Case Endings in Spoken Standard Arabic

Hallberg, Andreas

2016

Case Endings in Spoken Standard Arabic

Statistics, norms, and diversity in unscripted formal speech

Andreas Hallberg

Lund University
2016

LUND UNIVERSITY
Joint Faculties of Humanities and Theology

Case Endings in Spoken Standard Arabic.

Studia Orientalia Lundensia. Nova Series can be ordered via Lund University: *http://www.ht.lu.se/en/serie/sol/*. E-mail: *skriftserier@ht.lu.se*.

ISBN 978-91-87833-69-4 (print)
ISBN 978-91-87833-70-0 (pdf)
ISSN 0281-4528

Printed in Sweden by Media-Tryck, Lund University
Lund 2016

Abstract

Morphologically marked case is a salient Standard Arabic feature without parallel in Arabic dialects. As such it is a grammatical system learned by native speakers of Arabic through formal education. Case endings are traditionally regarded as an essential feature of Standard Arabic, but morphological case endings are used only sporadically in extemporaneous speech in formal situations where Standard Arabic is the expected variety. This study investigates how case endings that are used in speech are distributed in relation to morphosyntactic parameters with the aim of finding covert linguistic norms governing where case is and is not marked in speech. This is done by a quantitative analysis of a corpus consisting of 17 televised interviews of highly educated native speakers of Arabic. Only speech by the interviewees was analyzed, totaling 35000 words or 5 h and 22 min. Nouns and adjectives in the corpus are annotated for morphosyntactic features, including if and how the case ending is produced. The data show that the rate of case marking differs widely between speakers, but also that there are patterns, consistent between speakers, of how case endings are proportionally distributed in various morphosyntactic contexts. It was found that case endings are very rarely used in words with the definite article *al-*, in adjectival attributes, and on words at the end of utterances. Case marking is strongly favored on words where it would be orthographically represented in writing and on words with an enclitic pronoun. It was also found that these patterns are not the result of speakers relying on a set of fixed phrases to include case endings in their speech. The findings presented in this study have important implications for Arabic curriculum development, both in first and second language teaching, and also shed light on the role of the use of case endings in Arabic diglossia.

Contents

List of Tables

List of Figures

List of Abbreviations

1	first person
2	second person
3	third person
ACC	accusative
ADJ	adjective
AMB	ambiguous case marker
CA	Classical Arabic
COMP	complementizer
COP	copula
CS-N/C	construct state with annexed noun or clause
DEF	definite article
DUA	dual
ESA	Educated Spoken Arabic
f.	feminine
FUT	future tense
GEN	genitive
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed-Effects Model
HYP	hypercorrect case marker
IWA	Informal Written Arabic
m.	masculine
MSA	Modern Standard Arabic
NEG	negation
NOM	nominative
OLA	Oral Literary Arabic
PART	particle
PASS	passive
pl.	plural
PTW	per 1000 words
Q	interrogative particle
s.	singular
SA	Standard Arabic
SSA	Spoken Standard Arabic
VOC	vocative particle

Transcription

The transcription system in EALL (Reichmuth 2006:517) is followed here with the exception that *ẓ* replaces *ḍ̇* in all words and not only in proper names. Listed below is the Standard Arabic phonetic inventory as represented in Arabic script, EALL transcription, and the International Phonetic Alphabet.

ARA.	EALL	IPA	ARA.	EALL	IPA
أ	ʾ	ʔ	غ	*ġ*	ɣ
ب	*b*	b	ف	*f*	f
ت	*t*	t	ق	*q*	q
ث	*ṯ*	θ	ك	*k*	k
ج	*j*	dʒ	ل	*l*	l
ح	*ḥ*	ħ	م	*m*	m
خ	*x*	x	ن	*n*	n
د	*d*	d	ه	*h*	h
ذ	*ḏ*	ð	و	*w*	w
ر	*r*	r	ي	*y*	j
ز	*z*	z			
س	*s*	s			
ش	*š*	ʃ	ـَ	*a*	a/ɑ
ص	*ṣ*	sˤ	ا	*ā*	æː/ɑː
ض	*ḍ*	dˤ	ـُ	*u*	u
ط	*ṭ*	tˤ	و	*ū*	uː
ظ	*ẓ*	ðˤ	ـِ	*i*	i
ع	ʿ	ʕ	ي	*ī*	iː

Acknowledgments

I wish first and foremost to thank my supervisor Professor Lena Ambjörn for her advice and careful reading of everything from the first drafts to the finished manuscript. A task that was always carried out while providing good advice and firmly steering me away from many an ill-conceived idea. I also wish to thank Professor Marit Julien, my second supervisor, for helping me to straighten out thorny linguistic issues and for always guiding me towards better theoretical approaches and terminological choices. Professor Gunvor Mejdell gave me many valuable comments as the opponent at the 'final seminar' and generously gave me her notes with detailed remarks. I am also grateful to Dr. Maria Persson who has followed my progress in this project from the very beginning and attended the many seminars it generated and, in addition, she was always forthcoming with constructive comments. I want to give special thanks to Dr. Joost van De Veijer at the Humanities Lab for his excellent courses in statistics which were instrumental in shaping this dissertation, and for his generously spending time discussing and giving me advice on statistical matters and for helping me with solutions to tricky coding problems. Many thanks also to my fellow PhD-students at the neighboring Department of Communication and Media for making the solitary work of dissertation-writing much less lonely. Finally, and most importantly, I thank my wife Amani for being a miracle of support and patience and for putting up with a constant stream of requests for native-speaker intuitions.

Introduction 1

When I began my Arabic studies at Lund university, the morphological case endings, a set of grammatical endings added to nouns and adjectives, were presented as an essential part of the grammar. Exercises in using them correctly accordingly took up a large part of class time. Sure, they are for the most part not written, since vowels are not used in most Arabic texts, and speakers of Standard Arabic do not usually pronounce the case endings. However, any written text may have the vowel signs added to it, and the correct way of speaking is nevertheless to pronounce the endings, even if only very few speakers are able to do this consistently.

Any student of Arabic sooner or later has to come to grips with *diglossia*, the fact that formal written Arabic is very different from the kind of Arabic that speakers use in their every-day lives. Only in formal Standard Arabic, or *fuṣḥā*, as it is known in Arabic, is there a system of case endings. The dialects have no parallel system. My second year of Arabic studies was spent at the Jordanian University in Amman where I made friends who were happy to speak Standard Arabic which not all were native speakers are. I learned a great deal by imitating their way of speaking in which case endings turn up only here and there in certain expressions. When I asked about the case endings I got vague and contradictory answers: "We don't use them, but we should", "You need them for the sentences to make sense", "Its very difficult", "Only people who know proper Arabic use them correctly." Looking back now, having done the research for this thesis, such statements make sense within the context of Arabic language ideology, but at the time the situation was extremely confusing. I never got a clear answer or the understanding I was looking for as to how I was supposed to use the case endings in speech. Over time, I gradually came to accept and learn to navigate through this confusion while backgrounding it somewhat. As my skills in non-standard, dialectal Arabic improved, I used this

form of Arabic more and more. Since in the dialects there are no case endings, the whole thing naturally became less of an issue. Arabic speakers seemed not to be too bothered with it, so why should I be?

Having come back to Sweden and having finished my MA, I traveled to Damascus to continue my studies. I felt I still needed to improve my Arabic. During my first semester there I took classes in Arabic for foreign students, and in the second semester I attended classes in the Arabic literature program with Syrian students. This second semester was a very interesting experience, not only for what I learned in class, but also for what I learned about Syrian education. Arabic syntax constituted a major chunk of the program. Half of the lectures in this subject dealt with syntactic concepts and operations, and the other half consisted of grammatical analysis of classical poetry. I found three things particularly remarkable about these classes. The first thing was that the mode of grammatical analysis was completely dependent on the system developed by Medieval Arab grammarians. The second thing was the nature of the textual material being studied. Not a single modern text was analyzed or discussed. Most of the students would find jobs as Arabic teachers, teaching the young to use and understand their language effectively. Yet they were given no experience in working with modern texts. The third thing that struck me as remarkable was the ritualistic way in which grammatical analysis was performed. Normally the teacher would point to a line of verse and ask someone to analyze it, to 'do *iʿrāb*' on the sentence. The student would go through the verse word by word, stating its word class, syntactic position, and if and how this position is expressed in a grammatical ending, all with memorized, formulaic phrases. Although I never took the exam myself, it was explained to me that it consisted exclusively of poetry to be analyzed in this manner. I could confirm this by looking at exams from previous years. These exams were considered to be the most difficult in the program as it is in all other programs that include this course, and it was this exam that the highest number of students failed each year.

These discoveries made me revisit my questions about the case system in Arabic. One of the pieces of the puzzle fell into place: the reason Arabic speakers find the case system in their standard language so difficult is the way it is taught, detached from speech and detached from modern uses of the language. I consider myself a fairly proficient speaker of Arabic, and the use of case endings is one of the reasons why I still shun Standard Arabic in favor of dialectal Arabic, even when I am addressed in Standard Arabic or in other situations where the standard form of the language would be appropriate. I often find myself stopping in the middle of a sentence thinking "Should I put a case end-

ing on this word?" hesitating and stuttering as a result and loosing my train of thought. This lack of confidence in using the standard language is something I have in common with most native speakers, and the use of morphological case is a large part of the problem.

Coming back to Sweden after this second stay abroad, I started to do my first proper teaching at Lund University, an experience as terrifying as it was rewarding. I taught Arabic the way I myself had been taught: a couple of weeks of the alphabet, the writing system and pronunciation, then some simple nominal sentences, and soon thereafter sentences with verbs. The case system is introduced with the first simple sentences. Case is, of course, present even in the simplest of grammatical structures and students, so the reasoning goes, need to be exposed to complete, correct sentences early on. The Arabic case system is naturally difficult for native speakers of Swedish, a language which only has remnants of a case system, so it was to be expected that the it would take up a lot of class time. However, I found myself more and more often devoting precious class time to explaining and having students practice the use of grammatical endings that are not present in printed text and that they will not use themselves when writing, and that in speech are used only sparingly, and only in the most formal styles.

Looking for a model of case usage to teach by, I turned to the scholarly literature, which gave few answers. There are plenty of impressionistic remarks stating that case endings are only used sparingly in speech. This is true enough, but it is not very useful for finding ways of teaching it. What was needed was some sort of model of how case endings are used by speakers proficient in Standard Arabic, speakers capable of making on the spot decisions to pronounce case endings in some words and not in others. The only such model that exists to date is the traditional model prescribing endings on all nouns and adjectives, except before a pause. The problem with this model is of course that no one follows it. Having students, native or non-native, follow this model is counterproductive, and it instills habits that will have to be unlearned in order to speak 'normal' Standard Arabic, the way proficient native speakers of Arabic do.

It is this lack of a usable description of how case endings are used in Spoken Standard Arabic, a description based on observation of actual speech, rather than on the received rules of grammar books, that I hope to help remedy with this study.

1.1 Aims

The primary aim of this dissertation is to provide a thorough description of how case endings are used in extemporaneously spoken Standard Arabic, and, secondly, to identify a set of norms that regulate the use of these endings. Put in more practical terms, the primary aim is to present quantitative data on how case endings are used in formal spoken Arabic, and the secondary aim is to explain these data.

The explicit, prescriptive norm embodied in traditional grammars is to apply case endings all nouns and adjectives except before a pause (*waqf*, roughly a sentence boundary). It is, however, quite clear when listening to educated persons speaking in their most formal register that this is not the way Standard Arabic is spoken. Case marking in Spoken Standard Arabic (see below for definition) is sporadic and seemingly inconsistent. The overarching hypothesis in the present study is that case marking in Spoken Standard Arabic is in fact governed by a set of covert linguistic norms. Instead of seeing lack of or inconsistent case marking as mistakes or shortcomings of the speaker, this study sees the patterns of case marking in speech as part of system of preferences of where — and where not — to mark morphological case.

The notion of investigating spoken Standard Arabic as the product of a linguistic system different from that of traditional grammar is, of course, not new. Parkinson (1993:72), for example, writes about "overt and covert norms" in Standard Arabic, Maamouri (1998:61) discusses how "the norms and behaviors of the language community" differ from traditional grammar, and in the entry in *Encyclopedia of Arabic Language and Linguistics* on first language teaching, Wahba (2006:104) contrasts the "ideological standard" of grammar books with the "organic standard" observed in actual language use. This has, however, rarely been systematically applied to the formal register of Arabic, a form of speech often implicitly assumed to conform with, or to aim to conform with, prescriptive grammar. When it does not, and it rarely does, it is classified not as Standard Arabic but as a mixed or middle register. Indeed, much recent research has set to out to capture semi-formal speech, leaving the more formal end of the spectrum aside. The statement in Harrell's classical study (Harrell 1964:3) that "Spoken Classical [i.e. Standard] Arabic" is a variety "to which little specific attention has been devoted" is still very much true today.

1.2 Motivation

The primary motivation for carrying out this study, as illustrated in the prelude to this chapter, is to inform curriculum development, both in the Western and

Arabic educational systems. Several scholars have called for systematic studies of Standard Arabic in use, as opposed to Standard Arabic as gleaned from grammars, to inform the formulations of proficiency goals for Standard Arabic (Parkinson 1994b:202; Badawi 2006:xi; Ryding 2006:18; Wahba 2006:104; Nielsen 2009:150–1), but studies of speech in this variety is still sorely missing.

In a Western setting, the strong focus on case from the early stages of language instruction has been challenged in recent years. Several newly published educational materials postpone the introduction of the case system until the basics of syntax and morphology have been covered. While this is major step forward, it leaves open the question as to how the case system is to be taught at intermediate to advanced levels. As far as I am aware, there is to date no accurate description for students to follow as to how native speakers of Arabic use case endings in speech, or for teachers to base instructions on. Instructing speakers to use case endings at their leisure, seeing that this is what native speakers appear to do, is of course problematic. For example, there are situations where native speakers avoid case marking altogether, as will be shown in this study, most notably on nominals with the definite article. If students develop ways of speaking where case is marked on such nominals or that in other ways deviate from norms of case marking in speech, the result is distinctly nonnativelike speech.

In an Arabic educational setting, the underlying problem is the same: students are taught linguistic and grammatical features that do not correspond to real-life uses of the language. The problem in an Arabic setting is further complicated by a highly conservative language ideology, an issue that needs to be addressed before educational reforms can actually be implemented. These issue of ideology are discussed as background information in this study but is not its focus. Even if at present the prospects of an officially sanctioned language reform seem bleak, it is hoped that when and if a serious discussion on the issue takes place, results presented in this study will contribute to and inform that discussion.

That being said, the problem of Arabic language instruction in the Arab world is indeed acute. Tens of millions of children study Arabic as their first language in school. These children struggle with the system of case endings, much of which is archaic, taught in a cumbersome system of grammatical theory, and not in active use in the modern language. Teaching as 'correct' and 'proper' Standard Arabic linguistic features that do not correspond to what pupils hear and read generates confusion and linguistic insecurity, in what one scholar has called the "alienation of Arabs from Arabic grammar" (Uhlmann 2012:105). This alienation contributes to the exclusion of large parts of the pop-

ulation from participation in the public sphere in which Standard Arabic is often the required variety.

It should be noted that a reevaluation of the role of case endings in a Standard Arabic language reform does not imply promoting the dialect and abolishing Standard Arabic. The two are distinctly different varieties. Reducing or even omitting case endings from Standard Arabic does not entail transforming it into dialectal Arabic, as some Arab conservatives claim. One does not need to examine the dialects to find ample empirical arguments for such a reevaluation. Within Standard Arabic, case endings are syntactically redundant, are not marked in writing, and are systematically pronounced only in very specific genres and circumstances. Reevaluating the role of case endings is, in other words a matter of corpus planning rather than a matter of status planning. It is a reform within Standard Arabic that does not alter its status with regards to the dialects. Furthermore, it is a reform that in practice has largely taken place already outside of Arabic formal education.

A third area, after language instruction and language reform, where the results presented here will be of use, is linguistic theory, more specifically for the theoretical understanding of diglossia and diglossic variation. Case endings are a salient feature of the 'high' variety in Arabic diglossia and it differs from other diglossic features in not having a parallel in the dialects. Case endings cannot be realized through processes of substitutions of low variety elements. The processes of case marking are therefore expected to be different from processes for producing other high variety features. As will be shown in Chapter 8, case endings are in fact distributed amongst speakers differently from other diglossic features, and they appear to be weighted as markers of 'standardness' in different ways by different speakers.

1.3 Spoken Standard Arabic

Defining linguistic varieties is inherently difficult. Definitions of Modern Standard Arabic typically make reference to (*a*) its emergence with the *nahḍa* movement of the 19th century as a development of Classical Arabic; (*b*) its codification; (*c*) its role as the written language; (*d*) its use as the language of news media, both written and spoken, and (*e*) it being acquired by speakers through formal education (e.g. Ryding 2005:7; McCarus 2006:238; Holes 2013:5). There are several ways to approach analytically the spoken form of Standard Arabic. For the purposes of this study, Spoken Standard Arabic is defined as *the most formal register of Arabic extemporaneous speech by proficient, highly educated native speakers of Arabic.* Aspects of this definition are discussed in detail Chap-

ter 5. The linguistic variety under investigation is thus defined by language external factors; it is the language as used by certain speakers in certain situations. This is done in order to avoid presuppositions of what the features of this variety are or ought to be. The underlying assumption is that the way competent speakers use the language in formal situations represents the oral form of the standard language.

As mentioned above, the birth of Modern Standard Arabic as a development of Classical Arabic is usually said to have occurred with the *nahḍa* movement in Egypt and the Levant in the 19th century. This movement was initiated by contact with modern European culture brought about by Napoleon's occupation of Egypt in 1798. The period saw conscious efforts to adapt Arabic to modern society, particularly with the development of journalism (Badawī 1993: 5; Newman 2013:475). Arabic dialects had, of course, continued to develop during this period but had no legitimacy and were not seen as having the potential to be adapted to a language of high culture and learning.

The linguistic difference between Classical Arabic and Modern Standard Arabic that resulted from the societal changes of the 19th century is usually described as being primarily on the lexical and phraseological levels, while the basic syntax remained unchanged. In this study, the term 'Standard Arabic' is used to refer to the standard language as a whole, including both its classical and modern varieties. Using this wider term does away with the distinction between the modern and the classical language and is thus closer to the Arabic concept of *fuṣḥā*, in which no such distinction is made. This has the benefit in the discussions on prescriptive Arabic grammar in Chapters 3 and 4, that, although based on Classical Arabic, in the Arabic domestic context it is considered to have validity over the modern language as well. The wider term of Standard Arabic is thus better suited in discussing notions of correctness in the Arabic language and in the use of case endings. However, since this is a synchronic linguistic study of the modern language, this term will in most cases be equivalent to Modern Standard Arabic in the linguistically descriptive parts of this dissertation. However, the distinction between modern and classical Standard Arabic will be made where necessary. The non-standard varieties of Arabic will be referred to as 'dialects'.

The traditional grammatical description of Standard Arabic, though fundamentally based on Classical Arabic, has important implications for understanding the use of grammatical features in the modern context. The language as described in these grammars is held up as the ideal, and it is the goal of language instruction in the Arab countries and also in Muslim communities around the world where Arabic is taught. The Arabic language, or more accu-

rately, this idealized form of the Arabic language, is held in high regard in Arab culture and the ability to live up to this linguistic ideal is imbued with a moral dimension, as is often the case with standard languages. This is something to which speakers of Standard Arabic must somehow relate.

The choice of the term 'Spoken Standard Arabic' to describe the variety under investigation, and that is represented by the corpus constructed for this study, may be controversial. Standard Arabic is closely associated with writing and with traditional grammatical descriptions. Virtually all formal extemporaneous speech clearly falls short of this ideal, and many of the examples from the corpus given in various parts of this study will certainly strike the reader as very much non-standard. For these reasons, many scholars have preferred to refer to formal speech with terms capturing this non-standardness; 'formal spoken Arabic', 'sub-standard Arabic', 'Educated Spoken Arabic', and the like. Others describe it in terms of mixing or codeswitching. (See Chapter 2 for an overview.) The reluctance to refer to even the most formal observed forms of extempore speech as Standard Arabic on the grounds that it shows variation and influences of the speaker's dialect is, in the view of this author, an implicit acceptance of prescriptivism. Furthermore, this route in effect excludes Standard Arabic from the realm of extempore speech since all speech deviates from traditional descriptions of Standard Arabic. The route taken in the present study is instead to regard the most formal form of speech as the spoken form of the standard language. This reasoning is further developed in Chapter 3.

1.4 Linguistic examples

There are numerous linguistic examples in this study, most of them from the corpus. Examples are transcribed with the EALL transcription system (see *Transcription* on page *xii*) as pronounced by the speaker. Transcriptions of words in the examples thus often deviate form their normal written forms. A word-initial glottal stop /ʔ/ is for example in the examples transcribed according to the actual pronunciation in that particular example. The definite article is thus transcribed as *ʾal-* in examples if pronounced with initial glottal stop by the speaker. (For words given as examples in the running text, word- and stem-initial glottal stops are omitted.) Similarly, vowel length is transcribed as pronounced, that is, typically with shortened word-final vowels. Pauses and hesitations are marked with ellipsis. Vowels in word boundaries that could be interpreted as case endings, whether prescriptively correct or not, are transcribed as part of the preceding word and glossed according to their status as case endings. This status is often ambiguous, in which case they are glossed as AMB. Ambiguity is

here a technical term, described in 7.7.1, and the glossing of some endings as ambiguous may strike the reader as counterintuitive. For other abbreviations used in the glosses, see *List of Abbreviations* on page *xi*. Boldface in the examples highlights words that exemplify the phenomenon under discussion. Saliently dialectal words are underlined where relevant for the discussion.

Examples from the corpus are followed with the speaker's name and a time reference in minutes and seconds to the recording of the program. These programs are available on Al Jazeera's channel on the video hosting site YouTube (*www.youtube.com*) and, in the case of audio recordings, on Al Jazeera's homepage (*www.aljazeera.net*). The names and time references next to examples are in the electronic version of this document[1] clickable hyperlinks to the appropriate point in the videos. By clicking on the name or time reference next to the example the reader can play the recordings at the point where the example is uttered. The link leads to the beginning of the sentence from which the example is taken, and there may therefore be a few seconds of speech before the actual material in example is uttered. Four of the seventeen interviews in the corpus are only available as audio recordings and are hosted on Al Jazeera's homepage. Links to these interviews lead to the site with the recording but not to the specific point in the recording where the utterance is made. The reader must in these recordings manually navigate to the relevant point as given in the time reference to listen to the utterance. Time references to these interviews are marked with an asterisk. All URLs to recordings and transcripts as published by Al Jazeera are listed in Appendix A.

1.5 Arrangement

The remainder of this dissertation is divided into thee parts: Part I, *Background*, Part II, *Method*, and Part III, *Analysis*, each consisting of three chapters. Thereafter the results are summarized and discussed in a concluding chapter. Some care was taken to make it possible to read each chapter on its own, independently of other chapters. This means that there are numerous summaries throughout the dissertation of previously presented information. When this is done there will generally be a cross reference to the section where the topic is more thoroughly dealt with. Each chapter (apart from this one) ends with a summary of the main points or findings.

Part I follows this introductory chapter and provides background information on the Arabic language and on ideas about and uses of the Arabic system

[1] The electronic version of this dissertation is available as open access at *https://lup.lub.lu.se/search/publication/8524489*.

of case endings. As such, Part I is aimed at a broader readership than are Parts II and III and it is written partly with the reader without a background in Arabic linguistics in mind, although it also addresses field-internal issues. The first chapter in this part, Chapter 2, discusses Arabic in terms of diglossia; a language where the standard variety is no one's actual first language and differs widely from the language of everyday conversation. This chapter is primarily a presentation and critical review of the various models that have been presented to come to grips with the mechanisms of variation that diglossia entails. The chapter traces the theoretical development from structuralist conceptualizations of diglossia to more complex models that integrate dimensions other than the duality of 'high' and 'low' varieties. Chapter 3 discusses Arabic as an instance of a standard language, comparable to other modern languages in the language ideology of its speakers and in notions of linguistic correctness. These are issues with particular bearing on case endings, since the correct marking of case is often seen as the most important sign of proper, authentic, and correct Arabic. The last chapter in this part, Chapter 4, focuses on the particular linguistic feature of grammatical case. After a brief description of the Arabic case system from a formal linguistic perspective, the chapter focuses on case as set of practices; how it is described in traditional sources, how it is taught, and how it is used in writing and in various forms of speech. Many of the practices described in this chapter will strike both native and non-native speakers of Arabic as commonsensical, or even trivial. They are for this reason rarely explicitly described. It is, however, important to make them explicit for an understanding of what it means for a speaker to use or to not use case endings in their speech, and how this relates to other uses of the language. Chapter 4 also includes a review of the little research there is of the use of case endings in extemporaneously spoken formal Arabic.

Part II describes the methods used in gathering, formatting, and annotating the corpus that provides the primary material for this study. Chapter 5 describes the principles and procedure that lead to choosing the seventeen episodes of a televised political interview program that make up the corpus. They where chosen to represent the most formal register of highly educated native speakers of Arabic with experience in appearing in public. An important reason for choosing this particular television show for analysis was that episodes of it are available as recordings as well as in transcripts. The practicalities of adapting and editing these transcripts for linguistic analysis is described in Chapter 6. The chapter deals with the choice of transliteration system, the adaptation of the texts to formatting standards, and the marking up of material within the interviews that was excluded from analysis, such as quotations and

formulaic expressions. The bulk of Part II is taken up by Chapter 7. This chapter describes the coding scheme used to annotate the transliterated and formatted texts. The primary aim of the annotation is to encode in the text where and how the speakers produce case endings in order to make this information computationally retrievable. Each of the tags in the coding scheme and their related linguistic features, 37 in total, are described with as much detail as practically possible in terms of how they are defined and operationalized. As such, most of this chapter, sections 7.3–7.7, is best used as a reference for looking up details when reading the final parts of the dissertation. Here the reader may take advantage of the many cross-references. An extended extract from the final, formatted and annotated corpus text is provided in Appendix B.

The three chapters in Part III present analyses and discussions on data extracted from the corpus as to how speakers use case endings. Each of the three chapters does this from a different angle. The first of the tree chapters in this part, Chapters 8, deals with idiosyncratic variation in the use of case endings and the extent to which case endings are used by the different speakers, and how this interacts with their use of dialectal linguistic features. The first chapter also presents data that shows that case marking, even though only used sporadically, is a productive linguistic system not limited to specific phrases or parts of the discourse. In Chapter 9 case endings are analyzed according to morphological parameters. Case marking is shown to be structured according to inflectional paradigm, with some specific forms of endings being used more often, either because they are easier to use or because they are more important markers of standardness. Some of these patterns are shown to be highly influenced by orthography. Other types of words, for example words with the definite article, are very rarely pronounced with an ending and are more or less excluded from case marking. Similar analyses are done in Chapter 10 but according to syntactic parameters, most importantly according to case and the various syntactic positions and functions that govern a particular case form. Speakers are shown to have different ways of distributing case marking across syntactic positions, yet they do this in a way so that case markers are balanced between the cases. It is also shown that there is very limited case marking on adjectival attributes, and that all speakers very consistently use a pausal form at the end of utterances.

After Part III, the results presented in this study are summarized and discussed in Chapter 11. The most prominent patterns of case marking that are consistent between speakers are listed and commented on. Some of these patterns have very direct implications for Arabic language pedagogy in that some forms of case marking are so rarely used in speech and are not marked in writ-

ing and thereby in practice only appear in very specialized uses of the language, such as recitations and news-broadcasts, and then only on the basis of textual material. Finally, some directions for further studies are proposed.

Quotes from Arabic sources are given in English, translated here if not stated otherwise. For brief quotes the original Arabic wording is given in transcription in parenthesis, and for sentence length quotes the original is printed in Arabic script in the margin. It remains to be stated that any errors in this dissertation are solely the responsibility of the author.

Part I
Background

Arabic as diglossia
Arabic as a standard language
Case in theory, tradition, and practice

Arabic as diglossia 2

The main topic of this chapter and the following one is the Arabic language situation. First, this chapter is concerned with Arabic from the perspective of diglossia, a situation where there are two varieties of the language that are used side by side, one a formal variety that primarily is written and one an informal variety that is primarily spoken. The following chapter discusses Arabic as an instance of a standard language together with accompanying notions of language ideology and linguistic correctness.

The literature on diglossia is massive. Hudson (1992) lists 1092 entries on the subject published between 1959 and 1992. Only the major trends in studies that deal directly with Arabic will be discussed in this chapter. First, diglossia as it was originally conceptualized by Ferguson in 1959, 'classic diglossia', is described in section 2.1. Thereafter four models describing Arabic diglossia as a series of levels are discussed in section 2.2, followed in section 2.3 by a discussion on how the differences between written and spoken language interact with diglossic variation. Section 2.4 is concerned with the problem of inter-speaker and section 2.5 with how this problem can be addressed by viewing diglossia as a process resulting in variation, rather than as a fixed set of varieties. Section 2.6 contains a brief discussion on why variation in case marking cannot be adequately explained as codeswitching. Section 2.7 contains a summary of the chapter.

2.1 Ferguson's classic diglossia

The concept of diglossia was first introduced to the anglophone linguistic community in a seminal article by Ferguson in 1959. In the article, Ferguson sets out to describe a specific type of language situation, exemplified by the language communities in Greece, German speaking Switzerland, Haiti, and the Arabic

speaking countries. These are all characterize by diglossia defined as

> *a relatively stable language situation in which, in addition to the primary dialects of the language (which may include a standard or regional standards), there is a very divergent, highly codified (often grammatically more complex) superimposed variety that is the vehicle of a large and respected body of written literature, either of an earlier period or in another speech community, which is learned largely through formal education and is used for most written and formal spoken purposes but is not used by any sector of the community for ordinary conversation.* (Ferguson 1959:336, italics in original)

Diglossic language communities thus differ from non-diglossic language communities in having two distinct varieties of the language that are used for different functions. There is the 'high variety' (H), which is seen as 'proper' language and is the form taught in schools and used for writing. It is used in speech only in very formal settings. The language used in everyday, informal conversation is the 'low variety' (L). L is very different from H and regarded as a deviation or a corrupt form of H. Speakers of these languages are very conscious of the existence of two separate varieties and have specific terms for the two: *français* and *créol haïten* in Haiti; *Schriftsprache* and *Schwyzertüütch* in Switzerland; *kathavérusa* and *dhimotikí* in Greece (where the situation has changed significantly since the publication of Ferguson's article in 1959); and *fuṣḥā* and *ʿāmmiyya* in the Arabic speaking countries.

The complementary functional distribution of H and L is a key element in the theory. Each of the two varieties is used in largely mutually exclusive domains. This is illustrated in Table 1 on the facing page. All uses of H in this scheme are either formal or written, and all uses of L are informal and oral, with 'captions on political cartoons' and 'folk literature' as the only exceptions. This list was composed over half a century ago and a number of technological and cultural changes have since created new domains where the uses of H and L can be mapped out. Some examples of how Ferguson's original table can be expanded to contemporary Arabic culture are listed in Table 2. These examples are drawn primarily from Holes (2013:passim). Cartoons are primarily in H in order to expose children to this variety, although major companies have recently started to dub feature films in the dialect. Film and TV-series for adult audiences are in L when in a contemporary setting, in order to reflect everyday conversation, and in H when in a historical setting. Subtitles on foreign films are in H, as are voice-overs in documentaries and translations of non-Arabic speech. Television game and entertainment shows are invariably in L. Modern pop lyrics are most often in L and sometimes in H, typically to give associations to classical romantic poetry. Commercials utilize both H and L, depending on the message and the target audience, with H signaling authority and L signal-

Table 1: *Domains of H and L. Reproduced from Ferguson (1959:329).*

	H	L
Sermon in church or mosque	x	
Instruction to servants, waiters, workmen, clerks		x
Personal letter	x	
Speech in parliament, political speech	x	
University lecture	x	
Conversation with, family, friends, colleges		x
News broadcast	x	
Radio "soap opera"		x
Newspaper editorial, news story, caption on picture	x	
Caption on political cartoon		x
Poetry	x	
Folk literature		x

Table 2: *Domains of H and L in contemporary Arabic culture*

	H	L
Cartoons	x	
Film and TV-series		
— in contemporary setting		x
— in historical setting	x	
Subtitles	x	
Voice-over	x	
TV entertainment shows		x
Pop lyrics		x
Commercials	x	x
Texting and chatting		x

ing youth culture or solidarity with domestic work (see the quote on page 25). Texting, chatting, Facebook posts, and other forms of instant messaging are written in L, often in Latin script (Palfreyman and Khalil 2003:passim), being in effect transcribed everyday speech.

The social importance of this functional distribution of H and L can hardly be overestimated. Knowing when to speak which variety is a vital part of the communicative competence (Hymes 1967:passim) of Arabic speakers. Speaking H in situation where L is appropriate is nothing short of comical. Speaking L in situations where H is appropriate may give an impression of lack of competence or lack of legitimacy. For the majority of speakers this aspect of communicative competence in Arabic simply means that they do not speak H at all, except jokingly, since situations requiring the active oral use of H are limited to only a few professions and positions.

Several of the characteristics ascribed to H by Ferguson in the quotation above — codification, association with high culture and writing — are characteristics shared by standard varieties in languages that are not considered to be diglossic, including most modern European languages. What more clearly sets diglossia apart from non-diglossic situations with standard and non-standard varieties, is that H in diglossic language communities is not used for everyday communication and is not learned natively by any segment of the population. Only L is acquired in childhood, while H is learned later in life through formal education. In non-diglossic language situations, the standard variety is based on the speech of a group of the society, typically the urban upper middle class of the capital. This segment acquires the standard variety, or something very close to it, as their native language and accordingly uses it for every-day conversation (Hudson 2002:40; Ferguson 1996:52).

Contributing to the popularity of diglossia as an analytical concept is Fishman's (1967:30) suggestion that diglossia be expanded to include "societies which are multilingual in the sense that they employ separate dialects, registers or functionally differentiated language varieties of whatever kind." This wider definition of the term led to a rise in publications of diglossia-related research, often on language situations not covered by Ferguson's narrower definition, and to a wider debate about how the term itself is best used (Hudson 1992: 617). Fishman's broader definition has been criticized for being too wide to be useful, since all modern language communities have a range of registers for different functions and therefore are in some sense diglossic. Fishman's extended diglossia also includes situations where two (or more) varieties are are used in a community but are not considered by their speakers to be the same language. This is most common in countries with a colonial past where

the local language serves as L and the colonial language serves as H. This new use of the term led Ferguson to clarify his intention that 'diglossia' should be restricted to situations where H and L are related varieties and are regarded by its speakers to be one and the same language (Ferguson 1996:75). There are then at present two parallel uses of the term (Hudson 1992:617), with Ferguson's original definition sometimes relabeled as 'classic diglossia' to differentiate it from the later broader definition. In the present study, the term 'diglossia' will be used in Ferguson's original, restricted sense.

As a description of Arabic, the view of H and L as discrete and compartmentalized entities has been criticized for being overly simplistic. When real-life language data is investigated, the picture is more complicated and less clear-cut. Politicians and preachers can, for example, be heard to switch to L even when delivering speeches and sermons, both archetypes of formal situations. Furthermore, speakers of Arabic sometimes mix Standard Arabic and dialectal features within a sentence or even within words, producing something that is neither completely H nor completely L. It has been mentioned in Ferguson's defense that he does mention the existence of such an intermediate, mixed form of speech in his original article (Ferguson 1959:332), but its implication for diglossia as a binary system was not explored. There is, in other words, more variation, and more complex variation in Arabic than can be accurately captured in the two discrete categories of H and L.

Ferguson's conceptualization of diglossia has nonetheless proved to be remarkably resilient. It is still the standard way of describing the language in Arabic teaching materials and also in linguistic textbooks. More recently it has been defended from several perspectives; as an instrument of typology of language situations (Snow 2013; Hudson 1992), for its correspondence with Arabic folk linguistics (Suleiman 2013:266) and native speaker intuitions (Parkinson 2003:30), and for the predictions Ferguson made about when diglossia would begin to be perceived as a problem by its speakers (Walters 2003:82).

2.2 Multiple-level models

The observation that there are forms of Arabic speech that are neither completely Standard Arabic (H) nor completely dialectal (L) naturally leads to the assumption that there is one or several middle or mixed varieties sandwiched between the prototypical H and L. The four most influential such models are described in this section. Particular attention is paid to how the higher registers are described in the models and to what role case endings play in defining the various levels.

The multiple-level approaches to Arabic diglossic variation analyze speech collectively, assuming that all competent speakers speak roughly the same way under similar social circumstances. This assumption faces problems when data from several speakers is analyzed in parallel. These approaches also fail to make a clear distinction between the written language and read speech on the one hand, and the spoken language on the other. These two issues are addressed in sections 2.3 and 2.4 respectively.

2.2.1 Blanc's levels

Blanc (1964) is often regarded as the first description of Arabic diglossia as a serialized set of varieties. The study investigates the mechanisms by which the 'pure' dialect is made to fit more formal situations and situations of inter-dialectal conversation. He proposes two such mechanisms for stylistic variation: 'leveling devices' and 'classicizing devices'. Leveling devices refer to the replacement of linguistic features that "sound local or rustic" (ibid.:82) with those of a prestigious and more widely understood urban dialect. Blanc gives the example of rural Palestinian dialects replacing the rural [k] with the Jerusalemite [ʔ] as the realization of /q/, or replacing the local *halʾēt* 'now' with the more widely used *hallaʾ*. By applying a number of such substitutions, local features are leveled out to produce a more prestigious variant. Classicizing devices refer to the replacement of dialectal features with markedly Standard Arabic alternatives. Examples of this are replacing [ʔ] with [q], or the use of Standard Arabic vowel patterns such as *bilād* 'lands' instead of the dialectal *blād*. Classicizing devices may also be modifications on the syntactic level, for example introducing the Standard Arabic complementizer *an* where the dialects have no complementizer. These mechanisms are applied to varying degrees to generate five 'style varieties':

> (1) "plain colloquial" refers to any local dialect, within which the speaker may select "informal" or "mildly formal" features;
>
> (2) "koineized colloquial" is any plain colloquial into which leveling devices have been more or less liberally introduced;
>
> (3) "semi-literary" or "elevated" colloquial is any plain or koineized colloquial that is classicized beyond the "mildly formal" range;
>
> (4) "modified classical" is Classical Arabic with dialectal admixtures; and
>
> (5) "standard classical" is any of a variety of Classical Arabic styles essentially without dialectal admixtures.
>
> (Blanc 1964:85)

The last of these styles, "standard classical", stands out from the rest in that it is not a vehicle for extempore speech. It is "used almost exclusively when reading aloud" (ibid.:84).

The main part of the study is a detailed analysis of a conversation between two Iraqis, a Syrian, and a Palestinian, recorded, as Blanc himself puts it, "with all the customary safeguards thrown to the winds": the informants are Arabic teachers, know that their language is being studied, and they are discussing language. Blanc nevertheless claims that the material is representative of "partly 'koineized', partly 'semi-literary'" speech (ibid.:86). The speech is analyzed in detail in terms of leveling and classicizing devices with no further mention of the five 'styles'. Blanc finds, amongst other things, that the most modifications are done in the lexicon and that the four informants differ in what mechanisms they apply. He found no instances of case endings other than in adverbial forms also used in the dialects.

The study is of less importance for its results (which are in any case difficult to evaluate due to the methodological problems) than as a pioneering work exploring middle varieties of Arabic. It foreshadows later research in its focus on how the formal register of individual speakers differ. This only comes to the fore in the lengthy detailed analysis and no general conclusion are drawn from it in the study. The classicizing and leveling devices have proven to be useful conceptual tools and in some form or another play a role in most later studies on Arabic diglossic variation.

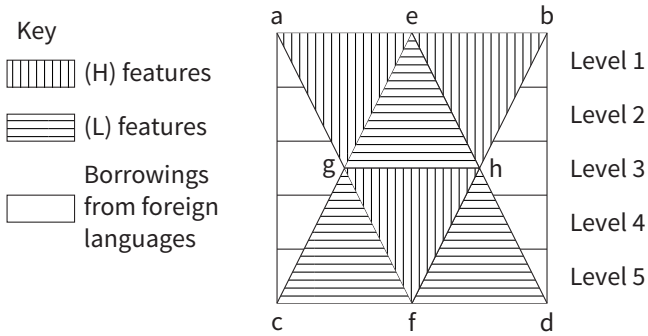
2.2.2 *Badawī's levels*

The five-level approach of Blanc was expanded upon by Badawī into more coherent, overarching system in his highly influential *Mustawayāt al-ᶜarabiyya al-muᶜāṣira fī miṣr* [*The Levels of Contemporary Arabic in Egypt*] (1973). Badawī's five levels are the following (with translations from Badawi 1985):

Level 1	Classical Arabic	*fuṣḥā t-turāth*
Level 2	Modern Standard Arabic	*fuṣḥā l-ᶜaṣr*
Level 3	Educated spoken Arabic	*ʾāmmiyyat al-mutaqqafīn*
Level 4	Semi-literate spoken Arabic	*ʾāmmiyyat al-mutanawwarīn*
Level 5	Illiterate spoken Arabic	*ʾāmmiyyat al-ummiyyīn*

The levels are defined by their domains of use, by their formal linguistic features, and by the segment of society that uses them. The domain of level 1 is almost entirely limited to religious authorities speaking on religious topics in preplanned situations, the prime example being sermons and TV-shows where re-

FIGURE 1: *Badawī's multiple-level model. Reproduced from Badawi (1985:17).*



ligious scholars answer questions from viewers on Islamic law. Level 2, Modern Standard Arabic, is used for written material dealing with topics related to modern culture or technology. Extempore speech at this level is rare. Extempore speech is found first in level 3, the language of intellectuals (*al-muṭaqqafūn*) discussing abstract and cultured topics, exemplified by panel discussions in broadcast media. Level 4 and 5 is speech on everyday topics by the literate and illiterate population respectively. Speakers often use two or more levels, ascending or descending between them in the same stretch of speech (Badawī 1973:92).

The levels are characterized linguistically by varying proportions of Standard Arabic and dialectal features on the one hand, and of foreign influences on the other. One example is the realization of /q/ which is pronounced [q] in levels 1 and 2 and is with increased frequency replaced by [ʔ] as one descends through the levels. On the syntactic level, VSO is said to be the dominant word order in level 1 and is gradually replaced by the SVO order that dominates level 5. There is, in this way, a gradual quantitative reduction of Standard Arabic and increase of dialectal features as one moves down the levels. The middle level, Educated spoken Arabic, is characterized by equal proportions of Standard Arabic and dialect features (ibid.:106). This is illustrated graphically in Figure 1. Badawī notes that these features are standard or dialectal only in terms of historical origin; each level is a distinct system (but see below) and a person speaking level 3, for instance, does not think in terms of mixing, but sees these features as a characteristic of the form of language he is currently using.

Figure 1 also shows the third type of element that make up the language in its various levels: borrowing. The foreign words in level 1 are were introduced in the classical era and have been assimilated. In level 2, there is a constant pressure to borrow new words since the language of this level deals with novel matters of foreign cultures and modern technology. At the same time, foreign borrowing in this level is limited by normative constraints (*quyūd luġawiyya*, ibid.:116). In level 3, the need for new words is equally pressing but the normative constraints are relaxed, and in this level therefore one finds the highest number of loans. In levels 4 and 5, the normative constraints remains minimal but the need for loanwords decreases. Loanwords found in these levels are those that have been assimilated in level 3 and then "trickle down" (*yatasarrab*) to the lower levels (ibid.:116).[2]

According to Badawī, the group most associated with level 1 is that of Islamic scholars. Levels 2–4 are the levels of the educated classes; level 2 for written purposes, level 3 for cultured discussion, and level 4 for the mundane and everyday. Level 3, Educated spoken Arabic, fills the same role as level 2 in that it is used for advanced or abstract topics, but is used when discussing these matters in extempore speech. It is the highest form of Arabic in which Arab intellectuals can speak comfortably. It has arisen in the gap between the lower levels, that, according to Badawī, are unfit for conversations on advanced topics, and the inability of the educated to use the higher levels for spontaneous conversation. There are a few individuals who master level 2 in extempore speech, but they are exceptions (ibid.:150). Levels 2 and 3 are described by Badawī as two sides of the same language; on the one hand the "contrived and written side" (*al-jānib al-maṣnūᶜ al-maktūb*) and on the other the "spontaneously spoken side" (*al-jānib al-tilqāʾī*, ibid.:171). This point is more explicitly made in a later article in English:

> Level 2 and level 3 share the same interests in society, namely modern life and culture. The fact that the one is spoken and the other is written does not put them in opposition as it does level 1 and level 5, rather they are in complementary distribution. A division of labor in relation to similar interests as well as being in close linguistic proximity within the scheme thus holds them together: whatever is written in level 2, which covers all aspects of modern society (science, art, technology, literature, etc.) is explained and discussed in level 3. (Badawi 1985:19-20)

[2] Badawī only discusses foreign influences in terms of lexical borrowing. Figure 1 suggests that this is seen as competing with Arabic H and L features that operate on all levels of linguistic description, not only the lexical level. It is, in other words, not clear why an influx of foreign words would lead to a reduction of Arabic morphosyntactic or phonological features.

The boundary between levels 2 and 3 is thus one of speech and writing. At the same time the line between levels 2 and 3 is said to be the line between Standard Arabic (*al-fuṣḥā*) and the dialect (*al-ᶜāmmiyya*). One of the definitions given for level 3 is that it is the level "in which the features of traditional Standard Arabic [as described by the classical grammarians] are reduced to a point where it is no longer possible to regard it as within the boundaries of Standard Arabic" (Badawī 1973:149). The borders between spoken and written language on the one hand, and between Standard Arabic and the dialect on the other, coincide between Levels 2 and 3. In other words, only that which is written is Standard Arabic proper.

هو الذي يصل تدهور صفات الفصحى التقليدية إلى حد يصبح معه من غير الممكن أن يبقى داخل العربية الفصحى.

Case endings occur only in the upper part of the scale.[3] One of the differences between levels 1 and 2 is the commitment to case endings in the former. In level 1, case endings are for example retained in proper names and in numerals. Speakers performing at this level are careful to retain case endings in these positions precisely because it distinguishes their speech from level 2. Badawī reports how the Egyptian Language Academy after a lengthy discussion officially permitted the omission of case endings in proper names (ibid.:125–6).[4] The use of case endings in proper names and numbers is thus a sign of linguistic conservatism surpassing even that of the Academy. Not producing case endings in proper names and numbers is permissible in level 2, in the sense that it is not noted as a mistake by listeners and is not subject to correction. Still, it is, according to Badawī, inherently more difficult to speak in a prescriptively correct manner in level 2 than in level 1, since the topics dealt with in level 2 are vastly more diverse, that is, not only religious, and the speaker must focus more on content, with less attention paid to formal features (ibid.:128). The degree to which the case system is followed in level 2 varies with individuals, topics, and situations, but there is near unanimity in not producing case endings on numbers and proper names (ibid.:143). Producing the *wrong* case ending on the other hand is a mistake and subject to correction in all levels (ibid.:135). In level 3, case endings are only used very rarely for emphasis and only in 'safe' positions, such as genitive markers after prepositions (ibid.:170). There are no case endings in levels 4 and 5 other than in lexicalized forms and formulaic expressions.

يجوز الوقوف بالسكون عند تتابع الأعلام في مثل «سافر محمد علي حسين» مع حذف «ابن» تيسيرًا على القراء والكتّاب، وتخلصًا من صعوبة الإعراب.

[3] Badawī uses the term '*iᶜrāb*' but only gives examples of case and not of mood, the other part of the *iᶜrāb* system. See 4.2 for a discussion of this term.

[4] The statement of the Academy, published in *Majallat majmaᶜ al-luġa l-ᶜarabiyya* [Journal of the of the Arabic Language Academy], nr. 20, p. 110 and quoted in Badawī (1973:126) reads: "It is permissible to use the pausal form in the sequence of a proper name, such as in 'Muḥammed ᶜAlī Ḥusayn traveled' with the omission of *ibn* ['son of'], in order to facilitate for readers and writers and to make away with the difficulties of *iᶜrāb*."

Throughout Badawī's study, the levels are described in two contradictory ways, each emphasized in different contexts. On the one hand, they are described as theoretical constructs slicing a continuous scale of linguistic variation and are, in this sense, arbitrary. On the other hand, they are described as qualitatively different entities, each with its own self-contained linguistic system. Badawī is very explicit about the fuzziness of the system. At one point, he likens the levels to the colors of the rainbow, stating that the fact that there are areas between the colors in the rainbow where two colors are mixed does not hinder us from identifying areas where the colors are clear and unmixed, which allows us to see it as a series of clearly defined colors (ibid.:94). In the description of individual levels, they are however often defined in relative quantitative terms, with one feature being used less or more frequently in neighboring levels. In effect even the clear colors are defined as mixes of the ends of the spectrum, in keeping with the analogy.

The levels developed by Badawī have a intuitive appeal in their direct applicability. Holes (2004:345) for example describes Badawī's model as "a useful picture both of how variation is structurally organized […] and what the social meaning of variation is." He applies it to an analysis of style shifts in ᶜAbd an-Nāṣir's speeches and in radio interviews. Badawī himself recently elegantly applied his levels in Arab television commercials:

> The highest level [level 1] is used only in admonitory commercials promoting water conservation or sexual chastity to combat AIDS, and the language is invariably Quranic. The second level down [level 2] is employed as the male 'voice of reason' or scientific voice, which often appears in voice-overs to banking or car advertisement. At the middle level [level 3], things are a bit more intimate: a family doctor talks to the camera, lecturing on the quasi-scientific properties of a brand of toothpaste, the protective quality of a baby diaper, or the curative powers of a certain medicine. But most ads on Arab national television are delivered on the next level [level 4], the 'standard colloquial,' of each region, which is reserved for object of indulgence such as food and fancy clothes. A small number of commercials use the pure, 'mother-tongue' colloquial language [level 5] for goods aimed at the lowest stratum of society.
>
> (reported in Hammond 2007:64)

An attempt to test Badawī's (1973) claims and to contrast them with Ferguson's (1959) classic diglossia was made in Elgibali (1985).[5] In the study, Elgibali computes the ratio of six sociolinguistic markers in each level to see whether there is a continuous or a stepwise transition. No stepwise difference was found and this fact was interpreted as evidence against the five-level model. The study has a number of problems, especially with regards to data collection. Data was

[5] See 4.8 for Elgibali's (1985) results regarding case endings.

collected to represent formal and informal speech and formal and informal writing in each of the five levels. This leads to some problematic cases, especially since formality is not explicitly defined. In level 1, formal speech is represented by recitations of the Quran and classical poetry, and informal speech by "Muslim religious oratory" (Elgibali 1985:38). Formal level 5 was "gleaned from conversations between employees and employers" (ibid.:40). What these two situations have in common that defines them as formal is not clear. Much of the analysis builds on comparisons of informal data from one level with formal data from one level down. The idea behind this is that it this is the smallest step where Badawī's model predicts a systematic distinction. Furthermore, the definition of what would constitute a discrete level builds on the assumption that Badawī's level 1 and level 5 correspond to Ferguson's H and L, while levels 2–3 are middle levels not covered in Ferguson's classical diglossia (ibid.:37). This is a misreading of both authors. Badawī's level 1 and 5 are both very narrow and cover only a limited segment of language use, of religious scholars and the illiterate respectively. It follows from Elgibali's reading that news broadcasts (level 4) are not included in Ferguson's H and that informal conversation amongst literate persons (level 2) is not L.

2.2.3 Educated Spoken Arabic

The middle variety that is neither Standard Arabic nor clearly dialect was investigated in a number of publications by El-Hassan (1977, 1978) and Mitchell (1978b, 1980, 1986) and Mitchell and El-Hassan (1994). They argue that there is a middle variety, labeled Educated Spoken Arabic (ESA) which is "a highly significant form of Arabic whose grammar is mostly shared by Egypt and the countries of the Levant" (Mitchell and El-Hassan 1994:2). The project is difficult to summarize in that descriptions of its theoretical assumptions, of ESA as a variety, and of the corpus on which these studies are based are scattered around the publications and do not always match. ESA itself is described in various ways. Mitchell (1986:10) sketches ESA as lacking Standard Arabic case and mood endings and as replacing the Standard Arabic negations, numerals and duals with their dialectal forms. ESA is said to be different from the multiple-level models of Blanc, Badawī, and Meiseles (see next section for a review of the latter): "We do not see it [ESA] as one of a series of separate varieties, on par with MSA and the vernaculars, but rather as created and maintained by the constant interplay of written and vernacular Arabic" (Mitchell 1990:12-13). Elsewhere, Mitchell (1985:56) states that "although it draws heavily upon both MSA and the vernacular, it nevertheless is its own variety of Arabic" (quoted in Ryding 2006:213). It is, in other words, difficult to get a clear idea of ESA from the

various publications. The presentations of the ESA of Mitchell and El-Hassan that follow is therefore more critical than are the other schemes reviewed in this chapter.

The ESA project is based on a large body of recorded material, known as the 'Leeds corpus', of conversations from Egypt, Syria, Jordan, the West Bank, and Kuwait between speakers of a wide range of occupations, from house wives and university students to diplomats and artists. Participants were asked to speak freely about their area of specialization (El-Hassan 1977:120, 1978:33). No conclusive description of the corpus and methods of data collection has been published. The following is gathered from remarks found in the various publications. El-Hassan (1977:120) describes the speakers in the corpus as falling in nine enumerated occupational categories that are meant to represent the educated classes, including high-school students, writers, diplomats and artists. A similar list in El-Hassan (1978:33) is slightly different and also includes house wives. The situations in which the recordings are done also vary considerably. The corpus consists of "discussions based on a wide range of inter-personal relationships" (El-Hassan 1977:120), including elicited speech where participants were asked to speak freely about themselves and their work, as well as discussions recorded from radio and television (El-Hassan 1978:33). El-Hassan (1977: 114) gives as an example of ESA a long extract from a Muslim sermon. Still, Mitchell and El-Hassan (1994:2) stress that material was selected to represent informal speech.[6]

In one of the first publications on the topic of variation, El-Hassan (1977: passim) criticizes previous descriptions of variation in Arabic, including those of Ferguson and Blanc, on the basis of counterexamples from the corpus described above. For example, Blanc's claim that there are no *iʿrāb* endings in the middle register and that negative particles are entirely dialectal is refuted on the basis of examples from the corpus. El-Hassan draws the conclusion that ESA must be studied with the variationist approach of Labovian linguistics. This approach was pursued in El-Hassan (1978) where the occurrence of dialectal and standard forms of demonstratives in a part of the Leeds corpus are quantitatively analyzed. He found a fair amount of variation between speakers but concludes that all national groups "used a fairly high percentage of the prestige-forms, which, among other factors, provides some justification for recognizing educated spoken Arabic as a social dialect". He also found that women are

[6] The extract is in fact interesting as an example of clear-cut switches between H and L. Statements are made in Standard Arabic very close to the prescriptive norm and are then repeated, and in the latter half if the extracts briefly expanded in the dialect. The text could be analyzed as codeswitching (see 2.6), rather than as an instance of ESA.

not initiators of linguistic change in Arabic since they use less standard forms than do men (El-Hassan 1978:53).[7] He also suggests, based on his data, that the "pan-Arab grammar of educated spoken Arabic" in Egypt and the Levant should contain the following six forms of the feminine demonstrative pronoun ('this') for objects near the speaker: *haaði(hi)*, *haazi(hi)*, *haadi*, *ha(a)y*, *haydi*, and *di* (transcription as in original), with specifications of in which proportion each is used in the different regions (ibid.:54).

Mitchell has a different, qualitative approach to the study of ESA. In two articles (Mitchell 1980, 1990), he sets out to distinguish dialectal stigmatized features and high-flown standard features that are not part of ESA, describing a filter, as it were, to what can be brought from the dialects and from Standard Arabic to ESA. Examples of stigmatized features are, for Egyptian Arabic, the aspects marker *tann-* 'again' and the use of the words *wād* 'boy' and *bitt* 'girl' instead of the standard, or non-local *walad* and *bint*. Examples from Palestinian Arabic are [tš] as realization of /k/, and for Syrian Arabic [g] as realization of /q/ (Mitchell 1980:90). One example of high-flown features is *iʿrāb* (case and mood endings) since "if they use ʔiʕraab then by definition they are not speaking ESA" (Mitchell 1986:19, underline in original). One of the conclusions drawn by Mitchell after having analyzed a number of examples, is that the grammar of ESA will necessarily be fuzzy and indeterminate to account for the variation found in the data (Mitchell 1980:105).

The Leeds project has met severe criticism. Haeri (1996:19) discusses the problem of categorizing speakers as 'educated' and downplaying the heterogeneity within this group with regards to religion, social class, educational background, and especially the difference in being educated in private schools, where a foreign language is the medium of education, or in a public school, where Arabic is the medium of instruction. Van Mol (2003:58-70) gives an extensive critique of the project and points amongst other things to the fact that little situational data is given about the recordings in the corpus and that some conversations may have been very informal and almost entirely in the dialect, whereas others are more formal, and that the same informant would have spoken very differently in these situations. Despite this, they are all regarded as instances of the same variety.

Of special interest here is the delimitation between ESA and Standard Arabic. As we saw above, Mitchell draws the line with *iʿrāb* endings, whereas

[7] This is an interpretation of the facts built on the identification with Standard Arabic as the prestigious variant in the Labovian sense, the target of linguistic adaptation of the middle class. This is a view that has been abandoned in later scholarship (Ibrahim 1986:121; Al-Wer 1997: 256–9).

El-Hassan (1977:120) regards these endings as part of ESA since they appear in their corpus. However, it is difficult to tell from the few examples discussed in the articles how representative the uses of case endings are of the corpus as a whole. Mitchell (1986:10) gives an impressionistic description of ESA as lacking case and mood endings, using dialectal negation forms and numbers and not using dual concord. In other words, ESA seems to be distinguished from Standard Arabic in that prescriptive grammar is not followed. Likewise in El-Hassan (1977), more than half of the twenty-six numbered examples of ESA are in fact prescriptively correct Standard Arabic except for the lack of case endings and minor phonetic variations. Deviations from the prescribed grammar, even omissions of case endings or traces of a regional accent, make speech non-standard and thus ESA. As in Badawī's scheme, this in effect means that virtually no spoken Arabic is Standard Arabic.

If ESA is delimited in the higher end by not complying with codified grammar, its lower limit is much less clear. As pointed out in Van Mol (2003:62), the fact that some stigmatized regional features are not present in the speech of the educated, as described by Mitchell, may be better explained as results of a general language change within the dialects towards the prestigious variety of the urban centers, rather than by proposing a separate, identifiable register. In other words, Palestinian [tš] and Syrian [g] are associated with socially less prestigious groups, and educated speakers either do not have them in their sociolect or drop them in accordance with well understood sociolinguistic patterns (Labov 1972:passim). With this very high upper limit and undefined lower limit, ESA in effect becomes *all speech of all educated Arabs*.

Despite its shortcomings, the influence of ESA as a concept has been substantial. ESA is now a commonly used term in Arabic linguistics, and has also inspired novel approaches to language instruction (e.g. Ryding 1991; Younes 2015).

2.2.4 Meiseles' 'quadrigossia' and Oral Literary Arabic

The wide scope of ESA as defined by the Leeds team was noted by Meiseles (1980). According to him, speech in the middle variety is always oriented towards either Standard Arabic (Literary Arabic (LA) in his terminology) or the speaker's dialect. He therefore proposed a model where the middle range is split into ESA as the lower segment and sub-Standard Arabic as the upper segment. With LA above and "basic or plain vernaculars" below this gives a set of four 'quadriglossic' levels:

1. Literary (or standard) Arabic
2. Sub-standard Arabic
3. Educated spoken Arabic
4. Basic or plain vernaculars

(Meiseles 1980:123)

Sub-Standard Arabic, the second level, has both a written and a spoken form. The written form, Informal Written Arabic (IWA) is "the extemporaneous writing the social circumstances of whose production do not pressure the writer into closely observing the language quality of his writing" (ibid.:125).[8] The spoken form of Sub-Standard Arabic, Oral Literary Arabic (OLA), is of particular interest for the present study. It is language "produced under such psychological and social circumstances as 'dictate' the prestige language. It is the language generally used for oral expression in the mass media communication, on all formal occasions and many a time also on semi-formal ones" (ibid.: 125).[9] These are situations where the speaker's intention is to use the form of the language. The notion that there is a way of actually speaking Standard Arabic that is different from merely speaking in a more educated or formal way is an important difference from previous models. Speakers may set out to speak Standard Arabic in a conscious switch to a different kind of language that is quite different from speaking in a more cultured form of one's dialect. This is especially true regarding the 'leveling devices' that have a regional prestige variety, the variety of a prestigious social group, as their target. Here unconscious sociolinguistic processes are set in motion. Speaking Standard Arabic, on the other hand, is not done in imitation of a social group and is not governed by the same social mechanisms. In fact, "the variety termed Sub-standard Arabic (=SsA) is defined exactly (in Ferguson's words) as an 'Arab's attempt to speak [or even write—G.M.] classical Arabic'" (ibid.:124, brackets in original). As will be discussed in the next chapter, the idea that speakers attempt (and consistently fail) to speak prescriptively correct Standard Arabic is problematic. The quote does, however, underline the importance of the speaker's intention in speaking a variety that is, for the speaker, distinctly different from their own or any other dialectal variety and that they identify as Standard Arabic.

According to Meiseles, OLA contrasts with Standard Arabic in that it contains dialectal admixtures, such as the *b*-prefix for imperfect indicative, exis-

[8] For detailed analysis of informal Arabic writing, see Meiseles (1979) and Belnap and Bishop (2003).

[9] 'Mass media communication' is here interpreted to include only news oriented genres and not entertainment. This distinction was less important when Meiseles' article was published, before the expansion of satellite television with its broad selection of entertainment shows (Hammond 2007:chap. 8) in which dialectal Arabic dominates.

tential *fī*, and the complementizer *innu*. Phonetically it is conditioned by the dialect of the speaker, which is a source of regional variation in OLA. It is distinguished from ESA on the other hand by classicizing devices, to use Blanc's term. One important such device is lexical choice in which the speaker "finds the easiest and surest way of demonstrating his desire to to 'dignify' his speech, using OLA rather than ESA" (ibid.:128). Meiseles exemplifies this with *raʾā* 'see' and *ḏahaba* 'go' instead of the dialectal *šāf* and *rāḥ*, and the relative pronouns *allaḏī* and its variants instead of the dialectal *illi*.[10] In cases where there are Standard Arabic lexical alternatives, the tendency in OLA is to choose a word that is not shared with the dialects, such as *istaṭāʿa* 'be able to' rather than the synonymous *qadara* that is also used in the dialects. Other examples of classicizing features that distinguish OLA from ESA are the use of dual concord, Standard Arabic passive forms, articulation of saliently Standard Arabic sounds, particularly [q], and the occasional use *iʿrāb* (case and mood) endings. The latter features is singled out by Meiseles as "a clear indicator of the speaker's desire to express himself in OLA" (ibid.:129).[11] These characteristics are said to be mere 'tendencies', leaving the question open as to how many of these features need be present and at what rate for a string of speech to cross the border from ESA to OLA. Furthermore, like the previously discussed authors, Meiseles cautions that there are frequent switches between varieties even down to the sentence level (ibid.:132), again making the model difficult to apply in practice. He also makes it clear that reading aloud is not OLA but more appropriately regarded as a form of written language (ibid.:125).

All this makes OLA a close equivalent to what in the present study is called Spoken Standard Arabic, and its domains of use matches closely the external criteria set up for the collection of material for the present corpus (see 5.2). Impressionistically, the description of OLA matches the style of most of the speakers in in the corpus. More problematic is the view held by Meiseles that in Sub-standard Arabic, of which OLA is a part, is "the reconstruction of the prescriptive LA model is the goal" (ibid.:125). The argument proposed here is instead that the spoken standard language has its own set of norms that differ from those of the codified grammar.

[10] Cf. Table 15 on page 177 for counts if these two variants of the relative pronoun in the present corpus.

[11] The use of *iʿrāb* in OLA is the subject of Meiseles (1979), reviewed in 4.8.

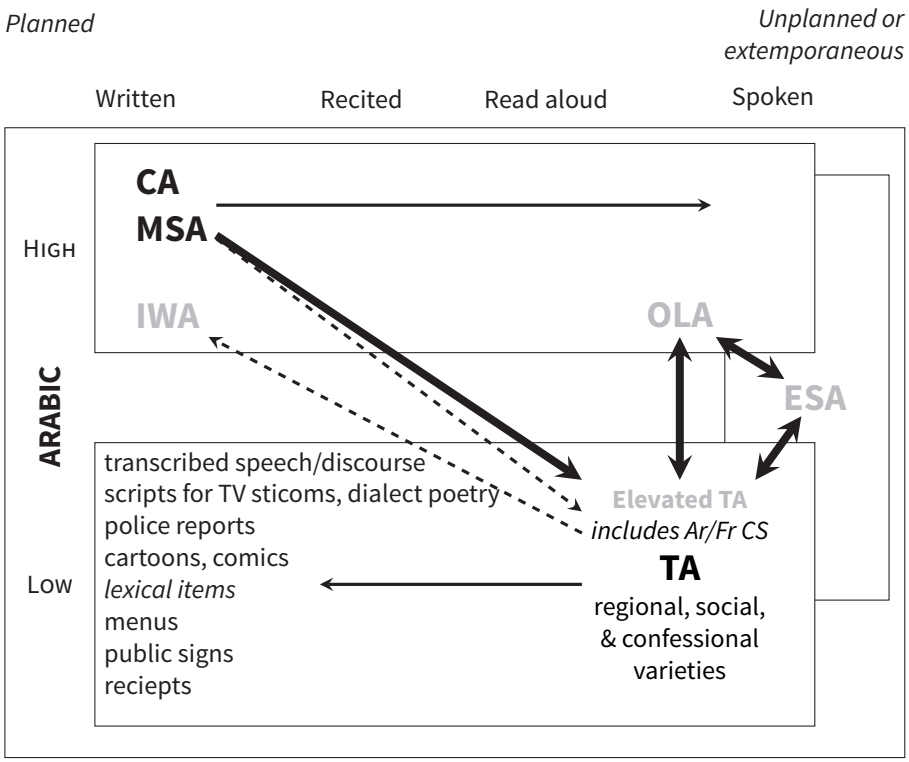
2.3 Speech and writing in diglossia

In the multiple-level models described above, the higher end of the spectrum is identified on the one hand with the language as represented in codified grammar, and on the other as the written language, the two being regarded as more or less identical. This section discusses the problems with the identification of the higher end of the spectrum with writing. The identification of the higher end with codified grammar is discussed in the following chapter.

In Blanc's scheme the highest style of oral Arabic is used exclusively for reading aloud, that is, an oral manifestation of writing. Similarly, Badawī draws a line between his levels 2 and 3 in that the former is written and the latter is spoken. In the ESA paradigm, the line is less clear, but the features associated with ESA — none or sporadic *iʿrāb* endings, use of dialectal forms and affixes — are characteristic also of virtually all forms of spoken Arabic, leaving only writing to Standard Arabic proper. Meiseles sees OLA as the attempt or intention to speak Standard Arabic (his LA), which itself is placed as a separate, written variety above it. All models relate speech to writing in the same scale of variation. They describe a serialized set of varieties, starting at the bottom with 'pure' dialect that further up becomes more similar to writing and finally becomes writing. Authors of the multiple-level approach also point to frequent shifts between levels. Presumably this cannot be done between the levels of speaking and the levels of writing unless by these shifts we mean that a speaker begins to read textual material aloud. In no event does a person who speaks in a more and more formal manner at some point start to read. There is in other words a clear discontinuation in the series of levels in terms of mode of production, and the above mentioned authors are clearly aware of this. It does however raise the question as to why spoken and written language is serialized on one and the same scale to begin with.

The discontinuation between speech and writing in these views of diglossia is elegantly illustrated by Walters (2003:97). He discusses what he calls 'post-diglossic' Arabic, a situation more complex than the binary scheme of classic diglossia due to widespread literacy resulting in changes in practices. He introduces a spoken-written axis as an additional second dimension of variation and practice, next to the dialect-standard dimension of previous models. In this two-dimensional space with H and L as the vertical axis and modality on th horizontal axis, different varieties of Arabic can be mapped out. This neatly summarizes the multiple-level models and also illustrates how they fail to take the difference between speech and writing into account. Modern Standard Arabic (MSA) and CA are accordingly represented in the H/written corner and OLA in the H/spoken corner. This is illustrated in Figure 2. in which Walter applies

FIGURE 2: *'Post-diglossic' Arabic. Reproduced from Walters (2003:97).*



[TA = Tunisian Arabic; CS=codeswitching.]

the model to the language situation in Tunisia (hence the prominent role of French-Arabic codeswitching). Movement up or down in the diglossic continuum is qualitatively different from sideways movement into another modality, a fact that one dimensional models of diglossia fail to account for. The power of this model is shown in the fact that it incorporates Ferguson's classical diglossia, Meiseles' OLA, as well as the ESA of El-Hassan and Mitchell. Note the wide scope of ESA in the figure which can be seen spanning most of the vertical axis.

Reviewing Ferguson's examples of how H and L are used in different situations (Tables 1 on page 17), it is clear that the situations where H (Standard Arabic) is used are either written, read aloud, or heavily planned, as in lectures an sermons, whereas L (the dialect) is used for extemporaneous speech. Writing is, as it were, the traditional domain of Standard Arabic, and speaking is the

traditional domain of the dialect. Most communication in Arabic takes either of these two forms, written Standard Arabic or spoken dialect, both of which are therefore highly conventionalized forms of the language. As one moves up towards the standard language in speech, or down towards the dialect in writing, more variation is introduced, with lower density of communication and less conventionalized forms. In the Figure 2 this is represented by OLA and ESA being printed in gray.[12] Social changes, especially widespread education and literacy, mean that the previously marginal areas of written L and spoken H have gained more prominence, a process that is reshaping the nature of diglossia. Hence the term 'postdiglossic' Arabic. As a consequence, written L and spoken H are becoming more and more conventionalized, but are still "best studied as a set of practices rather then being reified into varieties unto themselves" (Walters 2003:102).

Introducing the dimension of speech and writing makes it clear that the prescriptively correct form of Arabic that is most closely associated with writing is detached from speech. By moving upwards the scale of formality in spoken situations, one does not eventually reach prescriptively correct speech. Taking it as a point of orientation in the analysis of linguistic variation in speech is therefore problematic. A proposed solution to this problem is presented in the next chapter.

2.4 Inter-speaker variation

Under the multiple-level approaches to diglossic variation, data is analyzed collectively; data from a number of speakers are gathered and analyzed as a whole in an attempt to find the overarching system. This is done without considering the possibility that individual speakers may actually use different methods to reach some desired level of standardness. There are a large number of devices a speaker may use to make their language more standard: lexical, phonetic or morphemic substitution, syntactic alterations, or addition of Standard Arabic particles such as the emphatic *qad* and *inna* that do not alter the syntactic structure of the clause. Different speakers may use any number of these devices in different combinations and at different rates. The resulting outputs may all be speech that reaches similar levels of standardness, as impressionistically judged by a listener, but that are linguistically very different in that they consist of different combinations of standard and dialect features. The process that produces the higher registers may, in other words, differ substantially between individuals, both quantitatively in how consistently it modifies the language,

[12] The original figure uses outline text.

and qualitatively in what changes it applies. Schulz (1981:19) for example states that his material, being mostly radio broadcaster panel discussion, should be mapped into Badawī's level 3. The speakers are therefore expected to speak in more or less the same way and use features associated with that level. On the contrary, he continues, "we get a very broad spectrum of speech, ranging form what is almost 'pure' colloquial, to what is almost 'pure' classical." In the same vain, Mejdell (2006:47) describes multiple-level models such as those described above as tending to be "flawed by lack of, or only minimal, empirical support, and turn out to be difficult to apply to natural data".

That there would be such differences between individuals' oral production of Standard Arabic is to be expected, given how Standard Arabic is acquired, how it is perceived, and the limited extent to which it is used. The Arabic speaking child learns Standard Arabic as a second language when starting school, and most Arabs are never in a position to be asked to use it orally after finishing their education, since only a small portion of the population ever speaks in public or on formal occasions. Even those who do speak in public do not adhere to all the intricacies of Standard Arabic grammar that the educational system insists are crucial (see 4.3). Nor is there, as pointed out in Owens and Bani-Yasin (1991:25), a group of native speakers of Standard Arabic whose speech could constitute a model or norm. Every speaker has to find out for themselves how to speak Standard Arabic, and, if there are implicit conventions, it is up to every individual speaker to discern and generalize them. Furthermore, as in any other case of second language acquisition, individuals differ as to what extent and what parts of the new grammar they master. They will thus have different resources available for expressing themselves in this second language.

Apart from what linguistic devices speakers have at their disposal, there is also the question of how specific standard or non-standard features are judged to be, and, by extension, how appropriate they are seen to be for different contexts. Parkinson (1991) conducted a series of experiments in Cairo eliciting judgments of various spoken and written texts. In one experiment participants were asked to rank five written texts on different topics and with different amounts of dialectal features for suitability for the topic. Participants gave widely different rankings to the texts (ibid.:48). In a matched-guise test participants listened to seven different readings of the same text, every reading with different sets of standard and dialectal phonological and morphological features and with dialectal or standard renderings of numbers. Participants were asked to state whether or not each text is Standard Arabic (*fuṣḥā*) and then to rank them on a seven-graded scale from standard to dialect. For the first question, 75% of participants described all the texts as Standard Arabic indicating

that participants accept a wide range if variation within Standard Arabic. On the scale of standardness, phonological features, and not vowel endings, were shown to be the most important. No statistical difference was found in the ranking of texts with different realizations of ending vowels as long as traditionally correct Standard Arabic phonology was maintained (Parkinson 1991:58).

In interviews with Arab students about their style of writing in personal correspondence, Belnap and Bishop (2003:16) found "not only that various MSA features are hierarchically evaluated (some more formal, some less), but also that individuals differ in their ranking of some features." A Saudi informant for example said that he found the Standard Arabic passive forms too formal and avoided it, while a Jordanian said that she loves to use the passive (ibid.: 13).

Such differing judgments of individual items affect the target they have in mind, consciously or not, when developing proficiencies in Spoken Standard Arabic. Having different target outputs, speakers differ in the means, or 'devices', they use to reach a desired level of standardness and in the importance they ascribe to them. The situation is neatly summed up by Haeri:

> The extreme variation in the degree of knowledge in Classical Arabic on the part of the speakers, the kinds and frequencies of linguistic varieties that they get exposed to throughout their lives, etc., all contribute to non-uniform perceptions regarding the same 'piece of language'. (Haeri 1996:¶25)

Inter-speaker variation should be taken into account when analyzing formal Arabic speech, since it is quite possible that speakers modify the formality level of their speech in different ways. This approach was taken Schulz (1981) who in an impressive study counted a large number of features in material of "more than 1000 pages of transcription" (ibid.:14) from 49 speakers. The large number of speakers did however limit the amount of speaker time for each individual,[13] and the reliance on tabulating results by hand meant that the analysis is rather crude. More recently, small-scale studies with a focus on individual differences have been made by Parkinson (1994a) with material from four speakers, and Mejdell (2006), 7 speakers. These three studies have been important sources of inspiration for the present project and the results presented in them will serve as a points of comparison.

2.5 Diglossia as process

A problem facing the multiple-level paradigm of diglossic variation is, as described above, the amount of variation between the speech of individuals talk-

[13] As a very rough comparison, each of the 49 speakers in Schulz's corpus produces on average

ing in similar situations. The paradigm assumes that people speak in roughly the same way when acting under similar social circumstances. An alternative model of diglossic variation that may better account for the observed inter-speaker variation is what in Hawkins (1983) is called a 'process model' of diglossia. This model sees formal varieties of the diglossic continuum as the result of a transformational process within the individual speaker, very much in line with Blanc's 'devices'. Hawkins, in discussing the Greek diglossic situation, is critical of what he calls 'the static model' in which variation is stratified into levels or varieties. He gives the example of Householder (1962) who lists eleven levels of Greek, ranging form "Extreme (archaic) K[atharévousa]", that is upper H, to "Extreme (conversational) D[himikí]", namely lower L. The decision to use one of these labels for a certain text is, according to Hawkins, "either impossible or totally arbitrary" (1983:13). Instead, he proposes a process model of diglossia which "takes L as a basis and assumes that an indefinite number of varieties of H can be created from it, by the application of rules for 'purification'" (Hawkins 1983:12–13). H is by this model is not a different system from L, the native tongue, but a modification of it. Speakers use their native L as basis and by substitutions of words and phonemes, additions of grammatical markers, etc., approach H. A descriptive framework based on the process model would focus on these modification, the *means*, by which L is transformed into H, rather then describing and delimiting the output of these means as static levels.

The concept of diglossia as a process relies on the speaker's dialect being the underlying system also of semi-formal or formal forms spoken Arabic. Owens (2001:430–31), summarizing the literature, gives three basic arguments for this view. Firstly, the dialects are acquired natively while Standard Arabic is a second language. Secondly, affix material is dominantly dialectal, giving this the status of the 'matrix-language' (Myers-Scotton 1993:20). Thirdly, in states of high emotional tension, speakers tend to revert to the dialect thus revealing the cognitive primacy of this variety.

One might question whether the transformational rules by themselves explain the whole spectrum of non-standard — standard variation. Continuously applying transformational rules to produce Standard Arabic from a dialectal basis might be cognitively too costly a task to be viable. As a model for understanding variation in case marking in Spoken Standard Arabic, it is, however, a very useful approach. As will be further described in Part III, Spoken Standard Arabic is characterized by Standard Arabic word forms but with only sporadic use of case endings. Speakers can potentially speak in a high level of Arabic

112 sentences (calculated from Schulz 1981:206–14). The seventeen speakers in the present corpus produce an average of 199 sentences each.

without using case endings. The addition of case endings to words can be understood as a transformation of underlying unmarked word forms. The distributional structure of this transformation is the primary object of investigation in this study.

Further, in critique of the multi-level approaches of diglossia and their failure in accounting for complex variation, Hary (1996) proposes doing away with the notion of levels by applying a stricter view of diglossia as a continuum. In this stricter continuum model, speakers position themselves somewhere on the scale between the theoretical L and H poles, referred to by Hary as 'Variety A' and 'Variety C'. He proposes for the middle range between these poles a set of varieties labeled " 'Variety Bn', indicating the practically infinite variations in the middle register" (ibid.:72). In practice, all cases of spoken Arabic are Bn since the two endpoints of the scale represents theoretically 'pure' forms of language, while in reality there is always some form of mixture (ibid.:72). Instead of switching between varieties in a single stretch of speech, speakers slide along the continuum without crossing any borders. When approaching the upper end of the scale, speech starts to be perceived as Standard Arabic, and, when approaching the lower end, it starts to be perceived as dialect. What is of interest then is not where on the scale a speaker is positioned in some absolute sense but by what means he or she moves up and down the scale. This movement, namely the use of dialectal and standard features, is not random but follow implicational rules, making it theoretically possible to write a grammar of Bn. Hary's main argument for the systematic nature of Bn is that his native speakers informants were able to rank twelve variants of one individual word with different combinations of standard and dialectal features on a scale of standardness and they also discarded some combinations of features as unacceptable. This they did with a high degree of agreement, suggesting that the variation within Bn is governed by a hierarchical system.[14] Hary points out, with reference to Parkinson (1991), that when complete texts are rated in this way, a much more complicated task, there is little agreement.

The process model of diglossia is thus an attractive alternative to leveling approaches in investigating the use of case endings in Spoken Standard Arabic. It allows for analysis on the level of the individual speaker and for comparisons between individuals (Hawkins 1983:13).

[14] The fact that Hary's informants all have a background in linguistics may well explain their similar responses. Some are reported to have asked if an particular item was pronounced with a diphthong or a monophthong (Hary 1996:90n).

2.6 Codeswitching

Arabic diglossic variation has also been studied from the perspective of codeswitching. Codeswitching refers, broadly, to the alternation between two languages or language varieties in the same segment of speech with switches occurring either intra-sententially or inter-sententially. Dialectal Arabic and Standard Arabic, when seen as two separate linguistic systems, can be analyzed within this framework. Codeswitching can be studied with respect to its socio-pragmatic functions and in terms of its formal linguistic characteristics. Both approaches have been applied to Arabic diglossia. A forerunner in the socio-pragmatic approach was Holes (1993), who investigated how the Egyptian president Jamāl ᶜAbd an-Nāṣir used switches between the dialect and Standard Arabic in his speeches. He found that ᶜAbd an-Nāṣir employed Standard Arabic to relate the "abstract, idealized or metaphorical" and Egyptian Arabic for the "concrete and physical", often with the two in parallel so that principles conveyed in Standard Arabic are given concrete examples in the dialect (ibid.:33). Similar conclusions are drawn in Mazraani (1997:189–90) who apart from ᶜAbd an-Nāṣir also studied speeches of Ṣaḍḍām Ḥussayn and Muᶜammar al-Qaḍḍāfī. Albirini (2011) investigated the use of switches in various TV-shows and gives a long list of different functions that switches may fill, adding to the previous studies for example switches to the dialect to tell jokes and to downgrade the object at hand, and switches to Standard Arabic to adopt the role of an expert. Albirini also makes the important observation that the respective status of the two variants are not undermined by such practices but rather reinforced by them (ibid.:555).

The linguistic constraints of Arabic diglossic codeswitching have been studied by Eid (1982). She studied the occurrences of Egyptian and Standard Arabic items before and after 'focal points': relative pronouns, subordinators, tense markers and negative particles in their Egyptian and Standard Arabic forms. She found that switches occurred freely before these items but were constrained after them. No switch occurs after Standard Arabic focal points, and, after Egyptian Arabic focal points, switches occur freely except after negation. In Eid (1988), she argues that this is due to the Egyptian negation *miš* not marking tense and so must be followed by a tensed verb, while in Standard Arabic the negating particle carries tense and must be followed by an untensed verb. Bassiouney (2003) tested three theories of codeswitching constraints on Egyptian data and found that Myers-Scotton's *matrix language* hypotheses (Myers-Scotton 1993) best explained the data. This hypothesis states that in a segment of speech all 'system morphemes' are taken from one language, the matrix language, whereas the other *embedded language* provides content words.

The study of Arabic diglossia in terms of codeswitching is complicated by the similarity and the many instances of overlap between the two systems, making it difficult to categorize certain items as dialect or Standard Arabic and to pinpoint where a switch occurs. This problem is commented on by all authors mentioned above.[15] Eid (1988:56), for instance, ignores all observations that include words that are difficult to categorize. These and similar studies therefore focus on features in which there is a binary choice between a dialectal and a Standard Arabic variant, such as forms of negation or relative pronouns. Case endings, the object of the present investigation, is not a variable of this kind. The choice is between marking case, a Standard Arabic feature, and not marking case. The absence of case marking does not render a word dialectal and the two can therefore not be analyzed as oppositions in the way required by the codeswitching paradigm.

Examples (1) and (2) illustrate this point. They are examples of switches from marked to unmarked case in the same words without any sign of codeswitching to the dialect. In (1), the adverb *laḥẓa* 'for a moment' occurs twice after the verb *tawaqqaf* 'stop'. In the first instance, it is marked for case and in the second it is not. In (2), the topic *sūrīyā* 'Syria' twice takes the comment *maḥkūma* 'doomed' albeit with different modifiers. Only the first is marked for case. The repeated word without case ending in these examples does not occur in typical pause positions and there is no pause here as pronounced by the speakers. The changes from marked to unmarked case can thus not be be explained by the pause phenomenon.

(1) *naḥna lam natawaqqaf* ***laḥẓat-an*** *lā ʿan il-ʿamal…*
we NEG.PAST 1PL.stop moment-ACC NEG from DEF-work…
bal ʾana ʿala l-ʾaqall lam ʾatawaqqaf ***laḥẓa*** *ʿan al-ʿamal*
rather I on DEF-least NEG 1S.stop moment from DEF-work
'We have not stopped working for one moment, not … Or I at least have not stopped working for one moment.' (Ġalyūn, 19:14)

(2) *sūrīyā bi-l-ʾasās-i* ***maḥkūmat-un*** *bi-strātījiyyat-i-hā*
Syria in-DEF-foundation-GEN doomed-NOM by-strategy-GEN-her
bi-bunyat-i-hā ʾal- ʾal-ṭabīʿiyya ***maḥkūma*** *bi-ʾanna-hā*
by-structure-GEN-her DEF- DEF-natural doomed by-COMP-she
'Syria is fundamentally doom by her strategy, by her natural structure, sentenced by being …' (Tayzīnī, 24:01)

[15] This is less of a problem where there are clear cut inter-sentential switches such as in the speeches analyzed in Holes (1993). Clear switches seem however to be a characteristic primarily of oratory styles as opposed to conversational styles. In the corpus constructed for this study,

Examples similar to those above are ubiquitous in the corpus. They show the inability of codeswitching models to explain the variation in the use of case endings in Spoken Standard Arabic.

2.7 Summary

In this chapter, theoretical issues of Arabic as a case of diglossia were reviewed. Ferguson's original definition of the concept, while useful as an overall description of some aspects of the Arabic language community, is too general for detailed analysis of linguistic data. To be able to describe the more nuanced and less clear-cut interplay between the dialects and Standard Arabic, various models have been developed, posing intermediate levels between 'pure' dialect and Standard Arabic. The models of Blanc (1964), Badawī (1973), Mitchell and El-Hassan (various publications), and Meiseles (1980) were reviewed. These models were criticized from two perspectives. First, they do not make adequate distinctions between spoken and written language, placing them on the same scale and thus measuring speech as to how it compares with writing. Second, it was argued that these models do not take into account the variation found between speakers using a formal register.

Spoken Standard Arabic is not a highly conventionalized variety. There is no group of native speakers whose speech might constitute a norm, and speakers vary in the forms of Standard Arabic they are exposed to and with forms they regard as appropriate for different situation. All this means that it is very much up to the individual to develop forms of Standard Arabic suitable to the demands of extempore speech. The approach adopted here is therefore to study case marking in the formal register of Arabic as the result of a process of transformation from word forms that are unmarked, to word forms that are marked for case. The aim is then to see where these transformation are applied and how they converge and differ between individuals.

Lastly, it was argued that codeswitching is not useful framework for understanding the use of case endings, since case endings do not form a binary distinction between dialectal and Standard Arabic. Words that are not marked for case are not in and of themselves dialectal. Rather, variation in case marking is a variation within the variety Standard Arabic.

---

consisting of non-oratory modes of speech, there are very few clear-cut inter-sentential switches between Standard Arabic and the dialect. But see examples (101–103) on page 186, and (110) on page 207.

Arabic as a standard language

3

The previous chapter discussed Arabic as a case of diglossia, a language situation where the language used in everyday conversation differs considerably from the language of formal conversation and writing. This perspective sheds light on certain important features of the Arabic language situation that sets it apart from many other major modern languages. In the light of these facts, Arabic in this chapter is instead discussed as a case of a standard language. This perspective focuses on similarities between Arabic and other modern languages, particularly in terms of language ideology and notions of linguistic correctness. This change of perspective is useful in order to apply insights from studies of other languages, particularly studies of English, to such issues in Arabic. Much of the discussion in this chapter, therefore, takes English as its point of reference, as a prototypical case of a standard language.

This chapter is organized as follows: section 3.1 discusses how a language may be seen as a structured cluster of different varieties one of which is elevated above the others to the status of a standard variety. This standard variety comes to represent the language in its entirety and to be the point of reference for linguistic correctness. Section 3.2 discusses the role of codified grammar, the collection of rules with the double function of describing and regulating the standard variety, and how these rules ought not to be confused with the actual usage of the standard variety. The specific historical circumstances of the codification of Arabic grammar is described in section 3.3. Section 3.4 discusses the ideology that lends legitimacy to and maintains the elevated status of the standard variety. In section 3.5, it is proposed that the upper end of the Arabic diglossic continuum should not be identified with what traditionally is considered correct usage, but with how competent speakers actually speak Standard Arabic. Section 3.6 presents a summary the chapter.

3.1 Standard languages and linguistic varieties

Several perspectives have been presented as to how to define and distinguish between the notion of a *language* in contrast to the notion of a *standard language*. Kloss (1967:30) has proposed the term *Ausbau language* for "languages which have deliberately been reshaped so as to become vehicles of variegated literary expression." This term is used in contrast with *Abstand langugae*, a language defined as such by its linguistic distance to other varieties. Haugen (1966:passim) has described the historical process by which one variety becomes a standard language as involving four phases. The first phase in this process is the *selection* of one of several related varieties to be elevated above the others. The second phase is the *elaboration* of this variety to be used for various purposes. The third phase is the *codification* of the selected variety so as to be preserved and taught. The fourth and final phase is the *acceptance* of this variety as the standard by the society at large. The goal of this process is for the standard variety to have "minimal variation in form [...] maximal variation in function" (ibid.:931). Ferguson (1997), on the other hand, views the process of linguistic standardization as a less agentive process, where standardization is reached through a combination of avoidance of saliently local forms, functional specialization of varieties, and classicalization, that is, the adoption of older and written forms of the language that are perceived as representing the ideal.

A fourth approach, which is particularly useful for the purposes of this study, is the normtheroetical approach in which a language is seen as a cluster of related linguistic varieties, superseded by a standard variety as the norm of linguistic correctness (Ammon 1986; Bartsch 1989).[16] Varieties that are not elevated to become standard will here be referred to as dialects or non-standard varieties.

Under the normtheoretical approach, 'language' is defined as a cluster of varieties in which

> the relationship between the standard and the [non-standard] varieties is such that they are superseded by the standard, i. e. that in a number of situations, especially in situations of schooling, the use of some linguistic forms from varieties is rejected under the prescriptive[*sic*] of the standard as a measure of correction. (Bartsch 1989:200)

This definitions rests not on formal linguistic criteria but on social and political constructions as 'belonging' to a language and thus ascribing to its shared notions of correctness. Bartsch (ibid.:200-201) mentions three effects of this

[16] The following discussion is inspired by and builds on Mejdell (2008a), where this approach to standard languages is applied to Arabic.

definition of language. Firstly, it draws linguistically arbitrary borders within dialect continua. Two very similar dialects on either side of a national border may adhere to the normative standard of their respective countries and thereby be part of different languages. Maltese, for example, is from a purely linguistic standpoint clearly closely related to Arabic. It is from a normtheoretic perspective not Arabic, since it has its own national standard variety. Speakers of Maltese do not see themselves as speaking Arabic, and Standard Arabic is not 'correct' Maltese. Secondly, in regions where there is no superseding standard variety, even closely related varieties are not of the same language, since there is no overarching normsystem. Thirdly, the status of varieties as dialects or standard changes through history as varieties loose their status as standard, and when new standards arise. Mejdell (2008a:42-3) gives the examples of Scots loosing its status as a standard language around year 1700 to be included in English, Norwegian breaking away from the standard of Danish and establishing its own standard in the 19th century, and the recognition of Luxenbourgish as an official language in Luxembourg, breaking off from German as its own language.

This definition of language is closely related to the distinction between 'language communities' and 'speech communities', originally formulated in Silverstein (1998:407). Blommaert describes language and speech community in that

> the former are groups professing adherence to the normatively constructed, ideologically articulated 'standard' Language ('we speak English') and the latter are groups characterizes by the actual use of specific speech forms.
> (Blommaert 2006:246)

Thus, one can, in this sense, speak of Arabic as one language encompassing all the dialects as well as Standard Arabic, and refer to all Arabic speakers as belonging to the same language community despite the often considerable linguistic distance between the dialects. Speakers of the dialects on the other hand form smaller units of speech communities.

For purposes of linguistic description it may be of little relevance to encompass the Arabic dialect under the larger and linguistically arbitrary umbrella concept of 'language'. It is however crucial to understand the language ideologies and attitudes pertinent to formal speech.

3.2 Norm and codification

For the standard variety to effectively function as a model for correct language, it must be codified so that its grammar is described in grammar books and its words listed in lexica. Such works of reference are indispensable for language

instruction and for when one is uncertain of the appropriate form. The codified grammar, essentially a collection of rules, must not be confused with how the standard variety is used in actual speech and writing. It is often the case that a specific grammatical construction is described in a certain way in grammars that are generally recognized as authoritative, and is thus recognized as correct, even though this description may not represent how educated, proficient speakers and writers use the standard variety. The sentences in (3) are examples of this in English with bracketed variants representing the prescriptive form favored in grammar books.

(3) a. Here's (*Here are*) John and Mary.

b. Nobody likes having their (*his*) teeth drilled.

c. Here we are, just you and me (*I*).

(adapted from Schmidt and McCreary 1977:118-24)

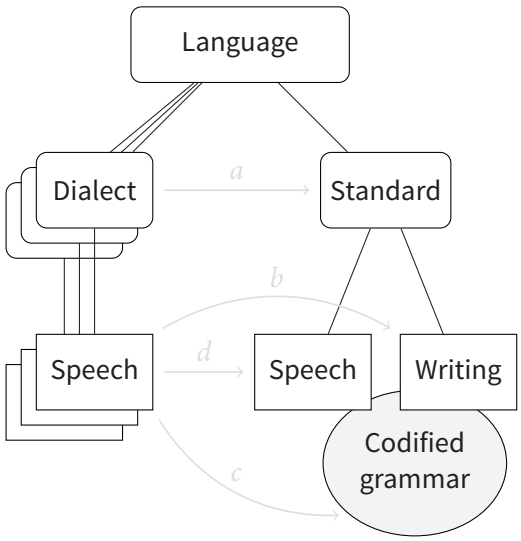
Such differences between the prescriptive forms and the forms representing actual use of the standard variety were a major theoretical concern for linguists of the Prague school.[17] The Prague linguists referred to the usage as 'norm', to distinguish it from prescriptive, codified grammar. This term is somewhat problematic in that codified grammar, its opposite in the theory, also embodies a set of norms, albeit of a different and more explicit kind. In the following, this concept will therefore be referred to as the *empirical norm*, as opposed to the *codified norm* of grammar writing (cf. Mejdell 2008a:47). In standard varieties, that are both codified and in active use, both these types of norms are in effect and are often to some degree in conflict. This is illustrated in Figure 3 and is further discussed below. (The gray arrows in the figure relate to the discussion in 3.5.)

The empirical norm differs from the codified norm in that it is observable in samples of the language in use and by the fact that it is subject to continuous linguistic change. In this sense, all varieties have a empirical norm, a system of linguistic conventions to which speakers relate. Dialects are not normally used for writing and have a norm only for the spoken language.[18] The standard vari-

[17] See Jedlička (1982) for an overview.

[18] When dialects are written, for example in text messages on mobile phones or to represent colloquial speech in novels, it follows *ad hoc* orthography representing speech rather than norms that have developed separately in non-standard writing. For Arabic, as with other languages, this form is undergoing rapid change due to the growing use of colloquial language in computer mediated communication. As this form of writing becomes more conventionalized there will be reason to describe the specific norms of the dialects in writing.

Figure 3: *Norm and codification in language*



ety of a language on the other hand typically has a long tradition of being both spoken and written, and has developed separate sets of norms for speech and writing. Structures accepted in spoken standard may be unacceptable in writing, and what is normal in writing may be awkward or high-flown in speech, even in very formal speech.

The codified norm is of a different ontological status than is the empirical norm. It is not observed in actual language use but in the metalingiustic descriptions in grammar books. It is an abstract set of formal rules, the result of an attempt to describe a perceived norm. As a description of language, the codification is historically dependent on the norm it aims to describe. While the empirical norm is in a state of continuous change, the codified norm remains fixed until actively changed by some authority. The extent and frequency to which the codified norm is updated depends on extralinguistic factors, such as how language authorities are organized and the ideals and ideologies by which they operate. The ideal situation according to the Prague linguists is one where the codification is continually reviewed through the scientific and systematic study of the empirical norm. The standard variety would thus reach a state of 'elastic stability', a situation where the codification reduces the rate of linguistic change but remains dynamically adaptable, not prescribing obsolete construc-

tions, nor seeking to suppress newer constructions that enter the standard variety (Havránek 1982:295).

In most accounts, a codified grammar is a particular and defining feature of a standard variety. While dialects may certainly be codified, in the sense that their structure is documented in grammars, these works do not have a prescriptive or normative force and are of no or marginal importance for the norm system of the language community. To correct someone for his or her use the *language* on the basis of grammar books is quite natural. The idea of correcting someone for his or her use of the local *dialect* on the basis of dialectologists' grammatical description is nothing short of comical. Hence the lack of codification for the dialects in Figure 3. In order to correct a speaker's speech in the dialect, one needs instead to make claims to other forms of authority, such as local traditions or the authenticity of local identities.

The codified norm of the standard variety is furthermore closely related to the written language, rather than to speech. Literary works held up as ideals of the standard variety are transmitted in writing, and the written word is, by its very nature, easier to observe, more readily providing source material for codification. Features characteristic of speech and not used in writing thereby often come to be considered as incorrect (Halliday 1994:97; Carter 1999:153–8). The contrasts between grammatical descriptions of Arabic case endings and the conventions that govern their use in speech as described in the following chapter is a prime example.

3.3 The codification of Arabic

In the European context, the norms that have formed the basis for the codification of the written standard variety have been that of the educated urban population in the early modern period, with the spoken form of the standard reaching a higher degree of standardization first in 19th century (Fisher 1996: 70). The codification of Arabic is very different in that it took place much earlier. With the emergence of Islam and the Muslim conquest in the seventh century, Arabic went from being a primarily oral language in the Arabian peninsula to the language of administration and learning of an empire. This changed role required standardization and codification of the language.

The first comprehensive Arabic grammatical work is Sībawayhi's (d. 798) *al-Kitāb*, the framework of which was built upon in the following two centuries until it reached its final stage of theoretical development in the tenth century. This grammatical codification was based on then already historical data: the Quran, pre-Islamic poetry, and sayings and proverbs (*amṯāl*, s. *maṯal*). In the

early stages of grammar writing this material was supplemented by data from Bedouin informants.[19] There is, in other words, a gap of 150 to 250 years between the codification and the empirical norm on which it was based (Ibrahim 1989:40). Indeed, linguistic examples (*šawāhid*, s. *šāhid*) from after the end of the Ummayad dynasty (150/750) were considered illegitimate as the basis of grammar writing. This was was early on adopted as a matter of methodological principle (Owens 1988:20; Bohas et al. 1990:18). Examples from after this period were deemed to be at too large a risk of being corrupted by foreign elements.

Grammarians of the 11th century onward were primarily concerned with systematizing this canonical grammatical description and refining the way concepts and rules are defined and formulated (Owens 1988:8). Short, concise, standardized formulations for any given grammatical rule were hammered out, and carefully constructed to account for known exceptions. This led to the development of a theoretically and terminologically highly homogeneous body of literature, characterized by "extreme coherence and systematicity" (Bohas et al. 1990:16). This period saw the production both of extremely condensed and expanded expositions with commentaries on previous, or, indeed, on ones own work, as well as grammars written entirely in verse. As Carter (1985:270) puts it, "the main preoccupation of grammarians after Ibn Bâbashâdh [d. 1077] was to find ever new ways of saying the same thing."

The system embodied in this later stage of theoretical development will be referred to in the following as Traditional Arabic Grammar, the canonical, formalized and prescriptive grammar still dominating formal language instruction in the Arabic speaking world (as well as in much Islamic education outside of the Arab speaking world). Traditional Arabic Grammar thus differs from the commonly used term 'Arabic Linguistic Tradition' as in, for example, Bohas et al. (1990), which may be taken to encompass the whole range of historical development including deviant theoretical interpretations. Traditional Arabic Grammar refers only to the endpoint of this theoretical development in its canonical formulation.

When Standard Arabic was revived during the Arabic renaissance, the *nahḍa* movement in the 19th century, after having been marginalized in the Ottoman empire as only to be used in religious domains (Chejne 1969:83; Newman 2013:473), it meant the revival of this very same grammatical tradition. There was a general agreement that wide-ranging lexical expansion was needed

[19] The *ḥadīṯ*, sayings of the Prophet are also normally included as sources but were, in practice, of little use as data for codification, since they were understood not to convey the exact original wording (Owens 1988:20; Suleiman 2011:9).

for Arabic to function as a modern standard language and neologisms and to lesser extent borrowings were encouraged (Stetkevych 2006:chap. 1–3; Holes 2004:44), but many of the syntactic changes that had taken place since the early codification[20] came to be seen as linguistic errors.

The present codified grammar of Arabic is thus based on a doubly archaic set of norms. The Arabic renaissance looked back on the classical codification that looked back at the early Islamic and the pre-Islamic eras. Some effects of the retention of the traditional grammar are quite striking for a non-Arab observer, especially with regards to education. Classical grammars from the 13th to 15th centuries are still used as standard textbooks in language instruction at the university level (Bohas et al. 1990:16).[21] The principle of relying only on historical data is also still applied in contemporary grammars and other prescriptive literature. An illustrative example of this is *Muᶜjam al-ʾaġlāṭ al-luġawiyya al-muᶜāṣira* [A Dictionary of Common Mistakes in Modern Written Arabic] (English title in original). It lists 2135 "mistakes" in modern Arabic as observed in modern media and claims to give the correct form based the Quran, *ḥadiṯ*, classical dictionaries (*ummahāt al-muᶜjamāt*), and poetry from the pre-Islamic to the Ummayad eras (al-ᶜAdnānī 1999:[page] ح). Another example is the main reference grammar and course book for subjects in Arabic grammar at Damascus University. In its introduction, the author states that he "has been careful — for the most part — [only] to give examples from the Quran — the foremost book of Arabic — , from the correct *ḥadiṯ* [sayings of the Prophet], and from the poetry of which the major grammar books are full" (Bayṭār 2007:1). It is, in other words, classical Arabic that is codified, while Modern Standard Arabic has yet to be described in Arab grammars and dictionaries (Al-Wer 1997:254; Newman 2013:480).

وقد حرصت ـ غالبا ـ على ضرب الأمثلة من القرآن الكريم ـ كتاب العربية الأول ـ وما صح من الحديث والشعر مما اتلأت به أمهات كتب النحو .

In such a situation, discrepancies between the contemporary empirical data and the codified norm, in analogy with the English examples in (3), are bound to be plentiful. Basing the codification, and thus language instruction, on norms now gone has the double effect of preserving archaic and obsolete features and banning features that have entered Standard Arabic since after the era of codification. Consequently,

> many hours of school time are spent on teaching archaic linguistic features which are mostly only encountered and exemplified through ancient texts in the classroom. Outside the school, the pupils are most unlikely to use or encounter examples of these grammatical forms. (Al-Wer 1997:255)

[20] See (Blau 1973, 1976) and (Newman 2013:480–2) for lists of such innovations.

[21] For example, the main course book on syntax (*naḥw*) at the Jordanian University in 2008 (when this author studied there) was Ibn Hišām al-Anṣārī's (d. 1359) *Šarḥ qaṭr an-nadā wa-ball aṣ-ṣadā* (Ibn Hišām 1990).

Conversely, students are corrected for using constructions that are now common in Standard Arabic texts. A situation where speakers learn that forms they often use and hear are incorrect is expected to lead to linguistic insecurity (Labov 1972:117). Linguistic insecurity has often been observed amongst Arabic speakers with regards to Standard Arabic (e.g. Alrabaa 1986:78; Haeri 2003:43; Maamouri 2006:77).

The problems of the mismatch between what is here called the empirical and the codified norm, and the possibility of language reform have of course not gone unnoticed in the Arabic speaking world. The issue of language reform was widely discussed in the first half the 20th century with the most radical suggestions involving the adoption of the dialect as a standard variety. Furayḥa (1955) and Mūsā (1964) are example of prolific proponents of this view. Such views were tightly connected with socialist ideals and fell out of favor with the rise of pan-Arabism in the 1950s that promoted the unity that the common language of Standard Arabic provides (Haeri 2003:135). Mejdell (2008b:116), in a recent study in Egypt, a country well known for taking pride in its dialect, found such suggestions now to be very rare in the contemporary debate.

Less drastic language reforms have been suggested, such as getting rid of the complicated[22] system of case endings, or to simplify some of the difficult inflectional classes. Such suggestions have been rejected after conservative opposition (Altoma 1974:295).[23] A third option that has been discussed is to reform how Arabic grammar is described without changing the standard language as such. This option has been discussed in the language academies without any significant results (Diem 1974:137; Abu-Absi 1986:339). According to Suleiman (1996:114-5), the failure of such proposals are due to the "naïve-realist" epistemological stance of the linguistic tradition that does not differentiate between the description of the language and the language itself. A change in grammatical theory and description would thus imply a change in the language that is ideologically unacceptable.

It should be pointed out that such mismatches between codification and linguistic practice are by no means unique to Arabic. Curzan (2014:69) lists a number of now questionable prescriptive rules for English that originated from a period of intense English grammar writing in the early 18th century that was often modeled on Latin grammar and that are suggested by the grammar check feature in widely used word processing software. This include rules concern-

[22] As described in Traditional Arabic Grammar (see 4.2.)

[23] One Arab linguist describes being "accused of treason for pointing out the need to re-standardize Arabic in order to incorporate the linguistic changes of the past twelve centuries. […] Worst of all, allegations such as these are frequently made by patrons of institutions of education and language academies" (Al-Wer 1997:125).

ing split infinitives, stranded prepositions, double negatives, and generic *they* (rather than the prescribed *he*, as in (3a) above). Thus what is claimed here is not that Arabic is qualitatively different from other languages in its discrepancies between the empirical and the codified norm. Rather, what is claimed is that in the Arabic situation, with the vast timespan since the completion of the canonical codification and the ideological opposition to revising it, the problem is particularly acute, and its consequences more severe.

3.4 Standard language ideology and prescriptivism

The standard variety of a language is one of several varieties that for historical, social, and political reasons has been elevated to the status of a standard. With this develops an ideology, in Arabic as in other language communities, that justifies and lends legitimacy to the standard variety and serves to preserve its status as the arbiter of correctness and as the representative of the language as a whole.

In their seminal study, Milroy and Milroy (1991:26) describe standard language ideology as being characterized by "intolerance of optional variability in language". Of two grammatically different ways of expressing the same, as in (3) above, only one is generally considered correct. One may debate which of two alternatives is correct, but that there would be two equally valid forms is seen as problematic. The standard language ideology is thus expressed in prescriptivist attempts to alter language use to minimize variability. Critical studies of this ideology often point to the impossibility of actually realizing this goal of homogeneity, thus regarding the mere existence of standard varieties, varieties in which this goal has been achieved, as a myth (Ricento 2006:20; Lippi-Green 1997:62). Milroy and Milroy (1991:22–3) similarly see standardization of language as a perpetual process of attempts to limit the intrinsic variability of human language, pointing to the function of variant forms as various forms of social markers, an aspect typically ignored in standardizing efforts (ibid.: 69).

Prescriptivist efforts to alter the use of language have often been dismissed by linguists as naive attempts to affect uncontrollable processes of linguistic change. This view has recently been challenged by Curzan (2014:chap. 2) who argues that prescriptivism, even though its effect on linguistic change may be small, should be seriously studied for its effect on linguistic attitudes, metalinguistic discourses, and linguistic insecurity. Curzan identifies four strands of prescriptivism, of which primarily the first two are of relevance for the present purposes:

- Standardizing prescriptivism: rules/judgments that aim to promote and enforce standardization and "standard" usage.
- Stylistic prescriptivism: rules/judgments that aim to differentiate among (often fine) points of style within standard usage.
- Restorative prescriptivism: rules/judgments that aim to restore earlier, but now relatively obsolete, usage and/or turn to older forms to purify usage.
- Politically responsive prescriptivism: rules/judgments that aim to promote inclusive, nondiscriminatory, politically correct, and/or politically expedient usage. (Curzan 2014:24)

Standardizing prescriptivism seeks to limit variation in the standard variety that is the result of its coexistence with non-standard varieties. It can be thought of as an attempt to keep apart the left and right hand side in Figure 3, or perhaps even to eradicate the left hand, non-standard side altogether. This is typically done under the assumption that the standard variety is the 'actual' language, and other varieties are understood as deviating from it (Milroy 2007: 136; Bailey 1991:4). If the standard variety is the actual language, not having learned it properly is not to have learned your own language and is a grave personal failure that provokes moral judgments (Trudgill 1975:27; Milroy and Milroy 1991:40; Bailey 1991:237). The fact that the standard variety is studied in the form of codified grammar, a coherent set of rules, enforces this perception. Non-standard varieties are thus perceived as illogical and unstructured since they are not presented as governed by rules and are often in conflict with the grammatical codification (Trudgill 1975:40).

There is much anecdotal evidence of the widespread view in the Arabic language community that the dialects have no grammatical rules and are illogical and unstructured. Typical descriptions of this view in the literature are Ibrahim (1983:512): "spoken Arabic is considered to be unworthy of the label 'language' but rather a nameless verbal corruption" and also Beeston (1970:81): "Arabs are often unwilling to recognize that the dialects have any grammar, but see them as chaotic deviations from the sacred norm." This strongly negative stance towards the dialects is dominant also amongst Arab linguists and policy makers. In a special issue of the Egyptian Language Academy's publication on the topic of diglossia, Šawqiy Ḍayf, former president of the Academy, writes that

> [the dialect] in all of its expressions is corrupted *fuṣḥā*. The dialect, therefore, needs to be seen to by the language academies in the Arab countries, by studying the corruptions that have entered it and have it return to the *fuṣḥā* in which it originates. (Ḍayf 2000:49)

[العامية] في جمهور ألفاظها فصحى محرفة، والعامية - لذلك - في حاجة إلى أن تعتني المجامع اللغوية في البلاد العربية بدراسة ما داخل ألفاظها من تحريفات وردها إلى أصولها الفصيحة.

Another author in the same issue speaks of the dialects as being the result of

ignorance (*jahl*) and backwardness (*taxalluf*) (Xalīfa 2000:123).

Uhlmann (2012:103) has described how the stakes for Arabic language instruction are particularly high. Not only is dialectal Arabic seen as a symptom of societal backwardness and of individual failure, but it is also seen as a concession to Western cultural dominance. The preservation of Standard Arabic is constructed as a struggle of national survival and independence thus giving standardizing prescriptivism in Arabic a strong sense of moral urgency.[24]

The second form of prescriptivism, stylistic prescriptivism, involves distinguishing between alternative forms *within* the standard variety. Curzan (2014: 33) describes it as "a nicety that distinguishes those who 'know better' from those who don't, but it does not distinguish standard English speakers from nonstandard English speakers." The use case endings in Standard Arabic (further discussed in the following chapter) falls into the category of stylistic prescriptivism. Case endings have a strong symbolic value as markers of Standard Arabic and are therefore commonly regarded as being a point of delimitation between the dialects and Standard Arabic. However, the non-use of case endings does not render a word, or a string of speech, as dialectal. As will be shown in the following chapters, variation in the use of case endings is quite clearly a stylistic feature within the standard variety.

Stylistic prescriptivism can be overdone. In situations where codified grammar deviates from the empirical norms, over-adherence to codified grammar may result in flaunting these norms, resulting in speech that is perceived as pretentious, pedantic, or overly formal (cf. ibid.:35). Speaking in compliance with codified grammar thus becomes a highly marked way of speaking. Some scholars have described this as a *super-standard* variety, a variety characterized by over-application of prescriptive forms. Schmidt and McCreary (1977: 415) classifies grammatical constructions as standard if they are "accepted and used by elite groups", and as super-standard if they are "either not followed by speakers or restricted to only the most formal level of style." They exemplify this with the bracketed variants in (3) above. Similarly, according to Bucholtz (2001:88), the super-standard can be contrasted with the standard variety "in its greater use of 'supercorrect' linguistic variables: lexical formality, carefully articulated phonological forms, and prescriptively standard grammar." Speakers producing speech of the super-standard variety can be understood as striving to comply with the codified norm, whereas speakers producing 'normal' standard can be understood as complying with the empirical norm.

[24] For an example of this discourse in action, see the documentary *Lisān aḍ-ḍād yajmaᶜunā* 'The language of *ḍād* unites us' produces by Al Jazeera and available at: http://www.aljazeera.net/programs/infocus/2013/12/4/لسان-الضاد-يجمعنا.

In Arabic, complete usage in speech of morphological case endings in accordance with prescriptive grammar is associated with formal read speech, and its use in extempore speech would result in form of a super-standard variety. It would deviate from the empirical norm of Spoken Standard Arabic, for example the norm not to mark case on words with a definite article (see 9.2). The speaker described in Parkinson (1994a:129, see 4.8.4) who uses case endings with a consistency that is "nothing short of breathtaking" is an example of over-use of case endings in super-standard Arabic, and the annoyance of the interviewer an example of social consequences of deviation from the norm.

3.5 Codified grammar and diglossic variation

The differentiation between the empirical norm and the codified norm in the standard variety has important implications for the study of diglossic variation in Arabic. The major theories and trends in this line of research, reviewed in Chapter 2, have seen this variation as a movement between the two poles of dialectal and Standard Arabic. Under the most simple interpretation, this is a movement between two varieties, illustrated with arrow *a* in Figure 3 (page 47). This conceptualization of diglossia fails to account for the differences between Standard Arabic as extemporaneous speech and formal read speech, or even writing. As was discussed in the previous chapter, the upper end of the diglossic continuum has often explicitly been identified with written Standard Arabic without accounting for the problems of having a continuous scale of variation across the spoken/written divide (arrow *b*). What is more, the primary source of information on the structure of Standard Arabic is the codified grammar (partly due to the lack of other descriptions of Spoken Standard Arabic). This has the effect that in practice the variation often comes to be analyzed as a movement between the dialects on the one hand and the codified norm on the other (arrow *c*). Examples of speech are then placed on a scale between an actual linguistic variety (the dialect) and a description of a variety (codified grammar) on the other. This may not have been a problem were it not for the fact that this description is to a large extent outdated and formulated with an explicitly prescriptive aim. By taking the traditional codified norm as the ultimate frame of reference, linguists often implicitly ascribe to a normative prescriptivist view of the Arabic language. Since Spoken Standard Arabic in many respects deviates from codified grammar as well as from the norms of written language, little, if any, observed speech is classified as Standard Arabic. The effect on analysis is that Standard Arabic as used in speech becomes an idealized target form, a theoretical construct that is never actually realized. This has also been formulated

as a methodological principle by some scholars (Badawī 1973:119; Hary 1996: 72). Hence the hesitation among some Arabists to classify linguistic extracts as Standard Arabic, preferring instead to describe them as "attempt[s] to speak classical Arabic" (Meiseles 1980:122) or as "approaching" or being "more or less" Standard Arabic (Versteegh 2001:194).

If linguistic variation in Arabic is described as levels or as a continuum where one pole is the speaker's dialect, then the other pole ought not to be an abstract set of definitive grammatical rules, nor the language of the classical past, but rather the way Standard Arabic spoken today. An alternative approach is to identify the upper end of the diglossic continuum not with the codified norm, but rather with the empirical norm of spoken Standard Arabic (arrow *d* in Figure 3). Under this approach, speaking Standard Arabic does not mean speaking in way that is traditionally regarded as 'correct'. Rather, it means speaking in accordance with the covert conventions of what is here called Spoken Standard Arabic, conventions that only partly overlap with the overt conventions of the codified norm.

The need then arises to identify what the norms are that govern Spoken Standard Arabic. This is an empirical question and can only be answered by observing speakers' oral use of this variety. This, in turn, can only be adequately done under the assumption that competent speakers produce a variety in situations where they are expected to speak Standard Arabic that is, indeed, the spoken form of Standard Arabic. The methodological implications of this approach — that are by no means trivial — are discussed in Chapter 5.

It is well known that competent speakers speaking in formal situations and thereby, by definition, are producing Spoken Standard Arabic, use Standard Arabic features at varying rates, forming a pattern of distribution of different proportional uses of a given feature or of a set of features. These rates can be mapped out and statistically described. The use of morphological case endings on nominals is a case in point. The corpus examined in this study was constructed to capture the most formal style of experienced speakers of Standard Arabic (see 5.2). It can be shown that these speakers mark case on average 7% of the time (see 8.1), with rates above 10% becoming increasingly rare. This is far below the prescriptive ideal of full case marking. Speech with complete or near complete use of case marking is thus not representative of Spoken Standard Arabic. It is 'better' or 'more' standard only from the perspective traditional prescriptivism. Indeed, 'overshooting' Standard Arabic features with respect to a given social situation, producing a 'super-standard Arabic', probably comes at a social cost in its deviation from normalcy (Parkinson 1991:59, 1994a:209).

Another example of a binary diglossic feature that may illustrate this point

is the use of the relative pronoun. The dialectal form of relative pronoun is *illi* in the major dialects (with minor variation). The Standard Arabic form is *allaḍī*, with variant forms marking number and gender. Speakers in the corpus of the present study use the Standard Arabic variant of the relative pronoun on average 77% of the time and otherwise the dialectal form (see Table 8.3 on page 176). This is normally interpreted as speakers mixing two varieties, namely their dialect and Standard Arabic. The problem with such an interpretation is that it is based on a presumption as to what (spoken) Standard Arabic is, a presumption informed by traditional grammatical codification where *illi* has no place. In observed use of Spoken Standard Arabic, some use of *illi*, as well as other dialectal features, is in fact the norm.[25] An alternative approach is to regard the use and non-use of case endings, the use of the dialectal and Standard Arabic form of the relative pronoun, and other diglossic variables as variation that is integral to Spoken Standard Arabic without resorting to an explanation of language mixing.

In informal situations, the usage level of markedly Standard Arabic features is expected to be significantly lower. Schulz (1981) and Mejdell (2006) report data on the use of the dialectal and Standard Arabic variants of the relative pronoun in situations of lower formality that that investigated here. Both give average usage usage levels of the Standard Arabic variant that are lower than the 77% reported here.[26] The data available for case marking is not directly comparable (see 4.8).

From the definition of Spoken Standard Arabic as the form of speech that speakers produce in formal situations, then, it follows that such speech includes a span of variation in the use of diglossic features, of which the production of case endings is a prime example.[27] This has two important implications for the study of Arabic diglossic variation. First, Standard Arabic ceases to be a mere theoretical concept and an, in reality, unreachable endpoint. Rather, it becomes an actual variety that can be empirically investigated and compared with other forms of Arabic. Secondly, even though it is certainly possible to trace much of this variation to the dialectal substrata, this does not necessarily mean that

[25] Cf. Mazraani (1997:195) who found that the dialectal form of the relative pronoun is "all pervasive" in speeches by three Arab heads of state.

[26] Schulz (1981:152) reports an average of usage of 47% of the standard variant of the relative pronoun, but comments that most speakers use one of the two more or less consistently. Mejdell (2006:375) reports numbers recalculated to an average of 42% usage of the standard variant. Schulz (1981:12) collected data from a variety of radio programs and Mejdell (2006:68) from an academic panel discussion, both in Cairo.

[27] The wide span of variation in Standard Arabic performances famously led Kaye (1972) to describe Standard Arabic as an "ill-defined" linguistic system. See Parkinson (1990:287-290) for a critique of this notion.

these features need to be analyzed as standard-dialect mixing. Such features could also be analyzed as variation within one variety. *Illi*, the dialectal form of the relative pronoun, can, for example, be argued to be an integral part of variation in Spoken Standard Arabic, comparable with the variation between the two prescriptively sanctioned forms *allātī* and *allawātī* for the relative pronoun in feminine plural.

One argument against this line of reasoning is that native speakers have clear and consistent notions about which variants represent Standard Arabic and which variants represent the dialect. Clearly, the vast majority of native speakers would, for example, agree that *illi* is a dialectal form and that *allaḍī* is a standard form. The argument could therefore be made that the use of *illi* in Standard Arabic contexts is indeed an intrusion from or mixing with the dialect. There are two problems with this argument. The first is that it is extremely difficult to elicit linguistic intuitions for forms perceived as non-standard (Labov 1972:213). Statements to the effect that certain features are standard typically reflect prescriptive notions and codified grammar as received through formal education rather than reflecting actual linguistic behavior or attitudes. Directly elicited native speaker intuitions on standardness thus tell us little more than the codified grammar from which such judgments are typically derived. The second problem with this argument is that such binary classifications of features as standard only relate to the structural distributions of such features on the formality scale so that the use of features classified as standard will increase as formality increases, and *vice versa*. But classifying a feature as standard or non-standard only tells us that it is likely to be used more in formal speech than in informal speech, but not how much more or at what point it becomes high-flown or super-standard.

Studying Arabic diglossic variation as a spectrum whose upper end is defined by what speakers actually do rather than what they ought to do according some predetermined set of rules, when combined with carefully controlling for situational factors, may prove to be a fruitful approach to make sense of some of the often bewildering variation in the use of Standard Arabic features observed in formal speech.

3.6 Summary

This chapter discussed the Arabic language as a cluster of linguistic varieties. As in other languages, one of these varieties is perceived as standard and the others come to be regarded as dialectal varieties of inferior status. The standard variety is seen as representing the correct form for the entire cluster of

varieties. The standard variety is furthermore internally complex, consisting of both a written and spoken form, each with its own set of linguistic norms. The standard variety has also been codified and this has been done generally on the basis of written language. It is the codified grammar that represents 'correct' or 'proper' language. In Arabic, the codification was developed in the Middle Ages and thus does not reflect changes in Standard Arabic that have occurred since then. This has led to a wide gap between Standard Arabic as it is codified and taught, and Standard Arabic as it is used.Linguistic prescriptivism, views, and judgments of how people ought to speak and write take the codification as measure of correct usage, even where it does not correspond to actual usage of the standard variety. These three sets of norms, the empirical norm of speech and writing respectively, and the codified norm, need be conceptually kept apart in how they relate to the variation observed in speech in the standard variety. Finally, an approach for studying Spoken Standard Arabic was proposed in which variation in the use of Standard Arabic features is seen as an integral part of this variety, rather than as mixing with the dialect or as failure to reach the target of prescriptive correctness.

Case in theory, tradition, and practice

4

The previous two chapters have discussed Arabic through the perspective of diglossia in Chapter 2 and as an instance of a standard language in Chapter 3. The present chapter is concerned with a particular morphological feature, namely the use of case endings. The chapter aims to map out linguistic and extra-linguistic factors and circumstances that have a bearing on how the case system is used in speech in order to provide a backdrop against which the results presented in Part III can be interpreted. As such, this chapter is partly written with the non-specialist in mind. It aims to enumerate and make explicit aspects of learning and using Arabic that are often taken for granted by Arabic speakers and that are therefore often not subject to conscious reflection by users of the language. Some of these aspects have important methodological implications for the study of the use of morphological case endings in speech, as well as for teaching and learning Arabic.

Case endings are first encountered by the native speaker of Arabic in formal education. It has no direct parallel in the grammar of the natively spoken dialects. This fact together with the strict and traditional methods by which Arabic grammar is taught and the variation of case marking in the linguistic input — topics dealt with later in this chapter — gives a situation where the individual perceives case marking as a particularly difficult and complicated aspect of Standard Arabic grammar. Case endings can thus be considered to be at the center of the difference between the empirical and the codified norm in Standard Arabic, discussed in the previous chapter. This mismatch between the codified and idealized form of the language and general linguistic practices is often interpreted as a lack of linguistic competence of speakers, since codified grammar is the standard against which competence is measured. The low level

of proficiency in the use of case endings is a topic of lament amongst Arab traditionalists and has been so since the early classical period (Ayoub 2006: passim). Western scholars have likewise often noted the poor command of case and mood endings even amongst the highly educated (Kaye 1972:43; Alrabaa 1986:76; Ibrahim 1989:42; Parkinson 1993:59). Indeed, Beeston (1970:53) clearly overestimates the general level of proficiency when stating that "at the present, only a small sector of the highest educated Arabic-speaking population is capable of using the [case] system with complete confidence in extempore, unprepared, diction." The view in the native Arabic linguistic tradition that case is an essential part of the Arabic language means that low proficiency in the use of the case system is seen as deeply problematic: if you do not master case you do not master Arabic. This perceived problem is something all speakers of Standard Arabic must relate to one way or another.

The widely held view of both native and non-native speakers of Arabic is that the case system is highly complicated and therefore difficult, and this notion is a recurring theme in this chapter. The idea of the difficulty of the case system is echoed throughout the linguistic literature, where it is often described as 'complex' or 'complicated' (e.g. Meiseles 1977:175; Haeri 2003:40; Saiegh-Haddad and Spolsky 2014:232). In this chapter, it is argued that the perception of the Arabic case system as complicated and difficult should be understood in the context of extralinguistic circumstances: educational practices and strong prescriptive notions of correctness, rather than as something pertaining to the grammar itself. The notion of the difficulty of the case system is a consequence of the discrepancy between common descriptions of the language and actual linguistic practices.

This chapter begins in section 4.1 with a brief linguistic description of the case system in Standard Arabic. This is followed by a discussion in 4.2 of how case is described in Traditional Arabic Grammar, the system by which the case system is taught in first language teaching[28] in the Arab world. The particulars of this method of teaching are the subject of section 4.3. The importance of case endings in Traditional Arabic Grammar and in language instruction stands in contrast with the *de facto* syntactic redundancy of the case system in Standard Arabic, a topic discussed in section 4.4. Section 4.5 describes the system of final vowel omission in pause (*waqf*) which often results in the omission of case endings. The rest of the chapter focuses on the linguistic practices of case marking in different modalities and genres. Section 4.6 describes how case marking is used in writing, where only certain forms of case endings are actually present,

[28] By 'first language teaching' is here meant the teaching of Arabic grammar and language to speakers of Arabic rather than to learners of Arabic as a second language.

since short vowels are not orthographically marked. Section 4.7 discusses conventions of case marking in reading aloud and recitation, and section 4.8 is a review of previous research on case marking in the formal registers of spoken Arabic. Section 4.9 is a summary of the chapter.

4.1 The case system in Standard Arabic

Syntactically, the Standard Arabic case system is straightforward, albeit with a few quirks.[29] There are three grammatical cases, referred to in Western traditional terminology as nominative, genitive and accusative.[30] Typologically, the Arabic case system is of the accusative type with the main distinction being between subject and object. This is the most common type of case system in the world's languages, represented for example by the case systems in the European languages (Blake 1994:119). It differs, for example, from ergative case systems where subjects of intransitive verbs take the same case as objects of transitive verbs.

The nominative in Arabic thus marks the subject of the clause. It is also used to mark the complement in equational clauses without a verbal copula (as in (4) below). The genitive marks all prepositional complements and adnominals in annexation (*iḍāfa*), semantically denoting possession or some other form relationship between entities. The accusative marks direct and indirect objects as well as other forms of verbal complements, including complements of explicit verbal copulas, known as '*kāna* and her sisters'. The accusative also marks constituents outside of the argument structure that are not covered by the genitive, such as adverbials, exclamations, and vocatives. The functions of the three cases in Standard Arabic can thus be described as follows:

NOMINATIVE	subjects; complements of null-copula
GENITIVE	prepositional complements; adnominals
ACCUSATIVE	everything else

This description captures the case distribution on the vast majority of nominals in any natural text. There are some situations that do not fit this description, however. One of the most common is the complement of a set of complementizers and conjunctions, known as '*inna* and her sisters', that

[29] The system is only briefly summarized here. See 7.6 for a detailed description of Standard Arabic case governance and how it is applied in the annotation of the corpus.

[30] But in Badawi et al. (2004) 'independent', 'dependent', and 'oblique'.

govern the subject of the following clause and assign accusative instead of the expected nominative case. Furthermore, cardinal numbers govern the case of the counted noun in peculiar and unintuitive ways. By and large, however, case governance in Standard Arabic is very basic as compared to case systems in many other languages.

Morphologically, case marking in Standard Arabic is more intricate however, with eight declensional classes (as listed in Badawi et al. 2004:50–53) exhibiting various patterns of syncretism. The case marking suffix interacts morphologically with number and definiteness in different ways in the paradigms which results in a large set of different endings. In (4), for example, the words *kalima* 'word' and *marfūᶜ* 'nominative' are both in the canonical triptote paradigm and in the nominative. Both take the ending *-u*, but only the second nominal is indefinite and therefore also takes a final *n*, the so-called nunation or *tanwīn*.

(4) *hāḏihi* *l-kalimat-u* *marfūᶜat-un*
this DEF-word-NOM nominative-NOM
'This word is [in the] nominative.'

It should be noted, however, that for most practical purposes, the morphological complexity of the case system as it is traditionally described is drastically reduced. In writing, only a small subset of these endings are actually represented, as detailed in 4.6, and it will be shown in Chapter 9 that the morphological complexity of the case system is reduced in speech since speakers tend not to mark case on words in difficult and infrequent paradigms.

As mentioned above, none of the modern Arabic dialects have a corresponding system of morphologically marked case. Case endings only appear in the dialects in formulaic expressions and on lexicalized forms borrowed from the standard language. The formulaic expressions with case endings are typically of religious text origin. Two well known examples are (5) and (6), both with a retained overt nominative ending *u* as in the word *al-ḥamd* 'the praise' in (5) and in *allāhu* 'God' in (6). Lexicalized forms with case endings are predominantly adverbs with the accusative ending *-an* such as *maṯalan* 'for example' and *dāʾiman* 'always'.

(5) *al-ḥamd-u* *li-llāh*
DEF-praise-NOM for-DEF-God
'God be praised' (al-Xuḍarī, 8:37)

(6) *ṣallā llāh-u ʿalay-hi wa-sallam*
pray DEF-God-NOM over-him and-preserve
'May God pray over him and preserve [him]'[31] (Badīʿ, 20:22)

4.2 *IʿRĀB*

In the theoretical framework of Traditional Arabic Grammar developed by the classical Arab grammarians, there is no direct terminological parallel for 'case'. The notion of nominals taking different forms depending on syntactic position or grammatical function is included in the wider concept of *iʿrāb*.[32] In the widely used pedagogical grammar *al-Ājurrūmyya* by Ibn Ājurrūm (d. 1327), it is, for example, stated that "*iʿrāb* is the change of word-endings due to the variation of operators which occur before them either explicitly or implicitly" (original and translation in aš-Širbīnī 1981:34). The term '*iʿrāb*' thus refers to the grammatical system as such. It is, however, also used to refer to the endings that are effected by the system (ibid.:37n). The technical term for the endings is otherwise *ʿalāmāt al-iʿrāb* 'markers of *iʿrāb*' or *ḥarakāt iʿrābiyya* '*iʿrāb* vowels'. In non-specialist contexts the endings are usually referred to simply as *taškīl* 'vowel signs' or *ḥarakāt* 'vowels'.

الاعراب تغيير اواخر الكلم لاختلاف العوامل الداخلة عليها لفظا او تقديرا

Iʿrāb is a unified framework for what in the Western grammatical tradition is regarded as two separate systems: case and mood.[33] A syntactic constituent, whether a verb or nominal, is thought of as occupying one of the four positions of *rafʿ*, *naṣb*, *jarr*,[34] or *jazm*. These correspond to cases or moods depending on whether the position is filled with a verbal or nominal constituent, as summarized in Table 3. *Rafʿ* corresponding either to 'nominative' or 'indicative' and *naṣb* either to 'accusative' or 'subjunctive'. Only nominals and not verbs can be in *jarr* 'genitive', and only verbs and not nominals can be in *jazm* 'jussive'. This unified terminology for verbal and nominal constituents reflects the formal overlap of case and mood markers in nominative/indicative with *-u*, and in accusative/subjunctive with *-a*.

The abstract notion of an *iʿrāb* position is referred to as *maḥall* (lit. 'position') or *ḥāl* (lit. 'situation', not to be confused with the specific accusative position of the circumstantial qualifier). A subject is, for example, said to be

[31] Said when mentioning the Prophet. Often translated as 'Peace be upon Him.'

[32] See Owens (1988) and Bohas et al. (1990) for more detailed descriptions of the *iʿrāb* system.

[33] Although in the Western Arabist literature, '*iʿrāb*' is often equated with case endings exclusively (e.g. El-Hassan 1977:120; Haeri 2003:40).

[34] Also referred to as *xafḍ*, e.g., in *al-Ājurrūmiyya* (aš-Širbīnī 1981:38).

Table 3: *Case and mood v.s.* iʿrāb

	RAF^{C}	*NAṢB*	*JARR*	*JAZM*
ENDING	*-u*	*-a*	*-i*	*-Ø*
CASE	nominative	accusative	genitive	—
MOOD	indicative	subjunctive	—	jussive

in *maḥall rafʿ* 'position of the nominative'. Alternatively, a constituent is described with the passive participle of the position name as *marfūʿ*, *manṣūb*, *majrūr*, or *majzūm* (lit. 'nominativized', etc). Particles and some other types of words that fall outside of the system are described with the phrase *lā maḥalla lahu min al-iʿrāb* 'it has no *iʿrāb* position'. Words whose ending varies according to syntactic position are *muʿrab* 'inflected' (passive participle of the verbal noun *iʿrāb*). Verbs in the imperfective conjugation and most nominals belong to this category. Words whose form does not change with syntactic position are *mabnī* 'invariable'. These include some forms of nominals and verbs in the perfective conjugation.

The centrality of *iʿrāb* in Traditional Arabic Grammar can hardly be overestimated. With the erosion of the case and mood systems in the spoken language *iʿrāb* became a mark of distinction of the elites.[35] This had important implications for the development of the Arabic linguistic sciences which came to focus on points of difference between the classical language and the dialects. *Iʿrāb* came to be seen as the most important such difference, and thus as a sign of eloquence and linguistic purity (Bohas et al. 1990:49–50). Thus, according to Versteegh,

> almost all Arabic accounts of the origin of Arabic grammar concentrate on the wrong use of the declensional endings, as if there were no other types of mistakes against the rules of the Classical standard. (Versteegh 1983:157)

In one such story, Abū l-Aswad ad-Duʾalī (d. 688), according to tradition the founder of the Arabic linguistic sciences, told Ziyād ibn Abīh, the governor in Basra, that the language of the Arabs had been corrupted due to their contacts with non-Arabs, and asked for permission to lay down the rules of correct speech. The Amir declined. Later a man came to court to complain to the

[35] It is disputed whether this loss of case endings in speech happened during or after the rise of Islam in the Arabic peninsula. See Versteegh (2001:46–51) for a summary of the debate.

Emir, saying "May God make the Amir prosper! Our father[acc.] passed away and left sons[nom.]" (az-Zubaydī 1954:14), with the subject in the accusative and the object in the nominative. The Emir was so upset by hearing this error that he summoned Abū l-Aswad and ordered him to do what he had previously forbidden him to do.

أصلح الله الأمير! توُفّى أبانا وترك بنون.

Various early Muslim authorities are attributed with statements emphasizing the importance of *iʿrāb*. The jurist Mālik Ibn Anas (d. 795), for example, is reported to have said that that "ʾiʿrāb is the jewelry of our tongue, so do not deprive your tongue of its jewelry" (ibid.:4, translated in Dévényi 2006:401). *Iʿrāb* is also a central theme in the *laḥn* literature, both classical and modern, concerned with common linguistic mistakes (Maṭar 1966:30; Ayoub 2006:630). Chejne (1969:50) neatly sums it up in that "these endings were looked upon in medieval and modern times as the most precious endowment of the language".

الإعراب حَلْى اللسان، فلا يمنعوا ألسنتكم حَلْيها

4.3 Case in the classroom

The focus on *iʿrāb* in the grammatical tradition is directly reflected in educational practices in the Arab world. There is, to the best of my knowledge, no detailed study of Arabic pedagogical practices, but remarks in the scholarly literature are uniform in their descriptions of the seemingly universal preoccupation with *iʿrāb* in education. Ibrahim, in a discussion on the difficulty learning the *iʿrāb* endings states that

> the teaching of Arabic grammar occupies such a major position that teaching/learning Arabic is often equated with teaching/learning grammar, which in turn is equated with teaching/learning the system of grammatical inflections.
> (Ibrahim 1983:511)

Maamouri (1998:53) states that the exercise of identifying the *iʿrāb* position of words in sentences "has become a central activity in the classroom", and Haeri (2003:40) notes that "grammar as a whole for most people comes to mean the case endings."

What is not mentioned by these authors, and is indeed rarely mentioned in the literature, is the highly formalized and uniform way in which *iʿrāb* is taught. The term '*iʿrāb*' refers, as described above, both to the system of syntactically dependent endings and to the endings themselves, but it is also used to refer to the strictly formalized practices of grammatical analysis (Uhlmann 2012:114). Derivationally, '*iʿrāb*' is a verbal noun and can thus be be interpreted as referring to an action or event. When used in this dynamic sense it takes the meaning of "performing formalized grammatical analysis." It can then be conjugated as a verb: *yuʿribu* 'he performs *iʿrāb* analysis', or in the imperative (a

form many students are more familiar with) *aʿrib* 'perform (2m.s.) *iʿrāb* analysis'.[36]

The practice of performing *iʿrāb* analysis is typically learned by imitating examples. In ad-Daḥdāḥ (1992:16), one of the few works where the practice is explicitly, if only partly, described, it is stated that "[to perform] *iʿrāb* on constructions is to mention the position of each and every part of the clause in the construction". For a nominal (*ism*), he continues, this means specifying

إعرابُ المركَّبات هو ذكرُ موقع كلّ جُزءٍ من أَجزاءِ الجملة في التَّركيب

(*a*) nominal type (participle, proper name, etc.);
(*b*) inflectional paradigm;
(*c*) case and case governance; and
(*d*) form of the case marker

Each of these features is a associated with a predefined and highly formalized set of phrases from which the student is expected to choose the appropriate phrase and to reproduce it verbatim. If, for example, a noun has been declared to be in the nominative in *c*, its case marker is specified in *d* with one of five possible phrases, one for each way the nominative may be marked in the different paradigms. If the nominative marker is *-u*, then *d* is to be realized with the exact phrase "and its nominative marker is the overt final vowel *u*" (ibid.: 16). Figure 4 on the facing page is an example of this procedure, taken from a Syrian school book for the seventh grade. It demonstrates how to perform *iʿrāb* on the sentence in (7).

وعلامة رفعه ضمَّة ظاهرة في آخره

(7) *nuqaddir-u* *lladayni* *yurabbiyāni* *abnāʾ-a-humā*
1PL.appreciate-IND who 3.raise.DUA children-ACC-their.DUA
tarbiyat-an *ṣāliḥat-an*
raising-ACC good-ACC
'We appreciate those [two] who raise their children well.'

A large part of language instruction consists of analyzing example sentences in this manner. Knowing this system, memorizing the set phrases, and knowing in what situations to use them is the primary way of demonstrating knowledge in Standard Arabic (cf. Uhlmann 2012:114). Students are called on to do *iʿrāb* on sentences orally in the classroom and in writing on exams. In primary school, this typically involves constructed sentences such as (7). At

[36] Lane's dictionary translates the verb *aʿraba* as "*He declined* a word" and "*He analyzed grammatically,* or *parsed* the sentence" (Lane 1863:vol.v, 1992, italics in original). Wehr's dictionary does not mention the use of the word to refer to analysis, but translates it as "to use desintential inflection, pronounce the *iʿrāb*" (Wehr 1994:703).

FIGURE 4: *Example of school book* iʿrāb. *Reproduced from al-Xayr and Muḥammad (2011:100)*

نقدِّر اللَّذَين يربِّيانِ أبناءَهُما تربيةً صالِحةً.

نقدِّرُ: فعلٌ مضارعٌ مرفوعٌ، وعلامةُ رفعهِ الضّمّةُ الظَّاهرةُ على آخرِهِ، والفاعلُ ضميرٌ مستترٌ تقديرُه " نحنُ ".

اللَّذَينِ: اسمٌ موصولٌ مبنيّ على الياء في محلِّ نصبٍ مفعولٌ بهِ لأنَّه مُثنَّى.

يُربِّيانِ: فعلٌ مضارعٌ مرفوعٌ، وعلامةُ رفعِهِ ثبوتُ النُّونِ لأنَّهُ منَ الأفعالِ الخمسةِ، وألفُ الاثنَينِ ضميرٌ متّصلٌ مبني على السكون في محلِّ رفعٍ فاعل.

أبناءَهُما: مفعولٌ بِهِ منصوبٌ، وعلامةُ نصبِهِ الفتحَةُ الظَّاهرةُ على آخرِهِ، و(هما) ضميرٌ متَّصِلٌ مبني على السكون في محلِّ جرٍّ مضاف إليه.

تربيةً: مفعولٌ مُطلَقٌ منصوبٌ، وعلامةُ نصبِهِ الفتحَةُ الظَّاهرةُ على آخرِهِ.

صالِحَةً: نعتٌ منصوبٌ، وعلامةُ نصبِهِ الفتحَةُ الظَّاهرةُ على آخرِهِ.

***iʿrāb* example**

We appreciate those [two] who raise their children well.

Appreciate (*nuqaddiru*) is an indicative imperfective verb and its indicative marker is the overt (*ẓāhira*) final *u*, and the subject is the pronominal verb inflection (*ḍamīr mustatir*), implicitly "we".

Who (*alladīna*) is an invariable (*mabnī*) relative pronoun with a final [letter] *yāʾ* in accusative direct object position.

Raise (*yurabbiyāni*) is an indicative imperfective verb and its indicative marker is the retention of [the letter] *nūn* because it is one of the 'five verbs'; the [letter] *alif* of the dual is a pronominal verb inflection with an invariable vowelless ending (*sukūn*) in nominative subject position.

Their children (*abnāʾahumā*) is an accusative direct object and its accusative marker is the overt final *a*, and 'their' (*humā*) is an enclitic pronoun with an invariable vowelless ending in genitive annexation position.

Raising (*tarbiyatan*) is an accusative absolute object and its accusative marker is the overt final *a*.

Good (*ṣāliḥatan*) is an accusative adjunct and its accusative marker is the overt final *a*. (al-Xayr and Muḥammad 2011:100)

later educational levels, it is performed on verses from classical poetry where *iʿrāb* becomes far more complicated than in the example above, often involving governance by and on constituents that are not overtly expressed in the sentence but that are implied (*muqaddar*) by grammatical theoretical. There are numerous books that give word for word *iʿrāb* analyses of the primary textual material in Arabic language instruction: of poetry (e.g. al-Maʿarrī 2007) and of the Quran (e.g. al-Ibrahīm 2007).

These educational practices have a palpable effect on attitudes towards Standard Arabic, and specifically towards the case system, the focal point of language instruction. This is captured by Haeri in her anthropological study on language practices in Cairo, in which she notes that her informants often complained about the difficulty of Standard Arabic (*fuṣḥā*). She continues:

> Every time I asked about what specifically people found difficult, they would give examples of problems with case endings. It is difficult to exaggerate Egyptians' attention to and fear of the case system. There is an ever-present and all-pervasive consciousness about them. Hence while everyone knows that the *tashkiil* [i.e. *iʿrāb* endings] are of utmost importance in reading the Qur'an, their active use in other contexts is feared and disliked, as in grammar classes or at exam and composition times. (Haeri 2003:42)

Uhlmann (2012:105) similarly describes these methods of grammar instruction as leading to the "alienation of Arabs from Arabic Grammmar."

The centrality of formal grammatical analysis in language instruction comes at the cost of time spent on developing other proficiencies. The *Arab Human Development Report*, published by the United Nations Development Program, while not mentioning *iʿrāb per se*, describes the situation of first language Arabic teaching as suffering from an "acute crisis" apparent in the

> concentration on the superficial aspects of teaching grammar and morphology, rather than on core concepts of texts and their respective holistic structures; inattention to semantics and meaning; neglecting the functional aspects of language use, such as improving linguistic skills in everyday use. (Bennani et al. 2003:125)

Wahba points to the contrast between how, in second language teaching in Arabic, the focus has shifted to oral proficiency and to the lack of such a parallel development in first language teaching. In first language teaching, he continues,

> the focus is the language lesson itself, rather then the development of a student's ability to use Arabic in class. The student's role is to listen, and memorize the lesson rather than to express, discuss and be critical in the language. (Wahba 2006:110)

The importance of these practices for our purposes is their role in reproducing ideas of linguistic correctness as they relate to morphological case. These practices may not be effective for developing proficiency in Standard Arabic and in using the case system, but they imbue students with a clear sense that knowledge of the Arabic case system is essential. A student who has gone through the entire educational system may not have learned to use morphological case well, but they will certainly have learned that they should have. The classroom, where virtually all native speakers spend or have spent hundreds of hours, is an important arena for reproducing ideas about language and linguistic correctness, and one of the core ideas that is communicated, simply by the time and effort that is invested in it, is that *iʿrāb* endings are an essential part of the Arabic language.

This has profound implications for what it means for a speaker to engage in active and creative speech in Standard Arabic. The negligence in the acquisition of oral skills via language instructions means that there is no readily available way of speaking Standard Arabic. Persons who find themselves in position where they are expected to perform extemporaneously in Standard Arabic have received no applicable model in their education for doing so. The centrality of the case system is clearly communicated, but not how it is to be applied in speech. The speaker therefore has to find or device their own way of using or not using case endings in speech.

4.4 Case markers and syntactic redundancy

Case and mood endings are in the Arabic tradition regarded both as a sign of the beauty of the Arabic language, and as indispensable for a sentence to be logically coherent (Bohas et al. 1990:5). az-Zubaydī (d. 989) writes, for example, in the introduction to his biographical encyclopedia of Arab grammarians that God gave each people their own language, of which Arabic is the most beautiful and the most clear. In this context he specifically singles out *iʿrāb*: "He [God] made *iʿrāb* an ornament of the tongue and a line of demarcation concerning the difference in meaning" (az-Zubaydī 1954:1, translated in Chejne 1969:15. In Arabic grammatical treatises this point is often made by referring to the etymological relationship between the word *iʿrāb* and the verb *aʿraba* 'express' or 'clarify', the implication being that *iʿrāb* is necessary of speech to be expressive and clear. This argument is found in classical grammars (e.g. aš-Širbīnī 1981: 34) and modern grammars alike (e.g. Bayṭār 2007:18).

وجعل الإعراب حليًا
للّسان،زماما وفضلا لما
اختلف من معانيها.

From a linguistic point of view, it is, however, quite clear that case markers are syntactically superfluous. The main role of case in systems of the accusa-

tive type, such as that in Arabic, is to differentiate the subject from the object, but the subject-object word order in Standard Arabic is, in practice, very rigid (Holes 2004:251). The word orders VSO and SVO, exemplified in (8) are both very common, and one is hard pressed to find alternative word orders such as the VOS and OVS word orders in (9) in which case markers would be needed to disambiguate the sentence.[37]

(8) a. *aᶜraba ṭ-ṭālib-u l-bayt-a*
analyzed DEF-student-NOM DEF-verse-ACC

b. *ṭ-ṭālib-u aᶜraba l-bayt-a*
DEF-student-NOM analyzed DEF-verse-ACC
'The student analyzed the verse.'

(9) a. *aᶜraba l-bayt-a ṭ-ṭālib-u*
analyzed DEF-verse-ACC DEF-student-NOM

b. *l-bayt-a aᶜraba ṭ-ṭālib-u*
DEF-verse-ACC analyzed DEF-student-NOM
'The student analyzed the verse.'

A situation where case endings may serve a disambiguating function in Modern Standard Arabic is in attributing adjectives to one of the constituents in a genitive construction that are of the same gender. If they are of different gender, the ambiguity is resolved by the gender agreement in the adjective. Badawi et al. give the example in (10). In print, where the case endings are not present, or if they are not pronounced when spoken, the adjective *jadīd* 'new' could be attributed either to *mašrūᶜ* 'project' or *qānūn* 'law', giving the two possible interpretations in the translation. In practice, as the authors note, contextual knowledge rarely leaves much doubt of the correct interpretation.

(10) *mašrūᶜ-u qānūn-i l-xidmat-i l-madaniyyat-i l-jadīd*
project.M-NOM law.M-GEN DEF-service.F-GEN DEF-civil.F-GEN DEF-new.M
'**the new project** for the law of community service' or less likely, 'the project for the new law of community service'
(Badawi et al. 2004:110, emphasis in original)

[37] Agius (1991) gives some examples of VOS word order in contemporary literature, but all are sentences in which the thematic roles of the arguments leaves the object-subject order as the only logical option. One example is *ġannā hāḏā š-šiᶜr-a ašhar-u l-muġanniy-īn* (sang this DEF-poetry-ACC DEF-famous-NOM DEF-singer-MPL.GEN) 'The most famous singers sang this poetry' (ibid.:44).

On a more practical level, the syntactic redundancy of case markers is also shown in its role in reading. According to Stetkevych (2006:84) and Bateson (1967:81–2), native speakers do not maintain a mental awareness of case endings when reading. In an experiment reported in Mughazy (2005:70) where participants where asked to read sentences aloud (in an example of what in 4.7.1 is termed informal read speech), no participant produced any case endings, even when the endings were explicitly printed as vowel signs. Furthermore, in a study with second language learner participants Khaldieh (2001) found no correlation between the ability to correctly add case endings and reading comprehension, indicating that one can be proficient reader with a poor knowledge of the case system.

There are two well known Quranic verses that are worth mentioning in this context because of the their role in popular discourse as arguments for the functional role of case endings: verses 9:3 (11) and 35:28 (12). Both of these verses would be theologically absurd had not the case endings canceled the default SO word order. Example (11) would be interpreted as God disassociating himself from the Prophet, as he does with the idolaters, were it not for the nominative ending on *rasūl* 'messenger'. This ending leads to the interpretation of *rasūl* as a second topic of the equational clause, and thus as the Prophet also 'being quit' the idolaters. Example (12) would in the SO word order be interpreted as God fearing *al-ᶜulamāʾ* 'the learned'. The accusative ending on the word *allāh* 'God' overrides this interpretation, marking it as object of the verb *yaxšā* 'fear', and the nominative ending on *al-ᶜulamāʾ* marks this word as the subject so that it is the learned who fear God

(11) *ʾanna allāh-a barīᶜ-un min al-mušrik-īna wa-rasūl-u-hu*
COMP God-ACC innocent-NOM from DEF-idolater-MPL.GEN and-messenger-NOM-his
'God is quit, and His Messenger, of the idolaters.'[38] (Quran 9:3)

أَنَّ ٱللَّهَ بَرِيٓءٌ مِّنَ ٱلْمُشْرِكِينَ وَرَسُولُهُۥ

(12) *ʾinnamā yaxšā llāh-a min ᶜibād-i-hi l-ᶜulamāʾ-u*
PART 3MS.fear God-ACC of servants-GEN-his DEF-learned.PL-NOM
'Even so only those of His servants fear God who have knowledge'
(Quran 35:28)

إِنَّمَا يَخْشَى ٱللَّهَ مِنْ عِبَادِهِ ٱلْعُلَمَٰٓؤُا۟

Quranic examples, of course, tell us little about usage in Modern Standard Arabic. In fact, constructions such as those above would most likely be considered poor style, if not outright wrong, in modern prose. Furthermore, accord-

[38] Translations of the Quran are from Arberry (1983).

ing to Holes (2004:17) "the number of cases [in the Quran] where inflection is meaning-bearing and in which there would be a real chance of, say, subject being confused with object, compared to those where it is not, is minuscule."

In summary, even if it is possible to find or construct sentences that are ambiguous and where case endings would resolve this ambiguity, the relatively fixed word order of Standard Arabic renders case marking redundant in the vast majority of real sentences.

4.5 Pause and pausal forms

In traditional descriptions of Arabic, and with particular importance for Quranic recitation, there is a system of omitting final short vowel and nunation before a pause (*waqf*). How this system is interpreted and applied has important implications for case marking in Spoken Standard Arabic since he omission of the final short vowel often means omission of the case ending. A distinction should be made here between this shortened *pausal form* on the one hand and a *pause position* on the other. A word may be realized in its pausal form with the final short vowel omitted without being followed by a pause and thus not being in pause position. The complete word form with the final vowel retained is referred to as the *context form*, and words not followed by a pause are said to be in *context position*. The pausal form is also the citation form, the form used when mentioning the word in isolation. The citation form is therefore unmarked for case in most paradigms (cf. Table 8 on page 127.)

The morphological implications of pause have been thoroughly described by the classical Arab grammarians, but there is little indication of *when* pausal forms are to be used. In Ibn ᶜAqīl's commentary on *al-Alfiyya*, for instance, the section on pause consists of a list of instructions introduced with *iḍā waqafta ᶜalā …* 'if you make a pause at …', followed by some class of nouns and a description of its pausal form. But it is not stated where pauses are to be made (Ibn ᶜAqīl 2005:513–17). In manuals on Quranic recitation (*tajwīd*), on the other hand, the question of where pauses must, may, and must not be made when reciting is strictly codified with the different levels of permissiveness of pause marked in the Quranic text of modern editions. The distribution of pause locations in the Quran is based on whether a pause would confirm or corrupt the meaning of the text. An obligatory pause, for example, is usually found in between verses, roughly corresponding to sentences, and optional pause between between clauses (Nelson 2001:29). Pause positions are thus codified only with regard to specific points in the Quranic text.

In the Western literature on Arabic, the question of where pausal forms

are used has not received much attention. In text books, the existence of pausal forms is typically mentioned only in passing. The definitions of pause one does find fall rather neatly into two groups: phonemic definitions, that is pause as silence, and syntactic definitions, pause as clause or phrase boundaries. In the first group one finds Holes (2004:62) for whom "pause is defined (rather vaguely) as an audible break in delivery." Similarly, Mitchell (1990:99) gives the advice that in discourse, "the principle should always be followed that, whenever a pause is made, the preceding word should be pronounced in pausal form". In the second group, one finds the majority of grammars, for example, Wright (2011:vol.iii, 368), Schulz et al. (2000:47), and Haywood and Nahmad (1993:511). These authors advise using pausal forms at the end of sentences when reading prose. The two definitions overlap, since in fluent speech phonemic pauses are normally produced between major syntactic units (see Griffiths 1991:349; Kormos and Dénes 2004:149, and references therein).

There is thus no clear prescriptive description of when pausal forms are to be used. An important aspect of the pause phenomenon for our purposes is that it provides, for most words, a prescriptively sanctioned, alternative form that is unmarked for case. Indeed, lack of case endings in speech is often interpreted as an extension of the pause system, an exploitation of an area where the traditional prescriptive sources are unclear, allowing speakers to make away with much case marking but still remaining, nominally, within the boundaries of 'correct' Standard Arabic.

4.6 Case in writing

Case endings are only sporadically represented in writing due to the characteristics of the Arabic writing system. In most Arabic texts, case is only marked in a limited set of words and in certain grammatical contexts. The script therefore provides only limited input when learning the case system. In this section, the Arabic writing system is first briefly described and thereafter follows a description of the specific situations in which case is marked in unvowelled Arabic text.

The Arabic script has two layers of orthographic representation: one basic layer of letters representing consonants, long vowels and punctuation, and a second layer of complementary orthographic elements positioned above and below the letters of the basic layer. The second layer represent vowels, absence of vowels and lengthening of consonants (*tašdīd*). Any text can thus be represented with a so-called deep orthography, which does not clearly reflect pronunciation, or in a shallow orthography, where the pronunciation is more or

less unambiguously represented. The phrase in (13) can, for example, be written in Arabic without short vowels, as in (14a), or with complete vowelling, as in (14b).

(13) *niẓām-u* *l-kitābat-i* *l-ᶜarabīyy-u*
system-NOM DEF-writing-GEN DEF-Arabic-NOM
'the Arabic writing system'

(14) a. نظام الكتابة العربي

b. نِظَامُ ٱلْكِتَابَةِ ٱلْعَرَبِيُّ

Most texts only have the first, consonantal layer of letters. The second layer of vowel signs is fully employed in the Quran, the Bible, and in other religious source texts in order to preserve the exact form and ensure correct recitation. It is also employed in children's literature to facilitate reading.[39] Texts can also be written with various degrees of vowelling. In some vowelled texts vowel signs that are always used with some letters, such the *a* preceding *tāʾ marbūṭa*, are not always printed. In unvowelled texts vowel signs are occasionally included to avoid ambiguity in homographs or to facilitate the reading of rare words. Vowels are also normally added in connection with orthographic *alif* in otherwise unvowelled texts (see below).

Since most case endings take the form of short final vowels they are not graphically represented in writing. In the example above, all case markers take the form of short vowels, and so are not represented in the unvowelled rendering. Only in certain paradigms and grammatical cases is the case marker graphically represented. These represent only a small proportion of all nominals in a text. A printed Arabic page therefore typically contains only sporadic overt case markers, typically one or two per paragraph.[40] Arabic text can thus

[39] This has the effect that children's literature is in this regard morphologically more complex than are texts for adults. The adult reader of unvowelled text can, and normally does, ignore case that is not orthographically represented (see 4.4), while in children's literature case endings are ever-present.

[40] In the present corpus, tokens that would have a graphically represented case marking in writing together represent 9% of the data. While not directly applicable to written texts, since word frequencies may differ between speech and writing, it suggests that roughly every tenth noun in an edited text has an overt case marker. Most of these endings need not be processed as case endings when reading the text, however. Indefinite adverbs may be considered to take the ending *-an* (ـً) as part of word formation rather than as part of the case system, and the accusative/genitive form of the sound m.pl. and the dual paradigms *-īn/-ayn* (ـين) are compatible with the dialectal caseless forms encoding only number. If these forms are excluded from the count of orthographically marked case the number is reduced to 6%.

Table 4: *Forms of orthographically marked case*

	INDEF. TRIPTOTE	INDEF. DEFECTIVE	SOUND M.PL.	DUAL	FINAL *HAMZA*[a]	THE FIVE NOUNS[b]
	'letter'	'judge'	'Egyptians'		'your part'	'father'
NOM	حرف	قاض	مصريون	مصريان	جزؤك	أبو
GEN	حرف	قاض	مصريين	مصريين	جزئك	أبي
ACC	حرفًا	قاضيًا	مصريين	مصريين	جزأك	أبا

[a] Only with an enclitic pronoun.
[b] Only in construct state.

to a large extent be read without attention to case endings (since they are not there in the first place), but occasionally words appear in graphically alternative forms in which case is marked. The types of words that take orthographically marked case in normal unvowelled texts are listed Table 4, with examples of their graphical forms in the three cases.

One of the most frequent situations where case endings are graphically represented is the orthographic *alif* in accusative indefinite forms in the triptote paradigm. Compare for instance the case endings for the word *ḥarf* 'letter' in the three cases. In nominative and genitive it takes the endings *-un* and *-in* respectively, endings not represented in unvowelled text. (The *n* of the nunation is represented by a doubled vowel sign and not a consonantal letter.) The word is in these forms written حرف. In the accusative it takes the ending *-an* that is represented in writing with the orthographic *alif* (ا) together with the vowel sign (ً), giving the graphical form حرفًا. The typographical convention of representing this ending with the vowel sign and not only with the letter *alif* makes this ending particularly visually prominent in unvowelled text.[41] Orthographic *alif* is not added if the word stem ends in *tāʾ marbūṭa* (ـة), *-āʾ* (ـاء), or *-aʾ* (ـأ).[42]

[41] Searches in arabiCorpos show that the vowel sign for *-an* is around ten times as common in newspaper text than is any other vowel sign. 99% of these instances of *-an* are occur in connection with orthographic *alif*. The convention of printing *-an* in connection with *alif* seems not to be followed in electronically published text. A quick check on the web pages of Al Jazeera, BBC Arabic, al-Ahram and Al Arabiya, showed that in news articles in these pages this vowel sign is not present.

[42] See Fischer (2002:8) and Schulz et al. (2000:333). Wright (2011:vol.i, 12) states that according to older orthography, final *hamza* always prohibits orthographic *alif* but that this rule is usually not followed.

Orthographic *alif* also applies in indefinite forms in the defective paradigm, in which case the stem final weak radical appears as the letter *yāʾ* (ـيـ). The word *qāḍin* 'judge', for example, is written قاضٍ in the nominative and genitive forms with the final radical omitted, but as قاضيًا in the accusative.

The other frequent situation where case is graphically represented is in sound m.pl. paradigm. The word *miṣrī* 'Egyptian' in the genitive and accusative plural forms take the ending *-īna* and is written مصريين. In the nominative, it takes the ending *-ūna* and is written مصريون. Similarly, the dual ending is *-ayni* in the genitive and accusative, giving the graphical from مصريين that is identical to the plural, while the dual nominative ending is *-ān*, giving مصريان.

The two final situations where case is graphically represented are much less frequent. The first is in the five nouns paradigm in construct state. The five nouns, as the name suggests, comprise a group of merely five different lexemes. Case is here orthographically marked only in construct state since the case marker then takes the form of a long vowel. The word *ab* 'father [of]' is for example *abū*, *abī*, *abā* in the three cases respectively, written أبو, أبي, and أبا respectively.

The final group of words where case marking is orthographically represented is on word stems with a word final letter *hamza* 'on the line', that is not resting on another letter (ء), for example in the word *juzʾ* 'part', written جزء. The *hamza* (pronounced as a glottal stop) is written on the line if preceded by a long vowel or by a consonant. If the stem has an enclitic pronoun the *hamza* is no longer word final and other rules for the spelling of *hamza* come into play. *Hamza* then takes as its 'chair' the letter that corresponds to the following case vowel. Thus while the word *juzʾ* takes the same form جزء in three cases, the graphical shape of *hamza* is different in the three cases if the word has an enclitic pronoun such as *-ka* 'your (m.)' (ـك): جزؤك in nominative, جزئك in genitive, and جزأك in accusative. It should be noted that these variable spellings of *hamza* are somewhat esoteric. Even Arabic specialists are often uncertain as to the correct spelling and there is some variation even in published, professionally edited texts (Parkinson 1990:276–7).

4.7 Case in read speech

Reading a text aloud or reciting it from memory is a very different activity from normal speech production. As bound by the written text, the speaker is freer to focus on formal aspects of their speech, such as case endings. Where these are represented in the text they can then be directly reproduced from the visual input. The writing system of Arabic and the way case is represented in it has

implications for what it means to read Arabic aloud. This is described below, followed by a discussion of the proposed limited role of Quranic recitation in learning the case system.

4.7.1 Orthography and read speech

In languages with shallow orthographies such the as Spanish or Finnish a person who has learned the alphabet and basic pronunciation can, in principle, read a text aloud by simply 'sounding out' the text, mapping letters or combinations of letters to sounds, without necessarily understanding what he or she is reading. This is not possible in the normal deep orthography of Arabic where short vowels are not represented. Reading a text aloud in Arabic necessarily involves adding phonemic information not present in the text. This addition of phonemic information is of two kinds. The first is the addition of word internal vowels that are specific to the lexeme. This is done by identifying the word from the consonant string, often in combination with the grammatical context. When the word has been identified the reader adds the appropriate vowels, either as stored in the mental lexicon or by deducing them from rules of word formation. This lexically based vowel addition is non-optional. A words cannot be phonetically produced without applying some vowelling pattern to the consonantal skeleton. The second kind of vowel addition is to add the syntactically dependent case and mood endings. It requires an awareness of the syntactical role and the inflectional paradigm of the words. This kind of vowel addition is not strictly necessary since the case and mood endings are syntactically redundant, as discussed above. Where word final vowels are needed for prosodic reasons, epenthetic vowels can be inserted in accordance with the dialectal system.

Read speech in Arabic can thus be divided into two kinds depending on how these two types of vowel addition are applied. First, in what may be termed 'casual' or 'informal read speech', a text is read in a situation where there is no external demand for linguistic correctness and where the aim is simply to transmit the content of the text. One example of such a situation would be reading a part of an article to a friend or reading instructions aloud from a manual. In this form of read speech, only the non-optional lexically based vowel addition is normally employed.

Second, in situations where there is an external pressure enforcing linguistic correctness, in what might be termed 'public' or 'formal read speech', the target is full realization of the syntactically dependent endings, except for words in pause position. Typical cases of formal read speech are news broadcasts, voice overs, and lectures and speeches (although in the latter two this is very much

dependent on the speaker). This is also the way pupils are trained to read aloud in school. It is generally regarded as the correct way of reading a text aloud, although few persons have reasons to perform it after finishing their education. Naturally, the extent to which the aim is reached depends on the skill of the person reading and on how well prepared they are. In the following, 'formal read speech' will refer to performances by professionals trained for the purpose who are producing an oral text that is largely prescriptively correct.

The one point where formal read speech typically deviates from codified grammar is in the pausal form of the accusative case ending *-an* in triptote nouns without *tāʾ marbūṭa*, that is in situations where it is written with orthographic *alif*. This is prescriptively and in Quranic recitation pronounced as a long final vowel *ā*, but in formal read speech it is pronounced as a short vowel with nunation, that is identical to the context form (cf. Harrell 1964:34). For example, the word *dāʾiman* 'always' in news broadcasts would be pronounced with nunation as *abadan* also in pause, but in Quranic recitation as *abadā*.[43]

Formal read speech is an important point of comparison for patterns of case marking in Spoken Standard Arabic. It is used in prestigious situations and regarded as linguistically correct. It is likely to be taken as a model or point of reference for persons developing skills in Spoken Standard Arabic.

4.7.2 Quranic recitation

Recitation of the Quran is a form of Standard Arabic that millions of people are exposed to daily in large amounts. Quranic recitation is heard on public transportation, in shops and in homes played on recordings or broadcast on radio or TV. On special occasions, the recitation is performed live by a professional reciter. The Quran is recited with full inflection of *iʿrāb* endings as fixed in the Quranic text and with observance of pausal forms. It is for many Muslim children the first form of Standard Arabic they are exposed to through devotional practices in the home or through Quran schools (Wagner and Lotfi 1983:115–8; Haeri 2003:passim).

While quantitatively a main part of many people's exposure to Standard Arabic and its system of case endings, there are reasons to believe that Quranic recitation has only limited effect on the peoples proficiency in the case system despite the importance ascribed to it in Muslim culture. It is in many ways set apart from other uses of language. First, it is recited publicly in the form of *tajwīd*, a highly developed artistic form of recitation, with minutely codified rules regarding, amongst other things, degrees of vowel length, nasalization,

[43] According to Holes (2004:65) the prescriptively correct pausal form of the ending is regularly heard in Omani news broadcasts.

and on *sandhi* phenomena specific to the genre (Nelson 2001:chap. 2). *Tajwīd* is also highly melodic, with "melodic contour, conventions of modulation and transposition, use of melodic cadence [...] dictated by the aesthetics of the improvisatory style of Arabic music" (ibid.:125). Second, the Quran as a fixed text is clearly set apart from all other samples of language in Muslim culture as the verbatim word of God. As Haeri puts it,

> the nonarbitrary relation between the forms and meanings of the Qur'an is articulated in strikingly similar ways by educated and uneducated people from widely different social and generational backgrounds. (Haeri 2003:13)

The mere *sound* is taken to be of the utmost significance, as is emphasized in the musical qualities of *tajwīd*. The case endings in the Quran are part of this fixed string of sounds. The Quranic text is normally first encountered as a text to be memorized and recited with correct pronunciation long before the pupil is expected to understand the syntax behind the varying forms (Wagner and Lotfi 1983:116). Third, the archaic language of the Quran sets it apart from modern texts. While the syntax has changed little between the Arabic of the Quran and Modern Standard Arabic, the vocabulary and stylistic patterns of the Quran are highly archaic. Linguistic patterns that are heard and perhaps even learned through recitation cannot be directly transferred and applied to other forms of Arabic.[44]

Tajwīd is thus performed in a form of Standard Arabic with complete case endings and adherence to pause rules, but it is largely detached from other forms of linguistic interaction. As an ever present feature of the linguistic surrounding, *tajwīd*, as pointed out in Walters (2003:99) "becomes a ready resource in interaction for speakers to index Islam, the Quran, recitation, or religion, whether in a serious or jocular manner", but this is in the form of quotes and formulaic expression, not as an integration of its linguistic structures into other forms of oral Arabic.

4.8 Case in speech: previous research

Extemporaneously spoken Standard Arabic is characterized by sparse use of case endings and relatively little is known about the extent to which the endings are used and according to what patterns. Grammars and textbooks typically only mention that there is such variation and that practices differ with genres and contexts (e.g. Badawi et al. 2004:33; Ryding 2005:166-7). Investigating the use of case endings in speech is the overarching aim of this study and results of this investigation are presented in Part III. In this section, previous research on

[44] But see (100) on page 185 for a possible counterexample.

the use of case in spoken Standard Arabic, two dissertations and three articles, is reviewed. The dissertations, Schulz (1981) and Elgibali (1985), investigate case ending as one of a series of diglossic variables. Of the three articles, Parkinson (1994a) and Meiseles (1977) investigate the use of final vowels generally while Magidow (2012) focuses exclusively on case endings. These studies are reviewed below in chronological order and only to the extent that they deal with case morphology. Studies of Arabic speech that is not formal (Palva 1969; Blanc 1964), not extemporaneous (Harrell 1964; Van Mol 2003), or that does not deal with case endings (Mejdell 2006) are not reviewed.

4.8.1 Meiseles (1977)

Meiseles discusses what he calls the "restitution of 'word-endings'" in Oral Literary Arabic (OLA) as represented by radio broadcasts recorded between 1969 and 1974. Of particular importance for the present study is his observation that it is in some situations impossible to determine whether a certain vowel is in fact a case ending or is produced by dialectal morphology. He mentions in this regard the vowel *i* before a definite article and the enclitic pronouns *-u(h)* 'his' and *-ak/-ik* 'your (m./f.)'. The methodological consequences of this are however not carried through. In the material he found that case markers are more likely to occur before *waṣla* and in the construct state followed either by an enclitic pronoun or a noun (Meiseles 1977:176). One of the examples provided of vowel ending in construct state is (15). The final vowel in *majlisi* could have been produced as a helping vowel in the dialects. It is therefore doubtful whether it is an example supporting the claim that case marking is preferred in construct state preceding the article. It begs the question as to what degree observations of this kind lie behind the general observation.

(15) *qarār* ***majlis-i*** *l-amn*
decision counsel-GEN/AMB DEF-security
'the decision of the security council' (Meiseles 1977:176)

Meiseles furthermore mentions two patterns of case marking in speech that differ from the traditional prescriptive rules and that are so common, he suggests, that they are best regarded not as mistakes but as conventions of OLA (ibid.:170–8). The first pattern is the use of case endings "in relative freedom" in pause position. The second pattern is the application of the basic triptote case marking paradigm on the less frequent diptote nouns. No numeric data are provided for any of these observations and they appear to be largely impressionistic.

4.8.2 Schulz (1981)

Schulz studied diglossic variation in the speech of 49 speakers appearing on Egyptian radio and he analyzed data with respect to a large number of features. Seventeen of the 49 speakers in his corpus made no use of case endings. Of the 32 that did, 21 marked case less than 5% of the time, pause positions excluded. Schulz investigates data of the remaining eleven speakers that mark case at rates over 5%. He breaks this data down as follows, with rates of case marking in percentage for each case:

Nominative	13%
Genitive	19%
Accusative	39%
N, G, A Combined	31%
Accusative Indefinite	67%
N, G, A Combined in Feminine Nouns	14%

(Schulz 1981:138)

These data show a difference between case marking frequencies in the three cases, with the accusative being the most frequently marked. This he attributes to orthographic *alif*, which forces the reader to develop and maintain awareness of this particular case ending (ibid.:139). This interpretation is supported by there being no difference in case marking between definite and indefinite nouns except in the accusative where only indefinite forms take orthographic *alif*. The higher rate of genitive marking than nominative marking can, according to Schulz, be explained by the fact that words in the genitive "always immediately follows the words that requires it to be in the genitive". He also cautions that counts of genitive marking may be inflated due to dialectal epenthetic *i* being counted as a case ending (ibid.:139).

4.8.3 Elgibali (1985)

The study by Elgibali on the coherency of diglossic levels in Egyptian and Kuwaiti Arabic was briefly discussed in 2.2. It is revisited here with a focus on the results it presents on the occurrence of case endings. As with Schulz (1981), case endings were only one of several variables investigated. The quantitative results that are of interest here are those for case marking in the higher end of the diglossic scale: classical Arabic, Modern Standard Arabic (MSA), and Educated Colloquial, levels 1 to 3 in Badawī's scheme (see 2.2.2). Data for

these levels is divided into informal and formal speech. The percentages of case marking, summarized from Elgibali (1985:77; 98), are as follows:

	CAIRENE (%)		KUWAITI (%)	
	formal	informal	formal	informal
Classical	85	20	88	24
Modern Standard	20	10	39	13
Educated colloquial	9	2	12	3

These data indicate that there is a more dense usage of case endings going up the diglossic continuum and going from informal to formal speech. As was mentioned in 2.2, there are problems with the data collection in the study, particularly in how the concept of formality is defined. The data for informal oral MSA is language used in "television and radio newscasting and in university literary lectures" (ibid.:39). It is not stated in the study why these contexts are to be categorized as informal. Data is also presented on contrasting *formal* oral uses MSA but the sources of this part of the material are not given. Furthermore, no distinction is made between spontaneous speech and reading aloud. A rate of case marking of 20% in informal MSA, which includes newscasts seems very low. It may however be due to the fact that possibly half of the material was from one single lecture, which may have been presented by an individual who uses very few case endings.[45] There are no indications of how ambiguous endings are interpreted or whether lexicalized adverbs are included.

The gradual decrease of case marking in the levels is disrupted in the formal register of levels 5, which refers to the illiterate colloquial level. For this level, a rate of case marking of 20% is recorded and is due to the inclusion of Quranic recitation in this data. Rather than controlling for read speech or recitation, the author chose to include it in the data of illiterate colloquial since it is frequent in formal situations in this level (ibid.:79). In other words, Quranic recitation is grouped in the same category of data as every-day conversations between illiterate speakers. Quranic recitation is also included in the category of formal Classical Arabic (ibid.:38). All this makes the results unreliable as presented.

[45] The sources used are not clearly presented in study. It is stated that there are two sources of data in this category (Elgibali 1985:39). A table on page 114 that summarize the data indicates however that there are ten sources, each of a length of between 1500 and 2000 words.

4.8.4 Parkinson (1994)

A different and more detailed approach is persued by Parkinson in an article that has been an important source of inspiration for the present study. He recorded interviews with a large number of speakers in Cairo in which informants were asked to speak *fuṣḥā*. Four of these interviews are analyzed in great detail for their use of vowel endings. The four speakers had different levels of education and were chosen to represent the whole range of performance levels as evaluated by native speaker assistants. The relevant data are as follows (adapted from Parkinson 1994a:183–4):

	SPEAKING TEST SCORE (1–5)	# CASE VOWELS	% CORRECT CASE VOWELS
Speaker 1	1.65	4	25
Speaker 2	2.00	35	37
Speaker 3	2.80	94	91
Speaker 4	4.50	>90%	>98

The main part of the article consists of an interpretation of these numbers through a detailed analysis of each speaker's oral text. The first speaker used only four ending vowels on nouns, three of which were prescriptively incorrect. Parkinson interprets this as the speaker having little grasp of these vowels as grammatical markers. Rather, speaker 1 includes more or less random vowel endings to mark his speech as Standard Arabic. Speaker 2 used more but still only sporadic ending vowels. He marked some indefinite objects of verbs correctly with *-an* but otherwise added *-a* to nouns and *-u* to unit numbers seemingly at random and without relation to grammatical case. He differed from the previous speaker in using *a* rather than the dialectal *i* as an epenthetic vowel in word boundaries. Only 3 of 35 ending vowels on nominals were on words with a definite article. Speaker 3 was more proficient in Standard Arabic and made more use of final vowels than the previous speakers and, when he used them, they were generally prescriptively correct, with incorrect vowels produced only in complicated constructions. Eleven of 22 nouns marked for case were words with an enclitic pronoun, clearly a favored position for case marking for this speaker. He correctly marked every single indefinite object with *-an*, but did not mark case on a single noun with a definite article. The fourth speaker analyzed by Parkinson used case endings in essentially the prescribed, classical manner, marking the correct case on all nouns and adjectives with adherence to pause rules. The speaker did however add case endings in pause in particularly

difficult constructions, apparently as a demonstration of proficiency. He did all this "with a fluency and ease that is nothing short of breathtaking" (Parkinson 1994a:201). One of the conclusions drawn from this analysis is that speakers act under

> a similar set of constraints, patterning their sprinkling [of final vowels] in a particular way, with some forms (defined both phonologically, morphologically and syntactically) being more likely to receive vowels than others. (Parkinson 1994a:208)

Testing this claim and identifying such constraints are the main aims of the present study.

Parkinson furthermore describes how one of the Egyptian assistants who conducted the interviews first spoke very 'correct' Standard Arabic with full use of final vowels, but since informants seemed to react negatively to this, he "struck a kind of compromise between the totally accurate prescriptive *fuṣḥā* and the social situations as he perceived it" (ibid.:209), ending up speaking in way similar to speaker 3 described above. This is contrasted with the fourth speaker's uncompromising correctness that, according to Parkinson, comes across as annoyingly pedantic. Even the Egyptian interviewer, who had explicitly asked him to speak Standard Arabic had to struggle to hide his annoyance (ibid.:209). In other words, speakers with good oral control over the case system need to negotiate between the overt norms of prescriptive correctness and the covert norms of what is appropriate for the situations, norms that are in conflict even in very formal situations.

A weakness in Parkinson's study is that speakers may have interpreted the instruction to speak *fuṣḥā* in different ways. It could be interpreted either as an instruction to speak prescriptively correctly (in accordance with the codified norm) or to speak the way Standard Arabic is typically spoken (in accordance with the empirical norm). These interpretations may yield different forms of speech from the same speakers, and it is not clear which of these two norm systems is in effect. An alternative approach is to have subjects react to an authentically formal situation, which is difficult to achieve in an experimental setting.

4.8.5 Magidow (2012)

In a recent article that deals specifically with case endings in spoken formal Arabic, Magidow argues that one of the uses of case marking in speech is to mark words for *saliency* and *individuation*. The data are taken from two to four minute segments of television programs, including three episodes of Al Jazeera's debate program *al-Ittijāh al-muʿākis* and its religious program

aš-Šarīᶜa wa-l-ḥayāt, as well as a program where "neither the station nor the name of the program is clear from video" in addition to a program with an "unknown (probably Egyptian) male who gives a lecture on Arabic grammar. Date and location unknown" (Magidow 2012:65–6). The principles behind this particular collection of material are not clarified.

A number of examples are given where case endings are said to mark saliency and individuation. Saliency is defined as "the importance of an entity to the discourse as judged by the speaker of that discourse." Furthermore, "any type of noun could be viewed as relatively more important than surrounding discourse and therefore have salience marking applied to it" (ibid.:69). Individuation on the other hand is defined as a cluster of properties, including agency, morphosyntactic and semantic definiteness, specificity and textual prominence (ibid.:71). As the author himself points out, "the level of individuation and salience are not always crystal clear, and it seems more likely that they represent a continuum rather than discrete levels" (ibid.:72). This vagueness translates into a circular argument in the analysis: case endings mark these properties, and the way we know that a word has these properties is that it is marked for case. More or less any constituent in an example can be interpreted as having the properties of saliency or individuation. One example of this argumentation is the claim that *mā* with the meaning of 'any' or 'some', the so-called *mā ibhāmiyya*, "may be considered a type of specificity marker [!] and as expected it correlates here with the use of case marking" (ibid.:78). Another example is the analysis of an example from a discussion of religious wars in Islamic history (ibid.:81). The first use of the word *ḥarb* 'war' in the example is not marked for case, as, according to the author, it is expected to be as the salient topic of the discussion. It is marked for case later in the example in the phrase *wa-lam takun ḥurūban ḥawla d-dīn* 'but they were not wars over religion'. The author explains this in that it is religious wars rather than wars in general that is the topic of discussion. It does not seem to matter which of the two instances of the word was marked for case for an argument to be formulated that this is done to mark the word as salient.

As in Meiseles (1977), the formal overlap between dialectal epenthesis and case endings is not taken into account in the analysis. There are two examples in the article of a word final *i* being interpreted as a case ending and analyzed as such for saliency and individuation, when this vowel is the expected epenthetic vowel in Egyptian Arabic (see 7.7.1): examples 13: *signi tadmur* 'the Tadmur prison', and 24: *ḥaqqi ʾaḥad* 'anybody's right' (Magidow 2012:77–8; 83). Both examples are produced by Egyptian speakers who would insert an epenthetic *i* also in dialect forms of speech. The vowels cannot be interpreted off-hand as marking case.

Some general observations of these reviews are, first, that reading aloud must be controlled for in order for the data to be properly interpretable. Read speech and extemporaneous speech are different in terms of what case endings the speaker must produce and what endings they can reproduce from a written text, as detailed in previous sections in this chapter. Second, attention must be paid to the ambiguities that arise in case marking due to interactions and formal overlaps with the dialects. When the dialectal system of epenthesis is not taken into account it risks leading to misinterpretation of data. Third, it is important to make sure that speakers strive at the empirical rather than the codified norm by having them react to authentically formal situations. Fourth, there are indications that case marking is structured by syntactic and morphological parameters such as gender (marked with *tāʾ marbūṭa*) and definiteness. These indications provide hypotheses to be tested in the present study.

4.9 Summary

In the this chapter, the case system in Arabic was described in general linguistic terms as being simple on the syntactic level despite widespread assumptions to the contrary and more intricate on the morphological level. The traditional Arabic description of the case system was briefly presented together with a description of how this system is reflected in formal language instruction. It was argued that teaching practices in the Arab world plays a large part in reproducing the notion that case marking is an essential and necessary part of the Arabic language. The role of case marking in 'correct' speech is complicated by the existence of pausal forms in which the case ending is often omitted and whose use is not clearly codified. The way case is marked in the Arabic script, where vowels are not represented, was described in some detail, followed by how this affects practices of case marking in read speech. It was argued that formal read speech, as heard for example in news broadcasts, serve as a model for speakers of Standard Arabic, while Quranic recitation plays a more marginal role. Finally, previous research on case marking in spoken Standard Arabic was reviewed, highlighting both methodological issues and observations to be tested in the present study.

Part II
Method

Material
Text preparation
Morphosyntactic annotation

Material

5

This is the first of three chapters describing the procedures by which the corpus on which this study is based was collected, formatted, and annotated. This chapter describes the choice of material for the corpus, the next chapter describes the how the texts were prepared and formatted, and the chapter thereafter describes the principles and categories of the coding scheme used for annotation.

The first section in this chapter, section 5.1, gives an overview of available Arabic text corpora. No sufficiently large corpus of spoken Standard Arabic was found and it was therefore necessary to construct a new corpus for the purposes of this study. The criteria for such a corpus are discussed in section 5.2. A selection of episodes of the interview program *Liqāʾ al-Yawm*, broadcast on the Al Jazeera news channel, was found to be a suitable pool of material to be included in the corpus. In section 5.6, the channel and the program are presented together with a discussion on how this material relates to the specified criteria and why specific episodes were deemed not to fit the criteria and were excluded. This left 45 episodes with interviewees of different Arabic nationalities. Speakers of three dialectal groups, 17 speakers in total, were selected for inclusion in the final corpus. The motivations behind this choice are discussed in section 5.5. Section 5.6 provides a description of the final corpus, and section 5.7 summarizes the main points of the chapter.

5.1 Arabic corpora

There exist a number of corpora of written and spoken Arabic. Modern electronic corpora consisting of millions of words of written material[46] cannot be used to answer the research questions about the spoken language raised here.

[46] See Al-Sulaiti and Atwell (2006) for an overview of corpora available up to 2006, and Al-Thubaity et al. (2013) for corpora published between 2006 and 2013.

There are numerous corpora of spoken dialectal varieties of Arabic (Behnstedt and Woidich 2013:313). These are also unusable here since they include a minimal use of case markers, if it all. Corpora of spoken Arabic in a formal register are much more scarce. A few studies have been published together with transcribed texts in formal Arabic, including Harrell (1964), Diem (1974), Mazraani (1997), and Mejdell (2006). These are either too short to be subject to quantitative analysis or do not represent extempore speech. Van Mol (2003) constructed a sizable annotated corpus of reports from news broadcasts, comprising 320000 words. As news broadcasts this primarily read speech. Furthermore, vowel endings were not transcribed used as annotations to mark grammatical function (ibid.:163). There is also the so-called Leeds corpus (El-Hassan 1977) that has not been published and whose method of collection has been heavily criticized (see 2.2.3).

Since no suitable corpus was found, a new one was constructed, tailored specifically to answer the research questions regarding Spoken Standard Arabic raised in this study.

5.2 Criteria

The material included in the corpus should comply with external criteria on the one hand regarding the situation, and on the other regarding the participant speakers. Arabic case endings are only used in the highest linguistic registers, and the material for the study therefore needs to be produced in situations where Standard Arabic is the expected variety. The situations in which the (oral) texts are produced should also vary as little as possible among themselves to minimize external variables that affect linguistic register. The text in the corpus should be produced in situations that (*a*) are formal; (*b*) consist of extempore speech, and (*c*) deal with similar topics. To make sure that all speakers have roughly the same relation to standard Arabic in terms of access and expectations, the corpus should only include speakers that are (*d*) native speakers of Arabic; (*e*) highly educated, and (*f*) public figures. These criteria are discussed in order below.

5.2.1 Formality

Formality is a common and intuitively appealing way of describing and categorizing social events. Ferguson's definition of the H variety as the variety that "is used for most written and formal spoken purposes" (1959:336) has become the most common characterization of Standard Arabic. This definition is found in one form or another in virtually all educational material and academic writ-

ings dealing with Arabic. It is mostly used as an over-arching concept in that it is taken to include or correlate with other factors that influence the choice to speak Standard Arabic. Such factors include the educational level of speakers and the 'seriousness' of the topic. When used in this way, 'formality' could be taken to include the other criteria discussed in this section.

The concepts of formality is, however, on closer inspection more vague. It has been used in a variety of ways that are not always made explicit. Uses of the term in sociolinguistics, ethnography of speaking, and related fields is discussed in Irvine (2001). First of all, she notes that either the *situation* or the *code* can be described as formal. In the case of Arabic, one might observe people employing the formal *code* of Standard Arabic and therefore categorize the situation as formal, or one might categorize the *situation* as formal and therefore expect participants to employ the formal code of Standard Arabic. There is a causal relation between a formal situation and a formal code, but it is not necessarily absolute. Educated speakers of Arabic for example may employ different registers of Arabic in reactions to identical situations (see 2.5). Irvine further identifies four different ways the term formality has been used in describing situations. The first type of formality is *increased code structuring* including additional rules or conventions that govern behavior only in formal situations. This might involve the linguistic code, but also the dress code, body language, the spatial organization of the event, or who is allowed to speak and when. The second type of formality is *code consistency* which means that there is little or no alternation between codes, for example for comical or rhetoric effects. The third is the *invocation of positional identities* where participants act as representatives of an opinion, a party, an organization, etc., and not as private individuals.[47] The fourth and last potential meaning of formality is *central situational focus* where participants focus their attention on one specific predefined topic without spontaneous changes of subject. These types of formality often, but not necessarily, coincide (ibid.:196). They provide four aspects by which a situation can be judged to be formal or informal.

A methodologically more specific definition of the word 'formality' is developed in Labov (1972:chap. 3) where the term is used interchangeably with the concept of 'attention paid to speech'. This is operationalized in the 'sociolinguistic interview' that is designed to elicit data consisting of casual speech, interview answers, reading of a text, and reading of a word-list, all from one and the same informant. These are four forms of oral production in which gradually

[47] Cf. this to Diem (1974:55), according to whom "Der höchste Formalitetsgrad ist dann erreicht, wenn in einer formellen Situation eine bedeutende Persönlichkeit über ein wichtiges Thema spricht, zumal sie dann nicht als Privatperson, sondern ex officio Stellung nimmt."

more attention is paid to speech. Reading a word-list is for example described as a more formal activity than reading a story (Labov 1972:80). Seen from this perspective, the corpus should consist of 'careful speech' — speech that is not read from a manuscripts but where the speaker is aware of his own way of speaker and keen to speak in a 'correct' and 'proper' way.

Furthermore, the authenticity of formality is important. Standard Arabic is a variety used only in certain types of social interaction. There are social norms that dictate when, where, and by whom Standard Arabic is spoken. Breaking these norms amounts to socially deviant behavior, even if the fact that there are different interpretations of this norm means that there is a middle ground where the decision to speak Standard Arabic or dialect is not clear-cut. It is the nevertheless the case that deviations from the norm come at a social cost.

Authentic formality is naturally very difficult to create in an experimental testing situation. For example, to ask an informant to speak Standard Arabic in a room together with a researcher, as in the experiment reported in Parkinson (1993), is to request of them to act formally in what is not in and of itself a formal situation. Participants in such a situation are in effect asked to act contrary to social norms. The risk is then that subjects perceive the situation as awkward and lower their linguistic register to better comply with the actual situation. Furthermore, experimental situations where informants are asked to speak in Standard Arabic are problematic in that it is not quite clear what is tested. Informants might interpret an explicit instruction to speak Standard Arabic as an instruction to follow the prescriptive norm of codified grammar (see 3.2), a testing situation they might be familiar with from school. This risks making the data less suitable for an investigation of covert linguistic norms.

5.2.2 Extemporaneity

The corpus of the present study should contain spontaneous speech so as to represent on the spot, unplanned language production. This is important for several reasons. First, this mode of language production is assumed to give the best representation of how the morphological case rules are internalized, since speakers cannot go back and edit the text. Second, the linguistic norms of reading aloud are different from those of speaking, particularly with regards to the inclusion of case endings (see 4.7) and it is only the norms of extemporaneous speech that are under investigation here. Third, some case endings are marked in writing (see 4.6) and are produced in read speech on the basis of visual input, rather than as the result of grammatical processing by the speaker. A speaker reading a text might, of course, also have added other vowel signs to the text to mark case. This influence of the written text will be present both if the speaker

reads directly from a written text or if they recite a text memorized from a written original.

Few situations are however completely spontaneous or unplanned. A common classification of modes of speech is that it may be

> *impromptue*, ie. delivered spontaneously without prior preparation, it may be *extemporaneous*, ie. planned in advance but presented freely, it may be *memorized*, ie. carefully prepared, committed to memory, and read by rote, or it may be *manuscript delivery*, ie. a speech read from a written manuscript. [...] The situation is purely impromptu only if the speaker has not at all been able to predict the speech situation nor the subject of the speech or discussion.
> (Lehtonen 1982:40, emphasis in original)

The criterion for inclusion of texts in the corpus is that they should represent extemporaneous speech. In this form of delivery, syntax and actual wording is planned simultaneously with delivery, whereas the textual structure and word choice is preplanned (ibid.:41). The case endings are on syntactic level and are thus produced online in extempore speech.

The criterion of extemporaneity puts a natural cap on the level of formality. A president's speech to the nation would arguably be more formal than a televised interview with the same president, but would not comply with the criteria of extemporaneity. The aim here is then is to find a material that is as formal as possible while still consisting of extempore speech.

5.2.3 Topic

The topic of a talk or conversation affects the linguistic register in several ways. In Arabic, the diglossic specialization of the dialects and Standard Arabic has led to their developing vocabulary for their respective domains. In discussions of Arabic diglossia, it is often stated that the dialects have underdeveloped vocabulary for scientific and other specialized topics. It is less often mentioned that Standard Arabic has an underdeveloped vocabulary for everyday and intimate or emotional topics. Everyday objects like pieces of clothing or household objects are normally not the subject of written texts or talked about in formal situations and there is thus little need to know the Standard Arabic words for them. While it is true that such word are found in Standard Arabic dictionaries, these are often not widely known, and speakers have to revert to dialectal terms if the need arises to refer to such objects in what is otherwise Standard Arabic speech. For words describing or expressing emotion or intimacy the detached, formal code of Standard Arabic is often felt to be inadequate or misplaced. Hence the often humorous effect when colloquial insults are translated with Standard Arabic equivalents in subtitled movies. Accordingly, scolding or

ad hominem attacks in debates tend to call for shifts to the dialect, as do jokes or talk of taboo subjects (Albirini 2011:55off). The importance of this for the present purposes is that corpus should therefore include speech on serious and abstract topics with minimal digression into the personal or mundane.

Furthermore, different learned discourses may be governed by different linguistic expectations and norms. The subject area of material included in the corpus is to be kept constant as far as possible to exclude this source of variation.

5.2.4 Native speakers

Standard Arabic is a language that is learned in formal education and has no native speakers as such. Being a native speaker of Arabic is to be a native speaker of one of the Arabic dialects. Since none of the dialects have a system of morphologically marked case similar to that of Standard Arabic, there is only a minimal difference in interference from the dialect on the use of case endings in speakers' spoken Standard Arabic. None of the dialects have morphological case that could be transferred (see 5.5 for further discussion).

5.2.5 Education

As a variety learned primarily through formal education, the level of competence in Standard Arabic is strongly correlated with the level of a speaker's education. The causal link between education and skills in Standard Arabic is not straight forward, however. Haeri (1997:passim) has described how primary education in Egypt is split between private 'language schools' where languages other than Arabic are used as the medium of instruction and Arabic is held in low regard, and public schools where all subjects are taught in Arabic. Persons of the upper and upper middle classes whose parents can afford to send them to private schools then often continue to receive their higher education either abroad or in academic fields that are taught entirely in English or French, all to the effect that "the higher ones social class, the less likely it is that one will learn [Standard Arabic] well" (Haeri 2000:68).

The strong link between levels of education and knowledge of the formal grammar of Standard Arabic, has, nevertheless, been demonstrated in a number of articles by Parkinson (1993, 1994b, 1996, 2003). Persons with at least secondary education have received formal instruction in Arabic grammar and have had to compose advanced texts in Standard Arabic to be judged and corrected. They have thereby at least had the opportunity to develop some measure of metalinguistic thought to be applied in linguistic self monitoring. This

ability hinders them from developing strategies for speaking Standard Arabic such as those observed with persons with low or no education, who insert Standard Arabic endings seemingly without relation to formal grammar (Parkinson 1994a:192).

5.2.6 Public figures

Being an educated native speaker of Arabic does not necessarily mean having developed oral skills in Standard Arabic. One might read and write Arabic extensively in one's profession or privately without ever being required to use it orally. Only persons appearing and speaking in public with some claim to authority, such as politicians and university professors, have a vested interest in developing skills in Spoken Standard Arabic. Montgomery (2008:261) defines public figures as persons that "hold institutional positions and by their official status are treated as 'having some locus' on the matter at hand." This definition is modified here to include persons who held institutional position also in the past. The importance of the concepts of public persons for the present study is that they are expected to have developed skills in oral performance of Standard Arabic.

5.3 Al Jazeera and *Liqāʾ al-Yawm*

The material used for this study was broadcast by Al Jazeera. Al Jazeera has been described as "the most powerful news-and-current-affairs channel in the Arab world" (*The Economist*, 12 January 2013). A Gallup poll conducted in 2002 in five Arab countries found Al Jazeera to be the preferred choice for television news in these countries, far ahead of its competitors.[48] The network was founded by the Qatari emir Ḥammād Bin Xalīfa in 1995 with capital in the form of a loan to the news network. When Al Jazeera did not turn a profit and could not repay the loan it was extended indefinitely. The channel recruited several employees of an Arabic language BBC news channel that closed down in 1996 after a brief existence, and thereby from its early history included staff trained in Western standards of journalism. Al Jazeera is generally regarded to have raised the bar for the quality journalism in the region and has stood out in dealing with topics that the mostly tightly controlled state-owned media in the region did not address. The controversial and hugely popular debate program *al-Ittijāh al-muᶜākis* 'The Opposite Direction' has been especially prolific in raising sensitive political, social, and religious issues. In this respect the program was instrumental in transforming Arabic satellite television into a

[48] *http://www.gallup.com/poll/5857/AlJazeera-Arabs-Rate-Its-Objectivity.aspx*

genuine public sphere by focusing on real, pressing issues and having them discussed in live television in a open and free manner (Lynch 2006:41). It regularly gives airtime to dissident and controversial political actors. Al Jazeera was, for example, the first Arab television channel to air Israelis speaking Hebrew (Miles 2005:37). This has led to numerous diplomatic crises between Qatar and other Arab states. Internal Qatari affairs, however, remain off limits for critique and discussion. The Qatari government and the representatives of Al Jazeera insist that editorial decisions are made with complete independence from the Qatari state, but critics claim that its international coverage is actively used as a propaganda tool by the Qatari government, particularly in conflicts in which the government is directly involved (Samuel-Azran 2013:passim).

Al Jazeera came to be widely known in the west after the 9/11 terrorist attacks when it received and broadcast video statements by the al-Qaida leader Usāma Bin Lādin and after its extensive coverage of the US-led invasions of Afghanistan and Iraq in the aftermath of the attack. Its role as an important source of global news was further established by its extensive coverage of the events of the so-called Arab Spring to which it had much better access than its competitors in international news. During these events, news on Al Jazeera was also made more accessible globally with its English language news channel that was launched in 2006.

Liqāʾ al-Yawm 'Todays Meeting' is a political interview program that is broadcast roughly weekly on Al Jazeera. Each episode features a guest somehow involved in Middle Eastern politics, leaders of a political parties, ministers, representatives of foreign powers or political bodies with involvement in the Middle East, or in some cases persons without political position but with first hand experience in some political event. There is no one single regular host in the program. Interviews are conducted by well known Al Jazeera news anchors or in remote areas by reporters. The format of the interview is a traditional 'news interview' (Heritage 1985:112; Clayman and Heritage 2002:passim) where the interviewer gives an introduction, addressing the viewers, and then proceeds to ask the interviewee a series of questions. Some guests are subject to very critical questions and interrupted and pressed to give answers, while others are allowed to speak more or less at their own leisure and to get their message across with few critical questions. Interviews also vary in length. The shortest of the episodes included in the corpus is 21 minutes in total broadcast time and the longest is 50 minutes.

5.4 Applying criteria

The criteria described above were applied to the original 118 episodes of *Liqāʾ al-Yawm* broadcast by Al Jazeera in 2010 and 2011. This temporary delimitation was arbitrarily chosen to encompass a large enough number of episodes that were recently broadcast when this project was started. The episodes in this timespan constituted a initial pool of episodes that were filtered through the criteria to be representative of spoken Standard Arabic. 73 episodes did not fit the criteria described above and were discarded. To avoid too much dialectal diversity that complicate analysis, three of the dialect groups of the remaining 45 episodes were chosen for inclusion, resulting in a corpus of 17 interviews.

The largest group of excluded material, 76 interviews, was that of non-native speakers of Arabic. Most of these interviews were broadcast with an overlay translation. Two were with nationals of non-Arab countries, a Turk and an Iranian, who were not translated but who do not speak Arabic natively. Another three were Iraqi Kurds who are likely to have Kurdish, and not Arabic, as their native tongue.

The interviews in *Liqāʾ al-Yawm* are in an every-day sense highly formal; they are discussions with prominent politicians about current affairs watched by a huge audience on one of the most respected Arabic news channels. In section 5.2 above, the four different definitions of formality were given above based on Irvine (2001): *increased code structuring*, *code consistency*, *invocation of positional identities*, and *central situational focus*. The spoken language in *Liqāʾ al-Yawm* is formal by all the these four definitions. First, as a news interview, the situation is regulated by an elaborate set of rules not found in normal conversation. Consider for example the simple fact that, in the news interview, only one person asks questions and only one provides answers, or that one of the participants begins the conversation with a lengthy introduction (Clayman and Heritage 2002:passim). A more subtle form of increased code structuring in news interviews is the lack of vocalized confirmations ('yeah', 'hmm' etc.) on part of the interviewer (Heritage 1985:99; Montgomery 2008:265). This is a notable feature in the corpus.[49] Second, formality in terms of code consistency is evident in the complete lack of code-switching with other languages, particularly English, that is otherwise common amongst educated Arabs. This is especially noteworthy since half of the speakers, nine of seventeen, have degrees from universities in English speaking countries.

Third, participants speak as political representatives or experts, not as pri-

[49] In Montgomery's four-part typology of the news interview, this is a characteristic feature only of the 'accountability interview' where the interviewee is held responsible for their actions or positions. All the interviews in the corpus fall into this category.

vate individuals. There is no attempt in the program to let the audience 'get to know' the interviewee or to temporally relax the atmosphere with more lighthearted questions and to approach the related genre of the 'chat show' (Clayman and Heritage 2002:96). Fourth, participants speak on a predefined topic as determined by the interviewer's questions. While it is a characteristic of the news interview that interviewees have strategies of avoiding answering some questions (ibid.:chap. 7), this does not normally take the form leaving the topic. In relation to the Labovian notion of formality, the interviews are obvious examples of *careful speech* since the situation is socially defined as an interview (Labov 1972:79).

The formality in the corpus should be consistent between interviews. The characterization given above holds for all the interviews. The introductions are overall very similar, structurally and linguistically; the presenter greets the audience, gives a brief background of the guest, greets the guest, and then proceeds to ask him or her the first question, all in very high level of Standard Arabic. While the format of all interviews is replicated, there is some variation in the physical setting. Some interviews are conducted in Al Jazeera's studio while others are in what appears to be official meeting rooms. Only one interview, with Riyāḍ al-ʾAsʿad, is in what is clearly a private living-room, with the interviewee seated on a sofa with a cupboard visible in the background behind the interviewer. Two interviews stood out in terms of the physical setting and were excluded. The first is an interview with Sādiq al-ʿAḥmar, head of the Yemeni Ḥāšid clan, that was conducted outdoors in a garden, probably at the interviewees home. The other is with the Palestinian Muhammad ʿAbd al-Raʾūf al-Mabḥūh who was sought after by Israeli authorities on terrorist charges at the time of the interview. The footage was taken ten months prior to broadcast and originally meant for a documentary according to the introduction on the episode, but was broadcasted in *Liqāʾ al-Yawm* on occasion of his assassination in Dubai by Mussad agents. The interview was conducted in a dark room without the interviewee revealing his face.

The level of formality is authentic in the sense that the guests in *Liqāʾ al-Yawm* are not responding to an explicit request to speak Standard Arabic, but are responding to the situation as such. This sets this corpus apart from experimental situations where participants in effect have to pretend to be an important person talking in a formal setting.

The topics covered in all of the interviews include politics and current affairs. There are brief excursions onto religious or academic discourses and these are typically present to build a political argument. All interviews deal with a combination of general and specific matters. Typically the guest is asked to

describe a general political situation and is then confronted with questions regarding his own or his party's or organization's statements, actions or opinions in the matter. The interviewers go straight to the question after the introduction and a very brief exchange of formal greetings. Most of the greetings are excluded as formulaic expressions (see 6.5). There is after this no further exchanges of pleasantries. Personal experiences and anecdotes are very rarely brought up by interviewees, and only to illustrate or exemplify the matter at hand. There is extremely little exchange not directly on topic.

The speech of the interviewees in the program is extemporaneous. Guests in the program can presumably foresee more or less which issues will be brought up and can to some extent plan their answers. The degree to which speakers are allowed to follow these planned answers varies between interviews however, depending on how much the interviewer intervenes to steer the conversation. On one end of the scale is the interview with Jamāl al-Xuḍarī who is allowed to speak freely on what his organization has done for the people in Gaza. On the other end is the interview with Burhān Ġalyūn who is frequently interrupted with critical questions, 33 times to be exact. No speaker uses memorized speech or manuscript delivery more than to make brief quotes that were excluded in the analysis (see 6.5). All but four interviews in the corpus were available as video where it could be ascertained that speakers are not reading aloud from written texts. Some of the interviewees do occasionally glance at papers in front of them, but not more than to suggest that these contain brief notes of talking points.

All the speakers in the corpus have at least a secondary education from Arabic countries and can therefore be assumed to have a working knowledge in formal Arabic grammar (see 4.3). All have university degrees, many from universities in non-Arab countries (see Appendix A). All interviewees included in the corpus are or have been public figures, most of them as politicians. By applying the criteria for inclusion in the corpus as described above, the original 118 interviews are reduced to 45.

5.5 Representation of dialects

Structurally, the dialects differ in how they relate to Standard Arabic. A linguistic feature that for a speaker of one dialect is a salient marker of Standard Arabic might not be so for a speaker of another dialect. An often quoted example is the demonstrative pronouns in Egyptian and Levantine Arabic. In Egyptian Arabic the demonstrative pronoun *da* (m.s.) appears after the noun. The Levantine Arabic counterpart *hāda*, or one of its variants, appears before the noun,

as is also the case with the Standard Arabic *hāḏā*. This is illustrated in (16).

(16) Egyptian: *il-bet da*
Palestinian: *hāda/ha- l-bēt*
Standard: *hāḏā l-bayt*
'This house'

When a speaker of Egyptian Arabic produces the Standard Arabic form, this involves a syntactic transformation of word order as well as a lexical substitution, and it is therefore perceived as a saliently Standard Arabic form. For speakers of Levantine Arabic, the standard form is achieved by a minor phonetic adaptation, and the form is therefore for them less saliently standard. Many other such examples could given from the phonetic, morphological, lexical, or syntactic levels of analysis.

When it comes to case endings, however, all Arabic dialects have the same structural relation to Standard Arabic; none of the dialects have a system of morphological case, whereas Standard Arabic does. Occurrences of case endings in speech are therefore a highly salient markers of Standard Arabic for speakers of all dialects. This makes it possible to compare material from speakers of different dialects in a study of the use of morphological case.

Differences between the Arabic dialects in linguistic structure may have an indirect effect on case marking. In the example with demonstrative pronouns above, a speaker of Levantine Arabic who wishes to clearly mark the sentence as Standard Arabic could add a case marker. Indeed, if their dialect produces *hāda l-bēt*, adding a case ending is the only option to mark the phrase as Standard Arabic, short of phonetic substitutions (*ē* to *ay* or *d* to *ḏ*). For a speaker of Egyptian Arabic on the other hand, the Standard Arabic *hāḏā l-bayt*, without a case ending, is already significantly different from their dialect, and they are in no need to add a case ending to mark the phrase as standard. Controlling for such effects would be an extremely complex endeavor and will not be attempted here. The effects are however probably fairly small and are here assumed not to have a significant affect on the use of case endings. The potential problem of a structural and indirect impact of dialectal differences in case marking were weighed against the aim of finding conventions of spoken Standard Arabic as a wider, regional standard language, an aim that requires comparisons of speech with different dialectal substrata.

What has to be taken into account in dialect variation, however, is how epenthesis in the dialects generate forms that are similar to case marked forms in Standard Arabic (see 7.7.1). Since the rules of epenthesis differ between di-

alects, this causes different patterns of overlap with Standard Arabic case endings that have to be accounted for separately for each dialect. The differences in epenthesis in the dialects mean that only material from speakers whose dialects have been thoroughly described will be included. Also, the number of dialects need to be kept limited so as not to make the problem of diverging rules of epenthesis overly complicated.

The 45 interviews that remained after applying the criteria represent fourteen nationalities. Nationality is here taken as a proxy variable for the dialect variety of the speaker. While this is a somewhat crude measure and dialects might, in some cases, be more appropriately divided by cities or smaller regions, it serves the purpose here as a practical way of dividing speakers in dialect groupings. The following list gives the number of interviewees from each country of the remaining after the criteria were applied as described in the previous section:

Algeria	2	Nigeria	1
Egypt	5	Palestine	7
Iraq	3	Sudan	2
Lebanon	2	Syria	5
Libya	10	Tunisia	3
Morocco	1	UAE	1
Mauritania	1	Yemen	2

The largest groups in the material after the criteria have been applied are Libyans (10), Egyptian (5), Palestinians (7) and Syrians (5). Speakers of Egyptian, Palestinian and Syrian Arabic were chosen for inclusion in the corpus. These are dialects that have been thoroughly described and with which the author is familiar. They add up to a total of seventeen interviews, which suites the time constraints of the project. These seventeen interviews form the corpus of this study.

5.6 The corpus

The total broadcast time of the seventeen episodes of *Liqāʾ al-Yawm* that were included in the corpus is 7 hours and 41 minutes. Of this only the speech by the interviewees is analyzed. This constitutes a total of 5 hours and 22 minutes of speech and around 38000 words, or 15000 annotated nouns and adjectives.

Summary information on the seventeen speakers in the corpus is given in Appendix A. The biggest weakness with this set of speakers is that they are

all males. In the episodes of the chosen period, 2010–2011, only two Arabic speaking women appeared; Maya al-Jarībī, general secretary of the Democratic Party in Tunisia (6 October 2010), and Dafna Birāq, a Turkish national and wife of a suicide bomber (5 February 2010). It is quite possible that the use of case marking is strongly gender coded and there would be no way of telling this from this corpus.

All speakers have some form of university degree and at least ten have a doctoral degree. Nine have a degree from a university in North America or Europe. Since they are persons with long careers behind them, the average age at the time of the interview is high, 65.1 yeas (SD=9.80).[50] The two oldest speakers are born in 1930 and were 80 and 81 years old when the interviews were recorded in 2010 and 2011. The youngest speaker, Riyāḍ al-Asᶜad was born in 1961 and was 50 years of age at the time of the interview.

5.7 Summary

This has explained how and according to what principles and from what material the corpus was constructed. Since no sizable corpus of formal and extempore Arabic speech was found, a new corpus was constructed specifically for this project. This corpus needed to comply with a set of criteria. The corpus should represent formal situations in which the speakers talk extemporaneously on similar and serious topics. The speakers should be native speakers of Arabic, educated, and public figures. Al Jazeera's interview program *Liqāʾ al-Yawm* provides a large amount of suitable material. From among an initial pool of episodes of this program from 2010 and 2011, seventeen were chosen for inclusion in the corpus. These are all interviews with Egyptians, Syrians, and Palestinians that fit the criteria. A weakness in this corpus is that it only includes male speakers.

[50] These numbers are calculated only on the year of birth, and not on date. For two speakers year of birth could not be established.

6 *Text preparation*

The empirical material of the study conducted for this dissertation consists of 17 episodes of the program *Liqāʾ al-Yawm* broadcast on the Al Jazeera television channel. The previous chapter described the principles behind choosing this program and the specific episodes for inclusion in the corpus. One reason for building the corpus from this particular material is that there are transcripts available published by Al Jazeera on their homepage. This chapter describes how the transcripts of these episodes were formatted and prepared for annotation and linguistic analysis. It is important to note that it is the audio from the broadcast and not the transcripts as such that are the source of data. The transcripts do however play an important practical role in the process of data collection. Having the Arabic text available drastically decreases the work-load of data collection as the basic work of transcription is already done and the work that remains consists of formatting, editing, and annotation. This makes it possible to work with a larger body of material than would otherwise be possible given the constraints of this project. The original transcripts are generally very accurate and represent many dialectal features and grammatical idiosyncrasies. However, there seems not have been any systematic policy with respect to transcription. At times the transcriber makes corrections to what is actually said and transcribes dialectal lexemes with its Standard Arabic counterpart or changes grammatical endings so as to be in accordance with codified grammar. In such cases the transcript was edited to match what was actually uttered.

Following this section, the CHAT format, the standard to which the transcripts were adapted, is first briefly described in section 6.1. For practical reasons, the texts were automatically transliterated to a system of Latin characters as described in section 6.2, and word stems of nominals were delimited from pre- and suffixes by principles explained in section 6.3. Section 6.4 describes segmentation on a higher level, namely that of conversational turn taking into

utterances. Some types of words are, for various reasons, problematic to include in a quantitative study of case marking in speech and were therefore marked up for exclusion from analysis. They are listed and described in section 6.5. Section 6.6 presents a summary of the chapter.

6.1 The CHAT format

In this section, the process of adapting the transcripts from Al Jazeera's web page to the CHAT (MacWhinney 2000:vol.i) is described. The CHAT format was originally developed for the study of child language but is well suited also for other forms of spoken language. It has a highly developed system for marking features typical of natural speech such as repetitions and syntactically incomplete utterances. More importantly it allows for annotation and analysis in the CLAN programs (ibid.:vol.ii). The availability of such an effective tool for manual annotation of the texts was crucial for the current project.

Figure 5 shows a minimal CHAT file. The file begins with a 'header' consisting of lines containing optional meta data such as the date of the recording, name, age, language, and other information on participants and their respective speaker code, `*ROS` and `*FAT` in the example. The header also contains a reference to the accompanying audio or video file. The lines containing the actual transcription, the so-called 'main tiers', begin with an asterisk and a speaker code and contain no more than one utterance (see 6.4 below). Each main tier ends with a time stamp referring to the beginning and end of the utterance in milliseconds in an accompanying media file. This makes the audio or video segment containing the actual utterance directly accessible from within the CHAT file. Each main tier is followed by optional 'dependent tiers'. These are lines beginning with `%` that contain various types of linguistic annotation. The dependent `%mor`-tiers in the example contains parts-of-speech tags. There may be several optional tiers for each main tier that contain different kinds of annotations or comments.

The main deviation from CHAT standards in this project is that all annotation, including morphological segmentation and morphosyntactic labels, is added in the main tier. The segmentation in tiers presupposes a work-flow where the transcript has been finalized before linguistic information is added. In the present project, a work-flow was developed where editing and annotation was done simultaneously making the division into main and dependent tiers impractical.

FIGURE 5: *Example of a CHAT file, adapted from an example distributed with CLAN.*

```
@Begin
@Languages: eng
@Participants: ROS Ross Target_Child @ID:
eng|samples|ROS|||||Target_Child|||
@ID: eng|samples|FAT|||||Father|||
@Media: clip, audio
@Situation: Breakfast table
*FAT: and <what do you mean> [//] what is it [//]
      what are you asking for ? •0_2312•
%mor: coord|and pro:wh|what cop|be&3S pro:wh|what
      aux|be&PRES pro|you part|ask-PROG
      prep|for ?
*ROS: alert [!] alert ! •2312_3715•
%mor: adj|alert n|alert !
*FAT: alert means like it's time for a fire alert
      . •3715_6937•
%mor: n|alert v|mean-3S conj|like
      pro|it cop|be&3S n|time prep|for det|a
      n|fire n|alert .
@End
```

6.2 TRANSLITERATION

One reason for choosing the program *Liqāʾ al-yawm* as material for this study was that there are transcripts available. The fact that these are orthographic transcriptions suits the needs of the project. A phonetically transcribed text would make much of the computer analysis extremely complicated. Lexical and morphological analysis via identification of character strings would be virtually impossible due to the many different ways one and the same lexical item can be pronounced by different speakers, or even by one and the same speaker.

It was found best to transliterate the Arabic transcripts to a Latin-based alphabetical system to avoid problems of text directionality. The transliteration system needed to fill two requirements. It had to have (*a*) a one-to-one relation to the Arabic characters, and (*b*) a high degree of human readability. The one-to-one relation was necessary to make sure that there is no loss of information

when the text is transliterated in order to allow for automatic transliteration and to retain the possibility of reverting the text to Arabic script after editing. The readability of the transliteration was important since most of the editing and annotation was to be performed on the transliterated text.

All commonly used systems for romanizing Arabic script are transcriptions and not transliterations as they represent phonemes and not Arabic graphemes. Systems such as that of Library of Congress that employ character combinations with *h* for fricatives, for instance *kh* and *th*, have a high degree of readability in vowelled texts, but lead to many ambiguities in unvowelled text since the *h* can also represent an independent phoneme. The requirement of a one-to-one relation also rules out systems in common use in the linguistic literature, including that of the *Encyclopedia of Arabic Language and Linguistics* (Reichmuth 2006), and of Deutsche Morgenländische Gesellschaft on which the former is based. Although they both have only one character for each phoneme, they do not make distinctions between the five different ways of writing the Arabic phoneme *hamza* (أ إ ئ ؤ ء) and the three ways of writing *alif* (آ ى ا). They also lack a character to represent *tāʾ marbūṭa* (ة).[51] The system used in French libraries, ISO 233, marks *tāʾ marbūṭa* and the various types of *ʾalif*, but not the five *hamza*s.[52]

The second requirement, that the text is to have a high degree of human readability, excludes the Buckwalter system, the standard system in Arabic computational linguistics.[53] This is a strictly graphemic transliterations designed for computational analysis, but its readability is hampered by only using ASCII characters such ⋆ for the letter *ḏāl* and & for the letter '*hamza* on *wāw*'. Whatever transliteration scheme is chosen, the result will be read only with some difficulty due to the unfamiliarity of unvowelled transliterated Arabic text.[54]

In the end, an extended form of the EALL system was developed for this project, with additional characters to unambiguously represent the five *hamza*s, the three *alif*s and *tāʾ marbūṭa*. This system will be referred to as EALLX, and is shown in Table 5. Vowels were added only as suffixes or as part of suffixes in so far as they are relevant to case marking. In addition to the three vowels

[51] *Tāʾ marbūṭa* is a suffix with the main function of marking feminine gender. It is represented in Arabic script with a special grapheme when word final and pronounced *-a* or *-at*.

[52] *http://guideducatalogueur.bnf.fr/ABN/GPC.nsf*

[53] *http://www.qamus.org/transliteration.htm*

[54] Very late in the this project, I discovered SAMPA, "a machine-readable phonetic alphabet" (*www.phon.ucl.ac.uk/home/sampa*) which has been adapted to represent Arabic. It uses combinations of ASCII characters and is highly readable. With some modifications it might have been a good option for this project.

Table 5: *eallx transliteration*

alphabetic				non-alphabetic[a]	
ا	ʼ	ض	ḍ	آ	ā
ب	b	ط	ṭ	ى	ý
ت	t	ع	ᶜ	ء	ɔ
ث	ṯ	ظ	ẓ	أ	á
ج	j	غ	ġ	إ	í
ح	ḥ	ف	f	ئ	ẏ
خ	x	ق	q	ؤ	ẇ
د	d	ك	k	ـَ	a
ذ	ḏ	ل	l	ـُ	u
ر	r	م	m	ـِ	i
ز	z	ن	n	ـً	aa
س	s	ه	h	ـٌ	uu
ش	š	و	w	ـٍ	ii
ص	ṣ	ي	y	n/a	e
				ـّ	n/a
				ـْ	n/a
				ة	p

[a] Not traditionally included in the Arabic alphabetical order.

of the Arabic script with their respective nunation forms, e was used to represent /ə/ whose phonemic status in Standard Arabic is ambiguous. The system uses Unicode character encoding.

6.3 Lemmatization

Word stems were separated from pre- and suffixes with a hash and a hyphen respectively. A 'stem' is here defined as a word without prefixes, a case ending, or an enclitic pronoun, and a word as a character string surrounded by white-space or the oral realization of that string. The purpose of lemmatizing

the transcripts was to isolate the case endings so as to make them analyzable according to their surface form and to make it possible to segment the data by lexical items.

Prefixes were delimited from the stem with a hash #. Possible prefixes on Arabic nominals are the following: (*a*) the conjunctions[55] *wa-* 'and', and *fa-*, sequential 'and'; (*b*) the prepositions *li-* 'for', *ka-* 'like', and *bi-* 'with' 'in', 'by', and (*c*) the definite article *al-*. All except the article consist of one Arabic letter. They can stack to a total of three prefixes, one of each type from the list above, and in that order, as in (17).

(17) *wa-bi-l-fiʿl*
and-in-DEF-deed
'and indeed' (al-Xuḍarī, 16:41)

The definite article combined with other prefixes was delimited with an automated process that identified word initial combinations of characters, such as ´l (ال), ll (لل), or w´l (وال) and excluded the relative pronouns *allaḍī*, *allatī*, etc. The delimiter added this way was then removed from a word in a list of false hits, such as *bāliġ* 'adult' and *iltibās* 'ambiguity'. Other prefixes were delimited during the course of the manual annotation.

Case endings and enclitic pronouns, if present, were separated from the stem with a hyphen (-) during the manual annotation. For words in the triptote, diptote, dual and the sound m.pl. paradigms the isolated stem is equivalent to the indefinite singular in pausal form (18). In plurals formed by infixes (19) and in sound f.pl. (20), the isolated stem is equivalent to the indefinite plural in pausal form. For words in the defective declension the final weak radical was also isolated from the rest of the stem (21). In these stems the last letter is, in some conditions, dropped, and excluding the final weak root makes the lexical item searchable (e.g. via the string q´ḍ) in all conditions.

(18) *al-filasṭīniyy-ūn*
DEF-Palestinian-MPL.NOM
´l#flsṭyny-wn (الفلسطينيون)

(19) *bi-ʾašyāʾ-in*
with-things-GEN
b#ášy´ʾ-ii (بأشياء)

[55] Prefixes *wa-* and *bi-* can also be 'particles of oath', but the functional distinction is not relevant for purposes of lemmatization.

(20) *al-filasṭīniyyāt-u*
DEF-Palestinian.FPL-NOM
´l#flsṭyny´t-u (الفلسطينيات)

(21) *al-qādī*
DEF-judge
´l#q´ḍ-y (القاضي)

The letter *tāʾ marbūṭa* (ة) is in Arabic orthography transformed into a regular *tāʾ maftūḥa* (ت) when the noun has an enclitic pronoun and *tāʾ marbūṭa* is no longer word final. In these cases, the *tāʾ maftūḥa* was reverted in the corpus to *tāʾ marbūṭa* when the enclitic pronoun was isolated. The transliteration of such words thus represent orthographically incorrect forms with non-final *tāʾ marbūṭa*, as exemplified in (22). This was done to make all stems with an original *tāʾ marbūṭa* identifiable.

(22) *kalimat-u-hu*
word-NOM-his
klmp-uhu
(كلمته)

6.4 Utterances

In the CHAT-standard, the text is divided into 'utterances'. Each utterance is represented by one line of text beginning with a speaker code. In the corpus, these are either *INT: for 'interviewer' or *INF: for 'informant'.[56] Lines marked this way as belonging to the interviewer were not further manipulated and are not part of the analysis. Each utterance represents one 'minimal terminable unit' or 't-unit' for short, defined as "one main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it" (Hunt 1970:4). This serves two purposes. First, it divides the text in segments of a suitable length for linkage with the media file, and, second, it allows for analysis of the case endings in pause position (in one of the several possible definitions of the term, see 4.5).

The t-unit is an alternative to the script-based concept of the sentence, defined as a stretch of speech which would end in a full stop in writing. The sentence is not a suitable unit for analyzing speech. Speakers for example often use

[56] There is no 'interviewee' speaker code in the CHAT standard.

conjunctions, typically 'and', between utterances to signal that they intend to keep their turn in the conversation, resulting in what would be extremely long sentences. The problem is further complicated in Arabic where conjunctions to a large extent replace punctuation as used in European languages.

The end of an utterances is in CHAT marked with a space followed either by a full stop or a question mark. Incomplete utterances are terminated with `+/.` for interruption, `+//.` for self-interruption, and `+...` for tailing off.

Since Arabic is a pro-drop language with optional pronominal subjects, a sequence of two verbs with only one explicit subject can be interpreted either as one and two main clauses. Example (23) could thus be interpreted in either of two ways. The first is as one clause with one subject and two verbs, giving the translation without the parenthesis. The other interpretation is as two main clauses, the second with no overt subject, giving the translation with the parenthesis. Such sentences were interpreted as one main clause unless there were prosodic cues or a pause to suggest a syntactic divide.

(23) *hāḍihi* *l-marḥala* ***ʾatat*** ***wa-ʾanhat*** *maẓāhir-a*
this DEF-phase came and-put.end.to.3FS manifestations-AMB
t-taqaddum
DEF-progress
'This phase came and (it) put an end to the manifestations of progress'
(Tayzīnī, 8:37)

Each utterance was linked to the corresponding segment of the accompanying media file. This was done using CLAN procedures that add a time stamp giving the beginning and the end time of the utterance in milliseconds at the end of each utterance. This makes it possible to directly access and listen to any given utterance in the transcript.

6.5 Internal exclusions

Some material in the interviews was for various reasons excluded from annotation and quantitative analysis. Such material was marked up using a set of 'final codes' () labels in square brackets inserted after word to be excluded. Segments longer than a word were marked for exclusion by enclosing it in angular brackets and adding the post code after the closing bracket. Final codes for internal exclusions are listed in Table 6 and described below. Only the first three of these are CHAT standards.

Some of the material excluded from analysis with final codes, notably proper names and numerals, are systematically unmarked for case by all speak-

Table 6: *Final codes for internal exclusions*

Retracing	
– without correction	[/]
– with correction	[//]
– with reformulation	[///]
Quotation	[”]
Formulaic expression	[F]
Proper names	[P]
Dialectal nominals	[D]
Numerals above 10	[N]
Unanalyzable	[U]

ers (see below) and their frequency is highly dependent on the topic. Sāʾib ᶜArīqāt, for instance, discusses negotiations between PLO and the Israel and relates detailed information on the size of areas for suggested land swaps and economic figures for the occupied territories. He makes use of numerals more than any other speaker in the corpus. Including numerals in the study, that no speakers marks for case, would unduly reduce Sāᶜib ᶜArīqat's proportional rate of case marking.

1. Retracing. Three types of retracing were marked up and excluded: retracing without change, retracing with correction, and retracing with reformulation. Retracing without change is simply a repetition of a word or phrase. Non-final iterations were excluded regardless of whether the repetition was an intentional rhetoric device, a filler in a thought pause, a hesitation, or some other type of disfluency. In retracing with correction, the speaker backtracks the sentence to change the form of a word to correct a grammatical mistake. In retracing with reformulation there is no apparent linguistic mistake to be corrected, but the speaker backtracks and reformulates the idea or abandons it for another one. Retracing with reformulation also includes false starts. On any of these types of retracing, only the final iteration was included and annotated.[57]

[57] Instances of retracing with a correction that include a change in a case marker were marked with so-called 'gems' (MacWhinney 2000:vol.i, 28, part.1) in order to be easily retrievable.

II. Quotations. The analyzed material is only to include extempore speech. Therefore, parts of the discourse that are word for word quotations were excluded from the material. There are examples in the corpus of quotations are from the Quran and sayings of the Prophet, as well as from books and statements of third person. Proverbs were also considered quotations, as well the 'mention', as opposed a 'use' of a word, such as the mention of the word *iṣlaḥāt* 'reforms' in (24). Furthermore, when an interviewee uses part of the wording of the interviewers question in a directly following answer, that segment of the answer was marked up as a quotation.

(24) *ᶜind-ha yumkin al-qawl **ʾiṣlāḥāt***
with-this 3MS.is.possible DEF-saying reforms
'Then it is possible to say "reforms".' (Kayālī, 6:43)

III. Formulaic expressions. Greetings and expressions of politeness or deference, or expressions of religious origin such as *al-ḥamd-u li-l-lāh* 'God be praised' were excluded. These formulaic expressions are pronounced with case markers also in the dialects and are frozen forms not part of an active case system.

Some speakers in the corpus use set phrases as fillers and produce them in the exact same way every time. Ḥātim ᶜAbd al-Qādir, for instance, says *bi-t-taʾkīd* 'for sure' 22 times in the interview, always without a case ending. This is for this speaker clearly a frozen phrase not affected by the case system. It was not excluded since it is a individual practice and not necessarily a reflection of wider linguistic usage.

IV. Proper names, titles and terms of address. Proper names often retain dialectal phonological features when used in discourse that is otherwise Standard Arabic and they are normally not inflected for case. The practice of not marking proper names for case is even endorsed by the otherwise conservative Arabic language academies in spite of traditional grammar (see note 4 on page 24). Since proper names are to a large extent excluded from the case system they were marked up and excluded.

Proper names were here taken to include also proper names of countries, cities, organizations, etc. after it was found in the preparatory studies that these also were systematically unmarked for case by all speakers. Proper names that are extended noun phrases are sometimes referred back to by a shortened form. One example of this is *al-majlis al-waṭanī* 'the National Council' being shortened to *al-majlis* 'the counsel' by Burhān Ġalyūn. Such shortened forms were not excluded since they are in and of themselves not proper names.

Also excluded were terms of address. Arab societies have highly developed systems of terms of address that index social relations and hierarchies or express attitudes towards the addressee. These terms are embedded in and derive their significance from their use everyday interactions (Parkinson 1985) within the domain of non-standard speech. Since Standard Arabic is not normally spoken in situations where social relations are maintained and constructed, it has no corresponding system of terms of address. Pronouncing a dialectal term of address with a case ending, and thereby marking it as Standard Arabic, would extract it from the contexts in which it derives its social meaning. The significance of this for our purposes is that if speakers appearing in the corpus wish or by social conventions are required to express respect via terms of address, they are bound not to mark them for case. This is clearly shown in the corpus by the word *ustāḏ*, a respectful term for addressing intellectuals. It is the preferred term for the interviewees to address the interviewer. The lexeme *ustāḏ* is uttered 41 times by interviewees in the corpus and is never marked for case. Example (25) is case in point. The sentence is utterly correct in Standard Arabic with all case and mood endings pronounced and with observance of pause rules with omission of the case ending in the final word. Only the term of address *ustāḏ* is not marked for case.

(25) *wa-taṣawwar* ***ʾustāḏ*** *ʾanna … ʾanna-ka tatruk-u*
and-imagine ustāḏ that … that-you 2MS-leave-IND
bayt-a-ka šahr-an dūna ʾiṣlāḥ-in wa-dūna htimām
house-ACC-your month-ACC without repair-GEN and-without care
'Imagine, *ʾustāḏ*, that … that you leave your house for a month without reparation and without care.' (Tayzīnī, 22:02)

When referring to third person, the same words are not technically terms of address, but might more appropriately be termed 'titles'. Titles were excluded when used in connection with the proper name of the person they refer to. There are 45 instances of the word *raʾīs* 'president' used in this way, exemplified in (26), where it is is not marked for case. This can be contrasted with (27), where the word *raʾīs* is not used in connection with a proper name and is marked for case. This latter use of titles were included for annotation.

(26) *ʾana* *bataʾammal* ***ʾar-raʾīs*** *maḥmūd* *ᶜabbās* *ʾan* *yaṣil* *ʾila* *ġazza*
I 1s.hope DEF-president Maḥmūd ᶜAbbās that 3MS-arrive to Gaza
'I hope that President Maḥmūd ᶜAbbās will arrive in Gaza.'
(ᶜAbd al-Qādir, 21:13)

(27) ***raʾīs*** *ʾal-* *ʾal-wilayāt-i* *l-muttaḥida* ***raʾīs-un*** *ḍaᶜīf*
president DEF- DEF-states-AMB DEF-united president-NOM weak
'The president of the … the United States is a weak president.'
(al-Qaddūmī, 16:38*)

v. Dialectal nominals. There is ample evidence that morphemes that are exclusively standard, such as case endings, are not used on dialectal stems (Schulz 1981:87; Owens and Bani-Yasin 1987:731; Hary 1996:80; Davies et al. 2013:340). Words with dialectal stems were therefore marked up and excluded. A stems was considered dialectal if it is (*a*) formed from a root that is exclusively dialectal, and (*b*) used with a meaning that is not associated with that stem in Standard Arabic. Put differently, only stems of a standard root used with a standard meaning were considered standard.[58] Only the underlying root was here taken into account and not its phonetic realization in the particular instance. This means that in this study a word can be realized within a wide range of phonetic variation and still be annotated for case marking as a Standard Arabic word. The Standard Arabic /q/, for example, is often realized as *ʾ*, a glottal stop, in the represented dialects. The standard word *qiṣṣa* 'story' would be included and annotated also when realized dialectically as *ʾiṣṣa*, even though it in this pronunciation would normally not appear with a case ending (e.g. *ʾiṣṣat-un*). This is a limitation of the present method. Low rates of case marking with certain speakers may be correlate with the use of dialectal phonemes in otherwise Standard Arabic words and this cannot be detected in this study.

The second part of the definition—Standard Arabic stems used with a non-standard meaning—applies for instance to the word *lāzim*. In Standard Arabic, this is an adjective meaning 'necessary', but in the dialects it is used as a modal auxiliary 'have to' and, when it is used in this way, it is marked up in the corpus as dialectal. Another frequent example is *wāḥid*, which in Standard

[58] This is also the approach in *A Frequency Dictionary of Arabic* where "dialect labels were applied to lemmas that were exclusively dialectal, as well as lemmas that were primarily MSA but also manifest certain dialect-influenced uses that show up in the data" (Buckwalter and Parkinson 2011:7).

Arabic is the numeral 'one', and in several dialects is an indefinite pronoun 'someone'. When used in the latter meaning, it was marked up as dialectal and excluded. In difficult cases Wehr (1994), Parkinson (2006), and native speaker informants of the respective dialects were consulted.

This narrow definition of dialectal stems is ultimately mandated by the orthographic transcription that does not record fine grained phonetics. The narrow definition, however, has the practical benefit of avoiding time consuming auditory judgments to be made on individual words. A phonetically based definition of dialectal as standard words would be complicated to operationalize. It is not clear how one could define a point in the phonetic variation of a word where it crosses the line between being standard in the sense that it can take standard morphemes, and becomes dialectal, and cannot take a standard morpheme. That point might also differ between dialects, individuals, or even between lexemes. Hary (1996:80), for example, reports mixed answers from informants on whether *raʾēt-hu* 'I saw him' a standard stem with a dialectal pronunciation and a standard pronominal suffix is an acceptable form. There is also the case of prescriptively non-standard regional pronunciations that in their respective regions are accepted as standard. A characteristic of Egyptians Arabic is the realization of /j/ as [g] instead of the standard [dʒ]. The Egyptians in the corpus more or less consistently use this Egyptian pronunciation and often combine it with case endings. See, for instance (109) on page 206. Another such example that features in the corpus is the pronunciation of the Standard Arabic interdentals /ṯ/ and /ḏ/ as the sibilants [s] and [z] respectively. Operationalizing a phonetic definition of dialectal stems is, in other words, a deeply complicated endeavor. It was discarded for the more blunt but more practical lexico-semantic definition given above.

Also marked as dialectal are words with prefixed *ʿal-* or *hal-*. In both Egyptian and Levantine Arabic, the preposition *ʿalā* 'on' and the definite article *(a)l-* are often contracted to the combined prefix *ʿa-l-* (28). Similarly, the demonstrative pronouns, in Standard Arabic *hāḏā* (ms.) or *hāḏihi* (fs.), are contracted with the article to *ha-l-*. Nouns with either of these prefixes were tagged as dialectal.

(28) *ʾinta* ***ʿa-l-ʾaqall*** *ṭaraḥt* *xamaṣṭʿašar* *qaḍiyya* *fī* *hāḏihi*
you on-DEF-least 2MS.put.forward fifteen issue in this
l-mudāxala *ya* *ʾax-i* *mḥammad*
DEF-interjection VOC brother-my Muḥammad
'You have put forward at least fifteen [different] issues in this interjection, my brother Muḥammad.' (Šallaḥ, 19:36)

VI. Numerals above ten. The Arabic numeral system as described in traditional grammars is notoriously complex. Briefly, the numerals one and two are adjectives and therefore appear after the counted noun and agree with it in case and gender. Three to ten have the opposite gender to the counted noun and appear before it. The noun is in plural genitive. Eleven to nineteen are compound words with the decade agreeing in gender with the counted noun while the unit has the opposite gender. Both parts are indeclinable for case (except for in the number twelve) and the following counted noun is in singular accusative. For even hundreds, thousands, etc., the counted noun is in genitive singular. The dialectal numbers are far less complicated and it is normally these that are used in spoken Standard Arabic, at least for numbers above ten (Harrell 1964: 49; Bateson 1967:86; Diem 1974:47; Badawi et al. 2004:256). Numbers above ten were therefore marked up and excluded from annotation. Included are, however, the numbers *miʾah* 'hundred' *alf* 'thousand', *milyūn* 'million', and *milyār* 'billion' in their singular and plural forms when not part of compound numbers such as 'three hundred'.

VII. Unanalyzable. Natural speech often generates sequences that are not analyzable by traditional means. Such unanalyzable material was excluded from annotation. This is the case, firstly, in slips of tongue causing grammatical oddities, as in (29). Here the speaker probably meant to say *wazīr al-xārijiyya* 'the minister of foreign affairs' which is a genitive construction and the word *wazīr* should therefore not have the definite article. Here it is pronounced with the article and it is therefore unclear how it is to be annotated for definiteness. It is also not clear if the expression is to be regarded as a genitive construction, and therefore case assignment becomes problematic, which also extends to the dependent adjective *al-turkī*. The whole phrase was marked up as unanalyzable.

(29) *naḥnu ʾayḍan qābalna yaᶜni duᶜīna ʾila liqāʾ maᶜa*
we also met.1PL PART invited.1PL.PASS to meeting with
l-wazīr-a l-xārijiyya ʾal-turki[59]
DEF-minister-AMB DEF-foreign.affairs DEF-Turkish
'We also met with and were invited to a meeting with the Turkish minister of foreign affairs.' (Ġalyūn, 16:38)

Constituents also become unanalyzable when parts of the sentence are inaudible due to unclear pronunciation, interfering sounds, or occasional poor sound quality. It is then often difficult to determine the syntactic positions of

[59] The *l* of the article in *al-turki* was in this case not assimilated to the following 'sun-letter'.

surrounding words. In (30) *xxx* represents inaudible material and words in boldface were marked as unanalyzable.

(30) *wa-lā tabqa baqiyya kāfiya* ***li-binā'*** *xxx* ***aṯ-ṯawra***
and-not 3FS.remain remnant sufficient for-building xxx DEF-revolution.
'And what remains is not sufficient to build xxx the revolution'
(Abū Majd, 7:08)

6.6 Summary

The 17 interviews in the corpus were formatted according to CHAT standards. To make the texts more easily able to be analyzed and to ensure that they are displayed consistently in different software, the transcripts were transliterated with a system designed to be both readable and to represent all Arabic graphemes. Prefixes and suffixes were delimited from the stem in all words to be annotated. This was done to make tokens with specific stems as well as specific forms of pre- and suffixes searchable. Each speaker transcript was segmented into utterances with each utterance on one line beginning with the speaker code. Utterances were linked to the corresponding segment of the accompanying media file. This makes the appropriate part of the audio or video file accessible from any given point in the corpus. Finally, certain classes of words, including proper names and numbers above ten, that are rarely if ever marked for case, were marked up in the corpus texts to be excluded from analysis.

Morphosyntactic annotation

7

This chapter describes the details of the coding scheme applied in the annotation of the corpus text. A *label* consisting of a string of characters encoding morphosyntactic features was added to each token in the corpus. The label consists of a series of *tags*, each encoding one particular feature. The most important feature encoded in the label is if and how the case ending was pronounced by the speaker. These labels, together with the token words, provide the data to by analyzed in Part III.

In this chapter, general methodological issues regarding the coding scheme are discussed in section 7.1. This is followed by a description of the concrete manifestation of the label as a string of characters in section 7.2. The main part of this chapter, sections 7.3–7.7, consists of definitions, descriptions, and examples of the linguistic features encoded in the annotation scheme. These are: headedness in section 7.3, inflectional paradigms in 7.4, definiteness in 7.5, case governance in 7.6, and case marking in section 7.7, where also the methodologically thorny issue of ambiguous endings is discussed. Section 7.8 describes some measures taken to reduce the number of errors in the corpus and how the annotated corpus text was used to generate a database for the analysis of case marking. Section 7.9 is a summary of the chapter.

7.1 General considerations

The point of departure when developing the coding scheme for the corpus was to encode as many grammatical features as practically possible that may affect the use of case endings in speech and then test which of the encoded variables correlate with case marking. This study is thus to a large extent exploratory, and statistical results should be interpreted from this perspective. Although several hypothesis will be formulated and tested within this larger exploratory approach.

7.1.1 Underlying principles

The first and most important aim of the coding scheme is to account for grammatical parameters that govern the surface form of the prescribed case morphemes. It should, in other words, be possible to deduce the prescribed case ending for a particular token by looking at the label alone. This can be achieved by encoding three features: case position, inflectional paradigm, and definiteness. For example, a noun in the genitive (case position) that is a triptote (paradigm) and has the definite article (definiteness) can only ever have the prescribed case ending *-i*. With three tags for these three parameters, together with a tag on how the case ending is realized by the speaker in this particular token, the morphological distribution of case endings in speech can be quantitatively analyzed. Apart from these tags, information on whether the token is a head noun or attribute and by what grammatical rule it is governed for case was encoded in the label, giving a total of six tags.

Several variables that could be trivially derived from the corpus were not encoded in the label but were instead generated together with the database as described at the end of this chapter (see 7.8). Such variables include the presence of *tāʾ marbūṭa* in the word stem, and whether the token occurs at the end of a utterance.

The coding scheme should also be as exhaustive as possible, meaning that all individual nominals in the corpus should fall into one of the predefined categories. Exhaustiveness was only practically possible for Standard Arabic constructions. Developing categories for all dialect constructions would be a very complicated endeavor because of the variation both between and within dialects. However, constituents in the most frequent non-standard structures were assigned a tag and were annotated (see 7.6.1). The annotation of these tokens is relatively uninteresting from the perspective of case marking, since they are consistently unmarked for case (see 10.3). Having them annotated and therefore quantifiable means however that they can be utilized as a rough measure of the amount dialectal syntax speakers employ in their speech.

The number of possible values for each variable encoded in the labels had to be kept limited so as not to make the coding scheme overly complicated. Some categories were therefore given an intentionally broad definitions with a range of possible, but unspecified, subcategories. The most notable examples of this are the case governance categories of adverb and miscellaneous accusative (see 7.6.3). This avoids having long lists of categories that apply only to a few actual tokens in the corpus.

Practical considerations thus put limits on the level of detail that can be captured by the coding scheme. However, once the corpus is finalized the coding

scheme could, with limited workload, later be expanded to cover deeper levels of detail in areas of specific interest. For instance, all instances of the 'adverb'-tag could be located and replaced with tags for different types of adverbs.

7.1.2 Grammatical description

The categories encoded in the label should conform with conventional descriptions of Standard Arabic grammar. This makes the results comparable with other studies and accessible to educators and linguists with different backgrounds. The categories developed here are primarily based on Badawi et al. (2004) and Ryding (2005), both standard scholarly reference grammars of Modern Standard Arabic. They represent what might loosely be labeled the 'Western' tradition of Arabic grammatical description. This tradition has its roots in the late 19th century works whose authors distanced themselves from grammatical thought inherited from the Arabic classical tradition and instead aimed to base their description on observed usage (Killean 1984:228). Traditional Arabic Grammar is nevertheless very much present in grammars in the Western tradition. This is especially clear in the description of grammatical structures that have no clear parallel in European languages and for which there is no established European terminology. In these cases, the Arabic terms are often borrowed from Traditional Arabic Grammar, directly or as calques, to the effect, according to (Van Mol 2003:152), that "almost all contemporary [Western] descriptions use both western and Arabic terms together, so that we might in a certain sense use the term 'traditional mixed grammar'." This accurately describes the consulted grammars.[60]

In contrast to the Western tradition that aims at contemporaneity, Traditional Arabic Grammar is based as a matter of methodological principle on historical data (see 3.3) and includes rules and categories in which some Modern Standard Arabic constructions do not fit comfortably. This is the system through which native speakers in the Arab world study Arabic and it is therefore the system used for metalinguistic thought and self-monitoring, and through this it may affect how case endings are produced in speech. Categories described in this section will therefore be related to their counterparts

[60] Badawi et al. (2004) and Ryding (2005) both base their descriptions on contemporary data. Still, one finds in them descriptions of constructions that are clearly not attested in their data. (See footnotes 75 and 66 for examples.) Badawi et al. (2004:3) write that "the logic of corpus-based grammar requires that missing items are deemed non-existent", but continues to state that this is unacceptable to the educated native speaker, and that in such cases they draw on the "traditional and contemporary literary and linguistic competence" of the Egyptian co-author. This approach risks reproducing constructions inherited from Traditional Arabic Grammar that are obsolete in Modern Standard Arabic and which adds undue complexity to the description.

in Traditional Arabic Grammar where this differs significantly from Western tradition. Such differences are primarily found in accounts of case governance. The primary source on the Traditional Arabic Grammar used in this study are Wright (2011), originally published 1859, and Ibn ᶜAqīl's (d. 1367) commentary on *al-Alfiyya* (Ibn ᶜAqīl 2005), a grammar written in verse by Ibn Mālik (d. 1274). It is highly representative of the canonical tradition at large and widely used as teaching material in university level language instruction in the Arab world (Bohas et al. 1990:16).

7.2 Formal aspects of annotation

As mentioned above, the labels attached to tokens consists of six tags, each encoding one grammatical feature. This is exemplified in (31a), which is an example of the transcription and coding scheme as applied to (31b). The tags are separated from the transcribed word with a dollar sign (`$`). They take the form of a string of three uppercase letters and are separated from one another by a colon. The case and the case governance tags are separated by a forward slash to make these two tags more visually connected in the label. This format makes for a verbose coding, with each label consisting of 24 characters, including delimiters. It does however have the advantage of being relatively readable since all tags are in effect abbreviations of grammatical terms.

(31) a. `b´l#tákyd$HED:GEN/PRE:TRI:ART:UNM`

b. *bi-t-taʾkīd*
with-DEF-certainty
'Definitely' (ᶜAbd al-Qādir, 20:32)

The first tag after the dollar sign in (31a), `HED` indicates that the word is not an attribute. `GEN` indicates that it is in genitive position and `PRE` that it is in genitive position as a prepositional complement. `TRI` is the tag for the triptote paradigm. `ART` indicates that the token has the definite article, and `UNM` that it was pronounced by the speaker as unmarked for case, with no vowel or other ending that could be interpreted as a case marker. The complete tagset is given in Table 7 on the facing page. The annotated features are described in detail in sections 7.3–7.7 below.

Labels were manually inserted in the corpus using the CLAN software's 'coder mode' (MacWhinney 2000:vol.i, 23, part ii). In coder mode, tags are inserted using a menu system that prompts the annotator to choose a tag from a predefined set. This system was structured hierarchically so that only tags applicable to the active position in the tag sequence could be chosen and inserted.

Table 7: *Annotation tagset*

CASE		HEADEDNESS	
Nominative	NOM	Head noun	HED
Genitive	GEN	Attribute	ATT
Accusative	ACC		
CASE GOVERNANCE		**PARADIGM**	
Nominative		Triptote	TRI
Subject		Diptote	DIP
— of vso	VSO	Sound f.pl.	SFP
— of svo	SVO	Sound m.pl.	SMP
— of unfinished clause	UNF	Dual	DUA
		Defective	DEF
Topic	TOP	Final *alif*	ALF
Comment	COM	The five nouns	FIV
Dialectal compl.	FII		
Accusative		**DEFINITENESS**	
Object	OBJ	Definite article	ART
Absolute object	ABJ	Construct state	
Compl. of *ʾinna* and her sisters	INN	— noun or clause	CON
		— enclitic pronoun	PRO
Absolute negation	NEG	Indefinite	IND
Compl. of *kāna* and her sisters	KAN		
		MARKING	
Number spec.	TMZ	Marked	MAR
Adverb	ADV	Unmarked	UNM
Miscellaneous acc.	MIS	Ambiguous	AMB
Genitive		Inaudible	XXX
Annexed	ABS	Not applicable	NOC
Prepositional compl.	PRE		

7.3 Headedness

Nominals that function as attributes to a head noun and agree with it in case were tagged as attributes. These are typically adjectives (32), but also quantifiers such as *kull* 'all'; 'whole'; 'every' (33) and adjectival determiners (Badawi et al. 2004:223) such as *ġayr* 'un-' or *šibh* 'semi-' (34). Tokens that are not attributes were tagged as 'heads'.

(32) *hunāk* *tadaxxulāt* ***ʾiqlīmiyya***
there interferences regional
'There are regional interferences.' (ᶜAbd al-Qādir, 19:48)

(33) *wa-kāna* *ḏālika* *mawjūd-an* *fi* *t-tārīx* ***kull-u***
and-was.3MS that present-ACC in DEF-history whole-its
'And that has been present in all of history' (Mursī, 6:52)

(34) *wa-hunālika* *taḥāluf* ***ġēr*** *…* ***ġēr*** *muᶜlan* *bayna* *l-burjuwāziyya*
and-there alliance not … not announced between DEF-*bourgeoisie*
wa-n-nizām
and-DEF-regime
'And there is an un… unofficial alliance between the *bourgeoisie* and the regime' (Kayālī, 12:50)

7.4 Inflectional paradigm

Tokens were tagged as belonging to one of eight different inflectional paradigms. These are listed in Table 8 on the facing page in definite, construct state, indefinite, and pausal forms. The paradigms only partially correspond to grammatical number: sound m.pl., sound f.pl. and dual represent their respective numbers, the five nouns are always singular, and words in the other paradigms include both singular and plural nouns. One and the same lexeme can belong different paradigms in different situations, depending on how it is inflected for number and gender. The eight paradigms are the following:

1. Triptote. The vast majority of nouns have different endings for the three cases and are known as triptote nouns. In this paradigm, only the indefinite accusative ending *-an* is orthographically marked with a letter and appears in writing. The dialect form has a null ending, identical to the Standard Arabic pausal form.

Table 8: *Case paragms*

	DEFINITE	CONSTRUCT	INDEFINITE	PAUSAL
	I. TRIPTOTE			'letter'
NOM	*al-ḥarf-u*	*ḥarf-u*	*ḥarf-un*	*ḥarf*
ACC	*al-ḥarf-a*	*ḥarf-a*	*ḥarf-an*	*ḥarf*
GEN	*al-ḥarf-i*	*ḥarf-i*	*ḥarf-in*	*ḥarf*
	II. DIPTOTE			'bigger (m.)'
NOM	*al-akbar-u*	*akbar-u*	*akbar-u*	*akbar*
ACC	*al-akbar-a*	*akbar-a*	*akbar-a*	*akbar*
GEN	*al-akbar-i*	*akbar-i*	*akbar-a*	*akbar*
	III. DEFECTIVE			'judge'
NOM/GEN	*al-qāḍī*	*qāḍī*	*qāḍin*	*qāḍī*
ACC	*al-qāḍiy-a*	*qāḍiy-a*	*qāḍiy-an*	*qāḍī*
	IV. THE FIVE NOUNS			'father'
NOM	*al-ab-u*	*ab-ū*	*ab-un*	*ab*
ACC	*al-ab-a*	*ab-ā*	*ab-an*	*ab*
GEN	*al-ab-i*	*ab-ī*	*ab-in*	*ab*
	V. FINAL *ALIF*			'bigger (f.)'
NOM/ACC/GEN	*al-kubrā*	*kubrā*	*kubrā*	*kubrā*
	VI. SOUND FEMININE PLURAL			'Egyptians'
NOM	*al-miṣriyyāt-u*	*miṣriyyāt-u*	*miṣriyyāt-un*	*miṣriyyāt*
ACC/GEN	*al-miṣriyyāt-i*	*miṣriyyāt-i*	*miṣriyyāt-in*	*miṣriyyāt*
	VII. SOUND MASCULINE PLURAL			'Egyptians'
NOM	*al-miṣriyy-ūna*	*miṣriyy-ū*	*miṣriyy-ūna*	*miṣriyy-ūn*
ACC/GEN	*al-miṣriyy-īna*	*miṣriyy-ī*	*miṣriyy-īna*	*miṣriyy-īn*
	VIII. DUAL			'Egyptians'
NOM	*al-miṣriyy-āni*	*miṣriyy-ā*	*miṣriyy-āni*	*miṣriyy-ān*
ACC/GEN	*al-miṣriyy-ayni*	*miṣriyy-ay*	*miṣriyy-ayni*	*miṣriyy-ayn*

II. Diptote. Some nouns, known as diptotes, deviate from the triptote paradigm in that the indefinite form does not take nunation and only has two forms, with the accusative and genitive both taking the ending *-a*. As with triptotes, the dialect form is a null-ending, which is identical to the Standard Arabic pausal form.

III. Defective. Participles and some patterns of broken plural formed from a root with a weak final radical (*wāw* or *yāʾ*) are inflected in accordance with the defective paradigm. In indefinite nominative and genitive the weak radical is dropped and replaced with the nunation *-in*. When definite or in construct state the weak final radical appears as *-ī*. It can then only take the accusative ending *-a*, while genitive and nominative case is not marked. The dialect form is a final *-ī* in all situations. Defective participles marked for feminine with *tāʾ marbūṭa* are inflected according to the triptote paradigm and were tagged as such.

IV. The five nouns. The words *ab* 'father', *ax* 'brother', *ḏū* 'owner', *fam* 'mouth', and *ḥam* 'father in law' are known as the five nouns.[61] Only the first three of these five nouns were found in the corpus. The case inflection of the five nouns deviates from the triptote nouns in that the case ending takes the form a long vowel in construct state, *-ū*, *-ī* or *-ā*, for the three cases respectively. The dialectal form has a final *-ū* in construct state regardless of case, and a null ending in absolute state.

V. Final alif. Stems with a final long *ā* represented in writing by the letter *alif* are not marked for case. The consulted grammars make a distinction between words where the final *-ā* represents an original weak radical, as in *hudā* 'guidance', and words where the final *ā* has been added in word formation, as in *kubrā* 'bigger (f.)'. The former is said to take the nunation *-an* in all cases, whereas the latter ends in *-ā* (Wright 2011:240; Badawi et al. 2004:52). No example of a stem with final *alif* pronounced with nunation was found in the corpus. Either way, they are not marked for case. No distinction was therefore made between the two paradigms of final *alif*.

[61] Sources on Traditional Arabic Grammar mention a sixth word in this category, *han* that are said to refer to taboo subjects in general or to female genitalia in particular. This sixth word is reported as being declined as the five nouns by the early grammarians' Bedouin informants. Some works in Traditional Arabic Grammar, for example Ibn Hišām's (d. 1359) *Šarḥ qaṭr an-nadā wa-ball aṣ-ṣadā* (1990:37), therefore refer to this category as 'the six nouns' (*al-asmāʾ as-sitta*). 'The five nouns' is however the dominant term in pedagogical grammars and is the term inherited by the Western tradition.

VI. Sound f.pl. The number suffix *-āt* is added to the stem in the sound f.pl. paradigm,[62] replacing *tāʾ marbūṭa* if present in the singular. The case marker is added to this ending, with syncretism between the accusative and genitive forms in *-i*, with or without nunation. The dialectal form is *-āt*, identical to the Standard Arabic pausal form.

VII. Sound m.pl. Case and number are encoded in one and the same suffix in the sound m.pl.; *-ūna* for nominative and *-īna* for genitive and accusative, both marked in writing. The dialectal form is *-īn*, which is identical to the Standard Arabic genitive/accusative endings in pausal form.

VIII. Dual. Case and number are encoded in the same suffix also in the dual: *-āni* for nominative and *-ayni* for genitive and accusative, both marked in writing. The dialect form is *-ēn*, which is similar to the Standard Arabic genitive/accusative pausal form (see 7.7.1).

7.5 Definiteness

As shown in Table 8 above, the form of the case ending differs with definiteness in several paradigms. Tokens were therefore coded for definiteness as either (*a*) indefinite; (*b*) definite; (*c*) construct state with annexed noun or clause (CS-N/C), and (*d*) construct state with an enclitic pronoun.

I. Definite article. Nominals in absolute state are marked as definite with the prefixed definite article *al-*.

II. Indefinite. The main morphological characteristic of indefinite nominals is that they take nunation in the two most frequent paradigms, triptote and the sound f.pl. If there is no case ending on which to attach the nunation then there is no overt morphological marker of indefiniteness. This is illustrated in (35) where the noun *amr* 'issue', 'matter' is marked for genitive case and as being indefinite with *-in*, whereas the adjective *kabīr* 'big', 'major', which is inflected by the same paradigm is unmarked for case and definiteness.

(35) *wa-l-ʾān naḥnu našku min … min* ***ʾamr-in*** ***kabīr***
and-DEF-time we 1PL.complain from … from issue-GEN major
'And now we complain about … about a major issue.' (Tayzīnī, 25:04)

[62] The 'sound' plural paradigms, as also the dual, are formed by adding a suffix to singular form. This is in contrast to the 'broken plural', forms produces by changing the vowel pattern internal to the stem of the singular.

III. Construct state with annexed noun or clause (CS-N/C). Nouns in construct state were given one of two different tags depending on the kind of constituent that is annexed to it, that is what form the possessor takes. A nominal in construct state with an annexed noun (36) or clause (37) were tagged as CS-N/C and thus distinguished from nouns annexed with an enclitic pronoun, described below.

(36) *bi-**quwwat-i** s-silāḥ*
with-force.AMB DEF-weapon
'with military force' (Ġalyūn, 0:53)

(37) *wa-lākin bi-**šarṭ** ʾan lā taqif ʾal-intifāḍa ʾiṭlāq-an*
and-but on-condition that not 3FS.stop DEF-uprising absolute-ACC
'but on the condition that the uprising does not stop under any circumstances.' (Kayālī, 20:15)

CS-N/C differs from the other types of definiteness in that it can be syntactically either definite or indefinite. This is illustrated in (38) where *nuqṭat* 'point' is in CS-N/C. It is syntactically indefinite since the annexed noun *taḥawwul* 'transition' is indefinite as it does not have an accompanying definite article. The fact that *nuqṭat* is syntactically indefinite is shown in that the adjective attributed to it, *ʿarabiyya* 'Arabic', is indefinite. In this example, *nuqṭat* was tagged for definiteness as CS-N/C and its adjective *ʿarabiyya* as indefinite.

(38) *kān hunāk yaʿni **nuqṭat** taḥawwul **ʿarabiyya***
was.3MS there PART point.F transition[.M] Arabic.F
'There was, so to speak, an Arabic point of transition.' (ʿArīqāt, 23:30*)

Coordinated nouns in construct state, as in (39a), are disallowed in Traditional Arabic Grammar (Wright 2011:vol.ii, 201).[63] It is often commented on as a widespread linguistic mistake. The prescribed construction is instead to have one of the heads after the genitive construction with an enclitic pronoun referring to the annexed noun. The prescriptively correct formulation of (39a) is given in (39b).

[63] The construction, known in Arabic as *iqḥām* (lit. 'intrusion'), is not explicitly mentioned in Arabic grammars, but as pointed out in Gully (1993:24), it is also not given as a viable alternative.

(39) a. *maṭālib* *li-ḍimān* ***ḥurriyyat*** ***wa-nazāhat-i*** *l-ʾintixābāt*
demands for-guarantee freedom and-integrity-AMB DEF-elections
'demands to guarantee the freedom and integrity of the elections'
(al-Barādiᶜī, 4:01)

b. *maṭālib* *li-ḍimān-i* ***ḥurriyyat-i*** *l-ʾintixābāt-i*
demands for-guarantee-GEN freedom-GEN DEF-elections-GEN
wa-nazāhat-i-hā
and-integrity-GEN-its
'demands to guarantee the freedom of the elections and their integrity'

Multiple heads in genitive constructions are however common in Standard Arabic, both in writing (Blau 1973:184; Gully 1993:23-30) and in speech as represented by tokens found in the present corpus. Multiple heads in genitive constructions were all annotated and tagged as CS-N/C, since it does not affect case assignment or the prescribed form of the case ending.

iv. Enclitic pronoun. The second type of construct state is where the annexed constituent is an enclitic pronoun (40). The reason for separating CS-N/C and construct states with an enclitic pronoun is that in the latter, the case ending is not in word-final position, but between the stem and the enclitic pronoun. Not producing the case marker in words with an enclitic pronouns results in prescriptively incorrect forms. This is quite different from nouns in CS-N/C where unmarked morphological case is identical to a Standard Arabic pausal form (see 4.5).

(40) *miṣr-i*[64] ***ʾumm-u-na***
Egypt-AMB mother-NOM-our
'Egypt is our mother.' (Badīᶜ, 35:32)

7.6 Case governance

Case governance is represented in the coding scheme as two tags: the first specifies the case as nominative, genitive or accusative, and the second the rule by which this case is assigned. The case tag is determined by encoded case governance and is thus analytically superfluous. It is included in the coding scheme

[64]The *i* in this position is an Egyptian epenthetic vowel and therefore an ambiguous case marker (see 7.7.1).

for practical reasons; the governance rules are too many to be quickly scanned by the human eye when it appears on the computer screen as a menu list during annotation. The role of the case tags is to split this list into three parts, so that after the case tag has been set, it limits the choices for the following case governance tag to those that are applicable to that particular case.[65] In the descriptions below the categories for case governance are listed under the headings of their respective case.

7.6.1 Nominative

The primary function of the nominative case is to mark the subject. The subject of a clause was tagged as belonging to one of three different types, depending on the type of clause it appears in, VS, SV, or TC (topic/subject-complement). Traditional Arabic Grammar considers the first two to be fundamentally different from one another whereas grammars in the Western tradition typically do not.

In Traditional Arabic Grammar, equational clauses and SV-clauses are jointly termed *jumla ismiyya* 'nominal clause' (clause beginning with a nominal). These are regarded as fundamentally different from VS-clauses or *jumla fiʿliyya* 'verbal clause' (clause beginning with a verb). The Arabic term for the subject in an equational or a SV-clause is *mubtadaʾ* (lit. 'starting point'). It is analyzed as being governed for nominative by an implicit constituent preceding it termed *ibtidāʾ* (lit. 'beginning'). The subject in VS is considered to be a completely different constituent, termed *fāʿil* 'agent', and governed for nominative by the preceding verb.

In Western grammars on the other hand, subjects VS and SV-clauses are generally considered to have the same syntactic relation to the verb, VS and SV being optional word orders. In this tradition, these two subjects are instead contrasted with the subjects in equational clauses where there is no verb. Subjects in equational clauses are sometimes referred to as 'topic'. The Arabic and Western grouping of clause types is illustrated in Table 9. The division in the coding scheme between three types of subjects allows for analysis that contrasts these types of subjects in various ways.

The categories of case governance for the nominative are the following:

[65] Several other such restrictions on tag choices were built into the menu system. When, for example, the case governance tag has been set to 'absolute negation', 'indefinite' becomes the only option for the subsequent definiteness tag. Another example is that tokens tagged for declension as 'final *alif*' could only be tagged for the case marking with 'not applicable'.

Table 9: *Analysis of subjects*

ARABIC			WESTERN
jumla ismiyya	{ TC		equational clause
	SV	}	verbal clause
jumla fiʿliyya	VS		

I. Subject in VS-clause. The first of the three kinds of subjects is that of an VS-clause (41), the default word order in Standard Arabic. Subjects appearing between an auxiliary and a main verb were tagged as a subject of VS-clause. An example of this is (42) where the subject *an-niẓām* 'the regime' appears between the auxiliary *kāna* and the main verb *badaʾ* 'begin'.

(41) *ʿindama tuqarrir **al-ʾidāra** ʾirsāl-a s-safīr*
when 3FS.decides DEF-administration sending-AMB DEF-ambassador
'When the administration decides to send the ambassador […]'
(al-MuʿallimI, 3:21*)

(42) *wa-law kāna **n-niẓām** badaʾ min bidāyat-i l-ʾaḥdāṯ*
and-if was.3MS DEF-regeme begun.3MS from beginning-AMB DEF-events
wa-qaddam hāḏihi l-ʾiṣlāḥāt
and-presented.3MS these DEF-reforms
'And had the regime started from the beginning of the events to present these reforms …' (Kayālī, 6:43)

II. Subject in SV-clause. The second type of subject is that of a SV-clause in which the subject precedes the verb (43).

(43) ***al-ḥukūma** tumāris dawr-a-ha*
DEF-government 3FS.performs role-ACC-its
'The government is performing its role.' (Fayyāḍ, 18:30*)

III. Topic (subject of equational clause). The third type of subject, here referred to as 'topic', is that of a verbless equational clause consisting of a topic and a comment (44).

(44) ***xaṣā'ir-na*** *naḥnu qalīla jidd-an*
losses-our we few seriousness-ACC
'Our own losses are very few.' (al-As^{c}ad, 3:51)

IV. Subject of an unfinished clause. When a clause for one reason or another is left unfinished, subjects are sometimes left hanging, and it is unclear whether they were to be the subject of an SV-clause or the topic of an equational clause. In such cases, the subject was then tagged as 'subject of an unfinished clause'. This situation arises for example in tailing off, as in (45), or in complicated sentence structures when the speaker does not resume the main clause. This is the case in (46), where the speaker introduces *aš-ša^{c}b is-sūri* 'the Syrian people' as the subject followed by a sequence of prepositional phrases and then a lengthy parenthetical clause (not reproduced). The utterance is then terminated with a drop in intonation, signaling a planned sentence termination, and the main clause with *aš-ša^{c}b is-sūriyy* as its subject is never completed.

(45) *hal* ***il-qānūn*** *… halla' ṭab 'ēh ^{c}āliyāt-i t-taṣḥīḥ*
Q DEF-law … PART PART what mechanisms-AMB DEF-correction
'Is the law … Well, now, what are the correcting mechanisms?'
(Hilāl, 5:18)

(46) *'ana bi-'i^{c}tiqād-i* ***'aš-ša^{c}b-i*** ***s-sūri*** *fī 'istimrār-u*
I in-belief-my DEF-people-AMB DEF-Syrain in continuation-its
bi-n-niḍāl fī daf^{c}-i l-mujtama^{c}-i d-dawli
in-DEF-struggle in push-AMB DEF-community-AMB DEF-international
wa-l-^{c}arab bi-šakl 'asāsi
and-DEF-arabs in-way primary
'It is my belief that the Syrian people in its continuing struggle to push the international community and especially the Arabs …' (Ġalyūn, 3:33)

V. Comment. By 'comment' is meant here the predicate in an equational clause. The unmarked word order for equational sentences is topic — comment, as in (44) above. This order is sometimes reversed to give emphasis to the comment as in (47). Comments can also constitute complete utterances when the topic is omitted as shown in (48).

(47) *wa-****mafhūm-un*** *'asbāb-u*
and-understood-NOM causes-its.AMB
'and its causes are [well] understood' (Abū Majd, 16:28)

(48) *ṣaḥīḥ*
correct
'[That is] Correct.' (ᶜArīqāt, 10:52*)

VI. Dialectal constituent. This category includes tokens that are governed by a set of dialectal structures and that therefore fall outside the Standard Arabic system of case assignment. The inclusion of dialectal constituents under the nominative case is somewhat arbitrary since these tokens by definition cannot be assigned case with Standard Arabic grammar, but they are parallel to topic and comment positions in Standard Arabiccase assignment. The category includes constituents governed by the following exclusively dialectal constructions:

(*a*) existential *fī* 'there is' and *lissa* 'is still' (49)
(*b*) nominal negators, such as *miš*, *mū* 'is not', *mā ᶜind-uh* 'he does not have', etc. (50)
(*c*) the complementizer *innu* (51)
(*d*) the pseudo-verbs *bidd-uh* and *ᶜāyiz* 'want' (52)

(49) *fī* ***mawḍūᶜ*** *ʾana* *ʾurīd* *ʾurakkiz* *ᶜalē-h*
there.is subject I want 1S.focus on-it
'There is a subject I want to focus on.' (al-Xuḍarī, 13:19)

(50) *il-qānūn* *miš* ***sahl***
DEF-law not easy
'The law is not easy.' (Abū Majd, 16:44)

(51) *naḥna* *nataṣawwar* *ʾinnu* ***n-nizām*** *māt*
we 1PL.imagine that DEF-regime died
'We see it as if the regime has died.' (Ġalyūn, 2:30)

(52) *wa-l-kull* *bidd-u* ***muṣālaḥa***
and-DEF-all wanting-his reconciliation
'Everyone wants reconciliation' (al-Miṣrī, 18:39)

The complementizer *innu* is often used in the corpus instead of its Standard Arabic counterparts *anna* or *inna*, as in (51). The subsequent nominal was then tagged as a dialectal constituent. Had the speaker instead used the standard complementizer *anna* in this example, then the following noun, *an-nizām* 'the

regime', would have been tagged as an accusative complement of *inna* and her sisters (see 7.6.3). Here, that is not possible since *innu* is not listed as an accusative governing particle and has no place Standard Arabic case assignment.

In Syrian and Palestinian Arabic, the normal word for 'want' is *bidd-*. This word is particularly problematic for the coding scheme developed here. It is inflected for person by an enclitic pronoun and not by normal verbal inflection as in *bidd-u* 'he wants' (lit. 'his wanting'), and is thus morphologically closer to a noun than to a verb. It does however take objects with an underlying accusative as can be demonstrated with pronominal objects that do have an object form also in dialect (53). It is also negated as a verb (e.g. *mā biddi* 'I do not want') and can be modified with adverbs (e.g. *dāyiman biddu* 'he always wants'). One possibility is to treat the nominal object of *bidd-* as a verbal object and tag it as such. The construction is, however, saliently dialectal and its nominal object is never expected to be marked for case in the corpus as with the other dialectal constituents.

(53) *bid-na* *ʾiyyā-hum* *yittafiqu* *ʾawwal* *ši*
wanting-our OBJ-them 3MPL.agree first thing
'We want them to agree first of all.' (al-Miṣrī, 14:18)

The corresponding word in Egyptian Arabic, *ʿāyiz*/*ʿāwiz* is a participle, and its complement could less problematically be analyzed as an accusative object. It is clearly a dialectal word, however, whose objects are expected never to be marked for case. Objects of *ʿāyiz* were also tagged as dialectal constituents to balance the tagging of *bidd-* for Palestinian and Syrian speakers. The Levantine *bidd-* and the Egyptian *ʿāyiz* are assumed to be to be used at the same frequency in the dialects and to have the same degree of dialectal saliency.

7.6.2 Genitive

The genitive is the most straight forward of the three cases in that it marks only annexation and prepositional complements.

1. Annexed genitive. The genitive construction (*iḍāfa*) is extremely frequent in Arabic. It denotes possession and other kinds of relations between entities and is often used for what in English is expressed through compound nouns. It consists of a nominal in construct state (see above) followed by a second nominal in the genitive.

Several function words that are formally nouns, for example *kull* 'all' and *ġayr* 'not', 'un-' (54) form genitive constructions with a following noun and these were then tagged as an annexed genitive.

(54) *hāḍā ġayr-u* ***mawjūd-in*** *fī sūrīyā*
this not-NOM existent-GEN in Syria
'This is nonexistent in Syria.' (Tayzīnī, 15:23)

Complex chains of annexed genitives with three or more constituents, as in (55), are common. In such chains of annexation all but the first term were tagged as annexed genitives.

(55) *huwa qāʾim ʿala ʾasās* ***ʾistikmāl*** ***ʿamaliyyat-i***
it stand.3ms on foundation continuation process-AMB
l-bināʾ
DEF-construction
'it rests upon the foundation of the continuation of the construction process.' (Fayyāḍ, 2:35*)

Certain numerals take annexed genitives as complements: in plural after numbers 3–10, and in singular after even hundreds, thousands, etc. (56).

(56) *hunāk ḥawāli miʾat-ēn ʾalf* ***mustawṭin*** *dāxil*
there around hundred-AMB.DUA thousand settler inside
madīnat-i l-quds
city-AMB DEF-Jerusalem
'There are around two hundred thousand settlers inside of the city of Jerusalem.' (ʿAbd al-Qādir, 3:40)

11. Prepositional complement. All prepositions in Arabic take complements in the genitive. Although Arabic propositions are diverse in their form and can be divided into at least three subgroups, no such distinction was made in the coding scheme.

Western grammars typically divide Arabic prepositions into three groups: true prepositions, semi-prepositions, and compound prepositions.[66] True prepositions are 'true' in the sense that they are a closed word class. Many of them do not have a trilateral root and they are, synchronically, not derived from nouns. Three of them are prefixed to the noun. Badawi et al. (2004:174–98) list the following twelve true prepositions:[67]

[66] The terminology in Ryding (2005:366) is followed here. Badawi et al. (2004) refer to the first two groups as 'prepositions' and 'prepositionals', respectively. These designations are problematic in that they leave no term for a category on the higher level including both groups.

[67] Ryding lists 10 true prepositions. She does not include *ladā* and *maʿa* 'with'. She excludes

ᶜalā	on over; above
ᶜan	about; away from; on (topic)
bi-	in; at; with; by
fī	in; within; into
ḥattā	until
ilā	to; as far as
ka-	like; as
ladā	with; amongst; upon
li-	belonging to; for the purpose of
maᶜa	with; in spite of
min	from; of
munḏu	ago; since

The second group of preposition is semi-prepositions. These are an open word class, derived from nouns, that has grown significantly in Modern Standard Arabic as compared with Classical Arabic (Procházka 2006:701). Badawi et al. (2004:199–219) list 44 semi-prepositions that all take the vowel ending *-a*, as in the word *amāma* 'in front of' in (57):

(57) *ʾ**amāma*** *l-barlamanāt*
in.front.of DEF-parliaments
'in front of the parliaments' (al-Xuḍarī, 8:55)

In Traditional Arabic Grammar, there is no relationship between semi- and true prepositions. True prepositions are referred to as *ḥurūf jarr* (s. *ḥarf jarr*), literally 'particles of the genitive'.[68] Only true prepositions are regarded as directly governing a complement for genitive. Semi-prepositions are analyzed as nouns and are assigned accusative case by being locative or temporal adverbials (*ẓurūf*, s. *ẓarf*), forming a genitive construction with the following noun. This following noun is then analyzed as being in the genitive as an annexed noun and not as a prepositional complement. This difference in the Western and the Arabic tradition in the analysis of semi-prepositions is illustrated in Table 10.

True prepositions are combined with semi-prepositions to form a third type: compound prepositions. Two frequent examples are *fī dāxil* 'in' (lit. 'in

maᶜa from the list of true prepositions with the argument that Sibawayh, the 8th century grammarian "cites the phrase *min maᶜ-i-hi*, showing that *maᶜ-a* can sometimes be the object[*sic*] of another preposition" Ryding (2005:393n).

[68] There is no consensus in this tradition exactly which words are to be regarded as as *ḥarf jarr* (Procházka 2006:699).

Table 10: *Analysis of semi-prepositions*

WESTERN	*ʾamāma*	*l-barlamānāt*
	in.front.of	DEF-parliaments
	preposition	prepositional complement
	case n/a	genitive
ARABIC	*ʾamām-a*	*l-barlamānāt*
	front-ACC	DEF-parliaments
	adverbial (*ẓarf*)	annexation (*muḍāf ilay-hi*)
	accusative (*manṣūb*)	genitive (*majrūr*)

inside)' and *bi-dūn* 'without' (lit. 'by without'). The change of the vowel ending of the semi-preposition from *-a* to *-i* when preceded by a true preposition supports the Arabic analysis, since the semi-preposition here follows nominal case morphology. At least some of the compound prepositions seem to be in free variation with their non-compound counterparts. In the corpus, al-Qaddūmī for example uses both *dāxil al-muxayyamāt* and *fī dāxil al-muxayyamāt* 'in the camps'.

In this study, no difference was made between true, semi- and compound prepositions. Their complements were all tagged as prepositional complements and the prepositions themselves, of whatever type, were not regarded as nominals and were not annotated.

7.6.3 Accusative

The accusative case is functionally the most diverse. Grammars in the Arabic tradition feature long lists of constructions that govern the accusative, often under the heading *al-manṣūbāt* 'the accusatives'. Many of these constructions are not part of the core argument structure and are, as such, at times difficult to precisely define and operationalize in the way required for linguistic annotation. Some of these problems were sidestepped here by having two categories, adverbs and miscellaneous accusative, that encompass much of this diversity and whose subcategories are difficult to disentangle from one another.

1. Object. This category includes objects of verbs, dialectal and standard, and, less commonly, of participles and verbal nouns. No distinction was made be-

tween the two objects of ditransitive verbs. In (58), for example, the first objects *al-quwwāt* 'the forces' and *aš-šabīḥa* 'the henchmen', as well as *libās madani* 'civil clothing' as the second object of the ditransitive verb *yulabbis* 'dress', were both tagged as objects.[69]

(58) *bada*ʾ *yilabbis al-jēš wa-l-quwwāt-i l-*ʾ*amniyya*
started.3MS 3MS.dress DEF-army and-DEF-forces-AMB DEF-security.ADJ
wa-š-šabīḥa ***libās madani***
and-DEF-henchmen clothes civil
'And it [the regime] started to dress the army and the security forces and the henchmen in civil clothes' (al-Asᶜad, 10:42)

Case assignment for objects of verbal nouns and participles that have both verbal and nominal characteristics is more complicated. The underlying object of these verbal elements surfaces either in the accusative, as an annexed genitive, or marked with the preposition *li-* 'for', in which case it was tagged as prepositional complement. For verbal nouns, the object surfaces as an annexed noun in the genitive if it is the only overt argument, as in (59). If both the underlying subject an object are overtly expressed, as in (60), the subject surfaces as a annexed genitive or as an enclitic pronoun, and the object as an accusative. In verbal nouns of ditransitive verbs, one object surfaces as an annexed genitive noun or as an enclitic pronoun, and the other as and accusative object (61) (Badawi et al. 2004:257; Wright 2011:vol.ii, 57).

(59) *takwīn-a* ***d-dustūr***
formation-AMB DEF-constitution
'the formation of the constitution' (Mursī, 12:03)

(60) *fuqdān-**u*** ***š-šarᶜiyya***
loss-his.AMB DEF-legitimacy
'his loss of legitimacy' (Ġalyūn, 13:53)

(61) *bi-waṣf-i-**hi*** ***ḥadaṯ-an*** *ʾijtimāᶜiyy-an wāsiᶜ-a* *ʾad-dalāla*
in-description-GEN-its event-ACC social-ACC wide-ACC DEF-implication
'in describing it as a social event with wide implications' (Tayzīnī, 1:35)

With transitive participles, the underlying object surfaces as an annexed genitive when the event referred to is past perfective or a state (62), and as

[69] Traditional Arabic Grammar describes also triply and quadruply transitive verbs (Wright 2011:vol.ii, 51). Badawi et al. (2004:380) found no instances if this in their data, nor were any found in the corpus of this study.

accusative when referring to non-past events or historical present (63) (Wright 2011:vol.ii, 63; Cantarino 1974:vol.ii, 412).

(62) *kull-u*[70] *wāḥid lābis* ***qinā***^{c}
every-AMB one wearing mask
'Everyone is wearing a mask' (Abū Majd, 23:48)

(63) *wa-*^{c}*ind-u musā*^{c}*id* ^{ʾ}*amni fi l-ġurfa l-mujāwira māsik*
and-with-him assistant security.ADJ in DEF-room DEF-next holding
^{ʾ}***awrāq*** *fī-ha listāt maṭlūb-īn*
papers on-them lists wanted-PL.AMB
'And he has a security assistant in the next room, holding papers with lists of wanted persons.' (Šallaḥ, 31:21)

The situation is further complicated by the fact that dialectal forms have a preference for participles in some situations where finite verbs are preferred in Standard Arabic (Holes 2004:220–23). The participles in (62) and (63) would probably be perceived as dialectal, making the question of their case governance moot. While the use of the participle in these examples is arguably not Standard Arabic in a stylistic sense, it is nevertheless formally correct Standard Arabic and was treated as such for purposes of tagging.

11. Absolute object. Arabic verbs can take their own verbal noun as an accusative argument, a so-called absolute or internal object. The absolute object is in both Traditional Arabic Grammar (Wright 2011:vol.ii, 43ff; Ibn Hišām 1990: 217) and Western grammars (Cantarino 1974:vol.ii, 445ff; Badawi et al. 2004: 145) described as having three different functions. These functions are (*a*) unspecified to emphasize or assert the truth of the statement[71] (64); (*b*) specified with an adjective or annexed genitive to give various adverbial meanings (65), and (*c*) inflected for number or specified with a numeral to give iterations of the action (66).

(64) *kāna d-damm-u yanṣabbu min ḏirā*^{c}*-i-hi nṣibāba-an*
was DEF-blood-NOM pour from arm-GEN-his pouring.NOUN-ACC
'the blood was **pouring** from his arm [**a pouring**]'
(Badawi et al. 2004:145, boldface in trasnslation in original)

[70] The *-u* ending in *kullu* is a possible epenthetic vowel in Egyptian Arabic and is therefore not interpreted as a nominative case ending (see 7.7.1).

[71] Interestingly, no example of this was found in the corpus.

(65) *wa-fahimat* ***fahm-an*** ***xāṭiʾ-an***
and-undersood.3FS understanding-ACC faulty-ACC
'and they [the group] misunderstood' (Badīᶜ, 13:05)

(66) *naxṭu* ***xuṭawāt*** *li-tarsīx* *mumārasāt-i* *d-dīmuqrāṭiyya*
1PL.step steps to-establish practices-AMB DEF-democracy
'We take steps to establish democratic practices' (Hilāl, 23:36)

In Traditional Arabic Grammar, phrases such as *šukran* 'thank you' are analyzed as absolute objects (*mafᶜūl muṭlaq*) of an implied verb: [*aškuruka*] *šukran* '[I thank you a] thanking'. In Traditional Arabic Grammar, absolute objects may also be substituted for another noun with similar meaning (*nāʾib mafᶜūl muṭlaq*). One traditional example is *jalasa quᶜūdan* 'he sat [a sitting]'. In this study, only when the verbal nouns echoes the root of an overt verb was it considered an absolute object. When it does not echo the root of a verb, it was instead tagged as an adverb (see below). This was done to make the category of absolute object more straightforwardly operationalized and to avoid some thorny issues of delimiting absolute state from circumstantial qualifiers. Absolute object, as the term is operationalized here, thus constitutes a much more narrow category than the corresponding *mafᶜūl muṭlaq* of Traditional Arabic Grammar.

III. Complement of inna *and her sisters.* *Inna* and her sisters are a set of five particles that assign accusative case to the subject in the clause they introduce. They are:

inna	verily; complementizers of *qāla* 'say'
anna	complementizer
lākinna	but
layta	if only
laᶜalla	perhaps

The two first introduce subordinate clauses as complementizers and the others introduce main clauses. All but *laᶜalla* are represented in the corpus. *Inna*, the first item in the list, needs some further comment. It is described in the consulted grammars as having two functions: introducing a main clause to give emphasis and as complementizer introducing indirect quotes following the particular verb *qāla* 'say' and its derivatives. Badawi et al. (2004:713) comments that *qāla* is occasionally used with the complementizer *anna*, contrary to traditional grammar. In the present corpus, this is in fact the dominant pattern

(see Table 17 on page 185) exemplified here in (67). This is thus not a prescriptively correct Standard Arabic use of *anna*.[72] It is however also not dialectal, since the dialects would normally here use the dialectal complementizer *innu*, as in (68). *Anna* is in other words a saliently standard lexical item, and it was here regarded as being able to govern case also when used as a complementizer to *qāla*. Thus in (67), for instance, *al-muqāwama* 'the resistance' was tagged as 'complement of *inna* and her sisters'.

(67) *naḥnu naqūl* ***ʾanna*** *l-muqāwama xayār-ha wāḍiḥ*
we 1PL.say that DEF-resistance choice-its clear
'We say that the choice of the resistance is clear.' (Šallaḥ, 28:50)

(68) *ʾana ʾaqūl ʾinnu ʾil-* ***wasāʾil-i*** *l-ʾiʿlām*
I 1S.say that DEF- means-AMB DEF-media
'I say that the … the mass-media …' (al-Xuḍarī, 24:13)

IV. Absolute negation. Complements of the absolute negation *lā* 'there is no' take the accusative ending *-a* without nunation. Such a complement also does not take the definite article, making it an oddity in the Standard Arabic nominal morphology, being neither marked as indefinite with nunation, nor as definite with the definite article. In this study it was regarded as indefinite for purposes of annotation.[73] Absolute negation is most common in the set phrases *lā šakk(a)* 'no doubt' and *lā budd(a)*, 'definitely' (lit. 'no escape'). Of the 85 instances of absolute negation in the corpus, 57 are either of these two phrases, with or without the case ending. Absolute negation is also used productively, as exemplified in (69).

(69) *lā* ***ṯiqat-a*** *l-ī fī ʾal-mawqif-i l-ʾamīrikī*
no trust-ACC for-me in DEF-position-AMB DEF-American
'I have no trust in the American position' (al-Qaddūmī, 16:33*)

[72] According to traditional Arabic Grammar, *anna* is used in the complement of *qāla* only if it is used in the sense of *ẓanna* 'believe' (e.g. in Ibn ʿAqīl 2005:134).

[73] The consulted grammars allow for the complement of absolute negation to be specified with an adjective. Only grammars representing Traditional Arabic Grammar explicitly allow it to have an annexed genitive (Ibn ʿAqīl 2005:150; Wright 2011:vol.ii, 96). Neither of these options are represented in the corpus; all complements of absolute negation in the corpus are lone indefinite forms without any form of specification.

v. Complement of kāna *and her sisters.* '*Kāna*, and her sisters', or 'verbs of being and becoming' (Badawi et al. 2004:399–414), is a set of verbs that take a complement in the accusative, as in (70). The consulted grammars vary in how many verbs are listed in this category, depending on how many rare and archaic verbs are included. The number of verbs listed ranges from nine in Ryding (2005:634) to nineteen in Wright (2011:100).

(70) *ʾawudd ʾan ʾakūn* ***ḥurr-an***
1S.want COMP 1S.be free-ACC
'I want to be free.' (al-Barādiᶜī, 6:07)

Complements of *kāna* and her sisters could in principle be grouped with objects as verbal complements. They are however, in Traditional Arabic Grammar, and often on Western grammars as well, regarded as a separate set of verbs operating on an underlying equational clause, rather than as simply a verb taking a complement.

vi. Numbers specification. Nouns that are complements of the numerals 11 to 99 are in singular accusative, as the word *ᶜāmm* 'year' in (71). In Traditional Arabic Grammar this position is referred to as *tamyīz*, a term that is further commented on below.

(71) *maḍa fi l-ḥukm ʾaksar min salāsīna* ***ᶜām-an***
remained.3MS in DEF-power more than thirty year-ACC
'He remained in power for more than thirty years.' (al-Barādiᶜī, 20:36)

vii. Adverbs. In this study, adverbs constitute a broad category defined as words in the accusative that modify verbs (72), adjectives (73), other adverbs, or a whole clause (74).

(72) *ʾana ʾattafiq maᶜ-ak* ***tamām-an***
I 1S.agree with-you complete-ACC
'I agree with you completely' (Fayyāḍ, 4:14*)

(73) *wa-hāḍa yaᶜni šēʾ muhimm* ***jidd-an***
and-this PART thing important very-ACC
'And this is, you know, a very important thing' (Ġalyūn, 7:28)

(74) ***ʾawwal-an*** *ʾaškur* *al-gazīra*
first-ACC 1S.thank Al Jazeera
'Firstly, I thank Al Jazeera.' (Badīᶜ, 1:01)

There are relatively few 'pure' adverbs in Standard Arabic. Most are derived from nouns and are formally in accusative case. As the examples above illustrate, many adverbs are consistently marked for case, and the case ending is better explained as word formation than as case inflection (Badawi et al. 2004: 56). Indeed, the adverbs in (72)–(74) rarely if ever appear without case endings. On the other hand, there are many instances in the corpus of adverbs without the accusative ending, as in (75) and (76).

(75) *wa-mā* *wagadna* *fī-hā* *šayʾ-an* *fuhim* ***xaṭaʾ***
and-what found.1PL in-it thing-ACC understood.3MS mistake
ṣaḥḥaḥna-h
corrected.1PL-it
'That which we found to be misunderstood in it we corrected.'
(Badīᶜ, 16:11)

(76) *daᶜ-ni* *ʾāxuḏ-ha* *bi-xtiṣār* *šadīd* ***band*** ***band***
let-me 1S.take-it in-summary strong paragraph paragraph
'Let me take it very briefly, paragraph by paragraph' (ᶜArīqāt, 21:37*)

The lexicalization of adverbs in Modern Standard Arabic has rendered less usable the analytical tools of Traditional Arabic Grammar (Badawi et al. 2004: 160), a system developed for analysis of Classical Arabic where this class was less productive. The case marking of lexicalized adverbs that are frequent in the corpus (see 10.4) are accounted for in Traditional Arabic Grammar in different and sometimes roundabout ways. In a book on difficult cases of *iᶜrāb*, al-Maᶜarrī (2005:passim), for example, analyzes *jamīᶜan* 'all' as a circumstantial qualifier (*ḥāl*); *abadan* 'ever' as a temporal adverbial (*ẓarf zamān*), and *jiddan* 'very' as "an absolute object of an omitted verb". Other adverbs are said two have two possible analyses. The word *ayḍan* can, for example, be analyzed as either an absolute object of an omitted verb, or as a circumstantial qualifier.

مفعول مطلق لفعل محذوف

The Western grammars consulted are quite explicit regarding the difficulties of categorizing Arabic adverbs (and adverbs in general) onto subcategories on formal or semantic grounds (Badawi et al. 2004:160; Ryding 2005:276). The subcategories they do deal with, derived from Traditional Arabic Grammar (circumstantial qualifier, *tamyiz*, accusative or reason, etc.) are fuzzy concepts with a number of borderline cases where arguments could be made for different

categorizations.

The approach taken here was to annotate as adverbs all adverbs with nominal morphology, even those that are clearly frozen forms, such as *jiddan* 'very'. The adverb category thus defined includes what in Traditional Arabic Grammar is described as a number of unrelated constructions, as exemplified above. These are the following, with reference to Wright (2011:vol.ii) in parenthesis:

tamyīz	specification (122)
mafᶜūl lahu	purposive object (121)
ẓarf	temporal or locative adverb (109, 111)
ḥāl	circumstantial qualifier (112)
nāʾib mafᶜūl muṭlaq	absolute object substitute (54)

The first item in the list, *tamyīz* is, in Traditional Arabic Grammar, used for a number of disparate functions. It can refer to the complement of numbers 11–99 and to specifications of verbs and adjectives, in which case it is here included under adverbs. Other functions of *tamyīz* are here included under 'miscellaneous accusatives', as described below.

viii. Miscellaneous accusative. In Arabic grammars one finds several other constructions that govern the accusative but are archaic or otherwise rare in Standard Arabic. This is a potentially long list of constructions (see below) but only two were found in the corpus and tagged as miscellaneous accusative: *al-munādā*, vocative, and *al-ixtiṣāṣ*, pronoun specification.

The vocative, firstly, is a noun preceded by either of the two vocative particles *yā* and *ʾayyuhā*. The noun is either in nominative or accusative case. The rules governing this as described in Traditional Arabic Grammar are fairly complicated. In brief, the vocative is in nominative, without nunation, only if it the addressee is addressed directly (*munāda maqṣūda*) and not further specified with a genitive annexation, an object, or a prepositional phrase. If the addressee is somehow specified or is addressed metaphorically (*munāda ġayr maqṣūda*) then it is in the accusative. These rules are, in any case, more or less factious since the vocative is hardly ever marked for case (Badawi et al. 2004: 287). The vocative is not infrequent in the corpus, but it almost exclusively governs proper names or terms of address, which were excluded from annotation (see 6.5). There were only one occurrences in the corpus of a vocative that is not a proper name or term of address, a metaphorical address (77). It is not marked for case.

(77) *yā* ***ʾamrīki*** *lā* *tuṣaddiq*
VOC American NEG 2MS.believe
'Oh American, do not believe [it].' (Šallaḥ, 33:00)

Second, *al-ixtṣāṣ*, pronoun specification (Wright 2011:vol.ii, 76), is a noun appearing after a pronoun to specify its meaning, as in (78). In Traditional Arabic Grammar, this noun is analyzed as being the object of the implicit verb *axuṣṣu* 'I specify' (Ibn ʿAqīl 2005:410).

(78) *wa-bi-t-tāli* *ʿalay-na* *naḥnu* ***l-muslim-īn*** *wājib*
and-in-DEF-following on-us we DEF-muslim-PL.AMB duty
kabīr *gidd-an*
very-ACC
'This means that there lies a heavy duty on us Muslims.'
(Abū Majd, 15:19)

The following are structures found in Traditional Arabic Grammar that govern the accusative but not represented in the corpus, with references to Wright (2011:vol.ii) in parenthesis.

istiṯnāʾ	exemptive (335)
at-taḥḍīr wa-l-ʾiġrāʾ	construction of incitement (74)
mafʿūl maʿhu	accompanying object (83)
tamyīz kam	specification of *kam* (124)
mā/lā al-ʿāmila ʿamal laysa	*mā/lā* with *laysa*-governance (104)
ʾin an-nāfiya/lāta	negating particles (105)

The exemptive particle *illā* 'but', 'except for' merits some further comment. In the consulted grammars, it is described as governing the accusative in affirmative clauses, but not in negative clauses. *Illā* in affirmative clauses seems to be extremely rare, however.[74] No example of *illā* in an affirmative clause was found in the corpus. In the corpus, it is only ever used in negated sentences, giving the meaning of 'only', as in (79).[75]

[74] Except expressions giving the time, e.g. *al-xāmisat-u ʾillā rubʿ-a* 'a quarter to five' lit. 'The fifth except a quarter'.

[75] Badawi et al. (2004:671) apparently also did not find any example of *illā* in an affirmative clause in their data of modern written Arabic. They give the traditional grammar-book example *qāma l-qawm-u illā zayd-an* (stood.up.3MS DEF-people except zayd-ACC) 'Everyone stood up except for Zayd' (cf. Ibn ʿAqīl 2005:234).

(79) *hāḍa lā yaḥduṭu* ***ʾilla*** *fī ʾumma tuʾminu bi-l-gahl*
this not 3MS.happens except in community 3FS.believes in-DEF-ignorance
'This can only happen in a community that believes in ignorance.'
(Abū Majd, 9:20)

Of the 'fossilized exemptive verbs' (Badawi et al. 2004:683), *mā ᶜadā*, *mā xalā*, and *ḥāšā*, that govern either accusative or genitive (both being considered correct), only the first was found in the corpus, twice, but not in a position were it governs a nominal.

7.7 Case marking

The variables described above together determine the prescribed Standard Arabic case ending. The final tag in the coding scheme encodes if and how this prescribed morpheme is actually produced by the speaker. One major methodological obstacle in quantifying case marking are endings that can be interpreted both as dialectal — and thus as not marking case — and as Standard Arabic case markers. This problem is analyzed in detail below after the categories of case marking as represented in the coding scheme are listed and described.

I. Marked. A token was tagged as marked for case if articulated with a case ending that (*a*) is in accordance with Standard Arabic rules of case assignment, and (*b*) has a form that is not attested in dialectal grammar. Both these requirements have to be satisfied for the token to be tagged as marked for case. A token with a form that does not fit either of these criteria was tagged as belonging to one of the categories of case marking listed below.

II. Unmarked. Tokens that do not have an ending that could be interpreted as a case marker were tagged as unmarked for case. Tokens in paradigms with endings that include long vowels or diphthongs — the sound m.pl., dual, and the five nouns in construct state — cannot be unmarked for case by this definition and were thus never given this tag.

III. Ambiguous. Tokens with endings that could have been produced by the dialectal system and that coincide with some form of the Standard Arabic case markers, whether or not prescriptively correct, were tagged as ambiguously marked for case. The patterns of such morphological overlap between the dialectal and the Standard Arabic system is fairly complex. The various situations where it arise are described in detail in 7.7.1 below.

iv. Hypercorrect. If a noun is unambiguously marked for case, but with an ending that is prescriptively incorrect for its syntactic position, it was tagged as hypercorrect. This is illustrated in (80), where the word *muraššaḥ* 'candidate' is marked for accusative with the ending *-an*. This ending could not have been produced by the dialectal grammar in this position and is therefore not ambiguous. As subject of the verb *yakūna* 'is' it is however in a nominative position and the prescribed ending is *-un*; the case ending is thus hypercorrect.

(80)

ʾaᶜlan	*bi-wuḍūḥ*	*ʾanna-na*	*lan*	*yakūna*	*la-na*
announced.3MS	with-clearity	that-we	NEG.FUT	3MS.be	for-us
muraššaḥ-an					
candidate-ACC					

'It [the council of the Brotherhood] has clearly announced that we will not have a candidate.' (Mursī, 9:11)

v. Indistinguishable. Words that are not clearly audible in the recordings were transcribed in the corpus as xxx and excluded from annotation. When the token word itself is audible but the ending is not, the word was fully transcribed and annotated but tagged as having an indistinguishable case ending. Endings are indistinguishable due to rapid or unclear articulation, speech that overlaps with that of the interviewer, or due to poor sound quality in the recording.

vi. Indeclinable. Most declensions when suffixed with the 1s. enclitic pronoun *-ī*, and words in the 'final *alif*' declension, cannot take case endings in Standard Arabic. These tokens were tagged as indeclinable.

7.7.1 Ambiguity

Ambiguous case markers are endings that could have been produced both as Standard Arabic case endings, whether prescriptively correct or not, and as dialectal endings or epentheses. Ambiguous case endings present a problem for quantitative analysis in they cannot be reliably regarded as marking or not marking case. They are excluded in most of the analysis presented later in this thesis. Below, previous studies and how have they dealt with this problem are first briefly described. Thereafter the problematic nature of ambiguous case endings is discussed and situations where ambiguity arises are listed.

In previous research where case markers in speech is studied, the issue of ambiguous endings is not explored in much depth, granted that case endings are not the sole focus of these studies. Schulz (1981:139n8) touches on the matter

only briefly, stating that it is in practice not difficult to tell the epenthetic vowel *i* from genitive case endings but does not mention how that distinction is made. He also states, however, that there is a possibility that numbers for case markers in the genitive "are slightly inflated, since I may have counted some transitional vowels as genitive endings" (Schulz 1981:139). Elgibali (1985) does not discuss the problem of ambiguous case endings. A more principled approach is found in Parkinson (1994a:181). He does not count *i* in epenthetic position for a speaker who consistently uses *i* as epenthetsis. For speakers who sometimes pronounce *a*, *u* or *i* in these positions, all vowels are treated as attempts at Standard Arabic vowelling. This method is problematic since speakers who do use case markers do so inconsistently. Even if a speaker does mark certain words for nominative with *-u*, words that are not marked by the speaker would still be subject to epenthesis and for this reason sometimes be pronounced with a final *-i*. The method thus assumes that if a speaker sometimes uses case endings, dialectal epenthesis is suspended, which is unlikely. The most detailed discussion on ambiguous endings is found in Meiseles (1977:184). He identifies several of the types of ambiguous endings listed below. According to him, it is for example often impossible to tell whether the *i* preceding the article, or a vowel connecting the enclitic pronouns *-ik*, *-ak*, or *-uh*, are Standard Arabic case endings or part of the dialectal form.

In the present study, a strict form of Meiseles' more sceptical view was employed as a matter of methodological principal. This principle is that *all nominal endings that could have been produced by the dialectal system are regarded as ambiguous* and tagged accordingly.

An ambiguous case ending can relate to the case system of Standard Arabic in three different ways. First, it may be of a form that corresponds with Standard Arabic but that is also the expected dialectal form: it is completely standard and completely dialectal. In (81), the final *i* in *al-mawqif-i* 'the position' is the correct Standard Arabic case ending that marks genitive after the preposition *ᶜalā*. But the nominal phrase *al-mawqif-i l-ᶜarabi* 'the Arabic position' would be produced with *i* in the same position also in a dialectal form as epenthesis preceding the article of the second word (and then more appropriately transcribed as *al-mawqif il-ᶜarabi*).

(81) *ʾurāhin* *ᶜala* ***l-mawqif-i*** *l-ᶜarabi* *naᶜm*
1S.bet on DEF-position-AMB DEF-Arabic yes
'I am betting on the Arabic position, yes.' (Ġalyūn, 15:54)

In the example above, it is not possible from this utterance alone to tell if

the inter-word *i* is a grammatical marker of case or part of the article. Counting such tokens as unmarked case would remove from analysis vowels that may have been intended as case markers, not allowing, as it were, speakers to mark case in these situations. Counting them as instances of marked case would undoubtedly wrongly include as case markers many instances of inter-word vowels that are in fact epenthesis and unrelated to the case system.

The second way ambiguous endings may relate to Standard Arabic grammar is when the dialectal system produces forms that taken in isolation are possible Standard Arabic forms, but that do not match the syntactic position of the word. In (82), the sound m.pl. ending in *al-lībiyyīn* 'the Libyans' can be interpreted as a Standard Arabic marker of genitive or accusative case, and then does not match its syntactic position as a nominative subject. The *-īn* ending is also the dialectal form of sound m.pl. Tokens with such endings could be regarded as unmarked for case, but this would fail to account for the possibility of hyper correction in that the speaker of (82) may have erroneously intended *-īn* as a genitive marker. Hypercorrection may seem unlikely in this example, but judgments on the likelihood of hypercorrection necessarily involve far-reaching assumptions as to how the case system is internalized by speakers, an approach best avoided.

(82) *ṭabᶜ-an* ***ʾal-lībiy-īn*** *iᶜtarafu* *fī-na*
nature-ACC DEF-Libyan-PL.AMB recognized.3MPL in-us
'Of course, the Libyans have recognized us.' (Ġalyūn, 8:15)

The third way ambiguous ending may relate to the Standard Arabic case system is when the ending itself can be interpreted as a case marker, but where it yields a word form that is not formally possible in Standard Arabic. In (83), there is an inter-word *a* after *ḥājāt* 'needs'. The word is in accusative position as the object of the verb *yurāᶜī* 'see to' and *-a* would be the prescriptively correct ending of the more frequent triptote paradigm. This word is however the sound f.pl., a paradigm that unlike the triptote takes the accusative ending *-i*. The speaker may have applied the triptote declension to this word. The *-a* ending can therefore be interpreted as hypercorrect, but also as part of the article (see below).

(83) *yajib* *ʾan* *yurāᶜi* ***ḥājāt-a*** *ṭ-ṭaraf-ayn*
must COMP 3MS.see.to needs-AMB DEF-side-DUA.AMB
'He must see to the needs of both sides.' (Šallaḥ, 41:23)

In all these three situations, the token was tagged as having an ambiguous case marker and they are glossed as such in the examples. Ambiguity in case marking is thus, in this study, a technical term, often encompassing endings that may seem very unlikely to be intended as case markers by the speakers. Some glosses of endings as ambiguous in the examples may thus be counterintuitive. These endings are problematic when case marking is quantified. They are excluded in most analyses, leading to loss of data, (see 8.1.1). This was nevertheless deemed a more sound approach than subjectively evaluating individual instances of vowel endings for the likelihood of them being intended as case markers.

Below, situations where ambiguous endings appear are mapped out and discussed. Five such situations were identified: (*a*) preceding *hamzat al-waṣl*; (*b*) epenthesis in consonant clusters; (*c*) vowel initial enclitic pronouns; (*d*) case endings with long vowels or diphthongs, and (*e*) the specific form *ayyi* 'any'.

1. Preceding hamzat al-waṣl. The most frequent situations for ambiguous case endings to occur is at word boundaries where the second word begins with *hamzat al-waṣl*, a word initial glottal stop that is omitted if preceded by a vowel. By far the most frequent case of *hamzat al-waṣl* is in the definite article *al-*, as in (81) and (83) above. *Hamzat al-waṣl* also appears in verbs in the perfect tense and verbal nouns of forms VII-X (e.g. *intixābāt* 'elections', and in a small number of underived nouns (e.g. *ism* 'name'[76], and in the relative pronoun *allaḍī* in its various forms.

The definite article is in its dialectal form normally *il-* or *el-*. The vowel of the article often fills the slot of the case ending of the previous noun, resulting in ambiguity. In higher registers of spoken Arabic the article is also often realized as *al-*, regardless of case, in what appears to be a stylistic variation of the purely dialectal forms (cf. Parkinson 1994a:193).[77] This is probably due to *ʾal-* being the citation form of the article and the Standard Arabic phrase initial form, causing it to be perceived as the more correct form. This free variation is illustrated in (84), where the words *majāl* 'area' and *ʿamal* 'work' are both in genitive position and followed by a word with the definite article. One has the ending vowel *-i* and the other *-a*.

[76] These words are normally pronounced with glottal stop (*hamzat al-qaṭʿ*), contrary to traditional accounts (Kaye 1991).

[77] In the corpus there are 727 triptote tokens with a final *a* (that does not represent *tāʾ marbūṭa*) that are followed by a definite article and where this *a* could be part of the article. Of these, 608 are in nominative or genitive positions and so are prescriptively incorrect. This suggests that *a* preceding the article is produced largely independent of case.

(84) *fī* ***magāl-i*** ***l-ᶜamal-a*** *s-siyāsi*
in area-AMB DEF-work-AMB DEF-political
in the area of political work (Mursī, 4:37)

The ambiguity of case markers in connection with the definite article and other instance of *hamzat al-waṣl* is resolved in three situations. The first is in connection with *tāʾ marbūṭa* realized as *a*, as in all four words in (85). The token is then unmarked for case. The second situation is when *hamzat al-waṣl* is realized as a glottal stop, as in (86). The vowel is then part of the article and cannot be a case marker.[78] The third and final way ambiguity preceding *hamzat al-waṣl* is resolved is on tokens with a masculine singular *nisba* ending realized as an unstressed final *-i*, as in the word *siyāsi* 'political' in (87). The word is then necessarily unmarked for case since the *nisba* ending must receive stress for a case ending to be added. If the *nisba* ending is stressed, ambiguity applies as normal, as in (88)

(85) *al-ʾistrātījiyya* *l-kāmila* *li-l-marḥala* *l-muqbila*
DEF-strategy DEF-complete for-DEF-phase DEF-coming
'the complete strategy for the coming phase' (al-Asᶜad, 21:45)

(86) *hunāk* ***ḥaqq*** ***ʾal-lāji-ʾ*** *fī* *ʾan* *yaxtār*
there right DEF-refugee in COMP 3MS.choose
'There is the right of the refugee to choose.' (ᶜArīqāt, 11:46*)

(87) *wa-l-mašhad-a* ***s-siyāsi*** *l-maṣri*
and-DEF-sceneAMB DEF-political DEF-Egyptian
'and the Egyptian political scene' (Abū Majd, 6:12)

(88) *al-mašrūᶜ-a* ***n-nahḍawíyy-i*** *t-tanwīríyy*
DEF-project-AMB DEF-renaissance.ADJ-AMB DEF-enlightenment.ADJ
'the Enlightening-Renaissance project' (Tayzīnī, 8:03)

11. Epenthesis in consonant clusters. Ambiguous case endings also arise when in Egyptian Arabic vowels are inserted in between words to brake up consonant clusters. Traditional Arabic Grammar allows for a limited set of possible sylla-

[78] The phenomenon is also noted by Salib (1980:59) who attributes it to "the attention and prominence speakers tend to give to individual L[iterary] A[rabic] words". This feature is not dialectal, nor is it prescriptively standard; it seems to be particular to the spoken Standard Arabic of at least some individuals. In the present corpus, it is particularly noticeable in al-Xuḍarī's interview.

ble structures and triple consonants are never allowed. In this system, clusters of three consonants over word boundaries after nouns are avoided with the obligatory vocalic case ending. In Spoken Standard Arabic where case endings are normally scarce, dialectal epenthesis mechanisms come into effect instead, generating ambiguous endings.

Arabic dialects vary in their tolerance of consonant clusters and in how epenthetic vowels are inserted to avoid them. In the dialects represented in the corpus, a sequence of three consonants over word boundaries in Syrian and Palestinian Arabic is either tolerated or resolved by inserting an epenthetic vowel after the first consonant, producing CVC-C, with hyphen representing word boundary (Shahin 2006:528; Cowell 2005:25; Klimiuk 2013:89). These epenthetic vowels are weakly pronounced and have not been transcribed in the glossed examples (except in (90) below). What is important for our purposes is that no vowel is inserted in the word boundary in the position of the case ending.

In Egyptian Arabic on the other hand, consonant clusters are broken by inserting an epenthetic vowel after the second consonant, producing CC-V-C. The epenthetic vowel then typically falls in the position of a potential case ending, resulting in ambiguity. Compare (89), uttered by an Egyptian, to (90), uttered by a Syrian. The word *amr* 'matter' occur here in similar phonotactic contexts, but only in Egyptian Arabic does it result in ambiguity.

(89) *wa-ᶜindama yaʾti* ***ʾal-ʾamr-i*** *ʾila miṣr*
and-when 3MS.come DEF-matter-AMB to Egypt
'But when it comes to Egypt …' (al-Barādiᶜī, 7:17)

(90) *ᶜaskarat-i s-sawra haḍa* ***ʾamər*** *marfūḍ*
militarization-AMB DEF-revolution this matter rejected
'Militarization of the revolution—this is unacceptable.' (al-Asᶜad, 6:40)

One other form of epenthesis in word boundaries that was not described in the consulted works on Egyptian Arabic was found in Egyptian speech in the corpus, a form of epenthesis that arises in the intersection between Standard Arabic and dialectal phonotactics. The sequence VVC-C, a long vowel followed by two consonants is not a possible sequence in Egyptian phonotactics. It is resolved by shortening the long vowel VC-C (Aboul-Fetouh 1969:16; Mitchell 1978a:13). When speakers opt not to shorten the vowel in order to keep close to the Standard Arabic pronunciation, the sequence is instead resolved by epenthesis at the word boundary. If the preceding word is a noun, the epenthetic vowel fills the slot of the case ending, resulting in ambiguity. Examples (91)

and (92) illustrate this. The word *aṣḥāb*, 'owners' in the first example and *māl*, 'capital' in the second were both pronounced with a final *i*, and in neither example is this the prescribed case ending. Such tokens were tagged as ambiguously marked for case for Egyptian speakers.

(91) ***ʾaṣḥāb-i*** *ruʾūs-i* *l-ʾamwāl-i* *l-fāsida* *yuḥāwilūn*
owners-AMB heads-AMB DEF-money-AMB DEF-corrupt try.3MPL
'The owners of corrupt capital try [to stage a counter revolution].'
(Mursī, 19:52)

(92) *wa-lā* *yurīdūna* *ʾan* *yaʾxuḏu* ***māl-i*** *ʾaḥad*
and-not want.3MPL to take money-AMB anyone
'and they do not want to take anyones money.' (Badīᶜ, 27:01)

III. Vowel initial enclitic pronouns. In the dialectal enclitic pronouns in 2m.s. and 2f.s., *-ak* and *-ik*, the vowel marks gender. In the Standard Arabic counterparts, this same vowel position is occupied by the case ending and a vowel following the *k* marks gender: *-uka*, *-aka* and *-ika* in 2m.s. in the three cases. When the word is in pausal form, this final Standard Arabic gender vowel is dropped, producing the forms *-uk*, *-ak*, and *-ik*. The latter two are identical with the dialectal forms and thereby ambiguous. This is so with the word *bayt-ak* 'your house' in (93), where the *a* in the pronoun can be interpreted either dialectally as marking gender, or as a Standard Arabic case marker.

(93) *ʾaᶜṭī-ni* ***bēt-ak*** *il-yōm*
give-me hous-AMB.your DEF-day
'Give me your house today.' (Šallaḥ, 45:16)

The 3m.s. enclitic pronoun is *-u* in the represented dialects. Its Standard Arabic counterparts are *-uhu*, *-ahu*, and *-ihi* in the three cases, giving the pausal nominative form *-uh* which is often indistinguishable from the dialectal form and thus ambiguous, as in (94).[79]

(94) *yaᶜni* *da* *suʾāl* *ᶜana* *mā* *ʾaᶜraf-š* ***ʾigābt-u***
PART that question I NEG 1S.know-NEG answer-AMB.its
'Well, that is a question to which I do not know the answer' (Hilāl, 10:31)

[79] One speaker in the corpus, Muḥammad Kayālī, a Syrian, pronounces this pronoun as *-a(h)* consistently in all case positions, and never *-u(h)* (see 9.2.2), a form used in the geographically neighboring Baghdad Arabic (Jastrow 1980:149; Abu-Haidar 2006:226). It was here taken to be

In Egyptian Arabic ambiguity arises also with other enclitic pronouns when the addition of an enclitic pronoun results in the consonant cluster CCC. An epenthetic vowel, normally *a*, is then inserted (95). Before the enclitic pronouns *-hum* (3m.pl.) and *-kum* (2m.pl.), the epenthetic vowel is *u*, coinciding with the Standard Arabic nominative marker (96) (Aboul-Fetouh 1969: 87; Abdel-Massih 1975:143; Woidich 2006:326). Such vowels were considered ambiguous for Egyptian speakers only.

(95) *bidāyat-an* *bi-t-taṣwīt* *bi-r-raqam-a* *l-qawmi*
beginning-ACC with-DEF-voting with-DEF-number-AMB DEF-national
wa-ġayr-a-ha *mina* *l-maṭālib-i* *s-sabᶜa*
and-other-AMB-her from DEF-demands-AMB DEF-seven
'Beginning with votation with the national number, as well as the other seven demands.' (al-Barādiᶜī, 4:24)

(96) *lammā* *yataʾāl* *il-wuzarāʾ-i* *l-muraššaḥ-īn*
when 3MS.mentioned.PASS DEF-minister-AMB DEF-candidating-PL.AMB
nafs-u-hum
self-AMB-their
'When the candidating ministers themselves are mentioned …' (Hilāl, 2:29)

IV. Case endings with long vowels or diphthongs. In the dual, sound m. pl., and the five nouns, the case ending includes either a long vowel or a diphthong. These paradigms do not have Standard Arabic pausal forms where case is unmarked and they include one form that is identical with the dialectal form, which and then ambiguous.

In sound m.pl., the genitive/accusative pausal ending *-īn* is identical to the dialectal form and ambiguous. In the dual, the genitive/accusative ending is *-ayn*. The Standard Arabic diphthong *ay* corresponds phonemically to the monophthong *ē* in the dialects. It is often difficult to distinguish between the two in rapid speech. Furthermore, what is clearly pronounced as the diphthong *ay* in spoken Standard Arabic may be the result of a phonetic substitution of a dialectal *ē* and thus unrelated to the case system (cf. Hary 1996:81). It is therefore not clear that a dual ending *-ayn*, even when clearly pronounced as a diphthong, marks case. It was, similarly to the *-īn* in sound m.pl., tagged as

a dialectal feature in Kayālī's speech and was tagged as ambiguous in his interview. This is the only case in the study where speech patterns of an individual were taken into considerations in the application of the coding scheme.

ambiguous.

The five nouns take a long vowel in construct state, *-ū*, *-ā*, and *-ī* in the three cases respectively. In the dialects the ending on these words in construct state is *-ū*, a form that is therefore ambiguous. The word *ḏū* 'owner' of this paradigm is exclusively standard and does therefore not have a direct dialectal counterpart. It was nevertheless coded as ambiguous when in the nominative form in analogy with *axū* 'brother [of]' and *abū*, 'father [of]'.

v. Ayy-i. The word *ayy* 'which'; 'any' is often pronounced with a final *i*, the Standard Arabic genitive ending, seemingly regardless of case. The word occurs 51 times in the corpus with final *-i*, evenly distributed between the case positions, and often in situations where epenthesis is not expected, as in (97). This ending on this particular word was therefore taken to be either lexically or phonetically conditioned and tokens in this form were tagged as ambiguous.

(97) *wa-lan* *yastaṯni* ***ʾayy-i*** *ṭāʾifa*
and-NEG.FUT 3MS.make.exception.for any-AMB sect
'And it will not make exceptions for any sect.' (al-Asᶜad, 15:34)

7.8 Code and transcription checking

After the corpus had been annotated as described above, an *R* script (R Core Team 2013) was run on it to generate a database of all annotated tokens in the corpus. Apart from the annotated information, the script also extracts a number of other variables for each token including prefix, word stem, suffix, speaker, beginning and end time of the utterance, preceding and succeeding word, and sequential token number. Figure 6 is an example of how the data is represented in the database.

To improve the consistency of the annotation and transcription, the annotated text went through a checking procedure in which a number of incorrectly transcribed or annotated tokens were identified. This was done in three steps. First, prefixes, suffixes, and labels were checked for typing errors. There are a limited number of possible prefixes, suffixes, and tags, and deviations from these sets were located and corrected. A script was then run on the corpus that checks whether each of the six tag positions is filled with a tag of the appropriate variable so that, for example, the first tag designating headedness is either `HED` or `ATT`.

In the second step, the dependency between word stems and the paradigm tag was checked. This was done by generating word lists from the corpus

FIGURE 6: *Database extract. The example shows ten of the variables in the database generated from the first ten tokens in the corpus extract in Appendix B.*

pref	word	suff	headedness	case	casegov	definiten	paradigm	marking	speaker	timestamp
	áwl	ʿaa	HED	ACC	ADV	IND	TRI	MAR	badi	61301_63530
ʿl	ḥqyqp		HED	ACC	ADV	ART	TRI	UNM	badi	66768_80021
ʿl	ftrp		HED	GEN	PRE	ART	TRI	UNM	badi	66768_80021
ʿl	mʿḍyp		ATT	GEN	PRE	ART	TRI	UNM	badi	66768_80021
	tjrbp		HED	ACC	KAN	IND	TRI	UNM	badi	66768_80021
	ṯryp		ATT	ACC	KAN	IND	TRI	UNM	badi	66768_80021
	mlʿḥẓʿt		HED	NOM	VSO	IND	SFP	UNM	badi	80021_84191
	mrʿjᶜp		HED	GEN	PRE	IND	TRI	UNM	badi	84191_90105
l	bᶜḍ	a	HED	GEN	PRE	CON	TRI	AMB	badi	84191_90105
ʿl	mwʿqf		HED	GEN	ABS	ART	DIP	UNM	badi	84191_90105

with stems tagged for each paradigm. These lists were manually checked to find stems not belonging the respective paradigm. Sound m.pl. and dual were checked together with their suffixes since their stem itself is identical with stems in the triptote paradigm. The word list for the triptote paradigm was too long to be checked manually. Instead, this list was checked for word stems occurring in the manually checked word lists of the other paradigms.

In the third step, dependencies between affixes and the label were checked by tabulating tags in the label to prefixes and suffixes. For instance, the nunation *-un* (in the corpus text uu) should only occur with the case tag for nominative, the declension tag for triptote, the definiteness tag for indefinite, and the case marking tag for marked, or alternatively with the tag for hypercorrect case marking and any combination of the other tags. Tokens appearing in grammatically impossible positions in the tabulations as when the suffix *-un* was tagged as marked for case in the genitive were located and corrected.

When these corrections had been done, the database was generated anew. This procedure does not eliminate all errors, but only reduces their number. It also does not find errors in dependencies across word boundaries, such as mismatches in case annotation between head nouns and attributes, or in the headedness tag itself. Errors in these tags are therefore expected to be more frequent. Several errors were also found in the preliminary analyses. These were corrected in the corpus text, and the database was re-generated from the corpus before the calculation in the final analyses.

Finally, when the transliterated text had been finalized, a 'reader tier' with Arabic script, marked with `%ara:` was generated from each main tier. This tier was stripped of all annotation and pre- and suffix delimiters and the *tāʾ marbūṭa*s were reverted to *tāʾ maftūḥa* where appropriate and then re-transliterated to Arabic script. This produces a tier of a more readable Arabic text edited by the principles outlined in Chapter 6. See Appendix B for an example.

7.9 Summary

This chapter described how the corpus was annotated. The annotation is based on widely used Western grammatical descriptions of Arabic. This was compared to the native grammatical tradition where relevant. The annotation label consists of six tags, each in the form of three uppercase letters. Tokens — nouns and adjectives — were annotated for headedness, case, case governance, definiteness, inflectional paradigm, and realization of the case marker. Of these, case governance is the most diverse, with sixteen different possible values: six for nominative, two for genitive, and eight for accusative. This coding scheme makes it possible to analyze the data according to a number of different grammatical parameters.

A major methodological problem is formal overlap between dialectal and Standard Arabic morphology that sometimes makes difficult or impossible to asses whether a certain ending is in fact intended as a case marker. A detailed description of situations were considered where such overlap occurs and where the token was tagged as having an ambiguous case marker. As a matter of principle, the syntactic context was not taken into consideration for determining whether a case ending is ambiguous, but only the morphological overlap between the dialectal and Standard Arabic systems as it pertains to words in isolation.

When the corpus had been annotated, measures were taken to reduce the number of errors in the annotation. This included generating and manually checking word lists of the inflectional paradigms coded for, and tabulating forms of prefixes and suffixes according to occurrences with tags in the label. A database was then generated from the annotation in corpus text. This database is the primary object of analysis in the following chapters.

Part III
Analysis

Global measures and idiosyncratic variation
Case marking and morphology
Case marking and syntax

Global measures and idiosyncratic variation

8

This chapter explores differences between individuals in their use of case endings and aims at giving a description to the role played by case endings in their respective *speech styles*. Such speech styles are defined as *a set of linguistic practices and patterns characteristic of an individual's speech.* In this chapter, data from each interview is analyzed as a whole and is taken as a sample of how the individual speaks Standard Arabic and thus as an example of his speech style. The next two chapters explore the effect of morphosyntactic features on case marking with variation between speakers controlled for, whereas in this chapter variation between speakers is what is investigated. As was discussed in 2.5, there is considerable variation in the linguistic background of educated Arabic speakers in terms of the form of education and language exposure. This, in combination with the lack of formal regulation of Spoken Standard Arabic in the educational system, means that variation in formal speech styles is to be expected. The extent of some aspects of that variation is explored in this chapter. The speech styles as described here can thus be seen as a first step towards an inventory of possible ways of speaking Standard Arabic, or more specifically, of the range of possible variation observed in situations where Standard Arabic is the expected variety. Furthermore, the speech styles in the corpus are styles of powerful people, and their speech is broadcast by an influential new channel. They are thus likely to be to be taken as role models for speakers developing proficiencies in Spoken Standard Arabic.

The first step in exploring the role of case in the speech styles represented in the corpus is to calculate the proportion of case marking for each speaker. This is done in section 8.1. The consistency of case marking throughout each interview is then investigated in section 8.2. It is shown that speakers generally have

an even use of case endings throughout the interview. The interplay between case endings and dialectal linguistic features is investigated in section 8.3 where, in addition, speech styles of a few selected speakers are discussed in more detail. Section 8.4 is concerned with the use of case endings in fixed expressions, which are shown to be very few. Section 8.5 gives a summary of the findings presented in the chapter.

Some of the results presented in this chapter, particularly the even use of case endings over time and the very limited use of fixed phrases with case marking, are important in evaluating findings presented in the next two chapters. The patterns of case marking in morphological and syntactic situations described there are not conditioned by fixed or set phrases and are representative only of the interviews as a whole.

8.1 Overall rates of case marking

The first step in exploring variation in case marking is to calculate the proportional use of case endings for each speaker. This is done by dividing the number of case markers with the total number of potential case markers. Before this is done, however, there are some potentially problematic effects of the coding scheme on these case scores that need be controlled for. The way this is done is described in 8.1.1, and in 8.1.2 the resulting overall rates case marking are presented and discussed.

8.1.1 The disambiguated dataset

Tokens with ambiguous, hypercorrect and indistinguishable endings, and with word stems that cannot take case endings were filtered out from the calculation of case scores, together with some other items as explained below. The resulting set of data will be referred to as the *disambiguated dataset* and will be the basis of a number of analyses in this and the following two chapters.

Raw counts and proportions of categories of case marking (described in detail in 7.7) are listed in Table 11. The main part, 69.8%, of the almost 15000 annotated tokens[80] are unmarked for case and 9.8% are marked for case. Almost twice as many, 17.0%, are ambiguous as to case marking. The ambiguously marked tokens are problematic in that they represent forms that may have been intended as either marked or unmarked for case. Another 2.6% of the tokens are nominals that are grammatically indeclinable for case. 43 tokens, 0.3% of the data, are produced with prescriptively incorrect case endings and were coded as hypercorrectly marked for case. This small number is surprising seeing the

[80] The type count of word stems is 3075.

Table 11: *Frequencies of types of case marking*

	UNM.	MAR.	AMB.	INDECL.	HYP.	INDIST.	TOTAL
#	10466	1474	2552	391	43	65	14991
%	69.8	9.8	17.0	2.6	0.3	0.4	100[a]

[a] Percentages are rounded and may not add up to exactly 100.

perceived difficulty of the case system. It indicates that speakers only insert case endings when they are certain that they can do so correctly. The final 0.4% of the data are tokens with an ending that is inaudible or otherwise indistinguishable in the recording.

When calculating overall case scores, indeclinable tokens that cannot take a case marker or that have an indistinguishable ending were excluded. Hypercorrect case markers increase the distance from the dialects, but as failed attempts to produce correct case their role as markers of Standard Arabic is ambiguous. The choice was made here not to include them in the calculation of the case marking rate and to instead analyze them separately (see 10.6). Tokens with indeclinable, hypercorrect, and indistinguishable endings together make up a total of 2.3% of the data and would only have marginal effects on the scores if included.

More problematic is the exclusion of tokens with ambiguous endings (see 7.7.1). These are endings that can be interpreted both as Standard Arabic case endings and as dialectal and therefore caseless forms. Two common examples are *-a* and *-i* preceding the definite article, and the sound m.pl. ending *-īn*. Such forms were coded as ambiguous regardless of syntactic position. These tokens constitute 17% of the total data and are, in effect, an analytical blind-spot and were thus excluded. However, excluding all tokens with ambiguous endings when calculating proportions of case marking risks yielding misleading results, since some categories of tokens cannot be technically unmarked for case, only marked or ambiguously marked. Compare the coding of endings in the triptote and the sound m.pl. paradigms in the three case positions, schematically presented in Tables 12 and 13. In the triptote paradigm, the options are normally between a marked and an unmarked form, whatever the case, with hypercorrect forms being very rare. For tokens in the sound m.pl. the situation is very different. Only in nominative can case be unambiguously marked, and in no grammatical case can it be clearly unmarked for case. This means that using words in the sound m.pl. can only raise the rate of case marking as it is counted

Table 12: *Case marker coding for indefinite triptotes*

	NOM.	GEN.	ACC.
-un	**Mar.**	*Hyp.*	*Hyp.*
-in	*Hyp.*	**Mar.**	*Hyp.*
-an	*Hyp.*	*Hyp.*	**Mar.**
-Ø	Unm.	Unm.	Unm.

Table 13: *Case marker coding for sound m.pl.*

	NOM.	GEN.	ACC.
-ūn	**Mar.**	*Hyp.*	*Hyp.*
-īn	Amb.	Amb.	Amb.

here, and never lower it when tokens with ambiguous endings are excluded. Similar situations arise in the dual and in the five nouns where the dialectal form (with final *-ēn* and *ūn* respectively) is similar or identical to one Standard Arabic case marked form. The same problem arise in words with the 3m.s. and 2m./f.s enclitic pronoun. The dialectal forms of these endigns (*-u*, *-ak* and *ik*) are identical to Standard Arabic case marked forms in pause.

Categories of words such as the sound m.pl. that do not have an unmarked realization were excluded from the calculation of case marking altogether since including them would risk giving misleading results. Tokens excluded from the disambiguated dataset for this reason are:

(*a*) sound m.pl. (e.g. *muslimūn* 'Muslims')
(*b*) dual (e.g. *muslimān* 'two Muslims')
(*c*) the five nouns in construct state (e.g. *axū-nā* 'our brother')
(*d*) tokens with the 3m.s. or 2m./f.s. enclitic pronoun (e.g. *kitāb-uh* 'his book'; *kitāb-ak/-ik* 'your (m./f.) book')[81]

A weakness of this method is that it negatively affects the scores for speakers who rely on the sound m.pl. to include case endings in their speech. al-Muᶜallim for example marks case a total of thirty times (not counting adverbs). Five of these are in the form of *-ūn* in sound m.pl. and are excluded in the case score. If the sound m.pl. were included, al-Muᶜallim would have a case marking of 5.2% instead of his current 4.1% and would take the seventh, rather then the eighth rank in overall case marking (see below).

[81] The 3m.s. and 2m/fs. enclitic pronouns are not separated from other forms of enclitic pronouns in the coding scheme. They were filtered out by excluding tokens with suffixes of certain forms in the corpus text (`uh`, `ak`, `ik`, etc.). Theoretically, 3m.s. and 2m/fs. enclitic pronouns

Furthermore, indefinite adverbs are generally lexicalized with an accusative case ending. The ending on these adverbs is therefore better seen as lexicalization and not as part of the case system. Some adverbs however do not take accusative endings as part of word formation but show similar variation in case marking as other nominals. These include circumstantial qualifiers (*al-ḥāl*) and elative specification. The coding scheme does not allow for a separation between adverbs with varying and fixed case marking and all are therefore excluded here. Adverbs and several other types of tokens excluded from this dataset are analyzed separately in Chapters 9 and 10.

The disambiguated dataset that is the result of this filtering contains 11109 tokens. This is the sum of tokens coded as marked and unmarked for case, minus the tokens of categories *a* to *d* above. This dataset is the basis for calculations of overall case marking as well as several other analyses in this and the next two chapters.

8.1.2 Calculated overall case rates of case marking

After filtering out the items described above, rates of case marking for each speaker were calculated on the resulting disambiguated dataset. Table 14 lists marked case as counts, percentages, and log-odds, together with the standard errors of the log-odds for each speaker, calculated with a Generalized Linear Model (GLM) on the binomial data of marked and unmarked case in the disambiguated dataset. The total number of tokens for each speaker is also listed. al-Barādiᶜī for example, the speaker with the median rate of case marking, marks case 26 times in the disambiguated dataset. This represents 3.9% of all potential case markers in the interview, which is equal to a log-odds of –3.20 with a standard error of 0.20. These numbers are based on 878 tokens. In the following, most of the statistical modeling is performed on the log-odds, while percentages are generally more interpretable. Frequencies of case marking will be reported in both units.

The overall case marking is a measurement developed here for purposes of comparisons between speakers. It should not be taken as an absolute measure of case marking, for two reasons. First, dialectal word stems were excluded at the annotation stage (see 6.5) and so are not part of the data. A speaker can therefore potentially use a large number of dialectal words that cannot take case endings without this effecting the overall case marking, as long as they mark case on other words. Such a speaker would have a high case score even if they use relatively few case markers in relation to the number of words spoken.

could be realized with unmarked case, such as *kitāb-hu* and *kitāb-ka/-ki*, in analogy with other enclitic pronouns, but no such forms occur in the original transcripts.

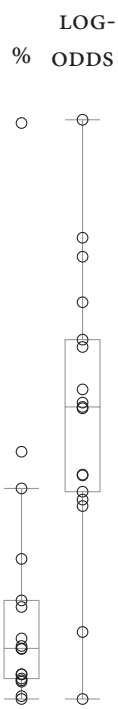


Table 14: *Overall rates of case marking*

	# mar.	%	LOG-ODDS	SE	*n*
Tayzīnī	218	42.2	−0.31	0.09	656
Badīᶜ	130	18.3	−1.50	0.10	1010
al-Qaddūmī	74	15.6	−1.69	0.13	609
Abū Majd	62	10.5	−2.15	0.13	837
Mursī	40	7.4	−2.52	0.16	783
Kayālī	44	6.9	−2.60	0.16	857
Fayyāḍ	38	4.6	−3.02	0.17	1108
al-Muᶜallim	18	4.1	−3.16	0.24	582
al-Barādiᶜī	26	3.9	−3.20	0.20	878
ᶜAbd al-Qādir	25	3.9	−3.21	0.20	896
al-Xuḍarī	17	2.0	−3.88	0.25	1051
Hilāl	13	2.0	−3.89	0.28	921
al-ʾAsᶜad	9	1.7	−4.05	0.34	757
Šallaḥ	19	1.6	−4.13	0.23	1583
Ġalyūn	10	1.5	−4.20	0.32	908
ᶜArīqāt	3	0.4	−5.46	0.58	956
al-Miṣrī	1	0.2	−6.14	1.00	598
Mean	44	7.5	−3.24	0.27	882
Median	25	3.9	−3.20	0.20	878

Second, pause (see 4.5) is not taken into consideration here. Of all tokens in the corpus, 15% occur before an intentional utterance termination, that is, in an utterance that is not interrupted or tailed off. The prescribed form for words in this position is the pausal form, which for many types of word have no case ending. This means that a person who speaks prescriptively correct Standard Arabic, marking case wherever possible except in pause position, would not have an overall case marking of 100% by this measure, but of approximately 85%.

The scores are not very informative in and of themselves in that they tell nothing of *how* case is marked; how case markers are distributed within the interview or how they correlate with morphosyntactic features. The numbers do however give an orientation on the kind of idiosyncratic variation there is with regard to case marking, and they provide a measure of case usage to which

other variables can be related. In the following and where applicable, data on speakers will be presented as ordered by these numbers.

This variation between speakers in case marking, plotted in the margin next to Table 14, is indeed one of the most striking features of the case scores. Scores ranges from al-Miṣrī's 0.2%, to Tayzīnī's 42.2%, the former representing one single observation of marked case in the disambiguated dataset. There is a concentration of speakers in the range from one and a half to five percent, with nine of the seventeen speakers falling within this range, and two speakers below it. Above this range five speakers are spread out more or less evenly up to the speaker with the second highest score, Badīᶜ, with 18.3%. The highest scoring speaker is Tayzīnī with 42.2%, more than double the score of any other speaker.

A much larger number of speakers would be needed to draw any definitive conclusions about speaker background and case marking. The correlation between nationality and case marking was tested on the disambiguated dataset a Generalized Linear Mixed-Effects Model (GLMM) with speaker as random effect.[82] The differences between the predicted rates for the three nationalities are not significant ($p>.79$).

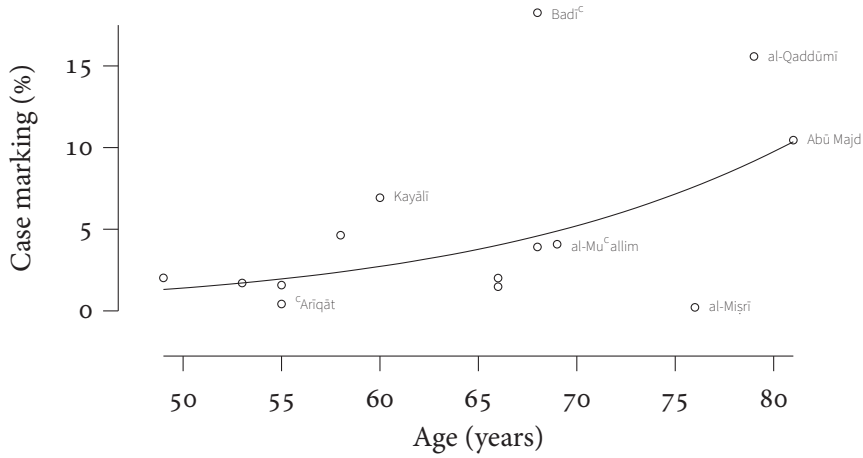
There is however a significant effect of age on case marking. This was tested in a GLMM as that above. It predicts an increase of 0.07 log-odds in case marking for each one year increase in age ($p=.047$). This is plotted in Figure 7, with the predicted rates of case marking transformed to percentage. There are several possible interpretations of this pattern. One is that older speakers with more experience in appearing in public have developed a higher degree of proficiency in using Standard Arabic in general, and its case endings in particular. Another is that speakers of the older generation grew up in an era where more emphasis was put on 'correct' Arabic as an expression of authority and legitimacy.

The widespread assumption that religiously conservative persons are particularly proficient in Standard Arabic is partially confirmed. The two speakers affiliated with the Muslim Brotherhood, Badīᶜ and Mursī, both have high scores, ranking second and fifth of the seventeen speakers. They are however nowhere near the score of Tayzīnī, a radical Marxist, and are matched also by several speakers representing secular organizations.

In sum, the numbers presented above illustrate the very different ways individuals face the challenges of being expected to speak Standard Arabic ex-

[82] GLMMs were conducted with the *lme4* package in *R* (Bates D. et al. 2014) and *p*-values were extracted with the *lmerTest* package (Kuznetsova et al. 2014). All GLMMs are logistic regressions on the binomial variable of case marking, and random effects refer to random intercepts only. See Jaeger (2008) for an introduction to logistic GLMMs.

FIGURE 7: *Case marking and age. The data point representing Tayzīnī at 42.2% is omitted in the plot. Aga data is missing for two speakers.*



temporaneously, and the very different approaches people take in dealing with the case system. Some speakers make dense use of case endings while others use only a few case endings in the entire interview. Most speakers have a style of speaking where case endings are clearly noticeable but sporadic.

8.2 Consistency in case rates of marking

Several researchers have noted that speakers in formal settings often start off with a high level of Standard Arabic and then gradually slide into dialect. As noted in Walters (2003:92), such a pattern can be interpreted in two ways. The first is that speakers begin with establishing their legitimacy by demonstrating that they control the standard variety, and then introduce more and more dialect to signal solidarity with their audience. Secondly, it can be interpreted as speakers not having the proficiency required to maintain Standard Arabic for any length of time and they are therefore forced to revert to their dialect. Meiseles (1977:189n) noticed this in his material especially when it comes to case and mood endings, but explains this as speakers going from scripted to unscripted speech. Schulz (1981:156-160), who termed this phenomenon 'discourse drift', investigated three long monologues in his material and found that for six features that could be analyzed as binary variables of standard or dialectal rep-

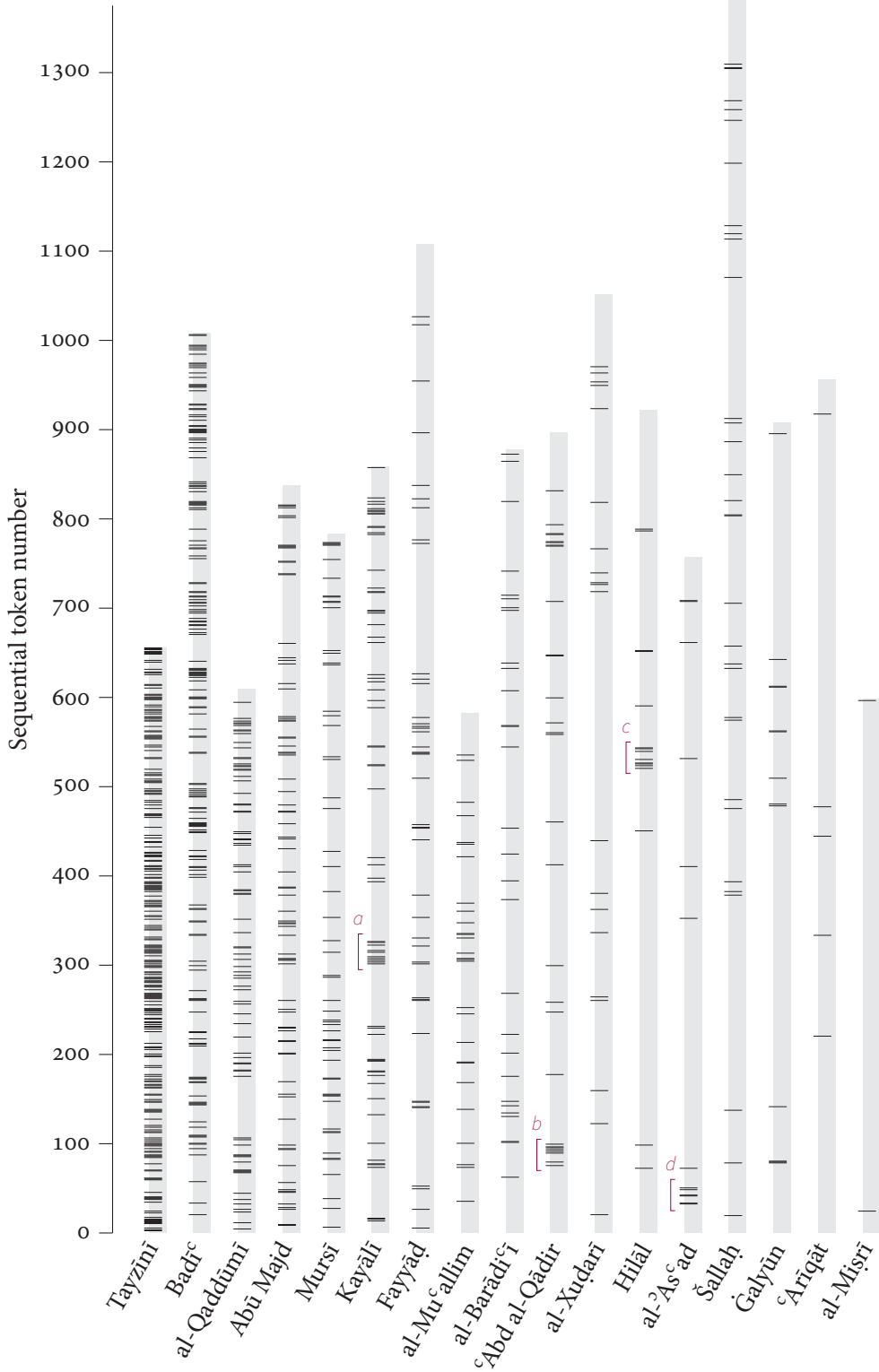
resentation, the relative rate of Standard Arabic variants markedly decreased in the second half of the monologues. Mejdell (2006:195), on the other hand, found no evidence of gradual transition to dialect in her material, except with one speaker who at one point abruptly shifted from Standard Arabic to dialect.

Case markers are a salient marker of Standard Arabic and are thus an important tool for establishing one's control of Standard Arabic, but are also considered difficult and unwieldy by Arabic speakers. If it assumed that there is a general pattern of speakers performing in Standard Arabic gradually drifting to dialect, this is expected to have an effect on the use of case endings. If the drift is explained by lack of proficiency, case endings would be one of the first features to be affected, being a layer of grammatical marking that is applied to one's speech and that requires attention to formal linguistic features. If, on the other hand, discourse drift is explained by speakers first establishing their mastery over Standard Arabic as a claim to legitimacy and then shift to signal solidarity by using dialectal forms, then case endings, with their symbolic value as markers of Standard Arabic (see Chapter 4.4) are expected to be used in the first part and then be dropped by the speaker so as not to come across as pedantic. However discourse drift is explained, it is likely to affect case endings.

To investigate discourse drift in the corpus, the temporal distribution of case endings was plotted for each speaker, giving Figure 8. Time is represented not in absolute terms but by sequential token number in the interview. This is because the database has no reference to absolute time of individual tokens, only to whole utterances. The gray bars represent the length of the interview in terms of annotated tokens and the black lines mark instances of marked case. All unambiguously marked case endings are plotted except for endings in indefinite adverbs. Note that this includes some tokens excluded in the disambiguated dataset. Speakers are ordered from left to right by decreasing case scores. Thus, Tayzīnī's heavy use of case endings is represented by the bar in the far left with roughly half of it covered with in black lines with his 42% case marking. Case marking becomes increasingly sparse further to the right as overall case marking decreases, with al-Miṣrī's bar on the far right showing only two case markers, one at the beginning and one at the very end of the interview.

Visual inspection of the figure shows little sign of speakers starting off with more case marking and then transitioning into speech with less case marking. The only such pattern is in the beginning of al-As^{c}ad's interview, in a cluster of case markers labeled *d* in the figure. In the interview, he is first asked to describe the Free Syrian Army, of which he is the commanding officer. He begins with invoking blessings for the victims of the war and with praising the courage of the Syrian people and the soldiers in the Free Syrian Army, all in a formalized

FIGURE 8: *Temporal distributions of case markers*



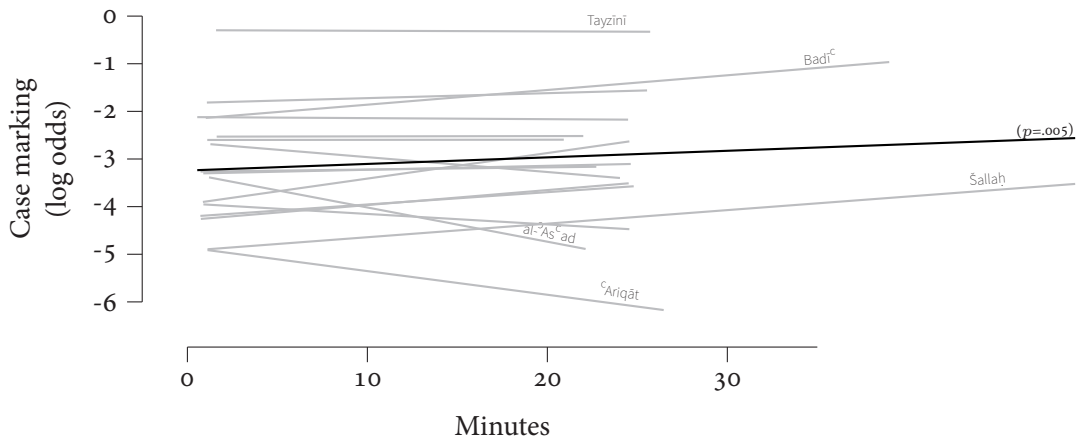
way typical of this kind of interview. In this first segment, there are no case endings. He then starts to speak more freely to answer the question, and this is where he produces seven case markers in close sequence, on the first seven instances of tokens with an enclitic pronoun. There were no enclitic pronouns in the previous introductory segment. The eighth and ninth tokens with case endings are hypercorrect, both occurring in the same utterance (98). Here the noun *azlām* 'men', 'henchmen' occurs twice with the erroneous nominative ending *-u*. The first is in genitive position after the preposition *min* 'from', and the second is in accusative position after the complementizer *anna*.

(98) *naštari min* ***ʾazlām-u*** *n-nizām baᶜḍ-a l-ʾasliḥa ʾallati*
1PL.buy from men-HYP DEF-regime some-AMB DEF-weapons which
naḥtāj ʾilay-ha li-ʾanna n-nizām yaᶜrif bi-ʾanna
1PL.need on-it for-COMP DEF-regime 3MS.know by-COMP
ʾazlām-u-hu *yabīᶜūna-hu bi-d-darāhim-a l-qalīli*
men-HYP-its sell.3mpl-it for-DEF-dirhams-AMB DEF-few
'We buy some of the weapons we need from the men of the regime,
because the regime knows that its men are selling it for a few dirhams.'
(al-Asᶜad, 3:24)

After this sentence, al-Asᶜad abandons this way of speaking with consistent case marking on words with an enclitic pronoun. In the rest of the interview, he only marks case six times and these instances are evenly spread throughout the discourse. This first part with dense case marking was probably an attempt to imitate the speech style of speakers with medium to high overall case marking who fairly consistently mark case on words with an enclitic pronoun (see 9.2). al-Asᶜad attempts to live up to this norm, maintains it for roughly a minute but then makes mistakes and abandons it. This is the only clear case found in the corpus of a speaker aiming at a rate of case marking he is then not able to maintain.

The presence of discourse drift in the corpus and its effect on case marking was further tested for each speaker in a GLM on the disambiguated dataset according to the point in the interview of elapsed time in minutes as the predictor. Since there is no reference to the exact time of specific tokens, only to whole utterances, time elapsed refers the beginning of the utterance in which the individual token occurs. In this test, a negative effect of elapsed time on case marking would indicate the presence of discourse drift. A positive or no effect would indicate that there is no discourse drift. Figure 9 shows the regression lines of the model as applied to each speaker except for al-Miṣrī', whose single instance of case marking in the disambiguated dataset makes his interview

FIGURE 9: *Change in rates of case marking over time. Gray lines represent individual speakers and the black line the overall prediction.*



unsuitable for regression analysis. The figure shows no clear general effect of time elapsed on speakers' use of case endings. Some speakers use more case endings as the interview progresses and some less. Of all these effects only the increase of case marking in Badīᶜ's interview is significant ($p<.001$). There may of course be true effects that cannot be verified in this test due to lack of data. The sharp downward slope in al-Asᶜad's speech for example did not yield a significant result in the linear regression model, but as was shown above, a closer inspection of his interview clearly shows that there is a change of case marking in his speech. Even with such unverified effects, the different directions of the slopes suggest that there is no strong overall effect of elapsed time on case marking. This overall effect was tested in a GLMM on the entire dataset and with the speaker variable as a random effect. The random effect adjusts for individual speakers for whom there is more data disproportionally effecting the estimation. This model showed a statistically significant ($p=.005$) small positive effect of 0.01 log-odds per minute. This is represented by the black line in Figure 9. Recalculated to percentage, the predicted case marking is increased by around half a percentage point every 10 minutes. This is too small an effect to be of much interest in and of itself, but it does show that as far as case markers are concerned there is no general pattern of speakers shifting away from Stan-

dard Arabic during the course of the interviews. Rather, the effects that can be statistically verified show the opposite. This result is in line with the findings in Mejdell (2006:195) and contradicts both popular belief and most academic opinions.

Another possible pattern of changes in case marking is that case markers are clustered in different parts of the interview. This might, for instance, be so if a topic is dealt with in which the speaker takes a particularly authoritative stance and signals this with elevated language. If such clusters are somewhat evenly spread out they would not be captured by the linear regression models described above. Figure 8 above shows a few suspected clusters. Naturally, speakers need to produce a moderate to low rate of case markers for clusters to be distinguishable from surrounding material. Four suspected clusters, labeled *a–d* in the figure, were investigated in the transcripts. Cluster *a* is a repetition of identical phrases and case endings but with different head nouns (99). Clusters *b* and *c* are similar rhetorical repetitions in direct sequence of identical or very similar phrases with the same case ending. These repetitions are described in more detail in 8.4. The repetitions in these constructions mark a forceful rhetoric stance, and having case markers as part of the repetition enforces that effect. It does not seem to be that case endings are limited to certain parts of the interviews, or that there are specific themes or topics that call for a shift of style and are characterized with especially dense case marking. This can be interpreted as speakers consistently speaking in their highest register with no shift further upwards.

(99) *ʾinna-ha jamāhīr muʾmina* ***bi-qadiyyat-i-ha*** *muʾmina*
PART-it people believing in-cause-GEN-their believing
bi-ʾinsāniyyat-i-ha *muʾmina* ***bi-qawmiyyat-i-ha*** *muʾmina*
in-humanity-GEN-their believing in-patriotism-GEN-their believing
bi-rabb-i-ha *muʾmina* ***bi-dīn-i-ha***
in-God-GEN-their believing in-religion-GEN-their
'They are masses who believe in their cause, believe in their humanity, believe in their patriotism, believe in their Lord, believe in their religion.' (Kayālī, 7:35)

There is thus no general pattern in the corpus of speakers reducing their level of case marking as the interview progresses. Speakers generally maintain a consistent rate of case marking throughout the course of the interview, whether this rate be high or low. This can be interpreted to mean that part of developing a speech style in Standard Arabic is to find a level of case marking that can be comfortably maintained for a longer stretch of continuous speech. The one ob-

servation of discourse drift that was found was in the interview with al-As^{c}ad, a speaker with relatively little experience of appearing in public. He only recently became a public figure, formally forming the Free Syrian Army in July 2001,[83] only five months before this interview took place. In this limited time, he may not have acquired the experience in speaking Standard Arabic required to know what degree of case marking he can maintain. All other speakers in the corpus, none of whom show a pattern similar to that in al-As^{c}ad's interview, have long political or academic careers behind them. The lack of discourse drift and the even distribution of case endings in the interviews suggest that the speakers in the corpus have identified a level of case marking, be it with dense or sparse, with which they are comfortable and that they can maintain.

8.3 Dialectal features

In this section, the interplay between case markers and dialectal features in the various speech styles represented in the corpus is investigated. The extent to which speakers mix case endings with saliently dialectal features is statistically explored, with direct morphosyntactic constraints between the two variants controlled for. It is shown that while there is a negative correlation between the use of dialectal elements in a speech style and case marking, there is wide variation in the interplay between case endings and dialectal features in their speech styles.

The use of case endings in speech is a salient marker of Standard Arabic. It is, in some respects, grammatically incompatible with dialectal features. On the morphological level, standard affixes are not attached to dialectal stems, as has been consistently shown in previous research (see 7.6.1). In the present corpus no instance of a case marking was found on a stem coded as dialectal. As an example of a constraint on the syntactic level, Eid (1982, 1988) has shown that in Egyptian-Standard codeswitching, dialectal verbal negators are not followed by saliently standard words, whereas there are no such constraints after dialectal relative pronouns or complementizers. An example of syntactic constraints on the combination of standard and dialectal features from the present corpus is that nominals coded as dialectal constituents (as defined in 7.6.1) are not marked for case even if the nominals themselves are of a Standard Arabic form. Of the 307 dialectal constituents in the corpus only one is marked for case (see example (103) on page 187).

The use of case endings is thus, in some situations, directly constrained by the presence of dialectal features. The question investigated in this section

[83] *http://www.aljazeera.com/indepth/features/2012/10/201210131150295831 26.html.*

Table 15: *Measures of dialectal usage*

	NOMINAL		CONST.		REL. PRON.	
	#	PTW	#	PTW	#	%
Tayzīnī	0	0	0	0	0	0
Badīᶜ	0	0	0	0	1	2.6
al-Qaddūmī	0	0	0	0	1	3.6
Abū Majd	48	23.1	40	19.3	5	41.7
Mursī	4	2.3	11	6.2	6	30.0
Kayālī	9	4.7	3	1.6	0	0
Fayyāḍ	6	2.4	26	10.6	9	32.1
al-Muᶜallim	0	0	0	0	0	0
al-Barādiᶜī	0	0	0	0	0	0
ᶜAbd al-Qādir	1	0.5	28	12.7	1	6.2
al-Xuḍarī	0	0	4	1.7	2	6.1
Hilāl	42	20.8	52	25.7	16	76.2
al-Asᶜad	10	5.8	8	4.7	8	26.7
Šallaḥ	20	4.8	17	4.1	8	17.8
Ġalyūn	4	2.1	27	13.9	12	48.0
ᶜArīqāt	10	3.7	19	6.9	3	17.6
al-Miṣrī	89	51.4	64	37.0	15	75.0
Mean	—	7.1	—	8.5	—	22.6
Median	—	2.6	—	4.7	—	17.6

PTW = per 1000 words

is whether the presence of dialectal features affects case marking beyond the direct morphosyntactic constraints, or, to put it differently, whether there is apart from the grammatical constraints a stylistic constraint that limits how the dialect can be combined with case endings in Standard Arabic speech styles. To test this, the scores of case marking in the disambiguated dataset were compared with measures of three dialectal features in the speech of each of the seventeen speakers in the corpus, listed in Table 15 with speakers ordered by decreasing overall case marking.

In instances where there are grammatical dependencies between dialectal features and case marking, the data was filtered to avoid them. This way, there is minimal grammatical interaction between the predictor (dialectal feature) and

the dependent variable (rate of case marking). The most important such filtering is that case scores are for these models calculated with dialectal constituents excluded, since this is one of the dialectal features. The scores of overall case marking are therefore slightly different from those presented in Table 14 above. What is tested is thereby not whether dialectal features restrict the possibility of marking case on the grammatical level (which they do), but to what extent case endings and features of the dialect can coexist in one and the same speech style.

The three dialectal features investigated are (*a*) nominals; (*b*) constituents, and (*c*) relative pronouns. These features were chosen mainly because they can be automatically counted in the transcripts and in the database. Other important features that are not investigated here are, for example, dialectal verbs and particles, the verbal *b*-prefix, and the realization of /q/ or of the interdental fricatives /ḏ/ ans /ṯ/. Extracting data for these features would require further annotation and techniques that are beyond this project.

The first of these three features, dialectal nominals (see 6.5), are nouns and adjectives either formed of non-standard roots or patterns (e.g. the Egyptian *ḥāga* 'thing' instead of the standard *šayʾ*) or stems used with a meaning that is not associated with that stem in Standard Arabic (e.g., *wāḥid* 'one' used as an indefinite pronoun). Phonological variations of Standard Arabic stems were not considered dialectal. This means that nominals counted as dialectal are words that clearly and unambiguously stand out as such. Only nominals were coded as dialectal in the corpus and are therefore countable. The figures do therefore not include dialectal verbs or particles.

The second dialectal feature is dialectal constituent (see 7.6). These are constituents governed by (*a*) existential *fī* 'there is'; (*b*) dialectal negators (e.g. *miš* 'is not'), and (*c*) the complementizer *innu*.[84] These constituents are governed by non-standard constructions and thus have no prescribed case ending.

The third dialectal feature whose correlation with case marking is tested is the realization of relative pronouns. Relative pronouns were not coded as standard or dialectal in the corpus but were counted as character strings in the

[84] In the coding scheme there is also a fourth type of dialectal constituent, those of the dialectal pseudo verbs *bidd-* (Lev.) and *ᶜāyiz* (Eg.) 'want'. They are excluded from the count of dialectal constituent here since *bidd-* and *ᶜāyiz* were also marked up as dialecal nominals, giving an overlap between the two features. Consider the phrase *bidd-u muṣālaḥa* '[he] wants reconciliation'. This would be counted as a dialectal feature twice unless this measure was taken, since *bidd-* 'want' is both a dialectal nominal and takes a dialectal constituent. Only two speakers are effected by this filtering, al-Miṣrī's and Hilāl. Other speakers use *bidd-* or *ᶜāyiz* only as auxiliary verbs that do not take a nominal complement.

Table 16: *Forms of the Standard Arabic relative pronoun*

	SIN.	DUAL		PLURAL
		nom.	acc./gen.	
m.	*allaḍī*	*allaḍāni*	*allaḍayni*	*allaḍīna*
f.	*allatī*	*allatāni*	*allatayni*	*allātī/allawātī*

corpus text.[85] The relative pronoun is realized either dialectally as *illi* or as one of its Standard Arabic counterparts: *allaḍi*, *allatī*, etc, as listed in Table 16. The dialectal *illi* takes the same form regardless of gender, number and case. The Standard Arabic relative pronoun is more complicated in that it is marked for gender and number, and in the dual also for case. The one form of a dialectal relative pronoun thus corresponds to a set of eight different Standard Arabic forms.[86]

Data for dialectal features are listed in Table 15 above. Each feature is listed firstly as a raw count. Dialectal nominals and constituents are also listed as occurrences per 1000 words (PTW) to compensate for the difference in length of the interviews. Relative pronouns are a binary variable as either standard or dialectal realization, and therefore listed as percentages of dialect realization rather than as counts in PTW. In the interview with Fayyāḍ, for example, there are six dialectal nominals, which is calculated to 2.4 nominals PTW. (The low counts of dialectal nominals reflect the restrictive definition of dialectal word that is applied here.) There are 26 dialectal constituents in his interview which translates to 10.6 dialectal constituents PTW. Fayyāḍ realizes the relative pronoun as *illi* nine times, which is 32.1% of all relative pronouns in the interview. Other relative pronouns are realized in one of the Standard Arabic forms.

8.3.1 Case marking and the diglossic continuum

Table 15 above shows a pattern of increasing dialectal features the lower the case score as shown by higher numbers in the lower part of the table. The speaker

[85] I.e. `ʿlḍy`, `ʿalty`, etc. for standrard, and `ylly` for dialectal variants, using search patterns allowing for optional prefixed conjunctions and prepositions. Before counted, variations in spelling of *illi* in the corpus text were all changed to `ylly` (representing يلي), its most common orthographic form in the corpus.

[86] The relative pronouns *mā* 'what' and *man* 'who' are not accounted for here. These formally coincide with interrogative pronouns and cannot be trivially differentiated from these.

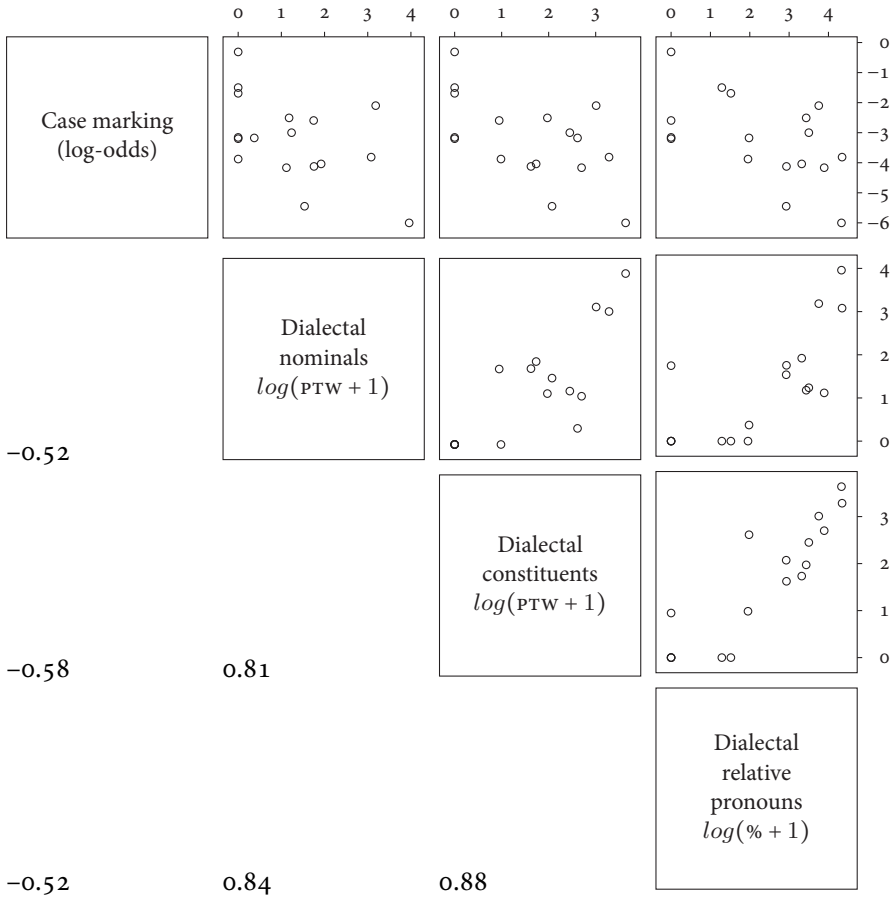
with the highest case score, Tayzīnī, uses none of the investigated dialectal features. The next two, Badīᶜ and al-Qaddūmī, use the dialectal relative pronoun *illi* once each and none of the other two features. Conversely, the speaker with the lowest case score, Miṣrī's, has high values on all three dialectal features. There is however no smooth transition from fewer to more dialectal features as the case score decreases. Abū Majd, for example, has the fourth highest rate of case marking but also the third highest score of dialectal constituents and of dialectal relative pronouns, and the second highest score of dialectal nominals. Two speakers in the mid range of case marking, al-Muᶜallim and al-Barādiᶜī, do not use any of the counted dialectal features. There is, in other words, some variation in how speakers mix case endings and dialectal features in their speech styles.

Table 15 above shows a strong correlation between the three measured dialectal features. If a speaker scores high on one, he tends to score high also on the other two. The dialectal features are closely bound up with one another. This is illustrated in the three bottom right scatterplots in Figure 10 where the data on the dialectal features from Table 15 are plotted, with each circle representing one speaker. Since the data are highly skewed towards the lower end, all values were logarithmic transformed after being incremented by one in order to avoid zero value observations. The strong correlation between the use of the three dialectal features can clearly be seen in the angular distribution of the plots. The Pearson correlation coefficients for the three pairings of dialectal features are all high: 0.81, 0.84, and 0.88 respectively.

This is in accordance with the 'inter-speaker consistency' discussed in Mejdell (2006:176). She found that even if speakers react differently to the same situation in terms of the amount of dialect and Standard Arabic they use, "speakers tend to rank equally high or low in their relative usage level of [Standard Arabic] and [Egyptian Arabic] variants on all features." Dialectal features are used to different degrees by speakers but they pattern roughly the same way. In other words, speakers place themselves somewhere within the diglossic continuum (see Chapter 2) and this position is roughly the same, as related to other speakers, with regards to grammatically unrelated linguistic features.

Now, in the various theories of Arabic diglossic variation, case endings have been argued to position speakers on the same diglossic spectrum as do binary standard-dialect features such as the use of dialectal words or affixes (see 2.2). A speaker may raise his speech towards the standard by ridding it of dialectal features, as well as by adding case endings. The use of case endings, however, shows much less inter-speaker consistency. This is shown in the top three plots in Figure 10 were the use of case endings is plotted against the

FIGURE 10: *Case marking and dialectal features: correlations*



use of the three dialectal features. There is, as expected, a negative correlation; speakers with more dialect features use less case endings. All three effects are significant however (p=.023; .005; .001). This should not be interpreted as three different results since the three scores of dialectal features are highly correlated. Rather, the three plots together show the trend of decreased case marking with increased use of the dialect as measured in three different ways. These correlations are, however, much weaker that those between the dialectal features themselves. This gives a wide spread of data points in the plots and low corre-

lation coefficients: –0.52, –0.58, and –0.52 respectively.

Thus, while dialectal features are tightly bound up with one another, so that the use of one dialectal feature strongly implies the use of the others, case endings are used in a way that is related to, but different from, dialectal features. This indicates that the use of case endings varies due to partly different factors than do other diglossic variables. The usage level of case endings cannot be accurately predicted from measuring the use of dialectal features.

This can be illustrated by comparing the overall case marking of speakers with very little dialect in their speech style. There are five speakers who either make no or one single use of one of the dialectal features. The hypothesis of inter-speaker consistency, if taken to include case endings, predicts that these speakers use more or less equal amounts of case endings. On the contrary, their case scores range from –3.20 to –0.31 log-odds, or 3.9% to 42.2%. Thus, the use of case endings is only vaguely related to the use of dialectal features.

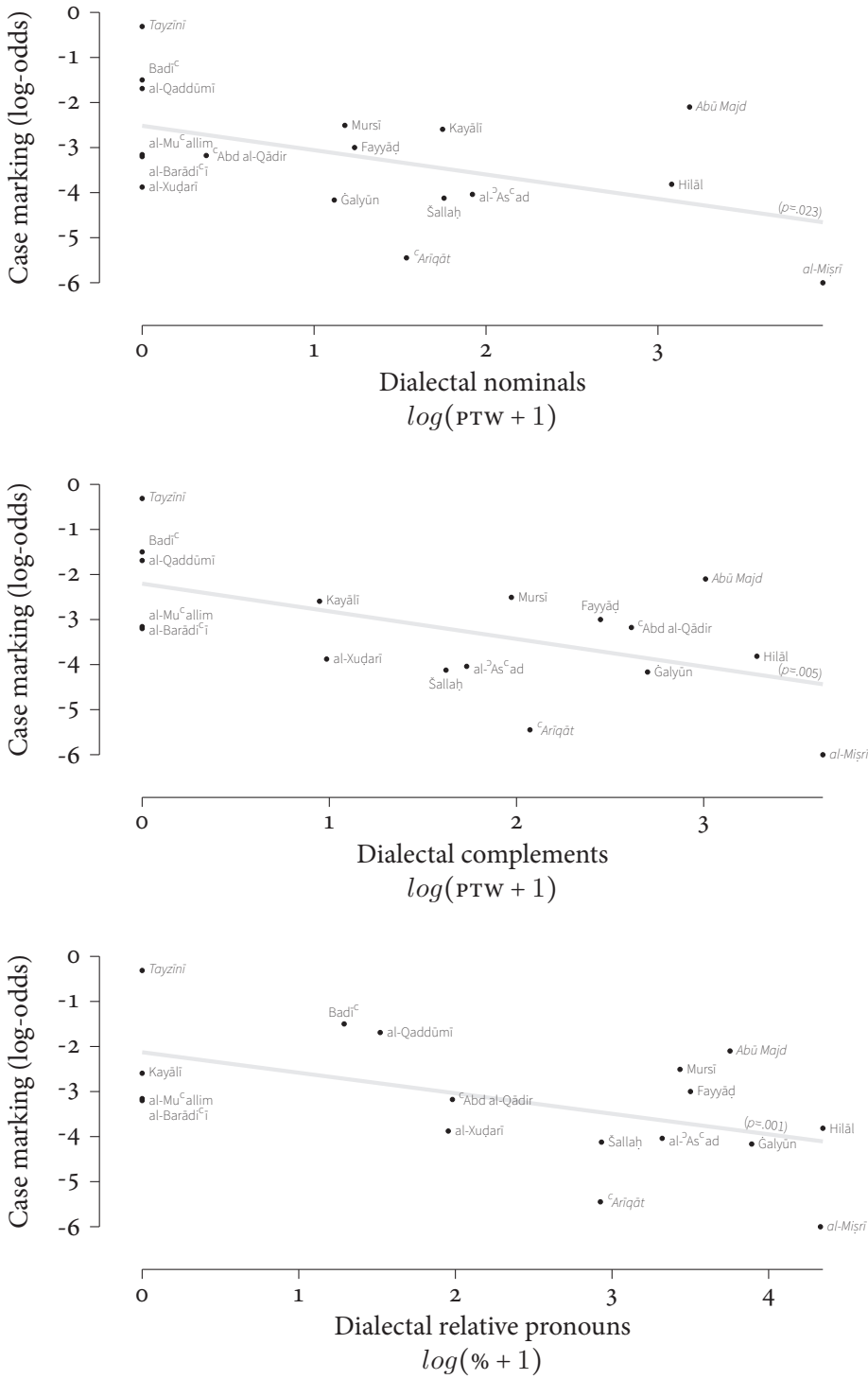
8.3.2 Case and dialect in individual speech styles

Since the use of case endings varies to a large extent independently from other diglossic variables, it provides a space in which speakers develop individual and original styles of speaking. This is illustrated in Figure 11 on the facing page. where overall case marking is plotted against the use of each dialectal feature. The three plots in the figure are equivalent to the three plots on the top row of Figure 10, enlarged to fit speaker labels and ordered vertically.

Each plot in Figure 11 shows a trend of increased use of the dialectal feature being accompanied by a decrease in case scores. The correlation between the dialectal features shows in speakers tending to have similar positions along the horizontal axis in all three plots. The overall rates of case marking are less related to the dialectal features, which shows in the vertical spread of data points on any given point on the horizontal axis.

Apart from showing the general trend in the data, the plots in Figure 11 also provide a visualization of individual speaker's speech styles and how they relate to the statistical norm. Each speaker is represented by one data point which can be related to the predicted score of the regression line. A speaker who is plotted near the regression line in the three plots speaks in a way that is predicted by the model with regard to the combination of dialect and case endings in his speech style. He may use few or many dialectal features but still speak in a statistically predicted way in that he uses case endings to a corresponding high or low degree. Hilāl, for example, is positioned to the far right on all three plots since he makes heavy use of the dialect. He is not far from the regression line and so uses expectedly few case endings. Similarly, Fayyāḍ is in the mid range

FIGURE 11: *Case marking and dialectal features: individual speakers*



both in his use of the dialect and in his use of case markers and is positioned close to the regression line on all three plots. He thus speaks very differently from Hilāl, but both can be described as having a normal mix of the dialect and case endings.

Speakers who are plotted some distance above or below the regression line have speech styles that deviate from the main trend. Four such speech styles will be discussed here in more detail: the styles of Tayzīnī, al-Miṣrī', Abū Majd, and ᶜArīqāt, each representing one direction of deviation from the main trend. It should be noted that the discussion below focuses on statistical deviations. They are as such not representative of the corpus as a whole. The statistically deviant styles are however particularly interesting in that they highlight the span of speech styles represented in the corpus.[87]

Tayzīnī, the first speaker to be discussed, does not make a single use of the measured dialectal features but is far beyond the predicted rate of case marking for dialect-free speech. His use of case endings is nothing short of exceptional, with more than double the proportional case marking of the second ranking speaker. This, in combination with the complete absence of the investigated dialectal features, makes for style of Spoken Standard Arabic that is strikingly bookish and high-flown. It is, however, not only with regards to case marking and avoidance of the dialect that Tayzīnī's speech is high-flown, he is also the only speaker to consistently use the prescriptively correct complementizer *inna* for the sentential complement of *qāla* 'say'. The Standard Arabic complementizer is normally *anna*, but after the verb *qāla*, in any of its forms, *inna* is the traditionally correct form (Ryding 2005:425; Badawi et al. 2004:713). The other speakers either use the saliently standard but prescriptively incorrect *anna*, or the dialectal *innu* for this function, with only a few instances of the prescribed variant. Counts of these three forms as used by the speakers in the corpus are listed in Table 17. The small number of observations for some speakers is due to their preference for not using any complementizer in this position (i.e. 'I said X' rather than 'I said *that* X'). The table shows that the two speakers below Tayzīnī in overall case marking, Badīᶜ and al-Qaddūmī, who use a very high form of Standard Arabic, both consistently use *anna*. Tayzīnī's use of *qāla inna* isThis is probably not perceived as incorrect by most listeners. in other words not representative of Spoken Standard Arabic as represented by the corpus, but rather of an archaic or prescriptivist speech style.

Tayzīnī at times also uses phrases that are of classical, or even Quranic style.

[87] The reader is encouraged to follow the hyperlinks in the URLs listed in Appendix A and listen to the interviews in order to get an impression of the individual speech styles discussed below.

Table 17: *Counts of complementizers of* qāla *'say'. Complements of the participle or verbal noun are not included.*

	innu	*inna*	*anna*
Tayzīnī	–	5	–
Badīᶜ	–	–	10
al-Qaddūmī	–	–	13
Abū Majd	–	1	–
Kayālī	–	2	7
Fayyāḍ	2	–	1
al-Muᶜallim	–	–	3
Mursī	2	–	1
al-Barādiᶜī	–	1	2
ᶜAbd al-Qādir	–	–	6
al-Xuḍarī	4	2	3
Hilāl	–	–	–
al-Asᶜad	–	–	1
Šallaḥ	1	–	5
Ġalyūn	1	–	1
ᶜArīqāt	12	–	–
al-Miṣrī	12	–	–
Total	34	11	53

One example is (100) with the phrase *in huwa illā* 'it is nothing but'. This phrase occurs 25 times in the Quran with different pronouns but is rare in Modern Standard Arabic.

(100) *wa-huwa ʾinna mā taqūlūna-hu min mā yaḥduṯu fī sūrīya*
an-it COMP what 2mpl.say-it of what 3MS.happens in Syria
ʾin huwa ʾillā *natījat-u muʾāmarat-in xārijiyya*
NEG it but result-NOM conspiracy-GEN foreign
'…and that is: that which you say about what is happening in Syria is that it is nothing but the result of a foreign conspiracy.' (Tayzīnī, 20:14)

Some examples of the prescriptive character of Tayzīnī's speech style that sets him apart from the other speakers in the corpus are mentioned in the following

chapters, including his occasional case marking on indefinite diptotes and on *nisba* adjectives (see 9.1.1 and 9.5 respectively).

At the other end of the case marking scale is al-Miṣrī with his one single case marker.[88] His rate of case marking is virtually zero and is matched only by ᶜAriqāt who will be discussed below. He is also the speaker in the corpus with the highest use of the investigated dialectal features, although Hilāl and Abū Majd are not far off. His combination of the use of no case endings and heavy use of the dialect makes his speech style unique in the corpus, as can be seen in his extreme position in the bottom right in the plots in Figure 11. Some particularly striking features in his speech that are not quantitatively investigated are his more or less consistent use of the dialectal variant of high frequency and saliently dialectal verbs (*rāḥ* 'went', *bidd-* 'want', etc.), of the indicative *b*-prefix, and the use of *iḥna* 'we' rather than the Standard Arabic *naḥnu*. His speech is by no means purely dialectal, however. He uses the Standard Arabic form of the relative pronoun instead of the dialectal form 25% of the time (see Table 15), and often pronounces /q/ in accordance with Standard Arabic phonology. All in all he would probably not be regarded as speaking Standard Arabic by most native speakers. He is nevertheless kept in the corpus since he fits the external criteria and thus represents the variation in speech styles used in formal, unplanned situations.

Tayzīnī and al-Miṣrī's are then extreme examples of the main trend of the dialectal features as being negatively correlated with case marking. Two speakers, Abū Majd and ᶜAriqāt have speech styles that are extremes as oppositions to the main trend. Abū Majd is positioned in the far right in all three plots and some distance above the regression line. This means that he combines a frequent use of case endings with a heavy use of the dialect, opposite to what is predicted by the model. Abū Majd's dense mix dialectal and standard features results in juxtapositions of the two within the same utterance in ways that are rarely seen with other speakers. This is exemplified in (101)–(103), in which dialectal words are underlined and Standard Arabic case and mood endings are marked with bold text. Example (103) is also the only instance in the corpus of a dialectal constituent being marked for case.

(101) *da* *maklama* *fi* *ʾaḥyān-**in*** *kaṭīra*
that talking.shop[89] in times-GEN many
'That [the political discussion] is often a [mere] talking shop.'
(Abū Majd, 10:52)

[88] al-Miṣrī's also produces a second case marker, the nominative ending *-ūn* in sound m.pl., which is not included in the calculation of case scores.

(102) *ʾana* *miš* *ʿāyiz* *qināʿ* *yukallim-**u*** *qināʿ-**an***
I NEG want mask 3MS.address-IND mask-ACC
'I do not want a mask talking to a mask.' (Abū Majd, 24:03)

(103) *ʾana* *šāyif* *fī* *bawādir* *bāliġat-**u*** *l-xuṭūra*
I seeing there.is indications considerable-NOM DEF-danger
'As I see it, there are indications of considerable danger.'
(Abū Majd, 7:38)

Tayzīnī and al-Miṣrī can thus be described as using speech styles that are at the extreme ends of the part of the diglossic continuum that is encompassed in this corpus. Abū Majd's style is more difficult to locate in the continuum. His dense mix of dialect and Standard Arabic defies most conceptualizations of spoken Arabic in that it cannot be straightforwardly positioned on a diglossic scale. He draws more or less simultaneously from all parts of the scale. All speakers mix to some extent features from their dialect and Standard Arabic when speaking in formal situations and Abū Majd's style can be described as an extreme variant of this mixing.

The three speech styles described so far are plotted in the upper left, upper right, and lower right in the plots in Figure 11. In the corpus, there is no speaker clearly representing the lower left, the fourth form of deviation from the predicted speech styles. This would represent what might be termed 'caseless Standard Arabic' — minimal case marking as well as minimal dialectal features. In a corpus with more speakers, such a speech style might have appeared, but the significant negative correlations between the use of the dialect and case scores indicates that it would be very rare. The speaker in the corpus who comes closest to a style of caseless Standard Arabic is ʿArīqāt. He combines an average score of dialectal features with very little case marking, in total four instances of unambiguously marked case elements. He speaks a fairly 'pure' Standard Arabic, but one that is not elevated by the use of case endings. His speech style is thus plotted in the bottom middle of the plots in Figure 11 and is located some distance from any other speaker. He comes close to the hypothetical caseless Standard Arabic in terms of case endings, but is further from it in terms of avoidance of dialect.[90]

To sum up, although there is a general pattern of case endings, a saliently

[89] *A Dictionary of Egyptian Arabic* (Badawi and Hinds 1986) labels *maklama* as derisive and translates it as "any place used for idle chatter".

[90] ʿArīqāt's lack of case endings may be due to his very fast paced speech of 138 words per minute to be comparade with the average in the corpus of 115 words per minute. This habit of

Standard Arabic feature, being used less often by speakers who use more dialectal features, there is a large room of variation in what degree case endings are combined with dialectal features. The three dialectal features investigated are more highly correlated with one another than any of them is with the use of case endings, suggesting that case endings relate to the diglossic continuum in partly different ways than do dialectal features.

8.4 Fixed phrases

The correct usage of case endings is by native speakers of Arabic considered to be complicated and difficult. For many speakers the use of case endings is even associated with fear and anxiety (see 4.3). One way to avoid having to engage in the use of case endings as a productive grammatical system but still being able to including case endings in extempore speech is to rely on fixed phrases. Fixed phrases with case endings, especially if they have very general semantic content, can be re-used in different contexts spread out in the discourse in order to elevate the level of the speech style. Two examples of a phrases with the potential to be used this way are (104) and (105). With their general meanings, 'with regards to' and 'to a large extent', they could be reused in more or less any context. In (104) the case governing constituent, the preposition *bi-*, and the case ending that is governed by it, the genitive marker *-i* on the word *nisba* 'relation', are both contained within the phrase. The form of the case marked word *nisbati* is independent of the variable element following it. It can therefore be inserted as a fixed phrase together with a complement in any syntactic environment. Similarly in (105), the geneitive marker *-in* on the word *ḥadd* 'border' is governed by the preposition *ilā*, and the whole phrase can be used in different contexts with the same case marked form.

(104) *bi-n-nisbat-**i*** *li-X*
in-DEF-relation-GEN to-X
'concerning X', 'as far as X is concerned' (Šallaḥ, 1:24)

(105) *ʾila ḥadd-**in** kabīr*
to border-GEN large
'to a large extent' (al-Qaddūmī, 21:13*)

If the majority use of case endings in a particular speech style occurs in recurring fixed phrases such as those above, it would imply that the case system

speaking at a high rate may have hindered him from developing proficiency in the case system in that there is little time for effective self monitoring.

is poorly internalized and that it is largely a collection of frozen forms of an inactive system. If, on the other hand, only a small part of all case endings appear in reoccurring fixed phrases, it would imply that the case system, to the extent to which it is used, is internalized and used productively. Because of the length of the interviews in the corpus, most of which are around twenty-five minutes long, any such fixed phrases are expected to appear several times if speakers are indeed relying on them.

To identify possible fixed phrases, combinations of word stem, suffix, and case governance produced three or more times by the same speaker were considered fixed phrases and were retrieved from the database. If, for example, the exact sequence *bayt-i* (house-GEN) occurs three times in the same interview and all are prepositional complements then they would be retrieved. If two are genitive complements and one is a genitive annexation, they would then not be retrieved. The reasoning behind this is that even if the word has the same form, using it in differently governed genitive positions requires operating with case as a grammatical system. If, on the other hand, the stem-suffix combination is only used in the same syntactic position, this can normally be achieved applying rules operating on a linear set of words or constituents. Indefinite adverbs were not included in the analysis. Also note that formulaic expressions, such as *al-ḥamdu li-llāh* 'God be praised' were excluded from annotation altogether (see 6.5).

The results of this operation are listed as glossed examples in Table 18 on the following two pages. The number of occurrences of each phrase and the total number of case markers for each speaker are also listed and where all instances of a certain item are preceded by the same or similar words these are glossed in the leftmost column. The information presented in the table is discussed below first in over-arching terms and then with regards to specific speakers.

8.4.1 Fixed phrases and overall case marking

There is surprisingly little use of fixed phrases in the corpus. Of the seventeen speakers, eight speakers produce no repeated phrases at all and are not represented in Table 18. Seven of the nine speakers that are represented produce only one or two repeated phrases. If the ubiquitous *lā budda* is excluded, only two of the eight speakers with the highest production of case marking have any fixed phrases at all.

Lā budda 'must' dominates the list of fixed phrases. It is used three or more times by five speakers. al-Qaddūmī is particularly fond of the expression and uses it eleven times. It is not a lexicalized form, however. Around half the times it is uttered in the corpus it does not have a case ending (*lā budd*) and five of the

Table 18: *Fixed phrases with case marking*

PRECEDING WORD(S)	STEM AND SUFFIX	CASE GOV.	#
	TAYZĪNĪ		
	qādir-an capable-ACC	SIS. *KĀNA*	4
lā there.is.no	*budd-a* escape-ACC	ABS. NEG.	3
	ḥālat-an situation-ACC	OBJECT	3
	ḥālat-in situation-GEN	PREP.	4
fī in	*ḥīn-i-h* time-GEN-its/his	PREP.	6
		Total marked=261	
	BADĪ^{C}		
fahimat/fahimū understood.3FS/MPL	*fahm-an* *xāṭiʾ-an* understanding-ACC faulty-ACC	ABS. OBJ.	6(3)
ʾuḥibbu [NAME] love.1S [NAME]	*ḥubb-an* *gamm-an* live-ACC complete-ACC	ABS. OBJ.	6(3)
namuddu/madadnā extend/extended.1PL	*ʾaydiy-a-na* hands-ACC-our	OBJECT	3
		Total marked=181	
	AL-QADDŪMĪ		
lā there.is.no	*budd-a* escape-ACC	ABS. NEG.	11
		Total marked=86	
	MURSĪ		
lā there.is.no	*budd-a* escape-ACC	ABS. NEG.	3
		Total marked=57	

Continued on next page.

Continued from previous page.

PRECEDING WORD(S)	STEM AND SUFFIX	CASE GOV.	#
	FAYYĀḌ		
lā there.is.no	*budd-a* escape-ACC	ABS. NEG.	4
		Total marked=50	
	AL-BARĀDIᶜĪ		
[NUM.] [NUM.]	*ᶜām-an* year-ACC	NUM. SPEC.	8
lā there.is.no	*budd-a* escape-ACC	ABS. NEG.	3
		Total marked=27	
	ᶜABD AL-QĀDIR		
	mazīd-an increase-ACC	OBJECT	5*
		Total marked=32	
	HILĀL		
naḥnu we	*mutamassik-ūna* clinging-mpl.NOM	COMMENT	3*
	intihāk-an violation-ACC	OBJECT	3*
		Total marked=18	
	AL-XUḌARĪ		
ʾabnāʾ sons	*šaᶜb-i-na* people-GEN-our	ABS. OBJ.	3
		Total marked=21	

*Rhetoric repetition.

thirteen speakers who use the phrase use it both with and without case ending.

There are no words with a definite article amongst the fixed phrases in the corpus. The most frequent type of reoccurring phrases generally are those with nouns that have an accompanying definite article and that are unmarked for case. al-Qaddūmī, for instance, uses the phrase *bi-l-fiʿl* 'indeed', 'in fact' (lit. 'in the deed') eighteen times, and ʿAbd al-Qādir uses the phrase *bi-t-taʾkīd* 'certainly' (lit. 'with the assertion') twenty-four times. Both these phrases could safely have been marked for case by adding the genitive ending *-i*, producing fixed phrases with case endings. The fact that this is not done is in accordance with the very consistent pattern not to mark case on nouns with a definite article (see 9.2). This pattern means that these and many similar phrases are not produced with overt case marking. On the other hand, there are frequent phrases with indefinite nouns and where this pattern does not apply but where this is nevertheless not done. The common adverbial construction *bi-šakl [ADJ]* 'in an [ADJ] way'; '[ADJ]ly', for example, is used 32 times in the corpus. The noun *šakl* could safely be marked for genitive with *-in* whenever this phrase is used since it is governed by the preposition *bi-* that is contained in the phrase. It is however only marked for case once in the entire corpus. What is striking in the material is not the exploitation of reoccurring phrasings to include case marking, but rather the many instances of case markers being added and then dropped in parallel phrases appearing in the same or in adjacent utterances (see for example (1) and (2) on page 40).

There are fewer fixed phrases with case marking than one would expect under the assumption that the case system is poorly mastered and that there is a social pressure to produce case endings in these interviews. Furthermore, many chances are not taken to utilize recurring self-contained phrases for case marking, indicating that speakers actually avoid repeating phrases with the same case marker. In languages where case marking is non-optional and is part of the native grammar, such a pattern would not be viable, since a large number of phrases are bound to be repeated in extended speech. In Spoken Standard Arabic, where case marking is used only sporadically, even very few repetitions of phrases with the same case endings would stand out as a markedly one-sided use of the case system, which may partly explain this pattern.

For the speakers in the corpus, fixed phrases play only a marginal role in overall case marking. Rather, most of the fixed phrases have specific rhetoric function, as will be shown below.

8.4.2 *The functions of fixed phrases with case marking*

The individual speakers' use of fixed phrases is analyzed in this section, except for those whose only fixed phrase is *lā budda*. It will be shown that fixed phrases with case marking serve different pragmatic functions for different speakers, and that only one speaker relies on fixed phrases to include a higher rate of case endings in his speech style. The more detailed discussion below should be viewed in light of the observation made above that fixed phrases are generally not used to raise the level of case marking. The space devoted to these items here should not be taken to mean that they play a quantitatively important role. Specific uses of fixed phrases are described here in some detail precisely to show that they have other functions than to regulate the speech style.

Tayzīnī, first of all, has the largest number of reoccurring phrases with identical case markers. This is to be expected giving his extensive use of case endings that is bound to naturally generate some repetition. The most frequent of his reoccurring stem-suffix combinations is *ḥīnih* 'its time' which is in all instances preceded by the preposition *fī* 'in'. In the other phrases the case governing word varies. That this phrase always takes this form is explained by Tayzīnī's remarkably consistent case marking of words with an enclitic pronoun. As will be shown in 9.2, he never uses unmarked or ambiguous forms in connection with an enclitic pronouns, and so the alternative ambiguous form of the phrase, *fī ḥīnuh*, does not to appear. This is therefore best interpreted not as a fixed phrase but as a result of other patterns in his production of case marking. In either instance, even if all Tayzīnī's repeated phrases are taken to be fixed phrases, the endings used in them only make up seventeen tokens, or 7%, of his total of 261 case endings.

Badīᶜ's reoccurring phrases are of a different kind in that each is used in a very specific context to ward off criticism. Firstly, he has two phrases consisting of an absolute object with an adjective. Each is repeated three times, giving six case markers each. The first of these two phrases, *fahimat/fahimū fahman xāṭiʾan* 'it/they misunderstood' is used in all three instances in the context of members of the Muslim Brotherhood misunderstanding the writings of the radical Islamist thinker Sayyid Quṭb as sanctioning violence. The other phrase with absolute object, *ʾuḥibbu [NAME] ḥubban gamman* 'I love [NAME] completely', is said about two prominent members of the Muslim Brotherhood who publicly criticized the process by which Badīᶜ was elected as Supreme Guide of the Brotherhood, revealing a schism within the organization.[91] The third use of this phrase is in connection with the previously mentioned Sayyid Quṭb. Badīᶜ's

[91] "Fault Line in Egypt's Muslim Brotherhood", *The Nation*. August 21, 2011.

third reoccurring phrase, *aydiyanā* 'our hands' is in all three instances directly preceded by *madadnā* 'we extended' either in past or present tense. In all three instances, it is said in the context of the Muslim Brotherhood metaphorically extending its hand to the Egyptian regime in an invitation to cooperate. These instances of *aydiyanā* are especially marked as Standard Arabic in that they make up three of the merely four occurrences in the corpus of words in the defective paradigm being marked for case (not counting eight instances of the lexicalized adverb *ṯāniyan* 'secondly'). The fourth marked defective nouns is also by Badīᶜ and in a very similar phrase (106). Here *aydin*, the case marked stem is marked for nominative.[92] While not the same phrase as above, the verb is of the same root as in the fixed phrase (*madda* and *imtadda*) and it uses the same metaphor.

(106) *šarīṭat* *ʾan* *tamtadda* *ʾilay-na* ***ʾayd-in*** *ṣāḍiqa* *muxliṣa*
condition COMP be.extended to-us hands-NOM honest loyal
'on the condition that honest, loyal hands are extended to us'
(Badīᶜ, 36:58)

Badīᶜ seems to have a repertoire of phrases with case endings to answer specific points of critique often leveled against the Brotherhood. Badīᶜ's fixed phrase only make up fifteen, or 8%, of his total number of 181 case endings.

The only speaker who clearly does use fixed phrases to include more case endings and to elevate his speech style is al-Barādiᶜī. He marks the word *ᶜām* 'year' for accusative as number specification eight times, which accounts for almost a third of his total number of produced case endings. These phrases are all used in the context of criticizing the ruling party by mentioning how many years it has stayed in power, how many years it has maintained martial law, etc. There are a total of eleven instances of number specification in Barādiᶜī's interview. The three that are not marked for case are two instances of the word *sana* 'year', a synonym of *ᶜām*, and one instance of *maqᶜad* 'chair'. In other words, whenever *ᶜām* is specified for number, it is marked for case, and when it is swapped for a synonym the case ending disappears, even though the endings would take the exact same form, i.e. *-an*. For Barādiᶜī, then, the construction *[NUM.] ᶜāman* is clearly a fixed phrase that plays a major part in his overall case marking. Together with *lā budda* it makes up 11, or 41%, of his total of 27 case endings.

ᶜAbd al-Qādir and Hilāl both have reoccurring fixed phrases that again

[92] The *-in* is here considered to give a case marked form as it contrasts to the pausal form with no ending on the one hand, and the dialectal ending *-ī* on the other.

are different in kind from the those of the previous speakers. The occurrences of the phrases are here used in direct sequence in structurally identical segments repeated for rhetorical effect. ᶜAbd al-Qādir produces two such repetitions close to one another in which the noun *mazīd* 'increase' is marked for accusative case as direct object. (These markings give rise to cluster *b* in Figure 8 on page 172.) In the first of the two sets of repetitions, he says that the occupation of Palestine each day brings "an increase (*mazīd-an*) in Judaization of Jerusalem, an increase in settlements, an increase in confiscations". In the first set of repetitions, only the first instance of *mazīd* is marked for case and in the second set all three instances of *mazīd* are marked for case.

Hilāl produces two such structures in which the same stem and case marking is repeated (both contained in cluster *c* in Figure 8). In the first, he states that "we [the ruling party] adhere (*mutamassikuna*, lit. 'are clinging') to the civil state, we adhere to a state built on citizenship, we adhere to [the principle] that Muslims and Christians enjoy the same rights." His second series of repetition occurs in the following utterance where he states that if any party opposes these principles "it is considered a violation (*intihākan*) of the constitution and the law — not (*miš*) a violation of [the principles of] the National [Democratic] Party — a violation of the constitution and the law." Both ᶜAbd al-Qādir and Hilāl use case endings to enforce what is already highly emphasized rhetorical devices, giving them more gravity and force.

For ᶜAbd al-Qādir, these case marked tokens represent five, or 16%, of his total of 32 case endings. For Hilāl, the case marked tokens in the repeated phrases represent six, or 33%, of his total of 32 case endings.

Finally, al-Xuḍarī repeats the phrase *abnāʾ šaᶜbinā* 'the sons of our people' three times. The second word *šaᶜbinā* 'our people' is marked for genitive as annexed to the first, while the first word *abnāʾ* 'sons' is unmarked for case. This means that al-Xuḍarī, a Palestinian politician, can reuse this phrase anywhere as a synonym for *al-filasṭīniyyūn* 'the Palestinians', a word he in fact never uses. Nor does he use the phrase *aš-šaᶜb al-filasṭīnī* 'the Palestinian people' that is frequent in the interviews with his countrymen in the corpus. It seems to be a part of al-Xuḍarī's speech style in Spoken Standard Arabic to substitute any word he would in dialectal contexts use to refer to his countrymen with *abnāʾ šaᶜbinā*. This is then a fixed phrase that raises his overall case marking. It only represent three, or 14%, of his 21 case endings however.

It is worth noting that the rhetorical functions of the use of case markers described above are made possible by using case endings only sporadically. A segment or phrase can only be marked with case endings for rhetoric effect if there are sufficiently few case markers produced for it to stand out. In a speech

style such as that of Tayzīnī with high usage of case marking, there is little room for variation. A phrase or a segment with 'extra' case endings in his speech would be indistinguishable from the surrounding usages. The communicative resource of signaling importance with case markers is not available to a speaker with that style.

In summary, fixed phrases with marked case make up only a small part of the speakers' total number of case endings. Of the seventeen speakers in the corpus, only al-Barādiᶜī and possibly Hilāl were found to rely heavily on a fixed phrase to extend the use of case marking. In general, speakers do not develop a repertoire of self contained expressions with case endings that are used to reach a target level of case marking. The fixed phrases that are used generally serve specific rhetorical rather than style-regulating functions.

8.5 Summary

This chapter described case endings as part of the Standard Arabic speech styles represented in the corpus. It was shown that speakers use case markers to very varying degrees despite producing speech in nearly identical situations and despite all having a background that favors proficiency in the highest formal register. The majority have a clearly noticeable but sporadic use of case marking. A few speakers produce a high rate of case marking with one or a few case endings in each sentence whereas a few other speakers mark case only on a couple of tokens in the whole interview.

It was further shown that speakers maintain a consistent level of case marking throughout the interview with little changes of speech style during the course of the interview. The observation that speakers slide into a less formal style after the first few minutes was shown not to hold for this corpus. This was interpreted to mean that the speakers have found a level of case marking that they can comfortably maintain.

Case marking was shown to be negatively correlated with dialectal features, so that the more dialect a speaker uses, the less case markers are produced. The variation within this pattern is fairly large, however, so that a wide range of the frequency of case marking is observed for speakers with the same measured value of dialectal usage.

Finally, the reliance on fixed phrases to produce case endings was investigated. Only one of the seventeen speakers was found to clearly rely of fixed phrases to include case endings in his speech style. Generally, speakers who use case markers do so in varying structures and on different lexical items, and the repetitions of identical phrases with case endings that were found serve

rhetorical rather than style regulating functions.

On the whole, case in Spoken Standard Arabic, at least as used in this corpus, is a more varied and living system than might first be assumed. The very different rates of case marking and the fact that it does not fluctuate significantly suggest that speakers who are faced with the need to develop a style of speaking Standard Arabic do so in very different ways. There is in other words no one way of speaking Standard Arabic; speakers develop individual speech styles in which the production of case endings may have central or marginal roles.

Case marking and morphology

9

The previous chapter described the production of case endings as a part of an individual's overarching speech style. The data was grouped by speaker to give descriptions of their different ways of using case endings in Spoken Standard Arabic. In this and the following chapter, data will instead be analyzed as grouped by grammatical variables. This chapter investigates the effects of the morphological parameters on production rates of case marking, and the following chapter examines the effect of syntactic parameters.

There are several ways in which morphological features may affect which words are marked for case in speech. The paradigm variable may have effects on case marking in that some paradigms are more frequent and central to the case system and are therefore more readily available to speakers, whereas others are more obscure, less frequent, and therefore less accessible. Grammatical case is also mapped to the paradigms with different patterns of syncretism. This is of particular importance for case marking in Arabic since it is learned primarily in formal education through processes of second language learning by all speakers. For native speakers, peripheral or irregular inflectional classes in their language are learned across time, but for the second language learner infrequent or formally deviant classes are more difficult. Since all speakers of Standard Arabic learn the language and its system of case marking only through formal language instruction, and in very similar forms of pedagogy (see 4.3), they are expected to have similar experiences of which types of words are easy to mark for case and which are more difficult.

The way case endings are marked in Arabic orthography (see 4.6) is also expected to affect case marking in formal speech. There are two aspects of this effect. The first is that case endings that appear in writing constitute a consistent

form of linguistic input for speakers learning the case system. When reading Arabic, one is exposed to case endings only in certain forms and in certain positions but more consistently so than in the varying practices of speech. The second aspect of the effect of orthography on speech is its role in shaping the perceived standardness of a text. When a native reader is confronted with a case ending in printed text, there is a direct clash with native dialectal syntax in which there is no parallel system system of marked case. This clash does not occur in the majority of nominals that lack overtly marked case in writing. This means that these specific case endings that are orthographically represented come to mark a contrast to dialectal forms and thereby become particularly salient markers of Standard Arabic. Speakers thus have stronger incentives to include them in their speech to mark that it is Standard Arabic rather than endings that are not represented in writing.

This chapter is organized as follows: case marking in the different paradigms is first analyzed in section 9.1. It is shown that some paradigms are marked to a high degree, whereas others are by and large excluded from case marking. In section 9.2, the effect of definiteness on case marking is investigated. It is shown that words with the definite article are hardly ever marked for case, while words with an enclitic pronoun are marked at a very high rate. It is suggested that this pattern can be traced to case marking in formal read speech as it is perceived by native speakers of Arabic. The section also includes a closer analysis of the 3m.s. enclitic pronoun, some forms of which are used as encoding different grammatical features by different speakers. Then follow investigations of three specific types of tokens: in section 9.3 of accusatives that take orthographic *alif*, in section 9.4 of words with *tāʾ marbūṭa*, and in section 9.5 of *nisba* adjectives. Tokens with prescribed orthographic *alif* are shown to be marked at a high rate while tokens with *tāʾ marbūṭa* and *nisba* adjectives are disfavored for case marking.

9.1 Paradigm

Tokens in the corpus were annotated as being inflected for case according to one of eight different paradigms (see 7.4). As mentioned above, some paradigms are less frequent and are expected to be more difficult to mark for case. This also has the effect that the eight paradigms are very unevenly represented in the corpus. This is shown in Table 20 where the frequencies of the eight paradigms are listed as raw counts and as percentages of the total data. The triptote paradigm is by far the most common with nearly thirteen thousand tokens, making up 86% of the data. The large amount of data on triptotes makes it pos-

Table 20: *Frequencies of paradigms*

	#	%
Triptote	12885	86.0
Sound f.pl.	825	5.5
Diptote	567	3.8
Sound m.pl.	277	1.8
Final *alif*	245	1.6
Defective	125	0.8
Dual	57	0.4
The five nouns	10	0.1
Total	14991	100

sible to make very precise estimates on case marking in this category and also to analyze narrow subgroups of triptotes, as is done, for example, in the analysis of orthographic *alif* and *nisba* adjectives later in this chapter. The sound f.pl. is the second most common paradigm and comprises 5.5% of the data; the diptote is the third most frequent with 3.8%. The other five paradigms each make up less then 2% of the data. The smaller amount of data for these paradigms means that the results presented for them are of a more preliminary character. For more conclusive results on case marking in these paradigms data would have to be collected from a much larger corpus or through some sort of elicitation task.

Paradigms also differ in the choices that are available to the speaker. In the triptote, diptote, and sound f.pl. paradigms, a speaker has the option of not adding any ending. He can then opt not to mark case if he is uncertain of the correct ending. In other paradigms, some form of ending is obligatory. The choice is then not between producing an ending or not, but between producing an ending according to the dialectal or the Standard Arabic system. This situation arises for example in the sound m.pl. and dual paradigms. For this reason, results for the triptote, sound f.pl., and diptote paradigms are presented together below separately from the sound m.pl. and dual paradigms. The defective paradigm and the five nouns are both few in number in terms of tokens and take case endings in specific forms that make them difficult to compare with other paradigms. They are therefore analyzed separately. The 245 tokens in the final *alif* paradigm cannot take case endings and will not be further commented on.

Table 21: *Predicted case marking on triptote, diptote, and sound f.pl.*

	%	LOG-ODDS	SE	*n*
Triptote	4.1	−3.15	0.33	10010
Sound f.pl.	2.7	−3.59	0.38	644
Diptote	2.0	−3.88	0.42	452

9.1.1 Triptote, diptote, and sound f.pl.

Triptote

	N	G	A
Ind.	*-un*	*-in*	*-an*
Art.	*-u*	*-i*	*-a*
Con.	*-u*	*-i*	*-a*

Sound f.pl.

	N	G	A
Ind.	*-un*	*-in*	*-in*
Art.	*-u*	*-i*	*-i*
Con.	*-u*	*-i*	*-i*

Diptote

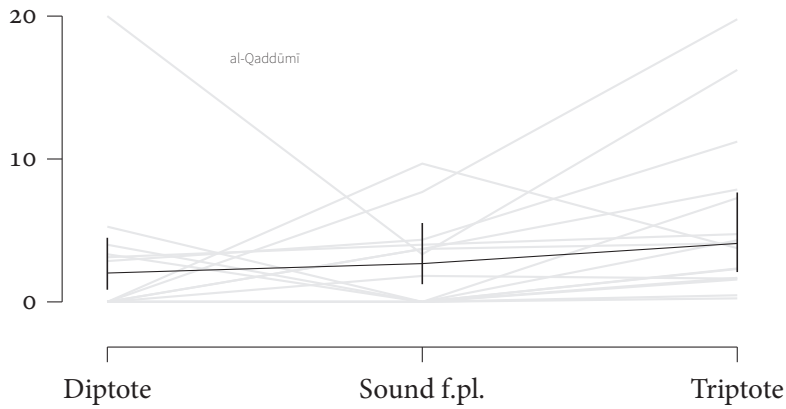
	N	G	A
Ind.	*-u*	*-a*	*-a*
Art.	*-u*	*-i*	*-a*
Con.	*-u*	*-i*	*-a*

The triptote paradigm is by far the most frequent and in pedagogical grammars it is presented as the canonical case marking paradigm. It is together with the five nouns the only paradigm where there is a morphological distinction between all three cases. Nominals in the sound f.pl. are easily identifiable due to the characteristic *-āt* ending. They differ from triptotes in their syncretism of the genitive and accusative forms which makes them structurally simpler with fewer forms. These patterns of syncretism are shared with the sound m.pl. and the dual. The diptote paradigm differs from the triptote only in indefinites with syncretism in the genitive and accusative forms and in not taking nunation. This gives a total of only three different forms in this paradigm. The traditional rules determining what nominals are diptote are however numerous and rather intricate (Wright 2011:vol.i, 239–46). For example, an adjective of the form *fa^{c}lān* is a diptote of it takes the feminine form *fa^{c}lā*, but triptote if they takes feminine form with *tā^{ə} marbūṭa*, i.e. *fa^{c}lāna*. This makes triptote nominals more difficult to identify than nominals in the sound f.pl. paradigm.

The hypothesis to be tested here is that tokens in the triptote paradigm are the least difficult to mark for case, being the most frequent and showing no syncretism. They are therefore expected to be marked for case more frequently than tokens of the sound f.pl. and the diptote paradigms. Diptotes are expected to be the least marked for case since they are the least frequent and are not as straightforwardly identifiable as words in the sound f.pl. paradigm. The likelihood of case being marked in these three paradigms was tested on the disambiguated dataset[93] with a Generalized Linear Mixed-Effects Model (GLMM) with the speaker variable as random effect. Table 21 shows the predictions of the model in percentages and log-odds, the standard errors of the log-odds, and the

[93] These three paradigms constitute the entire disambiguated dataset except for two tokens in the five nouns paradigm (see 8.1).

FIGURE 12: *Proportional case marking in triptote, diptote and sound f.pl. Gray lines represent individual speakers and the black the overall prediction.*



number of tokens. Triptote nominals are the most likely to be attested with case endings with a likelihood of 4.1%. The sound f.pl. is slightly more than half as likely to be case marked, at 2.7%, and diptotes are a little less than half as likely to be case marked, at 2.0%. The differences in case marking between triptote and the two other paradigms are significant (p=.029 and .008), whereas the smaller difference in marking between sound f.pl. and diptote is not (p=.387). The first hypotheses that tokens in the triptote paradigm are more often case marked was thus confirmed. The second hypothesis that the diptote paradigm is case marked less than the sound f.pl. was not confirmed.

There is considerable inter-speaker variation behind these numbers, as can be seen in Figure 12.[94] Four speakers mark some case forms on diptotes and none on sound f.pl., contrary to the overall pattern. al-Qaddūmī marks case at a rate of 20% case in diptotes (4 of 20 tokens) and 3.3% in the sound f.pl. (1 of 30 tokens).

The analysis above does not, however, take into consideration the fact that it is only for indefinite nominals that the diptote paradigm deviates from the triptote. Indefinite diptotes have been identified as a common source of pre-

[94] Error bars represent 95% confidence intervals. No confidence intervals are given for individual speakers so as not to clutter the figure, but these are naturally large. In this graph, the line representing Tayzīnī, which would be positioned high above the others, has been omitted to save space.

scriptive errors in oral performances of Standard Arabic. Kaye (1972:43) includes diptotes inflected as triptotes in a list of examples of prescriptive mistakes that are common when native speakers read texts aloud. According to Meiseles (1977:181), diptotes inflected according to the triptote paradigm "are so frequent that we have to regard them as (at least potentially) regular features of OLA [Oral Literary Arabic], and not as pseudo-corrections." These observations are not confirmed by the present corpus. The data on indefinite diptotes in the corpus is limited to 199 tokens. Of these, three are marked for case (two of which are produced by Tayzīnī), and none are hypercorrectly marked. In a GLMM with speaker as random effect, this gives a predicted rate of less than 0.01% (log-odds=−9.70, *SE*=3.21). The predicted rate of case marking on indefinite nominal tokens overall is fairly high at 8.86% (see 9.2). Thus, if speakers did treat diptotes as triptotes, several hypercorrect case endings on indefinite diptotes would be expected. On the contrary, there are no hypercorrect case endings. Speakers in this corpus seem to deal with the difficulties of marking case on indefinite triptotes, not by inflecting them as triptotes, but by excluding them from case marking altogether.

In summary, of the three most frequent paradigms, triptote, diptote and sound f.pl., the triptote paradigm is most often marked for case. At 4.1% it is roughly twice as likely to be marked for case as are the sound f.pl. and diptote paradigms. On the limited data available for indefinite diptotes, there was no indication of speakers marking it for case according to the triptote paradigm as suggested in previous research. Rather, most of the speakers opt not to mark it for case at all.

9.1.2 *Sound m.pl. and dual*

SOUND M.PL.

	N	G	A
Ind.	*-ūn*	*-īn*	*-īn*
Art.	*-ūn*	*-īn*	*-īn*
Con.	*-ū*	*-ī*	*-ī*

DUAL

	N	G	A
Ind.	*-ān*	*-ayn*	*-ayn*
Art.	*-ān*	*-ayn*	*-ayn*
Con.	*-ā*	*-ay*	*-ay*

The Sound m.pl. and the dual are similar to one another in having number and case encoded in the same ending. This means that the case ending cannot be omitted without changing the grammatical number of the word to singular. They therefore have no form that is overtly unmarked for case as do the three paradigms discussed above. Instead they have two basic forms, one that marks nominative case and one that marks genitive and accusative case and that is also a dialectal form where case is not encoded. The endings of both paradigms are marked in writing and are also stressed and therefore phonetically highly salient. For these reasons, they are hypothesized to be marked for case more than are tokens in other paradigms.

The number of occurrences for the two types of endings in the different case positions are listed in Table 22. In both paradigms, the accusative/genitive form is normally indistinguishable from the dialectal form (see 7.7.1). It is there-

Table 22: *Counts of case endings in dual and sound m.pl.*

ENDING		CASE POSITION		
		Nom.	Gen.	Acc.
Nominative	(*-ūn*, *-ān*)	47	0	3
Gen./Acc./Dial.	(*-īn*, *-ayn*/*-ēn*)	71	166	47

fore only in the nominative that there is a clear distinction between the case marked and the dialectal form. The likelihood of a sound m.pl. and dual being marked for nominative in nominative position is 42.4%, as predicted by a GLMM with a speaker as random effect (log-odds=−0.31, *SE*=0.45). This number cannot be directly compared to estimates given above of case marking in triptotes, diptotes, and sound f.pl. since it only includes tokens in nominative position and since ambiguous endings cannot be filtered out. The predicted rate of 42.2% is however remarkably high, and even though no direct numerical comparison can be made, it is clearly substantially higher then the rate for other paradigms which confirms the hypothesis. This is true also with the large standard error taken into account. (The lowest rate of case marking within the 95% confidence interval is 21.7%.) There are categories of grouped tokens that show rates of case marking on par with sound m.pl. and the dual, notably tokens that take orthographic *alif* (see 9.3), and nouns with an enclitic pronoun (see 9.2). But, when only paradigms are compared, sound m.pl. and the dual are by far the most favored for case marking.

There are other indications that case marking in the sound m.pl. and the dual are of a special status. Four of the speakers, Tayzīnī, Badīᶜ, al-Qaddūmī, and al-Muᶜallim, all speakers with high overall case marking, exclusively use nominative marked forms in nominative position. Together they produce 15 of the 47 correctly marked nominative tokens. Three of these speakers also produce one each of the hypercorrect nominative marked forms in accusative position as marked with gray in Table 22. In (107), the noun *ajandatān* 'two agendas' is in accusative position as the subject after *liʾanna*, but it is marked for nominative with *-ān*. In (108), the noun *multazimūn* 'committed' is in accusative position as the complement of *mā zilna* 'we are still', a sister of *kāna*, but it is marked for nominative with *-ūn*. In (109), *qaḍiyyatān* 'two agendas' is the direct object of the verb *aᶜṭaytu* 'I gave' but has the nominative ending *-ān*.

(107) *li-ʾanna hunāk* ***ajandat-ān***
for-COMP there agenda-DUA.HYP
'For there are two agendas.' (al-Qaddūmī, 19:36*)

(108) *li-ʾanna-na mā zilna* ***multazim-ūn*** *bi-natāʾij-a*
for-COMP-we not ceased.1PL committed-MPL.HYP by-results-AMB
l-qumma
DEF-summit
'For we are still committed to the results of the summit.'
(al-Muᶜallim, 9:45*)

(109) *ʾaᶜṭaytu kull-a magmūᶜat-in min-hum qaḍiyya ʾaw*
gave.1S every-ACC group-GEN of-them issue or
qaḍiyyat-ān *mina l-qaḍāya lati taš*ǧ*al miṣr*
issue-DUA.HYP of DEF-issues that 3FS.preoccupies Egypt
'I gave each group one or two issues that preoccupy Egypt.' (Badīᶜ, 36:33)

It is telling that three of the speakers with otherwise prescriptively correct use of the dual and sound m.pl. endings produce these hypercorrect forms. While for other speakers, the dialectal and ambiguous forms, *-īn* and *-ayn/-ēn*, are produced as unmarked for case, these four speakers appear to regard the ambiguous forms as marked for genitive/accusative and therefore as incorrect in nominative position. They do not allow themselves to have an unmarked option and must therefore always make an active choice of case marking. They can not avoid difficult situations which lead to errors. A similar pattern of non-optional case marking for some speakers is shown in 9.2.1 below with regards to the 3m.s. enclitic pronoun. In these words, for most speakers the form *-uh* is used as unmarked for case while a few speakers with high rates of case marking use it exclusively to mark nominative case.

It is interesting that the single metalinguistic comment found in the corpus (110) appears in the context of case marking in a sound m.pl. in nominative case. In this example, the speaker corrects himself, changing the sound m.pl. ending from *-īn* to *-ūn*, and then jokingly remarks that he is speaking dialect. He does this even though the corrected sentence is clearly marked as Standard Arabic with *ʾimma* 'either' and *ʾannahum* 'that they'. The speaker himself labels it as dialect only, apparently, on the grounds that he inflects the noun in sound m.pl. ending in accordance with the dialectal system. This inflection is however is routinely done on other types of words without it warranting self-correction or comments. It is only when a sound m.pl. is inflected dialectally that the speaker feels that an excuse is required for such a case form being produced. This is an indication of the special status of this paradigm as a perceived watershed between the dialect and Standard Arabic.

(110) *ʾimma ʾanna-hum kaḍḍāb-**īn** … yaᶜni aw ʾanna-hum*
either COMP-they liars-AMB … PAR or COMP-they
*kaḍḍāb-**ūn** … mnaḥki bi-l-ᶜāmmi ʾiḥna*
liars-NOM … 1PL.speak in-DEF-dialect we
Either they are liars(acc./gen.) … That is, or they are liars(nom.) …
(*Dialect*:) We are speaking dialect. (*Laughter*) (Šallaḥ, 34:53)

With limited data in the sound m.pl and dual, particularly with only tokens in nominative positions being counted, these findings should be taken as preliminary. They do however indicate that the sound m.pl. and the dual paradigms have special status as markers of Standard Arabic. This is shown in the high rate of case marking and in the fact that speakers with high case scores seem to treat case marking in these paradigms as compulsory, with no unmarked option. This status of the case endings sound m.pl. and dual as a marker of Standard Arabic may be traced to the phonetic saliency of these endings as stressed syllables and to them being marked for case in writing, a fact that forces the reader to develop a more direct awareness of case in these particular paradigms. This link between orthographic representation and case marking in speech is demonstrated more clearly and with improved accuracy in connection with orthographic *alif* (see 9.3).

As a marker of Standard Arabic the sound m.pl. and the dual case endings are quantitatively limited: the paradigms are relatively infrequent and case marking comes into play only in the use of nominative case. These paradigms are, however, highly salient and are proportionally very often marked for case.

9.1.3 Defective

	N	G	A
Ind.	*-in*	*-in*	*-iyan*
Art.	*-ī*	*-ī*	*-iya*
Con.	*-ī*	*-ī*	*-iya*

The defective paradigm is rather complicated in terms of prescriptive realizations of case. Indefinite nominals drop the final weak radical, *yāʾ* or *wāw*, in the nominative and genitive, and it is replaced with the nunation *-in*. In the accusative, the weak radical surfaces as *-ī*/*-iy* to which the accusative ending *-an* is added. In definites and construct state, the weak radical surfaces as *-ī* and can only be marked for case in the accusative with *-a*. The dialectal form of the defective paradigm has an invariable final long vowel *-ī*. This latter form is, with a few exceptions, used in all positions by speakers in the corpus, in effect excluding the defective paradigm from the system of case marking.

There are 128 tokens of the defective paradigm in the corpus and they are most frequently attested as the lexical items *tālin* 'following', *ṯānin* 'second', or *arāḍin* 'lands' that together make up 75 of these tokens. There are two instances of *ṯamāniya* 'eight', formally belonging to this paradigm, that are real-

ized dialectally as *tmān*, and eight instances of the lexicalized adverb *ṯāniyan* 'secondly' realized with a case ending as a lexicalized adverb. Badī^{c} is the only speaker to use case marked forms in the defective paradigm other than in the word *ṯāniyan*. This he does four times, three of which are in the fixed phrase *namuddu/madadnā aydīana* 'we extend/extended our hands', and the fourth is in the nominative as the subject in a similar phrase (see 8.4). The other 114 defective nominals in the corpus are produced with a final *-ī*, regardless of case and definiteness. The data for this paradigm is thus rather limited, but it indicates that defective nominals are excluded from case marking for the speakers in the corpus, taking an invariable final *-ī*.

9.1.4 *The five nouns*

	N	G	A
Ind.	*-un*	*-in*	*-an*
Art.	*-u*	*-i*	*-a*
Con.	*-ū*	*-ī*	*-ā*

Nominals in the five nouns paradigm are inflected as triptote nouns except in construct state where they take a case endings in the form of a long vowel, *-ū*, *-ī* and *-ā* in the three cases respectively. *Ax* 'brother' and *ab* 'father', the nouns in this class that are used in the dialects, take the *-ū* ending in construct state regardless of case. There are a mere ten tokens of the five nouns paradigm in the corpus, six *ax* and four *ḏū* 'owner of'. The four instances of *ḏū* all have the ending *-ū*. Two of these are in nominative syntactic position and two are in genitive position and thus have a prescriptively incorrect form. There are two instances of the word *ax* in construct state, *axūnā* 'our brother' and *axīka* 'your brother' both with prescriptively correct case markers. No conclusions can be drawn from these few observations.

9.2 Definiteness

In Standard Arabic, definiteness affects the form of the case ending in the triptote, diptote and sound f.pl. paradigms.[95] Together these represent 95% of the total data in the corpus. The following analysis will be restricted to the disambiguated dataset that is made up of tokens in these paradigms. Tokens in the corpus were coded as belonging to one of four morphologically defined categories of definiteness, described in more detail in 7.5. These categories are:

(*a*) definite article (e.g. *aš-šiᶜāru* 'the slogan')
(*b*) construct state with annexed noun or clause (CS-N/C)
(e.g. *šiᶜāru l-ḥizb* 'the slogan of the party'; *šiᶜāru llaḏīna* … 'the slogan of those who …')

[95] Case endings the five nouns are also affected by definiteness, but they will not be considered here due to lack of data (see 9.1.4).

TABLE 23: *Predicted case marking by types of definiteness*

	%	LOG-ODDS	SE	*n*
Definite article	0.3	−5.71	0.44	5137
CS-N/C	2.8	−3.56	0.44	1534
Indefinte	5.3	−2.88	0.42	3996
Enc. pronoun	43.9	−0.25	0.43	442

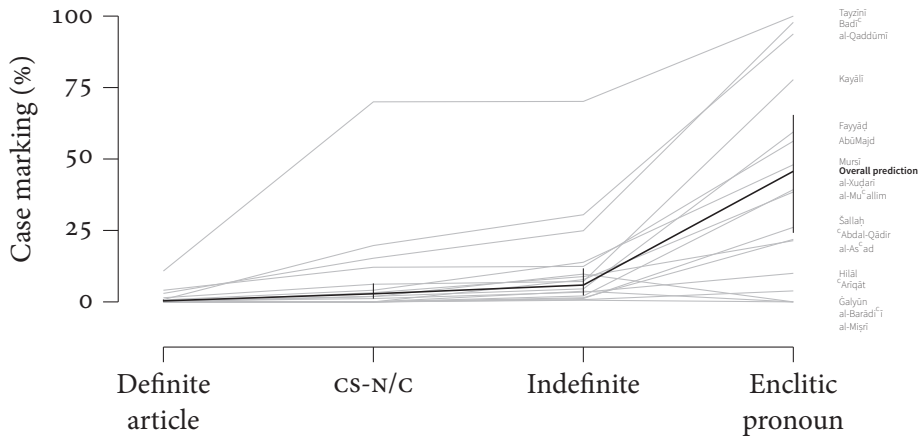
(*c*) indefinite (e.g. *ši^{c}ārun* 'a slogan')
(*d*) enclitic pronoun (e.g. *ši^{c}āruka* 'your (m.s.) slogan')

The analysis below shows that tokens with a definite article are very rarely marked for case whereas tokens in CS-N/C and indefinites are more often marked for case, and tokens with an enclitic pronoun are frequently case marked, almost half the time. It is also shown that this pattern is very constant across speakers.

The most important difference between the types of definiteness in terms of prescribed case marking is in their pausal forms. In words with definite article and in indefinites, the pausal forms are unmarked for case. Words in CS-N/C cannot normally be followed by a pause and thus have no prescriptive pausal form. In words with an enclitic pronoun, the case marker is 'protected' from final vowel omission in pause by the enclitic pronoun. In other words, prescriptive grammar generates forms that are unmarked for case, analogously with the dialectal forms, in words with the article and in indefinites, but not in construct state. It is hypothesized that tokens in constructs state (in CS-N/C and words with an enclitic pronoun) are more often marked for case in the corpus, reflecting the lack of unmarked pausal forms for these words in formal read speech.

The predicted likelihood of case marking in the four types of definiteness, calculated with a GLMM in the disambiguated dataset and with speaker as random effect, are listed in Table 23. The table shows that speakers in the corpus very rarely mark case in tokens with the definite article. Words with the definite article have a predicted likelihood of only 0.3% of being marked for case. Speakers mark case more often in CS-N/C at a rate of 2.8%, and still more in indefinites at a rate of 5.3%. Tokens with enclitic pronoun stand out from the rest with a predicted likelihood of 43.9% of being marked for case. The differences in case marking between each successive two types of definiteness are all

FIGURE 13: *Proportional case marking by types of definiteness. Gray lines represent individual speakers and the black line the overall prediction.*



significant (p<.001).

The four types of definiteness thus form a hierarchy of increased case marking in the order *a* to *d* as listed above. This hierarchy is remarkably consistent between speakers, as shown in Figure 13. In the previous chapter, it was shown that speakers mark case at very different overall rates that range from marking almost half of all nominals to only a few in the entire interview. Figure 13 shows that even though speakers vary widely in overall case marking rates, they have a very similar manner of distributing the case endings that they do use over different types of definiteness. All speakers mark case at a higher rate going up the hierarchy. To put it differently, the lines in Figure 13 are, with a few exceptions, either flat or have a positive slope. The definiteness hierarchy is so consistent that it may be considered to form a constraint on how case marking is structured in Spoken Standard Arabic. This constraint implies that if a speakers marks case at a certain rate in one type of definiteness, he will mark it at a higher rate in types of definiteness further up the hierarchy. Consequently, if for a given speaker there are no case markers on words with an enclitic pronoun, the topmost type in the hierarchy, there is no or minimal case marking elsewhere in that person's speech.

There are three counterexamples to this hierarchy. Badīᶜ and Ġalyūn both mark less case on tokens with an enclitic pronoun than in indefinites, but the clearest counterexample is al-Barādiᶜī's higher rate of case marking in CS-N/C relative to forms with an enclitic pronoun, 9.7% and 1.2% respectively. In 8.4 in the previous chapter, it was shown that al-Barādiᶜī relies heavily on a fixed phrase with a case marked indefinite noun (*[NUM] ᶜām-an* '[NUM] years') to reach his level of case marking. His frequent use of this phrase overrides the case marking hierarchy and accounts for his higher rate of marking in indefinite forms and than in words with enclitic pronoun.

In the hierarchy of case marking of types of definiteness, tokens with a definite article are at the lowest point of the hierarchy and are clearly disfavored for case marking. There is a total of 59 case marked tokens with the definite in the disambiguated dataset (roughly half of these, 27 tokens, are produced by one speaker; Tayzīnī). The relatively high overall rates of case marking for some speakers are thus reached by marking case on words other than those with a definite article, as can be clearly seen in Figure 13. Even Tayzīnī with his rather extreme rates of case marking follows this pattern. His rate of 10.8% case marking on tokens with the definite article is high, but much lower than his case marking on tokens in other types of definiteness. He has 100% case marking on words with an enclitic pronoun, but there are also other speakers with rates close to this. The categories of CS-N/C and indefinite forms are where Tayzīnī really stands out, with rates of case marking looming high above all other speakers. His rate of case marking on CS-N/C is 70.0%, as compared with the overall predicted rate of 2.8%. His rate on indefinite forms is 70.1%, the overall score being 5.3%. It is primarily in tokens in CS-N/C and in indefinite form that Tayzīnī's production of case marked items stands out. More importantly, he follows the general pattern of drastically lower case marking in words with a definite article.

If case marking is disfavored for tokens with the definite article, the opposite is true for tokens with an enclitic pronoun. These tokens are marked at a much higher rate than other types of definiteness. The increase of case marking in this category relative to other types of definiteness is very sharp for most speakers. Six speakers have case marking rates of over 50% in this category. The difference in case marking between tokens with an enclitic pronoun and indefinite tokens, the second most marked type of definiteness, is statistically significant for ten of the seventeen individual speakers ($p<.002$). Words with an enclitic pronoun are, in other words, at a level of their own in terms of case marking. High rates of case marking on words with an enclitic pronoun seems to be a linguistic norm in spoken Standard Arabic. This is also reflected in the

disproportional number of hypercorrect endings in this category: tokens with an enclitic pronoun make up only 6.0% of the total data, but half, 21 of 43, of the hypercorrect case endings (see 10.6). For most words, speakers seem to mark case only if they are certain of the right form, resulting in very few hypercorrect forms. For words with an enclitic pronoun this 'threshold of doubt' seems to be lower; it is in these words worth risking some incorrect endings in order to mark case more frequently. A particularly telling response to this norm was described in 8.2 in the previous chapter. There it was shown how al-As^{c}ad produces consistent case marking on words with an enclitic pronoun in the beginning of the interview. But soon he makes grammatical mistakes and abandons this style of speaking, producing only occasional case makers thereafter.

Returning now to the hypothesis formulated at the beginning of the section, it is clear that it is only partly confirmed. It was hypothesized that tokens in CS-N/C and those with an enclitic pronoun would be more often marked for case since neither of these has prescriptively sanctioned unmarked forms. Tokens with enclitic pronouns are indeed very often marked for case, but tokens in CS-N/C are not. They are marked slightly more than tokens with the definite article, but less so than indefinites.

One possible explanation for this hierarchy of case marking of in types of definiteness can be found in how case endings in formal read speech, one form of linguistic input, are filtered through listeners' dialectal syntax and reanalyzed as optional. This possibility is explored below.

9.2.1 Definiteness and perceptions of case marking in read speech

Prescriptively correct oral forms of Standard Arabic are used in what in a previous chapter was termed formal read speech, with news broadcasts as the primary example (see 4.7). It is a form of oral language production with high status and it is generally regarded as linguistically correct. It is therefore assumed to be one of the primary sources of input for speakers developing proficiency in Spoken Standard Arabic. The way case marking is conditioned by definiteness in the corpus may be explained by modeling how case marking in this form of speech is processed by speakers' native dialectal grammar. Assuming that Arabic speakers have a non-native, and therefore inefficient, capacity to process case markers online, they can be assumed to rely on their native, caseless syntax to process vowel endings in formal read speech. Since most pausal forms are not marked for case in Standard Arabic, many words are unmarked for case in formal read speech. Furthermore, formal overlaps between case endings and dialectal epenthesis mean that overt case endings can often be processed by a listener as an epenthetic vowel rather than as a case marker. Due to the syn-

tactic redundancy of case endings in Modern Standard Arabic (see 4.4), this reduction of case marking in the listeners perception comes with minimal loss of syntactic information and minimal risks of misinterpretation. A person who is developing proficiency in Spoken Standard Arabic with read speech as the primary form of linguistic input may thus (unconsciously) construct a grammar where case marking is optional in many situations.

The first thing to consider in modeling dialectal processing of case marking in read speech is pausal forms. In words with the definite article (111a) and in indefinites (111b), the case ending is omitted in the pausal form (disregarding here sound m.pl. and dual). Nominals with a definite article and indefinite nominals are thus often unmarked for case in formal read speech. Words in CS-N/C (111c) do not have a pausal form since they cannot normally be followed by pause. They are discussed separately below. For words with an enclitic pronoun (111d), the case ending is always retained in pause, being 'protected' from final vowel omission by the pronoun.

(111) a. *aš-ši^{c}ār-u* → *aš-ši^{c}ār*
DEF-slogan-NOM DEF-slogan
'the slogan'

b. *ši^{c}ār-un* → *ši^{c}ār*
slogan-NOM slogan
'a slogan'

c. ***ši^{c}ār-u*** *l-ḥizb*
idea-NOM DEF-ḥizb
'the slogan of the party'

d. *ši^{c}ār-u-ka* → *ši^{c}ār-u-k*
slogan-NOM-your.MS slogan-NOM-your
'your slogan'

Since the unmarked pausal forms appear in connection with a pause they are perceptually salient (Peter 1985:1039) and a listener is allowed time to process them in the following pause. Therefore, these forms may come to stand out as highly salient for the listener, even if they are quantitatively fewer than the case marked context forms. On the other hand, words with enclitic pronouns are always produced with case endings in formal read speech, even in pause. Based on this input, the listener may reinterpret case endings as optional for words with a definite article and in indefinites, but obligatory for words with an

enclitic pronoun. This would explain the strong tendency amongst speakers in the corpus to mark case on words with an enclitic pronoun and not on words with the definite article and on indefinites.

The situation for words in CS-N/C is different in that they are not normally followed by a pause in fluent speech. Therefore, they do not appear without case marking in formal read speech as do indefinites and words with a definite article. However, a large proportion of all words in CS-N/C are directly followed by the definite article. For these words, the ending vowels *-i* and *-a* can be processed by the listener as epenthesis, rather than as case marking. Consider (112) with the word *wazīr* 'minister' in CS-N/C in the tree cases followed by the definite article. In (112a), the nominative ending *-u* is clearly distinguishable as a case ending since it is not used as an epenthetic vowel in the dialects. In (112b) and (112c), however, the genitive and accusative endings *-i* and *-a* are identical to epenthetic vowels (see 7.7.1) and can be processed as such without connection to the case system. The word *wazīr* in (112b) and (112c) may then comes to be processed as unmarked for case. This means that a listener may perceive words in the genitive and accusative in CS-N/C to be to a large extent unmarked for case, even in prescriptively correct formal read speech.

(112) a. *waṣala* ***wazīr-u*** *l-māl*
arrived.3MS minister-NOM DEF-finance
'The minister of finance arrived.'

b. *taḥaddaṯū* *maᶜa* ***wazīr-i*** *l-māl*
spoke.3MPL with minister-GEN DEF-finance
'They spoke to the minister of finance.'

c. *daᶜū* ***wazīr-a*** *l-māl*
invited.3MPL minister-ACC DEF-finance
'They invited the minister of finance.'

This model of processing formal read speech through dialectal patterns of epenthesis thus offers an possible explanation as to why words in CS-N/C are not marked for case in the corpus to the same extent as words with an enclitic pronoun, despite both not having a prescriptively sanctioned unmarked form. The reinterpretation of case endings in CS-N/C as epenthesis may lead to them being reanalyzed as unmarked or optionally marked for case.

9.2.2 The 3m.s. enclitic pronoun

Tokens with the 3m.s. enclitic pronoun *-hu* 'its/his' are not included in the disambiguated dataset since one of its possible realizations, *-u*, is ambiguous with respect to case marking. It is however the most common form of the enclitic pronoun in the corpus, with 292 tokens. 266 of these are used in conjunction with triptote nouns and are analyzed in this section. It can be shown that the form *-u(h)* is used as unmarked for case in all case positions by most speakers in the corpus, while some speakers use it only in nominative position. Furthermore, there is in the corpus only one single use of the accusative pausal form *-a(h)*, indicating that this form is not part of most speakers' Spoken Standard Arabic.

The combinations of triptote case endings and the 3m.s. enclitic pronoun in pausal and context forms give a total of six different forms: *-uhu* and *-uh* in the nominative, *-ihi* and *-ih* in the genitive, and *-ahu* and *-ah* in the accusative. The dialectal form of the pronoun is *-u* in the represented dialects.[96] This form is in practice difficult to distinguish from the Standard Arabic nominative form *-uh* in normally paced fluent speech. They are regarded as identical in this study and will be here referred to as *-u(h)*.[97]

All instances in the corpus of the 3m.s. enclitic pronoun used in conjunction with triptote nouns are listed in Table 24 on the following page. Tokens are grouped vertically by speaker, ordered from top to bottom by decreasing overall scores of case marking, and horizontally by case position. The table shows that the form *-u(h)* becomes progressively more dominant in all case positions towards the bottom of the table for speakers with lower overall case marking. This indicates that this is a form that is unmarked for case for most speakers and is used analogously to the null-ending in other words. Accordingly, speakers with an overall case marking rate close to zero in the bottom of the table use this form fairly consistently. Four speakers, Tayzīnī, Badīᶜ, al-Qaddūmī, and Fayyāḍ produce only prescriptively correct forms of the pronoun (with the exception of one hypercorrection) which means that they produce *-u(h)* only in nominative position or not at all. For these speakers, then, there is no unmarked alternative for tokens with this pronoun. (For al-Qaddūmī this is based on one single observation.) This particular form of the pronoun appears to encode dif-

[96] The dialect of Kayālī differs from that of the other speakers in the corpus in that he uses *-a(h)* as the 3m.s. enclitic pronoun, and never *-u(h)*. The form *-a(h)* is evenly distributed between the three cases in his interview, just as *-u(h)* is for the other speakers. It was therefore coded as ambiguous for case marking in his interview and it is included in counts of *-u(h)* in the analysis conducted here, being its functional equivalent.

[97] Both are transcribed as –uh in the corpus text.

Table 24: *3m.s. enclitic pronouns on triptote nouns. Forms with ambiguous case markers are printed in red italics and hypercorrect forms are underlined.*

	NOMINATIVE	GENITIVE	ACCUSATIVE
Tayzīnī	uhu	ihi ihi ihi ih ih ih ih ih ih ihi ihi ihi ihi ih ih ihi ihi	ah ahu
Badīᶜ	ihi *uh uh uh* uhu uhu *uh uh uh* uhu	ih ihi ih ihi ih ih ihi ihi ih ih ih ihi ihi ihi ih ihi ihi ih ih ihi ihi ihi	ahu ahu ahu ahu ahu ahu
al-Qaddūmī		ih	
Abū Majd	*uh uh uh* uhu *uh uh uh uh uh uh*	*uh uh uh* ih *uh* ihi ih ihi ihi ih ih *uh uh uh* ih *uh*	ahu *uh* ahu *uh uh uh*
Mursī	ihi *uh uh uh* uhu *uh uh*	ih ihi ihi ihi ihi *uh* ih *uh uh* ihi	*uh* ahu ahu ahu *uh*
Kayālī	ahu *ah ah* uhu	*ah ah* ih ihi ih ih ih ihi ihi ih ih ih ih ih ih ihi ih ih ih *ah*	*ah* ahu ahu *ah* ahu *ah ah*[a]
Fayyāḍ	*uh uh uh uh uh uh uh*	ih ihi ih ih ih	
al-Muᶜallim	ihi	*uh* ihi *uh* ihi ihi ih *uh*	ahu *uh uh*
al-Barādiᶜī	*uh* uhu *uh*	*uh uh* ih	*uh uh uh*
ᶜAbd al-Qādir	*uh uh*	*uh*	*uh uh*
al-Xuḍarī	*uh uh uh uh uh uh uh uh uh uh uh uh*	ihi ihi	*uh*
Hilāl	*uh uh uh uh*	*uh uh uh uh uh uh*	*uh uh uh*
al-ʾAsᶜad	ahu ahu uhu	ihi ihi *uh uh uh uh uh uh uh*	uhu ihi *uh*
Šallaḥ	uhu *uh uh uh* uhu *uh uh uh uh uh uh uh*	ihi ihi *uh uh* ihi *uh* ihi *uh uh* ihi ihi ih ih *uh*	*uh uh* ahu *uh uh uh uh uh*
Ġalyūn	*uh*	*uh uh* ihi *uh uh uh uh uh uh uh* ih	*uh uh uh uh*
ᶜArīqāt	*uh uh uh uh uh uh uh*	*uh uh uh uh*	*uh*
al-Miṣrī	*uh uh uh*	*uh uh uh*	
Total[b]	*uh* = 7 uhu = 11	*uh* = 55 ihi = 48 ih = 48	*uh* = 37 ahu = 17 ah = 1

[a] For Kayālī *-ah* is the dialectal form and thus ambiguous.
[b] Hypercorrect forms are not included in the total.

ferent grammatical features for different speakers. For Tayzīnī, Badīᶜ, Fayyāḍ, and possibly al-Qaddūmī, it encodes nominative case, whereas for other speakers it does not encode case.

Although the data are limited, the table also indicates that the accusative pausal form of the 3m.s. enclitic pronoun, *-a(h)*, is rare. Of the 18 instances of accusative marked forms, there is only one single pausal form (113). There is also no instance of this form of the pronoun with tokens in paradigms other than the triptote and that would not be listed. This is a striking difference to the nominative and genitive marked forms where pausal and context forms are used interchangeably. For the nominative, the pausal form is naturally dominant since it is used also for unmarked case, and in the genitive, the context and pausal form are used 48 times each. The pausal form *-a(h)* thus seems to have a very marginal role in case marking in Spoken Standard Arabic.

(113) *bi-ḥayṯu ʾinnu ʾal-ʾāxar yaʾti li-yafᶜal-a* ***fiᶜl-a-h***
in-where COMP DEF-other comes to-make-IND deed-ACC-his
'… in order that the other comes to do his deed.' (Tayzīnī, 12:21)

The 3m.s. enclitic pronoun is thus similar to other forms of enclitic pronouns in having a form *-u(h)* that is used as unmarked for case, but it coincides with a Standard Arabic case marked form. It patterns as other unmarked forms in that it is produced more by speakers who generally produce less overall case marking. Speakers producing the most overall case marking use this form as marked for nominative, having no unmarked option for words with this pronoun. This is similar to the use of ambiguous forms of the sound m.pl. and the dual (see 9.1.2): speakers with high overall case marking use *-u(h)* as marked for case, and speakers with mid and low overall case marking rated use it as unmarked for case.

9.3 Orthographic *alif*

It was shown above that tokens in the sound m.pl. and dual paradigms are more likely to be marked for case than are nominals in other paradigms. It was argued that one of the reasons for this is that case markers in these paradigms appear in writing. The other frequent situation where a case marker appears in writing is in accusative indefinite triptotes without *tāʾ marbūṭa*. These are written with the letter *alif* (see 4.6 for details and examples) and are are also marked for case at a very high rate in the corpus. This pattern provides further evidence for the strong effect of orthographic case marking on speech.

Table 25: *Predicted case marking in indefinite accusative triptotes*

	%	LOG-ODDS	SE	*n*
W. ort. *alif*	41.2	−0.35	0.40	498
W/o ort. *alif*	3.6	−3.30	0.48	271

The accusative ending *-an* in indefinite triptotes is written with the phonetically void orthographic *alif* except on stems ending with *tāʾ marbūṭa*, *-āʾ*, or *-aʾ*. Tokens with the prescriptive case ending *-an* can thus be divided into two groups: those that would be written with orthographic *alif* and those that would not. Table 25 shows the predicted likelihood of case marking for tokens with prescriptive *-an* in these two groups as predicted by a GLMM on the disambiguated dataset with speaker as random effect (note that indefinite adverbs formed with an accusative ending are not included in this dataset). Tokens that would be written with orthographic *alif* are very likely to be marked for case, 41.2%. This is comparable to the rates for sound m.pl. and dual forms in the nominative as discussed above. Tokens with prescriptive *-an* that would not be written with *alif* have a predicted likelihood of only 3.6% of being marked for case. The difference of 37.6 percentage points is statistically significant (p <.001). It is worth emphasizing here that this large difference is in the rate of case marking of tokens in the same set of syntactic positions and with the same prescriptive ending. Therefore, the difference can not be explained by morphological and syntactic variables. Part of the difference can be accounted for by the negative effect of *tāʾ marbūṭa* on case marking (see below). It is however too small to account for a difference this large. Thus, this makes the effect of orthography on the rate of case marking very clear.

These data can be tentatively interpreted to mean that the accusative ending *-an* is primarily internalized by most speakers in connection with those words that take orthographic *alif*. These words come to be stored in the mental lexicon as possibly taking the *-an* ending since this possibility occurs in the written linguistic input. We can assume that not storing them lexically with this optional feature would make the appearance of orthographic *alif* in the text interrupt the reading process. Words with *tāʾ marbūṭa* do not change graphic shape with changes in case and can thus safely be stored as unchanging without affecting the reading process. What started out as an orthographic rule seems to have been reanalyzed as a grammatical rule in speakers' internal representation of the case system and thus dictates word forms produced in speech.

9.4 *Tāʾ marbūṭa*

The feminine marker *tāʾ marbūṭa* in Standard Arabic is realized as *-a* when word final and in absolute state and otherwise as *-at*. Adding a case ending to a word with *tāʾ marbūṭa* makes it no longer word final and thus requires the addition of *t* between the stem and the case ending (114). To mark case in words with *tāʾ marbūṭa* is therefore morphologically more complicated than for other words. The hypothesis to be tested here is that this has the effect that words with *tāʾ marbūṭa* are less likely to be marked for case.

(114) *al-jumla* → *al-jumla**t**-u*
DEF-sentence DEF-sentence-NOM
'the sentence'

Schulz (1981:138) found in his material that speakers mark case less on words with *tāʾ marbūṭa*. After excluding data from speakers with less then 5% overall case marking, he found that 31% of all nominals were marked for case, but only 14% of nominals with *tāʾ marbūṭa* were so marked. This calculation does not account for other variables with which *tāʾ marbūṭa* can be assumed to interact. Below it is shown that when other variables that affect case marking are controlled for, this negative effect of *tāʾ marbūṭa* is reduced.

There are three variables that may interact with the effect of *tāʾ marbūṭa* on case marking. The first is that *tāʾ marbūṭa* cancels the strong positive effect of orthographic *alif* as shown above. Part of the observed negative effect is thus likely to be a lack of the orthographic *alif* effect rather than being an effect of *tāʾ marbūṭa*. The second interaction is with the headedness variable (see 10.1). Not only attributes to words in feminine singular form have *tāʾ marbūṭa*, but this occurs also with attributes to non-human plurals. *Tāʾ marbūṭa* is therefore overrepresented in attributive forms which are less likely to be marked for case.[98] The third interaction is with the inflectional paradigm; only triptotes and diptotes can have *tāʾ marbūṭa*. A comparison between all tokens with *tāʾ marbūṭa* and all tokens without *tāʾ marbūṭa* is therefore also a comparison between paradigms, and part of the observed difference in case marking may therefore be a paradigm effect.

To test the independent effect of *tāʾ marbūṭa* on case marking, a GLMM in which these three interactions were controlled for was run on the disambiguated dataset. To control for the effects of orthographic *alif*, only tokens in nominative and genitive positions were included. The effect of the headedness

[98] In the corpus, 16.6% of all tokens, and 27.0% of tokens with *tāʾ marbūṭa*, are attributes.

Table 26: *Predicted case marking on triptotes with and without* tāʾ marbūṭa

	%	LOG-ODDS	SE	*n*
W/o *tāʾ marbūṭa*	1.0	−4.55	0.95	5511
With *tāʾ marbūṭa*	0.6	−5.08	0.95	2730

variable was compensated for by including it as a random effect together with the usual speaker variable. The paradigm effect was controlled for by only including data for triptotes. The likelihood of the incidence of case marking being predicted by the model are 1.0% for words without, and 0.6% for words with *tāʾ marbūṭa*, as shown in Table 26. The difference of 0.4% is statistically significant (p<.001).

The model predicts that, everything else being equal, if *tāʾ marbūṭa* is added to a word, its likelihood of being marked for case is reduced by roughly one third. Schulz's observation that words with *tāʾ marbūṭa* are less likely to be marked for case is thus confirmed, but the estimated effect is far lower than his observed halved rate of case marking for words with *tāʾ marbūṭa*.

9.5 The *nisba* ending

The *nisba* ending (*-ī* m./*-iyya* f.) forms adjectives from other parts of speech. The resulting adjective is inflected according to the triptote paradigm. The m.s. form of this ending is usually unstressed in oral Standard Arabic (Badawi et al. 2004:21; Ryding 2005:38–9), but in order for the word to be marked for case it must receive stress. Thus, the most common pronunciation of *miṣrī* 'Egyptian' is with the stress on the first syllable and this form cannot take a case ending without the occurrence of the additional operation of stress shift (e.g. **miṣriyy-un*). Only if the *nisba* ending is stressed can a case ending be added (e.g. *miṣríyy-un*). This can be accomplished a process whereby the speaker first shifts the stress to the last syllable and then adds the case ending. This is illustrated in (115). The intermediate form of this process, with a word final stressed *nisba* endings is fairly common in the corpus, as it is also in other forms of oral Standard Arabic (cf. Harrell 1964:34–5). If case endings are added to *nisba* adjectives through this process, then words with the m.s. *nisba* ending are expected to be marked for case to a lesser degree than are other triptotes because of the added morphological complexity.

Table 27: *Predicted case marking on triptotes with and without* nisba*-ending*

	%	LOG-ODDS	SE	*n*
W/o *nisba*	3.7	−3.27	1.25	5956
With *nisba*	1.9	−3.97	1.28	758

(115) *miṣ́ri* → *miṣrí́* → *miṣrí́yy-un*
Egyptian → Egyptian → Egyptian-NOM

The same end form may also be produced with the case marker being added to the stem, with the stress shift occurring as a phonological consequence, as illustrated in (116). The intermediate form with a case marker but no stress shift is not attested in the corpus. If this is the case then the m.s. *nisba* ending is expected not to have a negative effect on case marking, since the additional operation of stress shift occurs in an automatic phonological process.

(116) *miṣ́ri* → **miṣ́riyy-un* → *miṣrí́yy-un*
Egyptian → Egyptian-NOM → Egyptian-NOM

The hypothesis to be tested is that case is marked less in m.s. *nisba* than in other triptotes, which would indicate that it occurs through the first process and thus with stress shift logically preceding case marking. To test the hypothesis, tokens with the m.s. *nisba* ending *-ī*, 762 in total,[99] that were indexed in the database and their case marking compared to other triptote tokens in a GLMM with speaker as random effect. Only tokens in m.s. without *tāʾ marbūṭa* were included since only for these tokens is stress shift a prerequisite for case marking. Furthermore, very few *nisba* adjectives are in construct state (0.3% as compared 27.7% of the total data) or occur as heads (as defined in 7.3, 13.8% compared to the overall of 86.2%). These variables correlate independently with case marking and were therefore also included as random effects in the model. The results of the model are shown in Table 27. Tokens with the m.s *nisba*-ending have a predicted likelihood of 1.9% of being marked for case and occur as

[99] The total in Table 27 is smaller since it only includes tokens from the disambiguated dataset. *Nisba* adjectives were indexed by filtering out all triptote tokens with a stem final letter *yāʾ*. The stems represented in these tokens were listed to identify false hits such as *waᶜy* 'consciousnesses' and *ḥayy* 'alive', which were then excluded from the *nisba* index.

17 case marked tokens. Tokens without the ending are more than twice as likely to be marked for case at 3.7%. Tokens with the m.s. *nisba* ending are thus case marked at roughly half the rate as are other triptote tokens, a difference that is statistically significant ($p<.001$). This indicates that case marking on *nisba* adjectives is morphologically more complex than in other triptote nominals in that it involves an additional morphological operation.

9.6 Summary

In this chapter, data was presented and analyzed for case marking according to paradigms, types of definiteness, and with respect to a number of derivational forms with particular morphological properties. Of the inflectional paradigms in Standard Arabic, the triptote is by far the most frequent, followed by diptotes and sound f.pl. They are directly comparable in terms of rate of case marking in that they can all be unambiguously both marked and unmarked for case. It was shown that triptote nouns are marked at the highest rate, estimated to 4.1%, with diptote and sound f.pl. marked at approximately half that rate. This can be explained by triptotes being the canonical case marking paradigm and by the fact that this paradigm shows no syncretism. The sound m.pl. and the dual form another group of paradigms in that they do not allow for case to be unambiguously unmarked. They were shown to me marked at very high rate in the nominative, the only case were such a high rate of case marking can be observed. Even if this cannot be directly compared to the previous paradigms, the difference is such that it safe to say that tokens in the sound m.pl. and dual paradigms are marked for case to a higher degree than are tokens in other paradigms. Nominals in the defective paradigm have an invariable final *-ī*. The only exceptions to this, other than adverbs, are in a fixed phrase used by one of the speakers. There is no instance in the corpus of the ending *-in* in these tokens, as prescribed for indefinite masculine nominals in this paradigm. For the five nouns, there is too little data in the corpus for any conclusion to be drawn.

Case marking was shown to be hierarchically structured by the four types of definiteness defined in the coding scheme. Proportional case marking increases in types of definiteness in the following order: where a definite article is present, in CS-N/C, on the indefinite form, and on forms with an enclitic pronoun. This hierarchy is very consistent across speakers in the corpus. The increase is particularly sharp for tokens with an enclitic pronoun, which are marked for case almost half the time. It was suggested that these patterns of case marking may be accounted for by speakers reanalyzing case markers in these words as instances of epenthesis when such forms are heard in formal

read speech. Because of this, speakers may have reanalyzed these case markers as optional rather than obligatory. Furthermore, the data on the 3m.s. enclitic pronoun in its various forms indicate that speakers encode different grammar in the form *-u(h)*. Some use this form as marking nominative case and others use the same form as unmarked for case. The data as analyzed here also indicates that the pausal accusative form of this pronoun, *-a(h)*, is either marginal or non-existent in Spoken Standard Arabic.

An indefinite accusative form that, in writing, would be represented with orthographic *alif*, taken together with the high rate of case marking in nominative sound m.pl. and dual, give strong evidence for the influence of orthographically marked case on the production of inflected case forms in speech. Speakers seem to have internalized to a large extent the patterns of case marking through reading texts where such case marking occurs.

The presence of the feminine marker *tāʾ marbūṭa* in a word was shown be negatively correlated with case marking when other factors are controlled for. *Tāʾ marbūṭa* reduces the likelihood of the word it occurs with being marked for case by around one third and this could conceivably be due to the added morphological complexity of adding a *t* before the case ending in such words.

Finally, nominals with the *nisba* ending *-ī/-iyy* were shown to be marked for case at half the rate as other tokens under the same morphosyntactic conditions. It was suggested here that this is because adding a case ending on these words requires the additional operation of stress shift.

Case marking and syntax

10

The function of case markers in linguistic systems is to mark the grammatical or semantic role of constituents in a clause. The function of morphological case in Standard Arabic can be considered redundant due to the fixed word order that signals grammatical function in the clause (see 4.4). Rather, case markers in Spoken Standard Arabic primarily appear to serve a stylistic function. But for speakers to utilize this stylistic function, they need to maintain some awareness of the syntactic structure of the clause in order to choose the appropriate case ending for a particular word. This chapter explores how such syntactic features, and other features that operate above the word level, interact with the occurrence of case marking as it can be studied in the corpus data.

The first section of this chapter, section 10.1, investigates the difference in case marking between phrase heads and attributive adjectives that inherit case from the head noun. It is shown that case is very rarely marked on attributes, especially on adjectives. In section 10.2 it is demonstrated that speakers distribute case markers between the three cases in very similar ways. This segmentation of the data provides accurate predictions but does not account for the fact that each case is governed by a variety of often unrelated structures. This more fine-grained level of analysis is picked up in section 10.3, where case marking is analyzed for the particular grammatical structures by which they are governed. The data indicates that the speakers in the corpus have individually different preferences for what structures they mark case. This section also includes comments on pattern of case marking in some less frequent grammatical structures. Adverbs formed with the accusative differs from other forms of case marking in that they are independent of the syntactic surrounding and are often lexicalized. These are discussed in section 10.4. Section 10.5 shows that the pause system

is very much alive in Spoken Standard Arabic in that speakers do consistently avoid case marked context forms in utterance final position. Section 10.6 is an account of hypercorrect case markers in the corpus. These are fairly few but are distributed in ways that relate to observations to observations made elsewhere in this theses. Section 10.7 is a summary of the findings presented in the chapter.

10.1 Headedness

In Arabic, an attributive adjective follows the word it modifies and agrees with it in gender, number, and case. Case marking on attributives is thus often simply a matter of repeating the case ending of the directly preceding word, as in (117).

(117) *ʾaᶜṭi-ni* ***tanẓīm-an*** ***ṭawriyy-an***
give-me organization-ACC revolutionary-ACC
'Give me a revolutionary organization' (Kayālī, 17:38)

There are two things that may make case marking on attributive adjectives more complicated then copying the directly preceding ending. Firstly, the adjective may be of a different paradigm than the head noun and thereby take a different case ending. If for example the head noun is a non-human sound f.pl., then the adjective is inflected according to the triptote paradigm as feminine singular with *tāʾ marbūṭa*, as in (118). Secondly, the adjective may end up some distance from the governing element. This happens for example when the head noun is in construct state with an annexed genitive. The attribute is then added after the annexation, as in (119).

(118) *baᶜd* ***ziyarāt-na*** ***ʾal-kaṯīra***
after visitsFPL-our DEF-many.FS
'after our many visits' (al-Miṣrī, 14:18)

(119) *min ʾajl taswīq* ***mubādarat-i*** *s-salām-a* ***l-ᶜarabiyya***
for sake promotion initiative.F-AMB DEF-peace.M-AMB DEF-Arabic.F
'on order to promote the Arabic peace initiative' (al-Muᶜallim, 7:33*)

Since case marking on attributive adjectives sometimes require a new choice of case ending, and since this choice may occur at a some distance from the governing element in the clause, case marking in attributives is hypothesized to be marked for case at a lower rate than are tokens in head position. In Parkinson (1994b:53, 58), where results from a formal grammar test conducted

Table 28: *Predicted case marking by headedness*

	%	LOG-ODDS	SE	*n*
Head noun	2.7	−3.60	1.23	8838
Attr. adj.	0.6	−5.12	1.24	2212

by Egyptian native speakers are reported, it has been shown that case marking on attributive adjectives under these two circumstances is particularly difficult. Only around 20% of even highly educated speakers were able to add the correct case ending on words that are either of a different paradigm from the head noun or separated from it in the clause by other material.

The hypothesis that attributive adjectives are less likely to be marked for case then are phrase heads was tested in a Generalized Linear Mixed-Effects Model (GLMM) on the disambiguated dataset with speaker as random effect.[100] The variables of definiteness, pause, and *tāʾ marbūṭa* were added as random effects, in addition to the usual speaker variable, since the distribution of attributives is highly skewed in these variables and each of them independently affect case marking. The results are reported in Table 28. The results show that, all else being equal, head nouns are at 2.7% roughly four times as likely to be marked for case as are attributive adjectuves, marked at 0.6%. This difference is significant ($p<.001$) and confirms the hypothesis that attributives are less often marked for case than are head nouns. From the large difference we can also draw the conclusion that case marking on attributive adjectives is rare.

Furthermore, case marking on adjectival attributives seems to require that the head noun is also marked for case. There are a total of 45 case marked attributive adjectives in the corpus. They are all, with one exception,[101] preceded by a head noun that is also marked for case. Approaching the issue the other

[100] In the corpus, tokens that inherit their case from a head noun were coded as 'attributes'. Tokens that do not inherit their case from another noun were coded as 'heads' (see 7.3). Of tokens coded as attributes only those in absolute state were included for the purposes of this analysis. This was done in order to exclude non-adjectives tagged as attributes. These non-adjectives coded as attributes are primarily determiners and quantifiers with an enclitic pronoun, such as *kull(-uhu)* 'whole', 'all' and *nafs(-uhu)* 'itself'. This also excludes adjectives in the so-called 'false *iḍāfa*'. These are however very rare in the corpus. A total of 59 of the original 2271 tokens coded as attributes in the disambiguated dataset were excluded by this procedure.

[101] The exception is in the phrase *bi-l-mustawā l-lāʾiqi bi-ha* 'at the level appropriate for it' (Badīʿ, 25:50). Here the adjective *l-lāʾiqi* 'appropriate' is marked for genitive with the a final *-i*. The head noun *al-mustawā* 'the level' is in the final *alif* paradigm and cannot be marked for case.

direction, there are 95 head nouns in the corpus that are marked for case and are directly followed by an attributive adjective that take the same prescribed ending. These are in other words instances where the case ending could be copied from the head noun to the directly following adjective, as in example (117) above. Case marking on the adjective ought to be trivial in this instance. Yet only in roughly half of these 43 instances is the adjective marked for case thus giving a case marked head-adjective pair. In 29 of these case pairs, the case ending is *-an* with orthographic *alif*, as in (117) above. In fact, for every single head noun in the corpus with an prescribed *-an* ending that would be written with orthographic *alif* and that is directly followed by an adjective that is also takes a prescribed orthographic *alif* (84 noun — adjective pairs), both the noun and adjective are either marked or unmarked for case. Thus, while the pattern exemplified in (120), where we find case marking on the head noun but not on the adjective, is common in the corpus, there is no parallel example where both words take orthographic *alif*. Such phrases always display case marking either on both the noun and adjective, or with no case marking on either the noun or the adjective. There is no exception in the corpus of this pairwise case marking for words taking orthographic *alif*.

(120) *ʾanna-hum* ***juzʾ-un*** ***ʾaṣīl*** *mina* *n-nasig-a* *l-waṭani*
that-they part-NOM original of DEF-fabric-AMB DEF-national
'that they are an original part of the fabric of the nation'
(Abū Majd, 15:10)

To sum up, attributive adjectives are rarely marked for case and it can be calculated here that they are roughly one fifth as likely to be case marked compared with the frequency of case marking on head nouns. Case marking on attributive adjectives requires the preceding head noun to be also marked for case, further restricting case marking on these adjectives. The pattern of not marking case on attributive adjectives is offset by orthographic *alif*. In noun-adjective pairs where both words take orthographic *alif*, both words are either marked or unmarked for case.

10.2 Case

In this section, the data are analyzed as grouped by the three morphological cases: nominative, accusative, and genitive. The analysis on this level shows that tokens in the accusative are more often marked than are tokens in the other two cases and that there is no difference in frequency between nominative and genitive case marking. This pattern is consistent across speakers. When the

Table 29: *Case marking by case*

	%	LOG-ODDS	SE	*n*
Accusative	9.7	−2.23	0.33	1944
Nominative	3.0	−3.47	0.34	2981
Genitive	2.6	−3.62	0.33	6184

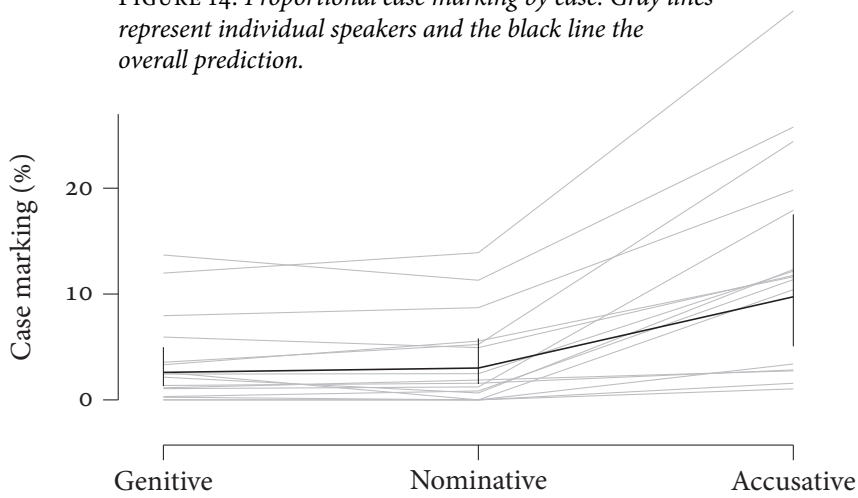
effect of orthographic *alif* is controlled for, there is no difference in marking between any of the three cases.

The analysis on this level does not account for the fact that the rules that govern each of the cases are diverse and that speakers are more likely to have case marking structured according to the rules by which case marking operates rather than according by the more abstract, overarching categories of the three cases. The more fine-grained analysis where data are segmented according to specific forms of case governance is presented in the next section.

The only previous statistical research that I am aware of investigating case marking in Arabic according to morphological case is Schulz (1981, see 4.8.2). Schulz found the highest rate of case marking to be in the accusative in his data, which he explained as an effect of orthographic *alif*. He also found higher rates of case marking in the genitive than in the nominative. This, he states, is due to the fact that a word in genitive position "always immediately follows the word that requires it to be in the genitive" (Schulz 1981:139). The analysis of the data from the present corpus corroborates his findings that nominals are more frequently marked for accusative than for the two other cases, but it was not found that the genitive marking occurs more frequently than nominative marking. Rather, genitives and nominatives are marked at equal rates in the present corpus.

Table 29 lists the rates of case marking for the three cases as predicted by GLMM on the disambiguated dataset with speakers as random effect. Tokens in the accusative are marked at 9.7%. This is more than three times the likelihood of case marking in the both the nominative, at 3.0%, and the genitive, at 2.6%. The differences between the rate of tokens marked with accusative and the rate of tokens marked with each of the other two cases are both significant ($p<.001$). This patterns holds for every individual speaker in the corpus; no speaker marks case at the highest rate in nominative or genitive, as illustrated

FIGURE 14: *Proportional case marking by case. Gray lines represent individual speakers and the black line the overall prediction.*



in Figure 14.[102]

The much higher rate of case marking in accusative is to be expected given the strong influence of orthographic *alif* on case marking (see 9.3). To test if the higher frequency of case marking in the accusative is also due to other factors the regression model reported above was run with prescriptive orthographic *alif* as an additional random effect. In this model, the higher rate of case marking in the accusative all but disappears and none of the differences between the three cases is significant (p=.522; .551; .109). In other words, the higher rate of case marking in the accusative is accounted for by the effect of orthographic *alif*.

Figure 14 further illustrates that the difference in the rate of case marking between tokens in the nominative and genitive is negligible. In the overall prediction, the small difference between these two morphological cases of 0.5% in the rate of case marking is not significant (p=.169). Tokens in nominative or genitive case are marked at the same rate. This contradicts the finding reported in Schulz (1981:139). It is worth noting that a listener who pays attention to case endings will indeed notice a large number of genitive case markers, due to the high frequency of words in the genitive. Indeed, of all overt case markers in the disambiguated dataset, 45% are in the genitive. This is then as shown above

[102] The line representing Tayzīnī is omitted in the figure to save space.

not because of genitives being proportionally more often marked for case, but simply because tokens in this case are very frequent.

To sum up, speakers distribute case endings in the three cases in very similar ways: in equal proportions for the nominative and the genitive, and with a higher rate of case marking in the accusative. This increase in case marking in the accusative can be accounted for by the effect of orthographic *alif*. The implications of this in terms of empirical norms in Spoken Standard Arabic are discussed in the following section.

10.3 Case governance

One of the main hypotheses of this study is that speakers internalize only some rules for case assignment and exploit them to mark case at some required rate. It has often been noted in the literature that case is more often marked on 'safe' positions in simple constructions where there is little risk of making mistakes. Badawī (1973:170), for example, suggests that nominative subjects of verbs (*fāʿil*) and genitives are often marked for case. As mentioned above, Schulz (1981:139) also makes this argument for the genitive. If this is so, then there should be clear differences in the rate of case marking on tokens in different syntactic positions with those tokens exhibiting higher rates of case marking in so-called "safe" positions.

Each token in the corpus was coded as filling one of fifteen different syntactic positions (see 7.6). The frequencies of these positions as counts and percentage of the total data are listed in Table 30. The positions were divided into three groups for the purposes of analysis of rates of case marking. The first group consists of eight *core syntactic positions* that occur extensively in the speech data of all speakers in the corpus. They are part of grammatical constructions that cannot be dispensed with in normal speech. The second group consists of *peripheral syntactic positions* that are part of grammatical constructions that for various reasons are relatively infrequent. Some are optional in the sense that they are stylistically marked alternatives of other constructions and are not used by all speakers. Case marking in these two groups is analyzed below. The third group consists of *other syntactic positions* and includes three positions defined primarily for methodological reasons, namely (*a*) subjects of unfinished clauses (see 7.6.1); (*b*) dialectal complements (see 7.6.1), and (*c*) miscellaneous accusatives (see 7.6.3) The first two of these latter three positions have one case marked token each and the third has none. Case marking in this group will not be further discussed. Adverbs form a special group and are not included in the following analysis. They are discussed separately in 10.4.

Table 30: *Frequencies of syntactic positions*

	CASE	#	%	GROUP
Prepositional comp.	G	5334	35.6	*core*
Annexation	G	2785	18.6	*core*
Object	A	1317	8.8	*core*
Topic	N	1025	6.8	*core*
Comment	N	978	6.5	*core*
Subject in VS	N	974	6.5	*core*
Adverb	A	911	6.1	n/a
Subject in SV	N	406	2.7	*core*
Comp. of *kāna* & sis.	A	320	2.1	*core*
Dialectal comp.	N	307	2.1	*other*
Comp. of *inna* & sis.	A	278	1.9	*peripheral*
Subject of unf. clause	N	141	0.9	*other*
Absolute negation	A	87	0.6	*peripheral*
Number spec.	A	63	0.4	*peripheral*
Absolute object	A	51	0.3	*peripheral*
Mis. accusative	A	13	0.1	*other*

10.3.1 Core syntactic positions

The core syntactic positions are the most frequent and are used by all speakers in the corpus in roughly equal proportions. The core syntactic positions together make up 87.6% of the disambiguated dataset. They include the following positions, grouped by morphological case (see the references in parenthesis for definitions):

NOMINATIVE (7.6.1)

(*a*) subject in SV-clause
(*b*) subject in VS-clause
(*c*) topic (subject of an equational clause)
(*d*) comment (predicate of an equational clause)

GENITIVE (7.6.2)

(*e*) annexed genitive
(*f*) prepositional complement

Table 31: *Case marking by core syntactic positions*

	%	LOG-ODDS	SE	*n*
Comp. of *kāna* & sis.	15.0	−1.73	0.37	281
Object	8.0	−2.44	0.35	1049
Comment	4.6	−3.04	0.36	828
Subject of SV	3.7	−3.27	0.42	287
Prepositional comp.	3.1	−3.45	0.34	3938
Subject of VS	3.1	−3.46	0.37	757
Topic	1.8	−4.03	0.40	765
Annexation	1.5	−4.18	0.36	2246

ACCUSATIVE (7.6.3)

(*g*) object (direct and indirect)
(*h*) complement of *kāna* and her sisters

Assuming that there is some form of linguistic norm governing in which of these position case marking is preferred, speakers are expected to order them roughly the same way in terms of case marking. If on the other hand there is no such norm, then speakers are expected to order them in individually different ways.

Table 31 shows the likelihood of case being marked in central case governance, as predicted by a GLMM on the disambiguated dataset with speaker as random effect. Both the accusative positions, complements of *kāna* and her sisters, and objects, are marked for case at the highest rates, at 15.0% and 8.0% respectively. The higher rates of case marking in these positions is explained by speaker preference for producing case marking in tokens that take orthographic *alif*. Comments are the third most marked position with 4.6%. Below comments in case marking frequency are subjects in VS- and SV-clauses and prepositional complements, all three marked for case at similar rates just above 3%. Topics and annexed genitives are the least marked for case at 1.8% and 1.5% respectively. The only statistically significant differences between successive categories of syntactic positions as ordered in the table are those between the first three ($p<.001$) and between subject of VS-clause and topic ($p=.039$).

The fact that only some of the successive pairings in Table 31 show a statistically significant difference indicates that this order of case marking frequency hides considerable inter-speaker variation. This variation is shown in Figure 15.

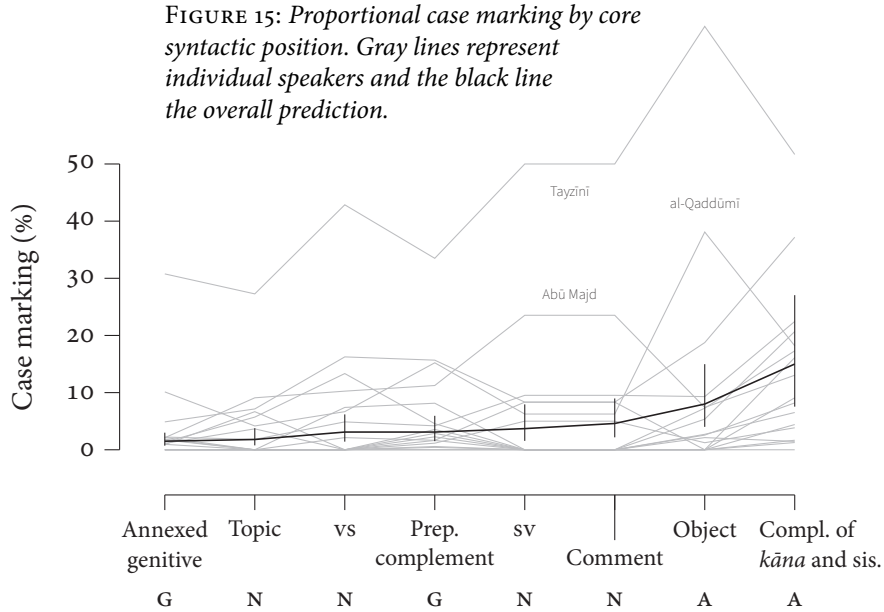


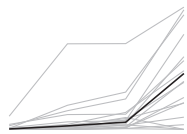
FIGURE 15: *Proportional case marking by core syntactic position. Gray lines represent individual speakers and the black line the overall prediction.*

The figure shows that there is no clear pattern in how speakers distribute case markers in central case governance. No speaker marks case in the core syntactic positions in the predicted order of case case marking frequency. The position of object is, for example, the second most likely to be marked for case given the overall prediction, yet six of the seventeen speakers never mark case on a token in this position. Some speakers have 'favorite' syntactic positions to mark nominals for case. Abū Majd, for example has a preference to mark case on subjects in SV-clauses and on comments, both of which are consistently unmarked for case by ten other speakers. Tayzīnī and al-Qaddūmī both have a preference for marking case on objects, contrary to the overall prediction of higher case marking in complements of '*kāna* and her sisters'. There is, in other words, no clearly consistent pattern of how case markers are distributed across tokens in various syntactic positions.

10.3.2 Norms of case marking in syntactic positions

The data presented above indicate that the correlation between case marking and syntactic position is not governed by covert linguistic norms in Spoken Standard Arabic. Speakers develop individual patterns of case marking of nominals in various syntactic positions, preferring to case mark nominals in certain

positions, and this may possibly reflect individual differences in proficiency in Standard Arabic. Conversely, most speakers have certain positions in which they completely ignore case marking on nominals and where other speakers may have a preference to mark case. This idiosyncratic variation can be compared with the neat pattern by which case marking is distributed on tokens according to definiteness as show Figure 13 on page 208 (reproduced here in the margin to the for ease of reference). Case marking is hierarchically structured by definiteness which may indicate that speakers act under a norm that regulate case endings by this property. There is no clear such norm in how case markers are distributed by core syntactic positions.



Some of this idiosyncratic variation in case marking with respect to syntactic position, namely the unevenness shown in Figure 15, can be attributed to the small amount of data for each data point in the syntactic position variable. Here the data is sliced by eight categories rather by four categories, as is so for types of definiteness. This results in smaller amounts of data for each individual data point, making case marking rates of tokens in specific positions for individual speakers less accurate. (Seven of the 136 data points in the figure represent fewer than ten tokens). It is possible that more data would have revealed a clearer pattern of distribution of case marking across syntactic positions. The data examined here, however, indicates that there is no overarching norm governing case marking in various syntactic positions.

It is furthermore striking that while speakers ignore case marking on tokens in some positions, they do this in ways that produce patterns of case marking that are balanced across the three cases as was shown in the previous section. If a given speaker for example never marks case on tokens in comment position, then he marks case on tokens in some other nominative position to compensate for the lower rate of nominative case marking. This results in the level lines between nominative and genitive case marking in Figure 14 above, despite the different ways of distributing case markers across syntactic positions. This pattern may be interpreted as a norm of equal representation of the three cases in case marking patterns.

To a large extent, equal marking of case endings means equal representation of surface morphological forms of case markers: 85.0% of tokens in the corpus are in the triptote paradigm and this paradigm is by far the most heavily marked for case in the disambiguated dataset (see 9.1.1). Equal distribution of case marking across the three cases thus to a large extent means equal distributions of surface forms of case markers. Conversely, over-representation of case marking of one case would entail over-representation of specific endings. If a speaker for example marks some percentage, even a very small percent-

age, of nominatives but no genitives, a listener would hear the occasional *-u* and *-un* endings of the triptote paradigm and also notice the absence *-i* or *-in*. This would give the impression that the speaker has only partially mastered Standard Arabic grammar. If a speaker however marks case in some nominative positions and not in others, or in one of the two genitive positions and not in the other, they can 'get away with' applying only parts of the case system and still give the impression of an rounded competence in Standard Arabic grammar.

The situation is only superficially different for the distribution of accusative case marking. The rates of case marking of accusative is higher for all speakers in the corpus and is increased by the influence of orthographic *alif*. However, comparison of the two accusative case positions in the far right of Figure 15 shows that speakers in the corpus tend to prefer to mark case in one or the other of these two positions, with large differences between the two resulting in steep slopes in the figure.

This can be summed up as follows: there is a pattern of case marking being evenly distributed across the three cases. This may result in very uneven distributions of case markers across syntactic positions for each distinct morphological case. Speakers have different patterns of case marking in different syntactic positions, but they differ in a way such that frequencies of case marking is roughly equal across the three cases. An effect of this is that surface forms of case markers are used at similar rates by speakers.

10.3.3 Subjects and topics

The figures presented in Table 15 show that speakers mark case at very equal rates for subjects in VS- and SV-clauses. These in turn differ from in their marking rate for topics.

Subjects preceding the verb are analyzed in radically different ways in the Western grammars of Arabic and in Traditional Arabic Grammar (see 7.6.1). In Western grammars, SV and VS are usually regarded as optional word orders of the same underlying structure. In Traditional Arabic Grammar, SV and VS are considered to be a fundamentally different clause types, with case assigned to the subject through different operations. The initial constituent in a SV-clause is considered not to be the subject of the verb in Traditional Arabic Grammar, but rather a topic (*mubtadaʾ*) being akin to the topic in a verbless equational clause. If speakers add case to subjects according to this scheme, then topics and subjects in SV-clauses would be expected to have similar rates of case marking. This could, for example, be achieved with a simple rule stating that clause initial constituents are nominative. If, on the other hand, subjects of VS- and

SV-clauses are marked with the same operation, by assigning nominative case to subjects of verbs, then these two are expected to have similar rates of case marking and a rate different to that of topics.

As was shown in Table 31 above, subjects of verbs in SV- and VS-clauses are marked at very similar rates: 3.65% and 3.05% respectively. The small difference between them of 0.6% is not significant in the model (p=.622). Topics are case marked at a lower rate, 1.75%. This is significantly different from the case marking of subjects in VS-clauses (p=.039) and from subjects in SV-clauses (p =.025).

With the reservation that there may be other factors affecting these numbers that are not accounted for here, these figures provide some indication that speakers process topics of equational sentences differently than they do subjects of explicit verbs, regardless of whether the verb precedes or follows the subject. This would contradict the theoretical presumptions of Traditional Arabic Grammar and thereby be in conflict with dominant methods of first language instruction.

10.3.4 Peripheral syntactic positions

The peripheral positions differ from the core positions discussed above in that they are optional or otherwise infrequent, with a total of 395 tokens occurring in these positions, or 3.6% of the disambiguated dataset. These peripheral positions are also used to very different extents by the speakers in the corpus. Three of the four positions are not used at all by at least two speakers. These positions, all governed for accusative, are the following (see 7.6.3):

(*a*) absolute object
(*b*) absolute negation
(*c*) complement of *inna* and her sisters
(*d*) number specification

Counts of these positions in the disambiguated dataset are listed in Table 32. Due to the limited data, case marking of tokens in the peripheral positions will not be analyzed statistically as is case marking of tokens in the core positions. The conclusions presented below, where case marking on tokens in the four peripheral positions are commented on in due order, are therefore of a more preliminary nature.

1. Absolute object. This construction conveys adverbial content and can be seen as an alternative formulation for other, more common forms of adverbials. The

Table 32: *Counts of case marking in peripheral syntactic positions, listing total number of tokens (tot.) and number of case marked tokens (mar.) in each position. Speakers are ordered by decreasing overall rates case marking. Highlighted cells are referred to in the text.*

	ABSOLUTE OBJECT		ABSOLUTE NEGATION		*INNA* & SIS.		NUMBER SPEC.	
	tot.	mar.	tot.	mar.	tot.	mar.	tot.	mar.
Tayzīnī	6	5	5	4	19	7	1	1
Badīᶜ	17	14	5	4	14	0	4	0
al-Qaddūmī	0	0	17	13	20	1	0	0
Abū Majd	13	8	3	0	8	0	3	0
Mursī	0	0	7	6	8	0	0	0
Kayālī	2	1	5	1	26	0	1	1
Fayyāḍ	0	0	12	5	3	0	0	0
al-Muᶜallim	0	0	4	3	23	0	0	0
al-Barādiᶜī	3	0	5	3	13	0	11	8
ᶜAbd al-Qādir	0	0	4	2	4	0	3	0
al-Xuḍarī	2	0	1	1	6	0	4	0
Hilāl	0	0	6	4	7	1	3	0
al-ʾAsᶜad	2	2	0	0	8	0	3	0
Šallaḥ	1	0	4	0	25	0	5	0
Ġalyūn	0	0	2	1	8	0	3	0
ᶜArīqāt	0	0	4	1	6	0	18	0
al-Miṣrī	2	0	0	0	3	0	3	0
Total	48	30	84	48	201	9	62	10

use of the absolute object is in and of itself a marked stylistic feature, regardless of case marking. In (121a), for example, the word *tāmman* 'complete' is part of an absolute object with the adverbial meaning of 'completely'. In (121b), uttered by the same speaker, an adverb of the same root, *tamāman*, is used with the same meaning.

(121) a. *wa-narfuḍu-hu* ***rafḍ-an*** ***tāmm-an***
and-1PL.refuse-it refusal-ACC complete-ACC
'And we refuse it completely.' (al-Asᶜad, 21:01)

b. *wa-naḥnu* *naᶜrif* ***tamām-an*** *wa-nasiq* *bi-ʾanna*
and-we 1PL.know complete-ACC and-believe in-COMP
'And we are certain (lit. 'we know completely') and believe that …'
(al-Asᶜad, 7:44)

Table 32 shows that absolute objects are used primarily by Badīᶜ, Abū Majd, and to a lesser extent by Tayzīnī who are all speakers with a high overall rate of case marking. When these speakers' results are taken together they produce 36 absolute objects, 27 of which are marked case. The other fourteen speakers in the corpus produce zero to three absolute objects each for a total of twelve tokens and three of these are marked for case. These numbers indicate that the absolute object is not widely used in Spoken Standard Arabic, at least as represented by this corpus, but that it is a stylistic feature used extensively only by speakers using a high level of formal correctness in their speech. These speakers also mark the absolute objects for case at a high rate.

11. Absolute negation. This is arguably the least difficult position to mark for case. It is marked by an invariable final *-a*, regardless of paradigm.[103] As with the absolute object, absolute negation can be seen as a stylistically marked alternative formulation for other grammatical constructions. In (122a), for example, the existence of 'an absolutely closed structure' is negated with absolute negation *lā wujūda* 'no existence'. In (122b), the directly following utterance, the existence of *al-muṭlaq bi-ḏātihi* 'the absolute itself' is instead negated with *ġayr* 'not', 'un-'.

(122) a. *ʾana* *ʾaᶜlam* *ʾanna* … *ʾanna* *bunyat-an* *muġlaqat-an*
I 1S.know that … that structure-ACC closed-ACC
bi-ṣūrat-in *muṭlaq* ***lā*** ***wujūd-a*** *la-ha*
in-way-GEN absolute NEG existence-ACC for-it
'I know that …that there does not exist [such a thing as] a completely closed structure.' (Tayzīnī, 14:54)

b. *ʾal-muṭlaq-u* *bi-ḏāt-i-h* ***ġayr-u*** ***mawjūd***
DEF-absolute-NOM in-self-GEN-its NEG-NOM existent
'The absolute itself does not exist.' (Tayzīnī, 15:02)

[103] Theoretically, a sound m.pl. could be governed by absolute negation, for example *lā*

Unlike the absolute object, however, the instances of absolute negation are distributed fairly evenly amongst the speakers in the corpus. It is marked for case in roughly half of these occurrences, seemingly independent of speakers' other uses of case marking. The proportion of case marking in absolute negation does not seem to decrease markedly with speakers further down the table who have lower overall rates of case marking, even if the construction itself is used less.

Furthermore, of the 84 instances of absolute negation in the disambiguated dataset, 55 are either of the two phrases *lā šakk(a)* 'no doubt' and *lā budd(a)* 'must', 'necessarily' (see also 8.4). These phrases both occur with and without case marking. Many speakers use both the marked and unmarked forms. al-Qaddūmī and Fayyāḍ stand out in Table 32 as heavy users of absolute negation, with seventeen and twelve instances respectively. They are however best described as heavy users of the phrases *lā šakk(a)* and *lā budd(a)* since these expressions represent all but two each of their total counts of absolute negation. If these two phrases are excluded, there are a mere 29 instances of absolute negation left in the corpus, 19 of which are marked for case.

While limited, these data suggest that case marking in absolute negation contexts does not covary with other forms of case marking. Speakers add the final vowel ending to absolute negation independently of other patterns of case marking. Also speakers with very little case marking sometimes mark case in this construction.

III. Complement of inna *and her sisters.* The third peripheral syntactic position is the complement position of *inna* and her sisters that are a set of particles that govern the accusative. Examples of these constructions occur in all interviews but are fairly infrequent. This is due to the widespread use of alternative particles or constructions: the use of the dialectal particle *innu* rather than the Standard Arabic *anna* and *inna*, a tendency to use direct rather than indirect speech thus dispensing with many complementizers, and the use of *lākin* 'but' that does not govern case in comparison with the conjunction *lākinna* which does.

The only speaker who frequently marks case on the complements of *inna* and her sisters is Tayzīnī who marks case on seven of nineteen complements. Two other speakers case mark complements once each. This gives a total of nine case markers out of a total of 201 tokens. Thus, with the exception of Tayzīnī, complements of *inna* and her sisters are by and large unmarked for case.

miṣriyyīn 'there are no Egyptians' (cf. Badawi et al. 2004:466), but there are no examples such as this in the corpus.

iv. Number specification. The number of occurrences of the final peripheral position under discussion, number specification — a counted noun governed for accusative by the numeral, is very much dependent on the topic of the conversation, that is whether or not the speaker needs to relate numerical facts. The position is most frequent in the interview with ᶜArīqāt who uses it primarily in descriptions of the geographical extent and the economical consequences of the occupation of Palestine. However, none of his eighteen number specifications are marked for case. The only speaker to make use of number specification and also mark it for case is al-Barādiᶜī. Of eleven number specifications, he marks eight for case. However, these are all in one and the same fixed phrase, *[NUMBER] ᶜāman* '[NUMBER] years' (see 8.4). Apart from al-Barādiᶜī's interview, there are only two instances of number specification marked for case in the corpus. Hence, it is not possible to draw general conclusions on case marking in number specification from these limited data.

To sum up, absolute objects are used primarily by a few speakers and are often marked for case. Absolute negation is used sparingly and its complement is often marked for case, even by speakers with low overall case marking. Complements of *inna* and her sisters are by and large unmarked for case. The data for case marking on numbers specification was too limited and its use too skewed by the topic of the discussion for any conclusions to be drawn.

10.4 Adverbs

Adverbs differ from the tokens in the core and peripheral positions discussed above in that they are mostly syntactically independent from the rest of the clause. Furthermore, the category of adverbs includes a large number of words with a lexicalized case ending so that they are always used with the case ending even in dialectal contexts. For these reasons, tokens coded as adverbs were not included in the disambiguated dataset (see 8.1.1) and were thus excluded in most of the previous analyses.

As was described in 7.6.3, sub-categorization of adverbs is complex and would require extensive theoretical development in order to handle borderline cases. In this study, this problem was bypassed by applying a broad definition of accusative marked adverbs. This avoids having to develop detailed methods of sorting the many borderline cases into different subcategories of adverbs and also reduces the complexity of the coding scheme. This does however severely limit the level of detail to which case marking on adverbs in the corpus can be analyzed, and the discussion here will therefore be limited to lexicalized adverbs.

Lexicalized case endings are by definition not subject to variation in case marking, which is the main subject of this study. They are however important for the present study in that lexicalized case markers constitute a large proportion of overt case markers in speech. More precisely, one third, 36%, of all overt case endings in the corpus occur on adverbs in indefinite accusative form, the vast majority of which are lexicalized with case endings. The occurrence of case endings on lexicalized adverbs makes impressions of case marking in speech highly unreliable. A large part of the case endings that are heard in speech are not the result of a productive case system but the result of word formation processes.

The following words, in decreasing order of frequency, occur ten or more times in the corpus, and are, with two exceptions,[104] consistently marked for case:

ayḍan	also
awwalan	firstly
abadan	never
maṯalan	for example
ḥaqīqatan	actually
fiᶜlan	indeed; actually
xāṣṣatan	especially
tamāman	completely; exactly
jamīᶜan	all
dāʾiman	always
ṯāniyan	secondly

These are only a fraction of all lexicalized adverbs. There are many more adverbs in the corpus that are consistently used with the *-an* ending but that have fewer than ten occurrences each in the corpus. There are also adverbs that exhibit variable forms, occurring both with and without a case ending. The adverb, *marra(tan)* 'once', for example, stands out in being a highly frequent indefinite adverb without being lexicalized with a case ending. It has 23 occurrences in the corpus, only five of which are pronounced with the case ending as *marratan*, and each by a different speaker. Note that this cannot be explained by it having *tāʾ marbūṭa* and thus not taking orthographic *alif*. Two of the words in the list above, *xāṣṣatan* and *ḥaqīqatan*, also have *tāʾ marbūṭa* but are

[104] The two exceptions are one adverbial use of *awwal* by Šallaḥ (45:09) in a heavily dialectal context, and one use of *xāṣṣa* by al-Qaddūmī (11:07*). The latter may be a way of raising the register by producing a pausal form that would normally only appear in formal read speech. The token itself is not in pause position.

nevertheless lexicalized with the ending.

Indefinite adverbs are a productive class primarily as formed from *nisba*-adjectives, such as *šahriyyan* 'monthly', from the adjective *šahrī*; or *dawliyyan* 'internationally', from the adjective *dawlī*. There are 42 tokens and 27 types of this kind of adverb in the corpus. Two particularly creative uses of *nisba*-derived adverbs are *ʿaqāʾidiyyan* 'with regards to belief' in (123) and *isrāʾīliyyan* 'with regards to Israel' in (124).

(123) *hunāk* *xilāf* *tāmm* *fikriyy-an* *wa-**ʿaqāʾidiyy-an***
there disagreement complete thought.ADJ-ACC and-belief.ADJ-ACC
'there is complete disagreement with regards to both thought and belief' (al-Barādiʿī, 24:13)

(124) *ʾal-matrūḥ* ***ʾisrāʾīliyy-an*** *huwa* *baladiyya*
DEF-suggested Israeli-ACC COP municipal
'The Israeli suggestion is [to form] a municipal' (ʿAbd al-Qādir, 14:16)

Adverbs are also formed from definite nouns, but much less frequently than from indefinite nouns, both in terms of types and tokens. There are a total of 167 tokens in the corpus of adverbials with the definite article. The vast majority of these, 145 tokens, are either of the two words *al-yawm* 'today' (lit. 'the day') and *al-ān* 'now' (lit. 'the time'). There are a further nine instances of *al-ḥaqīqa* 'actually', six instances of *hāḏihi l-marra* 'this time', and eight other types with one token each.

The only instance in the corpus of case marking on a definite adverb is in example (125) where *al-āna* has the accusative ending *-a*. It is in other words clear that even very frequent adverbs with the definite article are not marked for case. This is in accordance with the norm of not marking case on words with a definite article (see 9.2). This norm means that definite adverbs are not used with case endings and are not lexicalized in that form. Indeed, adverbs with the definite article could be interpreted as being lexicalized as *unmarked* for case. The word *al-āna*, while prescriptively correct, is an oddity and indeed unique in the corpus.

(125) *min* *ʾajl-i* *l-xalāṣ-i* *mim-mā* *naḥnu* ***l-ʾān-a*** *fī-h*
from sake-AMB DEF-salvation-GEN from-which we DEF-time-ACC in-it
'in order to save [us] from that in which we now find ourselves' (Tayzīnī, 16:06)

In summary, it is not possible with the annotation in its present form to conduct a detailed analysis of adverbs, and only lexicalization of adverbs has

been disused here. There are a large number of adverbs formed from indefinites that are lexicalized with case endings, but there are also frequent adverbs, for example *marra* 'once', that are not. Adverbs formed from definite nouns, such as *al-yawm* 'today' are consistently unmarked for case in the corpus, with one single exception.

10.5 Pause

In traditional rules of recitation, a final short vowel and nunation is omitted before a pause (see 4.5) and this often involves omission of the case ending. Each word thus has a context form, in which the ending is retained, and a pause form, in which the final short vowel is omitted and that therefore in most paradigms is unmarked for case (see Table 8 on page 127). Traditional Arabic Grammar specifies in detail what the pause and context forms are for particular types of words, but has little to say about where the pausal forms are to be used. This in effect leaves room for interpretation of how and where the pause rules are to be applied, and the lack of case marking in speech is often seen as an extension of the pause system. It is possible that case marked context forms are disfavored in pause position due to influences from formal read speech and recitation. On the other hand, speakers of Standard Arabic may opt to exploit syntactic and morphological situations where they are comfortable marking case and thus override pause rules as they do so. Pause would then have little or no effect on case marking. This latter view is held by Meiseles (1977:178–9), according to whom "contextual forms may occur in relative freedom before pause […] even at the end of a sentence." He reports 183 context forms in his corpus of which 24 are sentence final.[105] The analysis presented below shows that speakers, at least in the present corpus, do not use case marked context forms at the end of sentences, contrary to Meiseles' findings.

Pause can be defined either syntactically, as the end of syntactical units on the phrase or clause level, or phonemically, as a stop in the flow of speech. Neither phonemic pauses nor phrase boundaries were coded in the corpus and thus their effect on case marking cannot be tested here. The corpus text was however segmented into utterances that are defined as a main clause together with attached material (see 6.4). Thus, pause is defined as the end of an utterance for the purposes of this analysis. Tokens at the end of utterances terminated with an interruption, self interruption, or tailing off are not considered

[105] Meiseles (1977:178n) states that "adverbial *-an* […] is irrelevant to our discussion". This is here interpreted to mean that he excludes them from counts of case marking in pause position, as is also done here.

to be in pause. These are not a planned or voluntary sentence termination and would not be marked as such by the speaker. Furthermore, only tokens in the disambiguated dataset whose case endings are dropped in their pausal form are included in the following analysis (i.e. tokens in absolute state in the triptote, diptote, sound f.pl., and five nouns paradigms, n=9132).

The effect of pause, as defined above, on case marking in the present corpus is straightforward. There are 1865 tokens in the corpus that are in utterance final position and that are of a type in which the pausal form is unmarked for case. Of these 36 are marked for case. This corresponds to likelihood of 1.1% of words in pause being marked for case, as predicted by a GLMM with speaker as random effect (log-odds=−4.33, SE=0.39). The corresponding prediction for tokens in non-final positions is 3.0% (log-odds=−3.47, SE=0.36). The difference is significant (p<.001). Words in utterance final position are less likely to be marked for case than are other words. More importantly, of all case marked context forms in utterance final position, all but two are accusative triptotes without *tāʾ marbūṭa*, that is tokens that take the *-an* ending with orthographic *alif*. Words that take orthographic *alif* appear to be unaffected by pause. This was further tested by running the model only on tokens in the disambiguated dataset that take orthographic *alif* (n=498). In this model, there is no statistically significant difference in case marking on tokens in pause or in context position (p=.274). It is thus clear that words that take orthographic *alif* are unaffected by the otherwise very consistent negative effect of pause on case marking, as would also be the case in formal read speech. The pause system is very much alive in Spoken Standard Arabic as represented by examples in the corpus, at least in utterance final position.

10.6 Hypercorrect case markers

Tokens with endings that are not prescriptively correct and that are not attested in the dialectal grammar were annotated as hypercorrect in the corpus text (see 7.7). Note that this does not mean that such case markers are stylistically inappropriate, only that they are not in accordance with prescriptive grammar.

The number tokens with hypercorrect case endings in the corpus is small: 43 in total. They are evenly spread amongst the speakers. This number should be seen in relation to the 1474 tokens with overt case marking (934 if indefinite adverbs are excluded) and the 10466 tokens with unambiguously unmarked case. There are no tokens in the corpus that are hypercorrect due to having a form of case marking that is not part of the appropriate paradigm, such as a sound f.pl. with the ending *-an* or a diptote with the ending *-in*.

The hypercorrectly marked tokens form a heterogeneous group. The one pattern that clearly stands out is that tokens with an enclitic pronoun are highly over-represented. Tokens with an enclitic pronoun make up only 6.0% of the total data but half, 21 of 43, of the hypercorrect tokens. This indicate that speakers are more likely to mark case on these words even if they are not entirely sure of the correct form, reflecting the norm of high rates of case marking on these words (see 9.2). There is however no instance of backtracking to add a case ending on words with an enclitic pronoun, despite the unmarked forms of these words always being prescriptively incorrect.

The most common syntactic position among the hypercorrect tokens is the subject of a VS-clause, with 19 hypercorrectly marked tokens. Nine of these are subjects of *kāna* and her sisters marked for accusative rather than the prescriptive nominative, as in (126). Another example is (80) on page 149.

(126) *lan* ***yabqa*** ***ʾarḍ-an*** *natafāwaḍ* *ʿalay-ha* *lan* ***yabqa***
NEG.FUT remain land-ACC.HYP 1PL.negotiate over-it NEG.FUT remain
quds-an[106] *yumkin* *ʾan* *natafāwaḍ* *ʿalay-ha*
Jerusalem-ACC.HYP 3MS.be.possible COMP 1PL.negotiate over-it
'There will not remain any land to negotiate for, there will not remain any Jerusalem to negotiate for.' (ʿAbd al-Qādir, 17:06)

There are also very few instances of case repair in the corpus, that is instances where speakers backtrack to change case marking in an already uttered phrase or sentence. There are a total of eight instances in the corpus of backtracking to correct case marking. Three of these are corrections in the sound m.pl. in nominative position (see (110) on page 207 for an example). There is only one instance of case repair when the speaker backtracks to add a case marker where there previously was none.[107] This indicates that speakers do not perceive unmarked forms as errors to be corrected.

10.7 Summary

This chapter investigated how syntactic properties interact with case marking in the corpus. Phrase heads were shown to be four times as likely to be marked for case than are attributive adjectives that inherit case from the head.

[106] The word *quds* 'a jerusalem' is used here in the indefinite, as apposed to the proper name *al-quds*, and was therefore regarded as a regular (non-proper) noun for purposes of annotation.

[107] This example (al-Qaddūmī, 04:38*) is odd, not only for self-correction to add a case ending, but also in that this is done on a proper name *al-jāmiʿa(ti) l-ʿarabiyya* 'The Arab League', a category of words that are generally unmarked for case.

By and large, adjectives are unmarked for case. Furthermore, adjectives that are marked for case are, with one exception, preceded by a case marked head noun.

The bulk of this chapter dealt with how case markers are distributed across the three morphological cases under discussion as well as in the various ways in which each individual case is governed. It was shown that speakers have individually diverse ways of distributing case marking across grammatical constructions, preferring case marking in some positions and ignoring it in others. There appears to be no overarching linguistic norm governing how case is marked in different syntactic positions as there is in types of definiteness. However, the speakers in the corpus mark case on tokens in different positions in a way that is evenly balanced across the three morphological cases. This was interpreted to mean that the way the production of case markers is distributed in different syntactic positions is not governed by covert linguistic norms, but the way it is distributed across the three cases is. This means that many speakers in the corpus mark case primarily on tokens in a limited set of positions but still use a variety surface forms of case endings.

A large number of adverbs have been lexicalized with the case ending *-an* in Standard Arabic and these are clearly exemplified in the present corpus. There is, however, at least one highly frequent adverb, *marra* 'time', that has not gone through this process of lexicalization and is used both marked and unmarked for case. There is also a small number of highly frequent adverbs formed from definite nouns, most prominently *al-ān* 'now' and *al-yawm* 'today', that are consistently unmarked for case and that can be regarded as lexicalized in their unmarked form.

Speakers in the corpus consistently use pausal forms at the end of utterances and thus the pause system can be said to be very much alive in spoken Spoken Standard Arabic.

The number of hypercorrect case endings in the corpus is small, indicating that speakers only mark case if they feel confident that they can do so correctly. It is not possible to identify clear patterns in this small number of tokens other than that words with an enclitic pronoun are highly over-represented for hypercorrect case marking. This provides further evidence of a norm of a high rates of case marking in these words.

Summary and conclusion

11

In formal Arabic speaking contexts, one is expected to use the standard variety of the language. The topic this dissertation has been one important grammatical distinction that sets Standard Arabic apart from non-standard dialects, namely the presence of a system of morphologically marked case. Case endings are traditionally regarded as a crucial and defining aspect of Standard Arabic and their importance is strongly emphasized in formal schooling, the primary way of acquiring Standard Arabic. Very few speakers however, if any, have complete or native-like command of Standard Arabic, and it is typically spoken with incomplete and inconsistent use of case endings as defined by traditional descriptions, with case endings appearing only sporadically. The case system is widely considered to be difficult and complicated, and, for many speakers, learning and using the system of morphological case is associated with anxiety and feelings of linguistic inadequacy. Parkinson (1994a:207–8) poignantly describes the use of morphological case in Standard Arabic as speakers using case endings by "sprinkling them through their text at a certain rate to give the flavor of fuṣḥā [Standard Arabic] without making it so difficult on themselves they would not be able to speak at all." This sprinkling, he continues, is structured by morphosyntactic parameters. The overarching aim of the present study has been to find, describe, and quantify these patterns.

There are several different ways of approaching formal spoken Arabic as a linguistic variety. It has been defined in a number of ways in the scholarly literature depending on how the researcher attempts to account for the wide range of variation often found in this form of speech. For the purposes of this dissertation, the variety under investigation, Spoken Standard Arabic, was defined by language external criteria as the extemporaneous speech in the most

formal register of highly educated and competent native speakers of Arabic. Defined this way, Spoken Standard Arabic includes features traditionally regarded as non-standard or as incorrect Standard Arabic. The choice was nevertheless made to refer to this variety as Spoken Standard Arabic, which is based on the assumption that variation is a natural feature of spoken linguistic varieties and also of spoken standard varieties, and that conformity with traditional codified grammar is not in and of itself a good measure of standardness. The latter point is especially important in Arabic, where Traditional Arabic Grammar — which additionally wields a strong influence over modern Western grammars of Arabic — is highly prescriptive, harks back several centuries, and does not represent linguistic developments of the last centuries.

Due to the sporadic appearance of case marking in speech, patterns of case marking are not directly obvious, and finding such patterns requires systematic analysis of a large body of material from a number of speakers. To this end, a corpus of transcribed news interviews was constructed and annotated so that it could be searched for tokens of the use of case endings by speakers selected for inclusion in the corpus. Data from this corpus was then quantitatively analyzed for correlations of case marking with various morphosyntactic parameters. Several patterns of case marking were found and are summarized in this chapter. First the overarching characteristics of case marking in the corpus are described in section 11.1. This is primarily a summary of Chapter 8. In section 11.2, six particularly prominent patterns of the distribution of case marking in the corpus and their relationship to grammatical environments are listed and commented upon. Two types of patterns are identified: patterns specifying grammatical environments where case is generally unmarked on tokens, and patterns of how case marking is distributed on tokens in remaining environments. Some pedagogical implications of these findings are discussed in section 11.3. Finally, section 11.4 points to some questions raised by the results presented in this study and proposes avenues for further research.

11.1 General characteristics of case marking

One of the most striking aspects of the data at first glance is the range of very different extents to which speakers engage with the case system. In the measure of overall case marking described in 8.1, in which ambiguous endings and lexicalized forms are controlled for, rates of case marking for speakers vary between 0% and 42%, with an average of 7%. Eleven of the seventeen speakers fall within a range of 1.5% and 10%. The speaker in the corpus with the most case marking in his speech, 42%, stands out as a statistical extreme. Still, he

produces less than half of all potential case endings. At the other end of the scale, there are speakers who produce only a few case markers in each entire half hour interview. These rates of case marking are largely stable throughout each of the interviews (see 8.2).

As expected in the continuum model of diglossia, the use of case endings in the corpus negatively correlates with the amount of dialectal features found in an individual's speech as measured by three dialectal features (dialectal nominals, constituents, and relative pronouns, see 8.3.1). The usage rate of case endings was tested with constraints of these dialectal features on case marking controlled for. While the dialectal features showed strong internal correlations, each of them had a much weaker correlation with case marking. The use of one dialectal features is highly predictive as to how much a speaker uses other dialectal features, but less indicative of how frequently the speaker marks case. There are, in other words, many different ways of combining case marking and dialectal features in Spoken Standard Arabic.

Finally, it is important to note that the speakers whose speech is represented in the corpus generally do not rely on a set of fixed phrases to include case endings in their speech. In fact, the speakers seem to avoid reusing phrases with case marking (see 8.4.1). The case system, to the extent to which it is used by speakers in the corpus, is used productively.

11.2 Patterns of case marking

This study has identified a number of patterns as to how speakers in the corpus distribute case markers according to morphosyntactic parameters. Six of these patterns are particularly prominent and they are here divided into two groups. The first group is *patterns of unmarked case*, three situations in which case marking on tokens is very rare. The second group is *patterns of proportional case marking*, three patterns of distribution of case markers in morphosyntactic environments other than those that are generally unmarked for case as specified in the first group. A number of other patterns were found but are not listed here, either because of having a smaller impact or because they are less consistent across speakers.

The patterns listed below should be interpreted in the light of the overarching characteristics of case marking described in the previous section. The patterns are not predominantly found in any particular section of the interviews; the rate of case marking for each speaker is largely stable throughout the interview. Nor are they the result of speakers reusing a number of fixed phrases. Furthermore, these patterns are found within transcripts with varying rates of

case marking. These patterns describe how the case markers that are used in the corpus are distributed, largely independently of how frequently case is marked by the individual speaker.

11.2.1 Patterns of unmarked case

Since case is marked only sporadically by most speakers in the corpus, categories of words that are systematically unmarked for case blend into the background of unmarkedness, as it were, and are therefore difficult to detect through impressionistic observation. Identifying such categories is nevertheless important in that they constitute limitations on the reach of the system of case marking in Spoken Standard Arabic. The likelihood that words in these patterns are marked for case is very much lower than expected given the overall average of case marking and with interacting effects controlled for (see the cross-references below for details and quantitative estimates). Case marking in words in these positions can be regarded as exceptions in the corpus. Furthermore, they do not apply to words in which case is orthographically marked (see pattern *e* below). With these caveats, the patterns of unmarked case in the corpus are the following:

(*a*) *Case is not marked on words with the definite article* (e.g. *al-amr* 'the issue', see 9.2). This is the most important of the patterns of unmarked case. It is in direct conflict with prescriptive grammars and applies to a word form that is extremely frequent. This pattern is a characteristic of Spoken Standard Arabic that clearly sets it apart from formal read speech and recitation and is an important finding from this corpus study.

(*b*) *Pausal forms are used in utterance final position* (e.g. *amr [.]* 'issue [.]', see 10.5). Context forms in utterance final position (pause) are extremely rare in the corpus. This is, as opposed to other patterns listed here, in accordance with prescriptive grammar. The usage rate of the accusative *-an* ending is unaffected by pause, but only when this would be orthographically marked.

(*c*) *Case is not marked on attributive adjectives* (e.g. *amrun muhimm* 'important issue', see 10.1). This pattern only partly contradicts prescriptive grammar, depending on how the rules rules of pause are interpreted.[108] In noun-adjective pairs where both words take orthographic *alif* the adjective follows the head noun in terms of case marking.

[108] Since the adjective is normally final in the nominal phrase, it can be interpreted as being

These patterns constitute limitations on the reach of the system of case marking in Spoken Standard Arabic as represented in the corpus. The sentence in (127) was constructed to illustrate this. Example (127a) contains no saliently dialectal features and no case markers, which is typical of large parts of the corpus. With prescriptively correct case marking the sentence would be read as in (127b), assuming a pause after the final word. This sentence is typical of formal read speech, as in news broadcasts for example. Given the patterns of unmarked case listed above, case marking would, in Spoken Standard Arabic, typically only (and optionally) be added to the second of the four nominals in the sentence, namely to the prepositional complement *šakl*, as in (127c). The other nominals in the sentence each fall under one of the patterns of unmarked case: the subject *al-ᶜamal* 'the work' under *a* as having the definite article, the adjectival attribute *mubāšir* 'organized' under *b*, and the final nominal *šahr* 'month' under *c* as being utterance final.

(127) a. *badaʾa* *l-ᶜamal* *bi-šakl* *munaẓẓam* *baᶜd* *šahr*
start DEF-work in-manner organized after month

b. *badaʾa* *l-ᶜamal-**u*** *bi-šakl-**in*** *munaẓẓam-**in*** *baᶜda* *šahr*
start DEF-work-NOM in-manner-GEN organized-GEN after month

c. *badaʾa* *l-ᶜamal* *bi-šakl-**in*** *munaẓẓam* *baᶜd* *šahr*
start DEF-work in-manner-GEN organized after month
'The work began in a organized manner after a month.'

The patterns of unmarked case thus put a cap on the amount of case marking in Spoken Standard Arabic at a rate that is far below the prescriptive ideal. A hypothetical speaker who marks case on all nominals except those falling under the three patterns of unmarked case would have an overall rate of case marking of 48%, as calculated in the average proportion of tokens falling under patterns *a*–*c* in the interviews.[109] Complete, prescriptive case marking is not only a higher rate of case marking, it also deviates from patterns of unmarked case found in this corpus.

in pause position and thus prescriptively unmarked for case. It is thus possible that this pattern could be subsumed under a pause pattern in which pause is defined phonemically and/or as a phrase boundary, an option that was not pursued here (see 10.5).

[109] This number was calculated using the disambiguated dataset and is thus directly comparable to the overall rates of case marking listed in Table 14 on page 168.

11.2.2 Patterns of proportional case marking

The patterns of unmarked case listed above describe situations in which case marking is very rare. The case endings that are found in the corpus thus appear elsewhere. The patterns of proportional case marking listed below describe how they are distributed in these other situations. These patterns are the following:

(*d*) *Case is marked at higher rates in words with an enclitic pronoun than in other types of definiteness* (e.g. *amru-nā* 'our issue', see 9.2). The unmarked forms of these words (e.g. *amr-nā* 'our issue') are prescriptively incorrect regardless of case, and are so also in pause position. Speakers with high rates of case marking mark case at rates over 50% on these words. Speakers producing only a few case markers do so only in words falling under this and the following pattern.[110]

(*e*) *Case endings that are orthographically represented (in unvowlled text) are used at high rates*. This applies to:[111]

i. sound m.pl. and dual nominatives,[112] (e.g. *miṣriyyūn/-ān* 'Egyptians (m.pl./dua.)' see 9.1.2).

ii. indefinite triptote accusatives without *tāʾ marbūṭa*, that is words that would be written with orthographic *alif*, (e.g. *amran* 'issue' see 9.3)

Speakers with high rates of case marking mark case at rates over 50% in these words. Speakers producing only a few case markers do so only words falling under this and the previous pattern.

[110] al-Miṣrī and ᶜArīqāt are two such speakers appearing in the corpus. al-Miṣrī produces two overt case markers: in a nominative sound m.pl. (*e.i*) and in an indefinite triptote accusative without *tāʾ marbūṭa* (*e.ii*). ᶜArīqāt produces four overt case markers: in a word with an enclitic pronoun (*d*), in an indefinite accusative triptote without *tāʾ marbūṭa* (*e.ii*), in a dual (*e.i*), and in an indefinite noun governed by absolute negation. The latter is something of a oddity in the case system in that it is used in equal measure by all speakers regardless of other patterns of case marking (see 10.3.4). The occurrences of overt case markers for these speakers in Table 14 on page 168 are slightly lower then the ones listed here since they are based on the disambiguated dataset in which some categories of words are excluded.

[111] There are three further situations where case is orthographically marked (see 4.6), but these are very infrequent and no conclusions regarding their case marking can be drawn on the basis of the corpus.

[112] The genitive/accusative in these paradigms (with *-īn* and *-ayn* respectively) were here regarded as not marking case since they are very similar to or identical with dialectal forms.

(*f*) *Case marking is proportionally balanced in the three cases* (see 10.2). This pattern is not applicable to speakers who produce very few case endings. For speakers with middle to high rates of case marking, this patterns means that they use at least some of each of the three different case endings in the triptote paradigm. Many speakers produce these patterns by marking case in a way that is highly skewed in its distribution amongst the syntactic positions providing the contexts for each realization of morphological case. There is no clear evidence of some 'safe' positions being marked for case at higher proportions than other positions. The effect of orthographic *alif* (pattern *e.11*) is added to this patterns, to the effect that case is in absolute terms more often marked in the accusative.

The way case markers are distributed by speakers in the corpus is thus highly structured by morphosyntactic features, firstly in some categories of words being largely excluded from case marking, and secondly by case markers following certain patterns of proportional distribution. The patterns listed above are all highly consistent between speakers. These patterns (with the exception of pattern *b*) are not part of formal language instruction and are indicative of the existence of covert linguistic norms governing, and indeed limiting, the use of case endings in Spoken Standard Arabic.

11.3 Pedagogical implications

The primary motivation for conducting this study was to inform Standard Arabic pedagogy and curriculum development. The following is a brief, non-exhaustive discussion of some general points about how the findings presented above may inform teaching. These points are based on the assumption that Spoken Standard Arabic as it represented by the present corpus is an appropriate form of the language for students of Standard Arabic strive for. It entails partial abandonment of the prescriptive ideal and runs against commonly held notions of linguistic correctness in Arabic.

It is also assumed here that oral proficiency in Standard Arabic is a learner aim. This may not necessarily be the case seeing as how Standard Arabic is primarily written and as how the dialects fulfill most spoken language functions. In most institutions where Arabic is taught to non-native speakers, oral proficiency in Standard Arabic is nevertheless an explicit aim, sometimes paired with instruction in dialect (Nielsen 2009:151–2). Arabic teaching as a first language in the Arab world focuses exclusively on Standard Arabic but typically does not pay much attention to oral proficiency (see 4.3). There is however a strong argument to be made that it should, in order to provide access to a

formal spoken register and thereby facilitate participation in the public sphere for larger portions of the population. In short, wherever oral proficiency in Standard Arabic is a goal of teaching, these remarks may be relevant.

One first thing is that case marking is included in Spoken Standard Arabic through a process of optional addition. The data presented in this study show that for competent speakers of Standard Arabic word forms with unmarked case is the default. The average rate of case marking is 7%, and it is highly unlikely that this is reached by omitting 93% of all case endings, rather than by adding the 7% attested rate of case endings to an underlying unmarked form. Speakers of Standard Arabic typically use this variety without case endings and occasionally add case markers to words. However, the optionality of case marking in speech is in the pedagogical literature often described as an optional *omission* of case marking, rather than an optional *addition* (e.g. Ryding 2005: 166; Schulz et al. 2000:47). Such descriptions enforce — even if this is not the intention — the notion that case endings are present in the language in some sort of underlying or initial state, and that speakers omit them in the process of speaking.[113] On the basis of this assumption, educational practices are reproduced whereby learners are expected to actively control the system of case endings, since they "are there" even if not pronounced, and only then to take liberties with unmarked forms. It is therefore important to explicitly describe case endings in speech as an optional addition.[114]

The view of case endings as an optional addition implies that the case system is best introduced relatively late in a course on Standard Arabic when the basics of syntax and morphology have been covered and students have become familiar with Arabic without case endings. This route is taken by several recently published textbooks of Arabic as a second language (e.g. Wightwick and Gaafar 2005; Brustad et al. 2011). This approach may clash, however, with a lingering reluctance to instruct students to use unmarked forms that are traditionally regarded as incorrect. This includes unmarked word forms with an enclitic pronoun (e.g. *balad-ak* 'your (m.) country', *balad-u* 'his country', *balad-nā* 'our country'), the pausal form for indefinite triptote accusative, and the *-īn* and *-ayn* forms for the nominative of the sound m.pl. and the dual. These unmarked forms are typically regarded as dialectal and teachers may therefore hesitate to include them in language courses on Standard Arabic. It should be

[113] This is paralleled in descriptions of case markers in writing. Ryding (2005:166), for instance, states that "the case-ending system consists primarily of short, word-final vowels, *which are invisible* in conventional written Arabic texts" (emphasis in original), implying the presence of case endings even when they are not overtly marked. Compare this the more neutral description in Badawi et al. (2004:21) that case endings "seldom appear" in written texts.

[114] See Wightwick and Gaafar (2005:8) for an example of how this may be done.

remembered that they are regarded as dialectal on the basis of traditional, codified grammar, not on observed usage of speech. If the speakers in the corpus of this study are taken as examples of how Standard Arabic is normally spoken, one is perfectly justified to use these unmarked forms in Spoken Standard Arabic. Case markers are, in these situations, used at higher rates than are other forms of case endings, with a some speakers in the corpus using them very consistently, as described in patterns *d* and *e* above, but such case markings are nevertheless clearly optional for most speakers. The average speaker in the corpus uses case endings in the sound m.pl., the dual, and in words with an enclitic pronoun at rates of less than 50%. The marked and the unmarked forms of these words are in other words both representative of Spoken Standard Arabic.

A second implication of the results presented in this study is that in order for students to speak Standard Arabic in a way typical of highly educated, native speakers competent in the standard variety, there are words that ought not to be marked for case. These patterns of unmarked case, listed in 11.2 above, represent covert norms in the language community and it is not clear (with the exception of *b*) to what extent speakers are actually consciously aware of them. They could nevertheless be explicitly taught to students. The endings affected by these patterns are not used in extemporaneous speech and are not marked in writing. They are a feature of formal read speech and recitation but not of Spoken Standard Arabic. Most users of the language will only ever need passive knowledge of them to be fully functional in the language, even in very formal registers. Indeed, drilling students in an active use of these endings may not only be poor use of class time, but may in fact be counterproductive since it instills habits that need to be unlearned in order to speak Standard Arabic the way Standard Arabic is typically spoken. This point is particularly important for second language learners of Arabic who may not have the dialectal forms available to fall back on and thus risk establishing a habit of using the case marked forms of these words.

11.4 Suggestion for further research

This corpus study has investigated the Arabic case system as attested in Spoken Standard Arabic in natural speech production situations. It was shown how case marking in this variety of Arabic is structurally patterned in ways that are not immediately apparent from a surface examination. This study has been largely explorative, and studies replicating and testing the generalizability of its findings would be welcome, preferably on a greater amount of material in order to provide improved accuracy and reliability. Such studies could

include speech from other types of discourses and could control more accurately for speakers' educational background. Other programs and accompanying transcriptions published by Al Jazeera or other news outlets could be used. Al Jazeera publishes programs on religious and cultural topics that are similar to the news interviews analyzed here, both in terms of program format and the way they are published with video and accompanying transcription. The procedures of data collection and analysis developed in this study could be reproduced on materials from these programs. An analysis of case marking in the speech of guests appearing in the religious program *aš-Šarīᶜa wa-l-ḥayāt* would be particularly interesting in this regard because of the commonly held view that religious scholars have a better command of Standard Arabic than do other intellectuals. Such a study would have to carefully control for citations and paraphrases of the Quran and other religious texts.

The conclusions drawn in this study also raise a number of questions that have been addressed only partially or in passing here, some of which would require other methods of research in order to be properly addressed. There are two areas of research that were only partly investigated due to limitations in the annotation scheme and both could be more thoroughly investigated by extending the level of detail and scope of annotation in the present corpus. The first such area is that of case marking in adverbs. Adverbs were treated in the coding scheme as a very broad category. Extending the coding scheme with subdivisions of adverbs into kinds, defined either semantically, lexically, or syntactically, may reveal patterns of case marking within this group and possibly reveal a clearer division between productively derived and lexicalized adverbs. In practice, this would entail going through the adverb tags in the corpus and substituting them for tags of a more detailed scheme. The second area is pause, which for practical reasons was given a narrow definition in this study. Since other forms of pauses were not annotated, they could not be systematically investigated. Defining pause more broadly, for example as periods of silence at phrasal boundaries and encoding this in the corpus, is likely to reveal other patterns and give a more accurate description of how the pause system operates in Spoken Standard Arabic.

The corpus of this study could also be utilized to investigate patterns as to how mood, the other half of the *iᶜrāb* system, is marked in speech. Mood in Standard Arabic is very similar to morphological case in its formal characteristics and a parallel coding scheme could be developed for annotation of mood in the corpus. This could potentially reveal patterns of mood marking similar to those found for case and would together with this study provide a comprehensive description of the use of *iᶜrāb* endings in Spoken Standard Arabic.

The structured nature of case marking in Spoken Standard Arabic shown in this study raises the question as to what degree speakers are consciously aware of these structures. Case marking is particularly interesting in this regard, seeing as how metalinguistic awareness of it is fostered in formal education but according to principles that are undermined by linguistic practices and conventions of speech, as well as by the fact that awareness of case is not necessary for effective reading. The degree of awareness of these covert norms of case marking could be investigated by interviewing experienced users Spoken Standard Arabic, for example those featured in this study.

Another question raised by the findings in this study, and one that has only been addressed in passing here, is how deviations from the covert conventions of case marking would be perceived by speakers. One hypothesis is that such deviations, for example extensive case marking on words with a definite article, would be perceived as odd and result in negative judgments of the speaker. Some indications that this would be the case are found in Parkinson (1991:59), where informants in a matched guise test ranked the speaker as 'smarter' when case was only occasionally marked than when it was consistently marked. Similar tests could be done on hypotheses informed by the present study, measuring reactions to speech where case is marked partially with distributions of case markers corresponding to, and deviating from, the patterns described above.

Finally, the lack of research on educational practices in the Arabic speaking world is a hindrance to a deeper understanding of how norms of case marking, both overt and covert, are formed and reproduced. Several of the patterns of case marking found in this study contradict aspects of codified grammar and it is not clear how and if this is mitigated in educational practices. There is a growing body of research on the issue of diglossia and its effects on education and literacy (e.g. Alrabaa 1986; Amara 1995; Maamouri 1998; Khamis-Dakwar 2005; Saiegh-Haddad and Joshi 2014), but there are very few studies addressing more concretely the practical methods of teaching Arabic grammar in the Arab world. Section 4.3 gave a brief description of one aspect of Arabic grammar teaching that have yet to be thoroughly mapped out in future studies. Such studies could shed light on how ideas about case marking are formulated and transmitted to students, and whether covert norms of case marking such as those listed above are implicitly tolerated in the classroom or actively corrected. Studies of teaching materials from this perspective, of which Uhlmann (2012) is an interesting example, could also provide valuable insights.

The Arabic system of case marking is heavily loaded with language ideology and notions of correctness and it is used in different ways in different contexts.

The conventions of case marking in writing are different from those of reading aloud, and these are different from case marking produced in spontaneous speech. Systematic linguistic investigation is needed to map out these practices of case marking in order formulate an empirically based description of the role of case marking in various forms of Standard Arabic. The results of the research in this dissertation can provide one important step in this direction and provide a foundation for further research.

APPENDIX
Speaker information

The table on the following three pages provides a summary of information on the interviewees whose speech is analyzed in this study. This information has been gathered from a variety of sources, including bibliographical encyclopedias (Goldschmidt 2000; Mattar 2005; Fischbach 2008), news articles, reports, official home pages, and Wikipedia entries. It was not possible to find complete information for all speakers. Missing information is marked with question marks. For Muḥammad Kayāli, a former Syrian head of the military police and a politician, the only source of information is the presenter's introductory remarks preceding the interview and the interview itself. For each speaker the URL to the recording and transcript on Al Jazeera's homepage is provided.

NAME	NAT.	Y. OF BIRTH	UNIVERSITY DEGREE	OCCUPATION/ COMMENTS	BROADCAST DATE
Ḥātim ᶜAbd al-Qādir	Pa.	???	???	Minister of Jerusalem affairs	15 Oct 2010
http://www.aljazeera.net/programs/pages/3f8ea662-67a1-46bd-b5e0-ef16c605d75c					
Aḥmad Abū l-Majd	Eg.	1930	JSD, Cairo Uni.	Jurist; liberal Islamist thinker; former minister of youth, minister of media	19 Oct 2011
http://www.aljazeera.net/programs/pages/c3303e1d-7384-440c-a427-2ba4ccc30ebf					
Ṣāʾib ᶜArīqat	Pa.	1955	PHD, peace studies, Bradford Uni.	Minister of negotiations	28 Jan 2010
http://www.aljazeera.net/programs/pages/4efe140e-235d-4c58-bc3f-c2009e3c29e2					
Riyāḍ al-Asᶜad	Sy.	1961	???	Founder and commander of the Free Syrian Army; former general, Syrain army	19 Nov 2011
http://www.aljazeera.net/programs/pages/762654be-64b3-4690-893e-d85be23b688b					
Muḥammad Badīᶜ	Eg.	1945	Dr. of veterinary medicine, Uni. of Zagazig	Supreme guide of the Muslim Brotherhood	22 Jan 2010
http://www.aljazeera.net/programs/pages/2b4b24d2-ffod-43b2-ae55-9c7d4e47475a					
Muḥammad al-Barādiᶜī	Eg.	1942	JSD, international law, New York Uni.	Founder of National Association for Change; Nobel Peace Price laureate, formerly director genral of IAEA	10 Jan 2010
http://www.aljazeera.net/programs/pages/28a5593e-7b6d-445a-b093-80946b40b68f					
Salām Fayyāḍ	Pa.	1952	PHD, economics, Uni. of Texas at Austin	Prime minister	27 Aug 2010
http://www.aljazeera.net/programs/pages/7debf96c-9817-4a4e-8753-0df97f1a2c81					

Continued on next page.

Continued from previous page.

NAME	NAT.	Y. OF BIRTH	UNIVERSITY DEGREE	OCCUPATION/ COMMENTS	BROADCAST DATE
Burhān Ġalyūn	Sy.	1945	PhD, philosophy and humanities, Uni. of Paris VIII; PhD, political sociology, Sorbonne	Head of the Syrian National Council	26 Oct 2011
http://www.aljazeera.net/programs/pages/7ca5ee47-ed50-4b97-a7be-a5041892451b1					
[c]Alī d-Dīn Hilāl	Eg.	1944	PhD, political science, McGill Uni.	Media Secretary, Naitional Democratic Party; formerly minister of youth	26 Nov 2010
http://www.aljazeera.net/programs/pages/9bfe20ed-eb0d-4320-a0e1-b6dacdde2446					
Munīb al-Miṣrī	Pa.	1934	MA, geology, Uni. of Texas	Businessman; former minister of public works, Jordan; former MP, Fath	3 Sep 2010
http://www.aljazeera.net/programs/pages/03ecef22-abd3-4c42-b688-d854ba174b14					
Muḥammad Kayālī	Sy.	???	???, Egyptain Military Academy; PhD, political economy	Former head of military police	13 Jul 2011
http://www.aljazeera.net/programs/pages/958f6650-5173-4c0d-b313-e5f1005960f4					
Walīd al-Mu[c]allim	Sy.	1941	BA, economics, Cairo Uni.	Minister of foreign affairs	8 Oct 2010
http://www.aljazeera.net/programs/pages/d16aacb9-e166-44dc-91b7-d9b2a2bb9c6a					
Muḥammad Mursī	Eg.	1951	PhD, materials science, Uni. of Southern California	Chairman of the Freedom and Justice Party; affiliated with the Muslim Brotherhood	23 Jul 2011
http://www.aljazeera.net/programs/pages/d070e1f5-28d8-4f07-86fe-cce0b30a4607					
Farūq al-Qaddūmī a.k.a. Abū Luṭf	Pa.	1930	???, economics and political Science, Cairo Uni.	Cofounder of Fath	30 Jun. 2010
http://www.aljazeera.net/programs/pages/973d2988-2bbd-4064-99c2-6d15cb5e98e9					

Continued on next page.

Continued from previous page.

NAME	NAT.	Y. OF BIRTH	UNIVERSITY DEGREE	OCCUPATION/ COMMENTS	BROADCAST DATE
Ramaḍān Šallaḥ	Pa.	1958	PhD, economics, Durham Uni.	General secretary of Islamic Jihad	7 Apr 2010
http://www.aljazeera.net/programs/pages/47c02444-9d95-43cf-bcb3-9b8b20d39e20					
Ṭayyib Tayzīnī	Sy.	1938	PhD, philosophy Humboldt Uni.	Proffesor of Philosophy, Damascus Uni., Marxist thinker.	3 Aug 2011
http://www.aljazeera.net/programs/pages/43755484-db64-4af1-91a9-69c6eac510eb					
Jamāl al-Xuḍarī	Pa.	1955	BA, electric engineering, Cairo Uni.	independent MP; Head of the Popular Committee Against Siege of Gaza	19 Feb 2010
http://www.aljazeera.net/programs/pages/f4f1df06-f3a8-48e9-897c-57389963fa7f					

APPENDIX
Transcript example

The following three pages contain an excerpt from one of the corpus files representing the first three minutes of the interview with Muhammad Badīᶜ. The corpus text is based on transcripts retrieved from Al Jazeera's homepage. These were transliterated, formatted to CHAT standards, and edited in order to represent the speech as uttered in the recording more accurately (see Chapter 6). The first seven lines of each file begins with `@` and contain meta data. Each line of speech begins with a code representing the role of the interviewer (`*INT:`) and interviewee (informant, `*INF:`). Nouns and adjectives in the interviewee's speech are followed by morphosyntactic annotation unless marked up for exclusion, for example with a following `[F]` for 'formulaic expression'. Each of these lines end with a time stamp giving the start and end time of the utterance in milliseconds. Each line of transcription is followed by a duplicate line beginning with `%ara:` giving the utterance stripped of annotation and in Arabic script, which is intended to provide a readily readable version of the text.

B

```
@Begin
@Languages:     ara
@Participants: *INF Muḥammad_Badīᶜ Informant, *INT
               Ḥusayn_ᶜAbd_al-Ġanī Interviewer
@ID:           ara||INF||male|Egyptian||Informant||
@ID:           ara||INT||male|||Interviewer||
@Date:         10-JAN-2010
@Media:        100122muhammad_badi video
*INT: áhl´ bkm mš´hdyn´ . ´xt´rt jm´ᶜp ´líxw´n ´l#mslmyn
      ´l#dktwr mḥmd bdyᶜ lykwn mršd´ ṯ´mn´ lh´ áhl´ bkm
      fḍylp ´l#dktwr . •0_26711•
%ara: أهلا بكم مشاهدينا . اختارت جماعة الإخوان المسلمين الدكتور محمد بديع ليكون مرشدا ثامنا لها، أهلا بكم فضيلة الدكتور .
*INF: <áhl´ bkm>[F] . •26711_27575•
%ara: أهلا بكم .
*INT: dktwr bdyᶜ ántm x´rjwn mn ázmp kbyrp d´xlyp bᶜd msálp
      ´l#´ntx´b´t tṯ´yrt fyh´ ´l#´th´m´t wwṣlt ldý ´l#bᶜḍ ánh
      k´n hn´k tzwyr fy ´l#´ntx´b´t án hn´k tl´ᶜb´ fy hyỳp
      ´l#n´xbyn án hn´k ´ġṭṣ´b´ mn qbl mktb ´l#írš´d lslṭ´t
      mjls ´l#šwrý án hn´k qy´d´t lh´ wznh´ ´l#t´ryxy mṯl
      ´l#dktwr ábw ´l#ftwḥ áw lh´ wzn tnẓymy mṯl ´l#dktwr
      ḥbyb qd xrjt fy hḍ´ ´l#mwḍwᶜ . •27575_57008•
%ara: دكتور بديع أنتم خارجون من أزمة كبيرة داخلية بعد مسألة الانتخابات تطايرت فيها الاتهامات ووصلت لدى البعض أنه كان هناك تزوير في الانتخابات أن هناك تلاعبا في هيئة الناخبين أن هناك اغتصابا من قبل مكتب الإرشاد لسلطات مجلس الشورى أن هناك قيادات لها وزنها التاريخي مثل الدكتور أبو الفتوح أو لها وزن تنظيمي مثل الدكتور حبيب قد خرجت في هذا الموضوع .
*INF: <bsm ´l#lh w´l#ḥmd lillh w´l#ṣl´p w´l#sl´m ᶜlý rswl
      ´l#lh ṣlý ´l#lh ᶜlyh wslm>[F] . •57008_61301•
%ara: بسم الله والحمد لله والصلاة والسلام على رسول الله صلى الله عليه وسلم .
*INF: áwl-´aa$HED:ACC/ADV:IND:TRI:MAR áškr ´l#jzyrp[P] .
      •61301_63530•
%ara: أولا أشكر الجزيرة .
*INF: wáškrk y´ <áx ḥsyn>[P] +/. . •63530_65282•
%ara: وأشكرك يا أخ حسين +/. .
*INT: nḥn nškrk ᶜlý hḍh ´l#frṣp ´l#áwlý . •65282_66768•
%ara: نحن نشكرك على هذه الفرصة الأولى .
```

```
*INF: ´l#ḥqyqp$HED:ACC/ADV:ART:TRI:UNM ᶜndm´ áṯḥdṯ ᶜn
      ´l#ftrp$HED:GEN/PRE:ART:TRI:UNM
      ´l#m´ḍyp$ATT:GEN/PRE:ART:TRI:UNM wm´ ḥdṯ fyh´ áqwl ánh´
      k´nt tjrbp$HED:ACC/KAN:IND:TRI:UNM
      ṯryp$ATT:ACC/KAN:IND:TRI:UNM . •66768_80021•
%ara:       الحقيقة عندما أتحدث عن الفترة الماضية وما حدث فيها أقول أنها كانت تجربة ثرية .
*INF: nᶜm qd ykwn fyh´ ml´ḥẓ´t$HED:NOM/VSO:IND:SFP:UNM .
      •80021_84191•
%ara:                                                نعم قد يكون فيها ملاحظات .
*INF: nᶜm qd nḥt´j ílý mr´jᶜp$HED:GEN/PRE:IND:TRI:UNM
      l#bᶜḍ-a$HED:GEN/PRE:CON:TRI:AMB
      ´l#mw´qf$HED:GEN/ABS:ART:DIP:UNM
      w#bᶜḍ-i$HED:GEN/PRE:CON:TRI:AMB
      ´l#íjr´ˀ´t$HED:GEN/ABS:ART:SFP:UNM . •84191_90105•
%ara:                                  نعم قد نحتاج إلى مراجعة لبعض المواقف وبعض الإجراءات .
*INF: lknh´ slymp$HED:NOM/COM:IND:TRI:UNM
      mẏp$ATT:ACC/ADV:IND:TRI:UNM
      b´l#m´ẏp$HED:GEN/PRE:IND:TRI:UNM bᶜd
      ´stš´rp-i$HED:GEN/PRE:CON:TRI:AMB
      ´l#mstš´r-yn$HED:GEN/ABS:ART:SMP:AMB
      ´l#q´nwny-yn$ATT:GEN/PRE:ART:SMP:AMB
      ll#jm´ᶜp$HED:GEN/PRE:ART:TRI:UNM
      w#ᶜrḍ-i$HED:GEN/PRE:CON:TRI:AMB m´ tm ᶜlyhm .
      •90105_99443•
%ara:       لكنها سليمة مئة بالمائة بعد استشارة المستشارين القانونيين للجماعة وعرض ما تم عليهم .
*INF: wáqrw´ jmyᶜ-uhm$HED:NOM/VSO:PRO:TRI:MAR bán
      ´l#íjr´ˀ´t$HED:ACC/INN:ART:SFP:UNM
      slymp$HED:NOM/COM:IND:TRI:UNM
      tm´m-´aa$HED:ACC/ADV:IND:TRI:MAR . •99423_104541•
%ara:                                  وأقروا جميعهم بأن الإجراءات سليمة تماما .
*INF: ám´ ´l#´th´m´t-i$HED:NOM/TOP:ART:SFP:AMB ´lty tk´l lḥḍh
      ´l#jm´ᶜp$HED:GEN/PRE:ART:TRI:UNM fḥḍh lyst
      jdydp$ATT:ACC/KAN:IND:TRI:UNM . •104541_109621•
%ara:                                  أما الاتهامات التي تكال لهذه الجماعة فهذه ليست جديدة .
*INF: lkn ᶜlý mn ydᶜy án yqdm
      ´l#dlyl$HED:ACC/OBJ:ART:TRI:UNM . •109621_113829•
%ara:                                  لكن على من يدعي أن يقدم الدليل .
*INF: nḥn <bfḍl ´l#lh sbḥ´nh wtᶜ´lý>[F] lsn´
      jm´ᶜp$HED:ACC/KAN:IND:TRI:UNM
      sy´syp$ATT:ACC/KAN:IND:TRI:UNM km´ ytṣwr
      ´l#bᶜḍ$HED:NOM/VSO:ART:TRI:UNM . •113829_121631•
%ara:                  نحن بفضل الله سبحانه وتعالى لسنا جماعة سياسية كما يتصور البعض .
```

*INF: nḥn jm´ᶜp$HED:NOM/COM:IND:TRI:UNM tḥml hḏ´
´l#ísl´m$HED:ACC/OBJ:ART:TRI:UNM
b#šmwl-ih$HED:GEN/PRE:PRO:TRI:MAR . •121631_125141•
%ara: نحن جماعة تحمل هذا الإسلام بشموله .
*INF: whḏ´ ´l#ísl´m$HED:NOM/TOP:ART:TRI:UNM fyh
´l#xlq$HED:NOM/TOP:ART:TRI:UNM
w´l#qym$HED:NOM/TOP:ART:TRI:UNM tsbq
´l#ws´ẏl$HED:ACC/OBJ:ART:DIP:UNM
w´l#āly´t$HED:ACC/OBJ:ART:SFP:UNM . •125141_133520•
%ara: وهذا الإسلام فيه الخلق والقيم تسبق الوسائل والآليات .
*INF: fḥḏh ´l#āly´t-i$HED:NOM/TOP:ART:SFP:AMB ´lty njryh´ mn
´ntx´b´t$HED:GEN/PRE:IND:SFP:UNM
w#lw´ẏḥ$HED:GEN/PRE:IND:DIP:UNM wm´ ytbᶜh´ mn
āly´t-i$HED:GEN/PRE:CON:SFP:AMB
´l#tᶜ´ml$HED:GEN/ABS:ART:TRI:UNM
q´ẏmp$HED:NOM/COM:IND:TRI:UNM ᶜlý
qym$HED:GEN/PRE:IND:TRI:UNM
w#áxl´qy´t$HED:GEN/PRE:IND:SFP:UNM
ísl´myp$ATT:GEN/PRE:IND:TRI:UNM
ášylp$ATT:GEN/PRE:IND:TRI:UNM . •133520_149207•
%ara: فهذه الآليات التي نجريها من انتخابات ولوائح وما يتبعها من آليات التعامل قائمة على قيم
وأخلاقيات إسلامية أصيلة .
*INF: yjmᶜn´ fyh´ áwl-´aa$HED:ACC/ADV:IND:TRI:MAR
´l#xwf$HED:NOM/VSO:ART:TRI:XXX mn
´l#lh$HED:GEN/PRE:ART:TRI:UNM <sbḥ´nh wtᶜ´lý>[F]
w#mr´qbp-ihi$HED:NOM/VSO:PRO:TRI:HYP qbl
mr´qbp$HED:GEN/PRE:CON:TRI:UNM
áy-i$HED:GEN/ABS:CON:TRI:AMB
jhp$HED:GEN/ABS:IND:TRI:UNM . •149207_155613•
%ara: يجمعنا فيها أولا الخوف من الله سبحانه وتعالى ومراقبته قبل مراقبة أي جهة .
[…]

Bibliography

The definite article *al-* is ignored in the alphabetical ordering of names transcribed from Arabic script.

Abdel-Massih, Ernest T. (1975). *An Introduction to Egyptian Arabic*. Ann Arbor: MPublishing University of Michigan Library.

Aboul-Fetouh, Hilmi M. (1969). *A Morphological Study of Egyptian Colloquial Arabic*. The Hague: Mouton.

Abu-Absi, Samir (1986). "The Modernization of Arabic: Problems and Prospects". *Anthropological Linguistics* 28.3, pp. 337–348.

Abu-Haidar, Farida (2006). "Baghdad Arabic". In: *Encyclopedia of Arabic Language and Linguistics*. Ed. C. H. M. Versteegh. Vol. 1. Leiden: Brill, pp. 222–231.

al-ᶜAdnānī, Muḥammad (1999). *Muᶜjam al-aġlāṭ al-luġawiyya al-muᶜāṣira*. Beirut: Maktabat Lubnān.

Agius, Dionisius A. (1991). "Precedence of VOS over VSO in Modern Standard Arabic". In: *Semitic Studies. In Honor of Wolf Leslau on the Occasion of his eighty-fifth Birthday*. Vol. 1. Weisbaden: Otto Harrassowitz, pp. 39–55.

Albirini, Abdulkafi (2011). "The Sociolinguistic Functions of Codeswitching Between Standard Arabic and Dialectal Arabic". *Language in Society* 40.5, pp. 537–562.

Alrabaa, Sami (1986). "Diglossia in the Classroom: The Arabic Case". *Anthropological Linguistics* 28.1, pp. 73–79.

Al-Sulaiti, Latifa and Eric Steven Atwell (2006). "The Design of a Corpus of Contemporary Arabic". *International Journal of Corpus Linguistics* 11.2, pp. 135–171.

Al-Thubaity, Abdulmohsen, Marwa Khan, Manal Al-Mazrua, and Maram Al-Mousa (2013). "New Language Resources for Arabic: Corpus Containing More Than Two Million Words and a Corpus Processing Tool". In: *2013 International Conference on Asian Language Processing (IALP)*, pp. 67–70.

Altoma, Salih J. (1974). "Language Education in Arab Countries and the Role of the Academies". In: *Advances in Language Planning*. Ed. Joshua A. Fishman. The Hague: Mouton, pp. 279–313.

Al-Wer, Enam (1997). "Arabic Between Reality and Ideology". *International Journal of Applied Linguistics* 7.2, pp. 251–265.

Amara, Muhammad Hasan (1995). "Arabic Diglossia in the Classroom: Assumptions and Reality". *Israel Oriental Studies* 15, pp. 131–142.

Ammon, Ulrich (1986). "Explikation der Begriffe 'Standardvarietät' und 'Standardsprache' auf normtheoretischer Grundlage". In: *Sprachlicher Substandard*. Ed. Günter Holtus and Edgar Radke. Tübingen: Niemeyer, pp. 1–63.

Arberry, A. J. (1983). *The Koran Interpreted*. Oxford: Oxford University Press.

Ayoub, Georgine (2006). "Laḥn". In: *Encyclopedia of Arabic Language and Linguistics*. Ed. C. H. M. Versteegh. Vol. 2. Leiden: Brill, pp. 628–634.

Badawi, El Said M. (2006). "Arabic for Non-Native Speakers in the 21st Century: A Shopping List". In: *Handbook for Arabic Language Teaching Professionals in the 21st Century*. Ed. Kassem M. Wahba, Zeinab A. Taha, and Liz England. Mahwah, N.J: Lawrence Erlbaum Associates, pp. ix–xxiv.

Badawi, ElSaid M. (1985). "Educated Spoken Arabic: A Problem in Teaching Arabic as a Foreign Language". In: *Scientific and Humanistic Dimensions of Language: Festschrift for Robert Lado on the Occasion of His 70th Birthday on May 31, 1985*. Ed. Kurt R. Jankowsky. Amsterdam: Benjamins, pp. 15–22.

Badawī, M. M. (1993). *A Short History of Modern Arabic Literature*. Oxford: Clarendon.

Badawi, El-Said M., Michael G. Carter, and Adrian Gully (2004). *Modern Written Arabic: A Comprehensive Grammar*. London: Routledge.

Badawi, El-Said M. and Martin Hinds (1986). *A Dictionary of Egyptian Arabic*. Beirut: Librairie du Liban.

Badawī, as-Saᶜīd Muḥammad (1973). *Mustawayāt al-ᶜarabiyya al-muᶜāṣira fī Miṣr*. Cairo: Dār al-maᶜārif bi-Miṣr.

Bailey, Richard W. (1991). *Images of English: A Cultural History of the Language*. Ann Arbor: University of Michigan Press.

Bartsch, Renate (1989). "A Normtheoretical Approach to Functional and Status Types of Language". In: *Status and Function of Languages and Language Varieties*. Ed. Ulrich Ammon. Berlin, New York: Walter de Gruyter, pp. 190–215.

Bassiouney, Reem (2003). "Theories of Code Switching in the Light of Empirical Data from Egypt". *Perspectives on Arabic Linguistics* 15. Ed. Dilworth B. Parkinson and Samira Farwaneh, pp. 19–40.

Bates D. Maechler, M., B. Bolker, and S. Walker (2014). *lme4: Linear Mixed-Effects Models using Eigen and S4*.

Bateson, Mary Catherine (1967). *Arabic Language Handbook*. Washington, D.C: Center for Applied Linguistics.

Bayṭār, ᶜĀṣim (2007). *an-Naḥw wa-ṣ-ṣarf*. Damascus: Jāmiᶜat Dimašq.

Beeston, A. F. L (1970). *The Arabic Language Today*. London: Hutchinson.

Behnstedt, Peter and Manfred Woidich (2013). "Dialectology". In: *The Oxford Handbook of Arabic Linguistics*. Ed. Jonathan Owens. Oxford: Oxford University Press, pp. 300–325.

Belnap, R. Kirk and Brian Bishop (2003). "Arabic Personal Correspondence: A Window on Change in Progress?" *International Journal of the Sociology of Language* 163.1, pp. 9–25.

Bennani, F. et al. (2003). *Arab Human Development Report 2003: Building a Knowledge Society*. United Nations Development Program.

Blake, Barry J. (1994). *Case*. Cambridge: Cambridge Univ. Press.

Blanc, Haim (1964). "Stylistic Variation in Spoken Arabic: A Sample of Interdialectal Educated Conversation". In: *Contributions to Arabic Linguistics*. Ed. Charles A. Ferguson. Cambridge, Mass.: Harvard University Press, pp. 79–156.

Blau, J. (1973). "Remarks on Some Syntactic Trends in Modern Standard Arabic". *Israel Oriental Studies* 3, pp. 172–231.

— (1976). "Some Additional Observations on Syntactic Trends in Modern Standard Arabic". *Israel Oriental Studies* 6, pp. 158–190.

Blommaert, Jan (2006). "Language Policy and National Identity". In: *An Introduction to Language Policy: Theory and Method*. Ed. Thomas Ricento. Malden Mass: Blackwell Publ, pp. 238–254.

Bohas, Georges, Jean-Patrick Guillaume, and Djamel Eddine Kouloughli (1990). *The Arabic Linguistic Tradition*. London: Routledge.

Brustad, Kristen, Mahmoud Al-Batal, and Abbas al-Tunsi (2011). *al-Kitaab fī taᶜallum al-ᶜarabiyya: A Textbook for Beginning Arabic*. Washington, D.C.: Georgetown.

Bucholtz, Mary (2001). "The Whiteness of Nerds: Superstandard English and Racial Markedness". *Journal of Linguistic Anthropology* 11.1, pp. 84–100.

Buckwalter, Tim and Dilworth B. Parkinson (2011). *A Frequency Dictionary of Arabic: Core Vocabulary for Learners*. Routledge.

Cantarino, Vicente (1974). *Syntax of modern Arabic Prose*. Bloomington; London: Indiana University Press.

Carter, Michael G. (1985). "When did the Arabic Word NAḤW First Come to Denote Grammar?" *Language & Communication* 5.4, pp. 265–272.

Carter, Ronald (1999). "Standard Grammars, Spoken Grammars: Some Educational Implications". In: *Standard English: The Widening Debate*. Ed. Tony Bex and Richard J. Watts. London: Routledge, pp. 149–166.

Chejne, Anwar G. (1969). *The Arabic Language: Its Role in History*. Minneapolis: University of Minnesota Press.

Clayman, Steven and John Heritage (2002). *The News Interview: Journalists and Public Figures on the Air*. Cambridge: Cambridge University Press.

Cowell, Mark W. (2005). *A Reference Grammar of Syrian Arabic*. Washington, D.C.: Georgetown University Press.

Curzan, Anne (2014). *Fixing English: Prescriptivism and Language History*. New York: Cambridge University Press.

ad-Daḥdāḥ, Anṭwān (1992). *Muᶜjam qawāᶜid al-luġa l-ᶜarabiyya fī jadāwil wa-lawḥāt*. Bayrut: Maktabat Lubnān.

Davies, Eirlys, Abdelali Bentahila, and Jonathan Owens (2013). "Codeswitching and Related Issues Involving Arabic". In: *The Oxford Handbook of Arabic Linguistics*. Ed. Jonathan Owens. Oxford: Oxford University Press, pp. 326–348.

Ḍayf, Šawqī (2000). "Bayn al-fuṣḥā wa-l-ᶜāmmiyya". *Majallat majmaᶜ al-luġa al-ᶜarabiyya* 89.1, pp. 35–49.

Dévényi, Kinga (2006). "Iᶜrāb". In: *Encyclopedia of Arabic Language and Linguistics*. Ed. C. H. M. Versteegh. Vol. 2. Leiden: Brill, pp. 401–406.

Diem, Werner (1974). *Hochsprache und Dialekt im Arabischen: Untersuchungen zur heutigen arabischen Zweisprachigkeit*. Wiesbaden: Steiner.

Eid, Mushira (1982). "The Non-Randomness of Diglossic Variation in Arabic". *Glossa* 16.1, pp. 54–84.

— (1988). "Principles for Code-Switching Between Standard and Egyptian Arabic". *Al-Arabiyya* 21.1, pp. 51–79.

Elgibali, Alaa (1985). *Towards a Sociolinguistic Analysis of Language Variation in Arabic: Cairene and Kuwaiti Dialects*. Pittsburgh: Unpublished Ph.D. thesis.

El-Hassan, S. A. (1977). "Educated Spoken Arabic in Egypt and the Levant: A Critical Review of Diglossia and Related Concepts". *Archivum Linguisticum* 8, pp. 112–132.

— (1978). "Variation in the Demonstrative System in Educated Spoken Arabic". *Archivum Linguisticum* 9, pp. 32–57.

Ferguson, Charles A. (1959). "Diglossia". *Word* 15, pp. 325–340.

— (1996). "Epilogue: Diglossia Revisited". In: *Understanding Arabic: Essays in Contemporary Arabic Linguistics in Honor of El-Said Badawi*. Ed. Alaa Elgibali. Cairo: American University in Cairo Press, pp. 49–67.

— (1997). "Standardization as a Form of Language Spread". In: *Structuralist Studies in Arabic Linguistics: Charles A. Ferguson's Papers, 1954-1994*. Ed. R. Kirk Belnap and Niloofar Haeri. Leiden, Netherlands: Brill, pp. 69–80.

Fischbach, Michael R., ed. (2008). *Biographical Encyclopedia of the Modern Middle East and North Africa*. Detroit, Mich.

Fischer, Wolfdietrich (2002). *A Grammar of Classical Arabic*. New Haven, Conn.: Yale University Press.

Fisher, John H. (1996). *The Emergence of Standard English*. Lexington: Univ. Press of Kentucky.

Fishman, Joshua A. (1967). "Bilingualism With and Without Diglossia; Diglossia With and Without Bilingualism". *Journal of Social Issues* 23.2, pp. 29–38.

Furayḥa, Anīs (1955). *Naḥw ʿarabiyya muyassara*. Beirut: Dār aṭ-ṭaqāfa.

Goldschmidt, Arthur (2000). *Biographical Dictionary of Modern Egypt*. Boulder, CO: Lynne Rienner.

Griffiths, Roger (1991). "Pausological Research in an L2 Context: A Rationale, and Review of Selected Studies". *Applied Linguistics* 12.4, pp. 345–364.

Gully, Adrian (1993). "The Changing Face of Modern Written Arabic: An Update". *Al-Arabiyya* 26, pp. 19–59.

Haeri, Niloofar (1996). "Conceptualizing Heterogeneity in Arabic". *Égypte/-Monde arabe* 3.27. Ed. Madiha Doss and Catherine Miller, pp. 301–315.

— (1997). "The Reproduction of Symbolic Capital: Language, State, and Class in Egypt". *Current Anthropology* 38.5, pp. 795–816.

— (2000). "Form and Ideology: Arabic Sociolinguistics and Beyond". *Annual Review of Anthropology* 29.1, pp. 61–87.

— (2003). *Sacred Language, Ordinary People: Dilemmas of Culture and Politics in Egypt*. New York: Palgrave Macmillan.

Halliday, M. A. K. (1994). *Spoken and Written Language*. Victoria: Deakin Univ. Press.

Hammond, Andrew (2007). *Popular Culture in the Arab World: Arts, Politics, and the Media*. Cairo: American University in Cairo Press.

Harrell, Richard S. (1964). "A Linguistic Analysis of Egyptian Radio Arabic". In: *Contributions to Arabic LInguistics*. Ed. Charles A. Ferguson. Cambridge, Mass., pp. 1–77.

Hary, Benjamin (1996). "The Importance of the Language Continuum in Arabic Multiglossia". In: *Understanding Arabic: Essays in Contemporary Arabic Linguistics in Honor of El-Said Badawi*. Ed. Alaa Elgibali and El-Said M. Badawi. Cairo: American University in Cairo Press, pp. 69–90.

Haugen, Einar (1966). "Dialect, Language, Nation". *American Anthropologist* 68.4, pp. 922–935.

Havránek, Bohuslav (1982). "Zum Problem der Norm in der heutigen Sprachwissenschaft und Sprachkultur". In: *Soziolinguistik: Ansätze zur soziolinguistischen Theoriebildung*. Ed. Hugo Steger. Darmstadt: Wissenschaftliche Buchgesellschaft, pp. 289–296.

Hawkins, Peter (1983). "Diglossia Revisited". *Language Sciences* 5.1, pp. 1–20.

Haywood, John Alfred and Hayim Musa Nahmad (1993). *A New Arabic Grammar of the Written Language*. London: Lund Humphries.

Heritage, John (1985). "Analyzing News Interviews: Aspects of the Production of Talk for an Overhearing Audience". In: *Handbook of Discourse Analysis*. Ed. Teun Adrianus van Dijk. Vol. 3. London: Academic Press, pp. 95–117.

Holes, Clive (1993). "The Uses of Variation: A Study of the Political Speeches of Gamal Abd Al-Nasir". 5. Ed. Mushira Eid and Clive Holes, pp. 13–45.

— (2004). *Modern Arabic: Structures, Functions, and Varieties*. Washington, D.C: Georgetown University Press.

— (2013). "Orality, Culture, and Language". In: *The Oxford Handbook of Arabic Linguistics*. Ed. Jonathan Owens. Oxford: Oxford University Press, pp. 281–299.

Householder, F. W. (1962). "Greek Diglossia". In: *Georgetown Monographs*. Vol. 15, pp. 109–132.

Hudson, Alan (1992). "Diglossia: A Bibliographic Review". *Language in Society* 21.4, pp. 611–674.

— (2002). "Outline of a Theory of Diglossia". *International Journal of the Sociology of Language*, pp. 1–48.

Hunt, Kellogg W. (1970). "Syntactic Maturity in Schoolchildren and Adults". *Monographs of the Society for Research in Child Development* 35.1, pp. iii–67.

Hymes, Dell (1967). "Models of the Interaction of Language and Social Setting". *Journal of Social Issues* 23.2, pp. 8–28.

Ibn ᶜAqīl, ᶜAbd al-Lāh (2005). *Šarḥ Ibn ᶜAqīl ᶜalā Alfiyyat Ibn Mālik*. Ed. Anṭūniyūs Buṭrus. Tarablus: al-Muᵓassasa al-ḥadīṯa li-l-kitāb.

Ibn Hišām, Jamāl ad-Dīn (1990). *Šarḥ qatr an-nadā wa-ball aṣ-ṣadā*. Beirut: Maktabat Lubnān.

al-Ibrahīm, Muḥammad aṭ-Ṭayyib (2007). *Iᶜrāb al-qurᵓān al-karīm*. Beirut: Dār an-nafāᵓis.

Ibrahim, Muhammad H. (1983). "Linguistic Distance and Literacy in Arabic". *Journal of Pragmatics* 7.5, pp. 507–515.

— (1986). "Standard and Prestige Language: A Problem in Arabic Sociolinguistics". *Anthropological Linguistics* 28.1, pp. 115–126.

— (1989). "Communicating in Arabic: Problems and Prospects". In: *Language Adaptation*. Ed. Florian Coulmas. Cambridge: Cambridge UP, pp. 39–59.

Irvine, Judith T. (2001). "Formality and Informality in Communicative Events". In: *Linguistic Anthropology: A Reader*. Ed. Alessandro Duranti. Malden, Mass: Blackwell Publishers, pp. 189–207.

Jaeger, T. Florian (2008). "Categorical Data Analysis: Away From ANOVAs (Transformation or Not) and towards Logit Mixed Models". *Journal of Memory and Language* 59.4, pp. 434–446.

Jastrow, Otto (1980). "Das mesopotamische Arabisch". In: *Handbuch der arabischen Dialekte*. Ed. Wolfdietrich Fischer and Otto Jastrow. Wiesbaden: Harrassowitz, pp. 140–154.

Jedlička, Alois (1982). "Theorie der Literatursprache". In: *Grundlagen der Sprachkultur: Beiträge der Prager Linguistik zur Sprachtheorie und Sprachpflege*. Ed. Jürgen Scharnhorst and Erika Ising. Berlin: Akademie-Verlag, pp. 40–91.

Kaye, Alan S. (1972). "Remarks on Diglossia in Arabic: Well-Defined vs. Ill-Defined". *Linguistics* 10.81, pp. 32–48.

— (1991). "The Hamzat al-Waṣl in Contemporary Modern Standard Arabic". *Journal of the American Oriental Society* 111.3, pp. 572–574.

Khaldieh, Salim A. (2001). "The Relationship Between Knowledge of *Iʿraab*, Lexical Knowledge, and Reading Comprehension of Nonnative Readers of Arabic". *The Modern Language Journal* 85.3, pp. 416–431.

Khamis-Dakwar, Reem (2005). "Children's Attitudes towards the Diglossic Situation in Arabic and its Impact on Learning". *Languages, Communities, and Education*, pp. 75–86.

Killean, Carolyn G. (1984). "The Development of Western Grammars of Arabic". *Journal of Near Eastern Studies* 43.3, pp. 223–230.

Klimiuk, Maciej (2013). *Phonetics and Phonology of Damascus Arabic*. Warsaw: Katedra Arabistyki i Islamistyki, Uniw. Waszawski.

Kloss, Heinz (1967). "'Abstand Languages' and 'Ausbau Languages'". *Anthropological Linguistics* 9.7, pp. 29–41.

Kormos, Judit and Mariann Dénes (2004). "Exploring Measures and Perceptions of Fluency in the Speech of Second Language Learners". *System* 32.2, pp. 145–164.

Kuznetsova, Alexandra, Per Bruun Brockhoff, and Rune Haubo Bojesen Christensen (2014). *lmerTest: Tests for Random and Fixed Effects for Linear Mixed Effect Models*.

Labov, William (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Lane, Edward William (1863). *An Arabic-English Lexicon*. London: Williams and Norgate.

Lehtonen, Jaakko (1982). "Non-Verbal Aspects of Impromptu Speech". In: *Impromptue Speech: A Symposium*. Ed. Nils Erik Enkvist. Åbo: Åbo Akademi, pp. 33–45.

Lippi-Green, Ros (1997). *English with an Accent: Language, Ideology, and Discrimination in the United States*. London: Routledge.

Lynch, Marc (2006). *Voices of the New Arab Public: Iraq, al-Jazeera, and Middle East Politics Today*. New York: Columbia University Press.

Maamouri, Mohamed (1998). *Language Education and Human Development: Arabic Diglossia and Its Impact on the Quality of Education in the Arab Region.* Philadelphia: International Literacy Institute.

— (2006). "Literacy". In: *Encyclopedia of Arabic Language and Linguistics*. Ed. C. H. M. Versteegh. Vol. 3. Leiden: Brill, pp. 74–80.

al-Maᶜarrī, Šawqī (2005). *Iᶜrāb al-kalimāt wa-t-tarākīb al-muškila fī al-asālīb al-ᶜarabiyya*. Damascus: Dar al-ḥāriṭ.

— (2007). *Iᶜrāb miʾat bayt muxtāra*. Damascus: Dar al-ḥāriṭ.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

Magidow, Alexander (2012). "Explaining Inconsistent Case Marking in Spoken Formal Arabic". *Zeitschrift für arabische Linguistik* 55, pp. 62–99.

Maṭar, ᶜAbd al-ᶜAzīz (1966). *Laḥn al-ᶜāmma fī ḍawʾ ad-dirāsāt al-luġawiyya al-ḥadīṯa*. Cairo: al-Muʾassasa al-miṣrī al-ᶜāmma li-t-taʾlīf wa-n-našr.

Mattar, Philip (2005). *Encyclopedia of the Palestinians*. New York: Facts on File.

Mazraani, Nathalie (1997). *Aspects of Language Variation in Arabic Political Speech-making*. Richmond: Curzon Press.

McCarus, Ernest N. (2006). "Modern Standard Arabic". In: *Encyclopedia of Arabic Language and Linguistics*. Ed. C. H. M. Versteegh. Vol. 3. Leiden: Brill, pp. 238–262.

Meiseles, Gustav (1977). "Restitution of 'Word-Endings' in Modern Literary Arabic". *Israel Oriental Studies* 7, pp. 173–195.

— (1979). "Informal written Arabic: A Preliminary Evaluation of Data". *Israel Oriental Studies* 6, pp. 73–313.

— (1980). "Educated Spoken Arabic and the Arabic Language Continuum". *Archivum Linguisticum* 11, pp. 118–148.

Mejdell, Gunvor (2006). *Mixed Styles in Spoken Arabic in Egypt: Somewhere Between Order and Chaos*. Boston: Brill.

— (2008a). "Is Modern Fusha a 'Standard' Language?" In: *Linguistics in an Age of Globalization: Perspectives on Arabic Language and Teaching*. Ed. Zaynab Ibrahim and Sanaa A. M. Makhlouf. Cairo; New York: American University in Cairo Press, pp. 41–52.

— (2008b). "What is Happening to *lughatunā l-gamīla*? Recent Media Representations and Social Practice in Egypt". *Journal of Arabic and Islamic Studies* 8, pp. 108–124.

Miles, Hugh (2005). *Al Jazeera: How Arab TV News Challenged the World*. London: Abacus.

Milroy, James (2007). "The Ideology of the Standard Language". In: *The Routledge Companion to Sociolinguistics*. Ed. Carmen Llamas, Louise Mullany, and Peter Stockwell. London: Routledge, pp. 133–139.

Milroy, James and Lesley Milroy (1991). *Authority in Language: Investigating Language Prescription and Standardisation*. London: Routledge.

Mitchell, Terence Frederick (1978a). *An Introduction to Egyptian Colloquial Arabic*. Oxford: Oxford University Press.

— (1978b). "Educated Spoken Arabic in Egypt and the Levant, with Special Reference to Participle and Tense". *Journal of Linguistics* 14, pp. 227–258.

— (1980). "Dimensions of Style in a Grammar of Educated Spoken Arabic". *Archivum Linguisticum* 11, pp. 89–106.

— (1985). "Sociolinguistic and Stylistic Dimensions of the Educated Spoken Arabic of Egypt and the Levant". In: *Language Standards and Their Codification*. Ed. Douglas. J. Woods. Exeter: Exeter University Press, pp. 42–57.

— (1986). "What is Educated Spoken Arabic?" *International Journal of the Sociology of Language* 61, pp. 7–32.

— (1990). *Pronouncing Arabic*. Oxford: Clarendon.

Mitchell, Terence Frederick and S. A. El-Hassan (1994). *Modality, Mood and Aspect in Spoken Arabic: With Special Reference to Egypt and the Levant*. London: Kegan Paul International.

Montgomery, Martin (2008). "The Discourse of the Broadcast News Interview". *Journalism Studies* 9.2, pp. 260–277.

Mughazy, Mustafa (2005). "Reading Despite Ambiguity: The Role of Metacognitive Strategies in Reading Arabic Authentic Texts". *al-ᶜArabiyya* 38/39, pp. 57–74.

Mūsā, Salāma (1964). *al-Balāġa l-ᶜaṣriyya wal-luġa l-ᶜarabiyya*. Cairo: al-Maṭbaᶜa al-ᶜaṣriyya.

Myers-Scotton, Carol (1993). *Duelling Languages: Grammatical Structure in Codeswitching*. Oxford: Clarendon Press.

Nelson, Kristina (2001). *The Art of Reciting the Qurʾan*. Cairo; New York: The American University in Cairo Press.

Newman L., Daniel (2013). "The Arabic Literary Language: The *Nahḍa* (and Beyond)". In: *The Oxford Handbook of Arabic Linguistics*. Ed. Jonathan Owens. Oxford: Oxford University Press, pp. 472–494.

Nielsen, Helle Lycke (2009). "Second Language Teaching". In: *Encyclopedia of Arabic Language and Linguistics*. Ed. C. H. M. Versteegh. Vol. 4. Leiden: Brill, pp. 515–520.

Owens, Jonathan (1988). *The Foundations of Grammar: An Introduction to Medieval Arabic Grammatical Theory*. John Benjamins Publishing Company.
— (2001). "Arabic Sociolinguistics". *Arabica* 48.4, pp. 419–469.
Owens, Jonathan and Raslan Bani-Yasin (1987). "The Lexical Basis of Variation in Jordanian Arabic". *Linguistics* 25.4.
— (1991). "Spoken Arabic and Language Mixture". *Bulletin d'études orientales*, pp. 17–31.
Palfreyman, David and Muhamed al Khalil (2003). "'A Funky Language for Teenzz to Use': Representing Gulf Arabic in Instant Messaging". *Journal of Computer-Mediated Communication* 9.1.
Palva, Heikki (1969). "Notes on Classicalization in Modern Colloquial Arabic". *Studia Orientalia* 40.3, pp. 1–41.
Parkinson, Dilworth B. (1985). *Constructing the Social Context of Communication: Terms of Address in Egyptian Arabic*. Berlin, New York, Amsterdam: Mouton de Gruyter Berlin.
— (1990). "Orthographic Variation in Modern Standard Arabic: The Case of the Hamza". *Perspectives on Arabic Linguistics* 2. Ed. Mushira Eid and John McCarthy, pp. 269–295.
— (1991). "Searching for Modern Fuṣḥā: Real-life Formal Arabic". *Al-Arabiyya* 24, pp. 31–64.
— (1993). "Knowing Standard Arabic: Testing Egyptians' MSA Abilities". *Perspectives on Arabic Linguistics* 5. Ed. Mushira Eid and Clive Holes, pp. 47–73.
— (1994a). "Speaking Fuṣḥā in Cairo: The Role of the Ending Vowels". In: *Arabic Sociolinguistics: Issues and Perspectives*. Ed. Yasir Suleiman. Surrey: Curzon Press, pp. 179–211.
— (1994b). "Testing Native Speakers: Implications for Teaching Arabic to Non-Native Speakers". In: *Investigating Arabic: Linguistic, Pedagogic and Literary Studies in Honor of Ernest N. McCarus*. Ed. Ernest N. McCarus, Raji M. Rammuny, and Dilworth B. Parkinson. Columbus, Ohio: Greyden Press, pp. 191–203.
— (1996). "Variability in Standard Arabic Grammar Skills". In: *Understanding Arabic: Essays in Contemporary Arabic Linguistics in Honor of El-Said Badawi*. Ed. Alaa Elgibali. Cairo: American University in Cairo Press, pp. 91–101.
— (2003). "Verbal Features in Oral Fuṣḥa Performances". *International Journal of the Sociology of Language* 163.1, pp. 27–41.
— (2006). *Using Arabic Synonyms*. Cambridge University Press.

Peter, Ann M. (1985). "Language Segmentation: Operating Principles for the Perception and Analysis of Language". In: *The Crosslinguistic Study of Language Acquisition*. Ed. Dan Isaac Slobin. Vol. 2. Hillsdale, New Jersey: L. Erlbaum Associates, pp. 1029–1064.

Procházka, Stephan (2006). "Prepositions". In: *Encyclopedia of Arabic Language and Linguistics*. Ed. C. H. M. Versteegh. Vol. 3. Leiden: Brill, pp. 699–703.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reichmuth, Philipp (2006). "Transcription". In: *Encyclopedia of Arabic Language and Linguistics*. Ed. C. H. M. Versteegh. Vol. 4. Leiden: Brill, pp. 146–156.

Ricento, Thomas (2006). "Language Policy: Theory and Practice—An introduction". In: *An Introduction to Language Policy: Theory and Method*. Ed. Thomas Ricento. Malden Mass: Blackwell Publ, pp. 10–23.

Ryding, Karin C. (1991). "Proficiency Despite Diglossia: A New Approach for Arabic". *Modern Language Journal* 75.2, pp. 212–18.

— (2005). *A Reference Grammar of Modern Standard Arabic*. Cambridge, UK; New York: Cambridge University Press.

— (2006). "Teaching Arabic in the United States". In: *Handbook for Arabic Language Teaching Professionals in the 21st Century*. Ed. Kassem M. Wahba, Zeinab A. Taha, and Liz England. Mahwah, N.J: Lawrence Erlbaum Associates, pp. 13–20.

Saiegh-Haddad, Elinor and R. Malatesha Joshi, eds. (2014). *Handbook of Arabic Literacy: Insights and Perspectives*. 9. New York: Springer.

Saiegh-Haddad, Elinor and Bernard Spolsky (2014). "Acquiring Literacy in a Diglossic Context: Problems and Prospects". In: *Handbook of Arabic Literacy: Insights and perspectives*. Ed. Elinor Saiegh-Haddad and R. Malatesha Joshi. New York: Springer, pp. 225–240.

Salib, Maurice Boulos (1980). *Spoken Literary Arabic: Oral Approximation of Literary Arabic in Egyptian Formal Discourse*. Berkeley: Unpublished PhD thesis.

Samuel-Azran, Tal (2013). "Al-Jazeera, Qatar, and New Tactics in State-Sponsored Media Diplomacy". *American Behavioral Scientist* 20, pp. 1–19.

Schmidt, Richard W. and Carol F. McCreary (1977). "Standard and Super-Standard English: Recognition and Use of Prescriptive Rules by Native and Non□Native Speakers". *TESOL Quarterly* 11.4, pp. 415–429.

Schulz, David Eugene (1981). *Diglossia and Variation in Formal Spoken Arabic in Egypt*. Ann Arbor: Unpublished Ph.D. thesis.

Schulz, Eckehard, Günther Krahl, and Wolfgang Reuschel (2000). *Standard Arabic: An Elementary-Intermediate Course*. Cambridge: Cambridge University Press.

Shahin, Kimary N. (2006). "Palestinian Arabic". In: *Encyclopedia of Arabic language and linguistics*. Ed. C. H. M. Versteegh. Vol. 3. Leiden: Brill, pp. 526–538.

Silverstein, M. (1998). "Contemporary Transformations of Local Linguistic Communities". *Annual Review of Anthropology* 27.1, pp. 401–426.

aš-Širbīnī, Muḥammad Ibn Aḥmad (1981). *Arab Linguistics: An Introductory Classical Text with Translation and Notes*. Ed. Michael G. Carter.

Snow, Don (2013). "Revisiting Ferguson's Defining Cases of Diglossia". *Journal of Multilingual and Multicultural Development* 34.1, pp. 61–76.

Stetkevych, Jaroslav (2006). *The Modern Arabic Literary Language: Lexical and Stylistic Developments*. Washington D.C: Georgetown University Press.

Suleiman, Yasir (1996). "The Simplification of Arabic Grammar and the Problematic Nature of the Sources". *Journal of Semitic Studies* 41.1, pp. 99–119.

— (2011). "Ideology, Grammar-Making and the Standardization of Arabic". In: *In the Shadow of Arabic: The Centrality of Language to Arabic Culture*. Ed. Bilal Orfali. Brill.

— (2013). "Arabic Folk Linguistics". In: *The Oxford Handbook of Arabic Linguistics*. Ed. Jonathan Owens. Oxford: Oxford University Press, pp. 255–289.

Trudgill, Peter (1975). *Accent, Dialect and the School*. London: Arnold.

Uhlmann, Allon J. (2012). "Arabs and Arabic Grammar Instruction in Israeli Universities: Alterity, Alienation and Dislocation". *Middle East Critique* 21.1, pp. 101–116.

Van Mol, Mark (2003). *Variation in Modern Standard Arabic in Radio News Broadcasts: A Synchronic Descriptive Investigation into the Use of Complementary Particles*. Leuven; Dudley, Mass: Peeters and Departement Oostere Studies.

Versteegh, Kees (1983). "Arabic Grammar and Corruption of Speech". *Al-Abhath* 31, pp. 139–60.

— (2001). *The Arabic Language*. Edinburgh: Edinburgh University Press.

Wagner, Daniel A. and Abdelhamid Lotfi (1983). "Learning to Read by 'Rote'". *International Journal of the Sociology of Language* 1983.42, pp. 111–121.

Wahba, Kassem M. (2006). "First Language Teaching". In: *Encyclopedia of Arabic Language and Linguistics*. Ed. C. H. M. Versteegh. Vol. 2. Leiden: Brill, pp. 104–13.

Walters, Keith (2003). "Fergie's Prescience: The Changing Nature of Diglossia in Tunisia". *International Journal of the Sociology of Language* 163.1, pp. 77–109.

Wehr, Hans (1994). *A Dictionary of Modern Written Arabic (Arabic—English)*. Ed. J. Milton Cowan. New York: Spoken Language Services.

Wightwick, Jane and Mahmoud Gaafar (2005). *Mastering Arabic Grammar*. Basingstoke: Palgrave Macmillan.

Woidich, Manfred (2006). "Cairo Arabic". In: *Encyclopedia of Arabic Language and Linguistics*. Ed. C. H. M. Versteegh. Vol. 1. Leiden: Brill, pp. 323–333.

Wright, William (2011). *A Grammar of the Arabic Language*. New York: Cosimo.

Xalīfa, ᶜAbd al-Karīm (2000). "al-ᶜArabiyya al-fuṣḥā bayna lahjātihā w-ᶜāmmiyyātihā al-muxtalifa". *Majallat majmaᶜ al-luġa al-ᶜarabiyya* 89.1, pp. 119–125.

al-Xayr, Buṭayna and Laylā Muḥammad, eds. (2011). *al-Luġa al-ᶜarabiyya: marḥalat at-taᶜlīm al-ʾasāsī, aṣ-ṣaff as-sābiᶜ, al-faṣl aṭ-ṭānī*. al-Muʾassasa al-ᶜāmma li-ṭ-ṭibāᶜa.

Younes, Munther Abdullatif (2015). *The Integrated Approach to Arabic Instruction*. New York: Routledge.

az-Zubaydī, Muḥammad Ibn al-Ḥasan (1954). *Ṭabaqāt an-naḥwīyīn wa-l-luġawīyīn*. Ed. Muḥammad Abū l-Faḍl Ibrāhīm. Cairo.

General Index

Page numbers in italics refer to the primary discussion or definition of the entry. For abbreviations, see *List of Abbreviations* on page *xi*.

Author Index

The definite article *al-* is ignored in the alphabetical ordering of names transcribed from Arabic script.

Sudia Orientalia Lundensia. Nova Series

1. Lena Ambjörn, *Qusta ibn Luqa On Numbness. A Book on Numbness, Its Kinds, Causes and Treatment According to the Opinion of Galen and Hippocrates. Edition, Translation and Commentary,* 2000.

2. Maria Persson, *Sentential Object Complements in Modern Standard Arabic,* 2002.

3. Per Å Bengtsson, *Translation Techniques in Two Syro-Arabic Versions of Ruth,* 2003.