming that COLLEX and COLLMATCH consist of valid content in terms of adequacy. Even though no study was conducted in which language test experts were asked to analyse the test content, a method which seems to be one standard way of investigating content validity (see e.g. Brown 1983; Bachman 2004). In my opinion there is a strong case for arguing that the items in the tests are collocations, as defined in Chapter 2. Empirical support for this claim can be seen in the high reliability coefficients observed for the two tests, a fact which Weir (2005:23)[46] generally takes as evidence of consistency in terms of "content sampling". What seems to be lacking, though, when it comes to the content validity of COLLEX and COLLMATCH scores, is a more systematic and *proportional* selection of items in the light of a different model than the word knowledge framework based on individual words. Meara and Wolter (2004) have argued that word-centred approaches should be abandoned in favour of network-based approaches, where more holistic measures of mental lexicons are used (lexical organisation), rather than more and more detailed measures of individual words.

Another approach which could possibly take us a bit further would entail creating a frequency list of all verb + NP combinations, for example in the 100-million word BNC corpus[47], and then using a stratified random sampling technique for selecting test items, a technique commonly used for vocabulary size tests (see e.g. Schmitt *et al.* 2001). For example, 30 collocations could be sampled from each frequency band of a thousand word combinations between 1K and 5K, for a test of a total 150 items. Test formats involving different tasks, such as L2 to L1 translation and L2 collocation recognition, could be developed. Such an approach would in all likelihood presuppose a manual analysis of all the word combinations on the frequency list, so that, for example, pure idioms and free combinations could be discarded. The advantage of such an approach would be its desirable measurement characteristics. Just like scores on a vocabulary size test, the result on the sampled items from each frequency band level could then be extrapolated to roughly reflect knowledge of all the 1000 items in the frequency band.

[46] Weir prefers the term 'scoring validity' to reliability.

[47] Even more ideal would a sample which combines data from both a British English corpus, like the BNC, and an American English corpus, like the American National Corpus (ANC).

Even though a straightforward and immaculate extrapolation from COLLEX and COLLMATCH scores to universe verb + NP collocation scores is not possible, this does not mean that COLLEX and COLLMATCH scores are meaningless and without predictive power. We saw in several studies that groups of Swedish informants performed quite differently from each other, and significant differences existed between these groups of learners and native speakers of English throughout. For example, in study 4 (section 4.2) upper-secondary school students scored around 60 per cent on the 50-item COLLEX 4, whereas second and third term university students scored a mean corresponding to around 90 per cent (see Table 4.18). Almost the exact same pattern was repeated in the 100-item COLLMATCH 2 scores (see Table 4.21). In the same way, we saw in study 6 (section 5.2) that a sizeable group of native speakers of English (N = 34) performed significantly better than groups of Swedish students on both COLLEX 5 and COLLMATCH 3. On COLLEX 5, the native speaker group scored a mean corresponding to around 98 per cent, whereas the Swedish university student groups scored means of around 92, 85, 82, and 60 per cent, with scores decreasing as a function of lower learning level[48] (see Table 5.14). Again, the same pattern was visible in COLLMATCH 3 scores (see Table 5.17). It would thus be too defeatist to conclude that COLLEX and COLLMATCH scores do not reflect some sort of underlying receptive collocation knowledge, despite the above identified shortcomings.

Furthermore, in a concurrent validation analysis reported in Chapter 6, COLLEX 5 and COLLMATCH 3 scores were observed to correlate highly with another measure of receptive collocation knowledge scores (the syntagmatic part of the WAT, at r = .84 and .86). On the face of it, this was taken as construct validity support. However, the design of this criterion measure, the WAT (Read 1993, 1998) suffers from the same kind of flaws that COLLEX and COLLMATCH were identified with, in terms of content validity (see discussion section in Chapter 6). Furthermore, the frequency of certain words occurring in COLLEX, COLLMATCH, and the WAT are to some extent relatively low, which makes the potential influence of vocabulary size, at least theoretically, a tangible problem.

7.3.2.4 Answering RQ1

On the whole, both COLLEX and COLLMATCH produce highly reliable scores, as estimated through Cronbach's alpha (~ .90). This means that the amount of measurement error is acceptably low.

As to the construction of receptive tests of collocation knowledge, the findings in this thesis show that COLLEX and COLLMATCH appear to function well in the following respects: a) they are quick to sit, b) they involve simple test tasks, c) they appeal to test takers, d) they are easy to score, and d) yield minimally interval data. Samples of Swedish upper-secondary school students and university-level students, as well as university-level native speakers of English, were tested and both tests discriminated well between these categories of students. A problem, though, was experienced in terms of ceiling effects in COLLEX. This places restrictions on its power to discriminate between very proficient informants at the near-maximum score range.

Support for test validity was gathered in many different ways. Response validity was established through the observation of high reliability estimates. Prerequisites of construct validity were created through *a priori* theoretical and operational definitions of receptive

[48] A lower learning level means fewer terms of classroom exposure to English.

collocation knowledge as a construct. Empirical evidence was found in very high reliability estimates, interpreted as test item homogeneity, and through a non-equivalent groups design (Bachman 2004) where native speakers of English outperformed Swedish university students, and Swedish university students outperformed Swedish upper-secondary school students. Concurrent validity was established through strong correlations with the collocation part of the Word Associates Test (Read 1993, 1998), and through strong correlations between COLLEX and COLLMATCH themselves.

With regard to construct independence, certain problems were experienced when it came to observed concurrent validity values between COLLEX and COLLMATCH and other tests of lexical knowledge. Very strong correlations were observed with tests of vocabulary size and vocabulary depth. However, these strong correlations were believed to stem from construct overlap and interdependence, and properties of the design and characteristics of these tests, and it was concluded that this did not in effect pose a threat to the construct independence of COLLEX and COLLMATCH as tests of receptive collocation knowledge. A lower correlation between COLLEX and COLLMATCH, and reading comprehension scores was interpreted as a concurrent validity counter-claim.

Face validity was established through the administration of a questionnaire which collected informants' judgements on test appeal, test instruction clarity, test difficulty, and test construct, all of which gave satisfactory support. Prerequisites of content validity were created through careful definition of collocation as a linguistic unit, but the word-centred method of item selection raised questions to do with restricted extrapolation of test scores to universe scores.

In the light of these observations, I am now in a position to answer RQ1, which read:

Is it possible to develop receptive tests measuring English collocation knowledge as a single construct, capable of yielding reliable and valid scores, for use with advanced Swedish learners of English?

On balance, I would like to argue that the answer to RQ1 lies considerably closer to the affirmative than the negative. I have succeeded in constructing reliable tests of receptive collocation knowledge, capable of measuring this knowledge as a single construct with advanced Swedish learners of English. The degree of overall validity is judged to be fully acceptable for the use of COLLEX and COLLMATCH as proficiency tests, even though I did identify certain problems in terms of generalisability of test scores to universe scores, and ceiling effects. With validity being a perpetual process, improvements in this regard can be made.

7.3.3 Research question 2

7.3.3.1 Introduction

Research question 2 (RQ2) concerns the nature of the relationship between EFL learners' vocabulary size and their receptive knowledge of collocations. This question was addressed through correlating scores on COLLEX and COLLMATCH with scores on a vocabulary size test (VLT) in a series of studies.

7.3.3.2 Correlations with vocabulary size

A summary of the correlation coefficients observed in these studies is given in Table 7.3 below. Judging from the results of the correlation analyses, where significant, positive correlations ranging between .83 and .90 were observed, there is a strong relationship between vocabulary size, as measured in the VLT, and receptive collocation knowledge, as measured in COLLEX and COLLMATCH, respectively.

Table 7.3 Correlation coefficients (Pearson *r*) observed between different versions of COLLEX and COLLMATCH, and Vocabulary Size scores (The VLT).

Study	Collocation test versions	Vocabulary Size measure	Correlation	N
4	50-item COLLEX 4	VLT version 1	.87**	188
	100-item COLLMATCH 2	VLT version 1	.87**	
6	50-item COLLEX 5	VLT version M	.88**	269
	100-item COLLMATCH 3	VLT version M	.83**	
7	50-item COLLEX 5	VLT version 1	.90**	24
	100-item COLLMATCH 3	VLT version 1	.90**	

** Correlation is significant at $p < .01$, one-tailed.

The results shown in Table 7.2 give rise to a number of follow-up questions: a) is this a threat to the concurrent validity evidence discussed in section 7.3.2.2?; b) how can we explain the high correlations observed?; c) Are there any limitations to the data?

Relevant to the first question (a), Bachman has emphasized the need to show not only that a certain set of test scores, aimed at measuring a given language ability, correlate with other indicators of that same ability, but that they do not correlate with measures of other abilities (1990:250). In a strict interpretation of this decree, if we see vocabulary size as an example of a different ability, or as a measure of a different construct, and we observe correlations on a par with the levels summarized in Table 7.2 above, this would in effect be a counterclaim to existing concurrent validity of COLLEX and COLLMATCH as tests of receptive collocation knowledge. However, I will argue here that there is a strong reason for why this interpretation is precipitated. It could be argued that vocabulary size more or less determines receptive collocation knowledge in the written form, because single words are the "building stones" of collocations, in a strict orthographic sense. The inherent words of a collocation must be processed as meaningful linguistic units. Also, an informant who has a large vocabulary size can be assumed to have had a great deal of exposure to English. Thus, in the same way as reading comprehension has often been found to correlate highly with vocabulary size (Hazenberg & Hulstijn 1996; Qian 1999, 2002; Henriksen *et al.* 2004), as has general language proficiency (Meara & Jones 1990; Laufer 1997), arguably because single words are

the building stones of texts and discourse, it is not surprising that a relationship exists between single word vocabulary size and receptive collocation knowledge. Thus, vocabulary size is so closely related, conceptually, to collocation knowledge, that it does not serve as a cogent counterclaim. In this sense the relation is the same (or similar, rather) as that between vocabulary size and reading comprehension. The high correlation between these two constructs does not depreciate their respective existence as separate, but interdependent, constructs.

The question is however, how we can explain the correlation on a more technical level. One relevant factor is that COLLEX and COLLMATCH do contain a smaller number of words which are strictly not high-frequency. In reference to Appendices 5J and 5L, if we treat words lower than the 3K band as outside the high-frequency range, we find 14 words in COLLEX 5, and 20 words in COLLMATCH 3 which could be problematic[49], in the sense of a basic core meaning not being known, by the less advanced students for which the tests are aimed (i.e. upper-secondary school students). The high correlation could then be ascribed to the potential need to know some low-frequency words to be able to make an informed choice in the tasks used in COLLEX and COLLMATCH. However, even in cases where it is beyond reasonable doubt that students did know the words featured in COLLEX and COLLMATCH, for example in the case of the native speaker group in study 6, positive correlations were still observed between vocabulary size scores and COLLEX and COLLMATCH scores. With a mean of 143.6 on the VLT, out of a maximum of 150, significant correlations were observed at .43 and .57. These are admittedly lower than the ones observed for the different Swedish learner groups (see Chapter 5, Table 5.19), especially .43 for COLLEX, but the level is likely to have been reduced by the extremely small variance (M = 48.9, S.D. = 1.0) in the COLLEX scores. Consequently, these facts indicate that there is a relationship even when vocabulary size is controlled for. Further indications of the relationship can arguably also be seen in the correlations between vocabulary size scores and COLLEX and COLLMATCH scores for the most advanced Swedish informant groups, the third-term university students of English. This group scored a mean of 138.3 on the 150-item vocabulary size test, and there were correlations of .69 and .75 between COLLEX and COLLMATCH, and these VLT scores.

However, even if we have evidence to suggest that a large vocabulary facilitates verb + NP collocation recognition in COLLEX and COLLMATCH, it is not necessarily the case that knowledge of individual orthographic words leads to recognition of a collocation made up by these words. As has been shown in a recent study[50], Barfield (2006:199) found that knowledge of individual verbs and nouns did not in all cases entail recognition of their combination in a verb + noun collocation. Similar results have also been observed by Channel (1981) in a small-scale study of eight advanced students of English. Barfield (2006:342) suggests that his results could indicate that the L2 mental lexicon works in part from individual lexical items rather than lexical combinations, such as collocations. I will come back to this hypothesis in my discussion of the findings in relation to research question 3.

[49] The words in COLLEX 5 are: *commit, comply, employ, polish, sweep, pose, lodge, stroll, apologies, revenge, fuse, clench, fell,* and *heed*; The words in COLLMATCH 3 are: *impose, employ, commit, launch, assess, abandon, dismiss, justify, bind, sustain, cease, grace, objection, assistance, dispute, sin, approval, queue, thunder,* and *say.*

[50] This study appeared as the present thesis project was nearing its completion.

Even though we have established strong links between vocabulary size and receptive collocation knowledge, as measured in COLLEX and COLLMATCH, there might be certain limitations to my data. One issue is the use of the VLT as a test of vocabulary size, in particular with the student groups tested in the present thesis. Despite the fact that the VLT is generally seen as a proper size test, in its original versions (Nation 1983, 1990), it was not designed as such. No technical evidence of reliability and validity for this use was originally reported (Read & Chapelle 2001). For the versions used in the present thesis (Schmitt 2000; Nation 2001), however, such data have been presented (Schmitt *et al.* 2001). The drawback here, though, is the rather big gap between the word levels. With levels like 2K, 3K, 5K and 10K (+ academic word level), many informants at an advanced proficiency level performed well on the higher frequency word levels (2K and 3K). In many cases they also performed well on the 5K level. Their performance on the 10K level, however, indicated that they had problems with low frequency words. I could thus conclude that many students possessed a minimum vocabulary size of 5,000 words, but that they did not know all the words on the 10K level. What is interesting is to find out what is going on between the 5K and the 10K levels, a size range of as many as 5,000 words. With the current design, then, the VLT is a rather crude measure. Ideally, with the advanced learners tested in this thesis, a clearer picture of their knowledge of 6K, 7K, 8K, and 9K words is needed. Despite the fact that for most learners a sloping curve exists, with fewer words known for each lower frequency band, it is in theory possible that a slightly different pattern could emerge if scores on such levels were accumulated into an aggregate score. For example, two learners who score the same on the 10K level, could possess quite different levels of knowledge of the words between 5K and 10K, which would result in different total scores. This could come about, for example, as the outcome of special interests. Nation (2001:20) argues that beyond the high-frequency words of a language people's vocabularies grow as a result of their jobs, interests and specialisation. Many words in the 5-10K range are arguably fairly specialised in nature.

Another point that needs commenting on is the fact that the unmodified version (VLT 1) of the VLT was used in studies 4 and 7, and the modified version (VLT M) was used in study 6. The difference between the versions is shown in Table 7.4.

Table 7.4 comparison of VLT 1 and VLT M composition

Frequency band	VLT 1 number of words	VLT M number of words
2K	30	-
3K	30	30
AC	30	30
5K	30	45
10K	30	45
(Total)	(150)	(150)

As can be seen in the table, the 2K word level was dropped in the VLT M version, and instead levels 5K and 10K were increased in terms of number of items tested. On reflection, in VLT M, for the least advanced learners in terms of vocabulary size, the omission of the 2K level meant that words that they arguably would have had a good chance of knowing were exchanged for more words on the two lower frequency bands, 5K and 10K, with which they

arguably had greater problems. A probable effect of this may have been that in relation to scores on the VLT 1 test, scores on the VLT M test for lower level learners were slightly depressed. For the more advanced learners, however, the augmentation of the 5K and 10K levels meant that there was arguably a finer gradation in the rank order of the scores, with more words capable of discriminating better between these learners. The 2K level of the test would for these advanced learners provide very little discriminatory information. Importantly, though, on the whole the composition of the VLT M test is not believed to have had any compromising effects on the overall rank order of scores, in terms of invalid changes in position between lower ability and higher ability learners.

7.3.3.3 Answering RQ2

Research question 2 addressed the relationship between vocabulary size and receptive collocation knowledge. The question read:

What is the relationship between Swedish EFL learners' vocabulary size and their receptive knowledge of collocations?

In general, it was found that vocabulary size, as measured with the Vocabulary Levels Test (Schmitt 2000; Nation 2001; Schmitt *et al.* 2001), was strongly associated (r = .83 - .90) with receptive collocation knowledge, as measured with COLLEX (versions 4 and 5) and COLLMATCH (versions 2 and 3). If these correlations are squared, we find that the vocabulary size variable explained between 69 and 81 per cent of the variance in the receptive collocation knowledge scores. Since the existence of a small number of lower-frequency words (> 4K) in COLLEX and COLLMATCH may have had an inflating influence on these correlation values, the correlations between these variables for a group of native speakers of English (N = 34) were investigated. Lower but significant correlations were observed also for these informants (r = .43 - .57). Since it is beyond reasonable doubt that these native speakers knew the single words making up the word combinations in COLLEX and COLLMATCH, this was taken as evidence of the fact that a large vocabulary size facilitates the recognition of collocations. Since no previous studies exist in which correlations between vocabulary size and receptive collocation knowledge are investigated, no comparisons can be made.

7.3.4 Research question 3

7.3.4.1 Introduction

Research question 3 addressed the relationship between learning level and receptive collocation knowledge. Since COLLEX and COLLMATCH are capable of producing reliable scores, and, with certain reservations discussed above, also valid scores with regard to the construct receptive collocation knowledge, we have data which shed light on the development of L2 collocation knowledge, at least quantitatively. It was not practicable to collect longitudinal data, but the cross-sectional, pseudo-longitudinal data yield interesting results which merit discussion.

7.3.4.2 Differences in performance between learner groups

In several studies, I investigated whether differences existed in mean scores on COLLEX and COLLMATCH between different groups of informants, both Swedish students and native speaker of English. These investigations were made based on two types of independent variables: learning level, i.e. years of classroom exposure, and general English proficiency. The comparison of students at different learning levels was warranted from the lack of such comparisons in the literature.

In terms of the results in the present thesis, starting with level of study as the independent variable, a clear pattern that emerged in my data was that scores on COLLEX and COLLMATCH increased with a higher level of study (see studies 3, 4, and 6), which in turn reflects the length of classroom exposure to English. This means that first-year university students produced significantly higher scores than upper-secondary school students, and second-year university students produced higher scores than first-year university students (see percentages in last but one paragraph in section 7.3.1 above). However, the means of university-level student groups only one term apart did not always differ in terms of statistical significance. This is interesting in the light of Mochizuki's (2002) longitudinal study in which Japanese university students did perform better on a receptive collocation test over a period of nine months. However, a clear-cut comparison is not possible since my own studies were not longitudinal, but cross-sectional, or possibly pseudo-longitudinal. I will come back to this issue shortly. Despite this lack of difference in the present study, in general, students seemed to have had acquired better receptive collocation knowledge for each higher level of study they entered into. In other words, a higher level of study (more years of classroom exposure) implied higher general proficiency, which in turn resulted in better receptive collocation knowledge. However, the cross-sectional data are in this regard problematic for one main reason. In theory, the differences observed between, for example, third-term university students and first-term university students might not have come from the fact that the former group had progressed one year further in the education system, and through this year of study acquired a higher general proficiency and better knowledge, but rather that the difference in knowledge was perhaps there already in the first place. In order to be able to proceed to a higher level of study, students have to pass a number of proficiency exams, such as practical grammar, translation, pronunciation and oral fluency, as well as exams targeting knowledge of English linguistics and English literature. In a strict sense, only a longitudinal study could have revealed whether the improved performance (better receptive knowledge of English collocations) of the third-term students was a result of further study, or if these students possessed this level of knowledge already as first term students.

A caveat must also be expressed when it comes to the observed differences between the upper-secondary student groups and the university student groups. The former were pursuing studies in a number of different subjects, such as mathematics, history, Swedish, social science, and physical education. For these informants, English was just one out of many obligatory subjects. The university students, on the other hand, were full-time students of English, and as such they had made a conscious choice to study one single subject in higher education for at least one term. It is very likely that for many of these students, this choice was based on the fact that they enjoyed English as a subject, were highly motivated, and presumably also relatively advanced in terms of proficiency. It is of course also possible for upper-secondary students to be more proficient than university learners, a fact which makes inferential analyses of learning level groups vis-à-vis underlying populations problematic.

If we nevertheless assume that it is the further study, for example one more year of full-time English studies at university, that creates the difference in increased COLLEX and COLLMATCH scores, then this seems to imply that Swedish learners of English, with reservations for the restricted student sample in the studies, improve their receptive collocation knowledge implicitly through exposure to English in the various course modules that they take, and any additional exposure that they are subjected to on a day-to-day basis. In Sweden, exposure to English is amply available everyday outside the language classroom through TV, radio, music, and the Internet. For example, TV programmes in English are subtitled, not dubbed, which means that people in Sweden, provided that they watch TV regularly, are exposed to different varieties of English discourse, predominately American English. Through this exposure, increases occur in terms of vocabulary size, which we have identified as an important factor, and also, partly as a result of the increase in size, in terms of knowledge about the combinatory potential of words, for example verb + NP collocations. This is reminiscent of the explanations put forward by *inter alia* Vermeer (2001), Qian (1999), and Read (2004), holding that vocabulary size and vocabulary depth are related dimensions, and that a deeper knowledge of words (for example knowledge of word collocates) is the consequence of knowing more words.

At the time when the data for my studies were gathered, there was no indication that collocations were explicitly taught or otherwise targeted in the courses taken be these students. No specific vocabulary acquisition course was offered, and no course syllabi mentioned collocation knowledge as a specific learning outcome. Although this observation is not backed up by any empirical data, but is rather impressionistic, a fact that must be stressed, it is all the same relevant. The results of this study suggest that receptive collocation knowledge develops as a function of extended exposure to the target language.

7.3.4.3 Differences in relation to general proficiency

Because of the uncertainties associated with conclusions drawn about general differences in collocation knowledge between students at different levels of study, studies were also carried out with general proficiency as the independent variable. Thus, rather than assuming that with a higher level of study follows automatically a higher general proficiency, and with that a similarly higher receptive collocation knowledge, groups of students were formed based on general proficiency, irrespective of whether they were upper-secondary school or university students. Since no scores from a proper general proficiency test were available, I decided to use vocabulary size scores as indicators of general proficiency. This method is certainly not without problems, but empirical data exist which show high correlations between vocabulary size scores and general proficiency measures (Meara & Buxton 1987; Meara & Jones 1988; Laufer 1997). Consequently, in studies 4 and 6 I observed significant differences in COLLEX and COLLMATCH mean scores between LOW, MID and HIGH groups formed on the basis of vocabulary size scores (see Tables 4.27 and 5.24).

If the assumption of a stable relationship between vocabulary size and general proficiency is accepted and borne out, then my results suggest that there is a relationship between English receptive collocation knowledge and general English language proficiency. In the literature, conflicting results have been reported in this regard. Bonk (2001) reported a moderately high correlation (r = .73), and Gitsaki (1999) claimed that collocation knowledge develops as L2 learners' overall language proficiency develops. In contrast, Howarth (1996) found a very low correlation (r = .15) and a lack of correlation was reported by Barfield (2003). Consequently,

my results corroborate previous findings made by Bonk (2001) and Gitsaki (1999), but go against the findings of Howarth (1996) and Barfield (2003; 2006). This is intriguing, but a closer look at the methods used in the respective studies might create a clearer picture. Bonk (2001) used a mix of verb + noun, verb + preposition, and figurative use of verb phrases in productive sentence gap-filling, and receptive multiple-choice task items administered to university learners (N = 98, mixed East-Asian L1s), and Gitsaki (1999) tested three groups of Greek high school learners (13, 14, and 15 year-olds) which she classified as post-beginner, intermediate, and post-intermediate, through a guided essay writing task, a translation exercise (Greek > English) and a sentence-level cloze (cued production) test in which one part of an English collocation was deleted (N = 275, Greek L1). It should be pointed out that Gitsaki used measures like lexical density, target-like use of articles, and words per T-unit as indicators of proficiency, but not independently from the collocation measure itself, which makes her claim problematic. Howarth (1996) analysed verb + noun collocations used in written essay production of university learners (N = 10, mixed L1s). A potential problem with this study is the small number of informants, which potentially could have skewed his results. Finally, Barfield (2003) administered a discrete, receptive recognition test of verb + noun collocations (119 items). In terms of the test format, Barfield's study lies close to my own. The learners in his study were 93 Japanese university students, who where classified as ranging between low-intermediate, intermediate, upper-intermediate, and advanced, by an in-house English proficiency placement test. It is impossible, though, to know how these proficiency levels compare with those of the Swedish students in my study. On balance, with such mix of methods, the numbers, L1s, and proficiency levels of informants, a clear-cut comparison is difficult to make, and further research will hopefully create a more uniform state of knowledge with regard to the relationship between general proficiency and collocation knowledge.

7.3.4.4 The acquisition of English collocations by Swedish-speaking learners and potential causes of differences in performance between learner groups

Going back to the results observed for the Swedish learners, I concluded earlier that clear differences were observable between upper-secondary school students and university level students in terms of COLLEX and COLLMATCH scores. These differences are interesting to discuss because they were found consistently, and they were statistically significant. For example, on COLLEX 4, the tested groups of upper-secondary school students (10^{th} and 11^{th} graders) scored means around 30 (60%) out of the maximum score of 50, whereas university student groups scored means around 44 (88%) out of 50 (see section 4.2.4.3.1.2, Table 4.18). Similarly, on COLLEX 5, a different group of 11^{th} graders scored a mean of 29 (58%) out of 50, and university student groups scored means between 41 and 46 (82 – 92%) out of 50 (see section 5.2.3.3.3, Table 5.14). The same patterns were observed in COLLMATCH scores. If we accept the assumption that similar differences would be observed in other samples of Swedish students from these populations, the question is what makes the university students perform so much better at the COLLEX task, and COLLMATCH task for that matter, than the 10^{th} and 11^{th} graders.

I have earlier pointed at the potential influence of vocabulary size. In fact, if seen proportionally, the same differences observed above between the upper-secondary school informants and university informant groups for COLLEX and COLLMATCH scores were also present in the vocabulary size scores for these groups. Thus, vocabulary size is an

important factor, and size is often seen as a good indication of proficiency, so proficiency seems to be a factor.

A competing, or perhaps complementing explanation can perhaps be found in L1 interference in L2 word processing. A number of researchers have proposed that L2 learners at different proficiency levels may be more or less dependent on L1 mediation (Potter *et al.* 1984; Kroll & Stewart 1994; Jiang 2000, 2002). A central notion here is that the process of learning words in an L1 involves a simultaneous development of a semantic/conceptual system and a lexical store, whereas the process of learning words in an L2 implies an already existing semantic/conceptual system. Jiang has argued that L2 words are initially in the acquisition process mapped to L1 translations (lexical form), not to meaning directly (2002:619). By meaning is here meant an existing semantic or conceptual system. This model is generally referred to as the word association model, whereas a competing model, referred to as the concept mediation model (Potter *et al.* 1984) proposes that L2 words are connected directly to their meanings without L1 mediation. More recent research has suggested a developmental transition from word association to concept mediation, which led Kroll and Stewart to propose the Revised Hierarchical Model (1994). This model, which is shown in Figure 7.4 below, assumes a higher level of conceptual processing with increasing L2 skill, and it incorporates both previous models. Specifically, the model predicts that early in L2 acquisition, L2 words are linked to L1 translations, which in turn are linked to conceptual representations. This is indicated by the unbroken arrows. Consequently, strong lexical links map L2 onto L1. With increasing L2 proficiency, direct conceptual connections from L2 words to semantics will begin to develop. This is indicated by the broken arrow between the conceptual store and the L2 lexical store (Kroll & Sunderman 2003).

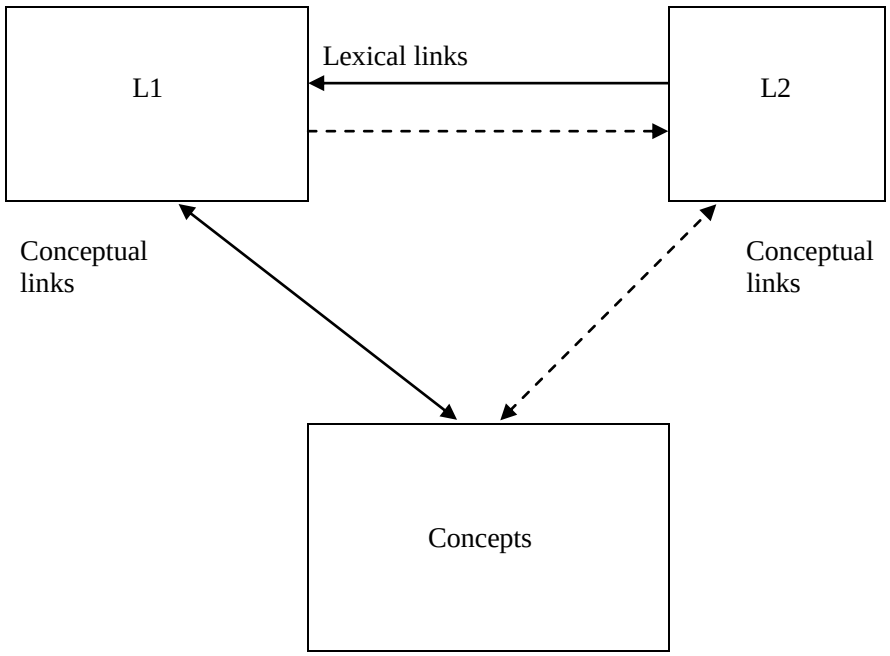

Figure 7.4 Revised Hierarchical Model (Kroll & Stewart 1994)

If we assume that the Swedish 10th and 11th graders have not yet started to process L2 words through direct conceptual connections, but are still dependent to a great extent on L1 word mediation, arguably this has potential consequences for their ability to process also English collocations in the receptive recognition task featured in COLLEX. Consider the test item taken from COLLEX 5 presented in Figure 7.5 below. In study 6 (section 5.2), the Item Facility of this item for a group of 11th graders (N = 26) was .65 compared to an average of .89 for the whole group of university students (N = 209), and .97 for a native speaker group (N = 34).

				a	b	c
1	a. make a conclusion	b. pull a conclusion	c. draw a conclusion			

Figure 7.5 Example items from COLLEX 5.

As was argued earlier in this thesis (see section 3.1.3), a test-taker faced with the COLLEX item task may resort principally to two cognitive strategies. Either, when processing the three word combinations presented in each item, a direct match can be made between one of the word combinations and a stored representation in the lexical mental lexicon (holistic approach). Or, a more analytic approach may be used in which the inherent elements of the word combinations, the L2 words, are processed separately (analytic approach).

In section 2.3.1 I accounted for Wray's (2002) postulation that collocations are formulaic sequences for native speakers, but they are essentially not so for non-native speakers. Wray's argument was that native speakers start with big units (collocations), and analyse them only as

necessary (into separate words), whereas collocations for L2 learners can be seen as separate items (words) which become paired (2002:211). In terms of teenage and adult L2 learners, Wray also emphasizes the tendency towards reliance on the word as a possible unit of linguistic processing. Tuition, she claims, that relies on the written medium, underlines the importance of small units over large ones (2002:206):

> All in all, after literacy, the second language learner is increasingly likely to deliberately aim to acquire a lexicon of word-sized units. The relative balance of words to formulaic word strings will be quite different from those [sic] of a native speaker.

This could be taken to mean that learners of low proficiency are prone to resort to analysis, whereas high-proficiency learners, in the sense of near-native speakers, may to a greater extent process word sequences holistically. This is also reminiscent of Barfield's (2006) hypothesis mentioned in section 7.3.3, which said that the L2 mental lexicon works in part from individual lexical items rather than lexical combinations. If we relate this to the Revised Hierarchical Model in Figure 7.4, this implies that high-proficiency learners tend to go from the L2 collocation form (combination of L2 words) directly to the conceptual store, with minimal influence from L1 forms, whereas lower proficiency learners go from the L2 collocation form via L1 translation equivalents of the individual orthographic words, through to the conceptual store.

If we assume that this is correct, then the implications in terms of processing involved in the COLLEX tasks become clearer. For example, in item 1 in Figure 7.5, a low-proficiency informant is assumed to process each word combination analytically, with strong L1 translation equivalent mediation. Thus, the L2 noun conclusion will be linked to the Swedish translation equivalent *slutsats*, which in turn will be linked to the concept associated with this abstract noun. Then, the respective L2 verbs *make*, *pull* and *draw* will each first be linked to potential L1 word equivalents, possibly *göra*, *dra* and *rita* or *dra* and then via these to conceptual representations. This is where it becomes interesting, for if the word-by-word L1 translations of the L2 English word combinations are juxtaposed, we get a. *göra en slutsats*, b. *dra en slutsats*, and c. *rita en slutsats* or *dra en slutsats*. In Swedish, **göra en slutsats* is infelicitous, and we are left with alternatives b and c, which both house an identical form, *dra en slutsats*, but where alternative c also invokes *draw* in the sense of 'to sketch'. The COLLEX task requires a single choice, and the choice is between alternative b. *pull a conclusion* and c. *draw a conclusion*. I would argue that out of the two competing forms, the L2 English verb *pull* is more strongly linked to the L1 Swedish verb *dra* than is the L2 English verb *draw*. In terms of word frequency, they are both 1K words according to the JACET 8000 list (Ishikawa *et al.* 2003), but the high frequency of *draw* is probably reflected in its sense 'to sketch something'. Indeed, the *Collins Cobuild* dictionary (Sinclair 2003) presents the meaning of *draw* as 'sketch' as sense 1, whereas the *draw* in the sense of 'deciding that a conclusion is true' is presented as the 17th sense. It is therefore not unlikely that with low-proficiency learners, the 'to sketch' sense of *draw* is more strongly evoked than the 'to decide' sense, which is clearly more formal and abstract. These learners are then prone to make the infelicitous choice of b. **pull a conclusion*, despite the fact that English *draw a conclusion* could be seen as a cognate of the Swedish *dra en slutsats*. In my data, out of the nine upper-secondary school informants who gave the wrong answer to this item, as many as

eight chose **pull a conclusion*. Thus, this reflects the strong L2 -> L1 lexical link mediation, which arguably lead them astray. It should be emphasized that the order of the above steps may not necessarily reflect the actual flow of events in an authentic situation.

In contrast, a high-proficiency learner facing the same COLLEX item is believed to have developed, by virtue of greater exposure to contextualized L2 input and acquired near-nativelike ability, a direct link between L2 lexical forms (single words or word combinations/collocations) and the conceptual store. In terms of exposure, Ellis (2002) suggests that language processing is intimately tuned to input frequency. Thus, on processing the word combination *draw a conclusion*, a mapping is made between this form and the conceptual representation of 'deciding that a particular conclusion is true' on the basis that learners have previously been exposed to this particular, or similar, form. This does not mean that L1 translation equivalents of the inherent words are no longer activated in processing, and that a certain amount of interference is not present in the processing activity, but that the strength of these connections is decreased in favour of the strengthened and more direct link between L2 form and concept which in a manner of speaking wins out. Thus, even though a similar process like the one accounted for above may effectively occur in a parallel fashion, the direct form-concept mapping overrides the potential interference. Jiang (2004a) argues that this stage implies more automaticity as well as idiomaticity, with less influence from L1 translations.

Even though L2 pedagogy and teaching pertinent to collocation knowledge is beyond the scope of this thesis, the findings do seem to indicate that Swedish learners of English, at upper-secondary school and university levels, possess relatively good receptive collocation knowledge. Many of the more advanced students performed equally good scores as informants from native speaker groups. I argued earlier that to the best of my knowledge, there is no structured teaching of collocations in the Swedish education system. The fact that Swedish students possess a relatively good knowledge despite the lack of specific instruction is interesting. It could be the case that, just like with L2 vocabulary, which is at least partly acquired incrementally through exposure, the type of verb + NP collocations featured in COLLEX and COLLMATCH are also acquired through exposure. Earlier, I pointed at the abundance of exposure to English in Sweden in addition to the classroom exposure that for most Swedish students starts in third or fourth grade in primary school. It is very likely that this extra-curricular exposure is a paramount consideration when discussing potential reasons behind the high performance levels. Relevant to this assumption, Nesselhauf (2005) found that that length of classroom exposure had no positive effect on collocation use, whereas length of exposure to the language (length of stays in English-speaking countries) had a slightly positive effect. Nesselhauf investigated collocation production in essays (German learners of English). Firstly, it is theoretically possible that collocation reception does not behave in the same way, and secondly, that we must remember that there may be a difference between classroom exposure and extra-curricular exposure, and that there might be similarities between the kind of language exposure Nesselhauf investigated and the extra-curricular exposure I referred to above. Moreover, Nesselhauf did not subject her data to statistical analyses, but interpreted the data rather impressionistically, a shortcoming which unfortunately places restrictions on her findings.

7.3.4.5 Answering RQ3

Research question 3 addressed the relationship between learning level and receptive knowledge of collocations. The research question read:

What is the relationship between the learning level of Swedish learners' of English as a foreign language (EFL) and their receptive knowledge of collocations?

The term 'learning level' was used to denote the formal progression in an education system, for example where on the one hand university students are on a higher learning level than upper-secondary school students, and on the other hand second-term university students are on a higher level than first-term university students. In terms of general language proficiency, more knowledge and better skills are assumed to be concomitant of a higher learning level. However, at the time when the present research project started, no previous studies had investigated whether receptive collocation knowledge increase as a function of a higher learning level. Schmitt (2000) has suggested that collocation knowledge is an advanced type of vocabulary knowledge, which could lead us to hypothesize that only the more advanced learners would be shown to have a good command in this area.

The findings in the present thesis show that receptive collocation knowledge, as measured in COLLEX and COLLMATCH increases as a function of higher learning level. In a series of cross-sectional studies, Swedish university students performed significantly higher scores than upper-secondary school students. Furthermore, significant differences were observed between university-level student groups one year apart in terms of learning level, but differences were not always significant between student groups only one term apart (4.5 – 6 months). This suggests that receptive collocation knowledge does not develop over such short period of time to a degree where it is measurable.

A potential explanation for the observed differences was seen in the hypothesized dominance in low-proficiency learners of L1 translation equivalent mediation in the links between L2 lexical forms (single words and word combinations) and conceptual representations, and the decreased role of this mediation in high-proficiency learners, something that supports the Revised Hierarchical Model (Kroll & Stewart 1994) as well as Wray's postulations about different processing in native speakers and non-native speakers (2002). More exposure to English was believed to facilitate collocation recognition, and a complementing explanation was also seen in the relation between vocabulary size and collocation knowledge, in that a deeper knowledge of words – for example knowledge of word collocates – is the consequence of knowing more words.

8 Conclusions, implications, and suggestions for further research

8.1 Introduction

Two overall aims guided the research project reported in this thesis. The first was to construct, use, and evaluate the effectiveness of tests of receptive collocation knowledge of L2 English, measured as a single construct. The second aim was to learn more about the level of receptive knowledge of collocations in advanced L2 learners, in particular in relation to vocabulary size and learning level. The time has now come to draw conclusions, to acknowledge some limitations, to consider the implications that follow from the conclusions, and to suggest areas of further research.

8.2 Main findings and conclusions

The test development project reported in this thesis has shown that it is possible to construct discrete tests of receptive collocation knowledge capable of yielding reliable scores when used with Swedish upper-secondary school and university-level students. Two tests were developed, aimed at complementing each other through slightly different test formats and tasks. The fact that two tests were used had the positive effect of making it possible to use one as concurrent validity support for the other. The COLLEX and COLLMATCH tests show good power of discrimination between test-takers at different proficiency levels and they are practical in being quick to sit and easy to mark. Furthermore, they seem to hold appeal with test-takers, and the monolingual test formats enable use with learners with different L1 backgrounds, as well as native speakers of English. Validation of the tests provided many kinds of evidence, which in an overall interpretation justify their use as proficiency tests, and as tests for diagnostic, placement or research purposes, but improvements are called for in terms of content validity, especially with regard to item selection and methods which would improve score generalisability.

Vocabulary size scores were observed to correlate strongly with receptive collocation knowledge scores in COLLEX and COLLMATCH, which implies that learners with large vocabularies are better at recognizing collocations than learners with smaller vocabularies. This pattern was observed also for native speakers of English, where it was beyond reasonable doubt that the single words making up the word combinations in the tests were known. These findings support explanations ventured by *inter alia* Vermeer (2001), Qian (1999), and Read (2004), who argue that a deeper knowledge of words, for example knowledge of word collocates, is the consequence of knowing more words.

The findings in the present thesis also show that receptive collocation knowledge, as measured in COLLEX and COLLMATCH, essentially increases as a function of higher learning level. Swedish university students performed significantly higher scores than upper-secondary school students. Furthermore, significant differences were observed between university-level student groups one year apart in terms of learning level, but differences were not always significant between student groups only one term, i.e. 4-6 months, apart. This suggests that the type of collocation knowledge measured in COLLEX and COLLMATCH does not develop over such short periods of time. Caution must be observed, however, when it comes to the sensitivity of the test tools: they may not be sensitive enough to pick up subtle

differences. Not surprisingly, but positive from a test validation perspective, native speakers of English outperformed all groups of Swedish learners. This was interpreted as support for the Revised Hierarchical Model proposed by Kroll & Stewart (1994), as well as Wray's (2002) postulation that collocations are formulaic sequences for native speakers, but essentially not so for non-native speakers. Furthermore, a great deal of exposure to English, both classroom exposure and extra-curricular, is believed to facilitate the acquisition of collocations, just like all other L2 skills.

8.3 Limitations

A number of limitations of the project reported in this thesis must be noted. Firstly, the targeted collocations in COLLEX and COLLMATCH were of two types of word combinations: adjective + NP and verb + NP. In their final versions, only the latter type was targeted. Even though a restriction on collocation types in the tests is likely to facilitate the interpretation of scores, and conversely that a use of a large number of types could make results difficult to interpret (cf. Gitsaki 1999), it also limits the generalisability of the test scores to receptive collocation knowledge in general.

Secondly, no in-depth analysis was carried out of the results of the test administrations with respect to the inherent semantic characteristics of the test items. Although rank-ordered lists of test item recognition levels were compiled (see Appendices), it was beyond the aim and scope of the present thesis project to investigate why certain collocations were recognised better than others by the different informant groups. Such an analysis is important if collocation difficulty is to be better understood.

Thirdly, the empirical studies conducted involved the gathering of cross-sectional data from Swedish learners of English at upper-secondary school and university levels. From a methodological perspective, this kind of approach should be complemented with longitudinal studies which can charter the development of receptive collocation knowledge in the same individuals.

Fourthly, I did not control for the degree to which the Swedish students in my studies had spent time abroad, in English-speaking environments. This kind of information would have provided an interesting variable to investigate.

A fifth limitation has to do with the fact that the corpus-based validation of test items in COLLEX and COLLMATCH was based on British English only. This was due to the lack of a suitable corpus of American English, similar to the BNC. Such a corpus is currently being developed (the American National Corpus (the ANC))[51]. The heavy reliance on British English in a test aimed at advanced Swedish learners is problematic since we know that much of the extra-curricular exposure that upper-secondary school students in Sweden experience is American English (Schepke 2007).

8.4 Implications

8.4.1 Testing

The series of empirical studies has shown promise for COLLEX and COLLMATCH as practical, appealing and reliable tests of receptive collocation knowledge, which also show evidence of different facets of validity in relation to their intended use. It was concluded,

[51] See www.americannationalcorpus.org

though, that further investigation is required into certain facets of validity, such as content validity. In their most current versions, COLLEX 5 and COLLMATCH 3 (Appendices 5I and 5K) may be used as diagnostic tests or proficiency tests in Swedish upper-secondary school and university settings. They may also be used as research tools, possibly as part of a test battery, as long as restrictions in test score generalisability, and potential ceiling effects are noted.

In terms of effectiveness, with regard to use with very advanced learners, COLLMATCH does not suffer from the same tendencies of ceiling effects as COLLEX, but the two tests make use of slightly different test tasks, and it is therefore recommended that they are used together in a test battery. The two tests complement each other, and since they are quick to sit and easy to mark, the administration of both tests in a test situation still affords a practical solution. It is perfectly possible for anyone interested in administrating COLLEX 5 and COLLMATCH 3 to reinstate some kind of control for self-indicated guessing, as was done in the earlier versions of the COLLEX format.

8.4.2 Learning and teaching collocations in a foreign language

The results suggest that 4-6 months of exposure to English, in a university-level setting, is not sufficient for receptive collocation knowledge to develop in Swedish students of English in a measurable way. There is evidence to suggest, however, that longer periods of exposure to English do facilitate the acquisition of collocations, as is the case with all aspects of the English language. Indeed, in an L1 acquisition setting, children learn language from exposure only. However, in an L2 setting, a complement to this exposure would be some sort of explicit learning of collocations. Because of their sheer number, it is probably unrealistic that collocations should be taught en masse in a structured way, just like it is unrealistic, mostly for lack of time, that teaching focuses on vocabulary material beyond the high-frequency words of the language (Nation 2001). The 'responsibility' for this type of learning probably has to lie with the learners themselves, but educators can draw students' attention to collocations, formulaicity and idiomaticity through classroom activities and teaching and learning materials. What is needed is first and foremost an awareness of collocations as linguistic items and the problems they may cause on the part of the learners (Howarth 1996; Hill 2000). Nesselhauf (2005:252) has suggested that:

> It is essential that learners recognize that there are combinations that are neither freely combinable nor largely opaque and fixed (such as idioms) but that are nevertheless arbitrary to some degree and therefore have to be learnt.

This raising of awareness could be accompanied by some teaching of collocations that are typically problematic, frequent and occur in a wide range (Nation 2001; Hill 2000).

An interesting anecdotal observation is that test use does lead to washback effects. Some of the data collection for this thesis was carried out as part of high-stakes vocabulary exams for students of English at Lund University, Sweden. Prior to the administration of COLLEX and COLLMATCH versions as part of end-of-term vocabulary exams at this university, the exam consisted solely of a 120-item vocabulary size test, in which students were required to match an English target word with one out of five possible Swedish translation equivalents in each item (see Gyllstad 2004). The effect that this test is believed to have had on students studying English is that they knew that L2 > L1 translation was the only type of vocabulary knowledge

tested in the exam, and nothing else. Hypothetically, there was no incentive for these students to pay much attention to collocations and other multi-word expressions[52]. When the L2 > L1 translation test was complemented with the L2 receptive recognition tests (COLLEX and COLLMATCH) it is very likely that the signal effect was that collocation knowledge is an important part of vocabulary knowledge, in addition to L2 > L1 translation. This means that the dominating hegemony of L2 > L1 translations was at least partly broken in favour of a focus also on the way words combine naturally in English. Evidence of this effect was seen in the many e-mails I received at this time from students about to sit a vocabulary exam in the near future, asking for resources (websites and literature) that may help them enhance their knowledge of collocations. Thus, if collocations are part of an exam, then students are likely to think that this type of lexical knowledge is important.

8.5 Suggestions for further research

Having investigated the performance of Swedish-speaking learners of English on COLLEX and COLLMATCH, validated through the performance of native speakers of English, it would be interesting to explore how learners with other L1s than Swedish perform on the tests. In fact, COLLEX 5 and COLLMATCH 3 are currently being used in two research studies. One involves Greek learners of English (Patrick McGavigan, University of Wales, Swansea, UK, personal communication), and the other involves French and Polish learners of English (Heather Hilton, Université de Savoie, France, personal communication). It will be interesting to see whether the levels of performance on COLLEX and COLLMATCH observed for Swedish learners are similar also to speakers of other languages.

The test development groundwork laid out in the present thesis could be expanded in many interesting ways. For example, investigations could be carried out with the aim of finding out what types of collocations pose particular difficulty for Swedish-speaking learners of English, and conversely, what types of collocations the same learners are potentially good at, and what the possible explanations may be (see Gitsaki 1999 for an example of such a study on Greek learners).

The receptive recognition formats used in COLLEX and COLLMATCH could also be further developed. One type of development would entail requirements on the part of the test-taker to not only recognize conventionalized English collocations in the test, but to also show that they know a suitable L1 meaning or translation equivalent of the L2 form. This could entail either a receptive multiple-choice format, or a requirement to supply the L1 form through receptive recall. Another expansion may entail the development of parallel and/or equivalent versions of the tests. Such development is more readily made for the COLLMATCH test than for the COLLEX test, since the latter is likely to require laborious analyses of suitable distractors, whereas the former is a yes/no-format that in a more straightforward fashion lends itself to these kinds of processes.

The large-scale cross-sectional comparisons of groups of learners carried out in the present thesis could be complemented with more qualitative study designs, in which individual patterns of collocation acquisition and use are investigated in-depth, through longitudinal

[52] The possible danger with this approach is that it may promote list learning of isolated words. List learning in itself does not have to be a poor strategy, but only list learning of single orthographic words is devoid of some of the context that those words normally appear in (see Hoey 2000). For example, *shed* as a verb commonly occurs together with noun objects like *weight, tears*, and *skin*.

methods. Such designs could comprise both measures of reception and production, and could potentially control for factors like type and amount of input, learner style and learner strategies, and motivation. It would be possible to investigate not only acquisition, but also attrition.

Making more structured observations of washback effects, for example through questionnaires or interviews with learners, could provide more substantial evidence than the anecdotal type presented here.

Finally, an interesting avenue to explore would be the development of computerized versions of COLLEX and COLLMATCH, capable of automatic scoring and possibly also measures of reaction times to items. Measures of reaction times could potentially shed light on the issue of whether learners store collocations holistically, as chunks, in the mental lexicon, or if their storage is essentially word-based.

REFERENCES

Aisenstadt, E. 1979. Collocability restrictions in dictionaries. In Hartmann, R. (ed.), *Dictionaries and their users*, 71-74. Leuven: Katholieke Universiteit.

Aitchison, J. 2003. *Words in the Mind*. Oxford: Blackwell.

Alderson, C. 1991. Dis-sporting Life. Response to Alistair Pollit's paper. In Alderson, C. and North, B. (eds.), *Language Testing in 1990s*, 60-67. London: Macmillan.

Alderson, C., Clapham, C. and Wall, D. 1995. *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Allén, S. 1975. *Frequency Dictionary of Present-Day Swedish*. Stockholm: Almqvist and Wiksell.

Altenberg, B. 1993. Recurrent Verb-complement Constructions in the London-Lund Corpus. In Aarts, J, de Haan, P. and Oostdijk, N. (eds.), *English Language Corpora: Design, Analysis and Exploitation*, 227-245. Amsterdam: Rodopi.

Altenberg, B. 1998. On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations. In Cowie, A. P. (ed.), *Phraseology, theory, analysis, and applications*, 101-122. Oxford: Oxford University Press.

Altenberg, B. and Granger, S. 2001. The Grammatical and Lexical Patterning of make in Native and Non-native Student Writing. *Applied Linguistics* 22(2). 173-195.

Amosova, N. N. 1963. *Osnovui anglijskoy frazeologii*. Leningrad: University Press.

Anderson, R. C. and Freebody, P. 1981. Vocabulary Knowledge. In Guthrie, J. T. (ed.) *Comprehension and Teaching: Research Reviews*, 77-117. Newark, DE: International Reading Association.

Anderson, R. C. and Freebody, P. 1983. Reading Comprehension and the Assessment and Acquisition of Word Knowledge. In Hunston, B. (ed.), *Advances in reading/language research, volume 2*, 231-156. Greenwich: JAI Press.

Arnaud, P. and Béjoint, H (eds). 1992. *Vocabulary and Applied Linguistics*. London: Macmillan.

Arnold, I.V. 1986. *The English Word*. Moscow: Vyssšaja škola.

Ashcraft, M. H. 2006. *Cognition*. New Jersey: Pearson Prentice Hall.

Aston, G. and Burnard, L. 1998. *The BNC Handbook*. Edinburgh: Edinburgh University Press.

Ayto, J. 2006. Idioms. In Brown, K. (ed.), *Encyclopedia of Language and Linguistics, volume 5*, 518-521. Oxford: Elsevier.

Bachman, L. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. 2004. *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

Bachman, L. and Palmer, A. 1996. *Language Testing in Practise*. Oxford: Oxford University Press.

Bahns, J. and Eldaw, M. 1993. Should we teach EFL students collocations? *System* 21 (1). 101-114.

Bailey, K.M. 1996. Working for washback: a review of the washback concept in language testing. *Language Testing* 13. 257-279.

Barfield, A. 2003. *Collocation Recognition and Production: Research Insights*. Tokyo: Chuo University.

Barfield, A. 2006. *An Exploration of Second Language Collocation Knowledge and Development*. Unpublished PhD Thesis. University of Wales, Swansea.

Barnbrook, G. 1996. *Language and Computers*. Edinburgh: Edinburgh University Press.

Bazell, C.E., Catford, C., Halliday, M.A.K. and Robbins, R.H. (eds.). 1966. *In memory of J.R. Firth*. London: Longmans.

Beglar, D. and Hunt, A. 1999. Revising and validating the 2000 Word Level and University Word Level Vocabulary Tests. *Language Testing* 16(2). 131-162.

Benson, M. 1985. Collocations and idioms. In Ilson, R. (ed), *Dictionaries Lexicography and language Learning*, 61-68. ELT Documents 120. Pergamon Press/British Council.

Benson, M., Benson, E. and Ilson, R. 1997. *The BBI dictionary of English word combinations*. Amsterdam: John Benjamins.

Berry-Rogghe, G.L.M. 1973. The Computation of Collocations and their Relevance in Lexical Studies. In Aitken, A.J., Bailey, R. and Hamilton-Smith, N. (eds.), *The Computer and Literary Studies*. Edinburgh: Edinburgh University Press.

Biskup, D. 1992. L1 influence on learners' renderings of English collocations. A Polish/German empirical study. In Arnaud, P.J.L. and Béjoint, H. (eds.), *Vocabulary and Applied Linguistics*, 85-93. London: Macmillan.

Bobrow, S. and Bell, S. 1973. On catching on to idiomatic expressions. *Memory and Cognition* 1. 343-346.

Bogaards, P. 2001. Lexical units and the learning of foreign language vocabulary. *Studies in Second Language Acquisition* 23. 321-343.

Bogaards, P. and Laufer, B. (eds.). 2004. *Vocabulary in a Second Language*. Amsterdam: John Benjamins.

Bonk, W.J. 2001. Testing ESL Learners' Knowledge of Collocations. In Hudson, T. and Brown, J.D. (eds.), *A Focus on Language Test Development: Expanding the Language Proficiency Construct Across a Variety of Tests. (Technical Report #21)*, 113-142. Honolulu: University of Hawai'I, Second Language Teaching and Curriculum Center.

Brown, F.G. 1983. *Principles of Educational and Psychological Testing*. New York: Holt, Rinehart and Winston.

Brown, J.D. and Hudson, T. 2002. *Criterion-referenced Language Testing*. Cambridge: Cambridge University Press.

Burton, R.F. 2002. Misinformation, partial knowledge and guessing. *Medical Education* 36. 805-811.

Butt, M. 2003. The Light Verb Jungle. *Harvard Working Papers in Linguistics* 9. 1-49. Department of Linguistics, Harvard University.

Cambridge ESOL examinations. Webpage available at: http://www.cambridgeesol.org/exams/cae.htm Accessed on 15 February 2007.

Cameron, L. 2002. Measuring Vocabulary Size in English as an Additional Language. *Language Teaching Research* 6(2). 145-173.

Carroll, J.B. 1968. The psychology of language testing. In Davies, A. (ed.), *Language Testing Symposium: A Psycholinguistic Perspective*, 46-69. London: Oxford University Press.

Carter, R. 1987. *Vocabulary: applied linguistic perspectives*. London: Routledge.

Carter, R. 1998. *Vocabulary: applied linguistic perspectives*. Second edition. London: Routledge.

Carter, R. and McCarthy, M. 1988. *Vocabulary and Language Teaching*. London: Longman.

Channel, J. 1981. Applying semantic theory to vocabulary teaching. *ELT Journal* 35(2). 115-122.

Chapelle, C. 1998. Construct definition and validity inquiry in SLA research. In Bachman, L. and Cohen, A. (eds.), *Interfaces Between Second Language Acquisition and Language Testing Research*, 32-70. Cambridge: Cambridge University Press.

Coady, J. and Huckin, T. (eds.). 1997. *Second Language Vocabulary Acquisition*. Cambridge: Cambridge University Press.

Cobb, T. 2006. *Vocabulary profile word list*. Available at: www.lextutor.ca/vp (Website accessed on 14 June 2006).
Cobb, T. 2007. *The Compleat Lexical Tutor*. Available at http://www.lextutor.ca/ (Website accessed on 1 February 2007).
Cowie, A.P. 1981. The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics* 2(3). 223-235.
Cowie, A.P. 1988. Stable and creative aspects of vocabulary use. In Carter, R. and McCarthy, M. (eds.), *Vocabulary and Language teaching*, 126-139. London: Longman.
Cowie, A.P. 1991. Multiword Units in Newspaper Language. In Granger, S (ed.), *Perspectives on the English Lexicon: A tribute to Jaques van Roey*, 101-116. CILL 17, 1-3.
Cowie, A.P. 1994. Phraseology. In Asher, R.E. (ed.), *The Encyclopedia of Language and Linguistics*, 3168-3171. Oxford: Pergamon.
Cowie, A.P. (ed.). 1998a. *Phraseology: theory, analysis, and applications*. Oxford: Oxford University Press.
Cowie, A.P. 1998b. Introduction. In Cowie, A.P. (ed.), *Phraseology: theory, analysis, and applications*, 1-20, Oxford: Oxford University Press.
Cowie, A.P. 1998c. Phraseological dictionaries: some East-West comparisons. In Cowie, A.P. (ed.), *Phraseology: theory, analysis, and applications*, 209-228. Oxford: Oxford University Press.
Cowie, A. P. and Howarth, P. 1996. Phraseological competence and written proficiency. In Blue, G.M. and Mitchell, R. (eds.), *Language and Education (British Studies in Applied Linguistics II)*, 80-93. Clevedon: Multilingual Matters.
Coxhead, A. 1998. *An Academic Word List*. Occasional Publication Number 18. LALS. Victoria University of Wellington, New Zealand.
Coxhead, A. 2000. A new academic word list. *TESOL Quarterly* 34. 213-239.
Cronbach, L. 1942. An analysis of techniques for diagnostic vocabulary testing. *Journal of Educational Research* 36. 206-217.
Cronbach, L. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16. 292-334.
Cronbach, L. 1984. *Essentials of Psychological Testing*. New York: Harper and Row.
Crowther, J., Dignen, S., and Lea, D. (eds.). 2002. *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press.
Cruse, A. 2000. *Meaning in Language*. Oxford: Oxford University Press.
Cruse, A. and Croft, W. 2004. *Cognitive Linguistics*. Cambridge: Cambridge University Press.
D'Anna, C. A., Zechmeister, E.B, and Hall, J.W. 1991. Toward a meaningful definition of vocabulary size. *Journal of Reading Behaviour: A Journal of Literacy* 23. 109-122.
Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara, T. 1999. *Studies in Language Testing 7: Dictionary of Language Testing*. University of Cambridge Local Examinations Syndicate. Cambridge: Cambridge University Press.
DeVellis, R.F. 1991. *Scale Development*. Newbury Park, NJ: Sage Publications.
Duden Oxford Grosswörterbuch. 2005. Oxford: Oxford University Press.
Ebel, R.L. 1979. *Essentials of Educational Measurement*. New Jersey: Prentice Hall.
Ellegård, A. 1960. Estimating Vocabulary Size. *Word* 16. 219-244.
Ellis, N. 1996. Sequencing in SLA: Phonological Memory, Chunking and Points of Order. *Studies in Second Language Acquisition* 18. 91-126.
Ellis, N. 1997. Vocabulary acquisition: word structure, collocation, word-class, and meaning. In Schmitt, N and McCarthy, M. (eds.), *Vocabulary: Description, Acquisition and Pedagogy*, 122-139. Cambridge: Cambridge University Press.

Ellis, N. 2002. Frequency Effects in Language Processing. *Studies in Second Language Acquisition* 24. 143-188.

Erman, B. and Warren, B. 2000. The idiom principle and the open choice principle. *Text* 20(1). 29-62.

Evert, S. and Krenn, B. 2003. *Computational approaches to collocations. Introductory course at the European Summer School on Logic, Language, and Information (ESSLLI 2003), Vienna*. Available at: http://www.collocations.de/EK/Articles/ESSLLI-1.ppt. (Website accessed on 27 May 2007).

Eyckmans, J. 2004. *Measuring Receptive Vocabulary Size*. Utrecht: LOT.

Farghal, M. and Obiedat, H. 1995. Collocations: A neglected variable in EFL. *International Journal of Applied Linguistics* 28(4). 313-331.

Field, A. 2005. *Discovering Statistics Using SPSS*. London: Sage.

Fillmore, C., Johnson, C. and Petruck, M. 2003. Background to Framenet. *International Journal of Lexicography* 16(3). 235-250.

Firth, J. R. (1957). A synopsis of linguistic theory. 1930-55. In Palmer, F.R. (ed.), *Selected papers of J.R. Firth*, 168-205.

Firth, J. R. 1951. Modes of meaning. In Palmer, F.R. (ed.), *Papers in linguistics 1934-1951*, 190-215.

Firth, J. R. 1968 [1952/3]. Linguistic analysis as a study of meaning. In Palmer, F.R. (ed.), *Selected papers of J.R. Firth*, 12-26.

Fontenelle, T. 1998. Discovering Significant Lexical Functions in Dictionary Entries. In Cowie, A.P. (ed.), *Phraseology: theory, analysis, and applications*, 189-207. Oxford: Oxford University Press.

Francis, W.N. and Kucera, H. 1982. *Frequency analysis of English usage*. Boston, MA: Houghton Mifflin.

Gitsaki, C. 1999. *Second Language Lexical Acquisition: A Study of the development of collocational knowledge*. San Francisco: International Scholars Publications.

Gläser, R. 1988. The grading of idiomaticity as a presupposition for a taxonomy of idioms. In Hullen, W. and Schulze, R. (eds.), *Understanding the Lexicon Meaning, Sense and World: Knowledge in Lexical Semantics*. Tübingen: Max Niemeyer.

Goulden, R., Nation, P., and Read, J. 1990. How large can a receptive vocabulary be? *Applied Linguistics* 11. 341-363.

Granger, S. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In Cowie, A.P. (ed.), *Phraseology: Theory, analysis, and applications*, 145-160. Oxford: Oxford University Press.

Green, D.M. and Swets, J.A. 1966. *Signal detection theory and psychophysics*. New York: Wiley.

Greenbaum, S. 1970. *Verb-Intensifier Collocations in English: An Experimental Approach*. The Hague: Mouton.

Greenbaum, S. 1974. Some verb-intensifier collocations in American and British English. *American Speech* 49 (1/2). 79-89.

Greidanus, T. and Nienhuis, L. 2001. Testing the quality of word knowledge in a second language by means of word associations: types of distractors and types of associations. *The Modern Language Journal* 85 (iv). 567-577.

Greidanus, T., Bogaards, P., van der Linden, E., Nienhuis, L. and de Wolf, T. 2004. The construction and validation of a deep word knowledge test for advanced learners of French. In Bogaards, P. and Laufer, B. (eds.), *Vocabulary in a Second Language*, 191-208. Amsterdam: John Benjamins.

Greidanus, T., Beks, B. and Wakely, R. 2005. Testing the Development of French Word Knowledge by Advanced Dutch and English-Speaking Learners and Native Speakers. *The Modern Language Journal* 89 (ii). 221-233.

Gyllstad, H. 2004. Testing L2 vocabulary: Current test formats in English as a L2 used at Swedish universities. In Heinat, F. and Manninen, S. (eds.), *The Department of English in Lund: Working Papers in Linguistics* 4, 21-40. Lund: Lund University. Available online at: http://ask.lub.lu.se/archive/00019717/01/gyllstad-wp-04.pdf

Haastrup, K. and Viberg, Å. (eds.). 1998. *Perspectives on Lexical Acquisition in a Second Language*. Lund: Lund University Press.

Halliday, M.A.K. 1961. Categories of the theory of grammar. *Word* 17. 241-292.

Halliday, M.A.K. 1966. Lexis as a linguistic level. In Bazell, C.E, Catford, C., Halliday, M.A.K. and Robins, R.H. (eds.), *In memory of J.R. Firth*, 148-162. London: Longmans.

Hazenberg, S. and Hulstijn, J. 1996. Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation. *Applied Linguistics* 17. 145-163.

Heatley, A., Nation, I.S.P. and Coxhead, A. 2002. *RANGE and FREQUENCY programs*. Available at: http://www.vuw.ac.nz/lals/staff/Paul_Nation (Website accessed in September 2004).

Heiman, G. 2006. *Basic Statistics for the Behavioural Sciences*. Boston: Houghton Mifflin.

Henning, G. 1987. *A Guide to Language Testing*. Boston: Heinle & Heinle.

Henriksen, B., Albrechtsen, D., and Haastrup, K. 2004. The Relationship Between Vocabulary Size and Reading Comprehension. In Albrechtsen, D., Haastrup, K. and Henriksen, B. (eds.), *Angles on the English- speaking World* 4, 129-140. Copenhagen: Museum Tusculanum Press.

Hill, J. 2000. Revising priorities: from grammatical failure to collocational success. In Lewis, M. (ed.), *Teaching Collocation*, 47-69. Hove: Language Teaching Publications.

Hoey, M. 2000. A world beyond collocation: new perspectives on vocabulary teaching. In Lewis, M. (ed.), *Teaching Collocation*, 224-245. Hove: Language Teaching Publications.

Hoey, M. 2005. *Lexical Priming: A new theory of words and language*. Abingdon, Oxon: Routledge.

Howarth, P. 1996. *Phraseology in English Academic Writing: Some Implications for Language Learning and Dictionary Making*. Lexicographica Series Maior 75. Tübingen: Max Niemeyer.

Howarth, P. 1998a. Phraseology and Second Language Proficiency. *Applied Linguistics* 19(1). 24-44.

Howarth, P. 1998b. The Phraseology of Learners' Academic Writing. In Cowie, A.P. (ed.), *Phraseology: Theory, analysis, and applications*, 161-186, Oxford: Oxford University Press.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Ijaz, I.H. 1986. Linguistic and cognitive determinants of lexical acquisition in a second language. *Language Learning* 36. 401-451.

Ishikawa, S., Uemura, T., Kaneda, M., Shimizu, S., Sugimori, N. and Tono, Y. 2003. *JACET 8000: JACET List of 8000 Basic Words*. Tokyo: JACET.

Jespersen, O. 1965. *A Modern English Grammar on Historical Principles, Part VI, Morphology*. London: George Allen and Unwin Ltd.

Jiang, N. 2000. Lexical representation and development in a second language. *Applied Linguistics* 21(1). 47-77.

Jiang, N. 2002. Form-meaning mapping in vocabulary acquisition in a second language. *Studies in Second Language Acquisition* 24. 617-637.

Jiang, N. 2004a. Semantic transfer and its implications for vocabulary teaching in a second language. *The Modern Language Journal* 88(3). 416-432.
Jiang, N. 2004b. Semantic transfer and development in adult L2 vocabulary acquisition. In Bogaards, P. and Laufer, B. (eds.), *Vocabulary in a Second Language*, 101-126.
Jones, N. 2001. *Reliability in UCLES' examinations*. Unpublished internal UCLES' report.
Jones, S. and Sinclair, J. 1974. English lexical collocations: A study in computational linguistics. *Cahiers de Lexicologie* 24. 15-61.
Jourdenais, R. 2001. Cognition, instruction and protocol analysis. In Robinson, P. (ed.), *Cognition and Second Language Instruction*, 354-375. Cambridge: Cambridge University Press.
Kamimoto, T. 2005. *The effect of guessing on vocabulary test scores: a qualitative analysis*. Paper presented at the 15th conference of The European Second Language Association EUROSLA, Dubrovnik, Croatia, 14-17 September 2005.
Kane, M., Crooks, T. and Cohen, A. 1999. Validating Measures of Performance. *Educational Measurement: Issues and Practice* 18(2). 5–17.
Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. London: Longman.
Kilgarriff, A. 1996. BNC database and word frequency list. Available at: http://www.itri.brighton.ac.uk/ ~Adam.Kilgarriff/bnc-readme.html. (Website accessed on 1 September 2004).
Kjellmer, G. 1984. Some thoughts on collocational distinctiveness. In Aarts, J. and Meijs, W. (eds.), *Corpus Linguistics*, 163-171. Amsterdam: Rodopi.
Kjellmer, G. 1987. Aspects of English Collocations. In Meijs, W. (ed.), *Corpus Linguistics and Beyond*, 133-140. Amsterdam: Rodopi.
Kjellmer, G. 1991. A mint of phrases. In Aijmer, K. and Altenberg, B. (eds.), *English corpus linguistics*, 111-127. London: Longman.
Kjellmer, G. 1994. *A Dictionary of English Collocations*. Oxford: Clarendon Press.
Klein-Braley, C. 1991. Ask a Stupid Question…: Testing Language Proficiency in the Context of Research Studies. In De Bot, K., Ginsberg, R. and Kramsch, C. (eds.), *Foreign Language Research in Cross-cultural Perspective*, 73-94. Amsterdam: John Benjamins.
Knutsson, R. 2006. *Formulaic Language in L1 and L2*. Unpublished Licentiate Thesis. Lund University.
Krishnamurthy, R. (ed.). 2004. *English collocation studies: the OSTI report*. London: Continuum.
Kroll, J.F. and Stewart, E. 1994. Category interference in translation and picture naming: evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language* 33. 149–174.
Kroll, J.F. and Sunderman, G. 2003. Cognitive processes in second language acquisition. In Doughty, C. and Long, M. (eds.), *Handbook of second language acquisition*, 104-129. Cambridge, MA: Blackwell.
Kruse, H., Pankhurst, J. and Sharwood Smith, M. 1987. A multiple word association probe in second language acquisition research. *Studies in Second Language Acquisition* 9(2). 141-154.
Källkvist, M. 1998. *Form-Class and Task-Type Effects in Learner English*. Lund Studies in English 95. Lund: Lund University Press.
Lado, R. 1961. *Language Testing: The Construction and Use of Foreign Language Tests*. London: Longmans.
Langer, S. 2005. A linguistic test battery for support verb constructions. *Linguisticae Investigationes* 27(2). 171-184.

Laufer, B. 1997. The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess. In Coady, J. and Huckin, T. (eds.), *Second Language Vocabulary Acquisition*, 20-34. Cambridge: Cambridge University Press.
Laufer, B. and Goldstein, Z. 2004. Testing Vocabulary Knowledge: Size, Strength, and Computer Adaptiveness. *Language Learning* 54. 399-436
Lewis, M. 1993. *The Lexical Approach*. Hove: LTP.
Lewis, M. 1997. Pedagogical implications of the lexical approach. In Coady, J. and Huckin, T. (eds.), *Second Language Vocabulary Acquisition*, 255-270. Cambridge: Cambridge University Press.
Linnarud, M. 1986. *Lexis in Composition*. Lund Studies in English 74. Malmö: Liber.
Lyons, J. 1977. *Semantics*. Cambridge: Cambridge University Press.
McCarthy, M. 1990. *Vocabulary*. Oxford: Oxford University Press.
McCarthy, M. and O'Dell, F. 2005. *English Collocations in Use*. Cambridge: Cambridge University Press.
McNamara, T. 2000. *Language Testing*. Oxford: Oxford University Press.
Meara, P. 1980. Vocabulary acquisition: a neglected aspect of language learning. *Language Teaching and Linguistics; Abstracts* 13. 221-246.
Meara, P. 1982. Word associations in a foreign language. *Nottingham Linguistics Circular* 11. 29-38.
Meara, P. 1987. *Vocabulary in a Second Language. Volume 2*. London: CILS.
Meara, P. 1990. A note on passive vocabulary. *Second Language Research* 6. 150-154.
Meara, P. 1996. The dimensions of lexical competence. In Brown, G., Malmkjaer, K. and Williams, J. (eds.), *Performance and Competence in Second Language Acquisition*, 35-53. Cambridge: Cambridge University Press.
Meara, P. 2005. *X_Lex: the Swansea Vocabulary Levels Test*. v2.05. Swansea: Lognostics.
Meara, P. and Buxton, B. 1987. An alternative to multiple choice vocabulary tests. *Language Testing* 4. 142-154.
Meara, P. and Jones, G. 1988. Vocabulary Size as a Placement Indicator. In Grunwell, P. (ed.), *Applied Linguistics in Society*, 80-87. London: CILT.
Meara, P. and Jones, G. 1990. *Eurocentres Vocabulary Size Tests 10KA*. Zurich: Eurocentres Learning Service.
Meara, P. and Milton, J. 2003. X_Lex: the Swansea Vocabulary Levels Test. Swansea: Lognostics.
Meara, P. and Wolter, B. 2004. V_Links: Beyond vocabulary depth. In Albrechtsen, D., Haastrup, K. and Henriksen, B. (eds.), *Angles on the English- speaking World* 4, 85-96. Copenhagen: Museum Tusculanum Press.
Mel′čuk , I. 1998. Collocations and lexical functions. In Cowie, A.P. (ed.), *Phraseology: theory, analysis, and applications*, 24-53. Oxford: Oxford University Press.
Melka, F. 1997. Receptive vs. productive aspects of vocabulary. In Schmitt, N. and McCarthy, M. (eds.), *Vocabulary: Description, Acquisition and Pedagogy*, 84-102. Cambridge: Cambridge University Press.
Messick, S. 1989. Validity. In Linn, R.L. (ed.), *Educational Measurement*, 13-104. Phoenix: The Oryx Press.
Messick, S. 1995. Validity of Psychological Assessment. *American Psyhologist* 50(9). 741-749.
Meyer, C. F. 2002. *English Corpus Linguistics*. Cambridge: Cambridge University Press.
Miller, G. 1956. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review* 63. 81-97.
Miller, G. 1999. On knowing a word. *Annual Review of Psychology* 50. 1-19.

Mitchell, T. F. 1971. Linguistic 'goings on': collocations and other lexical matters arising on the syntagmatic record. *Archivum Linguisticum* 2. 35-69.
Mitchell, T.F. 1966. Some English phrasal types. In Bazell, C.E, Catford, C., Halliday, M.A.K. and Robins, R.H. (eds.), *In memory of J.R. Firth*, 335-358. London: Longmans.
Mochizuki, M. 2002. Exploration of two aspects of vocabulary knowledge: Paradigmatic and collocational. *Annual Review of English Language Education in Japan* 13. 121-129.
Moon, R. 1997. Vocabulary connections: multiword-items in English. In Schmitt, N. and McCarthy, M (ed.), *Vocabulary: Description, Acquisition and Pedagogy*, 40-63. Cambridge: Cambridge University Press.
Moon, R. 1998. *Fixed Expressions and Idioms in English – A Corpus-based Approach*. Oxford: Clarendon Press.
Nagy, W. 1997. On the role of context in first- and second-language vocabulary learning. In Schmitt, N. and McCarthy, M. (eds.), *Vocabulary: Description, Acquisition and Pedagogy*, 64-83. Cambridge: Cambridge University Press.
Nation, I.S.P. 1983. Testing and teaching vocabulary. *Guidelines* 5. 12-25.
Nation, I.S.P. 1990. *Teaching and Learning Vocabulary*. New York: Heinle and Heinle.
Nation, I.S.P. 1993. Using dictionaries to estimate vocabulary size: essential, but rarely followed procedures. *Language Testing* 10(1). 27-40.
Nation, I.S.P. 1986. *Vocabulary Lists: Words, Affixes and Stems*. Occasional Publication No. 12. University of Wellington, Victoria: English Language Institute.
Nation, I.S.P. 1996. *Vocabulary Lists*. Occasional Publication No. 17. University of Wellington, Victoria: English Language Institute.
Nation, I.S.P. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.
Nation, I. S. P. 2006. How Large a Vocabulary is Needed for Reading and Listening. *Canadian Modern Language Review* 63(1). 59-82.
Nesselhauf, N. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics* 24(2). 223-242.
Nesselhauf, N. 2004. What are collocations? In Allerton, D.J, Nesselhauf, N. and Skandera, P. (eds.), *Phraseological Units: basic concepts and their application*, 1-21. Basel: Schwabe.
Nesselhauf, N. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.
Newell, A. 1990. *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.
Nurweni, A. and Read, J. 1999. The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes* 18. 161-175.
Oakes, M. P. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
Palmer, H. 1931. *First interim report on vocabulary selection*. Tokyo: Kaitakusha.
Palmer, H. 1933. *Second interim report on English collocations*. Tokyo: Kaitakusha.
Paribakht, T. and Wesche, M. 1993. The relationship between reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal* 11. 9-29.
Partington, A. 1998. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.
Pawley, A., and Syder, F. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In Richards, J. and Schmidt, R. (eds.), *Language and Communication*, 191-226. Harlow: Longman.
Pitt, D. and Katz, J. 2000. Compositional Idioms. *Language* 76(2). 409-432.
Potter, M., So, K., Von Eckardt, B. and Feldman, L. 1984. Lexical and conceptual representation in beginning and more proficient bilinguals. *Journal of Verbal Learning and Verbal Behavior* 23. 23-38.

Qian, D. 1999. Assessing the Roles of Depth and Breadth of Vocabulary Knowledge in Reading Comprehension. *Canadian Modern Language Review* 56(2). 282-308.
Qian, D. 2002. Investigating the Relationship between Vocabulary Knowledge and Academic Reading Performance: An Assessment Perspective. *Language Learning* 52(3). 513-536.
Read, J. 1993. The development of a new measure of L2 vocabulary knowledge. *Language Testing* 10. 355-371.
Read, J. 1998. Validating a test to measure depth of vocabulary knowledge. In Kunnan, A. (ed.), *Validation in Language Assessment*, 41-60. Mahwah, NJ.: Lawrence Erlbaum.
Read, J. 2000. *Assessing Vocabulary*. Cambridge: Cambridge University Press.
Read, J. 2004. Plumbing the depths: How should the construct of vocabulary knowledge be defined? In Bogaards, P. and Laufer, B. (eds.), *Vocabulary in a Second Language*, 209-227. Amsterdam: John Benjamins.
Read, J. & Chapelle, C. 2001. A framework for second language vocabulary assessment. *Language testing* 18 (1). 1-32.
Richards, J. 1976. The role of vocabulary teaching. *TESOL Quarterly* 10(1). 77-89.
Robert Collins English-French/French-English Dictionary. 1987. London: Collins.
Robins, R.H. 1961. John Rupert Firth. *Language* 37(2). 191-200.
Ruhl, C. 1989. *On monosemy: A study in linguistic semantics*. Albany: State University of New York Press.
Saeed, J. 2003. *Semantics*. Oxford: Blackwell.
Schacter, D.L. 1987. Implicit Memory: History and Current Status. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13(3). 501-518.
Schepke, P. 2007. *British vs. American English – Swedish upper-secondary school students' attitudes towards English varieties and their ability to distinguish between them*. Unpublished undergraduate paper. Malmö University College.
Schmitt, N. 1998a. Tracking the Incremental Acquisition of Second Language Vocabulary: A Longitudinal Study. *Language Learning* 48(2). 281-317.
Schmitt, N. 1998b. Measuring collocational knowledge: key issues and an experimental assessment procedure. *ITL Review of Applied Linguistics* 119-120. 27-47.
Schmitt, N. 1999. The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language Testing* 16(2). 189-216.
Schmitt, N. 2000. *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.
Schmitt, N. (ed.). 2004. *Formulaic Sequences*. Amsterdam: John Benjamins.
Schmitt, N. and McCarthy, M. (eds.). 1997. *Vocabulary – Description, Acquisition, Pedagogy*. Cambridge: Cambridge University Press.
Schmitt, N. and Meara, P. 1997. Researching Vocabulary Through a Word Knowledge Framework. *Studies in Second Language Acquisition* 20. 17-36.
Schmitt, N., Dörneyi, Z., Adolphs, S., and Durow, V. 2004. Knowledge and acquisition of formulaic sequences. In Schmitt, N. (ed.), *Formulaic Sequences*, 55-86. Amsterdam: John Benjamins.
Schmitt, N., Schmitt, D. and Clapham, C. 2001. Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing* 18(1). 55-88.
Schreuder, R. and Weltens, B. (eds.). 1993. *The Bilingual Lexicon*. Amsterdam: John Benjamins.
Shillaw, J. 1999. *The Application of the Rasch Model to Yes/No Vocabulary tests*. Unpublished PhD thesis. University of Wales, Swansea.
Siepmann, D. 2005. Collocation, colligation and encoding dictionaries. Part I: lexicological aspects. *International Journal of Lexicography* 18(4). 409-443.

Sinclair, J. 1966. Beginning the study of lexis. In Bazell, C.E, Catford, C., Halliday, M.A.K. and Robins, R.H. (eds.), *In memory of J.R. Firth*, 410-430. London: Longmans.
Sinclair, J. 1987. Collocation: a progress report. In Steele, R. and Threadgold, T. (eds.), *Language topics: Essays in honour of Michael Halliday*, 319-331.
Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Sinclair, J. (ed.). 2003. *Collins COBUILD Advanced Learner's Dictionary*. Fourth edition. Glasgow: Harper Collins.
Sinclair, J., Jones, S, and Daley, R. 2004 [1970]. English Lexical Studies: Report to OSTI on Project C/LP/08. In Krishnamurthy, R. (ed.), *English collocation studies: the OSTI report*, 2-204. London: Continuum.
Singleton, D. 1999. *Exploring the Second Language Mental Lexicon*. Cambridge: Cambridge University Press.
Söderman, T. 1993. Word associations of foreign language learners and native speakers: The phenomenon of a shift in response type and its relevance for lexical development. In Ringbom, H. (ed.), *Near-native proficiency in English*, 91–182. Åbo: Åbo Akademi University.
Stæhr Jensen, L. 2005. *Vocabulary Knowledge and Listening Comprehension in English as a Foreign Language*. Unpublished PhD Thesis. Copenhagen Business School.
Storrer, A. 2006. Corpus-based investigations on German support verb constructions. Manuscript. To appear in Fellbaum, C. (ed.), *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*. London: Continuum Press.
Stubbs, M. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.
Stubbs, M. 2004. A quantitative approach to collocations. In Allerton, D.J, Nesselhauf, N., and Skandera, P. (eds.), *Phraseological Units: basic concepts and their application*, 107-119. Basel: Schwabe.
Sweet, H. 1899. *The Practical Study of Languages: A Guide for Teachers and Learners*. London: Dent.
Swinney, D. and Cutler, A. 1979. The Access and Processing of Idiomatic Expressions. *Journal of Verbal Learning And Verbal Behavior* 18. 523-534.
Thorndike, E. and Lorge, I. 1944. *The Teacher's Word Book of 30,000 Words*. New York: Teachers College, Columbia University.
Thorndike, R.L. 1973. *Reading comprehension education in fifteen countries*. Stockholm: Almquist and Wiksell.
Vermeer, A. 2001. Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics* 22. 217-235.
Vinogradov, V. V. 1947. Ob osnovnuikh tipakh frazeologicheskikh edinits v russkom yazuike. In A.A. Shakhmatov, 1864-1920. *Sbornik statey i materialov*, 339-364. Moscow: Nauka.
Vives Boix, G. 1995. *The Development of a Measure of Lexical Organization*: The Association Vocabulary Test. Unpublished PhD thesis. University of Wales, Swansea.
Warren, B. 2001. Accounting for compositionality. In Aijmer, K. (ed.), *A Wealth of English – Studies in Honour of Göran Kjellmer*, 103-114. Göteborg: Acta Universitatis Gothoburgensis.
Warren, B. 2003. *The linguistic status of collocations*. Unpublished manuscript. Lund University.
Waring, R. 1997. A comparison of the receptive and the productive vocabulary sizes of some second language learners. *Immaculata* 1. 53-68.
Warring, R. 1999. *Tasks for Assessing Second Language Receptive and Productive Vocabulary*. Unpublished PhD Thesis. University of Wales, Swansea.

Weir, C. J. 2005. *Language Testing and Validation*. Houndmills, Basingstoke: Palgrave Macmillan.
Wesche, M. and Paribakht, T. S. 1996. Assessing second language vocabulary knowledge: depth versus breadth. *Canadian Modern Language Review* 53(1). 13-40.
West, M. 1953. *A General Service List of English Words*. London: Longman.
Wiktorsson, M. 2003. *Learning Idiomaticity: A Corpus-Based Study of Idiomatic Expressions in Learners' Written Production*. Stockholm: Almqvist&Wiksell International.
Wilkins, D. 1972. *Linguistics and Language Teaching*. London: Edward Arnold.
Wolter, B. 2001. Comparing the L1 and L2 mental lexicon: a depth of individual word knowledge model. *Studies in Second Language Acquisition* 23. 41-69.
Wolter, B. 2005. *V_Links: A New Approach to Assessing Depth of Word Knowledge*. Unpublished PhD Thesis. University of Wales, Swansea.
Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.
Xue, G. and Nation, I.S.P. 1984. A university word list. *Language Learning and Communication* 3. 215-229.
Zimmerman, J., Broder, P., Shaugnessy, J. and Underwood, B. 1977. A recognition test of vocabulary using signal-detection measures and some correlates of word and non word recognition. *Intelligence* 1. 5-13.

APPENDIX 3A: Item Facility (IF) and Item-Total Correlation (ITC) values of COLLEX 1

Item	Item pair	Item Facility (IF)	Item-total correlation (ITC)
1	break a record - strike a record	1.00	0.00
2	set the bed - make the bed	1.00	0.00
5	run a business - drive a business	1.00	0.00
8	lose faith - drop faith	1.00	0.00
9	finish a fire - put out a fire	1.00	0.00
10	make an objection - take an objection	1.00	0.00
11	keep a speech - give a speech	1.00	0.00
13	make progress - take progress	1.00	0.00
15	make a discovery - have a discovery	1.00	0.00
18	pay a promise - make a promise	1.00	0.00
19	pull a guess - take a guess	1.00	0.00
22	pay attention - show attention	1.00	0.00
23	make an apology - do an apology	1.00	0.00
26	lose patience - spill patience	1.00	0.00
28	drive a bicycle - ride a bicycle	1.00	0.00
31	speak a prayer - say a prayer	1.00	0.00
35	drop temper - lose temper	1.00	0.00
36	give birth - lay birth	1.00	0.00
40	do a sacrifice - make a sacrifice	1.00	0.00
43	solve a conflict - break a conflict	1.00	0.00
46	pull a conclusion - draw a conclusion	1.00	0.00
54	take a recovery - make a recovery	1.00	0.00
56	do a response - give a response	1.00	0.00
58	achieve a goal - solve a goal	1.00	0.00
4	do a favour - make a favour	.95	.34
7	make an escape - take an escape	.95	-.02
16	ride a car - drive a car	.95	-.02
27	button a belt - fasten a belt	.95	.19
41	pull a parallel - draw a parallel	.95	-.09
45	commit an error - conduct an error	.95	.55
49	break an issue - settle an issue	.95	-.09
25	brush shoes - polish shoes	.89	.08
32	break a habit - lay a habit	.89	.46
34	turn a key - twist a key	.89	-.03
39	catch a disease - receive a disease	.89	.29
53	hold a demonstration - lay a demonstration	.89	.24
55	make an estimate - draw an estimate	.89	.46
57	earn access - gain access	.89	.18
59	conduct a survey - commit a survey	.89	.18
3	drop count - lose count	.84	.62
12	lay the table - make the table	.84	-.11
24	run a fever - draw a fever	.84	-.11
42	write a draft - conduct a draft	.84	-.11
50	conduct a method - adopt a method	.84	.02
14	draw a watch - wind a watch	.79	.26
21	do a bow - take a bow	.79	.52
37	make a reminder - give a reminder	.79	.26
51	make a project - run a project	.79	-.10
20	reach a dream - realize a dream	.68	.20
29	hold a discussion - make a discussion	.68	.35
38	do damage - make damage	.68	.27
47	apply a formula - adopt a formula	.68	-.18
33	make a bath - run a bath	.63	.39
48	lose inhibitions - drop inhibitions	.58	.01
17	exercise rights - employ rights	.53	.44
52	perform a task - solve a task	.53	-.22
6	set a deal - strike a deal	.47	.39
44	lay a wound - dress a wound	.47	.50
60	employ a policy - pursue a policy	.32	-.39
30	fly a flag - run a flag	.21	.01
	MEAN	.86	.10

APPENDIX 3B: Item Facility (IF) and Item-Total Correlation (ITC) values of COLLEX 1 test administration

Item	Item pair	Item Facility (IF)	Item-total correlation (ITC)
2	run a business – drive a business	1.0	0.00
14	do an attempt – make an attempt	1.0	0.00
38	take a break – seize a break	1.0	0.00
1	set the bed – make the bed	.99	.10
8	crush a heart – break a heart	.99	.07
9	a heart beats – a heart strikes	.99	.07
11	hit a number – dial a number	.99	.30
32	tell the truth – say the truth	.99	.07
54	tell a lie – say a lie	.99	.09
12	make an effort – commit an effort	.98	.30
29	make a mistake – do a mistake	.98	.03
39	lose weight – drop weight	.98	.20
7	receive a disease – catch a disease	.96	.33
28	commit a crime – make a crime	.96	.20
5	bright future – light future	.95	.35
17	sweep the floor – brush the floor	.95	.15
26	pay a visit – do a visit	.95	.40
27	draw a conclusion – take a conclusion	.95	.43
47	keep a promise – hold a promise	.95	.22
44	the clock strikes – the clock beats	.94	.07
51	keep a diary – run a diary	.94	.23
57	make apologies – do apologies	.94	.16
18	drop charges – lay down charges	.93	.21
33	hold a speech – give a speech	.93	.16
41	keep one's breath – hold one's breath	.93	.19
56	brush shoes – polish shoes	.92	.19
16	put out a fire – turn out a fire	.90	.40
43	drop count – lose count	.90	.18
30	good chance – strong chance	.89	-.14
46	drop bombs – fell bombs	.89	.34
60	heavy smoker – big smoker	.89	.24
48	tell one's prayers – say one's prayers	.88	.55
58	make sacrifices – do sacrifices	.88	.17
13	walk the streets – run the streets	.87	.17
59	run an errand – make an errand	.87	.29
22	do somebody a favour – make somebody a favour	.86	.32
63	make amends – do amends	.86	.09
37	go on a journey – do a journey	.84	.45
61	drive a motorcycle – ride a motorcycle	.84	.15
36	shed tears – fell tears	.82	.50
15	fake gun – false gun	.80	.25
19	seize an opportunity – grab an opportunity	.76	.32
55	poor visibility – bad visibility	.76	.41
65	slim chance – slender chance	.76	.29
23	reach a dream – realise a dream	.74	.34
4	exercise one's rights – employ one's rights	.71	.43
50	fast asleep – hard asleep	.71	.55
20	heavy rain – hard rain	.70	.00
62	blow a fuse – strike a fuse	.69	.38
25	fair hair – light hair	.68	.41
64	pay heed – show heed	.67	.34
31	keep one's balance – hold one's balance	.65	.12
45	a clear conscience – a clean conscience	.64	-.08
34	pursue a career – do a career	.63	.51
42	kick a habit – undo a habit	.59	.30
6	hold discussions – make discussions	.58	.30
3	set a deal – strike a deal	.54	.45
24	do damage – make damage	.54	.29
49	take root – make root	.54	.16
35	false teeth – fake teeth	.51	.16
21	bring charges – run charges	.50	.28
52	dress a wound – lay on a wound	.31	.38
53	push a bicycle – lead a bicycle	.29	-.18
10	strong competition – hard competition	.25	.29
40	foul weather – poor weather	.22	.03
	MEAN	.80	.23

APPENDIX 4A: Frequencies and z-scores from the BNC for items in COLLEX 3

Item	Item pair	Z-score for the leftmost word sequence; span L0, R3		Z-score for the rightmost word sequence; span L0, R3	
		BNC co-occurrence f	BNC z-score	BNC co-occurrence f	BNC z-score
1	set the bed - make the bed	2	-4.0	121	1.2[1]
2	put out a fire - turn out a fire	54	4.6	11	-1.8
3	employ one's rights - exercise one's rights	0	0	207	96.3
4	hold discussions - make discussions	71	13.4	8	-7.7
5	bright future - light future	94	54.7	0	0
6	receive a disease - catch a disease	0	0	18	6.3
7	hit a number - dial a number	9	-2.3	84	91.1
8	make an effort - commit an effort	826	80.7	2	-0.4
9	set a deal - strike a deal	6	-2.1	61	36.6
10	strong competition - hard competition	54	19.7	3	-1.0
11	sweep the floor - brush the floor	21	18.9	5	5.0
12	drop charges - lay down charges	37	14.0	13	3.6
13	grab an opportunity - seize an opportunity	8	5.8	140	25.8
14	bring charges - run charges	67	9.9	0	0
15	false gun - fake gun	0	0	4	15.0
16	do somebody a favour - make somebody a favour	27	-7.5	19	-3.3
17	reach a dream - realise a dream	0	0	11	7.0
18	do damage - make damage	187	2.8	9	-6.5
19	pay a visit - do a visit	160	43.6	0	0
20	ripe fruit - mature fruit	27	86.8	2	3.7
21	draw a conclusion - take a conclusion	203	89.4	0	0
22	make a crime - commit a crime	23	-4.4	215	160.8
23	keep one's balance - hold one's balance	80	18.8	65[2]	15.2
24	hold a speech - give a speech	0	0	59	4.4
25	fast asleep - hard asleep	157	243.3	0	0
26	pursue a career - do a career	64	55.5	4	0.4
27	fell tears - shed tears	0	0	65	149.8
28	go on a journey - do a journey	0[3]	0	19	-7.4
29	do sacrifices - make sacrifices	2	-3.7	89	31.3
30	poor visibility - bad visibility	33	67.4	4	5.9
31	hold one's breath - keep one's breath	321	118.7	0	0
32	kick a habit - undo a habit	36	55.6	0	0
33	drop count - lose count	0	0	58	34.1
34	take root - make root	134	22.5	6	-4.3
35	heavy smoker - big smoker	51	107.3	0	0
36	tell a prayer - say a prayer	0	0	143	20.9
37	keep a promise - hold a promise	104	43.8	36[4]	14.2
38	lay on a wound - dress a wound	0	0	12	20.5
39	fell bombs - drop bombs	0	0	43	38.2
40	slender chance - slim chance	3	4.5	18	22.5
41	keep a diary - run a diary	111	55.6	0	0
42	brush shoes - polish shoes	0	0	11	35.5
43	make apologies - do apologies	86	28.4	0	0
44	lose weight - drop weight	394	140.6	4	0.7
45	false teeth - fake teeth	77	102.5	0	0
46	run an errand - make an errand	53	95.7	0	0
47	drive a motorcycle - ride a motorcycle	0	0	9	38.3
48	blow a fuse - strike a fuse	17	72.4	0	0
49	show heed - pay heed	0	0	59	142.3
50	wide awake - clear awake	39	230.2	0	0

1 Considering the low z-score, a phrase query was made in which 22 instances were found of the verb make as a lemma + the bed.

2 An analysis of concordance lines revealed that as many as 33 of these instances contained the phrase "hold [lemma] the balance of power". Only 1 instance contained the lemmatized verb *hold* + a possessive pronoun followed by the noun *balance*.

3 Considering the surprising lack of hits, a phrase query was made in which 11 instances were found of the verb go as a lemma + on a journey.

4 An analysis of concordance lines revealed that 19 of these instances contained the phrase: hold out a promise.

APPENDIX 4B: Z-scores from the BNC for items in COLLMATCH 1

Item no	Item	Z-score retrieved from the BNC; span L0, R3	Note
1	drop charges	23.2	
2	drop patience	-0.6	
3	drop weight	0.5	
4	drop hints	85.9	
5	drop anchor	46.8	
6	drop blood	-0.4	
7	lose charges	-2.1	
8	lose patience	90.2	
9	lose weight	158.1	
10	lose hints	-0.7	
11	lose anchor	-0.7	
12	lose blood	2.5	
13	shed charges	-0.5	
14	shed patience	-0.2	
15	shed weight	71.9	
16	shed hints	-0.2	
17	shed anchor	-0.1	
18	shed blood	21.2	
19	break a diary	-1.0	
20	break one's balance	-0.7	
21	break a promise	21.8	
22	break sway	-0.4	
23	break one's breath	-0.9	
24	break a secret	0.1	
25	hold a diary	-0.9	
26	hold one's balance	13.3	constr.: hold a/the balance of power
27	hold a promise	13.4	constr.: holds little promise, hold out a promise
28	hold sway	171.0	
29	hold one's breath	127.6	
30	hold a secret	8.8	constr.: only in V + a + Adj meeting
31	keep a diary	54.4	
32	keep one's balance	18.8	
33	keep a promise	32.3	
34	keep sway	-0.7	
35	keep one's breath	-2.5	
36	keep a secret	101.2	
37	say a prayer	14.5	form: say
38	say a language	-3.6	form: say
39	say a joke	-1.4	form: say
40	say farewell	23.4	
41	say a story	-2.6	form: say
42	say lies	-2.4	form: say
43	tell a prayer	-1.0	
44	tell a language	-4.9	
45	tell a joke	7.7	
46	tell farewell	0.6	
47	tell a story	112.5	
48	tell lies	54.2	
49	speak a prayer	1.4	
50	speak a language	39.6	
51	speak a joke	-1.2	
52	speak farewell	0.8	
53	speak a story	-3.0	
54	speak lies	-1.3	
55	beat time	-1.1	
56	beat a play	-1.2	
57	beat eggs	22.8	
58	beat a blow	-1.2	
59	beat a divorce	-0.7	
60	beat a miracle	-0.5	
61	strike time	-2.5	
62	strike a play	2.2	constr.: 'stroke play' in golf
63	strike eggs	-0.6	

64	strike a blow	51.6	
65	strike a divorce	-0.4	
66	strike a miracle	-0.3	
67	perform time	-1.3	
68	perform a play	2.5	
69	perform eggs	0.2	
70	perform a blow	-0.9	
71	perform a divorce	-0.7	
72	perform a miracle	19.9	constr.: ‘perform miracles’ = z-score: 99.8
73	throw conclusions	0.4	
74	throw a glance	15.7	
75	throw a party	6.0	
76	throw a breath	-0.3	
77	throw a vote	-1.4	
78	throw parallels	-0.3	
79	cast conclusions	-0.5	
80	cast a glance	30.5	
81	cast a party	-0.9	
82	cast a breath	-0.7	
83	cast a vote	62.5	
84	cast parallels	-0.3	
85	draw conclusions	131.8	
86	draw a glance	1.3	
87	draw a party	-3.6	
88	draw a breath	77.0	
89	draw a vote	-1.5	
90	draw parallels	56.7	
91	take amends	-0.9	
92	take headway	-1.1	
93	take attention	-3.0	
94	take a decision	13.7	
95	take precautions	75.3	
96	take a mistake	-2.7	
97	make amends	141.0	
98	make headway	111.5	
99	make attention	-7.0	
100	make a decision	76.2	
101	make precautions	-1.2	
102	make a mistake	181.9	
103	pay amends	1.9	
104	pay headway	-0.5	
105	pay attention	303.2	
106	pay a decision	-3.6	
107	pay precautions	-0.8	
108	pay a mistake	7.9	constr.: ‘pay for x’s mistake’
109	fair weather	15.9	
110	fair colour	-1.6	
111	fair hair	98.7	
112	fair eyes	-2.0	
113	fair paint	0.3	
114	fair skin	12.8	
115	blonde weather	-0.4	includes spelling: ’blond’
116	blonde colour	5.9	includes spelling: ’blond’
117	blonde hair	557.1	includes spelling: ’blond’
118	blonde eyes	3.3	includes spelling: ’blond’
119	blonde paint	-0.3	includes spelling: ’blond’
120	blonde skin	3.8	includes spelling: ’blond’
121	light weather	2.2	
122	light colour	20.8	
123	light hair	18.7	
124	light eyes	8.3	constr.: only with colour modification: ‘light blue eyes’
125	light paint	1.9	
126	light skin	0.9	
127	hard meat	-0.8	
128	hard drugs	18.0	
129	hard facts	32.4	

130	hard drinker	-0.3	
131	hard traffic	-1.5	
132	hard demand	-1.9	
133	tough meat	5.0	
134	tough drugs	0.8	
135	tough facts	-0.7	
136	tough drinker	-0.1	
137	tough traffic	-0.7	
138	tough demand	-0.9	
139	heavy meat	-0.9	
140	heavy drugs	1.5	
141	heavy facts	-1.1	
142	heavy drinker	109.0	
143	heavy traffic	96.6	
144	heavy demand	15.6	

APPENDIX 4C: COLLMATCH 1

COLLMATCH 1

INSTRUKTION:

I detta test finner du 8 stycken tabeller. I varje tabell finner du tre stycken ord till vänster, skrivna under varandra, och 6 stycken ord ovanför tabellen, uppradade bredvid varandra.

Din uppgift är att utifrån vart och ett av orden till vänster om tabellen ta ställning till om ordet går att kombinera med något av de 6 uppradade orden ovanför tabellen. Om du anser att en kombination finns i det engelska språket, d.v.s. används av infödda talare, sätter du ett kryss i den cell där orden möts.

Exempel:

9

	suicide	a problem	damage	a murder	someone a favour	justice
solve		X		X		
commit	X			X		
do			X		X	X

I exemplet ovan har angivits att följande ordkombinationer finns i det engelska språket:

”solve a problem”	(lösa ett problem)
”solve a murder”	(lösa ett mord)
“commit suicide”	(begå självmord)
“commit a murder”	(begå ett mord)
“do damage”	(ställa till skada)
“do someone a favour”	(göra någon en tjänst)
“do justice”	(skipa rättvisa)

Tabell 1-6 utgörs av verb + substantiv (nominalfraser)
Tabell 7-8 utgörs av adjektiv + substantive

1

	charges	patience	weight	hints	anchor	blood
drop						
lose						
shed						

2

	a diary	one´s balance	a promise	sway	one's breath	a secret
break						
hold						
keep						

3

	a prayer	a language	a joke	farewell	a story	lies
say						
tell						
speak						

4

	time	a play	eggs	a blow	a divorce	a miracle
beat						
strike						
perform						

5

	conclusions	a glance	a party	a breath	a vote	parallels
throw						
cast						
draw						

6

	amends	headway	attention	a decision	precautions	a mistake
take						
make						
pay						

7

	weather	colour	hair	eyes	paint	skin
fair						
blonde						
light						

8

	meat	drugs	facts	drinker	traffic	demand
hard						
tough						
heavy						

Kontrollera att du inte har hoppat över någon uppgift
Tack för din medverkan i denna forskningsstudie om ordkunskap!

APPENDIX 4D: Item Facility (IF) and Item-Total Correlation (ITC) values of COLLEX 3 test administration

Item no.	Item pair	Item Facility	Corrected Item-total correlation
22	make a crime - commit a crime	1.00	0.00
5	bright future - light future	.99	.07
11	sweep the floor - brush the floor	.99	.23
37	keep a promise - hold a promise	.99	.01
6	receive a disease - catch a disease	.98	.33
7	hit a number - dial a number	.98	.38
8	make an effort - commit an effort	.98	.18
12	drop charges - lay down charges	.98	.13
20	ripe fruit - mature fruit	.98	.22
21	draw a conclusion - take a conclusion	.98	.28
39	fell bombs - drop bombs	.98	.29
36	tell a prayer - say a prayer	.97	.06
44	lose weight - drop weight	.97	.19
1	set the bed - make the bed	.96	.33
29	do sacrifices - make sacrifices	.96	.33
33	drop count - lose count	.96	.08
16	do somebody a favour - make somebody a favour	.95	.29
47	drive a motorcycle - ride a motorcycle	.95	.01
19	pay a visit - do a visit	.94	.54
27	fell tears - shed tears	.94	.44
28	go on a journey - do a journey	.94	.31
50	wide awake - clear awake	.94	.28
42	brush shoes - polish shoes	.93	.24
41	keep a diary - run a diary	.92	.31
2	put out a fire - turn out a fire	.91	.63
24	hold a speech - give a speech	.91	.20
43	make apologies - do apologies	.90	.35
35	heavy smoker - big smoker	.88	.41
3	employ one's rights - exercise one's rights	.85	.28
31	hold one's breath - keep one's breath	.85	.27
13	grab an opportunity - seize an opportunity	.82	.39
17	reach a dream - realise a dream	.81	.33
30	poor visibility - bad visibility	.81	.61
49	show heed - pay heed	.81	.17
15	false gun - fake gun	.80	.04
40	slender chance - slim chance	.80	.20
23	keep one's balance - hold one's balance	.79	.14
25	fast asleep - hard asleep	.79	.28
26	pursue a career - do a career	.79	.56
46	run an errand - make an errand	.78	.18
48	blow a fuse - strike a fuse	.75	.28
32	kick a habit - undo a habit	.72	.32
9	set a deal - strike a deal	.67	.53
18	do damage - make damage	.67	.37
38	lay on a wound - dress a wound	.64	.42
34	take root - make root	.63	-.03
45	false teeth - fake teeth	.59	.36
14	bring charges - run charges	.55	.00
4	hold discussions - make discussions	.54	.31
10	strong competition - hard competition	.53	.33
MEAN		.85	.27

APPENDIX 4E: Item Facility (IF) and Item-Total Correlation (ITC) values of COLLMATCH 1 test administration

Item no.	Item	Item Facility	Corrected Item-total correlation
1	drop charges	1.00	0.00
13	shed charges	1.00	0.00
23	break one's breath	1.00	0.00
41	say a story	1.00	0.00
42	say lies	1.00	0.00
44	tell a language	1.00	0.00
47	tell a story	1.00	0.00
51	speak a joke	1.00	0.00
53	speak a story	1.00	0.00
56	beat a play	1.00	0.00
69	perform eggs	1.00	0.00
85	draw conclusions	1.00	0.00
86	draw a glance	1.00	0.00
87	draw a party	1.00	0.00
96	take a mistake	1.00	0.00
106	pay a decision	1.00	0.00
115	blonde weather	1.00	0.00
134	tough drugs	1.00	0.00
7	lose charges	.99	.07
43	tell a prayer	.99	-.08
52	speak farewell	.99	-.04
67	perform time	.99	-.07
73	throw conclusions	.99	-.01
79	cast conclusions	.99	-.03
93	take attention	.99	-.07
119	blonde paint	.99	.05
2	drop patience	.98	.35
9	lose weight	.98	.04
10	lose hints	.98	-.03
14	shed patience	.98	.05
19	break a diary	.98	.42
25	hold a diary	.98	-.01
38	say a language	.98	.42
40	say farewell	.98	.42
45	tell a joke	.98	.42
46	tell farewell	.98	.42
50	speak a language	.98	.42
84	cast parallels	.98	-.16
89	draw a vote	.98	-.13
102	make a mistake	.98	.35
108	pay a mistake	.98	.35
37	say a prayer	.97	.30
49	speak a prayer	.97	.30
76	throw a breath	.97	.43
99	make attention	.97	.21
120	blonde skin	.97	.34
136	tough drinker	.97	.43
31	keep a diary	.96	.34
39	say a joke	.96	.20
60	beat a miracle	.96	.39
62	strike a play	.96	.28
70	perform a blow	.96	.10
137	tough traffic	.96	.02
3	drop weight	.95	.01

30	hold a secret	.95	.13
58	beat a blow	.95	.33
68	perform a play	.95	.19
105	pay attention	.95	.20
107	pay precautions	.95	.20
117	blonde hair	.95	-.05
118	blonde eyes	.95	.04
6	drop blood	.94	.17
17	shed anchor	.94	.00
20	break one's balance	.94	.20
48	tell lies	.94	.30
81	cast a party	.94	.29
29	hold one's breath	.93	.20
36	keep a secret	.93	.11
100	make a decision	.93	.27
8	lose patience	.92	.23
66	strike a miracle	.92	.12
71	perform a divorce	.92	-.04
78	throw parallels	.92	.15
59	beat a divorce	.91	.24
90	draw parallels	.90	.16
91	take amends	.90	.13
95	take precautions	.90	.32
104	pay headway	.90	.27
77	throw a vote	.89	.46
82	cast a breath	.89	.43
101	make precautions	.89	.16
142	heavy drinker	.89	.05
27	hold a promise	.88	-.02
75	throw a party	.88	.43
127	hard meat	.88	.25
135	tough facts	.88	.10
141	heavy facts	.88	.13
64	strike a blow	.87	.35
103	pay amends	.87	.22
16	shed hints	.86	.62
35	keep one's breath	.86	.08
139	heavy meat	.86	.02
34	keep sway	.85	.13
72	perform a miracle	.85	.35
11	lose anchor	.84	.19
21	break a promise	.84	.32
54	speak lies	.84	-.01
61	strike time	.84	.24
113	fair paint	.84	.27
143	heavy traffic	.84	.07
22	break sway	.82	.09
130	hard drinker	.81	-.33
132	hard demand	.81	.13
24	break a secret	.80	.06
63	strike eggs	.80	.24
110	fair colour	.80	-.31
129	hard facts	.80	.19
131	hard traffic	.78	.01
122	light colour	.77	.02
123	light hair	.77	.07
83	cast a vote	.76	.35

116	blonde colour	.76	.18
55	beat time	.75	.15
114	fair skin	.74	.19
92	take headway	.72	.15
97	make amends	.72	.29
18	shed blood	.69	.33
5	drop anchor	.67	.15
121	light weather	.67	.15
4	drop hints	.64	.52
33	keep a promise	.64	.16
124	light eyes	.64	.19
88	draw a breath	.62	.48
65	strike a divorce	.61	.08
125	light paint	.60	.10
109	fair weather	.59	.13
80	cast a glance	.58	.05
126	light skin	.58	.06
12	lose blood	.57	.05
133	tough meat	.57	.15
112	fair eyes	.55	-.27
57	beat eggs	.54	.15
74	throw a glance	.54	.03
26	hold one's balance	.51	-.09
111	fair hair	.49	.34
32	keep one's balance	.46	.21
138	tough demand	.39	-.01
98	make headway	.36	.25
128	hard drugs	.33	.16
140	heavy drugs	.30	.04
144	heavy demand	.25	.18
28	hold sway	.21	.09
94	take a decision	.15	-.32
15	shed weight	.04	.19
MEAN		.84	.14

APPENDIX 4F: COLLMATCH 2

COLLMATCH 2

INSTRUKTION:

Denna testdel innehåller 20 (1-20) frågor. Varje fråga innehåller 5 engelska ordsekvenser. De 5 ordsekvenserna utgörs både av vanligt förekommande engelska ordkombinationer (kallas här: rätta), och ordkombinationer som inte förekommer naturligt i det engelska språket (kallas här: felaktiga).

Din uppgift är att välja ut de i engelska språket förekommande ordkombinationerna, genom att sätta ett kryss i rutan nedanför dessa. Observera att antalet ”rätta” och ”felaktiga” svar i varje fråga varierar!

Varje rätt besvarad sekvens i en fråga ger 0,5 poäng, och varje felaktigt besvarad sekvens ger 0 poäng.

MAXPOÄNG I DENNA TESTDEL: 50 poäng

EXEMPEL:

Nedan har ‘pay attention’, ’pay lip-service’ och ’pay fees’ markerats som naturligt före-kommande ordkombinationer i engelska språket och alternativ a) och c) har bedömts som felaktiga genom utebliven markering. Detta skulle resultera i full poäng på vår exempelfråga, följaktligen 5 x 0,5 = 2,5 poäng.

21	a. pay patience	b. pay attention	c. pay an assumption	d. pay lip-service	e. pay fees
	☐	☒	☐	☒	☒

1 a. have a say ☐ b. have a look ☐ c. have an experience ☐ d. have doubts ☐ e. have boredom ☐

2 a. lose one's temper ☐ b. lose pretence ☐ c. lose sleep ☐ d. lose fever ☐ e. lose weight ☐

3 a. do an effort ☐ b. do justice ☐ c. do harm ☐ d. do time ☐ e. do the trick ☐

4 a. draw the curtains ☐ b. draw a sword ☐ c. draw a favour ☐ d. draw a breath ☐ e. draw blood ☐

5 a. say a poem ☐ b. say farewell ☐ c. say grace ☐ d. say a riddle ☐ e. say a prayer ☐

6 a. break a heart ☐ b. break a journey ☐ c. break news ☐ d. break a habit ☐ e. break a reputation ☐

7 a. make a decision ☐ b. make an insult ☐ c. make sense ☐ d. make amends ☐ e. make a hug ☐

8 a. raise objections ☐ b. raise oaths ☐ c. raise a tackle ☐ d. raise suspicion ☐ e. raise money ☐

9 a. take face ☐ b. take precautions ☐ c. take progress ☐ d. take headway ☐ e. take drugs ☐

10 a. bear respect ☐ b. bear arms ☐ c. bear a call ☐ d. bear guilt ☐ e.bear witness ☐

11 a. give a speech ☐ b. give birth ☐ c. give advice ☐ d. give orders ☐ e. give place ☐

12 a. serve a purpose ☐ b. serve a sentence ☐ c. serve reason ☐ d. serve a crime ☐ e. serve apologies ☐

13 a. keep a diary ☐ b. keep approval ☐ c. keep one's balanc ☐ d. keep pets ☐ e. keep a secret ☐

14 a. catch a disease ☐ b. catch a bus ☐ c. catch a glimpse ☐ d. catch fire ☐ e. catch a look ☐

15 a. hold one's breath ☐ b. hold meetings ☐ c. hold one's calm ☐ d. hold trouble ☐ e. hold grudges ☐

16 a. pull a trigger ☐ b. pull respect ☐ c. pull punches ☐ d. pull rank ☐ e. pull a face ☐

17 a. run a danger ☐ b. run errands ☐ c. run a bath ☐ d. run a risk ☐ e. run a business ☐

18 a. throw a glimpse ☐ b. throw light ☐ c. throw hesitation ☐ d. throw a party ☐ e. throw importance ☐

19 a. set a failure ☐ b. set a trap ☐ c. set sail ☐ d. set an example ☐ e. set pressure ☐

20 a. drop bombs ☐ b. drop patience ☐ c. drop hints ☐ d. drop anchor ☐ e. drop one's memory ☐

APPENDIX 4G: Frequencies and z-scores from the BNC for items in COLLEX 4

Item	Item pair	Values for the leftmost word sequence; span L0, R3		Values for the rightmost wor sequence; span L0, R3	
		BNC co-occurrence f	BNC z-score	BNC co-occurrence f	BNC z-score
1	do damage – make damage	187	2.8	9	-6.5
2	put out a fire - turn out a fire	54	4.6	11	-1.8
3	lay a vote – cast a vote	0	0	93	99.2
4	hold discussions - make discussions	71	13.4	8	-7.7
5	bright future - light future	94	54.7	0	0
6	receive a cold - catch a cold	0	0	43	41.8
7	pay a visit - do a visit	160	43.6	0	0
8	strike a pose – hit a pose	15	51.0	0	0
9	fell tears - shed tears	0	0	65	149.8
10	strong competition - hard competition	54	19.7	3	-1.0
11	sweep the floor - brush the floor	21	18.9	5	5.0
12	employ one's rights - exercise one's rights	0	0	207	96.3
13	grab an opportunity - seize an opportunity	8	5.8	140	25.8
14	bring charges - run charges	67	9.9	0	0
15	false gun - fake gun	0	0	4	15.0
16	do somebody a favour - make somebody a favour	27	-7.5	19	-3.3
17	lodge a complaint – perform a complaint	24	60.5	0	0
18	set the bed - make the bed	2	-4.0	121	1.2[1]
19	hit a number - dial a number	9	-2.3	84	91.1
20	ripe fruit - mature fruit	27	86.8	2	3.7
21	draw a conclusion - pull a conclusion	203	89.4	0	0
22	perform suicide – commit suicide	0	0	309	505.2
23	tell a prayer - say a prayer	0	0	143	20.9
24	hold a speech - give a speech	0	0	59	4.4
25	fast asleep - hard asleep	157	243.3	0	0
26	pursue a career - do a career	64	55.5	4	0.4
27	set a deal - strike a deal	6	-2.1	61	36.6
28	go on a journey - do a journey	0[3]	0	19	-7.4
29	do sacrifices - make sacrifices	2	-3.7	89	31.3
30	poor visibility - bad visibility	33	67.4	4	5.9
31	hold one's breath - keep one's breath	321	118.7	0	0
32	direct an orchestra – conduct an orchestra	0	0	11	21.0
33	drop count - lose count	0	0	58	34.1
34	take root - make root	134	22.5	6	-4.3
35	heavy smoker - big smoker	51	107.3	0	0
36	keep one's balance - hold one's balance	80	18.8	65[2]	15.2
37	take one's revenge – make one's revenge	95	38.1	0	0
38	lay on a wound - dress a wound	0	0	12	20.5
39	fell bombs - drop bombs	0	0	43	38.2
40	slender chance - slim chance	3	4.5	18	22.5
41	keep a diary - run a diary	111	55.6	0	0
42	brush shoes - polish shoes	0	0	11	35.5
43	make apologies - do apologies	86	28.4	0	0
44	whip eggs – beat eggs	2	4.8	41	32.3
45	false teeth - fake teeth	77	102.5	0	0
46	make an attempt – do an attempt	534	48.4	227[4]	8.2
47	clench one's fist – tie one's fist	90	592.1	0	0
48	blow a fuse - strike a fuse	17	72.4	0	0
49	show heed - pay heed	0	0	59	142.3
50	wide awake - clear awake	39	230.2	0	0

1 Considering the low z-score, a phrase query was made in which 22 instances were found of the verb make as a lemma + the bed.

2 An analysis of concordance lines revealed that as many as 33 of these instances contained the phrase “hold [lemma] the balance of power”. Only 1 instance contained the lemmatized verb hold + a possessive pronoun followed by the noun balance.
3 Considering the surprising lack of hits, a phrase query was made in which 11 instances were found of the verb go as a lemma + on a journey.
4 An analysis of concordance lines revealed that these instances were made up of phrases like do attempt to do smtg, and do not attempt to do smtg.

APPENDIX 4H: COLLEX 4

COLLEX 4

INSTRUKTION:

Denna testdel innehåller 50 (1-50) frågor. Varje fråga innehåller två engelska ordsekvenser, den ena markerad med a) och den andra med b). Din uppgift är att välja en av de två sekvenserna i varje fråga.

Den ena av de två ordsekvenserna i varje fråga är en naturlig och vanligt förekommande sekvens i engelska språket medan den andra inte är det. Välj den ordsekvens som du bedömer är den naturligaste och vanligast förekommande genom att sätta ett kryss i högermarginalen i den kolumn som motsvarar ditt val.

Varje rätt besvarad fråga ger 0,5 poäng, och varje felaktigt besvarad fråga ger 0 poäng. Om du inte kryssar i någon av rutorna i en fråga, eller kryssar i bägge, får du 0 poäng.

EXEMPEL:

Nedan har ‘solve a problem’ markerats som svar på fråga 51, och ‘make a mistake’ har markerats som svar på fråga 52.

						a	b
51	a)	solve a problem	b)	break a problem		X	
52	a)	do a mistake	b)	make a mistake			X

						a	b
1	a)	do damage	b)	make damage			
2	a)	put out a fire	b)	turn out a fire			
3	a)	lay a vote	b)	cast a vote			
4	a)	hold discussions	b)	make discussions			
5	a)	bright future	b)	light future			
6	a)	receive a cold	b)	catch a cold			
7	a)	pay a visit	b)	do a visit			
8	a)	strike a pose	b)	hit a pose			
9	a)	fell tears	b)	shed tears			
10	a)	strong competition	b)	hard competition			
11	a)	sweep the floor	b)	brush the floor			
12	a)	employ one's rights	b)	exercise one's rights			
13	a)	grab an opportunity	b)	seize an opportunity			
14	a)	bring charges	b)	run charges			
15	a)	false gun	b)	fake gun			
16	a)	do somebody a favour	b)	make somebody a favour			
17	a)	lodge a complaint	b)	perform a complaint			
18	a)	set the bed	b)	make the bed			
19	a)	hit a number	b)	dial a number			
20	a)	mature fruit	b)	ripe fruit			
21	a)	draw a conclusion	b)	pull a conclusion			
22	a)	perform suicide	b)	commit suicide			
23	a)	tell a prayer	b)	say a prayer			
24	a)	hold a speech	b)	give a speech			
25	a)	fast asleep	b)	hard asleep			

						a	b
26	a)	pursue a career	b)	do a career			
27	a)	set a deal	b)	strike a deal			
28	a)	go on a journey	b)	do a journey			
29	a)	do sacrifices	b)	make sacrifices			
30	a)	poor visibility	b)	bad visibility			
31	a)	hold one’s breath	b)	keep one’s breath			
32	a)	direct an orchestra	b)	conduct an orchestra			
33	a)	drop count	b)	lose count			
34	a)	take root	b)	make root			
35	a)	heavy smoker	b)	big smoker			
36	a)	hold one’s balance	b)	keep one’s balance			
37	a)	take one’s revenge	b)	make one’s revenge			
38	a)	lay on a wound	b)	dress a wound			
39	a)	fell bombs	b)	drop bombs			
40	a)	slender chance	b)	slim chance			
41	a)	keep a diary	b)	run a diary			
42	a)	brush shoes	b)	polish shoes			
43	a)	make apologies	b)	do apologies			
44	a)	whip eggs	b)	beat eggs			
45	a)	false teeth	b)	fake teeth			
46	a)	make an attempt	b)	do an attempt			
47	a)	clench one’s fist	b)	tie one’s fist			
48	a)	blow a fuse	b)	strike a fuse			
49	a)	show heed	b)	pay heed			
50	a)	wide awake	b)	clear awake			

APPENDIX 4I: Item Facility (IF) and Item-Total Correlation (ITC) values of COLLEX 4 test administration

Item no.	Item pair	Item Facility	Corrected Item-total correlation
22	perform suicide – commit suicide	.98	.18
5	bright future - light future	.96	.28
11	sweep the floor - brush the floor	.96	.29
18	set the bed - make the bed	.95	.31
39	fell bombs - drop bombs	.95	.14
29	do sacrifices - make sacrifices	.94	.31
19	hit a number - dial a number	.93	.43
46	make an attempt – do an attempt	.93	.28
6	receive a cold - catch a cold	.92	.47
16	do somebody a favour - make somebody a favour	.92	.39
21	draw a conclusion - pull a conclusion	.92	.33
23	tell a prayer - say a prayer	.92	.43
35	heavy smoker - big smoker	.90	.34
41	keep a diary - run a diary	.90	.30
2	put out a fire - turn out a fire	.89	.31
15	false gun - fake gun	.89	.23
28	go on a journey - do a journey	.89	.35
42	brush shoes - polish shoes	.88	.32
40	slender chance - slim chance	.87	.35
43	make apologies - do apologies	.87	.33
7	pay a visit - do a visit	.86	.59
31	hold one's breath - keep one's breath	.86	.28
9	fell tears - shed tears	.84	.52
24	hold a speech - give a speech	.84	.38
33	drop count - lose count	.84	.30
8	strike a pose – hit a pose	.82	.50
20	ripe fruit - mature fruit	.82	.55
25	fast asleep - hard asleep	.81	.36
37	take one’s revenge – make one’s revenge	.81	.35
36	keep one's balance - hold one's balance	.80	.36
49	show heed - pay heed	.79	.35
30	poor visibility - bad visibility	.77	.58
50	wide awake - clear awake	.77	.63
4	hold discussions - make discussions	.74	.39
12	employ one's rights - exercise one's rights	.71	.49
32	direct an orchestra – conduct an orchestra	.71	.36
26	pursue a career - do a career	.70	.47
17	lodge a complaint – perform a complaint	.69	.59
13	grab an opportunity - seize an opportunity	.67	.52
3	lay a vote – cast a vote	.66	.41
38	lay on a wound - dress a wound	.65	.43
48	blow a fuse - strike a fuse	.65	.40
47	clench one’s fist – tie one’s fist	.63	.54
1	do damage – make damage	.61	.63
34	take root - make root	.58	.21
45	false teeth - fake teeth	.51	.33
14	bring charges - run charges	.48	.04
27	set a deal - strike a deal	.47	.54
44	whip eggs – beat eggs	.45	.33
10	strong competition - hard competition	.43	.39
MEAN		.79	.38

APPENDIX 4J Item Facility (IF) and Item-Total Correlation (ITC) values of COLLMATCH 2 test administration

Item no.	Item pair	Item Facility	Corrected Item-total correlation
1	have a say	.52	.66
2	have a look	.97	.12
3	have an experience	.65	.01
4	have doubts	.89	.56
5	have boredom	.98	.01
6	lose one's temper	.94	.46
7	lose pretence	.84	-.01
8	lose sleep	.40	.61
9	lose fever	.91	.25
10	lose weight	.97	.28
11	do an effort	.76	.48
12	do justice	.65	.36
13	do harm	.83	.48
14	do time	.63	.59
15	do the trick	.83	.32
16	draw the curtains	.71	.32
17	draw a sword	.86	.28
18	draw a favour	.94	.33
19	draw a breath	.43	.29
20	draw blood	.35	.52
21	say a poem	.86	.33
22	say farewell	.89	.28
23	say grace	.63	.63
24	say a riddle	.90	.15
25	say a prayer	.93	.29
26	break a heart	.91	.38
27	break a journey	.06	-.41
28	break news	.75	.24
29	break a habit	.88	.30
30	break a reputation	.78	.44
31	make a decision	.97	.26
32	make an insult	.73	.09
33	make sense	.96	.34
34	make amends	.60	.68
35	make a hug	.96	.25
36	raise objections	.57	.40
37	raise oaths	.92	.24
38	raise a tackle	.93	.11
39	raise suspicion	.67	.43
40	raise money	.93	.38
41	take face	.94	.24
42	take precautions	.74	.63
43	take progress	.81	.35
44	take headway	.86	.04
45	take drugs	.88	.11
46	bear respect	.77	.07
47	bear arms	.57	.35
48	bear a call	.98	.11
49	bear guilt	.49	-.01
50	bear witness	.74	.42

Cont.

Item no.	Item pair	Item Facility	Corrected Item-total correlation
51	give a speech	.87	.25
52	give birth	.91	.54
53	give advice	.95	.19
54	give orders	.97	.22
55	give place	.84	.16
56	serve a purpose	.84	.48
57	serve a sentence	.44	.38
58	serve reason	.90	.12
59	serve a crime	.65	.31
60	serve apologies	.92	.24
61	keep a diary	.91	.36
62	keep approval	.97	.06
63	keep one's balance	.86	.31
64	keep pets	.55	.49
65	keep a secret	.99	.21
66	catch a disease	.80	.42
67	catch a bus	.96	.27
68	catch a glimpse	.70	.70
69	catch fire	.70	.60
70	catch a look	.74	.41
71	hold one's breath	.92	.26
72	hold meetings	.74	.23
73	hold one's calm	.90	.19
74	hold trouble	.98	.25
75	hold grudges	.49	.61
76	pull a trigger	.91	.42
77	pull respect	.97	.25
78	pull punches	.14	.07
79	pull rank	.23	.39
80	pull a face	.48	-.04
81	run a danger	.85	.21
82	run errands	.78	.72
83	run a bath	.45	.48
84	run a risk	.67	.08
85	run a business	.95	.38
86	throw a glimpse	.53	-.12
87	throw light	.39	.02
88	throw hesitation	.95	.24
89	throw a party	.83	.53
90	throw importance	.96	.23
91	set a failure	.94	.33
92	set a trap	.90	.33
93	set sail	.72	.47
94	set an example	.84	.54
95	set pressure	.76	.36
96	drop bombs	.94	.39
97	drop patience	.88	.33
98	drop hints	.60	.57
99	drop anchor	.63	.58
100	drop one's memory	.82	.54
MEAN		.77	.32

APPENDIX 5A: COLLEX 5 – PILOT VERSION

COLLEX 5 - PILOT VERSION

INSTRUKTION:

Denna testdel innehåller 40 (1-40) frågor. Varje fråga innehåller tre ordkombinationer markerade med a), b) respektive c). Din uppgift är att välja **en** av de tre ordkombinationerna i varje fråga.

En de tre ordkombinationerna i varje fråga är en naturlig och vanligt förekommande sekvens i det engelska språket medan de andra två inte är det. Välj den ordsekvens som du bedömer är den naturligaste och vanligast förekommande genom att ringa i den.

INSTRUCTION:

This part consists of 40 test items (1-40). Each test item contains three word combinations marked a), b), and c). Your task is to choose **one** of the three word combinations in each item.

One of the three word combinations in each item is a natural and frequent word combination occurring in the English language, whereas the other two are not. Choose the word combination you think is the most natural and frequently occurring by ticking the box that corresponds to it in the right margin.

1	a.	do damage	b.	make damage	c.	run damage
2	a.	turn out a fire	b.	put out a fire	c.	set out a fire
3	a.	hold discussions	b.	make discussions	c.	set discussions
4	a.	receive a cold	b.	achieve a cold	c.	catch a cold
5	a.	do a visit	b.	hit a visit	c.	pay a visit
6	a.	strike a pose	b.	lead a pose	c.	hit a pose
7	a.	fell tears	b.	shed tears	c.	raise tears
8	a.	employ one’s rights	b.	exercise one’s rights	c.	conduct one’s rights
9	a.	grab an opportunity	b.	seize an opportunity	c.	catch an opportunity
10	a.	bring charges	b.	run charges	c.	push charges
11	a.	lend a complaint	b.	perform a complaint	c.	lodge a complaint
12	a.	make a conclusion	b.	pull a conclusion	c.	draw a conclusion
13	a.	commit a crime	b.	comply a crime	c.	conduct a crime
14	a.	tell a prayer	b.	say a prayer	c.	speak a prayer
15	a.	give a speech	b.	hold a speech	c.	perform a speech
16	a.	strike a deal	b.	set a deal	c.	step a deal
17	a.	go on a journey	b.	do a journey	c.	pull a journey
18	a.	keep one’s breath	b.	house one’s breath	c.	hold one’s breath
19	a.	direct an orchestra	b.	conduct an orchestra	c.	control an orchestra
20	a.	lose count	b.	drop count	c.	pass count

Fortsättning på nästa sida/continued overleaf

21	a.	take root	b.	make root	c. stick root
22	a.	hold one’s balance	b.	keep one’s balance	c. last one’s balance
23	a.	take one’s revenge	b.	make one’s revenge	c. obtain one’s revenge
24	a.	keep a diary	b.	run a diary	c. tend a diary
25	a.	brush shoes	b.	polish shoes	c. tidy shoes
26	a.	make apologies	b.	do apologies	c. lay apologies
27	a.	tie one’s fist	b.	fix one’s fist	c. clench one’s fist
28	a.	strike a fuse	b.	knock a fuse	c. blow a fuse
29	a.	show heed	b.	pay heed	c. spread heed
30	a.	make an escape	b.	take an escape	c. draw an escape
31	a.	lose faith	b.	drop faith	c. cut faith
32	a.	perform a survey	b.	commit a survey	c. conduct a survey
33	a.	push a bike	b.	lead a bike	c. press a bike
34	a.	send judgement	b.	pass judgement	c. set judgement
35	a.	say one’s mind	b.	speak one’s mind	c. talk one’s mind
36	a.	spoil the fun	b.	ruin the fun	c. destroy the fun
37	a.	earn a purpose	b.	win a purpose	c. serve a purpose
38	a.	make friends	b.	create friends	c. gain friends
39	a.	make measures	b.	take measures	c. stick measures
40	a.	speak shop	b.	say shop	c. talk shop

APPENDIX 5B: Word frequencies for COLLEX 5 – PILOT VERSION

JACET 8000 Frequency band	verbs	nouns
1K	achieve, bring, catch, control, create, cut, do, draw, drop, exercise, give, go, hit, hold, house keep, last, lay, lead, lose, make, pass, pay, press, pull, push, put, raise, receive, run, say, send, serve set, show, speak, spread, step stick, strike, take, talk, tell tend, turn, win	damage, fire, discussion, cold visit, tear, right, opportunity speech, deal, count, escape mind, purpose, friend, measure, shop
2K	blow, brush, conduct, destroy direct, earn, fix, gain, grab, knock, obtain, perform, tie	charge, conclusion, crime journey, breath, root, balance diary, shoes, faith, survey, fun
3K	lend, ruin, seize, shed, spoil	complaint, prayer, orchestra fist, bike, judgement
4K	commit, comply, employ, polish	pose
5K	lodge, tidy	apologies
6K		revenge, fuse
OFF LIST	clench, fell	heed

APPENDIX 5C: COLLMATCH 3 – PILOT VERSION

COLLMATCH 3 - PILOT VERSION

PROVDEL 1

INSTRUKTION:

Denna testdel innehåller 100 ordkombinationer. Din uppgift är att avgöra om ordkombinationerna förekommer i det engelska språket eller inte.

Om du bedömer att en ordkombination finns i det engelska språket, sätt ett kryss i rutan ’ja’.
Om du bedömer att en ordkombination *inte* finns i det engelska språket, sätt ett kryss i rutan ’nej’.

Kontrollera att du avgivit svar för samtliga ordkombinationer.

TEST PART 1

INSTRUCTION:

This part consists of 100 word combinations (1-100). Your task is to decide whether the word combinations exist in use in the English language or not.

If you think a word combination exists in use in the English language, tick the ‘yes’ box. If you don’t think a word combination exists in use in the English language, tick the ‘no’ box.

Please make sure you have answered all test items.

PART A

1	2	3	4	5
have a say	lose sleep	do justice	draw a breath	turn a reason
☐ ja	☐ ja	☐ ja	☐ ja	☐ ja
☐ nej	☐ nej	☐ nej	☐ nej	☐ nej
6	7	8	9	10
say grace	pick a glance	break news	make progress	claim trade
☐ ja	☐ ja	☐ ja	☐ ja	☐ ja
☐ nej	☐ nej	☐ nej	☐ nej	☐ nej
1 1	12	13	14	15
raise objection	bear witness	supply one's assistance	give a speech	serve a sentenc
☐ ja	☐ ja	☐ ja	☐ ja	☐ ja
☐ nej	☐ nej	☐ nej	☐ nej	☐ nej
16	17	18	19	20
stretch a regar	restore a favou	keep pets	catch fire	hold meetings
☐ ja	☐ ja	☐ ja	☐ ja	☐ ja
☐ nej	☐ nej	☐ nej	☐ nej	☐ nej
21	22	23	24	25
pull a face	run a bath	throw a party	shake a smile	set an example
☐ ja	☐ ja	☐ ja	☐ ja	☐ ja
☐ nej	☐ nej	☐ nej	☐ nej	☐ nej

PART B

26	27	28	29	30
fetch an illness	drop hints	play a trick	pay attention	meet a need
☐ ja	☐ ja	☐ ja	☐ ja	☐ ja
☐ nej	☐ nej	☐ nej	☐ nej	☐ nej
31	32	33	34	35
reach a conclusion	drag a limit	gather a matter	assume responsibility	suffer damage
☐ ja	☐ ja	☐ ja	☐ ja	☐ ja
☐ nej	☐ nej	☐ nej	☐ nej	☐ nej
36	37	38	39	40
cut a corner	fly a flag	realise a potential	sink speed	fit the bill
☐ ja	☐ ja	☐ ja	☐ ja	☐ ja
☐ nej	☐ nej	☐ nej	☐ nej	☐ nej
41	42	43	44	45
push one's luc	gain ground	perform a miracle	win one's memory	impose success
☐ ja	☐ ja	☐ ja	☐ ja	☐ ja
☐ nej	☐ nej	☐ nej	☐ nej	☐ nej
46	47	48	49	50
adopt an approach	clear one's throat	strike a blow	beat eggs	employ a technique
☐ ja	☐ ja	☐ ja	☐ ja	☐ ja
☐ nej	☐ nej	☐ nej	☐ nej	☐ nej

PART C

51	press charges	52	settle a dispute	53	swing a secret	54	grant permission	55	express a worry
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
56	rule an award	57	commit a sin	58	launch a campaign	59	stick one's mood	60	acquire a skill
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
61	deliver a speech	62	spread one's wings	63	assess damage	64	afford an opportunity	65	ride a storm
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
66	jump a queue	67	score problems	68	roll a look	69	exercise discretion	70	blow one's nose
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
71	rush rank	72	steal someone's thunder	73	dress a wound	74	pursue a career	75	challenge a view
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej

PART D

76	knock a concern	77	lay pressure	78	pack an affair	79	abandon ship	80	clean windows
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
81	dismiss an idea	82	shift gear	83	justify one's existence	84	bind blood	85	charge respect
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
86	cast a vote	87	kick one's heels	88	bend a rule	89	fill an aim	90	lend support
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
91	sustain an injury	92	hit approval	93	cease fire	94	snap one's fingers	95	shrug one's shoulders
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
96	stand an occasion	97	grab a hold	98	sit seed	99	fall a failure	100	file a report
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej

APPENDIX 5D: Word frequencies for COLLMATCH 3 – PILOT VERSION

JACET 8000 Frequency band	verbs	nouns
1K	have, lose, do, draw, turn, say, pick, break, make, raise, bear, supply, give, serve, keep, catch, hold, pull, run, throw, shake, set, drop, pay, play, meet, reach, suffer, cut, fly, realize, push, win, clear, strike, beat, press, express, rule, stick, spread, ride, jump, exercise, dress, challenge, lay, clean, fill, hit, stand, sit, fall	sleep, reason, news, progress, trade speech, sentence, fire, meeting, face, party, smile, example, attention, need, limit, matter, damage, corner, speed, ground, memory, success, approach, secret, worry, skill, opportunity, problem, look, career, view, concern, pressure, ship, window, idea, blood respect, vote, rule, support, fire, finger, shoulder, hold, report
2K	claim, stretch, drag, gather, assume, sink, fit, gain, perform, adopt, settle, swing, grant, acquire deliver, afford, score, roll, blow, rush, knock, pack, shift, charge, cast, kick, bend, lend, grab	justice, breath, glance, witness, regard, favour, bath, illness, conclusion, responsibility, potentia bill, luck, miracle, throat, blow, egg technique, charge, award, campaig mood, wing, storm, nose, wound, affair, existence, aim, injury, occasion, seed, failure
3K	restore, fetch, steal, pursue, snap, shrug, file	pet, hint, flag, permission, discretion, rank, gear, heel
4K	impose, employ, commit, launch, assess, abandon, dismiss, justify, bind, sustain, cease	grace, objection, assistance, dispute sin, approval
5K		queue, thunder
OFF LIST		say[1]

1 This word form did not exist in JACET 8000 as a noun, but the verb form was found in the 1K range.

APPENDIX 5E: Test validation questionnaire

Questions about the test that you just finished.

Put a cross in the scale under each question, and feel free to add any comments you might have on the dotted line under the scale.

1 How did you perceive the test instruction?

Very easy to understand	Easy to understand	OK	hard to understand	very hard to understand
<--------|----------|----------|----------|----------|------->				

Comment: …….

2 How did you perceive the difficulty level of the test?

very easy	easy	average	difficult	very difficult
<--------|----------|----------|----------|----------|------->				

Comment: …….

3 What feeling do you associate with the process of doing the test?

very appealing	appealing	OK	boring	very boring
<--------|----------|----------|----------|----------|------->				

Comment: …….

4 In your opinion, what kind of knowledge is measured in the test?

……

……

……

……

APPENDIX 5F: Item Facility (IF) and Item-Total Correlation (ITC) values of COLLEX 5 – PILOT VERSION test administration

Item no.	Item triple			Item Facility	Corrected Item total correlatio
4	receive a cold	achieve a cold	catch a cold	1.00	0.00
5	do a visit	hit a visit	pay a visit	1.00	0.00
6	strike a pose	lead a pose	hit a pose	1.00	0.00
24	keep a diary	run a diary	tend a diary	1.00	0.00
25	brush shoes	polish shoes	tidy shoes	1.00	0.00
31	lose faith	drop faith	cut faith	1.00	0.00
35	say one's mind	speak one's mind	talk one's mind	1.00	0.00
38	make friends	create friends	gain friends	1.00	0.00
12	make a conclusion	pull a conclusion	draw a conclusion	.96	.07
13	commit a crime	comply a crime	conduct a crime	.96	.20
14	tell a prayer	say a prayer	speak a prayer	.96	.20
26	make apologies	do apologies	lay apologies	.96	-.06
30	make an escape	take an escape	draw an escape	.96	.26
37	earn a purpose	win a purpose	serve a purpose	.96	.26
2	turn out a fire	put out a fire	set out a fire	.92	.34
3	hold discussions	make discussions	set discussions	.92	-.22
7	fell tears	shed tears	raise tears	.92	.39
17	go on a journey	do a journey	pull a journey	.92	-.04
18	keep one's breath	house one's breath	hold one's breath	.92	-.27
20	lose count	drop count	pass count	.92	.25
32	perform a survey	commit a survey	conduct a survey	.92	.01
34	send judgement	pass judgement	set judgement	.92	.44
39	make measures	take measures	stick measures	.92	.20
36	spoil the fun	ruin the fun	destroy the fun	.84	.02
23	take one's revenge	make one's revenge	obtain one's revenge	.80	-.19
29	show heed	pay heed	spread heed	.80	.44
28	strike a fuse	knock a fuse	blow a fuse	.76	.36
11	lend a complaint	perform a complaint	lodge a complaint	.72	.03
15	give a speech	hold a speech	perform a speech	.72	.14
19	direct an orchestra	conduct an orchestra	control an orchestra	.72	-.06
22	hold one's balance	keep one's balance	last one's balance	.72	.06
8	employ one's rights	exercise one's rights	conduct one's rights	.68	.24
9	grab an opportunity	seize an opportunity	catch an opportunity	.68	.45
27	tie one's fist	fix one's fist	clench one's fist	.64	.26
1	do damage	make damage	run damage	.60	.57
40	speak shop	say shop	talk shop	.60	.16
16	strike a deal	set a deal	step a deal	.52	.49
21	take root	make root	stick root	.52	.23
33	push a bike	lead a bike	press a bike	.48	-.03
10	bring charges	run charges	push charges	.16	-.11
			Mean	.83	.13

APPENDIX 5G: Item Facility (IF) and Item-Total Correlation (ITC) values of COLLMATCH 3 – PILOT VERSION test administration

Item no.	Item pair	Item Facility	Corrected Item-total correlation
1	have a say	1.00	0.00
3	do justice	1.00	0.00
9	make progress	1.00	0.00
14	give a speech	1.00	0.00
23	throw a party	1.00	0.00
25	set an example	1.00	0.00
29	pay attention	1.00	0.00
42	gain ground	1.00	0.00
47	clear one’s throat	1.00	0.00
54	grant permission	1.00	0.00
60	acquire a skill	1.00	0.00
62	spread one’s wings	1.00	0.00
70	blow one’s nose	1.00	0.00
5	turn a reason	.96	.15
12	bear witness	.96	.23
20	hold meetings	.96	.08
28	play a trick	.96	-.27
32	drag a limit	.96	.42
35	suffer damage	.96	.13
51	press charges	.96	.42
52	settle a dispute	.96	.10
57	commit a sin	.96	.10
58	launch a campaign	.96	.10
61	deliver a speech	.96	-.03
68	roll a look	.96	.15
74	pursue a career	.96	.10
82	shift gear	.96	.15
83	justify one's existence	.96	.10
94	snap one's fingers	.96	-.37
95	shrug one's shoulders	.96	.15
97	grab a hold	.96	-.16
98	sit seed	.96	.42
99	fall a failure	.96	.15
100	file a report	.96	.10
11	raise objections	.92	.00
16	stretch a regard	.92	.21
19	catch fire	.92	-.08
24	shake a smile	.92	-.02
31	reach a conclusion	.92	.13
44	win one’s memory	.92	.05
53	swing a secret	.92	.40
59	stick one’s mood	.92	.42
76	knock a concern	.92	.23
78	pack an affair	.92	.23
80	clean windows	.92	.21
81	dismiss an idea	.92	.32
93	cease fire	.92	.28
6	say grace	.88	.24
27	drop hints	.88	.16
41	push one’s luck	.88	.04

Cont.

Item no.	Item pair	Item Facility	Corrected Item-total correlation
43	perform a miracle	.88	.29
56	rule an award	.88	.19
71	rush rank	.88	.22
75	challenge a view	.88	.22
79	abandon ship	.88	.22
84	bind blood	.88	.06
88	bend a rule	.88	.50
18	keep pets	.84	.08
69	exercise discretion	.84	.24
8	break news	.80	-.23
10	claim trade	.80	.21
26	fetch an illness	.80	.52
30	meet a need	.80	.26
33	gather a matter	.80	.14
39	sink speed	.80	.22
50	employ a technique	.80	-.03
65	ride a storm	.80	.22
67	score problems	.80	.45
90	lend support	.80	.24
22	run a bath	.76	.16
46	adopt an approach	.76	.46
85	charge respect	.76	.13
86	cast a vote	.76	.34
87	kick one's heels	.76	-.09
13	supply one’s assistance	.72	.19
34	assume responsibility	.72	.25
38	realise a potential	.72	.24
92	hit approval	.72	.32
36	cut a corner	.68	.42
96	stand an occasion	.68	.19
2	lose sleep	.64	.32
7	pick a glance	.64	.32
48	strike a blow	.64	.52
91	sustain an injury	.64	.30
15	serve a sentence	.60	.21
45	impose success	.60	.33
63	assess damage	.60	.52
89	fill an aim	.60	-.02
17	restore a favour	.56	.20
40	fit the bill	.56	.08
4	draw a breath	.52	.61
21	pull a face	.52	-.14
55	express a worry	.52	.38
72	steal someone’s thunder	.48	.41
73	dress a wound	.48	.38
66	jump a queue	.44	-.14
77	lay pressure	.36	.47
37	fly a flag	.28	.33
49	beat eggs	.28	.23
64	afford an opportunity	.20	-.14
MEAN		.82	.17

APPENDIX 5H: Questionnaire responses with regard to COLLMATCH 3 and COLLEX 5 – PILOT VERSIONS

COLLMATCH 3 – PILOT VERSION

Informant	Comment
UL01	awareness of English collocations
UL02	Förståelse av vad ord betyder i olika sammanhang
UL03	Om man gillar poesi eller inte
UL04	One's spontaneous reaction as to whether certain words belong together
UL05	Ords möjlighet att kombineras med varandra; fraskunskap
UL06	Om man hört uttrycket tidigare eller kan tycka att det låter rätt
UL07	Ordkombinationer
UL08	Tyst och omedveten kunskap, ordkombinationer och fraser man "samlat på sig" genom åren
UL09	Förmågan att använda sig av uttryck i olika sammanhang
UL10	NO ANSWER
UL11	Allmän språkkunskap samt talförmåga i Engelska
UL12	Ordförråd, förmåga att koppla ihop ord
UL13	Hur mycket engelska man har läst, och vilken av den man själv skulle applicera i skrift
UL14	Meningen/syftet när ord tillsammans bildar en innebörd de enskilda orden ej kan representera. Syftar också till att placera meningarna i ett konkret sammanhang för att kunna tolka innebörden.
UL15	Möjligtvis en sorts "native" eller "vernacular" engelska. Känns inte som skolengelska utan mer engelska på riktigt i riktiga situationer.
UL16	Ordkunskap och ordspråk
UL17	Visar hur väl man eg. kan språket (verbal kompetens); sorterar dem som vistats länge i landet + infödda fr. dem som ej har förkunskaperna.
UL18	Intuitiv känsla för språket
UL19	Det handlar inte bara om ordförståelse utan också språkkännedom, även om man vet vad orden betyder så behöver inte orden tillsammans betyda något sammanhängande. Därför måste man nog ha flytet i språket inte bara ordkunskap.
UL20	Ordkunskap, kunskap om i vilka sammanhang ett ord används.
UL21	Det är kunskap som man främst tillgodogör sig i ett engelsktalande land. Kunskapen mäter främst idiomatiska uttryck i vardagligt tal.
UL22	Ordkombinationer, men också en fråga om jag stött på (hört) dem någongång.
UL23	I think it measures how attentive the person is when it concerns using words phrases. In real life it i always possible to express yourself in correct English and avoid words, phrases that sound strange.
UL24	Om man vet vilka kombinationer passar ihop så kan man en del engelska. Den här testdelen mäter det allmäna kunskapen i Engelska, inte bara ordförrådet, skulle jag tro.
UL25	NO ANSWER

Informant	Comment
UL01	Idiomatiska uttryck/kollokationer
UL02	NO ANSWER
UL03	Fraser or ordkombinationer; ordkunskap
UL04	As in COLLMATCH, one learns which words belong together, and those which clearly don't
UL05	Fraser ännu en gång
UL06	Fraser och ordval; man väljer det som låter bra eller som man hört förut.
UL07	Ordkombinationer, likt COLLMATCH
UL08	Olika fraser som används i olika situationer
UL09	Standarduttryck på engelska
UL10	Varierande
UL11	Allmän språkkunskap
UL12	Visualisera uttryck, koppla ord till varandra, ordförråd
UL13	Läsförmåga
UL14	Ens kunskaper i engelska när det gäller att uttrycka sig
UL15	"native", "vernacular"
UL16	Synonymkunskap och ordkunskap
UL17	Eftersom den även här testar ordkombinationer, visar den återigen vem som har riktig tal- och skrivvana, ej bara vocabulary.
UL18	Ordförståelse
UL19	Språkförmåga, hur väl man känner språket
UL20	Ordkunskap, kunskap om uttryck
UL21	Ordfraser som är viktiga att kunna framförallt interaktion/samtal
UL22	Ordkunskap (mest BR eng)
UL23	I think it measures the average level of vocabulary acquired. I think it is very useful. It was quite interesting for me to see what I know and what I don't. The second part of your test is more logical/understandible for me than the 1st part[COLLMATCH]
UL24	NO ANSWER
UL25	NO ANSWER

APPENDIX 5I: COLLEX 5

PROVDEL 3 A

INSTRUKTION:

Denna testdel innehåller 50 (1-50) frågor. Varje fråga innehåller tre ordsekvenser markerade med a), b) respektive c). Din uppgift är att välja **en** av de tre ordsekvenserna i varje fråga.

En de tre ordsekvenserna i varje fråga är en naturlig och vanligt förekommande ordkombination i det engelska språket, medan de andra två inte är det. Välj den ordsekvens som du bedömer är den naturligaste och vanligast förekommande genom att sätta ett tydligt kryss under motsvarande bokstav i rutan i högerkolumnen.

Exempel

								a	b	c
51	a.	do a mistake	b.	make a mistake	c.	run a mistake			X	

I exemplet ovan har alternativ b, 'make a mistake' valts som svar på fråga 51.

Varje rätt besvarad fråga ger poäng, och varje felaktigt besvarad fråga ger 0 poäng. Om du inte kryssar i någon av rutorna i en fråga, eller kryssar i två eller fler får du 0 poäng.

							a	b	c
1	a.	do damage	b.	make damage	c.	run damage			
2	a.	turn out a fire	b.	put out a fire	c.	set out a fire			
3	a.	hold discussions	b.	do discussions	c.	set discussions			
4	a.	receive a cold	b.	fetch a cold	c.	catch a cold			
5	a.	do a visit	b.	lay a visit	c.	pay a visit			
6	a.	strike a pose	b.	beat a pose	c.	hit a pose			
7	a.	fell tears	b.	shed tears	c.	raise tears			
8	a.	employ one’s rights	b.	exercise one’s rights	c.	conduct one’s rights			
9	a.	pull an opportunity	b.	seize an opportunity	c.	catch an opportunity			
10	a.	press charges	b.	run charges	c.	push charges			
11	a.	lend a complaint	b.	perform a complaint	c.	lodge a complaint			
12	a.	make a conclusion	b.	pull a conclusion	c.	draw a conclusion			
13	a.	commit a crime	b.	comply a crime	c.	conduct a crime			
14	a.	tell a prayer	b.	say a prayer	c.	speak a prayer			
15	a.	give a speech	b.	hold a speech	c.	perform a speech			
16	a.	strike a deal	b.	set a deal	c.	step a deal			
17	a.	go on a journey	b.	do a journey	c.	pull a journey			
18	a.	keep one’s breath	b.	house one’s breath	c.	hold one’s breath			
19	a.	direct an orchestra	b.	conduct an orchestra	c.	control an orchestra			
20	a.	lose count	b.	drop count	c.	pass count			
21	a.	take root	b.	make root	c.	stick root			
22	a.	hold a secret	b.	keep a secret	c.	last a secret			
23	a.	take one’s revenge	b.	make one’s revenge	c.	obtain one’s revenge			
24	a.	keep a diary	b.	run a diary	c.	lead a diary			
25	a.	brush shoes	b.	polish shoes	c.	sweep shoes			

26	a.	make apologies	b.	do apologies	c.	lay apologies	
27	a.	tie one’s fist	b.	fix one’s fist	c.	clench one’s fist	
28	a.	strike a fuse	b.	knock a fuse	c.	blow a fuse	
29	a.	show heed	b.	pay heed	c.	spread heed	
30	a.	make an escape	b.	take an escape	c.	draw an escape	
31	a.	lose faith	b.	drop faith	c.	cut faith	
32	a.	perform a survey	b.	commit a survey	c.	conduct a survey	
33	a.	push a bike	b.	lead a bike	c.	walk a bike	
34	a.	send judgement	b.	pass judgement	c.	set judgement	
35	a.	say one’s mind	b.	speak one’s mind	c.	talk one’s mind	
36	a.	spoil the fun	b.	break the fun	c.	destroy the fun	
37	a.	earn a purpose	b.	win a purpose	c.	serve a purpose	
38	a.	make friends	b.	create friends	c.	gain friends	
39	a.	make measures	b.	take measures	c.	stick measures	
40	a.	speak shop	b.	say shop	c.	talk shop	
41	a.	defeat a purpose	b.	break a purpose	c.	refuse a purpose	
42	a.	reply to the door	b.	respond to the door	c.	answer the door	
43	a.	lay birth	b.	give birth	c.	bring birth	
44	a.	close a habit	b.	break a habit	c.	lay a habit	
45	a.	earn access	b.	take access	c.	gain access	
46	a.	run the streets	b.	walk the streets	c.	stroll the streets	
47	a.	take harm	b.	do harm	c.	make harm	
48	a.	make progress	b.	take progress	c.	gain progress	
49	a.	let bombs	b.	drop bombs	c.	fell bombs	
50	a.	do sacrifices	b.	give sacrifices	c.	make sacrifices	

APPENDIX 5J: Word frequencies for COLLEX 5

JACET 8000 Frequency band	verbs	nouns
1K	answer, beat, break, bring, catch, close, control, create, cut, do, draw, drop, exercise, give, go, hit, hold, house, keep, last, lay, lead, let, lose, make, pass, pay, press, pull, push, put, raise, receive, refuse, reply, run, say, send, serve set, show, speak, spread, step stick, strike, take, talk, tell, turn, walk, win	damage, fire, discussion, cold visit, tear, right, opportunity speech, deal, count, escape mind, purpose, friend, measure, shop, secret, purpose, door, street, progress,
2K	blow, brush, conduct, defeat, destroy, direct, earn, fix, gain, knock, obtain, perform, respond, tie	charge, conclusion, crime journey, breath, root, diary, shoes, faith, survey, fun, birth, access, harm, bomb
3K	fetch, lend, seize, shed, spoil,	complaint, prayer, orchestra fist, bike, judgement, sacrifice
4K	commit, comply, employ, polish, sweep	pose
5K	lodge, stroll	apologies
6K		revenge, fuse
OFF LIST	clench, fell	heed

APPENDIX 5K: COLLMATCH 3

COLLMATCH 3

INSTRUKTION:

Denna testdel innehåller 100 ordsekvenser. Din uppgift är att avgöra om ordsekvenserna förekommer i det engelska språket eller inte. Om du bedömer att en ordsekvens finns i det engelska språket, sätt ett kryss i rutan 'ja'. Om du bedömer att en ordkombination inte finns i det engelska språket, sätt ett kryss i rutan 'nej'.

Kontrollera att du avgivit svar för samtliga ordkombinationer.

Exempel

101	catch importance	102	take precautions	103	shed attention
	ja	X	ja		ja
X	nej		nej	X	nej

I exemplet ovan har sekvens 102, 'take precautions' valts som förekommande i det engelska språket medan sekvenserna 101 samt 103 valts som icke förekommande.

Varje rätt besvarad fråga ger poäng, och varje felaktigt besvarad fråga ger 0 poäng. Om du inte kryssar i någon av rutorna i en fråga, eller kryssar i bägge, får du 0 poäng.

COLLMATCH 3

INSTRUCTION:

This part consists of 100 word combinations (1-100). Your task is to decide whether the word combinations are used in the English language or not. If you think a word combination is used in the English language, tick the 'yes' box. If you don't think a word combination is used in the English language, tick the 'no' box.

Please make sure that you have answered all test items.

Example

101	catch importance	102	take precautions	103	shed attention
	yes	X	yes		yes
X	no		no	X	no

In the example above, word combination 102, 'take precautions' has been chosen as an existing word combination in English whereas word combinations 101 and 103 have been chosen as not existing.

PART A

1	have a say	2	lose sleep	3	do justice	4	draw a breath	5	turn a reason
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
6	say grace	7	pick a glance	8	break news	9	make a move	10	claim trade
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
1 1	raise objection	12	bear witness	13	supply one's assistance	14	give a speech	15	serve a sentence
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
16	stretch a regar	17	restore a favou	18	keep pets	19	catch fire	20	hold meetings
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
21	pull a face	22	run a bath	23	throw a party	24	shake a smile	25	set an example
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej

PART B

26	fetch an illness	27	drop hints	28	play a trick	29	pay attention	30	meet a need
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
31	reach a conclusion	32	drag a limit	33	gather a matter	34	assume responsibility	35	suffer damage
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
36	cut a corner	37	fly a flag	38	realise a potential	39	sink speed	40	fit the bill
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
41	push one's luc	42	gain ground	43	perform a miracle	44	win one's memory	45	impose success
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
46	adopt an approach	47	clear one's throat	48	strike a blow	49	beat eggs	50	employ a technique
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej

51	press charges	52	settle a dispute	53	swing a secret	54	grant permission	55	express a worry
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
56	rule an award	57	commit a sin	58	launch a campaign	59	stick one's mood	60	acquire a skill
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
61	deliver a speech	62	spread one's wings	63	assess damage	64	afford an opportunity	65	ride a storm
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
66	jump a queue	67	score problems	68	roll a look	69	exercise discretion	70	blow one's nose
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
71	rush rank	72	steal someone's thunder	73	dress a wound	74	pursue a career	75	challenge a view
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej

PART D

76	knock a concern	77	lay pressure	78	pack an affair	79	abandon ship	80	clean windows
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
81	dismiss an idea	82	shift gear	83	justify one's existence	84	bind blood	85	charge respect
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
86	cast a vote	87	kick one's heels	88	bend a rule	89	fill an aim	90	lend support
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
91	sustain an injury	92	hit approval	93	cease fire	94	snap one's fingers	95	shrug one's shoulders
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej
96	stand an occasion	97	grab a hold	98	sit seed	99	fall a failure	100	file a report
☐	ja	☐	ja	☐	ja	☐	ja	☐	ja
☐	nej	☐	nej	☐	nej	☐	nej	☐	nej

APPENDIX 5L: Word frequencies for COLLMATCH 3

JACET 8000 Frequency band	verbs	nouns
1K	bear, beat, break, catch, challenge clean, clear, cut, do, draw, dress, drop, exercise, express, fall, fill, fly, give, have, hit, hold, jump, keep, lay, lose, make, meet, pay, pick, play, press, pull, push, raise, reach, realize, ride, rule, run, say, serve, set, shake, sit, spread, stand stick, strike, suffer, supply, throw, turn, win	sleep, reason, news, move, trade, speech, sentence, fire, meeting, face, party, smile, example, attention, need, limit, matter, damage, corner, speed, ground, memory, success, approach, secret, worry, skill, opportunity, problem, look, career, view, concern, pressure, ship, window, idea, blood respect, vote, rule, support, fire, finger, shoulder, hold, report
2K	claim, stretch, drag, gather, assume, sink, fit, gain, perform, adopt, settle, swing, grant, acquire deliver, afford, score, roll, blow, rush, knock, pack, shift, charge, cast, kick, bend, lend, grab	justice, breath, glance, witness, regard, favour, bath, illness, conclusion, responsibility, potentia bill, luck, miracle, throat, blow, egg technique, charge, award, campaig mood, wing, storm, nose, wound, affair, existence, aim, injury, occasion, seed, failure
3K	restore, fetch, steal, pursue, snap, shrug, file	pet, hint, flag, permission, discretion, rank, gear, heel
4K	impose, employ, commit, launch, assess, abandon, dismiss, justify, bind, sustain, cease	grace, objection, assistance, dispute sin, approval
5K		queue, thunder
OFF LIST		say

APPENDIX 5M: Item Facility (IF) and Item-Total Correlation (ITC) values of COLLEX 5 test administration

Item no	Item triple			Item Facility	Corrected Item-total correlation
22	hold a secret	keep a secret	last a secret	1.00	.00
13	commit a crime	comply a crime	conduct a crime	.99	.08
31	lose faith	drop faith	cut faith	.99	.23
43	lay birth	give birth	bring birth	.99	.25
4	receive a cold	fetch a cold	catch a cold	.96	.36
38	make friends	create friends	gain friends	.96	.24
44	close a habit	break a habit	lay a habit	.96	.19
49	let bombs	drop bombs	fell bombs	.96	.17
5	do a visit	lay a visit	pay a visit	.95	.37
48	make progress	take progress	gain progress	.95	.20
3	hold discussions	do discussions	set discussions	.94	.24
7	fell tears	shed tears	raise tears	.93	.40
24	keep a diary	run a diary	lead a diary	.93	.29
30	make an escape	take an escape	draw an escape	.93	.17
42	reply to the door	respond to the door	answer the door	.93	.35
50	do sacrifices	give sacrifices	make sacrifices	.93	.32
14	tell a prayer	say a prayer	speak a prayer	.92	.42
17	go on a journey	do a journey	pull a journey	.92	.25
18	keep one's breath	house one's breath	hold one's breath	.92	.08
25	brush shoes	polish shoes	sweep shoes	.92	.08
35	say one's mind	speak one's mind	talk one's mind	.92	.34
36	spoil the fun	ruin the fun	destroy the fun	.92	.42
45	earn access	take access	gain access	.92	.31
37	earn a purpose	win a purpose	serve a purpose	.91	.45
6	strike a pose	beat a pose	hit a pose	.90	.40
10	press charges	run charges	push charges	.89	.40
20	lose count	drop count	pass count	.89	.30
2	turn out a fire	put out a fire	set out a fire	.88	.55
12	make a conclusion	pull a conclusion	draw a conclusion	.88	.35
39	make measures	take measures	stick measures	.87	.49
47	take harm	do harm	make harm	.86	.26
26	make apologies	do apologies	lay apologies	.84	.46
15	give a speech	hold a speech	perform a speech	.78	.15
46	run the streets	walk the streets	stroll the streets	.78	.28
40	speak shop	say shop	talk shop	.77	.18
23	take one's revenge	make one's revenge	obtain one's revenge	.76	.34
9	pull an opportunity	seize an opportunity	catch an opportunity	.75	.56
19	direct an orchestra	conduct an orchestra	control an orchestra	.75	.40
28	strike a fuse	knock a fuse	blow a fuse	.71	.43
27	tie one's fist	fix one's fist	clench one's fist	.70	.51
34	send judgement	pass judgement	set judgement	.69	.62
1	do damage	make damage	run damage	.67	.52
8	employ one's rights	exercise one's rights	conduct one's rights	.65	.56
29	show heed	pay heed	spread heed	.65	.41
32	perform a survey	commit a survey	conduct a survey	.64	.45
11	lend a complaint	perform a complaint	lodge a complaint	.62	.51
16	strike a deal	set a deal	step a deal	.56	.54
21	take root	make root	stick root	.50	.37
41	defeat a purpose	break a purpose	refuse a purpose	.48	.44
33	push a bike	lead a bike	walk a bike	.33	.32
			Mean	.83	.34

APPENDIX 5N: Item Facility (IF) and Item-Total Correlation (ITC) values of COLLMATCH 3 test administration

Item no	Item pair	Item Facility	Corrected Item-total correlation
29	pay attention	1.00	.00
9	make a move	.99	.08
94	snap one's fingers	.99	.08
47	clear one's throat	.98	.21
62	spread one's wings	.98	.18
53	swing a secret	.97	.14
100	file a report	.97	.22
24	shake a smile	.96	.12
25	set an example	.96	.19
51	press charges	.96	.17
70	blow one's nose	.96	.22
98	sit seed	.95	.14
58	launch a campaign	.95	.21
12	bear witness	.95	.23
99	fall a failure	.94	.15
95	shrug one's shoulders	.93	.13
44	win one's memory	.93	.23
16	stretch a regard	.93	.22
5	turn a reason	.92	.25
23	throw a party	.92	.34
56	rule an award	.92	.11
78	pack an affair	.92	.22
57	commit a sin	.92	.16
33	gather a matter	.91	.27
59	stick one's mood	.91	.22
76	knock a concern	.91	.17
20	hold meetings	.90	.20
39	sink speed	.90	.27
54	grant permission	.90	.40
80	clean windows	.90	.13
14	give a speech	.89	.09
19	catch fire	.88	.29
31	reach a conclusion	.88	.20
26	fetch an illness	.88	.37
82	shift gear	.87	.27
52	settle a dispute	.87	.44
67	score problems	.87	.28
35	suffer damage	.86	.06
42	gain ground	.86	.23
71	rush rank	.86	.11
11	raise objections	.85	.23
32	drag a limit	.85	.39
79	abandon ship	.85	.34
85	charge respect	.85	.28
74	pursue a career	.85	.40
97	grab a hold	.85	.19
41	push one's luck	.85	.42
86	cast a vote	.85	.28
28	play a trick	.83	.22
43	perform a miracle	.83	.34

Cont.

Item no	Item pair	Item Facility	Corrected Item-total correlation
68	roll a look	.83	.23
60	acquire a skill	.83	.35
81	dismiss an idea	.82	.32
6	say grace	.81	.38
8	break news	.81	-.06
36	cut a corner	.81	.43
84	bind blood	.81	.18
83	justify one's existence	.81	.26
88	bend a rule	.79	.39
10	claim trade	.79	.24
3	do justice	.79	.02
92	hit approval	.79	.24
45	impose success	.78	.17
18	keep pets	.78	.30
65	ride a storm	.76	.06
7	pick a glance	.75	.40
93	cease fire	.75	.44
1	have a say	.75	.49
27	drop hints	.75	.32
48	strike a blow	.75	.29
87	kick one's heels	.72	.15
91	sustain an injury	.71	.29
89	fill an aim	.69	.23
96	stand an occasion	.67	.22
15	serve a sentence	.66	.27
69	exercise discretion	.65	.47
13	supply one's assistance	.64	.25
34	assume responsibility	.64	.37
38	realise a potential	.64	.14
90	lend support	.64	.37
2	lose sleep	.64	.44
21	pull a face	.63	.12
22	run a bath	.62	.35
61	deliver a speech	.61	.48
73	dress a wound	.61	.40
49	beat eggs	.58	.46
75	challenge a view	.58	.38
17	restore a favour	.57	.30
72	steal someone's thunder	.57	.31
66	jump a queue	.53	.32
77	lay pressure	.53	.25
50	employ a technique	.51	.45
63	assess damage	.51	.46
40	fit the bill	.50	.29
46	adopt an approach	.49	.43
30	meet a need	.46	.46
55	express a worry	.37	-.06
4	draw a breath	.35	.45
64	afford an opportunity	.25	-.12
37	fly a flag	.25	.34
MEAN		.78	.26