LUND
UNIVERSITY

# Formulaic language in academic writing

– Investigating the use of prefabs in linguistic abstracts

Phoebe Jönsson

ENGK01

Degree project in English Linguistics

Spring 2017

Supervisor: Henrik Gyllstad

## Abstract

In this study, the use of prefabricated language in academic abstracts is investigated, as well as the functions of prefabricated expressions (prefabs) in this type of text. By analysing ten abstracts taken from peer-reviewed articles in the field of Linguistics, the amount of prefabs is calculated, and their pragmatic roles investigated. The results are then compared to previous research within the subject. The study shows that prefabs are used slightly less in abstracts than in general English - though not to a statistically significant degree - and that about 49 % of the language in these abstracts is prefabricated. The prefabs used in the texts are also shorter than in general English, to a statistically significant degree. The reason for this could be the density of information needed in an academic abstract. This study helps to fill a gap in research in the subject of formulaic language in academic writing and offers suggestions for future research within the area.

**Key words:** formulaic language, academic writing, prefabricated language

# Table of contents



# Index of tables

# 1. Introduction

As Wray (2002) describes, there are chunks of language that appear repeatedly in daily life. When telling a story, greeting a friend or writing, these expressions, collectively known as *formulaic language* (FL), regularly recur. According to Schmitt (2010), previous research shows that rather than existing as single pieces, language functions largely in units of multiple words. One of the most well known principles within the field of FL is Sinclair's idiom principle, which states that rather than being constructed piece by piece in small sections (referred to as the open choice principle), language is constructed by putting together larger pieces of already formulated phrases. In natural speech and writing, a mix of these two principles is generally used (Sinclair, 1991). Erman and Warren (2000) write that this is also described in a lecture by Bolinger (1976), in which it is claimed that it would be more natural for the brain to store larger complex items rather than smaller ones, due to the extensive memory storage of the human brain.

Common examples of FL are idioms. These are noticeable since they are non-compositional, meaning that the unit cannot be understood from the component words alone. However, idioms are only one of several types of FL. Due to the large amount of different types, there are also a large amount of different terms used to describe the phenomenon. According to Wray (2002), there are over 50 different terms used in the literature. Examples of these terms are *chunks, prefabricated routines, collocations, holophrases* and *ready-made utterances* (Schmitt, 2010). In this study, the main terms that will be used are Erman and Warren's (2000) *prefab* as well as the more general *formulaic language*.

Formulaic language is common in both written and spoken language. According to a corpus study by Erman and Warren (2000), about 52 – 58% of language is formulaic, while Foster (2001) claims that the figure is 32%. The differing results are due to the researchers using different methods in their studies, which will be discussed extensively in this essay. FL also exists in several different languages, including Russian, French, Swedish, Hebrew, Chinese and many more (Schmitt, 2010).

One of the purposes of using FL is reducing processing effort (Wray, 2002). Wray (2002) writes that by using FL while speaking, it is possible to focus on other activities at the same time, like the ideas of the conversation or another unrelated task. However, there are several

other reasons for using prefabs. Chunks of FL make certain jobs easier, and are used avidly in for example sports commentaries, auctions and weather forecasts. They can also be used to "mark a style", meaning that a writer uses certain expressions in order to keep the text genre-appropriate. Prefabs are also used to lessen repetition and add variation in written text as well as to organise the text by marking discourse (Wray, 2002).

The purpose of this essay is to investigate how frequently FL is used in academic abstracts. The reason why abstracts are chosen for this study is because of their density of information – abstracts need to convey complex ideas in few words. Therefore it is interesting to investigate the role of prefabs in this type of text. This essay may help to fill a gap in the research of prefabs in academic text, since there have not been many studies made within this particular subject. The essay could also help shed light on how and why prefabricated language is used in academic English today. Abstracts taken from ten different peer-reviewed academic articles will be analysed in order to distinguish between formulaic expressions and "open choice" expressions according to Sinclair's idiom principle. The types of prefabricated expressions and their roles will also be analysed. As Simpson-Vlach and Ellis (2010) state, "there are so many [prefabricated] constructions that there is ever a need for prioritization and organization".

# 2. Background

## 2. 1 Theoretical background

As previously stated, formulaic language has many different uses (Wray, 2002). For example, it is used in order to make text more comprehensive and less repetitive, and to make spoken language more fluent. In short, it can be described as solutions to different linguistic problems. This, Wray (2002) states, can be further expanded upon by examining the reasons why these problems should be solved, that reason being the promotion of the speaker's interests. These interests include having easy access to information, expressing information in a fluent manner, being listened to in a serious way, having emotional and physical needs met, being provided with information and being perceived as an important member of the speaker's group.

Claude Shannon, one of the earliest scholars in the subject of the processing of FL, investigated the probabilistic nature of language in 1948. By stringing together letters in an order that followed the usual rules of distribution, he created words that were not real but still readable. Following this, sentences using words put together in the same way were created. This created an interest in the probabilistic way sentences can be created from psycholinguists. The study of these sentence approximations, however, was mostly put aside during the Chomskian era and did not reappear until the 1990s (Shaoul & Westbury, 2011).

Erman and Warren (2000) refer to Bolinger (1976) in their article, who has argued that it would be more natural for the human brain to store larger complex items rather than smaller ones. The basis for this comes from the field of information and learning theory in psychology. Expanding on these theories, Diesel (2007), using the ideas of learning theorists Anderson and Newell (1990), describes three psychological processes that occur in the brain when a language user is exposed to FL frequently. Increased frequency in the use of a word sequence strengthens "linguistic representations", meaning that a person's memory is reinforced by repeated exposure to the language chunk. This in turn means that the language chunk is more often activated and interpreted when using language. Increased frequency also "strengthens expectation". This means that the language user expects one word to follow another due to their frequent use together. This is referred to in previous research as *transitional probability* (Swinney & Cutler, 1979). Increased frequency also causes "automatization of language chunks", meaning that words that are frequently used together

create processing units and more compressed chunks of language. An example of this is compressing *going to* (two words very frequently used together) into *gonna* (Shaoul & Westbury, 2011).

According to Sinclair (1991), language can be constructed by using the idiom principle or the open-choice principle. The open-choice principle is connected to the so-called "slot-and-filler" principle. This means that language is the result of a large amount of small choices made by the language user. At each slot in a sentence, there are many options for the language user to choose between in order to create a sentence. This construction process is very complex, since every node point on a tree structure can symbolise an open choice. This principle is also the basis for most grammars (Sinclair, 1991). The idiom principle, however, states that the open-choice principle does not put enough restraints on the choices available when creating language. Instead of choosing from small, individual segments of language, the user chooses from a large number of pre-constructed phrases. These phrases often allow for variation in their semantic and lexical construction. Variation in the word order is also allowed – for example, *it is not in an academic's nature* is just as valid as *it is not in the nature of an academic.* According to Sinclair (1991, p. 114), this variation implies that users know "which way to interpret each portion of an utterance".

The two principles described are opposed, but still work together in natural language. Sinclair (1991, p. 114) writes that in normal text, switches between the two principles occur "when there is good reason". Furthermore, the two principles work better in different circumstances. For example, the open-choice principle would be more applicable in a legal statement, while a mix would be more appropriate in poetry (Sinclair, 1991).

There have been many different definitions of what a prefab is made by different scholars, such as Sinclair (1991) and Schmitt (2010). Schmitt (2010, p.117) describes them as "multiple word phraseological units". According to Erman and Warren (2000, p. 31), a prefab is defined is defined as:

> "(a) combination of at least two words favoured by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization."

Erman and Warren (2000) also write that at least one word of the prefab should not be able to be replaced by a synonym without changing the meaning, function or idiomaticity of the expression. As an example, *good friends* vs. *nice friends* is used. If *nice* is used instead of *good,* the meaning and idiomaticity of the phrase is lost. Additionally, some syntactic variability that is normally possible cannot be used in prefabs. Examples of these variations used by Erman and Warren (2000) are negation (*I guess* cannot be negated into **I don't guess*), loss of auxiliary (*It will do* versus **It does*) and reversed order.

In their 2011 article, Shaoul & Westbury quote Biber (1999), who writes that in order for a lexical bundle (described as an extended collocation, very similar to this study's prefab) to be considered formulaic, it must occur "at least ten times per million in a corpus for sequences between two and four words long, and at least five times per million for longer sequences" (Shaoul & Westbury, 2011, p. 2). It is stated that these rules are arbitrary, but are conventionally used today in the identification process of prefabs.

### 2.2 Sorting and categorisation of prefabs in previous research

Defining a chunk of prefabricated language is very difficult. According to Schmitt (2010), this is due to their large extent and diversity. However, a few different methods have been devised. Schmitt (2010) brings up the acquisition approach, which can be used to identify prefabs in spoken language. This approach is based on the psycholinguistic theory that FL is stored as fully formed chunks in the brain. This can be supported by the fact that while speaking a piece of prefabricated language, there should be no hesitation or pause in the middle of it. Additionally, Shaoul & Westbury (2011) write that eye movement tracking has proved that reading speed increases when reading familiar prefabs. However, this method of identification is restricted due to the large amount of variation in language (dialectal, verbal, lexical etc.). Another method of identifying prefabs is the phraseological approach, in which prefabs are defined by their "transparency and substitutability". This means that words are often restricted by their usual collocations. However, this method relies on analysis made by humans, and is therefore labour intensive and time-consuming. The most common method of identifying FL, Schmitt (2010) writes, is by corpus analysis. By identifying recurring sequences in corpora according to frequency, prefabs can be studied automatically. There are problems with this method as well, one of them being that the most frequent combination of one word is usually with a function word. This means that words will appear together only

because they are frequently used. Actual collocations can also be overlooked due to their infrequency (Schmitt, 2010).

Wray (2002) gives further suggestions on how to identify prefabs. The least scientific, but most commonly used method, she writes, is intuition. Formulaic language is closely connected to idiomaticity. When identifying an idiom, members of a language community can often tell when an expression "sounds right" and is considered as being a unit. The same, to an extent, is true in the case of prefabs. Though intuition is a common process of investigation, there is opposition to the use of it in research of FL, specifically by Chomsky (1965) and Sinclair (1991). Sinclair argues that intuition can only be used to gather information on intuition, not information about language. Instead, the only reliable source used in the identification process is corpora (Wray, 2002). However, Wray (2002) also states: "researchers, as members of their speech community, often are the self-appointed arbiters of what is idiomatic of formulaic in their data". In this essay, the intuition-based approach is supported by rechecking of the data against a corpus and the database Google Books, which is what determines whether an expression is a prefab or not.

Erman and Warren (2000) list further difficulties in identifying prefabs. One of the two main reasons, they state, is the fact that a prefab might not be seen as such to all members of the language community. Some are far more well known that others, while some are rarely used. The other difficulty is due to how easily overlooked prefabs can be. Only on further examination may they appear to be noncompositional or idiomatic. It is further stated that identifying "all and every" prefab in a text is, in practice, impossible.

Prefabs can be sorted and organised in many different ways. Some of these are genre-specific, while others are more general. While some researchers such as Hyland (2008), argue that the number of lexical bundles commonly used in general academic writing is not very large and therefore cannot be compiled (instead, Hyland (2008) argues that prefabs are often subject-specific and should be organised as such), others such as Simpson-Vlach and Ellis (2010) claim that there are in fact many prefabs that are specific to the overarching genre of academic writing. These, in turn, can be organised after their function in the academic text.

In their attempt to create a list of prefabs commonly used in Academic English, Simpson-Vlach and Ellis (2010) have organised the prefabs in three categories. This was done using a

corpus of 2.1 million words of academic text and speech. In this study, two-word phrases were excluded, and the authors focused on phrases between three and five words long. The phrases were sorted according to subject and function, which created three major categories of phrases used in academic speech and writing:

- **Group A: Referential expressions.** As the largest of the groups, this section includes the subcategories *specification of attributes*, *identification and focus*, *contrast and comparison*, *deictics and locatives*, and finally, *vagueness markers*. Members of the category "specification of attributes" are used as tangible or intangible framing devices. Examples of these are *based on the, in the context of, in such a way* (intangible) and *the amount of, the size of, the level of* (tangible). Quantity specifications are also included in this subcategory. The identity and focus-category include typically expository phrases. Examples of these are *such as the, referred to as* and *an example of.* The subcategory "contrast and comparison" include, simply, prefabs using comparison phrases, such as *as opposed to.* Prefabs in the category of "deictics and locatives" refer to physical locations or "spatial reference points in the discourse". Examples of these are *The University of Michigan* or *The United Kingdom.* The final category, vagueness markers, includes a very small amount of prefabs, though they are still important I Academic English. The most prominent example of vagueness markers is *and so on* (Simpson-Vlach & Ellis, 2010).
- **Group B: Stance expressions.** This group includes *hedges* and *epistemic stances,* as well as prefabs referring to *obligation and directive, ability and possibility* and finally, *evaluation.* The category of hedges includes prefabs that can have several different uses, but where the hedging is the most important function. Examples of hedging prefabs are *to some extent, there may be* and might *want to*. Epistemic stances deal with claims of knowledge or expressions of certainty or uncertainty. This group include prefabs such as *let's assume that*. The three final categories include prefabs like *tell me what* (obligation and directive), *going to be able* (ability and possibility) and *is consistent with* (evaluation) (Simpson-Vlach & Ellis, 2010).
- **Group C: Discourse organising expressions.** This group is split into four subcategories, the fist being *metadiscourse and textual reference.* This subcategory includes prefabs that are highly genre-specific to Academic English, such as *in the next section*. Another subcategory is *topic introduction and focus*, which are used to frame a large section of text. An example of this is *take a look at.* The third

subcategory, *topic elaboration,* are used to signal further elaboration of a topic. Examples are *it turns out, what happens is, in order to* and *as a result.* The final subcategory, *discourse markers,* include connectives (*as well as, in other words*) and interactive devices and formulas (*thank you very much*) (Simpson-Vlach & Ellis, 2010).

Others have had similar ways of organising prefabs used specifically in academic English. Hyland (2007) used the categories *research-oriented, text-oriented* and *participation-oriented* in his research. Additionally, Dontcheva-Navratilova (2012) states that this kind of categorisation may be misleading, since one prefab can have several different functions at the same time, as well as have different functions in different contexts.

When it comes to academic writing, previous research by Biber et al. (2001) suggests that writers of academic text improve their skills in a certain developmental sequence. According to Biber et al. (2001), academic writers start by using language similar to spoken language, and acquire the density of information and complex phrasal features needed in academic language over time (Staples, Egbert, Biber & Gray, 2016). Biber and Barbieri's (2007) study of lexical bundles in university settings shows that while prefabs are used in all types of university registers, they are most common in spoken academic English (here referring to lectures et cetera) and are more rare in written academic language (textbooks and similar course-related works). This study was made using a sub-component of the TOEFL 2000 Spoken and Written Academic Language corpus, with samples of spoken and written language taken from six major disciplines (Business, Engineering, Natural Science, Social Science, Humanities and Education), and focuses on lexical bundles, which may have some differences in structure and use that the prefabs described in the essay. Biber and Barbieri (2007) state that:

> "The extent to which a speaker or writer relies on lexical bundles is strongly influenced by their communicative purposes, in addition to general spoken/written differences. The explanation for the infrequent use of lexical bundles in the academic written registers (textbooks and academic prose) apparently lies in the restricted communicative goals of those registers—focused on informational communication—rather than the written mode per se." (p. 273).

In written university registers, referential expressions are the most common, while discourse organising expressions and stance expressions are more common in spoken university registers (Biber and Barbieri, 2007) When examining four-word prefabs used in MA theses written within the field of humanities, Dontcheva-Navratilova (2012) found that referential expressions are the most commonly used types of prefabs, while stance expressions are the least common.

In his Master's thesis, Vincent (2009) investigates the extent of which prefabs occur in a short text, using Sinclair's idiom principle as the foundation. The chosen short text is taken from a university textbook, and was chosen due to the fact that it was deemed typical of its genre. The prefabs in Vincent's study were identified with intuition, and then checked against the Bank of English corpus of 450 million tokens to determine whether they could be labelled prefabs or not. The result of the Master's thesis states that 92.3 % of the short text is "covered by the idiom principle" (p. 14), which is what will be investigated in this essay.

The mini-corpus used as a basis for the present study will be made up of a selection of ten abstracts, selected from published and peer-reviewed articles. The reason why abstracts are chosen for this study is because they are very short and dense in information. In order to summarise a research project in a small amount of words, the authors need to choose their words carefully and deliberately. It is therefore interesting to see what role prefabricated expressions play in this careful selection – are they still used in a text where the focus is not on prose but on clarity and information density? Additionally, there is a shortage of previous studies on the subject of prefabs in abstracts, and this essay will help fill a gap in the research within this subject. The research questions for this study are therefore as follows:

1. To what extent are prefabs used in academic text, specifically in abstracts taken from articles in the subject of linguistics?
2. What types of prefabs are used in the abstracts – what are their pragmatic roles?

# 3. Method and materials

## 3.1 Material collection and analysis

The abstracts for this study were taken from articles published in the *Journal of Linguistics* (Cambridge University Press, Cambridge Core) during the years 2015 and 2016. The articles were chosen by random selection – every third article published in an issue was selected until ten texts were chosen. The selection starts with the last issue of 2016 and moves backwards in the dates of publications to make sure that the articles are as relevant and recent as possible. The chosen journal is published three times per year and contain between four and six articles, along with other reviews and editorial material. After the selection, the prefabs were analysed using methods from Erman and Warren's (2000) study.

The chosen abstracts were taken from the articles listed below. The full abstracts along with their analysis and dates of publication can be found in the Appendix.

| Title | Author(s) |
|---|---|
| Lexical Bundles and Disciplinary Variation in University Students' Writing: Mapping the Territories | Durrant, Philip |
| Middle-passive voice in Albanian and Greek | Manzini, Rita M., Savoia, Leonardo M. |
| A constructional account of the 'optional' quotative marking on Japanese mimetics | Akita, Kimi., Usuki, Takeshi |
| Clitics: Separating syntax and prosody | Lowe, John J. |
| Canonical Gender | Corbett, Greville G., Fedden, Sebastian |
| Verbless predicative structures across Romance | Murano, Nicola |
| Exponence and morphosyntactically triggered phonological processes in the Russian verbal complex | Gribanova, Vera |
| Systematic mismatches: Coordination and subordination at three levels of grammar | Belyaev, Oleg |
| Information structure, (inter)subjectivity and objectification | van der Wal, Jenneke |
| Phonotactic constraints and sub-syllabic structure: A difficult relationship | Berg, Thomas., Koops, Christian |

*Table 1. Abstracts used in the study.*

The first part of the analysis, as described by Erman and Warren (2000), was made using a system of markings. When analysing a piece of text, each individual prefab in it was marked with a slash at the beginning and end. Every word that does not belong to a prefab was replaced by a dash. This was done to create a representation of the previously mentioned "slots" in the sentences, as well as isolating the formulaic pieces. Words inside prefabs that could be replaced with other words (though the slots they present still need to be filled) were made cursive and smaller. When counting the number of slots filled by prefabs, these words were ignored.

Non-obligatory extensions of the prefabs were put inside parentheses. If the extension was not seen as a part of the prefab, it was disregarded. If, however, they are common extensions of that particular prefab, they were included. In the cases where the extension slots were filled by other prefabs, these were put inside brackets. This creates a text that looks like follows:

> /This paper proposes/ /a (constructional) account of/ /the (longstanding) issue of *the optional quotative to-marking*/ - - - (- -) - -. /We argue that/ /this *optionality* comes from/ /the availability of/ - /morphological constructions/ - - - - - - - - - /a set of *mimetics*./

In the example above, the prefab "we argue that" is marked with a slash at the beginning and end. The following words in the sentence are also part of a prefab, and are therefore also marked with slashes. Analysing the text in this way makes the use of prefabs easy to overview. It also shows how much modification is possible inside the expressions. To demonstrate, the section above contains a total of 42 slots, of which 25 have been filled by a total of eight prefabs.

After the prefabs had been identified and the text structured, the number of prefabs per abstract were counted, as well as how many words each prefab consisted of. Using this, the average amount of words per abstracts was calculated, followed by the average amount of words in all of the abstracts in the mini-corpus. Erman and Warren (2000) also calculate the number of choices in each text, but this is not done in this essay. The reason for this is that this study choses to instead focus more heavily on the types of prefabs used in the abstracts.

Erman and Warren (2000) also divide the function prefabs from function prefabs, with either a grammatical or pragmatic function. This is not the categorisation that will be used in this essay and the method diverges from the one used by Erman and Warren (2000) at this point. The prefabs were instead categorised using the method created by Simpson-Vlach and Ellis (2010), discussed above, due to its focus on academic language, and the categorisation method by Erman and Warren was deemed less useful for a study focusing on academic writing. There were four categories of prefabs in Simpson-Vlach and Ellis (2010) study: referential expressions, stance expressions, discourse organising expressions and others. The category of "others" is necessary since there may be some prefabs that do not fit in with the other categories, these being very genre-specific to academic language. However, this

category was made as small as possible, with the absolute majority of the prefabs being sorted into one of the other three sections. The method used for the categorisation process was simple colour coding, with one colour for each group. After this was done, the number of prefabs in each category and abstract were counted and the average calculated.

As has been discussed above, the most difficult part of research concerning FL is the identification of the prefabs. While Foster (2001) states that intuition can be a reliable way of identifying a piece of FL (specifically by using a number of university teachers to help in the identification process, all native speakers of English and working without consulting each other), this strategy also has many flaws. These flaws include the fact that the data sets used in the research must be small in size, since the identification is done manually. Inconsistencies are also likely to appear, where a chunk has been identified as formulaic in one place and not in another (Wray, 2002). In order to increase the reliability of this study, intuition is not the only method used in the identification process. While some prefabs can be easily identified even by a non-native speaker of English (such as the author of this essay), the more uncertain cases have been checked against different sources. In the cases of uncertain "normal" prefabs, they have been checked against the Corpus of Contemporary American English (COCA) (http://corpus.byu.edu/coca/) to investigate whether the words appear together frequently enough to be called a prefab. This corpus was chosen due to its size (520 million words) and the fact that it is fairly recent, the texts in it being from the time period 1990 – 2015. In the cases of more technical expressions (here referring to scientific terms and highly subject-specific words), they have been checked against Google Books. By putting the term in quotes and searching for it in this database, its frequency of use can be determined. If it is a phrase coined by the author of the abstract and not commonly used in the research community, it can be concluded that the expression is in fact not a prefab. It would be equally sufficient to use Google Scholar or another database where published academic text can be found, but Google Books was chosen in this study due to its clear search mechanisms and its large amounts of collected material. While these rules may still appear abstract, they work to greatly increase the reliability of the study.

### 3.2 Statistical comparison

A one-sample t-test will be conducted in this study, which means that the collected data is compared to a hypothetical average (NCSS, n.d). This is done in order to investigate whether

there is a statistical difference between the means produced by this study and the means produced by Erman and Warren (2000). The average amount of prefabs in the abstracts used in this essay will be compared to the average amount of prefabs occurring in written English, which according to Erman and Warren (2000) is 52.3 %. A second test will be made, in which the average lengths (counted in words) of the prefabs found in this study are compared to the average length of the written prefabs found by Erman and Warren (2000), which is 2.80 words. In connection with the t-test, Cohen's D will also be calculated by dividing the difference between the two means with the standard deviation. This calculates effect size, and is independent of the scale of the study (Cahan & Gamliel, 2011).

The t-tests will be made using the program SPSS Statistics, and are calculated by $t = \bar{x} - \mu / s_{\bar{x}}$, where $\bar{x}$ is the average of the collected data, $\mu$ is the assumed general average (here, 52.3 %) and $s_{\bar{x}}$ is the standard error. By observing the p-value generated by the tests, the statistical significance can be determined. A p-value of less than 0.1 % gives a three star significance, a value of less than 1% (but bigger than 0.1%) gives a two star significance, and a p-value of between 1% and 5% gives a one star significance. A p-value of over 5% means that there is no statistical significance (Körner & Wahlgren, 2015). The t-tests are based on a null hypothesis ($H_0$), meaning that it is assumed that the difference between the "true" mean (52.3 % for the first test and 2.80 words for the second) and the mean produced in this study is zero. The alternate hypothesis ($H_1$) states that there is, in fact, a statistically significant difference between the two means (Djurfeldt, Larsson & Stjärnhagen, 2003).

The initial assumption is that there will be no statistical significance when using the average amount of prefabs (null hypothesis accepted) and a small significance when comparing the average lengths of the prefabs (alternative hypothesis accepted). Whether or not the significance exists, the test is still useful in the analysis of the figures produced by this study.

### 3.3 Validity and reliability

When conducting a research study, taking validity and reliability into account is important. Reliability refers to whether or not the same study can be repeated several times with similar results each time. The results should also be similar when conducted by different people. Additionally, the tools used in the study should be reliable (Hartman, 2004). The reliability of the method used in this study has been briefly discussed above, but it is necessary to note that

while the method is partially based on intuition (which means that the results may vary from person to person), the additional checking of the prefabs against corpora and Google Books (which are hereby judged as being reliable instruments) increase the reliability. Other ways of increasing the reliability could include rechecking the data after a period of time, or using an outsider's perspective for further inspection. However, this challenges the inter-rater reliability, since several people making the same analysis may have different views of the results (Bryman & Bell, 2003). The results of this study have been rechecked after a period of time (the exact same analysis as the one described above being applied a second time to two of the abstracts after two weeks had passed), with the same results as in the first analysis.

The validity of the study is measured in terms of whether or not the results of the study reflect the actual facts of the subject. Human error is one of the factors that can decrease the validity of a research project, wherein expectations of the results and other surrounding factors may affect the outcome of the study (Hartman, 2004). The problems with validity in this study include the fact that the examined abstracts only come from one specific journal in one specific area of study. It can therefore not be argued that the results of this study are generalizable for the entire genre of written academic English. However, this study does not claim to present any universal trends within the genre, rather only the tendencies that can be found in the mini-corpus used in this specific essay. In order to investigate a more significant section of written academic English, a much larger study would have to be made, which could be a subject for future research. A possible future study could entail analysing texts from many different disciplines in order to draw more general conclusions about prefabs in English abstracts. It can also be added that one mistake was made in the selection of the abstracts, wherein one of the texts was originally published in the *Journal of Applied Linguistics* rather than the *Journal of Linguistics*. This is not expected to affect the results of this study in any way, but this mistake must be acknowledged nonetheless.

## 4. Results

In this section, the results of the study will be presented. First, the average amount of slots filled by prefabs will be presented. In other words, this section describes how much of the abstracts are filled by prefabs. In the second part, the average number of words per abstract is shown. Following this, the different types of prefabs found in the abstracts are presented, as well as the number of prefabs per category. Lastly, the results of the statistical comparison made with a one-sample t-test are presented.

### 4.1 Amount of slots filled by prefabs

The descriptive results for each abstract are listed in *Table 2* in the same order as they appear in Appendix 1. The total amount of slots in the mini-corpus is 1770, and the average amount of slots per abstract is 177. The amount of slots filled by prefabs span between 31.5 % at its lowest and 55.55 % at its highest amount.

| Abstract nr. | Slots, total | Non-prefab slots | Filler slots | Prefab slots | Prefab percent |
|---|---|---|---|---|---|
| **1.** | 166 | 65 | 21 | 80 | 48.19% |
| **2.** | 184 | 80 | 46 | 58 | 31.5 % |
| **3.** | 162 | 56 | 16 | 90 | 55.55% |
| **4.** | 110 | 38 | 21 | 51 | 46.36 % |
| **5.** | 189 | 63 | 30 | 96 | 50.79 % |
| **6.** | 204 | 72 | 29 | 103 | 50.49 % |
| **7.** | 200 | 77 | 31 | 92 | 46.00 % |
| **8.** | 208 | 65 | 37 | 106 | 50.70 % |
| **9.** | 128 | 45 | 15 | 68 | 53.12 % |
| **10.** | 222 | 71 | 30 | 121 | 54.50 % |
| **Average:** | | | | | **48.72%** |
| **Std. dev:** | 36,01 | 13,48 | 9,62 | 22,19 | 6,83 % |

*Table 2. Amount of prefabs*

### 4.2 Length of prefabs

Below, the number of prefabs per abstract as well as the average length of the prefabs (in number of words) in each abstract is presented in *Table 3*. The average number of prefabs in total, as well as the average length of all the examined prefabs is also presented. The number of prefabs per abstract span between 16 and 40, and the average lengths of the prefabs span between 2.3461 to 2.8275 words.

| Abstract nr. | Number of prefabs | Average length of prefabs |
|---|---|---|
| **1** | 29 | 2.83 |
| **2** | 23 | 2.43 |
| **3** | 30 | 2.73 |
| **4** | 16 | 2.75 |
| **5** | 39 | 2.41 |
| **6** | 37 | 2.46 |
| **7** | 29 | 2.76 |
| **8** | 37 | 2.51 |
| **9** | 26 | 2.35 |
| **10** | 40 | 2.63 |
| **Average:** | **30.6** | **2.56** |
| **Std. dev.** | 7.73 | 0.17 |

*Table 3. Length of prefabs*

### 4.3 Organisation of prefabs

The found prefabs have been sorted using Simpson-Vlach and Ellis' (2010) mode of organisation. *Table 4* shows that the most common function is by far referential expressions, used about twice as often as the second most common function of discourse organisation.

| Abstract nr. | Referential expressions | Stance expressions | Discourse organisation | Other |
|---|---|---|---|---|
| **1.** | 12 | 5 | 10 | 2 |
| **2.** | 11 | 4 | 9 | 5 |
| **3.** | 12 | 3 | 6 | 3 |
| **4.** | 9 | 2 | 5 | 0 |
| **5** | 21 | 6 | 4 | 6 |
| **6.** | 13 | 2 | 4 | 5 |
| **7.** | 17 | 7 | 8 | 5 |
| **8.** | 21 | 4 | 10 | 2 |
| **9.** | 17 | 3 | 8 | 8 |
| **10.** | 16 | 3 | 6 | 4 |
| **Average:** | **14.9** | **3.9** | **7** | **4** |
| **Std. dev.** | 4.15 | 1,66 | 2,31 | 2,31 |

*Table 4. Organisation of prefabs.*

Stance expressions were initially expected to be the least common type of expressions, due to the information density of the abstracts, which was accurate. Discourse organising expressions were also expected to be less common in abstracts (and perhaps more common in the running text of the article), which also held true in this analysis. Some examples of the expressions found in the abstracts are as follows:

- Referential expressions: *in the form of, the basis of, an account of, easily determined, neither a… nor…*
- Stance expressions: *an inherent bias, high degree of, based on the assumption, best suited to, more or less*
- Discourse organising expressions: *in order to, at the outset, on (the) one hand, in combination with, in terms of*

### 4.4. Statistical comparison

The proportion of prefabs per abstract (in percent), in this study 48.72 %, has been compared to a test value of 52.3%, taken from Erman and Warren's (2000) study. This has been done using a one-sample t-test. There was no difference between the mean based on the 10 abstracts and the mean test value, $t$ (9) = -1.656, $p$ = .132, $d$ = 0.52. The calculation of Cohen's D ($d$) shows that the effect size in this study is medium sized.

The average length of the prefabs for each abstract (2.56 words) was also compared to Erman and Warren's (2000) average prefab length of 2.80 words. In this comparison, $t$ (9) = -3.904, $p$ = .004 and $d$ = -1.25. This means that there is a two star statistical significance in the difference between the average values of the two studies, and that the effect size is larger than one standard deviation.

## 5. Discussion

The purpose of this essay was to investigate the extent to which prefabs are used in academic text, specifically in abstracts taken from articles in the subject of linguistics. According to Sinclair (1991), switches between the open-choice principle (similar to the slot-and-filler principle) and the idiom principle occur "when there is good reason". Additionally, the two different principles are more suitable in different circumstances. Sinclair (19991) brings up the differences in writing style between legal documents and poetry, stating that the open-choice principle is more fitting for a legal document. Taking this into account, it could be said that the open-choice principle would be more suitable for academic writing. This study shows that formulaic language is used less in abstracts than in "general" English writing, when comparing the results of this study to the study made by Erman and Warren, though not to a statistically significant degree (2000).

The idiom principle by Sinclair (1991) puts a large focus on choice and choices made by the language user when speaking or writing. While the concept of choice may not be something that can be investigated fully in a study such as this one, the principle is still important when analysing the results of the study. The fact that roughly half of the abstracts consist of prefabs shows that, just as Sinclair (1991) has stated, both the idiom principle and the open choice principle work together despite being fundamentally opposed. Whether conscious or not, the authors of the abstracts have at many points in the short texts chosen to fill the empty slots with readily-made multi-word prefabs. These prefabs have in turn been chosen due to their functionality as referential expressions, stance expressions or discourse organising expressions in order to in the small space of an abstract convey the most important parts of a much larger scientific study.

While the comparison between the average amount of prefabs produced by Erman and Warren (2000) and the average produced by this study only shows a tendency to a statistical significance (p-value: .132), it can be said that prefabs are used to a slightly smaller extent in abstracts when compared to "general" text. This could be due to a number of factors, one being that the abstracts contain several more specialised technical terms that cannot be counted as prefabs. While there are genre-specific prefabs that are widely used by the research community, expressions that for example have been invented by the author do not count as prefabs according to this study, since they are not yet widely used. Due to the

information density required of an abstract, using a large amount of these expressions may lead to a lower average amount of prefabs in the text.

This essay also examined the length of the prefabs occurring in the abstracts. These were shorter than the ones commonly found in "general" English, to a statistically significant degree. The reason for this could be the same as the reason why there are fewer prefabs in abstracts than other texts – information density and a need to make the text as short as possible. Though the author may chose to use a prefab due to its function (discussed below), the prefab needs to very effectively convey meaning in the limited space of the abstract, and therefore a shorter than average construction would be fitting.

One of the purposes of using formulaic language is reducing processing effort, adding variation in written text as well as to mark discourse (Wray, 2002). The second research question of the essay concerned investigating the types of prefabs used in the abstracts, particularity what pragmatic role they play in the text. An interesting aspect of the genre of academic English is Simpson-Vlach and Ellis' (2010) Group C, discourse organising expressions. Out of the ten abstracts, eight start with a discourse organising expression, the most common being *this paper* or *in this paper*. This is also one of the most common written discourse organising expressions in the study by Simpson-Vlach and Ellis (2010). On average, there are seven of these expressions in each abstract, but the amount ranges between four and ten.

According to research by Simpson-Vlach and Ellis' (2010) and Dontcheva-Navratilova (2012), the most common type of prefab in academic writing is referential expressions. This also holds true in this study, wherein there are on average 14.9 prefabs of this type in each abstract, compared to 3.9 stance expressions and seven discourse organising expressions. Dontcheva-Navratilova (2012) also writes that stance expressions are the least common type of prefab in academic writing. Her statement that one prefab can hold several functions at the same time and have different functions in different context holds especially true for stance expressions.

While the method of identifying prefabs by intuition may seem like, as Wray (2002) states, "the least scientific" method, it was successful in this essay due to the rechecking against corpora and Google Books. Others have used similar methods in the past with success (for

example, Vincent (2009) used this method in his Master's thesis and Erman and Warren (2000) did the same in their article which is the basis for this study), but the reliability of the method still needs to be discussed. When using an intuition-based method such as this, it is important to acknowledge the fact that another analyst may produce different results. Erman and Warren (2000) state that a prefab may not be seen as such to all members of a language community, and that identifying every prefab in a text is in practice impossible.

Hyland (2008) claims that there are not enough lexical bundles commonly used in general academic language to be organised, and that they should instead be sorted by the subjects in which they are used. While this thesis follows the Simpson-Vlach and Ellis (2010) approach of organising the prefabs according to their function in the text, an alternative to this study could be investigating abstracts from many different subjects instead of just Linguistics. These could be organised according to subject, or a combination of the methods could be created wherein the prefabs are sorted by both subject and function. This could give a broader view of how prefabs are used in abstracts across different disciplines.

## 6. Conclusion

The first research question of this study asked to which extent prefabs are used in academic text, specifically abstracts from linguistic articles. The second question concerned what types of prefabs were used in the text and their pragmatic roles. The study concludes that prefabs are used to a slightly smaller extent in abstracts than in "general" text, the extent being 48.72% (compared to Erman and Warren's (2000) result of 52.3 %). However, the difference between the two numbers was not statistically significant. This small difference could be due to the information density of the abstracts and the technical expressions used in them. The results of the study also shows that prefabs are in general shorter in abstracts than in general text, with the average length of prefabs in abstracts being 2.59 words, compared to the general 2.80 words. There is a two-star statistical significance in the difference between these two results. Previous studies (Simpson-Vlach and Ellis (2010), Dontcheva-Navratilova (2012)) have stated that the most common function of a prefab in an academic text is as a referential expression. This is also supported by this study.

Previous research (Dontcheva-Navratilova, 2012) has investigated first/second language use of prefabs in academic text. This could be a possible subject for future research concerning how and to what extent L2 users of English use prefabs in their academic writing when compared to L1 users. Additional future studies within the field could include investigating different disciplines, such as Business, Natural Sciences or Engineering in a similar way to how texts from the subject of Linguistics have been investigated in this essay.

## References


Biber, D. & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, vol. 26, no. 3,* pp. 263 – 286. http://dx.doi.org/10.1016/j.esp.2006.08.003

Bryman, A. & Bell, E. (2003). *Företagsekonomiska forskningsmetoder.* Solna: Liber

Cahan, S. & Gamliel, E. (2011). First Among Others? Cohen's *d* vs. Alternative Standardized Mean Group Difference Measures. *Practical Assessment, Research & Evaluation, vol. 16, no 10.*

Cutler, A. & Swinney, D. A. (1979). The Access and Processing of Idiomatic Expressions. *Journal of Verbal Learning and Verbal Behaviour, no. 18,* pp. 523 – 524.

Djurfeldt, G., Larsson, R., Stjärnhagen, O. (2003). *Statistisk verktygslåda – samhällsvetenskaplig orsaksanalys med kvantitativa metoder.* Lund: Studentlitteratur

Dontcheva-Navratilova, O. (2012). Lexical Bundles in Academic Texts by Non-native Speakers. *Brno Studies In English, vol. 38, no. 2,* pp. 38-58.

Erman, B. & Warren, B. (2000). The idiom principle and the open choice principle. *De Gruyter, vol. 20, no. 1,* pp. 29-61. https://doi.org/10.1515/text.1.2000.20.1.29

Hartman, J. (2004). *Vetenskapligt tänkade: från kunskapsteori till metodteori.* Lund: Studentlitteratur

Hyland, K. (2008). As can be seen: Lexical Bundles and disciplinary variation. *English for specific purposes, vol. 27, no 1,* pp. 4-21. Retrieved from http://www.sciencedirect.com.ludwig.lub.lu.se/science/article/pii/S0889490607000233 2017-03-21

Körner, S. & Wahlgren, L. (2015). *Statistiska metoder*. Lund: Studentlitteratur

NCSS Statistical Software. (n.d). *One-Sample T-Test.* Retrieved from http://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/One-Sample_T-Test.pdf 2017-04-02

Schmitt, N. (2010). *Researching Vocabulary – A Vocabulary Research Manual.* Basingstoke: Palgrave Macmillan

Shaoul, C., Westbury, C. (2012). Formulaic sequences – do they exist and do they matter? In G. Libben, G. Jaerma and C. Westbury (Eds.), *Methodological and Analytic Frontiers in Lexical Research* (pp. 171 – 196). Amsterdam: John Benjamin's Publishing Company

Shu, L. (2013). Analysis of Prefabricated Chunks Used by Second Language Learners on Different Levels. *Theory and Pratice in Language Studies, vol. 3, no 9,* pp. 1667 – 1673.

Simpson, R. & Mendis, D. (2003). A Corpus-Based Study of Idioms in Academic Speech. *Tesol Quearterly, vol. 37, no. 3,* pp. 419 – 441. Retrieved from https://academic.oup.com/applij/article-abstract/31/4/487/191083/An-Academic-Formulas-List-New-Methods-in 2017-03-20

Simpson-Vlach, R. & Ellis, N. C. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics, vol. 31, no. 4,* pp. 487-512.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press

Staples, S., Egbert, J., Biber, D., Gray, B. (2016). Academic Writing Development at the University Level: Phrasal and Clausal Complexity Across Level of Study, Discipline, and Genre. *Written Communication, vol. 33,* pp. 149 – 183.

Sunderland, J. (2010). Research Questions in Linguistics. In L. Litosseliti (Ed.). *Research Methods in Linguistics*, pp. 9 - 28. London: Continuum

Vincent, B. (2009). Calculating the extent of the idiom principle through corpus analysis of a short text. (*Master's thesis, University of Birmingham, Birmingham*). Retrieved from: http://www.birmingham.ac.uk/schools/edacs/departments/englishlanguage/research/resources/essays/corpus-linguistics.aspx

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: CUP

# Appendix 1

**1. Lexical Bundles and Disciplinary Variation in University Students' Writing: Mapping the Territories**

DURRANT, PHILIP
DOI: https://doi-org.ludwig.lub.lu.se/10.1093/applin/amv011
Published online: April 2017
Volume 38, Issue 2
April 2015, pp. 165-193.

This paper describes disciplinary variation in university students' writing, as it is reflected in the use of recurrent four-word sequences. In contrast to previous studies, disciplinary categories are not assumed at the outset of the analysis, but rather emerge from an initial analysis of variation across all writers in the corpus. Variation is presented in the form of a visual map representing degrees of similarity and difference between individual writers. Emergent disciplinary groupings are then used as the basis for a qualitative analysis of distinctive lexical bundles. Analysis reveals four main disciplinary groupings. A primary distinction appears between hard (science/technology) and soft (humanities/social sciences) subjects, with two further groupings (life sciences and commerce) being intermediate between these two. Evidence is also found of cross-group disciplines, which draw on a variety of influences, and of particular disciplines which are internally heterogeneous. A qualitative analysis of bundles which are distinctive of 'hard' and 'soft' disciplines is presented in order to characterize the discourse functions which mark these categories.

/This paper describes/ /disciplinary variation/ - /university students'/ -, /as it is/ /reflected in/ /the use of *recurrent four-word sequences*/. /In contrast to [previous studies]/, /disciplinary categories/ /are (not) assumed/ /at the outset/ - - - - - /emerge from *an initial analysis*/ - /variation across *all writers*/ - - -. - - - /in the form of [a visual map]/ - - - - /difference between [individual writers]/. - - - /are then used/ - /the basis for [a qualitative analysis]/ - - /lexical bundles/. - - - - - -. - /primary distinction/ /appears between *hard (science/technology)* and *soft (humanities/social sciences) subjects*/, - - - - (/life sciences/ - -) - - - - -. /Evidence is (also) found/ - - -, - /draw on/ /a variety of *influences*/, - - - - - - - - -. /A qualitative analysis/ - - - - - - /which are/ /distinctive of *'hard' and 'soft' disciplines*/ /is presented/ /in order to *characterize*/ - /discourse functions/ - - - -.

**2. Middle-passive voice in Albanian and Greek**

M. RITA MANZINI, ANNA ROUSSOU and LEONARDO M. SAVOIA
DOI: https://doi.org/10.1017/S0022226715000080
Published online: 24 March 2015
Volume 52, Issue 1
March 2016, pp. 111-150

## Abstract

In this paper we consider middle-passive voice in Greek and Albanian, which shows a many-to-many mapping between LF and PF. Different morphosyntactic shapes (conditioned by tense or aspect) are compatible with the same set of interpretations, which include the passive, the reflexive, the anticausative, and the impersonal (in Albanian only). Conversely, each of these interpretations can be encoded by any of the available morphosyntactic structures. Specialized person inflections (in Greek and Albanian), the clitic *u* (Albanian) and the affix *-th-* (Greek) lexicalize the internal argument (or the sole argument of intransitive in Albanian) either as a variable, which is LF-interpreted as bound by the EPP position (passives, anticausatives, reflexives) or as generically closed (impersonals, in Albanian only). The ambiguity between passives, anticausatives and reflexives depends on the interpretation assigned to the external argument (generic closure, suppression or unification with the internal argument respectively). In perfect tenses, auxiliary *jam* 'be' in Albanian derives the expression of middle-passive voice due to its selectional requirement for a participle with an open position. Crucially, no hidden features/abstract heads encoding interpretation are postulated, nor any Distributed Morphology-style realizational component.

/In this paper/ /we consider/ /middle-passive voice/ - - - -, /which shows/ - - - /between *LF* and *PF*/. - - - (/conditioned by *tense or aspect*/) /(are) compatible with/ /the (same) set of *interpretations*/, /which include/ - - - - - - - - - (/in *Albanian* only/). -, /each of these *interpretations*/ - - /encoded by/ /any of the available *morphosyntactic structures*/. - - - (- - - - ), - - - (-) - - - - (-) - - - - (- - - - - - - - -) /either as a *variable, which is LF-interpreted as bound by the EPP position (passives, anticausatives, reflexives)* or as *generically closed (impersonals, in Albanian only)/*. /The ambiguity between *passives, anticausatives and reflexives*/ /depends on *the interpretation*/ /assigned to *the external argument*/ (/generic closure/, - - - /with the/ - - -). - - -, - - - - - - /(the) expression of *middle-passive voice*/ /due to *its selectional requirement*/ - - - - - - - -. - - /hidden features/ - - - - - -, /nor any *Distributed Morphology-style realizational component*. /

### 3. A constructional account of the 'optional' quotative marking on Japanese mimetics

KIMI AKITA and TAKESHI USUKI
DOI: https://doi.org/10.1017/S0022226715000171
Published online: 20 May 2015
Volume 52, Issue 2
July 2016, pp. 245-275

## Abstract

This paper proposes a constructional account of the longstanding issue of the optional quotative *to*-marking on manner-adverbial mimetics (or ideophones) in Japanese. We argue that this optionality comes from the availability of two morphological constructions – the bare-mimetic predicate construction and the quotative-adverbial construction – to a set of mimetics. On the one hand, the bare-mimetic predicate construction incorporates previously identified phonological, syntactic, and semantic conditions of the bare realization of mimetics.

This construction is instantiated by bare mimetics (e.g. *pyókopyoko* 'jumping around quickly') in combination with their typical host predicates (e.g.*hane*- 'jump'), and they behave as loose complex predicates with more or less abstract meanings. As with 'say'- and 'do'-verbs, these complex predicates involve quasi-incorporation, which is a constructional strategy for the morphosyntactic integration of mimetics into sentence structures. On the other hand, the quotative-adverbial construction introduces mimetics to sentences with a minimal loss of their imitative semiotics. This fundamental function is consistent with the wide distribution of quotative-marked mimetics.

/This paper proposes/ /a (constructional) account of/ /the (longstanding) issue of *the optional quotative to-marking*/ - - - (- -) - -. /We argue that/ /this *optionality* comes from/ /the availability of/ - /morphological constructions/ - - - - - - - - - /a set of *mimetics*./ /On (the) one hand/, - - - /previously identified/ /phonological, syntactic, and semantic conditions/ - - /bare realization of *mimetics*/. - - - /instantiated by/ - - ( - /' - - -'/) /in combination with/ - - /host predicates/ (- -), - /they behave as *loose complex predicates* / - /more or less [*abstract meanings*]/. /As with/ - - - -, - /complex predicates/ - -, /which is/ /a (constructional) strategy/ /for the *morphosyntactic* integration/ - - - /sentence structures/. /On the other hand/, - - - /introduces *mimetics* to/ - /with a minimal loss/ /- - - -. /This (fundamental) function/ - /consistent with/ /the wide distribution of *quotative-marked mimetics*/.

### 4. Clitics: Separating syntax and prosody

JOHN J. LOWE
DOI: https://doi.org/10.1017/S002222671500002X
Published online: 24 April 2015
Volume 52, Issue 2
July 2016, pp. 375-419

## Abstract

A problematic feature of clitic positioning attested in a number of languages is the ability of a clitic to appear inside a syntactic unit of which it is not itself a part, apparently due to prosodic restrictions on its positioning. The influence of prosody on syntax presents a challenge for any formal account, particularly any that strives to respect a modular view of the grammatical architecture. I present an account of clitic positioning within a recently proposed model of the syntax–phonology interface that aims at full modularity, showing that it is indeed possible in such an architecture, and showing where and how prosody and syntax interact in this model.

/A (problematic) feature of *clitic positioning*/ /attested in/ /a number of *languages*/ - /the ability of *a clitic*/ - - - - - - /of which it is *not itself a part*/, /(apparently) due to *prosodic restrictions*/ - - -. /The influence of *prosody* on *syntax*/ /presents a challenge/ - - - -, - - - /strives to *respect*/ /a (modular) view of *the grammatical architecture*/. - - /an account of *clitic*

*positioning*/ - /(a) recently proposed *model*/ - - - - - - - - -, /showing that/ /it is (indeed) possible/ /in such an *architecture*/, - - /where and how/ - - - - - - -.

## 5. Canonical gender

GREVILLE G. CORBETT and SEBASTIAN FEDDEN
DOI: https://doi.org/10.1017/S0022226715000195
Published online: 24 June 2015
Volume 52, Issue 3
November 2016, pp. 495-531

### Abstract

Nominal classification remains a fascinating topic but in order to make further progress we need greater clarity of definition and analysis. Taking a Canonical Typology approach, we use canonical gender as an ideal against which we can measure the actual gender systems we find in the languages of the world. Building on previous work on canonical morphosyntactic features, particularly on how they intersect with canonical parts of speech, we establish the distinctiveness of gender, reflected in the Canonical Gender Principle: In a canonical gender system, each noun has a single gender value. We develop three criteria associated with this principle, which together ensure that canonically a noun has exactly one gender value; we give examples of non-canonicity for each criterion, thus gradually building the typology. This is the essential groundwork for a comprehensive typology of nominal classification: the Canonical Typological approach allows us to tease apart clusterings of properties and to characterize individual properties with respect to a canonical ideal, rather than requiring us to treat the entire system as belonging to a single type. This approach is designed to facilitate comparisons of different noun classification systems across languages.

/Nominal classification/ /remains a *fascinating topic*/ - /in order to [make further progress]/ we need /greater clarity (of) *definition and analysis*. /Taking a *Canonical Typology* approach/, - - - - /as an ideal/ /against which/ - - - - - /gender systems/ - - - /the languages of the world/. /Building on [previous work]/ - /canonical morphosyntactic features/, /particularly on *how they intersect* / - - /parts of speech/, - - /the distinctiveness of *gender*/, /reflected in *the Canonical Gender Principle*/: - - - /gender system/, - - - - - /gender value/. - - - - /associated with *this principle*/, /which (together) ensure that/ - - - - - - /gender value/; - - /examples of *non-canonicity*/ /for each *criterion*/, - /gradually building/ - -. - - /the essential groundwork/ - /a comprehensive *typology* of/ /nominal classification/: /the *Canonical Typological* approach/ /allows us to *tease apart*/ - - - - - - /individual properties/ /with respect to *a canonical ideal*/, / rather than *requiring*/ - - - - - - - - - /a single type/. /This approach/ - /designed to *facilitate comparisons*/ - - - /classification systems/ /across languages/.

## 6. Verbless predicative structures across Romance1

NICOLA MUNARO
DOI: https://doi.org/10.1017/S0022226715000201
Published online: 01 June 2015

Volume 52, Issue 3
November 2016, pp. 609-637

## Abstract
This article develops an analysis of a verbless predicative structure attested throughout Romance: in this type of reduced clause the predicate linearly precedes the subject and is separated from it by a clear intonational break, while the missing verb is interpreted as a silent copula. I argue that this structure should be viewed as the result of three movement steps: the first step is to be identified with predicate inversion, that is, extraction of the predicate from the complement position of the predicative small clause to a higher specifier position thanks to phase extension, followed by raising of the predicate to the specifier of SubjP to check the EPP feature, and finally to the specifier of the left-peripheral projection FocusP in order to check a focus feature. The present analysis is based on the crucial, and independently motivated assumption, that the process of phase extension, produced by raising of the small clause internal relator R° to a higher functional head F°, is limited to small clauses associated with individual-level predicates. The verbless predicative structure is then compared to an analogous construction in which the preposed predicate is preceded by a *wh*-item, arguing that, despite their apparent similarity, the two structures should be clearly distinguished.

/This article/ - /an analysis of *a verbless predicative structure*/ /attested throughout *Romance*/: /(in) this type of [reduced clause]/ - - - /precedes the *subject*/ - - /separated from *it*/ - /a clear (intonational) break/, - - - - - /interpreted as *a silent copula*/. /I argue that/ /this structure/ /should be viewed/ - /the result of *three movement steps*/: /the first step/ - /to be identified (with) [predicate inversion]/, /that is/, - - - - - - - - - - /predicative small clause/ /to a higher (specifier) position/ /thanks to *phase extension*/, /followed by *raising of the predicate*/ - - - - - - - - - - -, /and finally/ - - - - - - - - - /in order to/ - - /focus feature/. /The (present) analysis/ - /based on the (crucial, and independently motivated) assumption/, - /the process of *phase extension*/, /produced by *raising*/ - - - - - - - - - - - /is limited to/ - - /associated with *individual-level predicates*/. - - - - /is then/ /compared to *an analogous construction*/ /in which/ - - - - /preceded by *a wh-item*/, /arguing that/, /despite their (apparent) similarity/, - - - - - /clearly distinguished/.

## 7. Exponence and morphosyntactically triggered phonological processes in the Russian verbal complex

VERA GRIBANOVA
DOI: https://doi.org/10.1017/S0022226714000553
Published online: 31 March 2015
Volume 51, Issue 3
November 2015, pp. 519-561

## Abstract
This paper examines a non-canonical morphophonological vowel alternation in the roots of Russian verbs that is conditioned by aspectual information (derived imperfectivization). This

aspectual morpheme is usually expressed as a suffix, but in the forms of interest appears as a vocalic nucleus in the root (whereas there is no vocalic nucleus in the perfective form). In a manner broadly compatible with Distributed Morphology (DM), I argue that this alternation is part of a more general phonological process – yer realization – special only in that it is triggered by morphosyntactic, rather than phonological, information. I propose an analysis of this pattern in which autosegmental representations – in this case, a mora – can be the exponents of morphosyntactic features. This approach obviates the need for DM readjustment rules, which have been criticized on empirical and theoretical grounds (Siddiqi 2006, 2009; Bye & Svenonius 2012; Haugen & Siddiqi 2013). I demonstrate that the required allomorphic interaction between the root and the derived imperfective morpheme is local, despite surface appearances: the intervening vowel is a theme vowel, inserted post-syntactically. This approach makes sense of broader patterns involving this theme vowel, and vindicates theories of allomorphic interaction that impose strict locality conditions (e.g., structural and/or linear adjacency).

/This paper examines/ - - - /vowel alternation/ /in the roots of *Russian verbs*/ - - /conditioned by *aspectual information*/ (- -). - /aspectual morpheme/ /is (usually) expressed as *a suffix*/, - /in the forms of *interest*/ /appears as *a vocalic nucleus*/ /in the root/ (- /there is/ - - - - - /perfective form/). /In a manner/ - /compatible with/ - - (-), /I argue that/ - - - /part of a (more general phonological) process/ – - - – - /only in that/ /it is triggered by *morphosyntactic, rather than phonological, information*/. /I propose/ /an analysis of *this pattern*/ /in which/ - - – /in this case/, - - – - - - - - /morphosyntactic features/. - - - /the need for *DM readjustment rules*/, - - - /criticized on (*empirical and theoretical*) grounds/ (- - - -). /I demonstrate that/ - - - -/between the *root* and the *derived imperfective morpheme*/ - -, /despite (surface) appearances/: - - - - - - - - - -. - - /makes sense of *broader patterns*/ - - - -, - - /theories of *allomorphic interaction*/ - - - - - (- - - - -).

**8. Systematic mismatches: Coordination and subordination at three levels of grammar1**

OLEG BELYAEV

DOI: https://doi.org/10.1017/S0022226714000450
Published online: 09 December 2014
Volume 51, Issue 2
July 2015, pp. 267-326

## Abstract

In this paper, I analyze two clause combining strategies in Ossetic that exhibit mixed properties between coordination and subordination. I argue that the 'mismatch approach' proposed by Culicover & Jackendoff (1997) and Yuasa & Sadock (2002) is best suited to account for their properties. However, in order to adequately describe the behavior of these constructions in terms of the mismatch approach, appealing to three levels of grammar is required instead of two levels (syntax and semantics) discussed in previous works. This provides a clear argument in favor of models of grammar such as Lexical Functional Grammar (LFG), where the syntactic level is split between constituent structure (c-structure) and functional structure (f-structure). The properties of semantic coordination and

subordination that have been proposed in earlier work mostly belong to the level of f-structure, and not semantics proper. I argue that the only substantial semantic difference between coordination and adverbial subordination is that the former introduces discourse relations between speech acts, while the latter introduces asserted predicates that link two propositions within the same speech act. I provide definitions of coordination and subordination at all the three levels of grammar formalized in terms of the LFG framework, and discuss the tests that can be used for each of these levels.

/In this paper/, - - /clause combining strategies/ - - - - - - /between *coordination* and *subordination*/. /I argue that/ /the 'mismatch approach'/ /proposed by/ - - - - - - /best suited to/ /account for *their properties*/. - , /in order to/ - - /the behavior of *these constructions*/ /in terms of/ /the mismatch approach/, /appealing to *three levels of grammar*/ /is required/ /instead of *two levels*/ (- -) - - /previous works/. - - /a clear argument/ /in favor of *models of grammar*/ /such as/ - - -, - - - - - /split between *constituent structure (c-structure)* and *functional structure (f-structure)*/. /The properties of *semantic coordination*/ - - - /have been proposed/ - /earlier work/ /(mostly) belong to *the level of f-structure*/, - - /semantics proper/. /I argue that/ /the only (substantial) (semantic) difference/ /between *coordination* and *adverbial subordination*/ is that the former introduces /discourse relations/ - /speech acts/, - /the latter/ - /asserted predicates/ - - - - - - - /speech act/. - - /definitions of *coordination and subordination*/ - - - - - - - - /in terms of/ - - -, - - - - - /can be used for *each of these levels*/.

## 9. Information structure, (inter)subjectivity and objectification

Jenneke van der Wal
DOI: https://doi.org/10.1017/S0022226714000541
Published online: 17 December 2014
Volume 51, Issue 2
July 2015, pp. 425-464

## Abstract

This paper discusses how information structure can be seen as a subjective and intersubjective concept in Verhagen's (2005) and Breban's (2010) definitions, though less so in Traugott's (2010) use of the terms. More difficult is the question of whether markers of information structure can be characterised as (inter)subjective; this is more easily determined for morphological markers than for prosody or word order. For unambiguous markers of information structure, I suggest that their emergence (e.g. copula > focus marker) is typically accompanied by (inter)subjectification, whereas their further development (e.g. topic marker > subject marker) displays objectification. The paper not only shows that grammatical items can undergo an increase as well as a decrease in (inter)subjectivity – thus denying strict unidirectionality – but also confirms that these processes are independent of grammaticalisation.

/This paper discusses *how*/ /information structure/ /can be seen as *a subjective and intersubjective concept*/ - - - - -, /(though) less so *in Traugott's*/ /use of *the terms*/. /More

difficult is [the question of]/ - /markers of/ /information structure/ - - /characterised as *(inter)subjective*/; /this is/ /(more) easily determined/ - /morphological markers/ - - - - /word order/. - - /markers of [information structure]/, /I suggest that *their emergence*/ (- - - -) - - /accompanied by *(inter)subjectification*/, - - /further development/ (- - - -) - -. /The paper/ /not only [shows that]/ /grammatical items/ - - - - /as well as/ - - - - - – - - - - – /but also/ /confirms that/ - - -/independent of *grammaticalisation*/.

## 10. Phonotactic constraints and sub-syllabic structure: A difficult relationship

THOMAS BERG and CHRISTIAN KOOPS
DOI: https://doi.org/10.1017/S002222671400022X
Published online: 18 June 2014
Volume 51, Issue 1
March 2015, pp. 3-39

## Abstract

Of late, a controversy has arisen over the internal structure of Korean syllables. While there is general agreement that non-phonotactic criteria argue for left-branching, Lee & Goldrick's (2008) left-branching phonotactic analysis is contradicted by Berg & Koops's (2010) claim as to a phonotactically symmetrical syllable structure. A comparison of the methodologies of the two studies, a revisit of the previous data and a new analysis cement the conclusion that there is neither a left-branching nor a right-branching phonotactic effect in Korean syllables. An investigation of the phonotactic structure of Finnish CVC syllables, which exhibit a psycholinguistic left-branching bias much like Korean, reveals that word-initial syllables possess a largely symmetrical organization whereas word-final syllables tend to show a right-branching slant. This curious set of results is consistent with the following three hypotheses: (i) The phonotactic criterion has an inherent VC bias. (ii) Symmetrical syllable structures represent a compromise between left- and right-branching effects. (iii) The strength of phonotactic constraints increases from earlier to later portions of words. The bottom line of this analysis is that, contra all previous claims, phonotactic constraints cannot be used as an argument for sub-syllabic constituency. We discuss the proposal that the basis of the left-branching bias in Korean syllables is instead to be found in the high degree of coarticulation between the onset consonant and the following vowel.

/Of late/, /a *controversy* has arisen/ - - /internal structure/ - - -. - /there is/ /general agreement/ - - - /argue for/ - - - - /phonotactic analysis/ - /contradicted by *Berg & Koops's (2010)* (claim)/ /as to/ - - - /syllable structure/. /A comparison of *the methodologies of the two studies*/, /a revisit of [the previous data]/ - - - - - /the conclusion that/ /there is/ /neither a *left-branching* nor a *right-branching*/ /phonotactic effect/ - - -. /An investigation of (the) [phonotactic structure]/ - - - -, /which exhibit/ - - - - /much like/ -, /reveals that/ - - - - - - - - - - /tend to show *a right-branching slant*/. - - /set of results/ - /consistent with/ /the following/ - -: (i) - - - - /an inherent (VC) bias/. (ii) - - - - /a compromise between *left- and right-branching effects*/. (iii) /The strength of *phonotactic constraints*/ - /from earlier to later *portions*/ of words. /The bottom line (of this analysis) is/ -, - -

/previous claims/, /phonotactic constraints/ /cannot be/ /(used) as an argument for *sub-syllabic constituency*/. - - /the proposal that /the basis of *the left-branching bias*/ - - - - - - /be found in/ - /high degree of/ - /between the *onset consonant* and the *following vowel*.