Formulaic language

Theories and methods

Edited by

Aleksandar Trklja and Łukasz Grabowski

Phraseology and Multiword Expressions 5

language science press

Phraseology and Multiword Expressions

In this series:

1. Manfred Sailer & Stella Markantonatou (eds.). Multiword expressions: Insights from a multilingual perspective.
2. Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.). Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop.
3. Yannick Parmentier & Jakub Waszczuk (eds.). Representation and parsing of multiword expressions: Current trends.
4. Schulte im Walde, Sabine & Eva Smolka (eds.). The role of constituents in multiword expressions: An interdisciplinary, cross-lingual perspective.

ISSN: 2625-3127

Formulaic language

Theories and methods

Edited by

Aleksandar Trklja and Łukasz Grabowski

language science press

Aleksandar Trklja & Łukasz Grabowski (eds.). 2021. *Formulaic language: Theories and methods* (Phraseology and Multiword Expressions 5). Berlin: Language Science Press.

This title can be downloaded at:
http://langsci-press.org/catalog/book/000

ISBN: no digital ISBN
no print ISBNs!
ISSN: 2625-3127
no DOI
ID not assigned!

Cover and concept of design: Ulrike Harbort
Typesetting: Sebastian Nordhoff, Felix Kopecky
Fonts: Libertinus, Arimo, DejaVu Sans Mono
Typesetting software: XeLaTeX

Language Science Press
xHain
Grünberger Str. 16
10243 Berlin, Germany
langsci-press.org

Storage and cataloguing done by FU Berlin

Freie Universität Berlin

Contents

Part I

New theoretical and methodological insights into formulaic language

Chapter 1

Predictability and prefab status: The case of adjective + noun sequences in English

Joan Bybee
University of New Mexico

Ricardo Napoleão de Souza
University of Helsinki

The main arguments of the chapter are that frequency or predictability effects are most appropriately applied within constructions, and that prefab status can lead to phonetic effects independently of frequency. The analysis uses tokens of adjective-noun sequences taken from conversational corpora. Study 1 applies diagnostics for conventionalization to establish prefab vs. free combination status in 239 bigrams. As expected from previous research, the criterion of non-compositional meaning gives clear results, while the criterion of fixedness runs into difficulties, especially as clusters of related prefabs are revealed. Study 2 examines the effect of prefab status vs. the effect of bigram frequency on the duration of the vowel in the adjective. Both prefab status and high bigram frequency lead to a shorter vowel, though the effect is not statistically significant.

1 Background

Of the many types of multiword expressions identified in the literature, the current study focusses on a small, structurally defined set of possible prefabricated sequences (prefabs), in particular, attributive adjective-noun sequences as used in American English conversation. The paper examines the role of frequency of

use (derived from various sorts of predictability measures) in contrast to the role of meaning and conventionalization in the formation and processing of prefabs. Both types of measures come with their challenges and our goal here is to discuss these challenges as we try to apply the two approaches to a small set of ten adjectives as used in excerpts from American English conversation.

Various tests have been applied to determine the role that multiword expressions play in linguistic processing. The experimental literature examines reaction times in acceptability judgment or reading tests (**EllisEtAl2008**; **GyllstadWolter2014**; **WolterYamashita2018**), as well as in priming experiments (**DurrantDoherty2010**). The latter authors argue that priming between words in a collocation indicates a mental record of their instances of co-occurrence. In another thread in the literature, researchers use spoken corpora to measure word duration and/or consonant and vowel reduction in word sequences that are in different probability relations (for example, Word 1 does or does not predict Word 2) (**BellJurafsky2009**). More recent research has shown that the phonetic properties of a word reflect its cumulative contexts of use over time (**Seyfarth2014**), leading to the finding that, for example, more predictable words grow shorter (**SóskuthyHay2017**).

In this contribution, we evaluate frequency and predictability-derived explanations of sequential processing vs. explanations that take into account syntactic context and meaning relations. Our data come from a study of vowel duration in adjectives which shows that the vowels in adjectives in attributive constructions are significantly shorter than those used in predicative constructions (**Souza2019**). Using the data on adjectives in attributive constructions, this paper first examines the results of corpus-based predictability studies and evaluates the likelihood that the results of such studies provide insights into storage and processing if they do not take into account grammatical factors such as the location of words within constructions of different types. Next, in Study 1 the discussion considers 237 adjective noun (AN) types taken from conversation in order to evaluate criteria for determining if a bigram has features of conventionalization as proposed in the literature (especially, **PawleySyder1983**; **ErmanWarren2000**; **Wray2002**; **CorriganEtAl2009**). We chose to use the term *prefab* (= prefabricated expression) in our discussion, because we view such conventionalized expressions as chunks for the purposes of storage, production and comprehension (**Bybee1998**; **Bybee2010**; **Bybee1998**). Using the resulting classification for prefab status, but controlling for frequency, in Study 2 we test whether the vowel in the adjective is shorter if it occurs in a prefab than in a novel pairing of AN. The results indicate that prefab status does have some impact on vowel duration in attributive constructions independently of frequency of use.

1.1 The role of frequency in the creation and processing of multiword expressions

One source of interest in multiword expressions (MWE) derived from corpus studies, which showed that certain strings of words tend to recur. Thus frequency of occurrence in a corpus can be an identifier of collocations (**JonesSinclair1974**) and lexical bundles (**BiberEtAl1999**). As texts and corpora are created by language users, it is proposed that such recurring expressions are also characterized as cognitive or production units that, as **PawleySyder1983** put it, have the effect of producing nativelike selection and nativelike fluency (see also **Siyanova-ChanturiaMartinez2(**
Only by being entrenched in memory storage can such units be recognized as conventional and at the same time serve to facilitate production and comprehension (**Langacker2008**). **Ellis1996** points out that such units are the result of the domain-general process of chunking by which memory is organized into recurring sequences. **Bybee1998**; **Bybee2002**; **Bybee2010** argues further that the sequential chunks of language use are the basis of constructions and constituent structure.

From a cognitive-processing perspective, it has not been established how much repetition is required to form a linguistic chunk. Certainly, high levels of repetition lead to routinization and the formation of constructions, to grammaticalization, to semantic/pragmatic change and phonological reduction (**Bybee2003**; **Haiman1994**; **Croft2000**). But it is undeniable that there are many word sequences recognized as conventionalized that are relatively low frequency. For example, the phrase *vanishingly rare* occurs only 13 times in the 600 million word COCA corpus (**Davies2008**) compared to another possible MWE *broad spectrum* which occurs 521 times. **Hoffman2004** argues that grammaticalization can also occur among phrases or constructions that are not of high frequency. He proposes that if a certain phrase is conventionalized as the preferred way of expressing a concept in a certain speech community, then it may be more salient than its frequency in corpora would predict. Conventionalization comes about by tacit agreement among speech participants, but only one or two repetitions may be enough to establish that agreement, as evidenced by the fact that language users command rare words and idioms that are infrequent, but widely known (**Wray2002**).

Thus frequency, while surely a factor in the processing of MWEs, is not the only factor that leads to entrenchment (**Wray2002**; **Schmid2017**). Rather text frequency interacts with cultural conventions established for the specific referents of MWEs, sometimes leading to special meanings. The two factors of text frequency and preferred means of expression may be partly independent since con-

ventionalized expression is possible in both high and low frequency phrases, and chunking in phrases of extreme high frequency may take place quite independently of the meaning of the chunk or its components. **Bybee2002** points out frequent repetition of items in sequence can lead to their phonological fusion even if the result is not semantically coherent. In many European languages, prepositions fuse with determiners (Spanish *de + el > del* 'of the [masc.]') and *a + el > al* 'to the [masc.]'), even though the result combines two units with completely independent meaning and function. Similarly, the contraction of auxiliaries in English creates phonological units that are semantically complex, including a pronoun and parts of modal or aspectual constructions: *I'm, he'd, they've,* etc.

The foregoing discussion suggests that it might be wise to distinguish MWEs whose only relevant property is frequency of use from those that have specific semantic or functional status. We return to this point in Sections 2–3, after considering how frequency interacts with MWEs and is used by researchers to understand the processing of these complex units.

1.2 Correlates of frequency and MWE status

Two well-documented correlates of token or text frequency of single words are speed of access in experimental settings (see **Ellis2002** for an overview) and phonetic reduction in natural spoken language. Recent research using eye-tracking technology has demonstrated that both of these correlates also apply to sequences of words (see **CarrolConklin2014**, Vilkaitė 2016, for reviews of the processing of MWE in eye-tracking experiments). The focus of our discussion below on vowel duration in adjectives concerns measures of predictability in corpora, but we begin here with a brief review of some of the experimental literature on lexical access of MWEs.

For lexical access, the robust finding is that high frequency words are accessed more quickly than low frequency words (**ScarboroughEtAl1977**; **GardnerEtAl1987** and others). To test whether such an effect is operative in access to MWEs, one must decide on a way to measure the frequency of such expressions. The various options include the token or text frequency of the MWE in a corpus or a measure that takes into account the frequency of the component words, since high frequency words may appear in sequence quite by chance. This measure, usually called *mutual information* (MI, **ChurchHanks1990**) compares the rate at which a pair of words would be expected to occur together by chance in a corpus and compares it to the actual frequency of the bigram. If a bigram occurs more frequently than would be expected by chance, then it might constitute a collocation.

This measure, of course, says nothing directly about the semantic or functional cohesion of the two-word sequence.

Testing strings judged as formulas, **EllisMaynard2008** found that high MI scores speed up the response times of native speakers in various reading tasks. A similar effect appeared in studies of adjective + noun (AN) bigrams and formulaic sequences. **WolterYamashita2017** found that native speakers (and advanced non-native speakers) respond more quickly to high frequency AN collocations than to lower frequencies ones when asked to judge the acceptability of the bigrams. This study compared the effects of the lexical frequency of component words, collocational frequency and MI scores all from COCA. **ÖksüzRebuschat2020** found effects of frequency in both L1 and L2 speakers in an acceptability study. These studies found effects of frequency (especially MI) but did not address the question of whether the collocates tested were associated just by frequency of co-occurrence or if there were effects determined by whether the collocates had special semantic or functional features.

In a set of priming experiments, **DurrantDoherty2010** attempted to distinguish what they call "psychological associates" from collocates that are frequent but have no special semantic relation. The results of a lexical decision task showed priming effects of the "associates" only if they are very frequent. In addition, when the prime was presented for only 60ms, facilitation was found for "associates" only, and not for other collocations. They argue that association and frequency effects are independent (**DurrantDoherty2010**:145).

1.3 Phonetic reduction and probability of occurrence

The other major research thread investigating the role of frequency in the processing of word sequences studies phonetic reduction in spoken corpora (**GregoryEtAl1999**; **JurafskyEtAl2001**; **BellJurafsky2009**; **Seyfarth2014**; **SóskuthyHay2017**). In contrast to experimental studies, corpus studies usually test all bigrams in a corpus (with some restrictions, such as words near pauses or dysfluencies). Frequency is addressed as the probability that a word will occur in a certain context in the corpus and includes token frequency, MI and a set of measures of transitional probability. The latter measures ask what the probability of the target word is given the preceding word or the following word. These measures are computed taking into account the frequency of the two-word sequence in the corpus, which is divided by the frequency of the individual words (see **GregoryEtAl1999**; **JurafskyEtAl2001** for formulae and details). Note that a high frequency word is not as good a predictor of surrounding words as lower frequency words, simply

because it is very likely to occur in a wide variety of contexts. This is especially true of function words, a point to which we return below.

Three of these studies on contextual predictability have found a significant effect on content word duration of the predictability of the target word given the following word: **BellJurafsky2009** found a shortening for words in this context, **Seyfarth2014** found that the association of word duration with informativity ("the average predictability of a word in context") is stronger given the following word, and **SóskuthyHay2017** report similar findings. That is, a content word is shorter if it is predictable from the next word. Given the existing theories of the role of predictability in determining the reduction or lack of it for words in context, this result is extremely puzzling. The influential theory of **Lindblom1990** proposed that speakers are aware of their audience's ability to comprehend words in connected speech and articulate those that are perhaps new or unexpected with greater accuracy than those that are predictable in the context. In a sequence of W1 W2, if W2 is to some extent predictable from W1, W2 could be more reduced because the listener has just heard the first word and can therefore have a chance at predicting the next word. Predictability from the following would mean that W1 can be more reduced because it is predictable from W2. In the context of Lindblom's theory, this is puzzling because the listener has not yet heard W2 and therefore has no basis for predicting W1.

Another theory from **BellJurafsky2009** is that words are reduced if they can be accessed from the lexicon more quickly. This theory has the disadvantage of seemingly conflating two very different processes – access from the lexicon and articulatory production. Even putting that objection aside, the finding that predictability from the following word affects the duration of the target word is still difficult to explain. Why is W1 more easily retrieved from the lexicon if it is predictable from W2? The answer would seem to be that both words are retrieved together, perhaps allowing W1 to be more reduced. This account would then point to a complex lexical representation for bigrams in which W1 is predictable from W2. Why then wouldn't bigrams in which W2 is predictable from W1 also be represented in lexical storage and undergo the same reduction effect?

1.4 The role of constructions

It is important to note that the studies discussed in the previous section are based on bigrams consisting of all types of words without regard for the grammatical constructions in which they are used. Puzzled by the results, **BybeeSouza2019** examined differences in vowel duration in one category – adjectives – in two different constructions, attributive and predicative. Ten adjectives were selected

based on phonological criteria (monosyllables containing a "lax" vowel, ending in /t/ or /d/) and 100 tokens of each adjective (from the Switchboard Corpus, **GodfreyHolliman1993**) were categorized as to whether they occurred in an attributive construction (*hot weather, dead cell phone*) or predicative (*it's so hot, my father is dead).* The duration of the vowel in each token was measured and the association of the vowel duration with a number of factors was examined. These factors included vowel quality (the lower lax vowels /æ/, /ɑ/ and /ɔ/ are longer than the mid vowels /ɛ/ and /ʊ/), voicing of coda consonant (the vowel is longer before a voiced stop), construction type (attributive vs. predicative), articulation rate, position before a pause, token frequency and two predictability measures, predictability given the preceding word and predictability given the following word.

Mixed effects modeling (with speaker and lexical adjective as random effects) revealed highly significant effects, as was expected, of vowel quality, coda voicing, position before a pause and articulatory rate. In addition, construction type was highly significant in predicting vowel duration as was predictability given the following word. Predictability given the preceding word only barely attained significance.[1]

Given the categorization by construction type, the same data can also be examined to determine the role of predictability for vowel duration in adjectives within a construction, a point we return to in Section 3.2. For now, consider a prominent factor that emerges from examining the raw data: adjectives in the two construction have very different contexts in terms of function vs. content words. Attributive adjectives have a very strong tendency to be followed by a noun, while predicative adjectives occur before a function word in 75% of the tokens. This skewing would mean that attributive adjectives are more predictable from the following word than predicative adjectives are because of the high frequency of the function words that tend to follow the predicative use. Also, as we reported above, adjectives in predicative uses tend to have longer vowels than in attributive uses. The higher predictability of attributive adjectives given the following word (usually the noun it modifies) then appears to be associated with vowel shortening, when in fact, it may be the construction type that conditions the shortening.

Given the puzzling finding that higher predictability from the following word leads to shorter word duration, we hypothesize that this result may be an artifact of the constructions that occur in English conversation. It appears that many

[1] A Random Forest analysis (**MatsukiEtAl2016**, see also **TagliamonteBaayen2012**) found lexical adjective and construction type to be the most important variables well above any of the predictability measures.

constructions create structures in which content and function words alternate, as shown in this typical utterance from the Buckeye Corpus (**PittEtAl2007**). Function words are underlined and content words are in italics.[2]

(1) because I *worked* with a *guy* that *was* a *cocaine addict* for a while and he couldn't *have* any *kind* of *caffeine otherwise* he'd *get* the *shakes* and

In this example, the only pairs of contiguous content words are *cocaine addict* and *caffeine otherwise*. In this example we see that there are many constructions that juxtapose function words and content words, such as Det + Noun, Prep + NP, and Aux + V. Constructions that juxtapose more than one content word occur less often. One of them – the AN construction – shortens the V of the first content word (**Morrill2011**). Perhaps other constructions have a similar phonetic effect. If that were so, then the finding that W1 is shorter if it is predictable from W2 might be attributable to the grammatical structure of English and not to predictability per se. A full test of this hypothesis is beyond the scope of this contribution; however, we can present a small pilot study that focuses on sequences of two or more content words in conversation. Five excerpts of approximately 200 words each were taken (randomly) from Buckeye (one of the corpora used by **BellJurafsky2009**). Matching as well as we could the criteria for function vs. content words given **BellJurafsky2009**, we found 85 bigrams consisting of two content words in this 1000 word sample.[3] Given that any stretch of speech has nearly as many bigrams as it has words (words before and after pauses belong to only one bigram while all others belong to two) it is quite revealing that so few bigrams (fewer than 10%) consist of two content words. That means that the vast majority of bigrams in any English utterance have at least one function word. Any measure of predictability for content words, then, is heavily biased towards having a function word as the preceding or following word. A function word context makes the target word relatively unpredictable (given the high frequency of function words). It turns out, then, that high predictability occurs largely in bigrams with two content words.

For this reason, it is instructive to examine what types of constructions are involved in the two-content-word bigrams found in the excerpts. The count revealed that the largest class of bigrams were AN constructions with 24 tokens,

[2] Criteria for distinguishing function from content words follows **BellJurafsky2009**, as discussed below

[3] Analysis was based on the transcript. Five content word bigrams probably occurred across intonation units and one was the result of a disfluency. In some cases bigrams occurred within bigrams, as in *really narrow minded*, which was counted as one bigram for *narrow + minded* and a second one for *really + [narrow minded]*.

or 28% of the bigrams. As numbers were counted as modifiers of a noun, the five tokens with numbers could be added to this, yielding 29 tokens of modifier + N (34%). The second largest class were also within NPs: NN sequences such as *summer class* with 11 tokens. Other bigrams within NPs were sequences of two adjectives (*big huge)* with 4 tokens and one miscellaneous noun modifier. Thus 45 (53%) of the bigrams occurred within a NP. We have already reported that adjectives within a NP have shorter vowels than predicative adjectives; in addition, **Morrill2011** found that in AN phrases and compounds, the vowel of the adjective is shorter than that of the noun. There may also be durational features of interest within NN sequences. While accent is the most important perceptual feature, sequences with compound stress (unaccented second noun) are shorter overall than those with phrasal accent (**Hirst1983**:fn 1) and the second noun is more predictable from the first (**BellPlag2012**).

Less is known about the duration or predictability of the other content word bigrams found in the Buckeye excerpts. Fifteen of these involved an adverb in different constructions; nine were Adverb + Adjective sequences, all in predicative constructions, such as (*was) pretty close,* in which it is very likely that the main stress is on the adjective and the adverb may be shortened. Similarly, in the case of the eight instances of Adverb + Verb sequences, such as *probably change*, the adverb is in a low prominence position and may be shortened. In the remainder of the bigrams, a verb and its argument are involved, as in the Subject + Verb bigram, *grandparents went,* or the Verb + Object bigram, *moving home*, and it is unknown whether shortening would occur in either element

The point of the pilot study is simply to raise the possibility that the relation between vowel or word duration and predictability found in other studies may be influenced by the construction types that have content word bigrams in English. If there is a general shortening of modifiers within a NP (perhaps including NN constructions) then at least part of the effect of predictability given the following word is due to modifiers in NPs and may be less a general principle of English spoken word sequences and more a specific property of certain constructions. Having proposed that the source of the shortening of a word due to its predictability from the following word occurs only in sequences in which both words are content words, future research can determine which constructions contribute to this effect.

2 Research questions

While most experimental studies of multiword expressions focus on perception/-comprehension, we consider the possibility that prefab status may affect the du-

ration of vowels in the constituent words. We have two questions to address:

RQ1: Does prefab status affect vowel duration independently of frequency of occurrence?

Given the discussion just prior, which strongly suggests that constructions may have an influence on both vowel duration and predictability measures, we focus only on AN sequences. As a further control, these AN sequences contain one of 10 adjectives chosen for their phonological shape (see above). Also, considering the lack of independence among predictability measures, such as predictability from the previous or following word, we chose just one – simple bigram frequency – as our predictability measure, that is, the frequency with which the AN sequence occurs in discourse, as measured by its frequency in the spoken portion of COCA.

The study also depends heavily on a workable definition of prefab, which all researchers admit is problematic (**ErmanWarren2000**; **Wray2002**, and others). Thus, our second research question, upon which the first depends, is:

RQ2: Can the various criteria proposed to identify prefabs or formulaic language be applied systematically to a set of AN sequences found in conversational discourse?

As mentioned here and in much of the literature, multiword expressions or formulaic language comprise many different types, which makes definition more difficult (**Wray2002**). By restricting our study to AN sequences, we have eliminated most pragmatic and grammatical prefabs and focus only on lexical prefabs (to use the terms of Erman & Warren [2000]). In the next section we discuss the methods and results of an attempt to hone a set of usable criteria for identifying (lexical) prefabs.

3 Methods

3.1 Study 1: Semantic and functional measures of prefab status

In this section we turn to an examination of MWEs that qualify as "lexical prefabs", and the criteria that can identify them. Some methods of identifying formulas or prefabs include searching published lists for English such as the *Oxford collocation dictionary* (**OxfordCollocationsDictionary2002**). The problems with such collections are that they are not sensitive to dialect differences and they cannot keep pace with cultural changes that create new prefabs (see **DurrantDoherty2010**).

Another approach queries a panel of speakers and asks them to categorize word sequences as formulaic or not, whether the phrase had "a cohesive meaning or function" and whether or not it was worth teaching to second language learners (**EllisMaynard2008**). That study found high agreement among raters on all three parameters (see also **CarrolConklin2014**). **Wray2002** raises objections to the use of native-speaker intuition to identify formulaic language, in particular, that there are no firm boundaries to formulaic language and different speakers may have different experience and judgments. These problems have less to do with particular methods and more to do with the nature of the object of study. For that reason, and for the lack of workable alternatives, our method unabashedly uses the intuitions of one native speaker (the first author) attempting to apply the various criteria that have been proposed in the literature. The result is a qualitative analysis rather than an experimentally induced set of judgments. The analysis has two goals: i) to evaluate the usefulness and applicability of proposed criteria, and ii) to classify a set of AN sequences into those that qualify as prefabs and those that are free pairings, a classification that can be used in the second study, which is quantitative. It is recognized and acknowledged in the analysis that some sequences are marginal and would likely be classified differently by a different analyst.

In the current study, we attempt to apply the criteria suggested by other researchers to 239 distinct AN sequences (types; 336 tokens) found in American English conversation from the Switchboard corpus. As mentioned before, adjectives were selected for this study based on their phonological shape (they are all monosyllables, have "lax" vowels and end in /t/ or /d/) with the additional criterion of sufficient frequency to yield one hundred tokens each in a Switchboard sample. Their selection was random from a semantic or functional perspective. All the adjectives selected were the first token of that adjective in a conversation, to control for duration effects of second mention. As a consequence, almost all sentences were spoken by different speakers.

The adjectives examined are:

(2) *bad, broad, dead, fat, good, hot, mad, red, sad, wet*

As the adjectives were chosen for their phonological shape they have a good range of properties: frequent (*good, bad*) and less frequent (*broad, wet*) adjectives, more or less concrete (*red, wet*) and more abstract (*good, sad*) and some that are predominately used predicatively (*mad*) and some that are often used attributively (*broad, dead, red*). On examination, we also find a range of prefab types appropriate for testing criteria for prefab status.

3.1.1 Application of criteria for AN sequences

Several different criteria for establishing the class of prefabs have been proposed but no one of them hands us all and only expressions considered prefabs. The following analysis is based on two criteria that are often mentioned in other studies: lack of compositionality and fixedness or restricted exchangeability.

Lack of compositional meaning (mentioned in **ErmanWarren2000**; **Wray2002**; **CorriganEtAl2009** and others) is a relatively easy criterion for native speakers to apply. It provides a sufficient reason for inclusion in the category of prefab. Non-compositional meaning is apparent in idioms and other expressions with metaphoric or metonymic meaning (see examples in 3.1.2). In the discussion below we also find special meanings associated with prefabs come from the cultural context (**Hoffman2004**).

Fixedness or restricted exchangeability is noted in **Sinclair1991**, **ErmanWarren2000**, **Wray2002**, **CorriganEtAl2009** and many others as a feature of prefabs. Fixedness is a feature of prefabs that have non-compositional meaning, but it is also a feature of totally transparent prefabs. In our data we have the phrase *broad shoulders*, which has compositional meaning, and yet is more idiomatic than *wide shoulders*.[4] This fixedness is a product of conventionalization, and as noted by others (**Pawley1986**; **ErmanWarren2000**; **Wray2002**), it is gradient, which makes it a difficult notion to apply.

The following sections describe the process by which 68 prefabs were selected from the set of 239 AN sequences. In the first step AN compounds were removed, as their distinct stress pattern affects the duration of the vowel in the adjective (**Morrill2011**). Some examples judged as compounds given the context were *fat days,* and *thin days.* While many definitions of prefabs rely on frequency of use as a criterion, we hoped to distinguish the role of semantic and functional criteria from frequency and therefore did not designate as prefabs AN bigram that had no conventionalization features except frequency of use. This decision does not mean that we do not think frequency is a major factor in conventionalization, rather that we hoped to distinguish its effects from other features. For example, many bigrams with the frequent adjectives *good* and *bad* were excluded. Bigrams such as *bad habit, bad news, bad weather* and *good idea* are relatively transparent, but may be prefabs because of their frequency. However, for this study they were not included as prefabs. The Appendices list the AN bigrams that were analyzed, with those categorized as prefabs in Appendix A and others in Appendix B.

[4] *Wide shoulders* is not impossible in English, but *broad shoulders* occurs 5 times more frequently in COCA than the other combination.

3.1.2 Metaphoric and metonymic bigrams

Starting with the most obvious examples of prefabs first, consider the list of the prefabs with metaphoric or metonymic uses found in the set of AN bigrams.

(3) Metonymic
(in) bad shape, (in) good shape (*shape* is used metonymically)
fat wallet (referring to a person's wealth)
good buys (meaning good value, not just the transaction)

(4) Metaphoric
(in a) bad way
broad daylight, broad humor, broad topic, spectrum and related terms (see 3.1.4)
dead place, dead period, dead week, dead wood, dead zone
hot check, hot point, hot topic, hot water (getting into trouble)
red alert, red badge, red carpet, red flag, red herring, red threat

These MWEs may have other properties of prefabs as well, but their non-transparent status is the clearest indicator of conventionalization.

3.1.3 Semantically transparent prefabs

Many of the prefabs in the sample are semantically transparent but still constitute the conventional way of expressing a notion. In these cases, restricted exchangeability might apply. As mentioned above, an example is *broad shoulders,* which passes the restricted exchangeability test, as *wide shoulders*, which is semantically transparent as well, is nonetheless non-idiomatic. For another example, *dead body,* restricted exchangeability does not easily apply. The expression itself is somewhat redundant, since *body* can mean 'corpse' on its own. However, since *body* is polysemous, *dead body* makes it clear what is meant. It is clearly a conventional way of referring to a corpse, but the criteria used here might not pick it out as a prefab.

This problem raises the issue of how to identify conventionalized expressions. **Pawley1986** describes the practice of dictionary-makers as including "any composite forms that is in common usage, i.e. if it is recognized by members of the language community as a standard way of referring to a familiar concept or conceptual situation". This criterion captures the idea that a prefab can serve as a lexical item and that the formation of prefabs is a lexicalization process (**BrintonTraugott2005**). It also captures the feeling that a prefab or formula can be the preferred means of expression of a concept within a community (**Wray2002**;

CorriganEtAl2009). Of course, the processes of conventionalization and lexicalization occur over time as a compositional phrase is used in a particular context in which it is associated with a particular entity or concept. Partly because the stock of such phrases changes over time, this criterion may also be difficult to apply in some cases. The following examples are found in our data.

(5) *good behavior* (a designation that earns a prisoner certain rewards)
hot fudge (a topping for ice cream)
hot shower (a rewarding, agreeable way to bathe)
mad rush (a crazy, frantic hurry [NB *a crazy rush* is not idiomatic])

Our examples also include a number of prefabs that apply to specific concrete entities, which in many cases are embedded within a cultural context, which endows them with extra meaning. A case in point is *red meat*, a phrase that does not simply mean any meat that is red, but rather designates a host of other properties as well. The following may fall into this category as well: *red beans* (which are actually brown when cooked), *red eyes* (a symptom of a disease), *red light* (a traffic signal), *red pepper* (a specific type of pepper), *red snapper* (a specific type of fish), *red spider* (a specific type of spider), *red wine* (a whole category of wine), *wet lab* (a lab outfitted to deal with hazardous chemicals) and *wet rag* (a handy thing to have on hand in the kitchen or elsewhere).

As mentioned above, we omitted many bigrams with *good* and *bad* that are transparent and probably count as prefabs due to high frequency. Some of these might also be conventionalized according to the criteria of designating a particular culturally salient concept, but it is very difficult to make that decision. Examples are *good experience, good reputation, good weather, bad experience, bad habit, bad influence.* Note, however, that most of these do not meet the "restricted exchangeability" criterion as *excellent* can be substituted for *good* and *terrible* for *bad* in these examples. It may be that such examples fall in the murky territory between prefab and free choice.

Two other bigrams lean more towards prefab status: *sad statement* and *sad commentary*. These two are interesting because they show some exchangeability in the noun, with *sad commentary* being much more frequent than *sad statement* (according to the COCA). Their prefab status emerges because there is not always an explicit statement or commentary in the context, and especially for *sad statement,* the meaning is about the same as *sad* alone, for example in the follow excerpt from COCA (**Davies2008**-):

(6) but it's a sad statement how little optimism I've seen so far

A few of the prefabs we identified were part of a larger prefab as we saw in the metonymic *in bad shape, in good shape* and the metaphoric *in a bad way.*

3.1.4 Prefabs as expandable

As we have seen, restricted exchangeability is not a property of some MWEs that appear to be prefabs. In fact, a number of diachronic studies have shown that many prefabs retain the analyzability of their internal structure and enough transparency of meaning to be expanded or in some cases rearranged for greater transparency. Prefabs figure in the creation of new constructions, as **Wilson2009** demonstrates for 'become' verb + adjective constructions in the history of Spanish. In that case, a single instance of a verb and adjective, *quedar solo* 'to be left alone', expands to use with other adjectives with similar meaning, becoming productive in certain semantic domains. **Bybee2014** shows that what appears to be a very fixed expression to indicate extreme poverty, as in *He hasn't got two nickels to rub together* (**COHA1971**), can expand in a variety of ways, but primarily in the NP *two* + N, where *coins, dimes, quarters, beans* and other nouns can occur. **BybeeModer2017** trace the changes in the prefab *beg the question* from its use to designate a type of fallacy in reasoning through the 15th to 19th centuries to a new interpretation now common in which it means 'raise the question'. In the latter use, *question* can be in the plural or modified, as in *So those high scores beg the politically incorrect question, are Asians naturally more intelligent?* (COCA).

In the following, we discuss two sets of AN sequences that seem to have prefab status and yet are not isolated, but rather incorporated into clusters that one might designate as small constructions.

3.1.4.1 Good + quantity

Because our sample was constructed around certain adjectives, we found a set of nouns forming a small construction-like expression with *good.* All the nouns designate a quantity and *good* indicates that this quantity is relatively ample. No other adjective can be substituted for *good* in this expression with the same meaning and the set of nouns used seems to be conventionalized. Here are the examples found in the data we examined.

(7) *good while*
good portion
good size
good bit

good deal
good money

3.1.4.2 Broad + noun

A further question that arises in the data concerns how to analyze the AN sequences with *broad.* Many of these clearly qualify for prefab status based on relative frequency, restricted exchangeability, or the feeling that the adjective "goes with" a particular noun. The two most frequent bigrams in our data with *broad* are *topic* (occurring 12 times in 100 tokens) and *spectrum* (occurring 8 times).[5] In both cases, substituting the near-synonymous adjective, *wide*, yields a much less idiomatic phrase. Some synonyms of *topic* also occur, such as *subject* (4 times) and *question* (twice), and these also are less idiomatic with *wide.* Another noun that occurs was *issue,* and while *broad question* seems to reach prefab status, *broad issue* perhaps does not. But here it becomes very difficult to distinguish prefab from free combination.

An alternative approach would be to attribute the choice of nouns to the meaning of *broad* rather than to a more arbitrary conventionalization process. Perhaps an even better alternative is to take into account three factors in explaining the selection of *broad +* noun. One would be the meaning of *broad*, determined by its prior usage with different nouns, a second factor related to that would be how similar a noun is in meaning to prior usage, and a third would be the frequency with which it is used with particular nouns, as indicative of the conventionalization of the AN sequence.

A way of incorporating these factors into an analysis would be to consider prefabs in clusters surrounding a high frequency prefab, much the way **BybeeEddington2006** analyzed combinations of verb + adjective in Spanish. The usage of *broad* (and the meaning derived from it) might consist of several related clusters. Based on the data we are considering here we can propose the following clusters:

- The center is *broad topic* and the related nouns are *subject* and *question* and the more marginal noun *issue* (not judged to be a prefab), and perhaps others that did not occur in the sample.
- The center is *spectrum* and related nouns are *range, scope* and *coverage.* These nouns refer to an abstract continuum.

[5]The Switchboard corpus was assembled from telephone conversations between volunteers who were given certain topics to discuss and they often make comments about the topic.

- In a third cluster the center is more difficult to discern, but the related prefabs describe a person's attributes and include *broad background, education, tastes* and *views,* all of which appear to be prefabs, and two more marginal members (which were judged not to be prefabs) *strengths* and *interest.*
- Related to *broad topic,* a fourth cluster would include *broad category, term* and *sense.* Two bigrams that were not judged to be prefabs were *broad definition* and *broad word.*

This rough analysis based on some 60 tokens of *broad* can be taken as an example even though other speakers or analysts might categorize more marginal types differently. It serves as a way of underscoring the fact that not only are prefabs syntactically analyzable, they are also not completely isolated in the lexicon, but come with many connections to other combinations. These facts make it more difficult to distinguish prefabs from more novel expressions.

As for the criteria one can use for identifying prefabs, the discussion calls into question restricted exchangeability as proposed by Erman and Warren. We also suggested another criterion, the concept of "cultural salience" as proposed by various authors (**Pawley1986** and **Hoffman2004**). This criterion would identify prefabs based on their designating a unique cultural concept, entity or situation. As noted in other works, no one criterion sets the boundaries of prefabs; rather a set of criteria working together comes closer to the goal.

3.2 Study 2: Vowel duration, prefab status and bigram frequency

BybeeSouza2019 showed that construction type affects the duration of the vowel in the adjective. Having argued here that predictability measures are heavily influenced by construction type, in Study 2 we analyze the effects of prefab status on adjective vowel duration. Van **LanckerCanter1981** demonstrated that speakers can produce phonetic differences that make it possible for listeners to distinguish the literal from the idiomatic meaning of word sequences, though the authors do not explain what phonetic features are involved. We hypothesize that vowel duration may be one of the phonetic differences that distinguish literal from idiomatic meaning. That is, the vowel in an adjective is shorter when that adjective is part of a prefab than when it occurs in a free pairing.

For this study, AN phrases were extracted from conversations in which speakers were seemingly making no special attempt to signal any specific type of meaning. Our goal is to determine if phonetic effects of prefab status occur even in these circumstances. The results may be quite subtle, because of a number of

factors: (1) our earlier study showed that the amount of variation in attributive adjectives is reduced compared to predicative adjectives, (2) the test is based on a small number of tokens (334 tokens suitable for analysis), (3) many other factors are at play: inherent vowel duration and final coda voicing, and (4) a large proportion of the prefabs have *broad* as the adjective and it contains the longest lax vowel.

To compare the effect of frequency on vowel duration, we also included bigram frequency as a separate variable in the analysis in order to determine whether prefab status has an effect on vowel duration that is independent of frequency. We opted for bigram frequency as the most straightforward measure of co-occurrence in a corpus. Bigram frequency scores, obtained from the spoken section of the COCA corpus, were then coded as high (21 or greater) or low (20 or lower).

The duration results are shown in Figure 1.1 below.

Figure 1.1: Comparison of vowel duration in adjectives in adjective-noun sequences given prefab status. The graphs on the left show raw values in milliseconds; log-transformed values appear on the right. Red boxes contain data for prefabs, blue boxes show data from free AN pairings.

Vowels in prefabs (shown in red in the graphs) were overall shorter than in free pairings (in blue), confirming our predictions. However, this difference failed to reach statistical significance in a Kruskal-Wallis test for categorical variables ($\chi^2 = 0.41, \mathrm{df} = 1, p = 0.52$). Since the data depicted in Figure 1.1 includes AN sequences of both high and low bigram frequency, it is possible that frequency may be influencing these results, especially for high frequency free AN pairings, which might have sorter vowels. The data shown in the interaction plot in Figure 1.2 below suggest that prefab status and bigram frequency play separate roles in determining vowel duration in our data.

Figure 1.2 suggests that the phonetic effects of prefab status may indeed be independent of bigram frequency scores, with the greatest difference occurring between low-frequency prefabs and low-frequency free AN pairings (top right). Even though this difference in duration failed to reach statistical significance, the trends in the distribution of the data warranted further exploration.

We then analyzed only the 182 low-frequency (i.e. bigram frequency of 20 or lower) AN sequences in an attempt to obtain a more detailed understanding of the differences in Figure 1.2. Thus, excluding high frequency sequences, we find once again that prefabs are overall shorter (mean = 120ms, SD = 44ms) than free AN pairings (mean = 124ms, SD = 51ms), though this difference is not statistically significant. This trend is depicted in Figure 1.3 below.

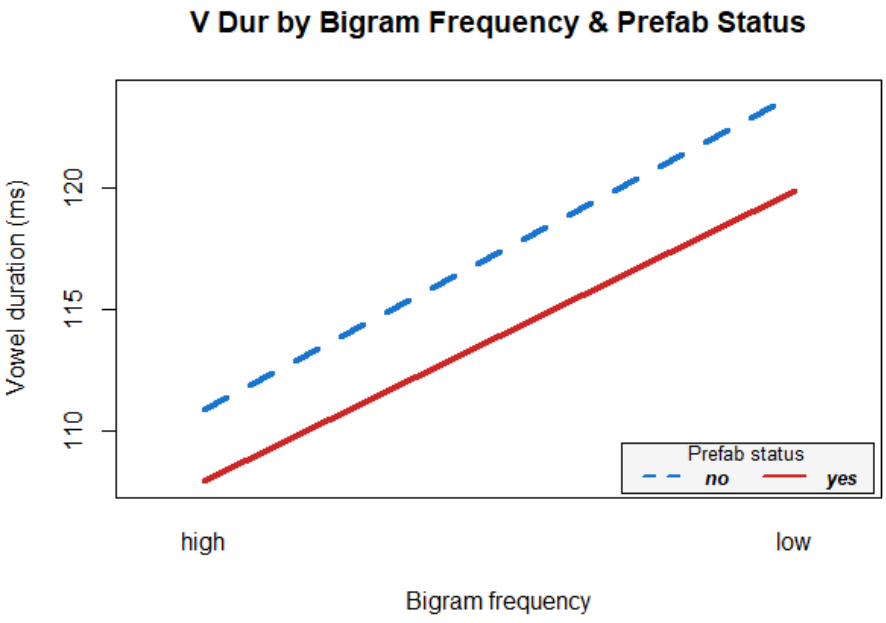


Figure 1.2: Vowel duration (raw values in ms) given prefab status and bigram frequency score distribution. Prefab data represented by the red line; free AN pairings represented by the dotted blue line.

Figure 1.3: Vowel duration in low-frequency AN sequences by prefab status. The graphs on the left show raw values in milliseconds; log-transformed values appear on the right. Low-frequency prefabs are shown in yellow; low-frequency free AN pairings shown in gray.

The current comparison indicates that the factors that lead to conventionalization of word sequences also play a role in determining the phonetic properties of words and constructions, one which may be independent of frequency (as suggested by **DurrantDoherty2010**). Although more data are necessary to strengthen the current findings, our results indicate that the lexical cohesion found in prefabs may lead to phonetic fusion and reduction in production. A larger number of AN tokens coded for prefab status may reveal that the trends we observed in our small sample in fact show a difference in duration that is statistically significant.

4 Discussion

4.1 Phonetic indicators of prefab status

The results reported here show that it might be productive to investigate further possible phonetic correlates of conventionalization. The many corpus studies cited in Section 1.4 used measures based on frequency and predictability. The phonetic variables investigated were word duration (which included lexical as well as phonetic differences) (**BellJurafsky2009**; **Seyfarth2014**; **SóskuthyHay2017**), t/d deletion in English, vowel reduction and duration (**JurafskyEtAl2001**). The more specific reduction processes specific to English – reduction of coronals and vowels – might be good candidates for phonetic features distinguishing prefabs. Our results suggest that vowel duration may be affected by prefab status. Of course, measures based on frequency are more objective and easier to calibrate than a measure based on prefab status, but numerous studies now show that prefab status may be discernible, especially when the data are restricted to a single construction

4.2 Does the adjective predict the noun or the noun predict the adjective?

Coming back to the discussion in Section 2 about the directionality of predictability, where we discussed the finding that for English the duration of a word is associated with its predictability from the following word, we can consider with AN sequences what it would mean to say that W2 predicts W1. As mentioned earlier, the studies, such as **BellJurafsky2009**, that find predictability from the following word is significant in predicting word duration, do not take into account construction type. Their conclusions, therefore, refer only to general processing mechanisms. Our proposal is that truly understanding the predictability data will require examining particular constructions. As AN constructions make up a major portion of bigrams consisting of two content words, they present a good starting point for understanding predictability relations.

In a AN attributive construction, the noun is considered the "head" syntactically and it also functions as "head" or grounding point semantically. When confronted with an adjective, a language user will not know how to interpret it until the noun is also known, as the abstract meaning of the adjective can only be made concrete when it is applied to a noun. A finding in **ÖksüzRebuschat2020** supports this supposition. They asked L1 and L2 speakers to judge whether English AN bigrams were in common usage or not, and measured their reaction times.

High frequency AN bigrams were judged more quickly than low frequency bigrams by both L1 and L2 speakers. In addition, the frequency of the noun affected the reaction time, with faster reactions for high frequency nouns. In contrast, the frequency of the adjective had no effect. These findings suggest that interpreting an AN sequence depends heavily upon accessing the meaning of the noun.

Viewed from another perspective, one could argue that nouns are associated with a restricted range of adjectives. Peppers come in black, white, red, green, yellow, hot, sweet or mild. Wine is red, white, or sparkling. A spectrum is broad or narrow. A while is good, long or short. To some extent it could be argued that the reverse relation also occurs, because *broad* could be said to predict certain nouns, such as *topic, spectrum* or *education.* Nouns selecting adjective may be more common than the reverse, but that has to be a topic for future research

5 Conclusion

The current contribution attempts to combine and reconcile quantitative and qualitative analyses, in a search for the factors influencing the properties of words in combination. We argue that predictability measures can be best understood in terms of particular constructions, rather than as general processing mechanisms. Within particular constructions, the relations between words are varied and may depend more on grammatical and functional factors than on frequency of co-occurrence (see **BellPlag2012** for noun-noun sequences). The hypothesis of this chapter is that certain phonetic distinctions, in particular, reduction of vowel duration, corresponds to conventionalization found in lexical prefabs. Our main finding is that prefab status affects vowel duration in adjectives independently of the frequency of the AN bigram. This result suggests that phonetic factors may in the future provide a diagnostic for the elusive notion of conventionality of expression.

A qualitative assessment of the relation between two content words in sequence is necessary to understanding the role of predictability in processing and storage. An attempt to determine the prefab status of AN bigrams reveals some problems in the criteria proposed in the literature. Examining a set of adjectives from conversation, two important properties of prefabs have emerged: one is that prefabs may be semantically compositional while referring to entities or concepts of some cultural importance and therefore constitute the conventional means of referring to these concepts. The second is that prefabs are not always isolated, but in some cases form clusters of semantically related word sequences. Thus, existing prefabs can be used to spin off new creative combinations demonstrat-

ing again how the conventional and the novel form a continuum and interact in language use.

Appendix A Adjective-noun sequences classified as prefabs

Adjective	Noun(s)
bad	*shape, way*
broad	*topic, question, shoulders, background, scope, subject, range, spectrum, humor, daylight, views, coverage, sense, term, education, category, tastes*
dead	*week, body, zone, bodies, wood, period, place*
fat	*wallet, dude*
good	*bit, deal, money, size, behavior, while, buys, condition, shape, portion*
hot	*check, shower, water, topic, fudge*
mad	*rush, dog*
red	*kidney beans, meat, herring, wine, meat, flag, eyes, threat, carpet, alert, lobsters, China, beans, light, snappers, pepper, spider*
sad	*statement, way, commentary*
wet	*labs, rag, rags*

Appendix B Adjective-noun sequences classified as free combinations

Adjective	Nouns
bad	*test, types, news, stuff, aspects, experience, thing, speller, cholesterol, team, mayor, points, things, example, weather, habits, signs, sides, knees, influence, meal, habit*
broad	*one, disparities, section, overstatements, word, interest, way, definition, segment, strengths, issue*
dead	*bride, cat, doctors, Iraqi, sport, birds, presidents, poets, shrubs, cat, people, cell phone, protein*
fat	*lady, women, intern, kid, American, Americans, dog, burger, people, kind, man, person, one, ones, kids, boy, cat, days*
good	*trees, movie, area, impact, shows, team, lunch, retirement, movies, idea, information, state, company, French fries, ones, nursing home, summer, ones, question, lesson, way, program, balance, reputation, job, experiences, point, player, weather, weekend, loan*
hot	*events, foods, jacuzzis, summertime, days, things, ones, weather, points, one, water, air, afternoons*
mad	*one*
red	*hat, dirt, insect, suspenders, dress, cat, Doberman, berries, sign, one, dye, ones, streak, house, strips, clay, salmon, Ferrari, badge, day, felt, face, star, birds, house, lightning, shirt, rats, fescue, azaleas,*
sad	*part, thing, things, joke, year*
wet	*cold, dry, food, kind, part, weather, stuff, cloth, June, hair, feed, heat, backyard, mess, towel, sponge, springs, paper towel*

Chapter 2

Cascading collocations: Collocades as correlates of formulaic language

Richard Forsyth

This chapter focuses on a technique for detecting, measuring and displaying traces of formulaic language. For this purpose, a suite of computational procedures has been developed in order to quantify the degree to which individual texts and text types incorporate inflexible sequences of words. This development is predicated on the assumption that, even if we have no precise definition of formulaic language, it is widely accepted that it is characterized by repetition of fixed sequences. The method involves compiling a *formulexicon* from a corpus of two or more text types and then using coverage by elements of that *formulexicon* as an index of the degree to which a text, possibly absent from the training corpus, is pervaded by formulaic sequences. The problem of deciding what lengths of *n*-grams are warranted by the data is dealt with by the simple expedient of binarizing coverage counts by *n*-grams of various lengths. Trials on a variety of text types show that this allows *collocades* – cascades of collocations, whose lengths are not pre-determined – to emerge from the data. Here the term *collocation* is used in its broader sense, as in "collocations are co-occurrences of words" (**Gries2009**). Software in Python 3 that implements this approach is available online under a Creative Commons licence. Examples of applying these procedures to a number of corpora illustrate some of the uses of this approach.

1 Introduction

Many linguists have celebrated the "unlimited creative potential" of human language (**Eggins1994**). Ron **Carter2004** has argued that creativity is an all-pervasive feature of everyday language, a point also emphasized by Chomsky.

> The normal use of language relies in an essential way on this unboundedness, on the fact that language contains devices for generating sentences of arbitrary complexity. Repetition of sentences is a rarity; innovation, in accordance with the grammar of the language, is the rule in ordinary day-by-day performance. (**Chomsky1972**)

On the other hand, others have noted the "deadly repetitiousness of language" (**Bolinger1965**). This refers to the fact that speakers and writers tend to reuse chunks of language, perhaps with slight variation, a phenomenon dubbed by **Sinclair1991** "the idiom principle". Presumably this reflects a natural tendency to save mental effort. As **Halliday2014** puts it, "repeated patterns require less brain power both to produce and to understand."

This apparent contradiction points to a dimension on which examples of language use can vary widely, from creative to routine. Research into *formulaic language* is at least in part an attempt to explore this polarity. However, the term refers to a wide variety of linguistic phenomena. Essentially, formulaic language is a negative concept: we recognize it when the creative potential of ordinary language, celebrated by Carter, Chomsky, Eggins and others, appears to be partially or completely restricted. Such restriction can happen for diverse reasons, which helps to explain why no precise, agreed definition exists of what exactly constitutes formulaic language, although many researchers are actively engaged in studying its manifestations, such as idioms, clichés, legal boiler-plate and apparently prefabricated lexical bundles.

It should perhaps be noted that, in contexts such as second-language learning, the use of multi-word units is sometimes viewed in a positive manner, as a sign that the learner is gaining phraseological competence. For instance, **GrangerBestgen2014** and **Leńko-Szymańska2016** explore the use of statistics relating to multi-word units as potential indices of learner competence. In this case the ability to deploy word groupings is evidence that the learner can operate with higher-level chunks.

In any case, whether or not formulaic language is viewed pejoratively or as valuable, we do not have a widely accepted method of assessing just where on the polarity from creative to formulaic a given text or corpus lies. A major objective of the present chapter is therefore to describe a computable index of linguistic flexibility/inflexibility which could serve to indicate of the degree to which a text or text type exhibits formulaic language. This problem, namely to what degree a particular kind of language is formulaic, is one of the key questions in the field (**Wray2002**), one that has not been answered in a comprehensive manner.

Strictly speaking, without a definition of our key term, it should not be possible to measure the extent to which a given text or speech is formulaic. Nevertheless, the present chapter sets out to describe, and apply, procedures that are designed to provide researchers with a quantitative index to associate with more impressionistic judgements; and to help identify sections within texts that can be further scrutinized as embodying mainly prefabricated segments. This development is predicated on the assumption that, even if we have no precise definition of formulaic language, it is widely accepted that it is characterized by repetition of fixed sequences.

In short, we don't know exactly what formulaic language is, and suspect that it has multiple causes, but we will attempt to measure it anyway. This attitude isn't quite as unscientific as it might seem. One can compare the situation in biology, for example, regarding the crucial concept *biodiversity*. Assessing biodiversity at various locations and trying to estimate the biodiversity of planet earth is something that scientists and many lay people agree is a matter of grave importance, although there is no single method for measuring it. Nevertheless, different researchers have proposed, and refined, a number of ways of quantifying the diversity of life forms in various habitats, which, between them, have helped to advance knowledge in this area (**Magurran2004**); consequently we can do better than mere guesswork when assessing whether and where biodiversity is increasing or decreasing.

In the sphere of linguistics, many studies have explored the phenomenon of collocation in a general sense, and several techniques have been developed to seek examples in corpora, using a variety of terms such as multi-word units, lexical bundles and others (e.g., **ShimohataEtAl1999**; **ZhangEtAl2009**; **KilgarriffEtAl2012**). Moreover, programs exist that are generally available, for example kfNgram (**Fletcher2012**) and Wordsmith Tools (**Scott2020**), which automate some aspects of the search for co-occurring linguistic units. These tools and techniques are primarily aimed, however, at throwing light on the linguistic behaviour of the speakers or writers of the texts in question. In other words, the multi-item units discovered are regarded as results in themselves. Some researchers also limit themselves to pre-specified grammatical functions, such as noun phrases (e.g. **Daille2003**; **ZhangEtAl2009**) and thus presuppose reliance on ancillary parsing or tagging software. For this reason they are difficult or impossible to adapt to the main purpose of the present approach, which is to quantify the pervasiveness of such multi-item units in various texts and/or text types.

The method described in this chapter involves compiling a *formulexicon* from a corpus of two or more text types and then using coverage by elements of that *formulexicon* as an index of the degree to which a text, possibly absent from the

training corpus, is pervaded by formulaic sequences. The problem of deciding what lengths of *n*-grams are warranted by the data is dealt with by the simple expedient of binarizing coverage counts by *n*-grams of various lengths. Trials on a variety of text types show that this allows *collocades* – cascades of collocations, whose lengths are not pre-determined – to emerge from the data. The extent to which a text is covered by such collocades can be quantified as an index of the degree to which that text is formulaic. Software in Python 3 that implements this approach is available online under a Creative Commons licence. (See Appendix.)

2 The formulib suite

To illustrate this approach, a small test corpus, consisting of seven subcorpora, has been compiled, briefly described in Table 2.1.

Table 2.1: Test corpora

Short name	Description
BEER	Texts from the back labels of beer bottles
EW	Short stories by Edith Wharton (1862–1937)
FEWREPS	Postings on Hong Kong Tripadvisor travel forum (2016) with fewer than 2 replies
LEAFLET	Information leaflets of medicines
MANYREPS	Postings on Hong Kong Tripadvisor travel forum (2016) with more than 10 replies
SRES	United Nations Security Council resolutions, 1999–2004
WINE	Texts from the back labels of wine bottles

Basic details of the numbers and sizes of these text collections are described in Table 2.2.

In Table 2.2 text lengths are given in tokens, which are normally words, although numbers, i.e. strings of numerals, also count as tokens. It will be seen that many of these documents are individually very short. The Hong Kong TripAdvisor postings can be as small as 17 tokens in length. The above texts are all in English. The system has been used with other languages, including Chinese, and can be applied to any language than can be encoded in Unicode (utf-8).

The suite of programs in Python 3 that constitute the formulib package and their functions are summarized in Table **??**.

Table 2.2: Sizes of text corpora

Short name	Texts	Tokens	Smallest	Median	Longest
BEER	118	15781	56	129	314
EW	44	365158	2271	8228	15682
FEWREPS	213	14898	17	56	359
LEAFLET	461	482373	180	946	5251
MANYREPS	610	65494	17	80	1468
SRES	275	248676	102	635	5452
WINE	86	11474	57	125	296

3 The trouble with *n*-grams

The initial program of this suite, outgrams, is relatively conventional. Its main output, referred to here as a *formulexicon*, is a list of the N most frequent *n*-grams for each category of the input text files, with $N = 80$ by default. A segment of its output for the SRES category, when the minimum and maximum *n*-grams lengths were specified as 3 to 6 (tokens), follows. The software allows punctuation to be ignored or preserved, and, independently, for upper case to be preserved or folded to lower case. In this and subsequent examples the option of removing punctuation was chosen, and lower case was enforced. Among other things, this illustrates why *n*-gram lists, in and of themselves, are not particularly enlightening.

```
# sres 154799 978607

1 (6, 175, 34, ('adopted', 'by', 'the', 'security', 'council', 'at'))
2 (6, 175, 30, ('by', 'the', 'security', 'council', 'at', 'its'))
3 (6, 153, 36, ('the', 'democratic', 'republic', 'of', 'the', 'congo'))
4 (6, 130, 33, ('the', 'charter', 'of', 'the', 'united', 'nations'))
5 (6, 121, 35, ('the', 'report', 'of', 'the', 'secretary', 'general'))
6 (6, 97, 28, ('of', 'the', 'charter', 'of', 'the', 'united'))
7 (6, 95, 36, ('decides', 'to', 'remain', 'actively', 'seized', 'of'))
8 (6, 93, 36, ('remain', 'actively', 'seized', 'of', 'the', 'matter'))
9 (6, 92, 32, ('to', 'remain', 'actively', 'seized', 'of', 'the'))
10 (6, 87, 34, ('report', 'of', 'the', 'secretary', 'general', 'of'))

[.... many lines omitted ....]

1 (3, 624, 21, ('the', 'secretary', 'general'))
2 (3, 579, 18, ('the', 'united', 'nations'))
```

Table 2.3: Programs of formulib and their functions

Program	Function
`outgrams`	The main input of this program is a collection of text files, normally divided into more than one text category. It compiles the components of a *formulexicon* by finding the most frequent n1-grams up to n2-grams in two or more categories of text files where n1 is 2 and n2 is 5 by default.
`formulex`	This program takes a collection of text files of more than one category, typically the same collection as input to outgrams, and applies the *formulexicon* already generated to compute coverage by collocades for each file and thereby indicate the extent of formulaic language in it. The program also produces a list of collocades, multi-word units whose lengths are derived from the data rather than being specified in advance (as will be explained below).
`taverns`	This program (Textual Affinity Values Employing Repeated N-gram Sequences) computes coverage of two or more categories of document not only by the collocades generated from their own category but by those of the other categories as well, thus identifying highly typical and highly atypical texts in each class. It also functions as a classifier by using coverage by collocades from all categories, to indicate likely category membership. Usually its input will be a holdout sample of text files which were not used in generating the *n*-gram lists.
`flicshow`	This program produces a colour-coded FLiC list (Formulaic Language in Context) designed to show how the collocades are distributed in various texts. (See section 5.)
`postflab`	This program takes a secondary output of formulex, the collocade list, and applies a 1-dimensional similarity scaling procedure to them, so that related sequences can be plotted in a way that reveals their inter-relationships. (See section 7.)

```
3 (3, 520, 20, ('the', 'security', 'council'))
4 (3, 319, 17, ('the', 'government', 'of'))
5 (3, 312, 13, ('of', 'the', 'united'))
6 (3, 245, 16, ('of', 'the', 'secretary'))
7 (3, 243, 20, ('secretary', 'general', 'to'))
8 (3, 237, 18, ('in', 'accordance', 'with'))
9 (3, 207, 21, ('the', 'implementation', 'of'))
10 (3, 202, 22, ('requests', 'the', 'secretary'))
```

Here we have the ten most frequent 6-grams and the ten most frequent 3-grams derived from the SRES subcorpus. The top line indicates that these *n*-grams are based on 154,799 tokens which amount to 978,607 characters. This is less than the size given in Table **??** because the full dataset has been split randomly into training and test sets, of 1117 and 690 text files respectively.

The first two 6-grams both occur 175 times in this corpus. The first is 34 characters in length, including blanks between tokens, and the second is 30 characters long. It seems a fair inference that they arise from a 7-gram, namely "adopted by the security council at its", but that is hardly obvious from the listing. Likewise, if we take items 7, 8 and 9 together, which occur 95, 93 and 92 times respectively, with a bit of background knowledge, we might arrive at the phrase "decides to remain actively seized of the matter", an 8-gram with which many of these Security Council resolutions sign off. However, this also is not immediately obvious from the listing (which is in fact designed more to be read by other programs than by people).

The point applies also with the shorter *n*-grams. The first three 3-grams (with the benefit of background knowledge) would seem to be natural units in their own right. But along with them, we find "secretary general to" and "requests the secretary", which are fragments of longer phrases.

What this illustrates is that in a standard frequency list of fixed-size *n*-grams, such as that above, longer *n*-grams tend to appear in the form of multiple fragments. It requires tedious inspection, along with background knowledge, to identify appropriate lengths for the fragmented pieces of repetitive phrasings. It is desirable for the computer to provide more help in that identification process.

4 From *n*-grams to collocades

The formulex program is designed for this purpose. The basic idea behind this program is very simple. The problem with *n*-gram lists is that they tend to contain multiple fragments of longer sequences, losing track of what might be considered the natural length of the sequences from which they are derived. So formulex

tries to put them back together by going back over the original texts to find out exactly which passages are covered by the items in the frequent *n*-gram list. The key concept here is *coverage*. The important point is that a text sequence is either covered or not: the number of *n*-grams that match a particular sequence of tokens doesn't matter, just whether any do or none.

To give an illustration of the covering process, suppose that you have gathered a corpus of political propaganda in which the phrase

> securing a better future for hardworking families

is repeated ad nauseam.

This could be regarded as a frequent 7-gram, but with the system's default settings, the longest *n*-grams to be saved will be 5-grams. Thus the *n*-gram list would probably include

> securing a better future for
> a better future for hardworking
> better future for hardworking families

as well as shorter subsequences, probably going down to 2-grams such as "a better", "better future" and "hardworking families". Suppose further that the program is processing the sentence shown in Table **??** (tabulated vertically, for convenience).

To keep things manageable, 4-, 3- and 2-grams have been ignored. This might well increase the totals in the column labelled "match count", but the main point is that coverage will be determined by whether this figure is greater than zero or not. The total number of matches isn't taken into account for this purpose. (It can be used in flicshow, as described in section 5.)

Sticking just to these 13 words (87 characters, including single spaces between words) and three 5-grams, the coverage would be 48/87 characters and 7/13 tokens, i.e. just the words that have a nonzero entry next to them under "match count". This would appear as 55.17% and 53.85% in the output. These two percentages tend to be highly correlated, meaning that similar conclusions are likely to be drawn from either. Character-coverage is placed first because I believe it is likely to be a slightly more sensitive indicator.

To summarize, the program works out coverage of tokens in this manner for each file separately using the *n*-grams from the same category as the text concerned (or the largest category if the text has an unseen class label) and also aggregates the coverage for each category. The texts are listed in descending order of character coverage.

Table 2.4: Computing coverage in formulex

(word) token	match count	covering *n*-gram(s)		
we	0			
are	0			
committed	0			
to	0			
securing	1	securing		
a	2	a	a	
better	3	better	better	better
future	3	future	future	future
for	3	for	for	for
hardworking	2		hardworking	hardworking
families	1			families
throughout	0			
britain	0			

The beginning of the main output file resulting from processing by formulex of a training sample of 1117 texts from the seven text categories described in Table 2.1 is shown below. In this case the formulexicon generated by outgrams contained the most frequent 3- to 6-grams from each category.

```
Wed Oct 16 15:28:36 2019
parafile: C:\keywork\parapath\grabchap.txt
metafile: c:\keywork\mets\grabchap_met1.dat
miniglen: 3
maxiglen: 6
topgrams: 80
1117 7
Category coverage % (characters, tokens) by frequent n-grams :
0 EW                 3.1528    4.0587
1 beer              23.6432   23.4533
2 fewreps            9.1539   10.3654
3 leaflet           13.3395   14.9340
4 manyreps           8.2614    9.6554
5 sres              17.6441   18.8619
6 wine              20.0487   20.2816
```

```
Document coverage % (characters, tokens) by frequent n-grams :
   1    526    85   57.50  58.82  beer    low_alcohol_czech_lager.
      txt
   2    607   103   49.34  49.51  wine    fabcab.txt
   3    789   136   48.86  47.79  beer    island_hopper_pale_ale.txt
   4    596   102   48.07  47.06  sres    S_RES_13212000-en.txt
   5    613   102   47.07  46.08  beer    ruddles_best.txt
   6    188    39   46.03  43.59  fewreps
      poor72995652_9302772_Getting_to_Kai_Tak_Cruise_Termin.txt
   7    940   154   45.70  42.86  sres    S_RES_15042003-en.txt
   8   1761   290   43.25  45.17  sres    S_RES_15552004-en.txt
   9   1030   165   42.97  46.06  sres    S_RES_14892003-en.txt
  10    237    45   42.86  46.67  manyreps
     good75047818_9504892_First_visit_to_hong_Kongwhere_t.txt
  11    525    96   42.78  40.62  wine    long_slim_chile.txt
  12    645   112   42.72  41.07  beer    youngs_bitter.txt
  13    543    90   42.10  42.22  beer    corona_extra.txt
  14    618   104   42.00  42.31  wine    coop_chilean_merlot.txt
  15    718   118   41.72  40.68  sres    S_RES_15052003-en.txt
  16   1259   193   41.67  44.04  wine
     lime_tree_cabernet_sauvignon.txt
  17    782   131   41.51  43.51  sres    S_RES_14852003-en.txt
  18    780   139   40.85  40.29  beer    battersea_rye.txt
  19    889   149   40.67  40.27  beer    salts_burton_ale.txt
  20    717   111   40.67  42.34  wine    coop_chianti_2013.txt

[.... many lines omitted ....]
```

According to either character-coverage or token coverage, the categories can be ranked from most to least as follows: BEER, WINE, SRES, LEAFLET, FEWREPS, MANYREPS, EW.

If we accept *n*-gram coverage as illustrated in Table **??** as an index of formulaicity, the most formulaic individual text is the beer back label for low alcohol Czech lager. The numbers on the line preceding that item indicate that this text consisted of 526 character containing 85 tokens. The 3- to 6-grams of the beer back label formulexicon covered 57.50% of the whole text in terms of characters, and 58.82% of its tokens.

Only texts from five of the seven categories appear in this top 20. The patient-information leaflet with the highest coverage was ranked 44th and the tale by

Edith Wharton with the highest coverage came in at position 918 with scores of 4.65% and 5.59% – in the last 100 of 1117 items. Clearly literary fiction does not contain long slabs of prefabricated language.

5 Collocades in context, and in colour

The formulex program identifies which text types are most pervaded by repetitive sequences, and thus most likely to contain a high level of formulaic language. It also identifies individual texts that are high or low in coverage by repetitive sequences, but does not identify which sequences occur where.

After examining the output and particularly after noting texts that are particularly high or low in collocade coverage, an analyst will naturally want to know more about which collocades are responsible for such differences. This is the purpose of the flicshow program, which is intended to show formulaic language in context.

It takes as input the formulexicon file produced by outgrams and applies it to a list of specified files. As output it writes a tokenized version of each input file in html. These outputs can be viewed in a browser such as Edge, Chrome or Mozilla Firefox.

In these output files the portions covered by the *n*-grams of the relevant category (i.e. the category of the text being processed, or the largest category if it has an unknown category label) are highlighted in colour, while the rest of the text is printed in black. Each token is written in a colour that depends on the length of the longest *n*-gram by which it is covered (by default) or the number of, potentially overlapping, *n*-grams that cover it (as an option).

Table 2.5: Colour-coding by longest *n*-gram that covers a particular token (default mode)

Score	Colour
6+	purple
5	red
4	orange
3	green
2	blue
1	cyan
0	black

An example, which is the text of UN Security Council resolution 1321, is shown in in Figure 2.1, to illustrate the kind of output generated. The colour scheme employed is detailed in Table 2.5.

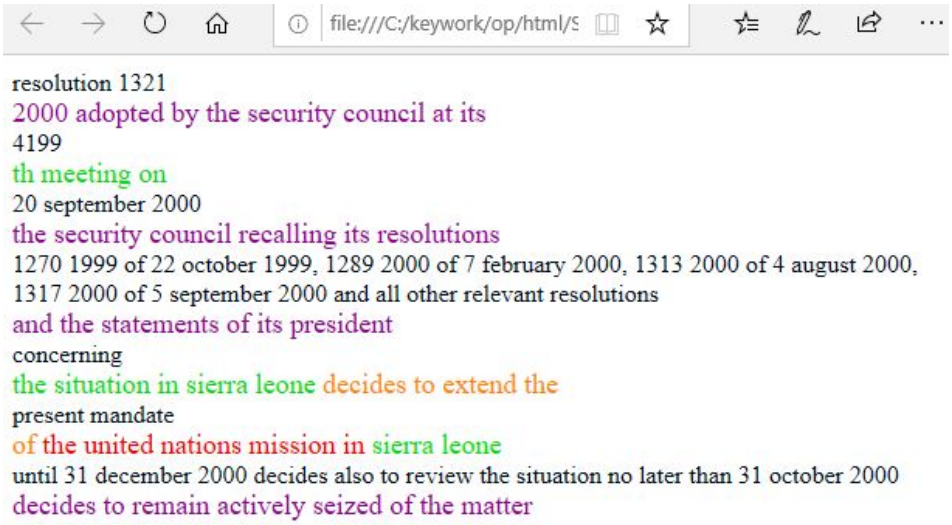


Figure 2.1: Text of UN Security Council resolution 1321, colour-coded by flicshow

The format here is that each change from covered to uncovered text starts a new line, so the original layout is lost; moreover, punctuation has been ignored, as in previous examples. The last line, ”decides to remain actively seized of the matter”, is an 8-gram. As already mentioned, the system was using up to 6-grams, so this indicates how fragmented portions of a sequence can overlap to give a better idea of the natural length of a repeated sequence.

More interesting is the third line from the bottom, ”of the united nations mission in sierra leone”. This begins with a single token ”of” in orange. Orange is the colour of 4-grams, but each token gets the colour of the longest n-gram that covers it; so this implies that the 4-gram ”of the united nations” was, in a sense, trumped, by the 5-gram, ”the united nations mission in”. Similarly, the last two words, ”sierra leone” are in green, indicating a 3-gram, but presumably the 3-gram ”in sierra leone” is also trumped by the same 5-gram, so that its leading token, ”in”, receives the colour of the longer sequence, in which it is the final token.

The intent of this colour scheme is to assist investigation of phraseological patterns by highlighting what might be termed a quasi-syntax, showing how longer

collocades are built up from shorter segments. Figure 2.2 shows another example, from the Hong Kong Tripadvisor postings, one that received many replies.

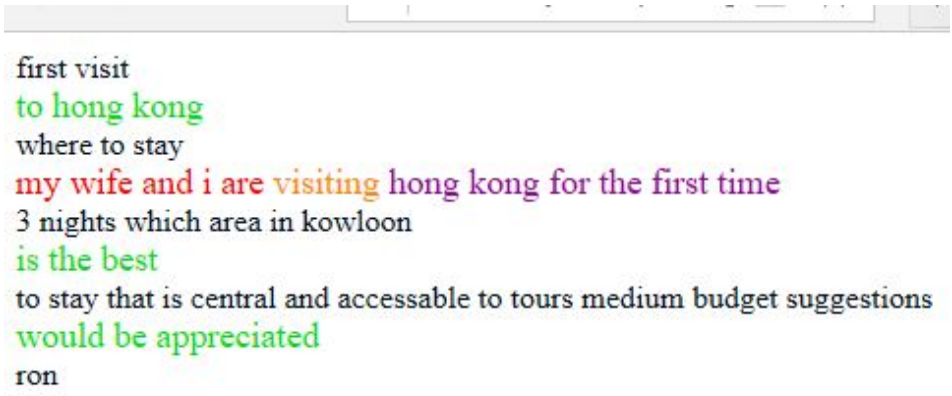


Figure 2.2: Colour-coded collocades in HK Tripadvisor forum posting

By contrast, Figure 2.3 shows output from flicshow when applied to a story by Edith Wharton called Venetian Nights. This extract exhibits long blocks of running text in black, with only a scattering of highlighted *n*-grams, most of which do not connect or overlap.

6 Classification via collocade coverage

The taverns program (Textual Affinity Values Employing Repeated N-gram Sequences) uses the formulexicon in a slightly different way, intended to indicate which texts are typical and atypical of their category and indicate how distinctive the categories are among themselves.

Inspecting individual texts can be a valuable opportunity to get close to the data, but in a typical corpus there is a huge amount of data to be inspected. The program taverns works in bulk mode and thereby gives an indication of which particular files might deserve the kind of close attention given to the output of flicshow. It goes a step further than formulex, using the same method, by computing coverage of each text specified not only by the *n*-grams of its own category, but by those of all the categories in the formulexicon file. Thus, in effect, it ranks each text file according to how typical it is of each category, including its own. Normally these texts are an unseen holdout sample, not used by outgrams to create the formulexicon.

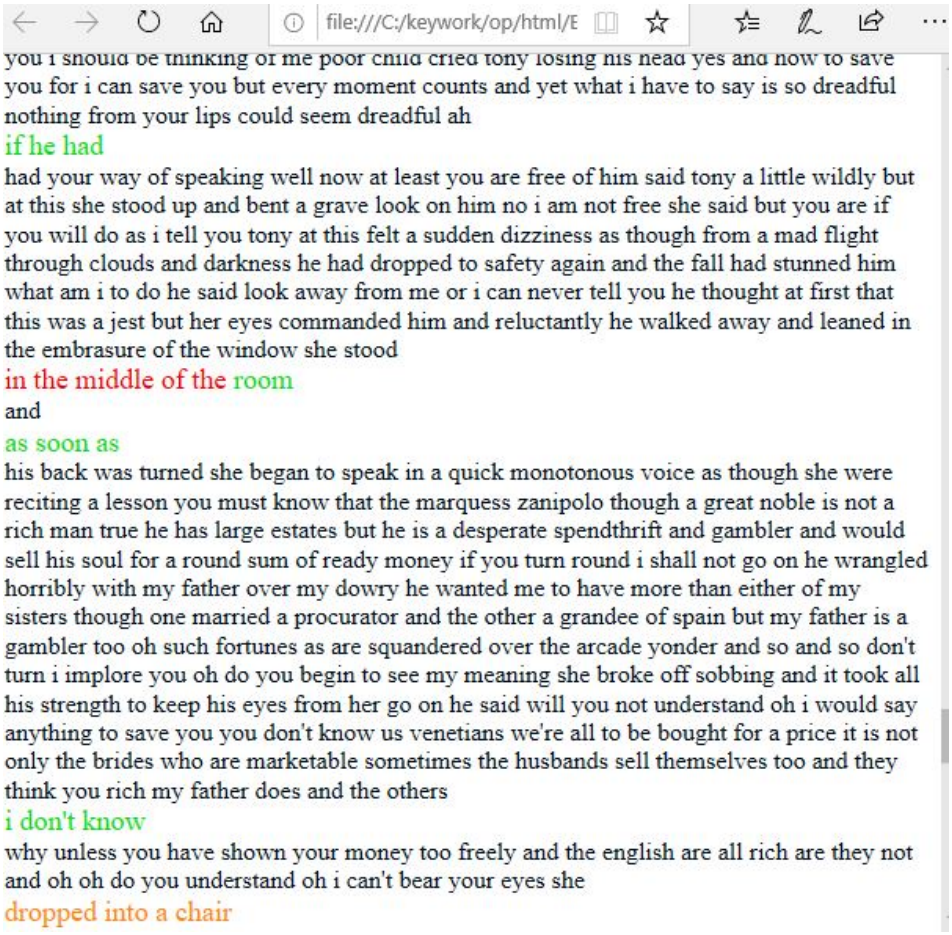


Figure 2.3: Extract from a tale by Edith Wharton

In addition, having done this, it performs text classification by assigning each text to the category which gives it the highest coverage score. It is not intended primarily as a text classifier, but the results in classification mode often shed light on the relationships between the text types involved, as well as identifying typical and anomalous texts.

The listing below shows the first 15 and last 15 lines of the taverns output for coverage by the BEER formulexicon.

```
Ranking by coverage of sequences from beer

   1    587    97   47.53  47.42  beer     youngs_hummingbird.txt
   2    586   100   44.37  43.00  beer     sol_cerveza.txt
   3    438    79   42.01  40.51  beer     budweiser.txt
   4    651   111   41.47  40.54  beer     tolly_english_ale.txt
   5    595   100   41.01  41.00  beer     wells_bombardier.txt
   6    636   105   35.69  36.19  beer     lancaster_blonde.txt
   7    709   123   34.41  33.33  beer     mcewans_amber.txt
   8    568    90   33.98  34.44  beer     marstons_burton_bitter.txt
   9    808   135   32.18  28.89  beer     blacksheep_venusmars.txt
  10   1101   186   31.06  29.57  beer     spitfire.txt
  11    683   120   29.28  28.33  beer     brains_sa.txt
  12    592   102   28.89  29.41  beer     wadworth_ipa.txt
  13    561    96   26.92  27.08  beer     yorkshire_gold.txt
  14    931   153   26.85  26.80  beer     weetwood_southern\_cross.txt
  15    571    92   26.27  27.17  wine     domaine_mandeville.txt

[.... many lines omitted ....]

 676   7397  1287    0.00   0.00  leaflet  Angitil_SR.txt
 677   5045   909    0.00   0.00  leaflet  Amoxil_Syrup.txt
 678   4384   789    0.00   0.00  leaflet  Amoxil_Capsules.txt
 679   6619  1144    0.00   0.00  leaflet  Algitec_Chewtab_Tablets.txt
 680   6198   946    0.00   0.00  leaflet  Adenocor.txt
 681  12823  2238    0.00   0.00  leaflet  Actrapid_Pen.txt
 682    762   123    0.00   0.00  wine     two_oceans_chardonnay.txt
 683   1072   180    0.00   0.00  wine     rina_ianca.txt
 684    483    75    0.00   0.00  wine     perlage_pinot_grigio.txt
 685    638   111    0.00   0.00  wine     paulmas_vinus.txt
 686    812   139    0.00   0.00  wine     la_chiave_2013.txt
 687    569    92    0.00   0.00  wine     finca_fabian.txt
 688    972   153    0.00   0.00  wine     era_puglia_falanghina.txt
 689    649   105    0.00   0.00  wine     domaine_begude.txt
 690    532    92    0.00   0.00  wine     doblez_garnacha.txt
```

The first item refers to the text of the back label of Young's Hummingbird ale. The first number is its rank, 1. The next two numbers give its size in characters and tokens, 587 and 97. The next two numbers show that 47.53% of its characters and 47.42% of its tokens were covered by 3- to 6-grams from the formulexicon of the BEER category. The last two columns give the actual category of the text and its file name. Note that this is a genuine holdout test, on 690 files that were not used by outgrams to create the formulexicon.

It will be seen that one WINE text creeps into the top 15 items as measured by typicality to the BEER category. In the bottom 15 items there are nine WINE texts. However, only 117 texts have more than zero coverage by the BEER formulexicon, so the order of the last 573 texts, with no coverage at all, is essentially arbitrary.

After listing each text as covered by *n*-grams from the formulexicon of each category (7 in this case) the program classifies each text according to how much of it is covered by each category's *n*-grams, taking maximum coverage to decide the assigned category. For the present example, the most confident 15 entries are listed below.

```
Results in classification mode:

      relative   actual       categories
rank  coverage% coverage%     pred : true      docname
   1    100.00    35.29       sres + sres  S_RES_12942000-en.txt
   2    100.00    30.10       sres + sres  S_RES_13362001-en.txt
   3    100.00    30.09       sres + sres  S_RES_15002003-en.txt
   4    100.00    29.86       sres + sres  S_RES_13162000-en.txt
   5    100.00    29.85       sres + sres  S_RES_14762003-en.txt
   6    100.00    29.63       sres + sres  S_RES_14432002-en.txt
   7    100.00    28.53       sres + sres  S_RES_13482001-en.txt
   8    100.00    27.94       sres + sres  S_RES_13882002-en.txt
   9    100.00    24.90       sres + sres  S_RES_14582003-en.txt
  10    100.00    24.86       sres + sres  S_RES_15182003-en.txt
  11    100.00    24.73       sres + sres  S_RES_14652003-en.txt
  12    100.00    24.10       sres + sres  S_RES_15482004-en.txt
  13    100.00    24.03       sres + sres  S_RES_13872002-en.txt
  14    100.00    23.73    leaflet + leaflet  SlowFe.txt
  15    100.00    23.62       sres + sres  S_RES_15302004-en.txt

The top line of this output

   1    100.00    35.29       sres + sres  S_RES_12942000-en.txt
```

signifies that Security Council resolution 1294 (from year 2000) has 35.29% coverage by SRES *n*-grams. The number in the second column, 100.00, indicates that

this 35.29% represents 100% of the total coverage by all 7 formulexicons, i.e. that *n*-grams from no category apart from SRES covered any of this text. These are the most confident classifications, 14 of the 15 being Security Council resolutions and one a medicine information leaflet.

The column labels ”pred” and ”true” stand for predicted and true category. The segment ”sres + sres” means that this file was predicted to belong to the SRES class and it was indeed from that class. The plus sign marks a correct decision: a minus sign would appear if it were incorrect and a question mark if the text's category were unknown.

Overall classification performance is summarized at the foot of the output listing by a confusion matrix, such as that for the present example, listed below.

```
Confusion matrix :

Truecat =          EW     beer  fewreps  leaflet manyreps  sres  wine
Predcat : EW       23        0        3        0        6    0     0
Predcat : beer      0       37        0        0        0    0     4
Predcat : fewreps   0        0       10        1       42    0     0
Predcat : leaflet   0        0        0      172        1    0     0
Predcat : manyreps  0        0       65        0      189    0     2
Predcat : sres      0        0        1        0        0  100     0
Predcat : wine      0        0        0        0        0    0    34

[As a monospaced text Listing.]
```

[As a Figure.]

Table 2.6: [As a Table.]

Truecat =	EW	beer	fewreps	leaflet	manyreps	sres	wine
Predcat : EW	23	0	3	0	6	0	0
Predcat : beer	0	37	0	0	0	0	4
Predcat : fewreps	0	0	10	1	42	0	0
Predcat : leaflet	0	0	0	172	1	0	0
Predcat : manyreps	0	0	65	0	189	0	2
Predcat : sres	0	0	1	0	0	100	0
Predcat : wine	0	0	0	0	0	0	34

Here the procedure makes 125 errors out of 690 decisions, about 18%, but only 18 of these mistakes, 2.6% of all 690 cases, arise from categories other than FEWREPS and MANYREPS. Essentially, this means that the system cannot distinguish between Hong Kong forum posts that receive few replies and those that receive many. Given how short these texts are (median sizes of 56 and 80 tokens) it would have been surprising, though interesting, if the two classes had been readily distinguishable by such a process. On the other hand, the other categories, even BEER and WINE, are well distinguished on this basis.

To give a point of comparison, the method described in **Wright2017** was implemented and applied to the same dataset. Wright obtained good results with this method in classifying messages from the Enron email corpus (**Cohen2009**) according to author. Each text was assigned the category with the highest similarity score based on the Jaccard coefficient (J) using sets of *n*-grams. The Jaccard coefficient divides the size of the set intersection by the size of the set union as in the formula

(1) $J = |A \cap B| / |A \cup B|$

were A is the set of *n*-grams in a single test text and B is the set of *n*-grams in the group of texts belonging to a particular category.

Wright's best results were found with tetragrams (4-grams) so that length was used on the same data as processed by the taverns program, above. Where taverns achieved a classification success rate of 81.88% (565/690), the Jaccard-similarity technique achieved 79.71% (550/690). When ignoring cases with zero similarity to any category, the rates were 83.77% (542/647) with taverns and 81.39% (503/618) with Jaccard similarity. This is only a single data point, but it does suggest that the present approach of using *n*-gram coverage gives results that are competitive with an established technique for this sort of application.

7 Clues from clusters of collocades

The previous sections have concentrated on analyses of individual texts or text categories; but a researcher in this field will typically not only be interested in how formulaic particular texts are, how typical or atypical they are of their class, and how similar or different a group of text types are amongst themselves, but also on how the repetitive sequences identified in this process relate to each other. In other words, it would be desirable for researchers into formulaic language to have a tool that helps to shed light on the patterns of phraseology that are responsible for high or low scores in terms of collocade coverage.

The program postflab is designed with this aim in mind. It uses a secondary output file of formulex, the flab listing (Frequently Assembled Lexical Bundles) as input and attempts to organize these frequent collocades in a manner that brings out their interrelationships.

An extract from a flab output file follows to illustrate the kind of data in question. This specimen consists of the first twelve lines from the patient information leaflet subcorpus.

```
3 leaflet 288 294415 1689519
 0.2868   285   16   3  tell your doctor
 0.2294   323   11   3  if you have
 0.2162   332   10   3  if you are
 0.1951   206   15   3  your doctor may
 0.1932   192   16   3  your doctor will
 0.1888   145   21   3  the active ingredient
 0.1749   197   14   3  you are taking
 0.1715   138   20   3  taking your medicine
 0.1545    30   86  16  if you have any questions or are not sure
    about anything ask your doctor or pharmacist
 0.1509    85   29   5  ask your doctor or pharmacist
 0.1442    84   28   6  out of the reach of children
 0.1406   198   11   3  do not take
```

The first line merely identifies the text category, and adds the information that it contains 288 files, comprising 294,415 tokens and 1,689,519 characters.

The next line shows that the collocade which covers the largest proportion of the subcorpus overall is ”tell your doctor”. This contains three tokens, is 16 characters in length and occurs 285 times altogether in that text category. The figure 0.2868 is a percentage, the percentage of the entire text of the subcorpus that is covered by this 3-token sequence. Further down the list can be seen the longest of these collocades ”if you have any questions or are not sure about anything ask your doctor or pharmacist”, a 16-element collocade that contains 86 characters and covers 0.1545% of the whole subcorpus, occurring 30 times.

A proportion of 0.1545% may seem tiny, but given that the commonplace triple ”one of the” is the most frequent collocade in the tales by Edith Wharton, covering a mere 0.0692% of the EW subcorpus, a 16-token sequence that accounts for even 0.1545% of a text corpus is worthy of attention.

A point to realize about such listings is that each item coverage is computed separately. For example, the eleventh item, ”ask your doctor or pharmacist”, occurring 85 times, is a substring of the item immediately preceding it. What this

implies is that the shorter string occurs 30 + 85 = 115 times altogether. It occurs 30 times preceded by ”if you have any questions or are not sure about anything” and 85 times without that preceding context. This avoids double counting.

The principle behind this mode of reckoning coverage can best be explained with reference to the digram ”your doctor”. This pair of tokens forms part of the items ”tell your doctor”, ”your doctor may” and ”your doctor will”, as well as the two longer items just discussed. As it happens, that particular word-pair occurs 2193 times in this training sample. What the figure of 206 next to ”your doctor may” tells us is that of these 2193 occurrences, 206 are followed immediately by ”may”; and likewise with the other collocades containing ”your doctor”.

The finding that shorter sequences often occur within longer collocades hints at a kind of network of phraseological possibilities surrounding a core component. However, because the flab output is listed in frequency order, it is very tedious to extract such information from the data as given. The program postflab is intended to alleviate this problem.

This program reads in the flab output produced by formulex and performs a 1-dimensional scaling on the collocades concerned, using string similarity as the value to be optimized. Multidimensional scaling (**UptonCook2006**) is a statistical optimization procedure which aims to reproduce as closely as possible a matrix of distances between items by assigning to each item coordinate values on a small number of dimensions. In this example, rather unusually, the algorithm is applied with just a single dimension. In postflab the process is taken to its minimal form, which means that the items are arranged in a single linear order that tries as far as possible to ensure that distances along the line correlate with distances derived from the entire matrix of inter-item similarities. In effect, the procedure adds, for strings, the concept of similarity order to the well-known concepts of alphabetic order and frequency order.

The derived ordering is written to a text file for inspection and also, more usefully, to a data file to be processed in the R package so that it can be displayed visually. Figure 2.4 shows the results of this procedure for the 36 most frequent collocades from the medical leaflet category.

In this diagram the vertical axis merely separates the items so that they do not overwrite each other. The horizontal axis represents the closeness of the items along a single dimension. The width of the blue lines is proportional to the aggregate coverage of all the collocades with the same score on the x-axis. These lines are intended to reveal the presence of certain groupings along the x-axis, including the main grouping which consists of a number of collocades containing the digram ”your doctor”. Hence the program has performed a clustering as a side-effect.

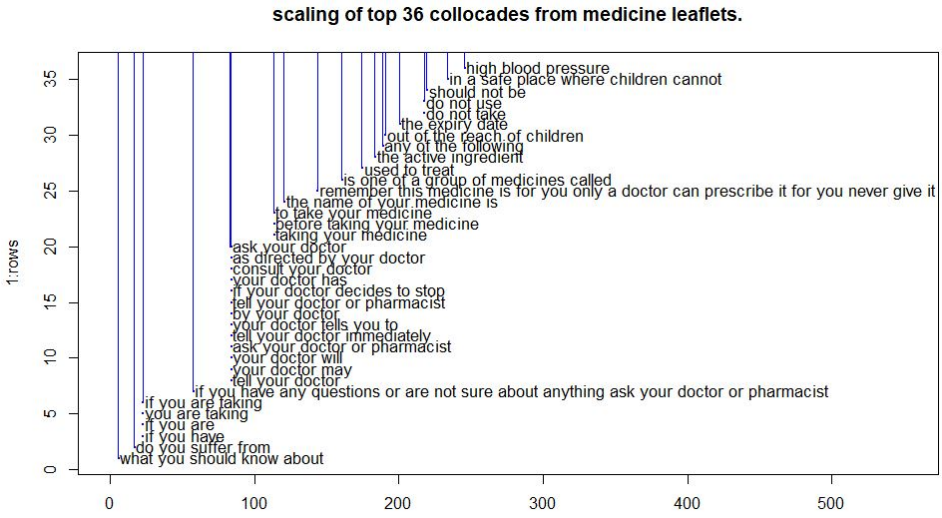


Figure 2.4: One-dimensional scaling of collocades from medical leaflet category

Although this graphic representation is based only on superficial string similarity, and has no semantic underpinning, it nevertheless makes it much easier for a researcher to find clusters of related phrasings than the text-based listing.

8 Concluding remarks

The formulib package implements one particular way of operationalizing the concept of formulaic language by using a traditional resource, the frequent *n*-gram list, in a slightly novel manner. It constitutes an innovative tool based on a simple idea, which offers the researcher informative ways of viewing repetitive phrasings in a corpus or collection of corpora. It takes further the work of **ForsythGrabowski2015** by providing estimates of how much formulaic language is found in individual texts as well as how formulaic certain text types are. Although the examples quoted in the present chapter are in English, formulib can be applied to any language, and it does not need pre-processing resources such as lexicons, parsers of taggers.

We cannot expect a single approach to cover all the different aspects of such a complex phenomenon. However, the argument of this chapter is that the concept that underlies the methods employed by the formulib software, namely collocade

coverage, offers a straightforward but relatively effective way of investigating some of the more important aspects of formulaic language.

Formulib cannot, of course, be regarded as an endpoint in the continuing attempt to explore patterns of formulaic sequences. Even staying with collocade coverage as a key indicator, there is room for further development. For instance, it would be highly desirable to find a more precise notation for integrating clearly related elements of the formulexicon such as

> please read this leaflet carefully before taking your tablets

and

> please read this leaflet carefully before you take your medicine

into a form that reveals their relatedness (a kind of micro-grammar). The two items above could be unified with the aid of a pattern-description language, such as the following.

> please read this leaflet carefully before [taking | you take] your [medicine | tablets]

However, to do this efficiently and reliably would require leading-edge artificial intelligence applied to the induction of a small-scale grammar, and the results would doubtless be hard to interpret without advanced data visualization techniques. Perhaps a reader may take up that challenge. At any rate it shows this approach opens up plenty of avenues for further research.

Appendix A Appendix

Better move this somewhere else, e.g. in the Introduction

The suite of programs in Python 3 that constitute the formulib package can be found at the address below.

http://www.richardsandesforsyth.net/software.html

Software and sample text files are freely available under a Creative Commons licence.

Chapter 3

Exploring the valency of collocational chains

Piotr Pęzik
University of Łódź

Whereas the prefabricated status of idioms or restricted collocations is relatively self-evident in their context of use, the "underlying rigidity" (**Sinclair1991**) of other types of phraseological units may only become evident through large-scale analyses of reference corpora. This chapter focuses on the identification of subtle lexico-grammatical petrification of multiword units in dependency-annotated corpora. More specifically, it investigates restrictions on the valency of binary collocations and their tendency to be regularly subsumed by larger collocational chains. For example, the binary collocation *deep breath* is almost invariably a direct object of a small set of verbs: *take*, *draw*, *let out*. This restriction can be contrasted with collocational chains in which other adjectival collocations of *breath* (e.g. *bad breath*) have a wider range of syntactic roles determined mainly by the potential valency of their head noun (i.e. its propensity to function as subject, object etc.). Apart from discussing examples of such constructions from Polish and English corpus data, the chapter also attempts to show how lexico-syntactic properties of multiword units can be systematically accounted for and explored using a dependency-based approach to phraseology extraction.

1 Introduction

Since phraseology is a field "bedevilled by the proliferation of terms and by the conflicting uses of the same term" (**Cowie1998**), it is not superfluous to clarify what is meant by the terms *collocations* and *collocational chains* in the context of this paper. Unless otherwise indicated, the term *collocation* is taken to mean a *binary lexical collocation*, i.e. a recurrent combination of just two content words

(possibly linked by a grammatical word) which remain in an explicit syntactic relation, e.g. *blind date*, *turn of phrase*. From the perspective of language production, collocations are assumed to be recalled from memory, either associatively or holistically, rather than recomposed in a completely spontaneous and uninspired manner. A review of different definitions of collocations by **Pęzik2018** shows that they usually appeal to three main types of identification criteria: formal, distributional of psycholinguistic. The so-called "restricted collocations" are combinations which consist of an "autosemantic" base and a "synsemantic" collocate (**HeidGouws2006**). They can be roughly classified into four major groups (**Mel'čuk2001**; **Pęzik2018**). Open binary collocations are composed of two largely autosemantic words, which makes them less obvious to recognize as phraseological units. One of their subtle characteristics as units of prefabricated language is a degree of *stereotyped recurrence* i.e. the tendency to occur in similar semantic, syntactic and pragmatic contexts (ibid.: 51). While both restricted and open collocations play a key role in the production of fluent, native-like language, restricted collocations may also cause reception problems for non-native learners of a given language. Some restricted collocations are in fact figurative idioms as they instantiate conventionalized metaphors, metonymies and other conceptual blends, e.g. *blind alley*.

ibid unclear

There is a wide spectrum of phraseological units which may consist of more than two words, such as pure and figurative idioms, proverbs, commonplaces, catch phrases, slogans etc. (**Cowie1998**). Although tens of thousands of idioms and collocations have been identified and recorded in dictionaries and combinatorial databases, there is a need for further research on some of the more subtle types of phraseological prefabrication. Among the less extensively researched phraseological phenomena are collocational chains, which are defined here as overlapping combinations of two or more lexical collocations.[1] As shown further in this paper, collocational chains can be composed spuriously or largely predetermined to occur in their entirety.

Defining collocations and other types of phraseological units (PUs) as word combinations linked by an explicit syntactic relation may come across as somewhat controversial, partly because syntactic idiosyncrasy of PUs is regularly enumerated as one of their most salient characteristics. In its extreme form it can be described as *ill-formedness* or deviation from grammatical regularity. For example, the fact that it is difficult to assign the constituents of the idiomatic expression *by and large* to modern day English morphosyntactic categories has earned

[1] Some definitions of collocational chains also distinguish between collocational chains and collocational clusters (**Hausmann2004**; **HeidGouws2006**).

it the name of "an ill-formed collocation" (**Moon1998**).[2] More often, phraseological units tend to be *petrified* in that they are mostly used in a limited subset of the morphological variants licensed by their otherwise regular syntactic structure. However, although syntactic idiosyncrasy testifies to the status of some PUs as clearly prefabricated constructions, one should not conclude that all or even most PUs are marked by syntactic irregularity. In reality, most idioms and collocations seem to be lexical realizations of regular syntactic patterns, even if their prototypical forms are petrified. The most obvious proof of this statement is the existence of dictionaries of idioms (**CowieMackin1975**; **CowieEtAl1993**) and collocations (**CrowtherEtAl2003**) whose macro- and microstructures are organized around a set of productive syntactic patterns of idiomatic expressions. Furthermore, the very fact that most subsentential PUs have to be embedded in the syntactic structure of a sentence means that they also have an "external valency" (**Burger2003**). To put it in the parlance of dependency syntax, PUs have typical syntactic roles as governors or dependents of other words and phrases in the sentence. Those two properties of PUs, i.e. their internal structure and external valency are implicitly recognized in combinatorial dictionaries as illustrated in the following entry for the phrasal verb *to back on to* from the Oxford Dictionary of Current Idiomatic English (**CowieMackin1975**: 10):

(1) back on to [A3] *have at its back, face at the back.* **S** house, shop; study, kitchen. **o**: court-yard; lane, alley

The internal structure of the phrasal expression is indicated by the label A3, which denotes intransitive verbs with a particle and a preposition, while its external valency is implied by the two lists of its typical subjects (S) and objects (o).

The assumption that most phraseological units have both a regular internal syntactic structure and an external valency specification opens up some perspectives of computerized explorations of their distribution as either self-contained or largely embeddable constructions. This point is elucidated at some length in the subsequent sections of this paper, but it can be illustrated right away with a simple example. In reference corpora of English, the seemingly independent binary collocation *profound effect* functions almost exclusively as a direct object of *have* as in *have a profound effect*. The latter construction is in turn subsumed by an even longer collocational chain with *on* as a fixed prepositional dependent and its open-ended prepositional nominal or pronominal object dependent as in

[2] Such highly idiosyncratic combinations are difficult to directly integrate in the standard dependency representation used for the proposed method of phraseology extraction.

to have a profound effect on + *NOUN/PRON*. The fact that such structures may be recursively recombined in multiple, possibly also prefabricated constructions has some practical implications for the design of phraseological dictionaries and databases.

This paper first discusses the problem of fragmentation of such collocational chains in dictionaries and automatic combinatorial databases. The phenomenon of syntactically restricting subsumption of shorter word combinations in longer recurrent constructions is then discussed in terms of potential valency restrictions. Finally, the paper presents a new software tool named Treelets[3], which showcases some applications of dependency-based phraseology extraction. The method of generating a combinatorial dictionary implemented in this tool uses special data structures called *subsumption graphs* to facilitate the search and visualization of embedded and overlapping phraseological constructions.

2 Relational phraseology extraction

2.1 Fragmentation of phraseological units

The degree to which various PUs can be expected to adhere to regular syntactic structure is important in the context of phraseology extraction (PE) – an area of corpus research which deals with automated identification of phraseological units in corpora through aggregation of word co-occurrences attested in reference corpora. PE techniques can be broadly categorized into positional and relational (**Evert2005**), although this distinction is sometimes blurred by practical considerations. Positional approaches rely on counting and weighing linearly related word co-occurrences in text. Relational PE techniques utilize explicit annotations of syntactic relations between constituents of PUs. As a result, the latter type of methods crucially depend on the syntactic predictability of PUs; word combinations which co-occur in syntactic configurations unpredicted by predefined syntactic patterns are ignored in the process of extraction. Because syntactic patterns used in the process of extraction have to a) conform to the particular treebank formalism used to annotate the working corpus and b) be consistently annotated by automatic syntactic parsers, the results of relational extraction may reveal only "details of language" covered by a particular syntactic theory (**Sinclair1991**) rather than the full spectrum of usage.

Another broad distinction can be made between "ad hoc" PE modules and extraction systems which precompute combinatorial databases with a dictionary-

[3]See http://pelcra.pl/new/treelets.

like macro structure. Ad hoc PE modules available in various corpus search engines usually perform positional extraction of binary collocations, n-grams or skip-grams for single and multiterm queries defined by users. For example, the collocation extraction of the MoncoEN corpus search engine[4] can be used to define a single- or multiword node expression for any corpus query formulated in its query syntax. Table 3.1 presents a list of adjectival collocates extracted from a sample of almost 70,000 occurrences of the noun *advantage* in data crawled from various English-language news websites. The results of the extraction query can be sorted by frequency or their strength of association, which is a variation of the Dice score in this case. The fourth column contains a frequency list of positions relative to the node word (which occurs at position=0) which is useful in identifying the predominant syntactic roles of the collocates. For example, the adjective *competitive* seems to mainly precede the noun *advantage*, which suggests that it is used as its adjectival premodifiers in this case. The last column lists word N-grams bounded by the node and collocate, which is meant to indicate some of the recurrent forms of each collocation as well as its higher-order constructions such as noun phrases with multiple adjectival modifiers, e.g. *unfair competitive advantage*.

Table 3.2 shows the top five adjectival collocates of the noun *advantage* recorded in HASK EN (Pęzik 2014)[5], a combinatorial database precomputed from the original edition of the British National Corpus (BNC). The remaining columns of the table show a selection of strength of association and dispersion scores.

Even though all of the top adjectives from the two lists seem to be genuine collocates of the noun *advantage*, both of the extraction systems illustrated above suffer from the problem of fragmentation: recurrent fragments of larger multiword expressions are represented as unrelated binary collocations. For example, taken at face value, both of the lists above might imply that *full advantage* is a self-contained intensifying binary collocation which could be used freely in a variety of syntactic roles predetermined by its head noun. However, in the first edition of the British National Corpus (BNC) more than 86% of the occurrences of *full advantage* function as part of the longer expression *take full advantage (of)*. In a 440 million word version of the Corpus of Contemporary American English (COCA) the same phrase is used as a direct object of *take* in over 96% of its attested usages. In other words, *full* seems to function as a collocational intensifier of *advantage* mostly when the latter is a direct object of the light verb construction *take + advantage* as illustrated in Ex. 2 below:

[4] See http://monco.frazeo.com.
[5] See also http://pelcra.pl/hask_en.

Table 3.1: Adjectival collocates of *advantage* retrieved with the MonCoEN search engine

#	Adjective	Frequency	Dice	Positions	Example N-grams
1	competitive	1260	0,0413	{-1=1239, -2=17, 2=4}	{competitive advantage=1078},{competitive advantages=161},{competitive business advantage=3}
2	full	1165	0,0382	{-1=1154, -2=9, 2=2}	{full advantage=1134},{fullest advantage=12},{full advantages=4},{fuller advantage=3},{full of advantage=3
3	big	900	0,0295	{-2=76, -1=789, 2=34, 1=1}	{big advantage=488},{biggest advantage=167},{big advantages=75},{advantage of big=14},{big an advantage=5},{big size advantage=5},{big fundraising advantage=4},{biggest home-field advantage=3}
4	unfair	643	0,0211	{-1=567, 2=2, -2=74}	{unfair advantage=525},{unfair advantages=42},{unfair competitive advantage=34},{unfair commercial advantage=5},{unfair trade advantage=4},{unfair competitive advantages=4}
5	great	571	0,0187	{-1=456, 2=41, -2=74}	{great advantage=320},{great advantages=92},{greater advantage=37},{advantage of great=23},{great comparative advantage=9},{advantage , great=7},{greater advantages=7},{great natural advantages=6},{great competitive advantage=5}

Table 3.2: Adjectival collocates of advantage retrieved recorded in HASK EN database

#	Collocate	Frequency	T-score	MI^3	G^2	JD
1	competitive	149.0	11.57	18.7144	613.96	0.82
2	full	166.0	8.43	16.28461	139.453	0.90
3	added	71.0	8.12	17.097	346.23	0.84
4	comparative	68.0	7.89	16.72091	308.08	0.72
5	unfair	67.0	7.69	16.20767	260.583	0.85

(2) Customers ***take*** **full advantage** of in-house electropolishing (...). [COCA, Physics Today]

Intended users of such tools and automatically extracted resources are therefore required to inspect the concordances underlying such tabular results to distinguish between mostly subsumed collocations and freely recombinable collocations such as *big advantage*. The latter combination is not restricted to occur in a single syntactic function. It is used as a direct object of verbs such as *have* or *give* in approx. 34% of its occurrences in COCA and as a nominal subject dependent (28% of occurrences) as in (3):

(3) The ***big*** **advantage** for the investor ***is*** that he can trade all his cryptocurrencies in one place. [MoncoEN, thenextweb.com]

Collocation dictionaries may also be affected by the problem of PU fragmentation. For example the *Oxford dictionary of collocations* (**CrowtherEtAl2003**) defines the noun bearing as a "way in which something is related" and lists three of its adjectival collocates: *direct*, *important* and *significant*. None of those collocations is likely to be used outside of the larger construction *have a direct/important/significant bearing on*. This information is only indirectly implied by the example sentence illustrating the use of the first of those collocations and a separate entry for the direct object lexical collocation of *have + bearing* and the grammatical collocation of *bearing + on*:

(4) bearing
1. *way in which sth is related*
ADJ. **direct**, **important**, **significant**
The rise in interest rates had a direct bearing on the company's profits.

VERB + BEARING **have**
PREP. ~ **on**

Of course, the coverage of this particular dictionary was by design limited to binary collocations and its space limitations preclude detailed usage notes. On the other hand, it could be argued that this collocational chain would probably be better represented as a single unit in this case. In OCD a special section labelled *PHRASES* is occasionally used to enumerate additional set expressions which do not conform to the four basic patterns of binary collocations covered by this dictionary.

Table 3.3 shows more examples of intensifying adjectival modifier collocations which are rarely, i.e. usually in less than 15% of cases used independently of larger, recurrent constructions.

The first example "collocation" in the second column of Table 3.3 is *upper hand*. In BNC it is always used a direct object of a handful of verbs shown in the last column (*have*, *gain*, *give*, hold) while in COCA there are sporadic instances of usage as implicit direct objects in elliptic headlines or in other syntactic roles. A similar level of subsumption is observed for the phrase *mental note* which is rarely used outside of the set expressions *make/take a mental note*. Non-direct object usage requires a creative context such as the piece of science-fiction writing illustrated in Ex. 5 below:

(5) Amber put it on her mental note pad. [COCA, Analog Science Fiction & Fact]

Some of the examples from Table 3.3, such as *have the upper hand* and *take a mental note* are simply multiword figurative idioms and thus the problem of their identification as phraseological units is purely technical. On the other hand, example phrases 5 (*important bearing*, which is invariably embedded in *have a direct bearing* in the two corpora), 6 (*profound effect*), and 7 (*significant role*) are restricted intensifying collocations. Examples 9 (*excellent job*) and 10 (*short laugh*) are open collocations comprised of two largely autosemantic constituents which simply happen to regularly form a longer structure with an overlapping direct object collocation. The typological dilemma with the latter examples is therefore whether they should be recognized as self-contained phraseological units or as more spurious and open-ended constructions. For the practical purposes of phraseology extraction, we might describe the frequently embedded phrases as *subsumed binary collocations*, depending on how unlikely they are to be used independently of the larger constructions. The subsuming constructions can be

Table 3.3: Examples of binary adjectival modifier collocations regularly embedded in larger collocational chains in BNC and COCA

#	Binary AMOD coll.	Frequency		Subsumption as DOBJ		Frequent verb governors in (COCA; BNC)
		BNC	COCA	BNC	COCA	
1	upper hand	122	670	122/122=1	608/670=0.91	have (36; 223), gain (41; 157), get(19; 100), give (8; 36), hold(8; 25) + the upper hand
2	little resemblance	59	208	57/59=0.966	181/208=0.93	bear (56; 181) + little resemblance
3	mental note	78	289	75/78=0.96	262/289=0.90	make (74; 228), take (1; 20) + a mental note
4	deep breath	664	4282	587/664=0.88	3872/4282=0.89	take (505; 3438), draw (72; 225), let out (7; 30) + a deep breath
5	important bearing	30	14	30/30=1	14/14=1	have (30; 14) + an important bearing
6	profound effect	164	460	156/164=0.95	420/460=0.91	have (150; 392) + profound effect
7	significant role	146	905	129/146=0.88	824/905=0.91	play (105; 692), have (15; 66) + a significant role
8	full advantage	170	423	146/170=0.86	406/423=0.96	take (142; 402) + full advantage
9	excellent job	58	326	49/58=0.85	306/326=0.94	do (46; 298) + an excellent job
10	short laugh	72	100	72/82=0.88	74/100=0.74	give (67; 43) + a short laugh

multiword idioms or simply recurrent collocational chains. Collocational chains which consist of subsumed collocations should be distinguished from spurious chains of independent collocations such as *my heart of stone is filled with pride*.[6]

As shown in Table 3.4 some of the *amod + dobj* constructions from Table 3.3 are highly likely to recur in larger recurrent structures which are also subtrees of the sentence dependency tree. For instance, more than 87 percent of the occurrences of the recurrent chain *play a significant role* in COCA have a prepositional object introduced by *in* as in *play a significant role in + POBJ*. The subsumption of the other four second-order chains shown in Table 3.4 in third-order chains is even higher in the two reference corpora.

Table 3.4: Subsumption of *amod + dobj* collocational chains in structures with a prepositional attachment

	Frequency		Used with a pobj dependent	
# amod + dobj collocational chain	BNC	COCA	BNC	COCA
1 play a significant role	105	692	92/105=0.88	616/692=0.89 (in)
2 bear little resemblance	59	181	52/59=0.88	175/181=0.97 (to)
3 have a direct bearing	37	54	33/37=0.89	54/54=1 (on/upon)
4 have profound effect	150	392	133/150=0.89	348/392=0.89 (on/upon)
5 take full advantage	142	402	132/142=0.93	380/402=0.95 (of)

2.2 Potential vs. activated valency of PUs

The subsumption of shorter dependency subtrees (including single word *subtrees*) in longer recurrent collocational or idiomatic structures may considerably affect the distribution of their syntactic roles. As an example, the ratio of the noun *fact* as a prepositional object dependent is much higher than the ratio of

[6] See https://www.youtube.com/watch?v=wvYRQ-sFMJw (2:18).

all nouns used as prepositional objects in reference corpora of English. However, the ratio of *fact* as a prepositional object "drops" approximately to the level observed for all nouns when *fact* is modified by an adjective (see the discussion of Tables 3.5–3.7 below). More generally, dependency type ratios vary considerably not only for different content words but also with respect to the higher order constructions in which they occur.

In dependency syntax, vertices representing words in the sentence dependency tree can be said to have a "passive valency" (**Mel'čuk1988**, cf. **Boguslavsky2003**; **Boguslavsky2016**; **Boguslavsky2003**). The passive valency of a dependency subtree (including single-word subtrees with their morphosyntactic roles such as nouns, verbs, adjectives etc.) can be defined as its default propensity to function as governors or dependents of a set of types. Since the terms "active" and "passive valency" have also been used to refer to the direction of the dominance relationship between words in a dependency tree (**Moroz2013**), to avoid confusion in this paper, the terms *potential* and *activated valency* will be used to describe the default and corpus-attested dependency patterns formed by words and phrases. For example, nouns have the rather obvious default potential of functioning as nominal subjects, objects of verbs or prepositions, nominal modifiers, etc. while verbs are typically sentence roots, auxiliaries, x-complements etc. The approximate activation of such potential valency roles (i.e. the activated valency of a word or phrase) can be estimated from manually annotated dependency treebanks or automatically parsed corpora. Neither of these options is ideal as treebanks are limited in size and parsers produce erroneous annotations, but even an approximate estimation of the activated valency of a word or phrase may throw some light on its actual usage.

As shown in Table 3.5, in both COCA and BNC, nominal dependents are usually prepositional objects (35–38% of all noun occurrences), direct objects (16%) and nominal subjects (12–18%).

The most common dependents of English nouns are determiners (24.6%), followed by adjectival modifiers (16.6) and prepositions (12.2). The proportions of dependency relation types for specific nouns, verbs or adjectives may be very different from such overall distributions. As shown in the last column of Table 3.5, the standard deviation of the prepositional object dependent type is 0.34 percentage points in COCA.

In studies of verb valency, it is taken for granted that different verbs can be classified into groups of similar *subcategorization frames*. In other words, different verbs require or *subcategorize* different types of configurations of their dependents. From the perspective of phraseology, it is also interesting to consider

Table 3.5: Ten most frequent types of nominal dependents in BNC and COCA

	BNC		COCA			
Dep. type	Freq.	Ratio	Freq.	Ratio	M/n[a]	SD
pobj	7 844 932	0.38	30 824 682	0.35	0.20	0.34
dobj	3 246 866	0.16	14 922 242	0.16	0.09	0.24
nsubj	2 416 031	0.18	10 870 549	0.12	0.10	0.25
compound	2 008 918	0.09	9 787 999	0.11	0.15	0.31
conj	1 550 954	0.07	5 509 718	0.06	0.08	0.22
attr	853 480	0.04	3 728 782	0.04	0.02	0.13
nsubjpass	579 565	0.03	1 474 911	0.01	0.06	0.21
ROOT	533 112	0.02	3 068 416	0.03	0.01	0.07
npadvmod	474 680	0.02	2 537 174	0.02	0.12	0.30
appos	305 757	0.01	2 215 588	0.02	0.02	0.12

[a]Mean per noun

the activated valency of nouns and other open-class *content words* such as adjectives or adverbs which function as headwords defining the entry structure of lexicographic resources. The activated valency of a dependency subtree (such as a word or phrase) can be defined as the set of dependent and governor types in which it is found in a reference corpus. Although corpus-based valency estimations can only be probabilistic and approximate in nature, they do shed light on the actual usage patterns of words and phrases, and they are especially revealing when such words or phrases tend to be embedded in larger recurrent constructions.

Table 3.6 shows the distribution of dependent types realized by the nouns *breath* and *fact* estimated from the syntactically annotated version of COCA used in this study. The use of *breath* as a direct object is considerably more frequent than the average value observed for nouns in this corpus (49.95 vs. 16.8%) whereas its frequency as a prepositional object is lower than the average (29.54 vs. 34.12). On the other hand, the noun *fact* is a prepositional object in over 65% of its occurrences, which is considerably higher than the average ratio of 34% observed for all nouns in this corpus. This example shows that the activated valency of those two words tends to differ considerably either from their potential valency as nouns or even from the overall or average rations observed for all nouns in a reference corpus.

Table 3.6: The nouns *breath* and *fact* as dependents in COCA

breath in COCA			*fact* in COCA		
Dep. type	Freq.	Ratio	Dep. type	Freq.	Ratio
dobj	12,830	0.50	pobj	108,026	0.67
pobj	7,589	0.30	nsubj	18,964	0.12
nsubj	2,344	0.09	dobj	16,150	0.10
compound	851	0.03	attr	7,336	0.04
conj	520	0.02	conj	3 657	0.02

It is not obvious whether the difference between the typical dependency types of the two nouns can be linked to their general semantic properties. What seems to be the case is that at least *some* of this variation is due to a handful of phraseological restrictions on the valency of those two nouns. For example, almost 72% of all the occurrences of *breath* as a direct object are governed by just four verbs: *take* (5194), *hold* (1896), *catch* (1428), *draw* (708). Taken alone, the support verb restricted collocation *take a breath* accounts for over 40% of the use of *breath* as a direct object. The syntactic distribution of the noun *fact* is even more biased by its formulaic usage: over 64% (69295) of its occurrences as a prepositional object *fact* are instances of a single discourse linking phrase: *in fact*.

The activated valency levels observed for a single word may change considerably once this word in used in a collocation. The previous section shows how the potential valency of binary collocations may also be restricted by the distribution of a small set of prefabricated higher order structures in which they are typically found. Table 3.7 shows frequencies of dependent types assumed by the nouns *fact* and *breath* when they are modified by adjectives as in *[access] to simple facts* or *have a bad breath*. The proportion of individual dependency types of the two nouns is only partly consistent with their overall dependency type distribution. Adjective-modified occurrences of *breath* are even more likely to be direct objects (68%), whereas instances of *fact* with an adjectival modifier are half as likely to be prepositional objects. Much of the first difference can be explained by the existence of the construction *take a deep breath*, which is used both literally as an established collocational chain and idiomatically as a figurative expression (see Table 3.3). The decrease in the ratio of prepositional object instances of *fact* observed when we consider its use with adjectival modifiers results to a large extent from the absence of *in fact* or a similar phrase in this ranking. There are some formulaic adjective-modified usages of *fact* as a prepositional object such as the

sentence initial discourse marker *in actual fact* (67 occs. in COCA), but they do not compensate for the absence of the much more frequent prepositional phrase *in fact*. The propensity of a word or phrase to be used as a dependent or governor of a larger structure may be significantly skewed by its use in a single higher-order phraseological construction such as *take a deep breath* or *in fact*.

Table 3.7: Top five dependent types *fact* and *breath* with a modifying adjective

amod(fact, x)			amod(breath, x)		
Type	Freq.	Ratio	Type	Freq.	Ratio
Pobj	3,333	0.32	dobj	6,033	0.68
Dobj	2,247	0.21	pobj	1,824	0.21
Nsubj	1,870	0.18	nsubj	314	0.04
Attr	1,332	0.13	ROOT	171	0.02
Conj	401	0.04	conj	163	0.02

In order for automatic combinatorial databases to account for activated valency patterns of phraseological units, they have to identify and represent such subsumption phenomena. The following section describes a method of storing extracted collocational structures which was designed to address the issue of recursive subsumption of PUs.

2.3 Subsumption graphs

Pęzik2018 describes an experimental method of extracting combinatorial databases from dependency-parsed corpora which keeps track of subsumption relations between overlapping constructions of different sizes. The working assumption of the PE approach used in this study is known as the continuity restraint (**OGrady1998**), which predicts that an idiom's obligatory lexical components form a subtree of the sentence dependency tree. The validity of this assumption depends on the exact dependency formalism used to represent PUs. Also, it seems to fail in the case of some variable idiomatic expressions such as *walk a thin|fine line|path between*. It is nevertheless a useful assumption in large-scale phraseology extraction. One of its advantages is that it covers collocational subtrees which are neither complete or single phrasal constituents such as *include such factors as*.

The extraction process starts with a set of headwords, which are simply part-of-speech typed lemmas of content words, and a set of dependency patterns in

which those headwords are expected to occur. Next, for each headword, the full set of lexically recurrent subtrees, "catenae" (**OsbourneEtAl2012**) or "treelets" is extracted from a reference corpus. Extracted subtrees are stored with some distributional and structural properties in a relational database. The headwords define the macrostructure of the resulting automatic combinatorial dictionary (ACD) and the set of patterns used determines the microstructure of each of its entries.

Recurrent subtrees containing a given headword are stored in a data structure called a *subsumption graph*. A section of a subsumption graph generated for the noun *effect* in COCA is shown in Figure 3.1. Its full version comprises 10 855 vertices representing subtrees containing this noun and occurring at least twice in this corpus. The vertices of the subsumption graph represent recurrent binary collocations and higher-order collocational subtrees whose syntactic structure matches one of the predefined patterns. The patterns can be defined manually as explained in Section 3.2 or derived in a weakly supervised manner from the corpus. The weighted directed edges indicate the subsumption relation. The value of the edge weights represents the frequency of subsumption observed in the reference corpus (it is in fact equal to the frequency of the subsumed combination). A loop edge is added to vertices without outgoing edges to indicate the frequency of the combination represented by that vertex. For example, the binary collocations *have + effect* and *profound + effect* have frequencies of 1 026 and 465 respectively as indicated by the weights on their loop edges. The subsumption ratio of a given collocation can be calculated to the extent it can be estimated from the set of patterns used as the sum of the frequency weights of edges incoming from other vertices divided by the total frequency of that node. For example, the subsumption score of the chain *have + profound effect on* in longer structures in this graph is 10 (*have + profound effect on people*) + 19 (*have + profound effect on life*) / 313 = 0.092. The number of the incoming edges (indegree) other than the loop edge reflects the *productivity* of a given subtree.

Complex restrictions on the potential valency of a given lexicalized subtree can be visualized as a syntactic subsumption graph similar to the one shown in Figure 3.2. As shown earlier in Tables 3.3 and 3.4, the subsumed intensifying adjectival modifier collocation *profound effect* is largely restricted to occur as a direct object (420/460 occurrences) of just 5 recurrent verbs, and when this is the case, it is in turn largely restricted to take a prepositional object (348/420). Such recurrent subsumption is conveniently represented as a subsumption graph.[7]

[7] The edge label *420, 5* means that *profound effect* occurs 420 times as a direct object of only 5 different verbs; The label *3,1* means that it is used only three times as a nominal subject with just one verb (to be), etc.

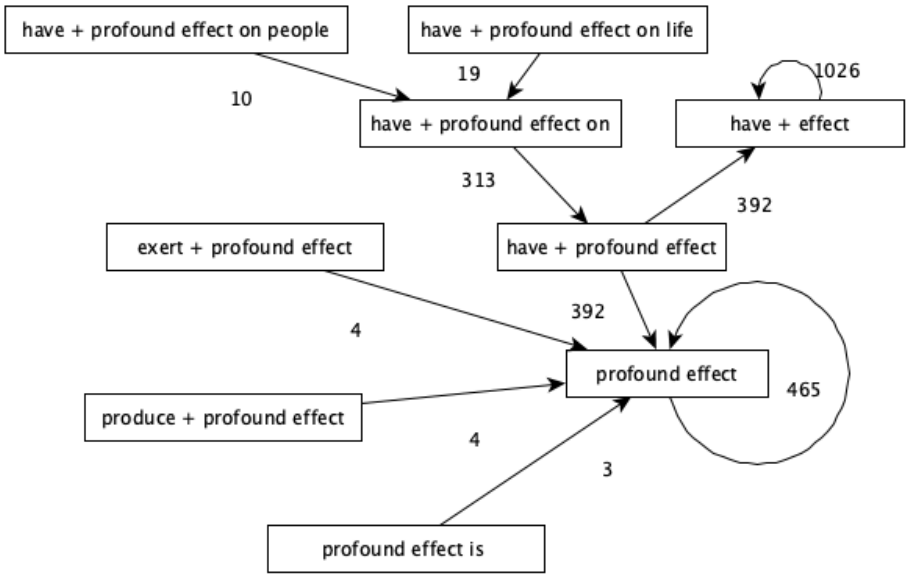


Figure 3.1: A simplified subsumption graph generated for recurrent dependency subtrees containing the noun *effect* in COCA. Only a subset of recurrent subtrees containing the noun is shown here.

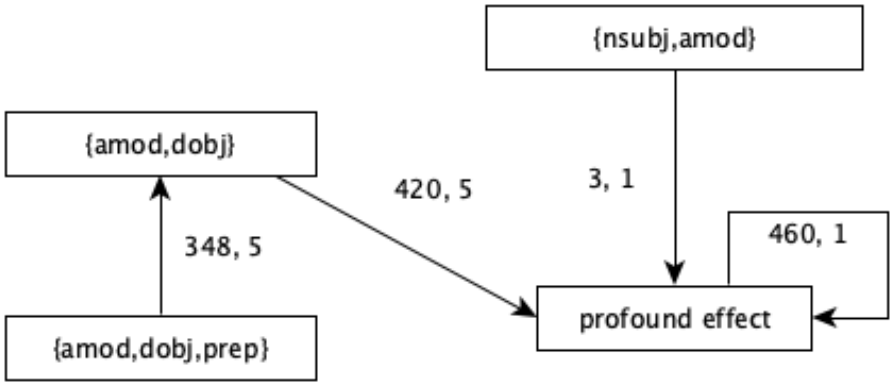


Figure 3.2: Corpus-attested valency patterns on the subsumed collocation *profound effect*

The examples discussed in this section show that both lexical and syntactic subsumption graphs provide an intuitive representation of such complex phenomena. As shown in the following sections, subsumption graphs can also be used to define the microstructure of entries in an automatic combinatorial dictionary.

3 Treelets

The last section of this chapter presents the first version of *Treelets* – a new application which implements the dependency-based phraseology extraction and visualization methods described above. The application is distributed freely as a Docker image and it can be used to extract one or more ACDs from a dependency-parsed reference corpus using user-defined dependency subtree patterns. The resulting ACDs can be searched through the built-in web application, exported or used directly as relational databases.

3.1 Corpus formats and metadata

The input formats currently supported by Treelets are: (1) plain text files with one text per line and (2) JSON Lines[8] format where each line contains a serialized dictionary with text metadata and contents. The metadata types supported include strings, floats, integers, text, dates and arrays of basic types and they can be explicitly imported into the corpus database using the second format. It is therefore possible to preserve the original structure of the imported corpus at the level of bibliographic annotation and use it to create filtering or aggregating queries against the corpus database (see Table 3.11). It is also possible to provide externally parsed texts in the CoNLL-U format. Plain text files can be dependency-parsed with one of the spaCy[9] or UDPipe models (**StrakaStraková2017**). The largest corpus indexed so far with Treelets contains 500 million words, but the database backend of the application is fairly scalable and it is possible to index larger corpora.

3.2 Defining extraction patterns

Once a dependency-parsed corpus database is created and indexed, it is possible to define a set of syntactic patterns to be used in the process of extracting a

[8] See http://jsonlines.org.
[9] See https://spacy.io.

combinatorial database. Table 3.8 shows the result of using different extraction rules predefined in Treelets. The last two columns of the table show the number of extracted treelets and their cumulative frequencies.

Table 3.8: A summary of 8 syntactic types of subtrees extracted from BNC

#	Pattern	Dependencies	Treelets	Occurrences
1	Adjectival modifiers	amod	405 385	4 090 711
2	Nouns with prep. objects	prep, pobj	368 357	1 906 280
3	Direct objects	dobj	275 037	2 562 868
4	Nominal subjects	nsubj	188 652	1 849 137
5	Adjectival modifiers as direct objects	amod, dobj	82 311	314 510
6	Nominal subjects with adj. modifiers	nsubj, amod	46 798	189 852
7	Adverbial mods. of adjectives	advmod	43 233	58 0021
8	Adjectival mods. as direct objects with prep.	amod, dobj, prep	20 082	70 545
9	Direct Objects with Prep. Objects	amod, dobj, prep, pobj	5 314	12 935

Custom extraction rules can be defined using the editor shown in Figure 3.3. In order to create a new extraction rule, which is essentially a dependency property subtree, it is necessary to define a directed tree graph as well as the aggregation keys of its vertices and edges. By default, the aggregation key is a combination of lemmas, part of speech tags and dependency types defined on the edges of the graph.

The rule shown in Figure 3.3 illustrates a special feature of Treelets which was implemented to deal with possible peculiarities of different dependency treebank annotation schemes. As hinted above, the validity of the continuity restraint, which assumes that phraseological units are lexicalized subtrees of the sentence dependency tree, may depend on the details of the dependency formalism. For example, in the current version of the Universal Dependency framework prepositions may function as case markers of their nominal heads. This means that

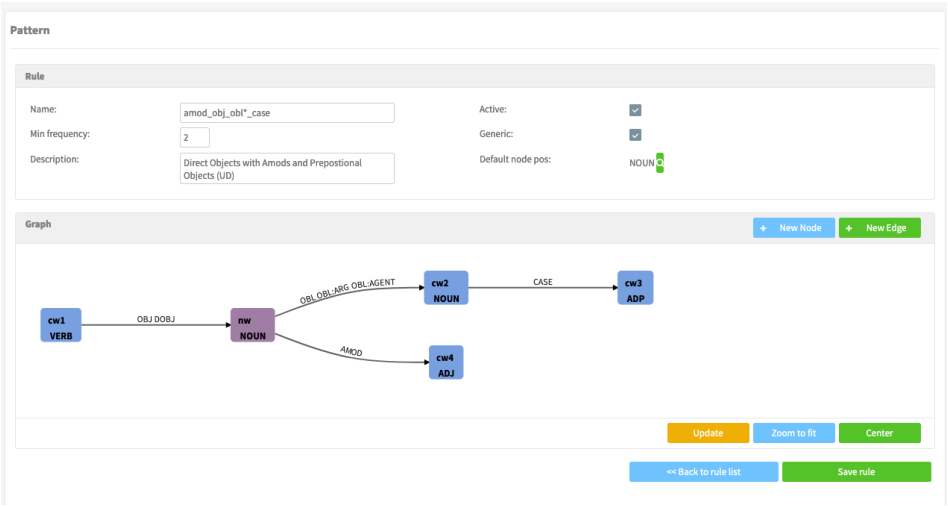


Figure 3.3: Designing extraction patterns in Treelets

the continuity restraint is not preserved for constructions such as *have a direct bearing on*. This is because the governor nominal node whose "case is marked" by the preposition *on* according to this representation is not an "obligatory" or "typical" lexical node of this expression.[10] To deal with such discontinuities, it is possible to use part-of-speech tags rather than lemmas to define the aggregation keys on whose values the extraction pattern is aggregated. In the example extraction pattern shown in Figure 3.3, the aggregation key of the vertex marked as *cw2* is therefore simply its part-of-speech tag (NOUN) rather than a combination of the tag and a lemma found on this vertex. In other words, the lemmas on the *cw2* vertex are ignored in the aggregation process and constructions such as *have a direct bearing on + NOUN* are counted as instances of the same lexicalized pattern. It is possible to test such rules on a selected reference corpus before using them for extraction. Once the ACD is generated it can be searched for both headwords such as *role* as a noun and specific treelets of arbitrary length in which it is found such as *play a specific role in the development of*. One of the results of a single query term for the string *role* in the ACD search field is a view similar to Table 3.9, which summarizes the syntactic types of the lexicalized treelets of the noun *role* in the ACD extracted from BNC[11].

[10] This example also shows that the results of relational phraseology extraction depend on the syntactic framework used to annotate a given corpus.

[11] Currently, syntactic variants of recurrent treelets such as *high hopes* vs *hopes were/are high* are not explicitly related in the underlying database. However, they are usually dynamically

Table 3.9: A summary of recurrent subtrees containing the noun *role*

#	Rule	Treelets	Examples
1	Nouns with prep. objects	1 182	*role of state, role in society, role in process*
2	Adjectival modifiers	504	*important role, major role, key role*
3	Adjectival mods. as direct objects	367	*play important role, play major role, play key role*
4	Direct objects	334	*meet role, play role, have role*
5	Adjectival mods. as direct objects with preps.	168	*play important role in, play major role in, play key role in*
6	Nominal subjects	106	*role be, role have*
7	Direct objects with prep. objects	91	*play important role in development, play central role in development*
8	Nominal subjects with adjectival mods.	53	*initial role be, former role be, final role be*

By clicking on a matching treelet, users are redirected to its dedicated page which currently consists of the following four sections:

- The concordances of the recurrent treelet in the reference corpus;
- The Statistics table with some statistical properties of the treelet, such as frequency, dispersion and strength of association;
- The dependency structure of the candidate construction;
- The Valency section, which features a tabular view of the directly subsumed and directly subsuming recurrent treelets. For example, the binary collocation *important role* is hyperlinked to the entry page for *play an important role*, which is linked to the entry for *play an important role in*, etc.

related in user queries. For example a search for the lemma *hope* will return both of the above-mentioned syntactic configurations of high + hopes in the summary table of results similar to Table 3.9.

The Valency section showcases a simple application of the subsumption graph structure of the ACD entries generated with Treelets. More sophisticated representations of the higher-order constructions detected with this application are discussed in the next section.

The current version of Treelets also supports extraction bases on untyped dependency tree patterns. In this mode, users only define lemmas for which all dependency subtrees up to a certain size (the current limit being six nodes) are extracted, aggregated and ordered by their frequency. In other words, only the *shape* of extracted subtrees, i.e. directed edges between the nodes, is predefined in this case. Table 3.10 shows the results of such ad-hoc extraction of recurrent subtrees containing the noun *factor* in BNC. Combinations of nouns joined by a preposition turn out to be the most productive pattern in which *factor* is found in this corpus with 254 distinct recurrent treelets identified. The largest number of instances is yielded by combinations of adjectives modifying this noun with 7962 occurrences identified.

Table 3.10: Weakly-supervised extraction of dependency subtrees

#	Structure	Subtrees	Mass	Examples
1	v2-prep-v1-pobj->v0	254	3104	number + of + factor one + of + factor depend + on + factor
2	v0-amod->v1	218	7962	key + factor important + factor major + factor
3	v1-dobj->v0	153	1087	take + factor identify + factor consider + factor
4	v1-nsubj->v0	120	2730	factor + be factor + include factor + influence
5	v2-pobj-v0-amod->v1	74	1292	of + other + factor by + other + factor of + important + factor

3.3 Exploring valency patterns

To illustrate the exploratory potential of subsumption graph visualizations, let us consider graphs generated for two entries from two different corpora. Figure 3.4 shows a subsumption graph generated for the noun *role* from BNC using the eight extraction rules mentioned above. Only subtrees which occurred in this corpus at least two times are shown in this graph. The two vertices with the highest indegree in this graph represent the direct object binary collocations *play a role* and *have a role*.

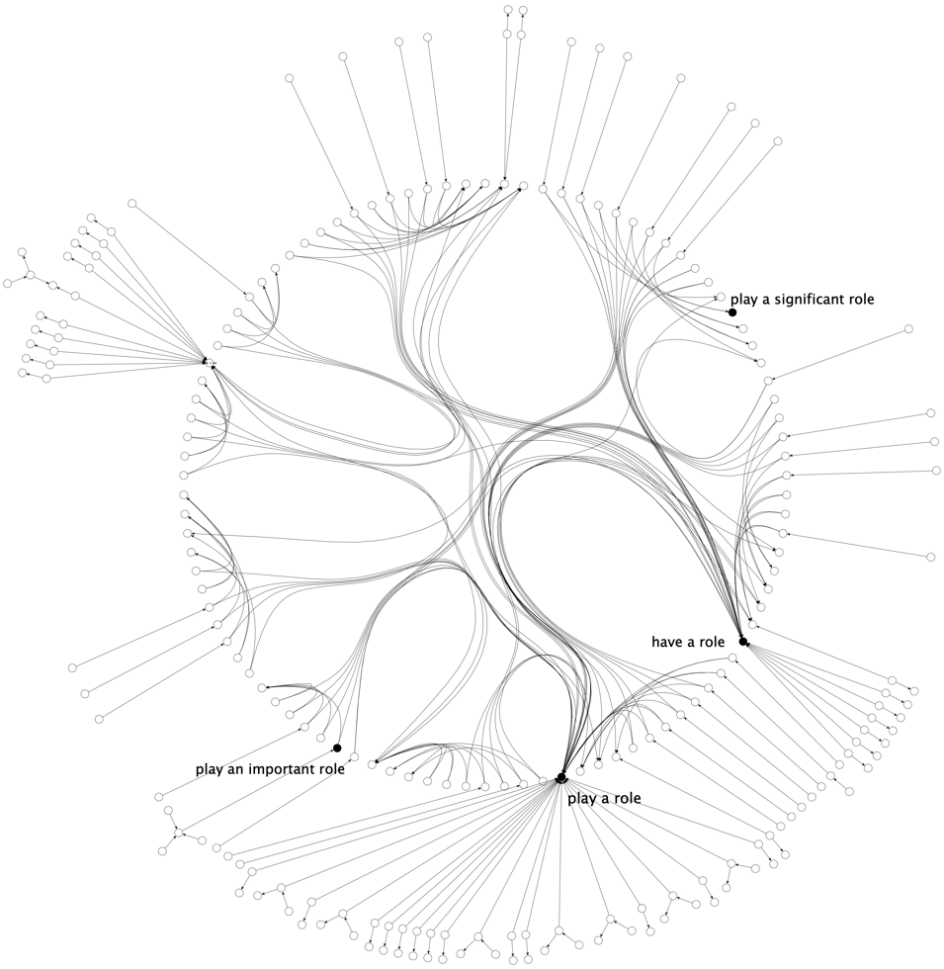


Figure 3.4: A subsumption graph of *role* as a noun generated from BNC. Only selected treelets are labelled

Table 3.11: Frequency and *productivity* of play/have a role in COCA and BNC

	Indegree		Frequency	
ACD	BNC	COCA	BNC	COCA
play a role	99	292	2 591	16 225
have a role	96	136	1 042	2 955

As shown in Table 3.11, *play a role* is considerably more frequent than *have a role* in COCA and only slightly more frequent in BNC. The indegrees of those vertices in the subsumption graph suggest that *play a role* is used in various larger constructions in COCA and BNC. The exact structure of the subsumption graph depends on the size and composition of the corpus, especially if a raw frequency threshold is used to select the nodes. In principle, it is also possible to use any conjunction of frequency, strength of association, dispersion, independence scores to create more sophisticated criteria of selecting the vertices of a subsumption graph. For example, a subsumption graph could contain only vertices representing subtrees whose average strength of association score is greater than some minimum *significance* threshold. Even a simple frequency-based subsumption graph may be helpful in formulating hypotheses to explain the differences between the varieties of English represented by the three corpora. One such hypotheses could be that *play a role* is more frequent in American English than in British English or simply in the registers and text types represented in the two reference corpora.

3.4 Source databases

The corpora and ACDs generated by Treelets are stored in a PostgreSQL database, which can be installed on any machine and in any location specified by the user. This means that more technical users can take full advantage of the dependency-parsed corpus database by querying it directly and from different client applications if necessary to obtain data views which have not yet been implemented in the Treelets web application. One example of such a query against the source corpus database is shown in Figure 3.5. The purpose of the query is to find dependency subtrees which consist of the preposition *of*, the noun *force* as its object and an unspecified adjectival modifier or compound noun dependent of this object. Furthermore, the search is limited to texts which are marked as “SPOK” (spoken

register in the imported COCA corpus). The results are aggregated on lemmas and dependency types, counted and limited to matching prepositional phrases which only occur in at least 20 different spoken texts in the corpus.

```
SELECT
wc1.head_lemma gov_lemma,  wc2.lemma dep_lemma,  wc1.lemma lemma,
wc1.dep dep, wc2.dep dep_dep, count(*) cnt, count(distinct(wc1.
    text_id)) texts

FROM
word_conllu wc1
JOIN word_conllu wc2 on wc2.sentence_id = wc1.sentence_id and wc2.
    head_id = wc1.id
JOIN text t on t.id = wc1.text_id

WHERE
wc1.lemma = 'force' AND wc1.head_lemma='of'
AND wc2.dep = ANY(ARRAY['amod','compound']) AND wc1.dep = ANY(ARRAY[
    'pobj'])
AND t.genre = ANY(ARRAY['SPOK'])
GROUP BY wc1.lemma,wc1.head_lemma, wc1.dep,dep_dep,dep_lemma
having count(distinct(wc1.text_id)) >= 50
order by cnt desc;
```

Figure 3.5: An SQL query used to extract and aggregate recurrent prepositional attachments from the Treelets corpus database

As shown in Table 3.12, the most frequent prepositional attachments identified with this query include *of armed forces*, *of military forces*, *of (det) work force*, etc. The query could be easily elaborated to identify more complex subtrees or return concordances of matching spans instead of aggregated summaries. It is also possible to relax the dependency joins defined in the query and use positional cooccurrence criteria in order to increase the recall of queries by matching unspecified or erroneously parsed dependency relations.

4 Summary and future work

The starting assumption of this paper was the vast majority of phraseological units have an internal syntactic structure and that subsentential PUs an external

Table 3.12: Recurrent prepositional attachments retrieved from the Treelets corpus database

gov_lemma	dep_lemma	lemma	dep	dep_dep	cnt	texts
of	armed	force	pobj	amod	197	167
of	military	force	pobj	amod	174	155
of	work	force	pobj	compound	134	122
of	task	force	pobj	compound	133	112
of	U.S.	force	pobj	compound	124	112

valency. Using a dependency-based phraseology extraction approach, the paper then demonstrated how those properties of PUs can be at least partly accounted for in automatically combinatorial databases. It is hoped that the software tool implementing dependency-based phraseology extraction may help lexicographers and phraseologists "deal with the enormous structural variety of English idioms" (**CowieEtAl1993**) and possibly also explore prefabricated collocational chains as a noteworthy type of phraseological units. Future versions of Treelets will also include phraseology detection features (**Pęzik2018**) to enable more advanced reference corpus-based indexing of idiomatic expressions.

Part II

Formulaic language in language learning contexts

Chapter 4

Paths to formulaicity: How do L2 speakers internalise new formulaic material?

Stephen F. Cutler
Cardiff University

How are new formulaic expressions acquired and stored by L2 learners? Defining formulaicity with respect to the individual speaker's storage and processing of a given expression as a single holistic unit (Myles & Cordier, 2017; Wray, 2002), two potential routes are explored: the 'fusion' over time of individual words and 'holistic acquisition', where an expression is internalised as a single unit from the start.

Two studies exploring the route to acquisition are reported. L2 speakers are presented with novel target expressions to memorise, and their ease of recall, accuracy and fluency over time is monitored. These delivery features are used in combination to indicate particular stages of acquisition that may be associated with each route.

Study 1 contrasts analytical and holistic methods for introducing the targets. Study 2 explores methods for determining the holisticity and processing automaticity of the target expressions in the learners' output. Drawing on the results of these, a model for the acquisition and storage of formulaic expressions based on the 'superlemma' model of **SprengerEtAl2006** is presented and discussed in relation to fusion and holistic acquisition.

1 Introduction

1.1 Background

Formulaic expressions are widely used by native speakers and have been shown to bring benefits in terms of fluency and speed of processing (Siyanova-Chanturia and Van Lancker **Sidtis2019**; **TowellEtAl1996**; Wray 2002). The holistic nature of such expressions is thought to contribute to efficiencies in processing that enable the fluent, connected multi-clause discourse of native speakers (Pawley & Syder 1983; **TremblayBaayen2010**). However, a variety of research suggests that, despite these and other benefits, L2 speakers of English do not use formulaic sequences to anything like the extent of native speakers (**Granger2019**; **Meunier2012**; **PaquotGranger2012**). Reasons given include a lack of sufficient exposure and a failure to notice that expressions may have a holistic nature. Other explanations (**Wray2019**) are related to the different ways that native and L2 speakers may approach language learning. For example, **DabrowskaLieven2005** have shown that children learn many multiword sequences as single units in their L1 and Wray (2002) suggests that native speaker continue to acquire formulaic expressions as whole expressions, only later breaking them down for analysis if the need arises. On the other hand, **WrayPerkins2000** have suggested that there may be a tendency for adult L2 learners to explicitly analyse any new expression in terms of its component parts.

The extent to which L2 learners use such an approach and the effect this has on the way that formulaic expressions are internalised has not been widely researched. There are some studies (e.g. Myles, Hooper and Mitchell, 1998) which have shown that L2 learners in a classroom situation do learn and use some formulaic expressions as whole units without initially attending to their component parts. However, **SchmittCarter2004** suggest that formulaic expressions are not always learnt in an all-or-nothing way. For example, studies in which L2 learners have specifically memorised sequences as whole units (**BoersLindstromberg2012**; **WrayFitzpatrick2008**) have shown that on-line reconstruction of the learned expression frequently takes place during recall and production, at least during some stages of the acquisition process. Research by **Bardovi-Harlig2019** in the field of second language pragmatics proposes an acquisition process for L2 learners whereby conventional expressions go through stages of becoming more target-like in terms of form and appropriacy to context.

These findings suggest that the way that speakers internalise new sequences can vary. In particular, two broad routes may be hypothesised: 'holistic acquisition', whereby a common sequence appears to be learnt and processed as a

single holistic unit immediately, and 'fusion' whereby an often used expression, initially constructed, becomes formulaic by regular usage to join the components into a single whole and fine-tune usage in terms of accuracy or appropriacy.

This chapter explores these different possible processes for internalising the sequences through two exploratory studies in which L2 speakers memorise target multi-word sequences. The first empirical study compares two different methods of learning and measures the effect these have on how expressions become formulaic for a speaker over time. The second study further explores how formulaicity may be identified in the context of an explicit model of internal representation for formulaic expressions. Findings from these studies are brought together in a discussion of possible models of acquisition.

1.2 Internal formulaic expressions

Many different terms and definitions have been used for formulaic language reflecting the different requirements of different schools of enquiry. An important distinction highlighted by **Wray2008** is between externally-defined sequences that are considered to be formulaic 'in the language' (such as idioms and high frequency multiword units) and those which may be 'psycholinguistic' units in the lexicon of the individual speaker. Some researchers (**Dahlmann2009**; **Erman2007**) have shown that these are not necessarily the same, particularly for L2 speakers. For example, an L2 speaker may know of a particular idiom (which is formulaic in the language) but not be able to use it smoothly. At the same time, a specific non-idiomatic expression (such as 'I'm an actuary in the Finance Department') may become psycholinguistically formulaic for that speaker (because it is relevant and often-repeated) while not being considered generally formulaic. Tabossi, Fanari, & **Wolf2009** suggest formulaicity of an expression for an individual speaker depends on the degree of familiarity or experience with the sequence and the way it has been learned.

Formulaicity in this sense therefore relates primarily to the way a particular expression has been internalised by the individual. A useful definition for this internal formulaicity is given by **MylesCordier2017**. They define an internally formulaic expression (which they term a 'Processing Unit') as:

> a multiword semantic/functional unit that presents a processing advantage for a given speaker, either because it is stored whole in their lexicon or because it is highly automatised.

This definition highlights the processing advantage of a formulaic expression (compared to a sequence constructed on-line) and defines the source of this to be

either holistic storage in the lexicon or automaticity. The concept of holistic storage, while potentially useful as a way of representing the unitary nature of formulaic expressions, has been challenged on empirical grounds (**Siyanova-Chanturia2015**). A key challenge is the finding from a variety of studies that when formulaic expressions are processed, their component words and structures are also accessed. This has been shown for idioms (**SprengerEtAl2006**) and frequent multiword expressions (Arnon and Cohen **Priva2014**). For example, **SprengerEtAl2006** ran a series of priming experiments that analysed response times for producing idioms. These showed that idiomatic (non-compositional) sequences (e.g. *'hit the road'*) both primed and were primed by constituent words in the sequence (e.g. '*road'*) and that the literal word meaning of the component word becomes active during idiom production. In order to accommodate this, they propose a model where the formulaic expression is represented by a 'superlemma' which "is a representation of the syntactical properties of the idiom that is connected to its building blocks, the simple lemmas" (p.176) by associative links in memory. In this way, the selection and processing of an idiom is similar to the processing of a single word in terms of lexical competition and co-activation. At the same time, it retains the idea that formulaic expressions have a syntactic structure related to the individual constituents at the lexico-syntactic level. This model therefore provides a good starting point for exploring the acquisition of internal formulaic expressions and is described in more detail in 3.2.

1.3 Exploring acquisition through targeted memorization

A useful way to investigate the acquisition of internal formulaic expressions by L2 speakers is through the targeted memorisation of novel expressions. Variations on such an approach have been used by **Wray2004** and **FitzpatrickWray2006** although not with a specific focus on internal formulaicity. In order to extend the targeted memorisation approach to investigate different paths to formulaicity, it is necessary to establish a way of identifying formulaicity in spoken output. Although it is not possible to observe cognitive attributes such as holistic storage or automatization directly, a common approach for identifying formulaicity has been to use sequence fluency. For example, in studies by **Erman2007** and **Dahlmann2009**, the absence of disfluency markers (such as pauses, hesitation, and repetition) was used as a criterion for formulaicity. More recently, **MylesCordier2017** have developed a set of criteria for the internal formulaicity of a sequence whereby fluency (indicating phonological coherence) along with evidence of its unitary nature (such as grammatical irregularity or semantic opacity) are the two necessary conditions. A sufficient criterion for satisfying the sec-

ond condition is that the learner has experienced the sequence as a unitary form with a given meaning. Therefore, in the specific case of targeted memorisation, the approach of **MylesCordier2017** effectively equates internal formulaicity with consistent fluency of delivery of the sequence at the time of testing.

The memorisation study by **FitzpatrickWray2006** highlighted considerable individual differences between participants in how they approached the process of memorising target sequences. Choosing how to control the input method is therefore an important consideration, since it is likely to have a significant effect on the learning outcome. For example, the principle of Transfer Appropriate Processing (**RoedigerEtAl2002**) proposes that any processing strategy is linked to a particular outcome. **Craik2002** states that encoding and retrieval are integrated in such a way that the initial processes determine the qualitative nature of the trait encoded. Barcroft (2002, 2006), exploring processing specificity, has shown that semantic, formal and mapping components are three separate and dissociable processes, and focusing on any one may take resources from the others. In general, elaborative approaches (strategies that facilitate an increased evaluation of an item with respect to particular features such as its meaning or structure) have been shown to increase learning with respect to that feature. For the intentional learning of formulaic expressions, different forms of semantic or formal elaboration have been suggested. These include: drawing attention to L1 congruence (**ConklinCarrol2019**), analysing component words and structure through matching or cloze style activities (**BoersEtAl2014**); linking metaphorical meanings of non-compositional idioms (**BoersEtAl2007**); and utilizing imageability (**SteinelEtAl2007**). These may lead to learning benefits in terms of long-term recall and accuracy, but their effect on fluency is not clear.

Insofar as internal formulaicity is defined in terms of holisticity and identified by delivery features such fluency, approaches to memorisation that are geared towards this outcome may be more effective in promoting 'holistic acquisition'. A key means of achieving fluency in a targeted sequence has been shown to be oral repetition. For example, **Nelson1977** demonstrated that repetition "at the phonemic depth of processing" facilitates memory for cued and un-cued recall and for recognition. **YoshimuraMacWhinney2007** showed that oral production fluency increases with the number of repetitions. The way in which the repetition is conducted is also important. Research into the effective learning processes of Chinese students (**AuEntwhistle1999**) suggests that rote memorisation is more effective if it is accompanied by a link with meaning. A study by **Ding2007** reported that a learning task involving the memorisation of a film script by copying a DVD was effective because the learners were being fully attentive to an imitation process. **NoiceNoice2006** researched how actors are able to learn their lines.

They showed that, for the non-actors participating in their study, the strategy of 'actively experiencing' the line as it was being spoken was more effective for accurate, fluent recall and reproduction than other memorising strategies. These kinds of repetition strategy may therefore be appropriate for achieving accurate acquisition of the complete phonological form while at the same time providing a strong automatic link to overall meaning and context.

2 Study 1: Comparing paths to formulaicity

2.1 Overview

The first study explores possible routes towards internal formulaicity by having L2 speakers memorise new target sequences via two different approaches. The first, Dynamic Repetition (DR), focuses on accurate and fluent reproduction of the sequences, while the second, Semantic-Formal Elaboration (SFE), is a more elaborative approach focusing on meaning and form. The effect of these initial processing strategies on formulaicity is assessed over time in terms of the fluency and accuracy with which the expressions are recalled.

Following the approach of Myles and Cordier (2017 as outlined in §**??**, internal formulaicity is indicated by the fluent delivery of the target sequence on recall. On this basis, it was hypothesised that the DR approach to learning was more likely to induce 'holistic acquisition' (as indicated by a target becoming internally formulaic immediately after initial learning).

After the initial learning phase, accuracy and fluency of recall were also tested after one and three weeks using a controlled series of recall tasks. As well as a means for checking internal formulaicity over time, these were designed to provide additional practice of the targets in a consistent way, allowing for the possibility of acquisition by 'fusion' (as indicated by a target becoming formulaic at a later stage).

- – **2.2 Method**

2.2.1 Participants

Ten Japanese speakers of English (JSE) at an intermediate/advanced level of English were recruited. There were nine females and one male, with ages from 28 to 45 and recent TOEIC scores (**ETS2019**) ranging from 760 to 940. All were

working adults chosen based on availability, level and because they were interested to take part. Full ethical procedures were followed in the collection of data and pseudonyms used when reporting on individual contributions.

- **2.2.2 Design**

The target sequences to be memorised are listed in Table **??**. All were verb phrases of 4 or 5 words selected from the Phrases in English (PIE) on-line corpus (**Fletcher2011**). Each had high frequency lexical words (with no repetition of these across the sequences), was non-congruent with the L1 Japanese. The sequences were confirmed to be unknown to the participants via an on-line check which involved them completing a cloze-style test and a check of recognition. The sequences were embedded in 4 stories (each of about 150 words) and the stories were paired to form two sets (AB and CD) of 6 sequences each. Sequences were balanced across the sets for length (words and syllables). Each story was assigned a suitable picture as a visual cue.

Table 4.1: List of target sequences

Set 1 (AB)	A1	turned a blind eye to
	A2	came to a head
	A3	breathed a sigh of relief
	B1	run the risk of
	B2	go a long way towards
	B3	like the sound of
Set 2 (CD)	C1	set his sights on
	C2	stood the test of time
	C3	get the hang of
	D1	knew better than to
	D2	toyed with the idea of
	D3	remains to be seen

To mitigate against the possible confounding effect of differences between participants or sequence memorability, a cross-over design was used whereby participants, sequences and order of learning were balanced across the two conditions. To facilitate this, participants were randomly assigned to one of four groups, as shown in Table **??**.

Table 4.2: Ordering of sequences and conditions by participant group

	1st	**2nd**
P1	AB (DR)	CD (SFE)
P2	AB (SFE)	CD (DR)
P3	CD (DR)	AB (SFE)
P4	CD (SFE)	AB (DR)

- **2.2.3 Procedure**

Each participant listened to a story (A or C) without any script, but while looking at the picture (to provide a cue for later). The three sequences in that story were introduced for learning either using DR or SFE (described below). The process was repeated for the second story (B or D), using the same method. The six sequences were then tested for recall (see §**??**). Next the procedure was repeated for the other two stories, with the sequences learned using the other method. The time given for memorisation of targets was the same for both conditions (18 minutes for 6 sequences). After all sets had been learnt and assessed, the participants listened to each story once more. Following a 10-minute break, there was a further assessment to establish performance at the end of the learning session (W0). After one week and three weeks, participants were given further assessments (W1 and W3 respectively).

The input sessions varied according to the condition as follows:

2.2.3.1 DR input

DR input approach focusses on consistent repetition of the expression with an emphasis on accurate imitation of prosody, intonation and rhythm, and 'active experiencing' of the sequences. The basic meaning of the expressions is provided by the story and the translations but is not further elaborated on. For each sequence, participants listened to the full sentence containing it and read a translation to check meaning. They then did a series of repetitions of the sequence following the exact intonation and rhythm of the model provided. (Where necessary this was slowed down to ensure accuracy). They interspersed this with repeating the whole sentence and also practised responding quickly to the Japanese translation of the sequence (as a cue card). Participants were encouraged to mimic the exact prosody and intonation of the delivery whenever they repeated each expression and "to imagine they were performing in a radio play". All engaged willingly with the process and appeared to enjoy doing it.

2.2.3.2 Semantic/Formal Elaboration (SFE) input:

SFE consisted of a generative exercise followed by some form-meaning tasks relating to the components, structure and meaning of the sequence. After listening to the story, participants were given a gap fill exercise based on the story script to try to generate the sequences. After finishing, they corrected this using the answer script and repeated each sequence out loud. They then did exercises looking at the structure of each sequence (count the verbs and nouns) and compared the sequence with its Japanese translation by rating their 'closeness' (in terms of words used). They were also asked to consider what might help them remember each sequence (e.g. particular words or images) and wrote example sentences for each which were then corrected if necessary by the researcher.

2.2.4 Assessments and measures

The same set of assessment tasks was applied at all stages:

1. ***Context* recall**: Given the picture and title, the participant retells the story trying to use the target expressions.
2. ***Cued* recall:** Cue cards (featuring the L1 translation of each sequence) are presented in random order and the participant recalls the appropriate sequence out loud. If they cannot do so, the researcher says the first word as a further cue.
3. ***Written* recall:** Participant writes down the expressions given the L1 translation
4. ***Read* out loud:** Each target is presented on a computer screen in random order and the participant repeats it.

The assessments were recorded, transcribed and analysed to calculate a variety of measures for each participant-sequence. For reasons of space, the current report focuses only on the context and cued recall tasks and on the following measures:

- ***Recall:*** The sequence was deemed to have been recalled if over 70% of the words matched the target on either of the recall attempts (context or cued).
- ***Accuracy:*** The sequence was considered 'fully accurate' if it exactly matched the target on either of the recall attempts.

- ***Fluency:*** For each recall attempt, any pause (>0.2s), reformulation, filler or hesitation was marked as a dysfluency. The sequence was considered 'consistently fluent' if it was delivered with no dysfluencies and with consistent form across the recall attempts.

The context and cued recall tests provide two different opportunities for the participants to recall and speak the expressions. The measures here are based on the combined responses to both tasks.

2.3 Results and key points

2.3.1 Summary of results

Overall, the ten participants, each learning six sequences via DR and six via SFE provided 120 participant-sequence combinations (60 for each condition). The numbers of sequences that are recalled (R-#), fully accurate (A-#) and consistently fluent (F-#) by condition and assessment phase across the two conditions at each of the assessments are given in Table **??**. The table also gives the cross-participant mean proportion of recalled sequences that were consistently fluent (Mean-F)

Table 4.3: Recall, accuracy and fluency by condition and assessment phase

Phase	Cond	R-#	A-#	F-#	Mean-F
W0	DR	47/60 (78%)	39/60 (65%)	19/47 (40%)	0.458 (sd=0.244)
	SFE	52/60 (87%)	39/60 (65%)	11/52 (21%)	0.215 (sd=0.152)
W1	DR	39/60 (65%)	33 (55%)	15/39 (38%)	0.398 (sd=0.314)
	SFE	37/60 (62%)	23/60 (38%)	7/37 (19%)	0.167 (sd=0.236)
W3	DR	49/60 (82%)	40/60 (67%)	21/49 (42%)	0.448 (sd=0.233)
	SFE	50/60 (83%)	39/60 (65%)	12/50 (24%)	0.258 (sd=0.262)

sd = standard deviation

The initial effect of the two input methods can be seen in the results immediately after learning (W0). These show that recall is slightly better for sequences learnt via SFE, while accuracy is similar across the two condition. For fluency, the proportion of recalled sequences that are consistently fluent (F-#) is higher in the DR condition (40%) than in SFE (21%). This difference is also evident in the Mean-F scores. A Wilcoxon Signed Ranks test indicated that the proportion of fluent sequences for the DF condition at W0 was significantly higher than for the SFE condition (Z=-2.801, p=.00256).

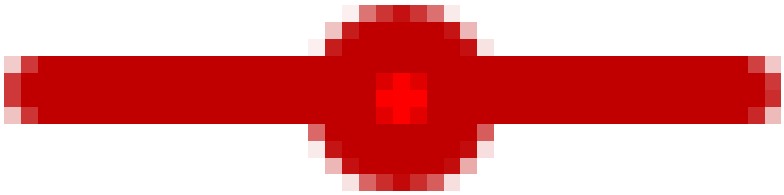
For the subsequent assessments, the general pattern of results for recall, accuracy and fluency is for a dip from W0 to W1 followed by a return to earlier levels at W3. This can be seen graphically in Figure 4.1.

2.3.2 Evidence of holistic acquisition and fusion

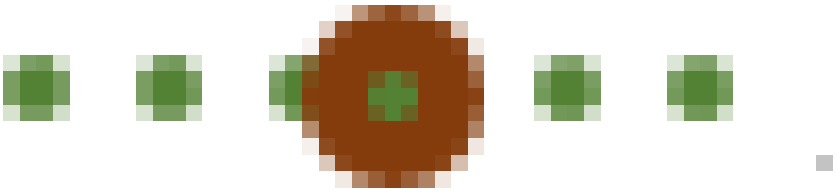
Following the approach of **MylesCordier2017**, fluent, consistent delivery of a sequence is considered a potential indicator of its internal formulaicity for that speaker. F-# therefore provides a count of such sequences. At W0, immediately after learning, the proportion of fluent targets was significantly higher for sequences learnt via DR than for those learnt by SFE, suggesting that the DR input did support holistic acquisition and may result in more expressions becoming formulaic for the speaker straight away. At the same time, this method did not appear to have a detrimental effect on recall or accuracy of the learnt expressions. However, while around 80% of the sequences were recalled at W0, even in the DR condition only about 40% of these were fully fluent. This may indicate limits on the numbers of sequences that can be memorised holistically in the given time period.

The results for the subsequent assessments suggest that similarities and differences between the conditions tended to remain over the three weeks. In particular, the proportion of fluent sequences for DR was still greater than that of SFE at W3. Although the overall numbers are small, the trend seems to suggest that the beneficial effects of DR are maintained over the longer term. On the other hand, the fact that the actual number of fully fluent sequences did not change much between W0 and W3, suggests that few additional sequences became formulaic for the speakers over the three weeks. For most targets, a reconstructive approach continued to be applied during recall. A typical pair of responses is given in Table **??**, where there is increased fluency and accuracy at W3, but without yet being sufficient for the target to be considered internally formulaic for the speaker.

Recalled (R-#)	Accurate (A-#)	Fluent (F-#)



DR



SFE

Figure 4.1: Proportions of recalled, fully accurate and consistently fluent sequences by condition and assessment phase

Table 4.4: Example 1 (Context recall responses by Kentaro for 'breathed a sigh of relief')

W0	he breathe on / breathe a / this one he breathe something / breathe / sigh / of relief / that …
W3	he breathed / a / bre- breathed a sigh of relief / because …

/ indicates points of dysfluency

It may be that the earlier assessment tasks at W0 and W1 (the only way the speakers had to 'practice' the expressions) were not sufficient to move more sequences into formulaicity at W3, but further practice could do so.

2.3.3 Overall recall performance over time

The overall pattern of performance in recall and accuracy of the target sequences was for a reduction in week 1 followed by an improvement in week 3. Since all participants confirmed that (as instructed) they had not reviewed the targets between tests, the overall reduction in performance at W1 is a not unexpected decay. However, the increased recall accuracy and fluency at W3 is more surprising. Since the only additional learning or review of the sequences following the initial input session was the W1 assessment check, the week 3 results suggest that this influenced the long-term learning. This interpretation supports work on spaced retrieval (**KornellEtAl2015**) which suggests that recall of learnt items (e.g. words learnt via flash cards) is enhanced by each attempt to retrieve them, and this effect occurs whether or not that attempt is successful, provided the correct answer is subsequently given. Although their work was not specifically on the learning of sequences, the retrieval conditions in the assessments used here were comparable. So, the repeated assessments may have supported the enhanced performance at week 3 as this was the fourth time the sequences were retrieved. Further, since the two retrieval attempts preceding W3 (W0 and W1) were spaced by a week while those preceding W1 were only spaced by 20–30 minutes in the initial session, the results may support research (**KornellVaughn2016**) that claims increasing retrieval spacing has a beneficial effect on learning.

3 Study 2: Exploring fluency and holistic automaticity

3.1 Introduction

While using fluency as an indicator of formulaicity follows a precedent set by previous research, it may be argued that fluency alone does not always imply holistic storage or automatic processing of the sequence (the defining features of internal formulaicity in the definition of **MylesCordier2017**). For example, **Segalowitz2010** argues that automaticity is more than a simple speeding up of cognitive processes; it involves a qualitative change in the way a process is organised or structured. Establishing a form of internal holisticity may represent such a qualitative difference for formulaic sequences when compared with simply constructing the sequence more and more fluently through repetition. Segalowitz

describes a type of automaticity linked with qualitative restructuring, terming it 'ballistic automaticity', based on the idea of automatic processing being unstoppable or involuntary.

Study 2 explores the idea that fluency may be a staging post towards formulaicity rather than necessarily the destination. Drawing together the ballistic automaticity described by **Segalowitz2010** and the representation of internal holisticity given by the model of **SprengerEtAl2006**, a psycholinguistic test for 'holistic automaticity' was used to determine the formulaicity of target sequences more explicitly. This was then applied to new data from the same group of 10 Japanese speakers of English who took part in Study 1. The aim was to determine the extent to which target sequences that are delivered consistently and fluently can also be shown to be automatic and holistic in the mind of the speaker. It also offered the opportunity to further explore the route to formulaicity of the original sequences.

3.2 Holistic Automaticity Test

In Holistic automaticity (HA), when the first word of a target sequence is activated (by hearing the word as an auditory prime), the speaker cannot help but process the whole sequence for potential speech production. In particular, subsequent words in the sequence will be activated and, given a suitable cue, preferentially selected over other candidate words in a word response test. The reasoning for this draws on the amended hybrid model of speech processing of **SprengerEtAl2006** introduced in §**??**

Meaning relationship
Associative link / co-activation

Figure 4.2: Adapted version of 'Superlemma' Model (**SprengerEtAl2006**, p.1760)

Figure 4.2 shows a simplified version of the model as applied to a target sequence from the current study. If the sequence is formulaic, the contention is that a superlemma (*get-the-hang-of)* exists which is linked to both to its conceptual meaning directly and to the lemmas of its constituent words via associative link. When the identity prime ('get') is heard, the lemma for *get* is activated which then activates the lemma for *get-the-hang-of*. This in turn activates the other constituent word lemmas, including the lemma for *hang*. When the letter cue ('h_') is then seen, it triggers a search for words beginning with 'h'. Since *hang* is already active, selection of this word is facilitated above other candidates.

3.3 Method

3.3.1 Target sequences

Along with the 12 sequences previously learnt by the participants in Study 1, six new control sequences were introduced. These were selected using the same principles as the originals and confirmed to be unknown to the participants.

For the HA testing, the initial verb of the sequence was taken as the prime and one of the key lexical words in the remainder of the sequence was the target word. For example, for the sequence *get the hang of*, *'get'* was the prime word and *hang* the target. Each sequence was to be presented twice: once with a cue letter corresponding to the target word (T-cue), once with a cue letter unconnected to the sequence (NT-cue). The list of sequences, primes and cue letters is given in the Appendix.

3.3.2 Procedure

3.3.2.1 Fluency assessment

To determine the current state of acquisition of the 12 sequences each participant undertook the same assessment (context recall and cued recall) given in Study 1. On this basis, participant-sequences were categorised into one of the following:

1. No recall (NoRec): The sequence was not recalled with sufficient accuracy in either task

1. Major dysfluency (D-major): Major or multiple disfluencies in either task
2. Minor dysfluency (D-minor): Only one minor disfluency in one or both tests
3. Fluent - low recall (F-low): Recalled on one test and fully fluent in that one
4. Fluent - high recall (F-high): Recalled on both tests and fully fluent and consistent in both

This categorisation was chosen to separate out those sequences that were judged phonologically coherent for that speaker at that time (4, 5) from those that were not (1, 2, 3). In addition, it enabled exploration of the extent to which ease of recall (of the whole sequence) and the 'degree' of fluency of a sequence may be relevant to automaticity. A minor dysfluency was defined as a single short pause (between 0.2s and 0.5s) occurring in one or both of the tests.

3.3.2.2 Brief review of the sequences

Following the assessment, the six control sequences were read out to the participant and then shown on a written list with a Japanese translation. The participant read each one out loud once to ensure it could be said smoothly with no pronunciation difficulties. After a short break, the participant was presented with all 18 target sequences in random order and asked to read each one out loud (to integrate the controls into the set of targets).

3.3.2.3 Introduction of response word controls

To provide some degree of control over the possible responses, a set of 40 words was introduced before the test. This was considered necessary to reduce the possibility that target words are chosen simply because of exposure to the target sequences during the earlier stages of the experiment. The 40 words contained 8 different starting letters which matched the range of cue letters of the test. All 18 target words were included along with 22 high frequency dummy words of similar form, resulting in 5 words for each initial letter.

Participant were presented with the words one-by-one on cards in random order. After repeating each one out loud, they performed a simple grouping exercise based on initial letter and repeated them again. After a break and immediately prior to the holistic test, a brief check was done in which the participants were presented with each cue letter and asked to say out loud any word they could think of. The purpose of this was to ascertain whether target words were preferentially in mind before the test.

3.3.2.4 HA test and analysis

The computer-based HA test consisted of 36 items (two for each target sequence). For each item, there was a fixation point on the screen accompanied by a beep. After 2.5s an auditory prime of the cue word (the first word of a sequence) was played and a further 750ms later, the cue letter appeared. Each spoken prime lasted between 500–600ms, leaving a short gap (150–250ms) before the letter cue was shown. The 36 items were presented in pseudo-random order to ensure that: (a) the two occurrences of each sequence were well separated, (b) the same cue letter was not repeated sequentially, and (c) cue letters did not follow presentation of a prime word with the same beginning letter. This was to minimise cross-item interference. Participants were given the following instruction:

> You will hear a word. You will then see a letter. Say a word beginning with that letter as quickly as you can. NOTE: You may like to use one of the

words introduced earlier but you don't have to. The aim is to respond as quickly as possible

The aim was to encourage participants to choose words from the list but without compelling them to think too consciously about it. Each test was recorded, and the participant response and response time (RT) noted for each item. To determine whether the target word had been activated and spoken quickly and in preference to other possibilities, a set of criteria was applied for each target sequence:

- The expected target word must be chosen in response to the T-cue.
- The RT for this word should be faster than that for the NT-cue word for the same prime.
- If there are other occasions when the same target word is given (i.e. as an NT-cue response to a different prime), all of these should also have slower RTs.

If all criteria were satisfied for a sequence for the participant, it was marked as a 'holistic hit'. To illustrate, Table **??** gives a typical example of a possible set of participant responses involving the prime 'get' and the response *hang* (for testing the sequence *get-the-hang-of*). In this example, the appropriate target response is given, and its RT is faster than any other response involving the prime 'get' or the response *hang*. So, it would be marked as a holistic hit.

Table 4.5: Example holistic test responses

Prime	**Cue**	**Response**	**RT (??)**
get	h_	hang	1.125s
get	b_	boy	1.491s
lie	h_	hang	1.662s
rolls	h_	hang	2.010s

3.4 Results

Across the ten participants, a total of 49 of the original sequences were deemed to be formulaic (23 low recall and 26 high recall), while 56 were non-formulaic (38 with major dysfluencies, 18 with a minor dysfluency) and 15 were not recalled at

all. This information was used to divide the results into categories for subsequent analysis.

In the word check test, 61% of responses were from the list of 40 control words given at the start of the session. Of these, 34% were target words from the original sequences and 16% were target words from the control sequences. These figures are close to the percentages expected if the words were chosen at random (12/40=30% and 6/40=15%, respectively). This was the anticipated result and confirmed that the target words were not preferentially activated before the test compared to other possible choices of words.

Table 4.6: Proportion of holistic hits over main categories

Sequence type			**N**	**Holistic Hits**	
Control			60	9 (15%)	
Not recalled			15	2 (13%)	
Dysfluent	Major	38	56	8 (21%)	15 (26%)
	Minor	18		7 (39%)	
Fluent	Low recall	23	49	12 (52%)	31 (63%)
	High recall	26		19 (73%)	

Table **??** gives the numbers and proportions of holistic hits across the sequence categories. As the table shows, the memorised sequences deemed formulaic by the criteria had a much higher percentage of holistic hits compared with non-formulaic learned sequences. The control sequence results are similar to those of the original sequences which were not recalled. Excluding the No Recall group, a Chi-square analysis comparing counts for Control, Non-F and Formulaic groups shows that the differences are significant (χ^2=25.257, p<0.00001) with Cramer's V=0.28, suggesting a medium to large effect (**Cohen1988**).

Looking at the more detailed categories, the proportion of holistic hits rose steadily from major dysfluency to minor dysfluency to fluent. Within sequences categorised as fully fluent, it rose from low recall to high recall. The results are suggestive that the likelihood of a sequence being holistically automatic increases the more fluent it appears to be and the more easily it is recalled. Figure **??** summarises the results, showing the continuous rise in holistic hits (representing holistic automaticity) through the categories (representing increasing degrees of fluency).

Figure 4.3: % of 'Holistic hits' per sequence type

3.5 Key points from Study 2

As would be expected if fluency is a necessary indicator of formulaicity, the fluent sequences had a significantly higher proportion of holistic hits than the dysfluent and control sequences. The proportion of hits rose steadily through the categories, suggesting that holistic automaticity may be sensitive to the relative fluency of the sequences and the ability to recall them. The results also showed that not all fluent sequences resulted in holistic hits. Qualitative analysis of the responses given in these cases suggests that this was not due to interference from an alternative expression the participant already knew (e.g. *come home*). Non-target response words were always control or other words without any obvious connection to the prime word (e.g. *hike* for the prime 'come' or *teeth* for 'stood'). The results therefore lend some support to the idea that automaticity may be a 'stronger' condition than fluency on the road to formulaicity, with some fluent sequences yet to have reached the holistic automaticity stage.

The HA test is necessarily probabilistic and, based on random (but appropriate) choices from the 40 control words and under the criteria for a holistic 'hit', the predicted false positive rate would be just under 10%. The percentage of hits for the Control group was higher than this and, although the numbers are small, may suggest that other factors may cause false positives. For example, it may be that some primes and targets are linked associatively (e.g. because they have been heard together before) even though the overall sequence is not formulaic. The rates of holistic hits for the dysfluent groups are discussed further in §**??**

An important finding from the initial assessment of the 12 original target sequences is that 105 (88%) of the participant-sequences were recalled and 49 (47%) of these were classified as fluent by virtue of being delivered fluently and consistently. This shows that the overall numbers for recall and fluency rose in the two months between the end of Study 1 and the start of Study 2 (S2). While it is possible that some participants experienced the sequences during the two months, this increase may be further evidence of a spaced retrieval effect as described in §**??** Regarding overall fluency change over time, the mean proportion of fluent sequences across participants rose from 0.298 at W0 to 0.446 at S2, and a Wilcoxon Signed Rank test showed that this increase was significant (z=-2.0896, p= .01831).

It is also interesting to note that, of the 49 fluent sequences at S2, 33 were originally learnt by DR and 16 by SFE. For the 31 fluent sequences that also had holistic hits, that ratio was consistent (21 to 10). This suggests that the long-term benefit (in terms of fluency and formulaicity) of the DR input is maintained.

4 Discussion: acquisition of targeted expressions over time

4.1 Patterns of acquisition

In the two studies, L2 speakers of English were given new expressions to learn and these were assessed at various points in time to determine the extent to which the expressions had become internally formulaic for the speakers. Overall, 31 participant-sequences (26%) were both fluent and demonstrated holistic automaticity at the time of Study 2. Assuming that these are cases where internal formulaicity has been attained, a closer look at them suggests some different potential routes to becoming that way. As found in Study 1, some expressions appeared to become formulaic straight away, particularly when learnt via the DR strategy. Throughout the assessments, these sequences remained more or less fluent and accurate, but varied in how consistently they were recalled.

Table **??** shows a typical example of such cases. The target is delivered fluently and accurately in context and cued tasks at W0. However, at W1 the participant requires a first word cue to deliver the expression, and at W3, she may be repeating the cue herself before delivering the sequence fluently.

Table 4.7: Example 2 (Responses from Kaori for 'run the risk of')

	Context recall	**Cued recall**
W0	now / run the risk of / losing staff	run the risk of
W1	<no recall>	RUN => run the risk of
W3	company / run? / run run /run the risk of / losing staff	run the risk of
S2	<no recall>	run the risk of

RUN=> indicates that the researcher gave the first word RUN as cue
/ indicates a point of dysfluency

Other sequences were not recalled fluently initially but became formulaic over time. This appeared to be facilitated by the practice and retrieval afforded by the regular assessments and suggests that some kind of fusion is taking place. Illustrative examples are given in Tables 8 and 9.

In each case, there is a mixture of fluent and dysfluent production (with the cued responses tending to be more fluent) and evidence of reconstruction at the

Table 4.8: Example 3 (Responses from Tetsuko for 'toyed with the idea of')

	Context recall	**Cued recall**
W0	he / he toyed / the idea of buying a new one	toyed with / with / the idea of / toyed with the idea of
W1	<no recall>	TOYED => / the idea of
W3	he / toyed / toyed with the idea of	toyed with the idea of
S2	he / toyed with the idea of buying a new one	toyed with the idea of

Table 4.9: Example 4 (Responses from Sachiko for 'set his sights on')

	Context recall	**Cued recall**
W0	he set his / sight on / inventing	SET=> set his / sights on
W1	he set his / he set his mind / of / creating a new game	set his / set his / mind / set his / target / it's not target
W3	then he set his / sights on / in- inventing new / games	set his sights on
S2	he / he set his sights on / inventing a new game	set his sights on

earlier stages. Example 3 shows how some words (*toyed*) and sub-sequences (*the idea of*) may be known and linked as part of the expression. In joining these together during reconstruction, non-lexical words ('*with*') may get missed out. Other examples from the studies include *turned blind eye* and *breathed sigh of relief*. Example 4 illustrates how existing knowledge, such as lexical associates of the component words (e.g., *mind*) or lemmas associated with the meaning (e.g. *target*), may interfere with reconstruction process. In these examples, the retrieval and corrective feedback of the assessments facilitated accurate fluent reproduction of the forms eventually. However, repetition without feedback could potentially lead to fossilisation of non-target formulaic forms.

4.2 Fluency, recall and 'degrees of formulaicity'

While the general trend was towards increased formulaicity over time, there was some inconsistency. For example, in Study 1, it was not always the case that fluency was maintained from one stage to the next. In Study 2, although the results showed that the more consistently fluent an expression was the more likely it was to also show holistic automaticity, there were still some dysfluent expressions which appeared to have holistic automaticity.

Apart from the likelihood of some false positives in the HA test (as described in §??), it may also be possible for some sequences to be formulaic for a speaker but sometimes delivered in a non-fluent way. In natural discourse, such pausing or hesitation may be for planning speech while holding one's turn (**Wray2019**) or for socio-pragmatic reasons, such as appearing sincere (**Bardovi-Harlig2019**). While these particular reasons for pausing are unlikely in the current context, they do highlight that speakers may choose to pause within formulaic material. A more common situation in the current studies is where the apparent dysfluency occurs because the speaker is trying to self-cue their recall of the whole sequence, as in Example 2 above. It may also be possible in cases such as "*blind eye / a blind eye // he turned a blind eye to her behaviour*" in which the self-cue is a sequence (*blind eye*) within the expression. Such responses were marked as dysfluent due to the reformulations (which indicate breaks in the sequence). However, it could also be that the sequence is holistically stored but not easily recalled on this occasion. This may parallel the tip of the tongue (TOT) phenomena (**EckeHall2013**) where aspects of a word can be recalled (e.g. the first letter) but not the whole word (even though the word is presumably holistically stored). The self-cue word or phrase may act as a label to the full expression, which may be linked phonologically (as in the case of TOT for words where the contributing part is a letter or phoneme) or via some other mnemonic.

This may be supported by the finding in Study 2 that holistic hits were far more likely when a sequence was easy to recall. An explanation is that, for low recall formulaic sequences, a 'superlemma' may exist but not (yet) be well-established in the lexicon (i.e. its connections with associated concepts and lemmas are still relatively few and weak). This could result in a lower level of activation in the HA test, making it more susceptible to interference from other more activated candidates. This reasoning could be extended to 'partially' formulaic sequences where a weakly established lemma may exist, but there still remains the possibility of a speaker reconstructing the sequence in situations where the whole lemma cannot be accessed from the given cue. In such a model therefore, identification of a sequence as 'formulaic' (via fluency or holistic automaticity) may depend not only on the existence of a holistic lemma but also on the strength and type of connections that that lemma has. This idea can explain the variation in holistic automaticity across categories, and also provides a way of understanding apparent 'degrees' of formulaicity within a holistic storage model such as that of Sprenger et al.

4.3 Modelling the routes to formulaicity

The model of Sprenger et al provides a useful way of showing how formulaic expressions may be represented in the mind, but it does not specifically address acquisition. While there do not appear to be any models of FL acquisition based on the 'superlemma', there are some more general models of vocabulary acquisition that may be adapted. For example, De Bot, Paribakht, & **Wesche1997** provide a structure for describing and explaining aspects of L2 word acquisition based on Levelt's model of speech processing (**??**). Levelt highlights the idea that the lemma has distinct elements including syntactic and semantic components which are, in turn, separate from the morphological and phonological components of the lexemes to which the lemma is linked. De **BotEtAl1997** suggest that when a learner encounters a new word, an 'empty' lemma structure is created. The learner then uses semantic and syntactic information from context (and morphological information from the lexeme depending on their experience of the language) to fill in this structure. This idea is extended by **Jiang2000** in his lemma mediation model of L2 vocabulary acquisition. He suggests that, in the initial stages of acquisition, the phonological (or written) form of the word is stored and a lexical entry created. The semantic and syntactic (and morphological) information is initially provided via associated links to the L1 translation or definition. This model has been applied to formulaic expressions in a study

by **YamashitaJiang2010** which applied the lemma mediation model to the acquisition of collocations by Japanese EFL and ESL speakers. In the context of the model, they took collocations to be holistic units with their own entry in the mental lexicon.

4.3.1 Modelling holistic acquisition

In terms of the models of De Bot et al and Jiang, an outline hypothesis is that the DR approach helps to create a holistic phonological form of the target expression in the mind of the speaker, facilitating the creation of an 'empty' lemma to which this lexeme is linked. The basic lemma structure is linked to the meaning (e.g. via the given L1 translation) and the context of the learning (via the story and the episodic memory of engaging with it). Holistic acquisition is achieved when there is sufficient targeted oral repetition of the sequence to create the (holistic) phonological form in memory and automate its retrieval given the appropriate elicitation cue. Accurate memorisation of the sequence in a fixed holistic form may then serve as a stable building block for further learning, to integrate semantic, syntactic and morphological aspects of the expression lemma.

Meaning relationship Associative link / co-activation
Dotted arrows indicate weaker links / dotted boxes indicate empty elements

Figure 4.4: Simple model of holistic acquisition for L2 speakers

Figure 4.4 presents a highly simplified model of this process, showing possible initial and final stages in the holistic acquisition of a formulaic expression. Initially, hearing the expression in context and seeing the L1 translation help set up the conceptual meaning. The holistic phonological form is established through the DR process and linked to the concept (and the strength of this may vary, as shown by the dotted arrow). The phonological form may also be linked associatively with phonological forms of words and sub-sequences, but direct links to their meanings are discouraged. As the target is retrieved and repeated over time, the link between the concept and the lemma are strengthened along with associative links to the lemmas of the component words and sub-sequences. This consolidates the holistic sequence lemma in memory and helps make it easier to recall

4.3.2 Modelling fusion

There were also cases of apparent 'fusion', where a sequence was initially reconstructed to some extent before later becoming formulaic. In many of these cases,

components and sub-sequences (e.g. *breathed* and *sigh of relief*); *turned* and *blind eye*) appear to be combined on-line, with dysfluencies marking their joins. In some cases, errors occur at the joins (*breathed his sigh of relief*; *turned blind eye*) usually involving less salient function words (e.g. *a*, *to*), or occasionally with the wrong choice of lexical word (e.g. *set his mind on*). There were also examples of morphological changes to the key lexical words (*breathe*; *turn*) compared to the given target. The morphological and lexical changes suggest that the meanings of the component words were being accessed during the reconstruction. Fusion therefore seems to involve a combination of the chunking together of known components and the correcting of erroneous or missing words. To some extent this process mirrors the latter stages of a sequence postulated by Bardovi-Harlig (2019, p.110) for the pragmatic L2 acquisition of 'conventional expressions':

> nontargetlike response [F0E0?] target-like response but non-target-like lexical resources [F0E0?] target-like lexical core [F0E0?] full conventional expression

In the fusion cases of the current study, the targeted learning of given expressions appears to move learners quickly to the 'target-like lexical core' stage, but further development is required to become fully formulaic. A possible model for this fusion process is given in Figure 4.5.

Meaning relationship Associative link / co-activation
Dotted arrows indicate weaker links / dotted boxes indicate empty elements

Figure 4.5: Simple model of fusion for L2 speakers

In the initial learning stage, while a conceptual meaning for the target expression may be established, it is not linked to a holistic lemma or single phonological form. To recreate the expression therefore, it is necessary to access the lemmas of the component words and sub-sequences which have been linked to the context and L1 translation (possibly via their conceptual meanings). So, while an 'empty' expression lemma may be created, it takes further retrieval and repetition to facilitate the chunking up and correcting required to develop a fused phonological form.

4.4 Conclusion

The two studies showed clear differences in the effect of the different learning approaches on internal formulaicity, along with useful insights into the acquisition

process. However, it should be acknowledged that the number of participants and target sequences tested is relatively small and representative of a specific type of learner and formulaic expression. For this reason, the research presented should be seen as exploratory and the results and conclusions would ideally be verified through larger scale studies and different types of learner. It is also important to emphasise that the approach and discussion focus on a particular definition of internal formulaicity and specific means of identifying it.

With those caveats in mind, the studies do nevertheless demonstrate that both holistic acquisition and fusion (as described here) are two possible routes for a target expression to become internally formulaic for a speaker. They further suggest that the method by which targets are memorised influences which of these routes is taken. Figures 5 and 6 show a possible way of modelling these route which are consistent with some existing models of lexical acquisition. They also show how apparently 'partial' formulaicity may be compatible with a model based on the idea of holistic storage. A particular implication is that, in the case of fusion, the meanings of component words and sequences are accessed in order to construct the expression, while this is not necessary for holistic acquisition. Fusion is therefore likely to be more susceptible to interference based on the speaker's existing knowledge of the component words or sub-sequences. Examples of this from the studies include cases where words may be strongly linked to other similar expressions (e.g. *like the idea of* for *like the sound of*) or when synonyms replace component words (e.g. *set his target on*). It also suggests that part of the benefits (for formulaic acquisition) of an approach such as Dynamic Repetition (DR) is that it de-emphasises the meanings of the component words. This is certainly beneficial in expressions where, like the targets in the studies, the whole is not (semantically) the sum of the parts.

With DR, the focus on repetition of the whole (delivered with sufficient intonation and feeling) may help to establish holistic storage of sequences early on, and maintain fluency and formulaicity of output over time. Further, because the whole is sufficiently linked to a particular example context and meaning, the simple repetition does not appear to impact negatively on recall or accuracy compared to the semantic-formal elaboration (SFE). While it was not designed as a pedagogic tool, DR in combination with certain elaborative approaches such as drawing attention to prosodic features of target sequences (**BoersEtAl2012**) may be a useful way of promoting formulaic acquisition. The studies also support the idea of regular (spaced) retrieval and simple corrective feedback as a way of consolidating recall and formulaicity of the learnt sequences.

Along with the way expressions are memorised, there are likely to be many other factors which could influence the extent to which target expressions will

become formulaic for a speaker and the route taken to do so. Indeed, despite the controlled choice of participants and sequences in the studies, there was still considerable variation in performance across participants and sequences. However, rather than any systematic trends for particular features of sequences (e.g. length) or participant (e.g. proficiency level), any variation appears more likely to be a complex interaction between these, related in part to the speaker's particular experience with the words and sub-sequences of each sequence. Further research which manipulates known or unknown component words and sub-sequences within target sequences when being learnt by L2 speakers may be useful to explore the routes to formulaicity further. It would also be interesting to investigate how other variables (such as length, prosodic features, imageability and L1 congruence) may affect these routes.

5 Appendix

List of target sequences, primes, cues and target words for the HA Test (Study 2)

	Sequence	Prime	T-cue	Target	NT-cue
A1	turned a blind eye to	turned	b	blind	f
A2	came to a head	came	h	head	b
A3	breathed a sigh of relief	breathed	r	relief	t
B1	run the risk of	run	r	risk	l
B2	go a long way towards	go	l	long	s
B3	like the sound of	like	s	sound	t
C1	set his sights on	set	s	sights	t
C2	stood the test of time	stood	t	test	i
C3	get the hang of	get	h	hang	r
D1	knew better than to	knew	b	better	r
D2	toyed with the idea of	toyed	i	idea	l
D3	remains to be seen	remains	s	seen	l
E1	look on the bright side	look	b	bright	f
E2	rolls off the tongue	rolls	t	tongue	h
E3	scared the life out of	scared	l	life	h
F1	walk on thin ice	walk	i	ice	s
F2	reserve the right to	reserve	r	right	i
F3	lie at the	lie	h	heart	f

Note: A1-D3 are the original sequences; E1-F3 the new controls

References

Bell, Melanie J. & Ingo Plag. 2012. Informativeness is a determinant of compound stress in English. *Journal of Linguistics* 48. 485–520.

Brinton, Laurel J. & Elizabeth C. Traugott. 2005. *Lexicalization and language change*. Cambridge: Cambridge University Press.

Bybee, Joan. 1998. The emergent lexicon. In Katherine M. Gruber, Kenneth Olson & Tamra Wysocki (eds.), *Papers from the 34th annual meeting of the chicago linguistics society*, vol. 2, 421–435. Chicago: Chicago Linguistic Society.

Bybee, Joan. 2002. Sequentiality as the basis of constituent structure. In Talmy Givón & Bertram F. Malle (eds.), *The evolution of language from pre-language*, 109–134. Amsterdam: John Benjamins.

Bybee, Joan. 2003. Mechanisms of change in grammaticization: The role of frequency. In Richard Janda & Brian Joseph (eds.), *Handbook of historical linguistics*, 602–623. Oxford: Blackwell Publishers.

Bybee, Joan. 2014. Analytic and holistic processing in the development of constructions. In Inbal Arnon, Marisa Casillas, Chigusa Kurumada & Bruno Estigarribia (eds.), *Language in interaction: Studies in honor of eve v. Clark*, 303–313. Amsterdam & Philadelphia: John Benjamins Publishing Co.

Bybee, Joan & David Eddington. 2006. A usage-based approach to Spanish verbs of 'becoming'. *Language* 82. 323–355.

Bybee, Joan & Carol Lynn Moder. 2017. Chunking and changes in compositionality in context. In Marianne Hundt, Sandra Mollin & Simone E. Pfenninger (eds.), *The changing English language: Psycholinguistic perspectives*, 148–170. Cambridge: Cambridge University Press.

Carrol, Gareth & Kathy Conklin. 2014. *Eye-tracking multi-word units: some methodological questions. Journal of Eye Movement Research* 7(5). 1–11.

Corrigan, Roberta, Edith Moravcsik, Hamid Ouali & Kathleen Wheatley. 2009. Introduction: Approaches to the study of formulae. In R. Corrigan, Edith Moravcsik, Hamid Ouali & Kathleen Wheatley (eds.), *Formulaic language*, vol. 1: Distributiona and historical change (Typological Studies in Language 82), xi–xxiv. Amsterdam/Philadelphia.

Croft, William. 2000. *Explaining language change*. Harlow, England: Longman Linguistic Library.

Davies, Mark. 2008. *The corpus of Contemporary American English: 450 million words, 1990–present*. http://corpus.byu.edu/coca/.

Durrant, Philip & Alice Doherty. 2010. Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory* 6(2). 125–155.

Ellis, Nick C. 1996. Sequencing in SLA: Phonological memory, chunking and points of order. *Studies in Second Language Acquisition* 18. 91–126.

Ellis, Nick C. 2002. *Reflections on frequency effects in language processing. Studies in Second Language Acquisition* 24(2). 297–339.

Erman, Britt & Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text* 20. 29–62.

Gardner, Michael K., Ernst Z. Rothkopf, Richard Lapan & Toby Lafferty. 1987. The word frequency effect in lexical decision: Finding a frequency-based component. *Memory & Cognition* 15(1). 24–28. DOI: 10.3758/BF03197709.

Godfrey, John J. & Edward Holliman. 1993. *Switchboard-1 release 2 LDC97s62*. Philadelphia: Linguistic Data Consortium.

Gregory, Michelle L., William D. Raymond, Allan Bell, Eric Fosler-Lussier & Dan Jurafsky. 1999. The effects of collocational strength and contextual predictability in lexical production. *Chicago Linguistic Society* 35. 151–166.

Haiman, John. 1994. Ritualization and the development of language. In William Pagliuca (ed.), *Perspectives on grammaticalization*, 3–28. Amsterdam: John Benjamins.

Hirst, Daniel J. 1983. Structures and categories in prosodic representation. In Anne Cutler & D. Robert Ladd (eds.), *Prosody: Models and measurement*, 93–110. Berlin: Springer Verlag.

Hoffman, Sebastian. 2004. Are low-frequency complex prepositions grammaticalized? On the limits of corpus data - and the importance of intuition. In Hans Lindquist & Christian Mair (eds.), *Corpus approaches to grammaticalization in English*, 171–210. Amsterdam: John Benjamins.

Jones, Susan & John Sinclair. 1974. English lexical collocations: A study in computational linguistics. *Cahiers de Lexicologie* 24. 15–61.

Jurafsky, Dan, Alan Bell, Michelle Gregory & William D. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In Joan Bybee & Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 229–254. Amsterdam: John Benjamins.

Langacker, Ronald W. 2008. *Cognitive grammar: A basic introduction*. Oxford: Oxford University Press.

Lindblom, Björn. 1990. Explaining phonetic variation: A sketch of the H&H theory. In William J. Hardcastle & Alain Marchal (eds.), *Speech production and speech modelling*, 403–439. Dordrecht: Kluwer Academic Publishers.

Matsuki, Kazunaga, Victor Kuperman & Julie A. van Dyke. 2016. The random forests statistical technique: An examination of its value for the study of reading. *Scientific Studies of Reading* 20(1). 20–33.

Morrill, Tuuli. 2011. Acoustic correlates of stress in English adjective-noun compounds. *Language and Speech* 55(2). 167–201.

Pawley, Andrew. 1986. Lexicalization. In Deborah Tannen & James E. Alatis (eds.), *Languages and linguistics: The interdependence of theory, data and application*, 98–120. Washington, D.C.: Georgetown University Press.

Pawley, Andrew & Frances H. Syder. 1983. Two puzzles for linguistic theory: Native-like selection and native-like fluency. In Jack C. Richards & R.W. Schmidt (eds.), *Language and communication*, 191–226. London/New York: Routledge.

Pitt, Mark A., Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth Hume & Eric Fosler-Lussier. 2007. *Buckeye corpus of conversational speech (2nd release) [www. Buckeyecorpus. Osu. Edu]*. Columbus: Ohio State University.

Scarborough, Don L., Charles Cortese & Hollis S. Scarborough. 1977. Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance* 3(1). 1–17. DOI: 10.1037/0096-1523.3.1.1.

Schmid, Hans-Jörg. 2017. A framework for understanding linguistic entrenchment and its psychological foundations. In Hans-Jörg Schmid (ed.), *Entrenchment and the psychology of language*, 9–35. (Language & the human lifespan). Berlin: Walter de Gruyter.

Seyfarth, Scott. 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition* 133. 140–155.

Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Siyanova-Chanturia, Anna & Ron Martinez. 2015. *The idiom principle revisited. Applied Linguistics* 36(5). 549–569.

Sóskuthy, Márton & Jennifer Hay. 2017. Changing word usage predicts changing word durations in New Zealand English. *Cognition* 166. 298–313.

Tagliamonte, Sali A. & Harold R. Baayen. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178.

Wilson, Damián V. 2009. From "remaining" to "becoming" in Spanish: The role of prefabs in the development of the construction quedar(se) + ADJECTIVE. In Roberta Corrigan, Edith Moravcsik, Hamid Ouali & Kathleen Wheatley (eds.), *Formulaic language*, vol. 1 (Typological Studies in Language 82), 273–295. Amsterdam: John Benjamins.

Wolter, Brent & Junko Yamashita. 2018. Word frequency, collocational frequency, L1 congruency, and proficiency in L2 collocational processing: what accounts for L2 performance? *Studies in Second Language Acquisition* 40(2). 395–416.
Wray, A. 2002. *Formulaic language and the lexicon.* Cambridge: CUP.

Chapter 5

Formulaic sequences with ideational functions in L1 student and expert academic writing in English

Ying Wang
Karlstad University, Sweden

Corpus studies have revealed that formulaic sequences are prevalent in academic discourse in English. The predominant trend in this research area is to take a frequency-based approach (e.g., lexical bundles, *n*-grams), relying on the computer to retrieve continuous word sequences that occur frequently in a given corpus. Such an approach has helped bring to light a rich repertoire of FSs with textual or interpersonal functions (e.g., *on the other hand*, *it is possible to*) that characterises successful academic writing. However, the use of formulaic language that is central to the construction of disciplinary knowledge has received relatively little attention partly due to the limitations of the identification method. Through manual identification and annotation of FSs in context, the present study examines successful L1 student and expert writing. The results reveal that both are highly formulaic in quantitative terms, and ideational FSs account for approximately 70% of all FSs identified. However, each has its own distinct features in terms of the variety of FSs used. In general, the student corpus employs more everyday FSs which are often highly idiomatic, whereas the expert counterpart yields more FSs associated with research and reasoning processes. It is also argued that knowledge of conventional usage patterns for what seem to semantically transparent and syntactically flexible FSs in academic discourse is not necessarily an inherent part of native speakers' linguistic competence, but needs to be acquired incrementally through formal instruction and training by non-native and native students alike.

1 Introduction

Formulaic sequence (FS) is defined by Wray (2002: 9) as "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar." FS is used in the literature as an umbrella term to mean anything from idioms, phrases, collocations, to clusters or multi-word units/expressions. Generally speaking, what makes a word sequence appear to be prefabricated can be either its high frequency of occurrence in a given situation, or the internal fixedness of the form, or sometimes both (**Siyanova-Chanturia2013**). Depending on the type of FS under investigation, different methodologies are used in previous studies to identify target sequences. The predominant trend in formulaic language research so far is to take a frequency-based approach (e.g., lexical bundle, *n-gram*), relying on computational tools to automatically identify frequently occurring word sequences in large text corpora. While this approach has the advantage of being methodologically straightforward and efficient, its inherent limitations have also been increasingly recognised (see Ädel & **Erman2012**; **Wang2018**). Among other things, some highly salient FSs tied up with a particular communicative context are difficult capture due to their relatively low frequency of occurrence and/or internal variability. More importantly, from a pedagogical perspective, such an approach often results in a large number of incomplete structural or semantic units (e.g., *although it is, that can be*) that are of limited use to language learners and novice writers, for whom the key information about FSs is rarely which sequences are the most frequent per se, but what functions they fulfil and what forms they tend to employ as well as the degree of variation allowed in a given context (Durrant & Mathews-Aydınlı 2011). In short, as **Biber2009** suggests, there is still a need to embrace new and complementary methodological approaches. The present study is a step forward in that direction by incorporating a primarily manual approach in identifying word sequences, continuous or discontinuous, in an attempt to provide empirical evidence on what may have been missed in frequency-based studies and what those overlooked FSs can tell us about formulaicity in language use.

Over the past decade, corpus studies, often utilizing a comparative approach, have revealed that FSs are prevalent in academic discourse[1] and offer an important means of differentiating disciplinary practices and groups of writers – the

[1]In this paper, terms such as 'academic discourse' and 'academic writing' are used to mean 'academic discourse/writing in English', and the claims made about them may not apply to other languages.

appropriate choice of a FS among a range of alternative expressions marks the writer as a member of the discourse community (e.g., **BiberEtAl2004**; **Cortes2004**; **Hyland2008a**; **20122008a**; **Durrant2017**). To date, differences between non-native (L2) and native (L1) or expert production have received most attention, often with the aim of outlining the difficulties experienced by L2 writers (either students or novice academics) (e.g., **Hyland2008b**; **ChenBaker2010**; Ädel & **Erman2012**). L1 novice writers, as **Hyland2016** points out, have been largely marginalized in studies of academic writing. Indeed, in studies focusing on FSs, L1 student writing, if involved, often serves as the benchmark against which non-native data are evaluated, with the assumption that the use of FSs is part of native speakers' inheritance (**WrayPerkins2000**). While this is true for everyday language use, it has been increasingly realised that academic English is no one's first language and formulaicity in academic writing may not be an inherent skill but require prolonged formal education and training (**FergusonEtAl2011**; **Pérez-Llantada2014**). The present study addresses this somewhat neglected line of research by putting L1 students under the spotlight. Through comparing successful L1 student disciplinary writing with published expert writing, the study aims to shed some light on the development of formulaicity specific to academic discourse among native speakers.

The frequency-based approach has helped uncover a rich repertoire of lexical-grammatical resources available for writers to organise their texts (e.g., *on the other hand, in addition*), take a stance towards its content (e.g., *it is possible to*), and to engage with the readers (e.g., *note that*). While such FSs with textual or interpersonal functions have received considerable coverage in previous studies, those that are associated with the propositional content typical of a given discipline, including core disciplinary concepts (e.g., *positive rights, position vectors*), methodologies and research procedure (e.g., *scale up to, at low/high stresses*), norms for reasoning (e.g., *rule out, a plausible explanation for*), have been largely neglected. In the few studies which do involve what they call 'research-oriented' expressions, only a handful of roughly defined sub-categories have emerged, e.g., location (e.g., *at the beginning of*), quantification (e.g., *a wide range of*), attribute (e.g., *the structure of the*), and procedure (e.g., *the use of the*) (**Cortes2004**; **BiberEtAl2004**; **Hyland2008**). This imbalance in coverage may be partly due to the limitations of the identification approach. Textual and interpersonal FSs tend to be longer word combinations – textual FSs in particular are likely to be invariable word sequences (**Wang2019**), which means they are more easily captured by automatic retrieval methods than FSs with ideational meanings which often involve two or more core lexical items with a great deal of formal variability.

Using a partly manual approach in the identification of FSs and a more comprehensive classification framework derived from Systemic Functional Linguistics (SFL) (cf. §??), the current study is part of an on-going project that sets out to investigate the use of FSs that distinguishes successful L1 student and expert academic writing, while at the same time exploring the potential and feasibility of the proposed methodology. The results of textual and interpersonal FSs can be found in **Wang2018** and **Wang2019**, respectively. This paper focuses on ideational FSs, and by comparing the results with those of textual and interpersonal FSs, it will also provide an overall picture of the distribution of the three categories of FSs in L1 student and expert academic writing.

2 Data and procedure

2.1 Data

The present study used the same data as used in Wang (2018, 2019), involving two small corpora of approximately 100,000 words, representing successful L1 student and expert writing, respectively (see Table 7.1).

Table 5.1: Data used in the study

	No. of texts	**No. of words**
Student corpus	15	46,722
Expert corpus	11	52,626
Total	26	99,348

The student texts were randomly drawn from the BAWE corpus (**NesiGardner2012**) as long as they fulfil the following criteria: belonging to the same genre ('essay'), written by L1-English students in their final year of undergraduate studies, awarded the 'distinction' grade. The texts are also evenly distributed across a number of disciplines so that they should provide a broadly representative sample of successful L1 student writing at the chosen level.

It is extremely difficult, if not impossible, to find a control corpus containing texts that are exactly equivalent to student writing (**Callies2015**). In the present study, the keywords that occur in the titles of the student texts were used to search for published research articles in order to minimise the effect of topic on lexical features (**CainesButtery2017**). In addition, all the articles were drawn from SCI indexed journals to ensure the quality of writing is reasonably high. In

terms of genre, while the published articles may be considered as representing a homogenous text type to a great extent, the 'essay' genre in the BAWE corpus is by definition quite broad, where the students "are expected to develop ideas, make connections between arguments and evidence, and develop an individualized thesis" (**NesiGardner2012**, p.38). An examination of the selected student essays revealed that indeed there can be variations across and within disciplines, but most of the essays seem to bear a great deal of resemblance to the expert counterparts in terms of the structure of the text and the type of arguments and evidence involved (e.g., empirical or theoretical). That said, student assignments are by nature different from published articles with regard to communicative purposes; therefore the comparison between the two must be treated with caution.

2.2 Identifying formulaic expressions

The present study aims to be as inclusive as possible in the identification of FSs. Therefore mixed criteria were adopted, given the rationale that "most examples will be captured one way or another" (**Wray2008**). If a multi-word sequence satisfies one of the following criteria, it was regarded as formulaic. [2]

(1) **Grammatical irregularity and/or semantic opacity (Wray2008; SchneiderEtAl2014; Herbst2015)**

This means that as long as some aspect of the form or meaning of a word sequence is not strictly predictable from its component parts or from regular grammar, the expression is a FS, e.g., *take place*, *account for*, *run through*. Note that there is a continuum of fixedness, ranging from those resulted from a grammaticalisation or lexicalisation process (e.g., *as opposed to*, *with respect to*) to those that allow a certain degree of compositional freedom and semantic transparency (e.g., ***in** a similar way*, ***in** this **way***, *the **way** **in** which*). In the present study, dictionaries (primarily the *Oxford Learners' Dictionaries*)[3] and the list of phrasal expressions provided by **MartinezSchmitt2012** were regularly consulted to avoid subjective judgement. If a word sequence is highlighted in the dictionaries (either as a separate entry or emphasised in bold type) or occurs on the list, it was considered to contain some kind of irregularity and therefore a FS.

[2] Some sequences may satisfy more than one of the criteria.

[3] This is an online source (https://www.oxfordlearnersdictionaries.com), which is home to the following dictionary and grammar reference titles:*Oxford Advanced Learner's Dictionary (9th edition)*,*Oxford Advance American Dictionary*,*Practical English Usage*,*Oxford Learner's Dictionary of Academic English*, and*Oxford Collocations Dictionary*.

(2) **Underlying frame (Wray2008)**

This refers to a formulaic frame that involves open slots to be filled, often by items of similar characteristics, e.g., *in the YEARs*, *in the Nth century*, *from YEAR to* YEAR,

(3) **Situation/register/genre-specific formula (Wray2008, Buerki2016)**

Expressions of this type are considered formulaic not because of their internal semantics or syntax, but rather the fact that they are the normal ways (judged by frequency of occurrence) of saying things in a particular situation. In the case of ideational expressions in academic discourse, some examples are *the nature of, the structure of, research methods, public opinion polls*. To identify such FSs, the present study relied on an online tool, IdiomSearch (**Colson2016a**; see also **Colson2016b**). This program uses a built-in list of frequently occurring multiword phrases (ranging from bigrams to sevengrams), derived from a multimillion-word reference database, to identify FSs in any given stretch of text. It has an advantage over the more commonly used tools such as AntConc particularly when dealing with small corpora where some FSs simply cannot reach the frequency threshold to be extracted. Clearly, one limitation of IdiomSearch is the difficulty in identifying FSs that are highly specific to a particular social practice or academic discipline (e.g., *Kant's critical philosophy, fluent aphasia*). However, such FSs are normally salient enough to be spotted manually and can be easily checked using either AntConc (whether they occur repeatedly in the given corpus) or Google Scholar (whether the same terms are used by other scholars).

The sequences identified by IdiomSearch were then manually sifted through to remove structural fragments without a clear meaning or function, such as *to be the*, *will give*, *is not a*, *we have a*. In some cases, an automatically identified sequence may contain more elements than needed for a complete semantic unit (e.g. *involves* in *involves the development of*) or only part of a semantic unit (*sequence of* in *an exact sequence of*, *a better way* in *in a better way*) (see **MartinezSchmitt2012** and **Buerki2016** for the idea of semantic units). Human intervention means that the FSs identified will be self-contained semantic units (e.g., *the development of*, *an exact sequence of*, *in a better way*) that can be of utility for language teaching and learning purposes. There are also some cases that were not identified by IdiomSearch but were nevertheless included in the analysis because they contain the same core elements as in those that have been identified by the program, albeit with some formal variations. Take the combination of *ask* and *question* for instance; while *asked questions about* was identified by the program, those involving changes of word order or form, or intervening elements as

in *the questions asked*, *asked 10 blocks of questions*, *asking questions*, *asking knowledge questions about*, *the question being asked* were all missed by the computer as the exact sequences may not be frequent enough in the reference database. However, the exclusion of such variations would risk overlooking potentially important features of a given discourse community, and a manual approach was applied exactly to identify those non-contiguous FSs.

2.3 Classification of ideational functions

The classification of functions in the current study was based on Systemic Functional Linguistics (SFL), developed by Halliday (see **Halliday2014**). SFL focuses on the underlying communicative functions of language and the systemic choices that are made available by the language system (**Gledhill2011**). Central to the theory is the notion of three kinds of metafunctions – ideational, interpersonal, and textual – which underlie the organisation of language. In previous studies of lexical bundles such as **Hyland2008** and **BiberEtAl2004**, the functional framework used was all based loosely on SFL. As discussed in the introduction, while textual and interpersonal functions have been extensively investigated in previous research, the ideational – also called as 'research-oriented' – functions are less well defined, often containing only a fairly small number of options. For a more comprehensive study of FSs with ideational functions, the present study turned to the original SFL framework for the purpose of deriving a workable annotation taxonomy.

The ideational metafunction in SFL is concerned with the construction of knowledge or human experience, represented as a configuration of a process (a type of action or event), participants in that process (an actor or object), and circumstantial elements such as time, place and manner. Each of these three components gives entry to a more specific system with a variety of options. Table 7.1 presents a slightly simplified version of the original system of ideational functions (see **Halliday2014**), excluding those that either are not normally associated with FSs or rarely occur in the type of discourse under investigation, such as the category of behavioural processes. Some of the functions as well as their explanations have been tailored to the discourse in hand and its features. For instance, verbal FSs in the present study are often related to reference to previous research (i.e., what other scholars say about something), definition, explanation, and argumentation. The circumstantial elements in the original framework were merged into a few main sub-categories. Among them, manner encompasses a number of elements, such as angle and role, which are treated as separate sub-categories parallel to manner in the original taxonomy. The remaining sub-categories (e.g.,

matter, accompaniment) were put under the 'other' category due to their low frequencies. Terminology was added to the framework to address the large number of specialist terms occurring in the data under investigation.

The corpus data were manually gone through to identify FSs based on the criteria presented earlier. The UAM Corpus Tool (O'**Donnell2013**) was used for the annotation of functions according to the functional taxonomy presented above.

3 Results and discussion

This section presents the overall frequencies of ideational FSs in the two corpora before offering a more detailed analysis of a number of major sub-categories of FSs found in the two corpora.

3.1 An overall picture

Altogether, 9,558 FSs with ideational functions were identified in the two corpora. Table 7.3 presents both raw and normalised frequencies (per 10,000 words) of ideational FSs in each corpus. To give an overview of the distribution of FSs associated with all the three metafunctions, the results from **Wang2018** and (**??**) regarding interpersonal and textual FSs are also presented in Table 7.3; see also Figure 8.1 for a graphical representation of the distribution. The log-likelihood test was conducted throughout the study to calculate whether a difference between two raw frequency counts is due to chance or to a statistically significant difference between the two corpora.

As mentioned in the introduction, it is the textual and interpersonal metafunctions that have attracted most attention in previous studies of lexical bundles, *n*-grams and other types of FSs. However, as shown in Figure 8.1, the two categories of FSs together only account for approximately 30% of all the FSs identified in each corpus, whereas ideational FSs make up for the remaining 70%, which, for some reason, have not been investigated systematically. Out of the 9,558 ideational FSs retrieved from the corpora, only 3,103 (32%) were captured by the frequency-based program, while 3,618 (38%) were completely missed; the remaining 2,829 (30%) instances were partially identified by the program in the sense that some of them may be part of a complete formulaic unit and some may contain elements outside a complete unit (cf. §**??**).

In terms of overall frequencies, Table 7.3 reveals a great similarity between the two corpora regarding ideational FSs; in fact, a statistically significant difference was only found in textual FSs between the two corpora. In other words,

Table 5.2: Sub-categories of the Ideational Metafunction

Process	
Material	Doing: action, movement, research procedure, e.g., *tidy X up to, turn away from, the operation of, search for, an examination of*
Mental	Perception, cognition, emotion, reasoning process, e.g., *make sense of, the understanding of, be expected to, take into account*
Verbal	Saying (normally associated with the reporting of previous research), explaining, defining, argumentation, e.g., *put forward, assert/proclaim that, as X put it, argue against, an explanation of*
Relational	Attributing, identifying, e.g., *consist of, be linked to, interaction with*
Existential	Existing, happening, e.g., *there be, there remain, the emergence of, take place*
Circumstance	
Location	Place, time, e.g., *in the world, in the Nth century, at the end of*
Manner	Means, comparison, degree, extent, angle, role e.g., *as a means of, quickly and easily, as opposed to, to the extent that, from the perspective of, in the form of*
Cause and contingency	Reason, purpose, condition, concession, e.g., *because of, as a result of, for the purpose of, in case of, in the absence of, in spite of*
Other	Matter, e.g., *with respect to*; accompaniment, e.g., *instead of, as well as*
Participant	
Attribute	Descriptive property, e.g., *the nature of, the character of*
Quantification	Quantity and category specification, e.g., *a small number of, a lot of, the majority of, a piece of, a type of*
Human or non-human entity	Normally non-specified, e.g., *human beings, ethnic minorities, a large audi-*

Table 5.3: Raw and normalised frequencies of FSs associated with the three metafunctions Distribution of the three types of metafunction in each corpus

	Student corpus			Expert corpus		
	Ideational	Textual	Interpersonal	Ideational	Textual	Interpersonal
No. of FSs	4484	631	1189	5074	890	1243
per 10,000 words	960	135	254	964	169	236

Loglikelihood test results:
Ideational: G2 = 0.05, $p > 0.05$
Textual: G2 = 18.88, $p < 0.0001$
Interpersonal: G2 = 3.38, $p > 0.05$

both student and expert texts are highly formulaic. Previous studies such as **ChenBaker2010** and Ädel & **Erman2012** have observed a lack of formulaicity in L2 undergraduate students' academic writing, in comparison to either L1 student writing of the same academic level or expert writing. The results presented above suggest that successful L1 student writing is fairly close to expert writing in terms of formulaicity, at least quantitatively. This in turn may lend support to the advantage that native speakers have over non-native students in the use of ready-made multi-word expressions, which are considered part of native speakers' linguistic competence that non-native speakers have limited access to (**20082002**; Wray 2002; **Kecskes2016**).

The following sub-section looks more closely at the ideational FSs based on the distribution of the three main categories of the ideational metafunction as well as the sub-categories as presented in Table 7.2.

3.2 FSs of different ideational functions

Table 7.4 presents the frequency counts and proportions of FSs with different ideational functions in each corpus.

The two corpora resemble each other again in the distribution of the three broad functional categories. The identified FSs are most likely to be involved in processes (43%), followed by circumstances (31–33%) and participants (24–26%). However, within each category, significant differences between the two corpora

Table 5.4: Distribution of FSs with different ideational functions in each corpus

		Student corpus		Expert corpus		
Categories	Sub-categories	No.	%	No.	%	Log-likelihood (G2)*
Process	Material	589	30	747	34	4.65, p<0.05
	Mental	443	23	466	18	1.06
	Verbal	392	20	491	22	2.47
	Relational	330	17	350	16	0.61
	Existential	187	10	211	10	0.00
	TOTAL	**1941**	**100**	**2201**	**100**	**0.05**
Circumstance	Location	451	31	497	31	0.11
	Manner	670	46	674	43	4.29, p<0.05
	Cause and contingency	270	18	268	17	2.15
	Other	79	5	139	9	10.36 p<0.01
	TOTAL	**1470**	**100**	**1578**	**100**	**1.76**
Participant	Quantification	287	27	267	21	5.06 p<0.05
	Attribute	479	45	535	41	0.02
	Terminology	268	25	432	33	21.74 p<0.0001
	Human and non-human entity	39	4	61	5	2.62
	TOTAL	**1073**	**101**	**1295**	**101**	**2.80**

*Only significant p values are shown in the table.

were found in some sub-categories: material processes, manner and 'other' circumstantial elements, as well as quantification and terminology. In what follows, some of these sub-categories will be examined further with examples drawn from the dataset.

Starting with material processes, as shown in Table 7.4, the expert writers used significantly more FSs than did the students. Table 5.5 gives some examples of such FSs, divided according to their structural make-up.

A few observations can be made from Table 5.5. To begin with, the FSs associated with material processes are made up of three main structural types: verb + preposition, verb + noun, and nominalisation + of. With regard to the first type, there are clearly more verb + preposition combinations, or phrasal/prepositional verbs, in the student corpus than in the expert counterpart. As can be seen in Table 5.5, some of the phrasal/prepositional verbs are shared by both corpora, e.g., *deal with*, *carry out*, *find out*, which are often used in academic writing to introduce a research topic, procedure, or a finding. However, the majority of the phrasal/prepositional verbs occur exclusively in the student corpus. Many of them seem to involve some kind of bodily movement and/or a figurative sense (e.g., *run away*, *storm out*, *fiddle with*, *trawl through*). As illustrated in the following examples, the use of such multi-word expressions is often associated with a narrative approach taken by the students in their essays.

- *The camera also* ***zooms*** **out** *to offer a wide shot of the four women, this serves to show how Miranda is surrounded and cornered by the others.* (BAWE_-3160b)
- *Having* ***trawled*** **through** *the archives the historian's next task according to him was to corroborate and compose a critique of the evidence at hand.* (BAWE_0255h)
- *However the difficulties with complex structures could be related to the suggestion that Broca's, and other non-fluent, aphasics* ***struggle*** **with** *comprehension of unfamiliar, less frequent and longer word retrieval,…* (BAWE_6206c)

Multi-word lexical verbs are more commonly seen in conversation and fiction than in academic prose (**BiberEtAl1999**: 409); the frequent occurrence of such verbs in the student corpus may thus also be taken as suggesting an informal style, which has been attested as a feature of student writing in general, regardless of L1 background (**GrangerRayson1998**; **GilquinPaquot2008**).

When it comes to verb + noun collocations, there seems to be a great deal of similarity between the two corpora. Some of them (e.g., *wage wars*, *commit*

Table 5.5: Examples of FSs representing material processes

	Student corpus	Expert corpus
Verb + Preposition	*expand on, look for, find out, deal with, trawl through, strip down, storm out, run away, go after, dress up, waltz into, fiddle with, move away, engage in, carry out, cover it with, break into, work on, sweep out, cut down, interfere with, trick sb into, force upon, suffer from, prevent/protect sb from, benefit from*	*delve into, build on, deal with, find out, engage explicitly with, was (calmly) engaged in, search for, work with, bring about, carry out, set back, interfere with, prevent sb from*
Verb + Noun	*make a detailed analysis, research conducted into, take a quick look at, tackle the problem, commit crimes against, commit an (earlier) error, wage wars, launch a media campaign, make some changes to, make more sales, make profit*	*the original research undertaken, overcome barriers, gain momentum, the murders/crimes/errors committed, wage wars, data was collected, take parental leave, take care of, meet their/the buyers' needs, impose limitations on, restrictions imposed by, further restrictions are imposed on*
Nominalisation + of	*an examination of, (the/a) study of, the development of, an/the analysis of, a wide shot of, the/an engagement of*	*scientific assessment of, the comprehensive collection and analysis of, the return of, a/the (thorough) development of, his engagement with*
Adjective + Noun	*scientific research, sexual abuse, marketing efforts, human endeavour*	*empirical study, further investigation/research, recent developments, scientific discoveries/advances, genetic modification, (in) previous research*
Other	*bought and sold, distri-*	*widely used, newly gen-*

crimes) occur in both corpora, prompted by the same topic or subject area. Other topic-related collocations were also found, such as *take parental leave*, *meet someone's needs*, *impose restrictions on* in the expert corpus and *launch a media campaign*, *make more sales*, *make profit* in the student corpus. What remains are research-related collocations (e.g., *conduct + research*, *collect + data*, *make + analysis*), which, again, can be found in both corpora. An additional point to be made here is that verb + noun collocations often show a great deal of formal variability in terms of word order and intervening elements (e.g., *impose limitations on*, *further restrictions are imposed on*). Such formal variations mean that the core lexical items are not always contiguous and therefore are likely to be missed by automatic retrieval methods; in other words, for both methodological and theoretical reasons, this is an area that is worth further exploration using large corpora.

Nominalisations are a well-established feature of academic writing, used to pack more information into a single sentence. In the present study, the frame nominalisation + *of*, with or without an article *a/an* or *the* before the combination, is fairly common in both corpora. However, as shown in Table 5.5, nominalization + *of* constructions in the expert corpus often also contain adjectives (e.g., *scientific assessment of*). While the *of*-frame represents a grammatical construction, which is considered formulaic on the grounds of its high frequency, there is a strong collocational tie between the two core lexical items involved. A similarly strong collocational link is also apparent in most FSs of the next two categories drawn from the expert corpus (e.g., *empirical study*, *widely used, fully developed*), many of which are associated with research processes. The student corpus, in sharp contrast, is still dominated by processes related to subject areas (e.g., *sexual abuse, bought and sold*).

Moving on to FSs associated with mental processes, although the two corpora display no statistically significant difference in terms of frequency, a close examination of the FSs themselves provided some interesting insights. As can be seen in Table 5.6, which contains examples identified from both corpora, most of this group of FSs involve two or three key components, which, again, are not always contiguous. Apart from verb + noun and adverb + verb collocations, most of the FSs involve a combination between a noun/adjective/verb and a preposition. Semantically, a great number of the FSs in both corpora are associated with awareness, understanding, decision-making, and opinion. However, the expert corpus yielded more FSs representing a reasoning process (e.g., *derive from*, *draw conclusion*, *make observation*, *the verification of*).

In contrast, the students seemed more inclined to emploit another type of FSs, associated with an emotional state, as illustrated in the following examples.

Table 5.6: Examples of FSs representing mental processes

	Student corpus	Expert corpus
Nominalisation + Preposition	*the same commitment to, one's (a full) understanding of* (F), *one's conception of, the feeling of, sb's thoughts about, an awareness of, his view(s) on/about/towards, intentions towards*	*an/the (full) understanding of, awareness of, confidence in, the views of, greater attention to, the thought of, sb's thoughts on, the perception of, the comprehension of, the verification of, be seen from*
Be + Adjective + Preposition/*that/to*-infinitive	*be (un)aware of, be reluctant to, be inclined to, be anxious to, be interested in, be very conscious about, be mindful that, be highly appreciated by, be wary of, be expected from*	*be (more/not) aware of/that, be opposed to, be concerned with, be prepared to, indifferent to*
Verb + Preposition	*conceive of, take into account, extend to, come up with, rule out, from this emotion we derive*	*derive from, know about, wonder about, hope for, take into consideration/account*
Verb + Noun	*have sympathy for, make judgements, make sense (of), have no sense of, get some sense of, get an/the rough idea of, take the decision to, decisions were taken, the decisions taken, make informed decisions for, make strategic decisions, decisions regarding which market segments to target can be made, generalisations made, bring to light, give a proof (of)*	*have little (to no) knowledge about/of, the decision should be made, make sense of, take stock of, the choices parents make, the major conclusions that can be drawn from, come to these conclusions, make that/this/more final observation(s)*
Other		*considered carefully, well understood, easily overlooked, better under-*

- *The Führer* ***was*** **only satisfied with** *forming a Protectorate rather than outright annexation when Hàcha unexpectedly co-operated.* (BAWE_0318e)
- *She* ***is*** **anxious to** *hear Nicholas say she looks beautiful and forces him to say so, this infantile behaviour matches her personality and role as a Gothic heroine.* (BAWE_3160b)
- *Exporters need to* ***be*** **wary of** *using the same promotional strategy in the UK as in their home country.* (BAWE_0222a)

This tendency seems to mirror the students' use of multi-word lexical verbs associated with material processes as discussed earlier, evincing characteristics of a narrative approach and everyday language in the student essays.

As in the case of FSs associated with mental processes, quantitatively, there is no statistical difference between the two corpora with regard to verbal FSs. Yet a few comments need to be made, nevertheless, about the particular FSs involved. Table 5.7 gives a list of examples from the dataset. What the two corpora have in common is the use of FSs to offer an explanation or to raise or answer a question, particularly in the expert corpus, with a range of lexical and syntactic variations (e.g., *answer the question, answer 10 blocks of questions, an answer to the question, the questions asked, ask objective-knowledge questions, ask a follow-up question, ask him a question*). In addition, topic-related FSs can be found in both corpora (e.g., *give + consent*).

The main difference between student and expert writing in this regard can be seen in the number of FSs associated with arguments and debates as well as elaboration in the expert corpus (e.g., *the justification for, an objection against, elaborate on*) versus that of those expressing actual verbal behaviour in the student corpus (e.g., *cheer someone up, laugh at, raise one's voice*). As Example (**??**) shows, the latter, most of which are highly idiomatic (e.g., *take/hold the floor*), seem to be prompted by, again, a need to narrate what is being analysed – a conversation in this case.

- *At line 11, B* ***makes*** **a closing kind of statement**. *It is not very meaningful to the discussion and B is therefore indicating that she has nothing further to add. Speaker A and C both respond with a backchannel, and even though C's is quite long, (line 13), neither* ***take*** **the floor**. (BAWE_6009b)

Table 5.8 provides some examples of FSs representing the most common circumstantial sub-category, namely manner. Most of such FSs are prepositional phrases. As can be seen in Table 5.8, the expert corpus yielded a more limited

Table 5.7: Examples of FSs associated with verbal processes

Student corpus	Expert corpus
a direct answer to the question, ask the question of,	*answer a list of questions, the questions asked,*
a more plausible explanation for, be explained by, a plausible explanation, an explanation for, have an explanation of,	*(the) argument(s) for, argue directly against, the justification for, the postulation of, claims that he set out, objections to, three objections that have been raised against, a final objection against,*
the narration of, a discussion about, be said bout, an excellent/objective account of,	*explanations of, some/an explanation of, have no plausible naturalistic explanation, any deeper explanation of, account for,*
be called upon to, be accused of, speech made, cheer sb up, talk about/to, laugh at, raise one's voice, hold/take the floor, make a closing kind of statement, consent to, give informed/full consent	*give a plausible account of, elaborate on, a summary of, a description of, a brief overview to, go into the fine details of, research reports,*
	give their informed consent to, commonly called, point out, enquire about, talk about, the repetition of, the utterances of, verbal communication, science communication

range of FSs, mostly in association with the way (manner, fashion, means) in which a process takes place, than did the student corpus. Some of the FSs occurring exclusively in the student corpus, again, involve emotional states such as *in admiration, with tolerance, in anger, without any major headaches.*

We have thus far witnessed a tendency, which is distinct of the student corpus, to involve FSs related to emotion as well as verbal and bodily behaviour, regardless of discipline. Together, they may suggest that we are dealing with two different genres here: narrative versus argumentative. However, given that the student essays are academic assignments given to last-year university students in the UK and that they are structured in a similar way to that of the published papers, it may be fair to say that the students at this stage are expected to produce work of a similar genre, albeit limited in scope and depth in comparison to

Table 5.8: Examples of FSs associated with manner

Student corpus	Expert corpus
in such strict dichotomy, with tolerance, in isolation, in Nazi rhetoric, in such a way that, by chance, in the same fashion, in a straightforward manner, the detail in which, in detail, quickly and easily, in anger, in admiration, at rest, in the form of, without any major headaches, positively or negatively, with difficulty, long/short term, in equilibrium, on the macro scale, at this fundamental level, at resonant specific frequencies, in 26 space-time dimensions, at a speed, at 100%, to a minimum	*in an existential manner, in an easy-to-read and understandable manner, in this manner, in a somewhat Hobbesian fashion, the ways in which, by way of, in this strange way, a political means through which, at the global level, at the macro level, in the conventional form, in detail, under the guidance of, in abstract/ADJ terms*

published ones. Or, to put it in another way, they can be regarded as novice writers in training. Indeed, bearing in mind that the two corpora also share a great number of FSs, it is unlikely that they represent two completely different genres of writing. Rather, a more reasonable explanation for the differences observed between the two corpora may be put down to the students' lack of awareness of genre conventions in terms of the way disciplinary knowledge is constructed and the style of delivery.

The students' lack of awareness of genre conventions can also be detected elsewhere. Take, for instance, FSs containing the word *way*. Altogether, 25 tokens with 16 different types were found in the expert corpus, and 42 tokens with 36 different types in the student corpus. Some examples are given in Table 5.9.

A few points can be made here. First of all, again, there is a great deal of variability in form, with fixed and variable slots occurring in a particular order. Three main patterns emerged from the examples about the use of FSs containing the word *way* in expressing means. All of them involve pairings of functions words (prepositions *in* and o*f*) with at least one variable slot: *(a) X way of/to*, *in a X way*, (*the*) *X way in which*. As noted by **Biber2009**, while conversation prefers continuous fixed sequences, the written discourse prefers FSs with internal variable slots. As can be seen here, more often than not, the fixed elements are not adjacent to each other. This is obviously another area where the automatic retrieval methods may be of limited use and which would benefit from a more systematic

Table 5.9: FSs with the key word 'way'

Student corpus	Expert corpus
a quicker way to	*the way in which*
a simple way of	*its/his/the nurse' way of*
an invasive way of	*in a way that*
its own way of	*by way of*
the German way of thinking	*his way of*
the most effective way of	*in this strange way that*
the most suitable way to	*this way*
the only way to	*a different way to do*
the ways of	*one way that*
in a better way	*in this way*
in a mechanical way	*the way in which*
in a purely mathematical way	*in such a way that*
in a rather abstract way	*in the same way*
in a similar way	*in a natural way*
in a simple way	*no way of doing*
in a sustainable way	*in a deterministic way*
in a very physical way	
in a way	
in an unsustainable way	
in complex ways	
in quite a simple way	
in this way	
in the way	
in the way of	
in such a way that	
through its unobstrusive way of	
different ways in which	
one of the ways in which	
the way in which	

investigation involving a larger dataset to generate possibly new understanding of features of formulaicity in language use in general and in academic discourse in particular.

As Table 5.9 shows, the student writers appeared to be less restrained in filling the variable slots than did the expert writers. The same can be said of the student writers' use of FSs associated with quantification. As shown in Table 7.4, the student writers employed this category of FSs more frequently than did the expert writers, the difference between the two corpora being statistically significant. Table 5.10 gives some examples of such FSs, which show a wide range in the student corpus, in contrast to a limited set in the expert counterpart.

Table 5.10: Examples of FSs associated with quantification

Student corpus	Expert corpus
(quite) a few, lot of, a bit more, a pair of, a piece of, a (very small) number of, a great swathe of, one/some/all/none of, a vast amount of, a piece of, a collection of, a wide array of, a series of	*a (limited/small/large) number of, a wide range of, some/many/most/either/one of, a multitude of, the vast majority of*

Some of the expressions, which occur exclusively in the student corpus such as *a bit more* in Example (??), testify again to an informal register that is said to be typical of learner writing as a whole, including both L1 and L2 writing. The use of *harmless* in Example (??) illustrates a trend that has been observed throughout the current study, namely the extent of liberty or 'creativity' that the student writers seemed to assume in filling the internal variable slots of FSs, without realising that some of them may be subject to certain restrictions in a given discourse community.

- *Poincare duality follows after* ***a*** **bit more** *work.* (BAWE_0049b)
- *The patient inhales* ***a*** **small, harmless amount of** *radioactive gas which then attaches itself to red blood cells in the blood...* (BAWE_6206c)

It is generally accepted that successful academic writing is marked by a high degree of formulaicity, but what is perhaps less well recognised is that even those seemingly transparent and syntactically flexible word sequences may have established particular patterns of usage that are adhered to, consciously or not, by the members of the discourse community (**Pérez-Llantada2014**; **Wang2018**). In this

case, even though the use of *harmless* is not semantically or grammatically deviant, in academic prose at least, it is not common to have another intervening adjective together with *small* in the FS *a X amount of*.[4] Although native speakers have available to them a large repertoire of everyday formulaic language (Sinclair 1991), the degree of liberty that the student writers seemed to take here, and in many other cases as shown in the study, suggests that the restrictions such FSs are subject to in academic prose may not be readily accessible to L1 students.

4 Conclusion

The present study set out to explore the potential of a computer-assisted manual approach in identifying and annotating formulaic language in academic writing, with a focus on ideational, or research-oriented, FSs. The first important finding is that ideational FSs account for 70% of all the FSs identified, a considerable proportion that would certainly warrant more serious attention than they have hitherto received. Most of such FSs contain two core lexical items or one lexical and one functional item in fixed slots, with the possibility of variable slots in between and change of word order, making it a particularly challenging task to automatically identify them. However, given their importance in understanding the nature of formulaicity in language use, these are the areas that would certainly benefit from a more vigorous investigation in future research.

Both student and published papers were found to be highly formulaic, particularly in quantitative terms. Indeed, the main differences between the two corpora are of a qualitative nature – that is, the two sets of texts seem to be formulaic in different ways. To start with, FSs associated with research and reasoning processes are conspicuously abundant in the expert corpus, whereas those expressing emotional states as well as verbal and bodily behaviour stand out in the student counterpart, suggesting the students' lack of awareness of genre conventions in terms of knowledge construction and language style.

Throughout the analysis, we also saw that the student writers seemed to be less restrained in filling the variable slots than were the expert writers. The results suggested that academic writing may not be as 'creative' linguistically as the students might have assumed. Rather, many seemingly transparent and syntactically flexible word sequences may have their preferred or conventional patterns of usage in academic discourse, just as members of a particular speech community have preferred ways of saying things (Wray 2002; **Kecskes2016**). It

[4] A search of *small +amount of* in the academic subset of British National Corpus (BNC) returned no instance involving any other adjective in between.

was argued that knowledge of such patterns of usage, which are probably not psychologically salient enough, may not be readily accessible to native speakers, echoing the claim that success in academic writing is "never guaranteed by generics or birth right alone" (Rajagopalan's 2004: 116), but "is acquired rather through lengthy formal education" (**FergusonEtAl2011**: 42) (see also **Hyland2016**).

The SFL framework for the classification of FSs has proved particularly useful in pinpointing areas of difference between student and expert writing. From a pedagogical point of view, these areas of difference would benefit from more targeted awareness-raising activities in the training of novice writers.

To conclude, through capturing and addressing discontinuous and less frequent - but nevertheless formulaic- FSs that have been largely overlooked in previous research, the approach taken in the present study clearly has potential to contribute to both the understanding and the teaching of FSs in disciplinary writing. However, more data are needed in order to draw more informative and definitive conclusions. As manual identification and annotation can only be carried out to a certain extent, to proceed, there is a need to explore the possibility of at least semi-automated methods for recognising and annotating entities in a large text corpus. Given that most of the ideational FSs identified in the present study involve two core node words, it may be promising to start from individual lexical items, either through a keyword analysis (see, for instance, **SolerWang2019**) or with a list of pre-selected node words (see **Römer2019**), to retrieve FSs and their recurrent usage patterns in an effective and consistent way.

References

Bolinger, Dwight. 1965. The atomization of meaning. *Language* 41(4). 555–573.

Carter, Ronald. 2004. *Language and creativity: The art of everyday talk*. Routledge.

Chomsky, Noam. 1972. *Language and mind [enlarged edition]*. New York: Harcourt Brace Jovanovich.

Eggins, Suzanne. 1994. *An introduction to systemic functional linguistics*. London: Pinter.

Forsyth, Richard & Łukasz Grabowski. 2015. Is there a formula for formulaic language? *Poznań Studies in Contemporary Linguistics* 51(4). 511–549. DOI: 10.1515/psicl-2015-0019.

Granger, S. & Y. Bestgen. 2014. The use of collocations by intermediate vs advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching* 52(3). 229–252. DOI: 10.1515/iral-2014-0011.

Magurran, Anne Elizabeth. 2004. *Measuring biodiversity*. Oxford: Blackwell.
Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Wray, A. 2002. *Formulaic language and the lexicon*. Cambridge: CUP.
Wright, David. 2017. Using word *n*-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. *International Journal of Corpus Linguistics* 22(2). 212–241.

Part III

Descriptive research on formulaic language

Chapter 6

Reading discourses through their phraseology: The case of Brexit

Andreas Buerki
Cardiff University

That social, cultural and political events leave their mark on the language of the communities shaped by those events is well-established in the area of the lexicon (cf. 'cultural keywords'). More recently, it has been shown that phraseological phenomena like common turns of phrase or usual expressions of a speech community are also moulded by significant events in the life of that community. This chapter investigates the extent to which social discourses can crystalize and therefore become readable, in phraseology. The example used is that of the 2016 referendum on the United Kingdom's membership of the European Union which was an event that has resulted in a prolonged and on-going period of social, cultural and political change and uncertainty in the UK. The investigation is based on a large corpus of UK media texts and includes the presentation of a methodology for the identification of phraseological expressions in texts, and their comparison across time and topics. Findings in terms of aspects of the Brexit discourse revealed in its phraseology are discussed and comments relevant to phraseological theory are made, including a demonstration of pro-tem phraseology and the speed of phraseological development. Finally, three theses are put forward to progress the field of discourse analytical research: 1) phraseological patterns allow a deep and insightful reading of discourses, because 2) discourses crystallise in phraseology, and 3) phraseological theory explains why this is the case.

1 Introduction

On 23 June 2016, a referendum was held in the United Kingdom (UK) to decide whether the country should remain a member of the European Union (EU).

By a margin of 1.89% of votes cast (The Electoral Commission, 2019), the result favoured the leave option and triggered a prolonged period of social, economic and political change and uncertainty: within the approximately three and a half year period following the referendum covered by this study, two prime ministers resigned, two general elections were held, the currency devalued notably (fullfact.org, 2019a), reported hate crime shot up (fullfact.org, 2019b) and while the country remained part of the EU, whether it would leave and on what terms remained uncertain.

In this chapter, we identify and analyse recurrent phraseological patterns used in the public discussion, within the UK, of the topic of the British exit from the EU (Brexit) in the period between 20 **February2016** (when the referendum was announced) and 31 **October2019** (when the UK missed the second expected exit date). We do this in order to see how these patterns allow us to feel the pulse of what is happening within society, what the issues of public concern are, what is contested or settled – in short, what phraseological patterns allow us to learn about the discourses of Brexit, and by extension other discourses. We also consider why phraseology is particularly suited to this task and what this means for discourse analysis.

In the following, we start out by looking at what exactly is meant by phraseological patterns and why they might be expected to be useful to the task of reading discourses. We also review some of the growing work in other areas of language and Brexit. Subsequently, the data used in this study are introduced, as well as the procedures employed to extract phraseological patterns relevant to Brexit from the source corpora. In the penultimate section of the chapter, the unearthed phraseology of Brexit is presented and grouped into various types of phenomena, before a discussion of the relevance of findings is embarked upon. We end on three concluding theses regarding the reading of discourses through their phraseology.

2 Background

2.1 Phraseology

The essence of the phraseology of a language is understood somewhat differently by different theorists today. However, most would agree that the following examples should be classed as phraseological expressions: *better safe than sorry!* (typically classed a proverb), *pushing up the daisies* (an idiom), *open letter* and *free trade agreement* (multi-word terms), *please hold* and *yours sincerely,* (formulae of spoken and written genres), *sign a contract* and *strong coffee* (collocations where

bases, e.g. *contract/coffee*, attract particular collocates), binomials (*well and truly* and *ins and outs*) and indeed many other usual sequences such as *never heard of it!* and *in recent years*. Although there are phraseological expressions that are entirely fixed (such as *yours sincerely*), many other expressions allow or even require a certain amount of modification. Such modification can take the form of an insertion of elements (*thank you for your [kind/speedy/prompt] reply*), the specification of schematic elements (such as the *X* in *at the end of the Xth century*) or the variability of verbal inflection (e.g. *make/makes/made a mistake*). In addition, many phraseologists now argue that there is 'a graceful transition from idiom-like […] phrases to fully abstract […] constructions' (Dominey, 2006:137), so that the difference between largely fixed phraseological expressions, phraseological patterns of a less lexically specific type (like *the Xer the Yer*) and yet more general patterns like the transitive construction (SUBJ VERB OBJECT) is one of degree of lexical specificity rather than a categorical difference (Buerki, 2016).

This difference in degree means that what is phraseological is occasionally difficult to delimit from what is not. In traditional phraseology, the essence of the phenomenon has typically been identified with the criterion triplet of polylexicality (expressions consisting of more than one word), fixedness patterns (there are restrictions on modifications) and idiomaticity (semantic or structural irregularity) (e.g. Burger, Häcki Buhofer and Sialm, 1982; **BurgerEtAl2007**), though it is acknowledged that the last of these may only apply to phraseology 'in a narrow sense' (Burger, Dobrovol'skij, Kühn and Norrick, 2007: 11). In other thinking, the essence of phraseology has been located in the manner of mental processing, namely that phraseological expressions are or appear to be processed holistically (e.g. Sinclair 1991:110; **Wray2002**:9).

The strand of thinking followed in this investigation, however, sees the locus of phraseology not primarily in items that show semantic irregularity nor at the level of an individual's language processing, but in expressions that have become conventionalised among a speech community to the degree that they have become common turns of phrase (e.g. Bybee, 2010:35; Buerki, 2020: ch. 1). These turns of phrase could be said, with **Bourdieu1977**, to be part of the collective *habitus* of the members of communities that use these expressions. Such common turns of phrase are conventional not only in the ordinary Saussurian sense in which all of language is based on arbitrary conventions negotiated among the speakers of a language (de Saussure, 1974[1916]:65–70), but additionally in the sense that although alternative ways of expression are possible, conventionalised turns of phrase are the usual ways of putting things among the community in question (**ErmanWarren2000**:30; Buerki, 2020). Two important additional insights flow from this. First, in order for there to be a usual way of putting some-

thing, that *something* has to be a meaning that is of sufficient salience and frequency of expression to develop its own usual way of being put. For example, in communities where telephone calls are frequent, the situation regularly arises that one party has to interrupt the interaction in order to either connect the caller to someone else or to carry out a task away from the phone. To make it clear to the conversation partner that this situation has arisen, the expression *please hold (the line)* is typically used in English, although there clearly would be alternative expressions that might (in the absence of a conventional phrase) be used to communicate such a meaning (perhaps *please wait*). One diagnostic of this type of conventionalisation is that other speech communities might use different usual expressions in this same situation, as **Allerton1984** points out: 'Ne quittez pas' (i.e. *don't leave*) in French, or 'Bleiben Sie am Apparat' (*stay at the apparatus*) in German. It follows that it is mainly meanings of a certain salience and regularity of occurrence in a community that develop usual ways of being put. The second insight is that conventional expressions of this type carry added value in terms of information about situations, intentions, wider understandings and conceptualisations that aids speed and accuracy of understanding (Feilke, 1994:238): *Please hold*, for example, not only communicates narrowly propositional information (that it is requested of the listener to wait, staying put), but also immediately conjures up situational information, information about participants, next turns, etc. In this sense, a conventional expression is more information rich than a novel or creative expression. In locating the essence of phraseology in usual ways of putting things in a speech community, we assert these important attributes of phraseological expressions and allow them to facilitate the reading of discourses occurring in their speech communities.

2.2 Phraseology as a barometer of societal goings-on

Phraseological expressions have featured in discourse analytical treatments of various types, but have typically remained without specific identification as phraseology. A good example is **Hart2017**'s (**Hart2017**) application of Critical Discourse Analysis (CDA) to reporting on the London riots of 2011: among key expressions identified are *fan the flames, spread [quickly] across X, take hold of Y* (2017:284–285), all of which are phraseological patterns that can conventionally be used without reference to literal fires (i.e. as dead metaphors in the sense of Lakoff and Johnson, 1980), as Hart shows. Although single words and more abstract constructions are also used by discourse analysts, phraseological expressions feature prominently in this domain (Stubbs, 2002). Attention has also been drawn

to the involvement of phraseological expressions in authorial stance and evaluation (e.g. Hunston, 2011; Biber, 2006) and the relevance of the latter in terms of building up a value system that 'is a component of the ideology which lies behind every text' (Thompson and Hunston, 2000:6).

Phraseological patterns are also noted for their particularly close links to the (changing) social and cultural issues of the communities among whom the expressions are conventionalised. As Stubbs points out, "the study of recurrent wordings is […] of central importance in the study of language and ideology, and can provide empirical evidence of how the culture is expressed in lexical patterns" (1996: 169). Studies in this area include **Linke2001**'s (**Linke2001**) analysis of the phraseology of death notices that she links to changing attitudes in society and **Wierzbicka2007**'s (**Wierzbicka2007**) study of salient cultural frameworks linked to phraseology in which she describes the expression *reasonably well* as a "whole cloud of culture condensed in a drop of phraseology" (2007:50). Handford, in his study of the language of business meetings, points out that "institutionalized clusters […] can shed light on the specific conventions of a community of practice; they demonstrate the particular approach to problems and the common communicative tools preferred by the community in question" (2010:144). Similarly, **Mair2007**'s (**Mair2007**) investigation of seven world Englishes shows "idiomatic and collocational preference are the most direct reflection of a community's attitudes and pre-occupations in linguistic structure" (2007:439). **Buerki2020**, in a comprehensive study of phraseological change across 20th century German, shows that social and cultural shifts are the single largest motivator of phraseological change (where motivation could be established), suggesting that phraseological expressions are indeed uniquely at the pulse of a community's preoccupations.

Links between phraseology and the issues that concern the community whose phraseology is studied have therefore been evidenced in a range of situations and locations. The underlying reason for the existence of these links is also apparent if, as suggested above, the essence of phraseology lies in its nature as common turns of phrase within a community: where one can detect conventionalisation of phraseological expression, it is necessarily the work of the speech community that, through repeated communicative events, starts to form common, agreed turns of phrase that facilitate communication in this area. As highlighted above, phraseological expressions also carry the added value of ready-made "pre-agreements of understanding" (Feilke, 1994:367, my translation), that is, situational, conceptual and other pragmatic associations that facilitate mutual understanding in communication and are there therefore rooted in the life of a community. Intriguingly, Seidlhofer showed that the phraseological facilitation

of 'common understanding' (2009:205) is so important that where no existing phraseological expressions are available, pro-tem phraseology is created to plug the gap. In the case of Seidlhofer's study, the lack of agreed phraseology was due to participants being in a lingua franca communicative situation, but the possibility of pro-tem, rapid creation of new phraseology must mean that similar new phraseology could arise in little time from a speech community having to engage with new situations, such as those created by the Brexit referendum or, as **Szerszunowicz2015** shows, by a root and branch change of system in a society such as occurred in Poland after 1989. It seems clear, therefore, that the phraseology of discourses has the potential to tap into the concerns of a community and make them readable, and that phraseology develops and adapts to reflect the concerns of communities, at least in some cases, within very short spaces of time.

2.3 Brexit language

Despite Brexit being a comparatively recent phenomenon, a range of treatments of Brexit and language have already emerged. Some are engaged with questions of language policy and ideology, particularly the role of English in a post-Brexit EU (e.g. Jacobsen, 2017; Kelly, 2018; Modiano, 2017), or have used corpus data to predict or analyse the outcome of the referendum using opinion mining (e.g. Celli et al., 2016; similarly Simaki et al., 2017). Others have looked at Brexit language from a lexicological and morphological perspective (e.g. Fontaine, 2017; Lalić-Krstin and Silaški, 2018) but a number have also taken various discourse-analytic approaches: **Buckledee2018** seeks to characterise how the different sides of the Brexit divide communicated their messages in the run up to the referendum and what effects these choices might have had. **Achilleos-SarllMartill2019** show how particularly the discourses of leave-supporting groups were "dominated by […] toxic masculinity […] first through the deployment of language that was associated with deal-making and, second through the deployment of language associated with militarism"(2019:15). Koller et al., (**??**) present a multi-authored collection of studies on a range of aspects of the Brexit discourse, some focussed on sub-topics, some on discourses in particular media like Wikipedia or Twitter. The volume also features what appears to be the only expressly phraseological treatment on Brexit: **Musolff2019**'s (**Musolff2019**) insightful study of the single proverb *having your cake and eating it* which has come to unexpected prominence in the Brexit discourse. Other analyses of the discourse on Brexit focus on metaphors employed: **Islentyeva2019**, looking at texts from the British right-wing press, identifies such underlying metaphors as RELATIONSHIP WITH EU-

ROPE AS A (BROKEN) MARRIAGE (evident in expressions like *the divorce bill*) and **Charteris-Black2019**'s (**Charteris-Black2019**) book-length treatment of Brexit metaphors also uncovers an impressive range of source domains evident in the Brexit discourse, from sinking ships to distrust and betrayal, to war and invasion, many of which are conveyed via phraseological patterns. A possible master metaphor, suggests Charteris-Black, involves a nostalgic regression back to an idealised past: Brexit as time travel. **Mair2019** presents a programmatic discourse-analytic paper on British Euroscepticism since 1945 using the Hansard and News on the Web corpora. Going 'beyond the lexicographical level [to include] the many new ways in which existing words are combined' (2019:2), encompassing what Mair regards as essential 'historical time-depth' (2019:2), the paper presents fascinating nuggets of language that trace the outline of British discourses on Europe. These include relative frequencies over time of empire-related, Commonwealth-related and Europe-related words as well as various tropes around concepts like control (*take/want (one's country) back* and *take back control*) that are traced across decades of use to show how current Brexit discourse is connected to earlier Eurosceptic discourses.

Phenomena that could be classed phraseological are part of some of the analyses of Brexit language discussed. However, express treatments of the phraseology of Brexit discourses are to date largely missing, despite the clear potential for such analyses to add important aspects to the current understanding of Brexit discourses.

2.4 Questions

It is therefore pertinent to ask what a phraseological reading the discourses of Brexit might reveal. To investigate this, it is necessary first to establish the phraseology of Brexit in all its facets: casual observation as well as the existing studies suggests that Brexit has occasioned the creation of new or newly prominent multi-word terms, slogans, and other expressions. It appears also to have brought fresh prominence to some phraseological dinosaurs (*cherry picking* / *having one's cake and eating it*). But do these familiar high-profile expressions reflect all of the phraseology of Brexit, or are there other, perhaps less noticeable turns of phrase that are important? Is there one phraseology of Brexit, or have different phases of Brexit produced their own distinct phraseological repertoires? Is the Brexit discourse more peppered with phraseology than comparable discourses? Once we have a good grasp of answers to these questions, we will be able to read what the phraseology says about the Brexit discourse in the UK and consider what this

might mean for a phraseological discourse analysis going forward. In the following section, we turn to the data and methods used to derive the phraseological inventory of the discourse on Brexit.[1]

3 Data and methods

To derive the phraseology of the Brexit discourse in the UK, a data-led and comprehensive corpus-linguistic method was chosen. Much of Brexit phraseology, like other aspects of Brexit, has made such a forceful entry into public consciousness that it appears not only observable, but almost unavoidable in seemingly any set of texts discussing the topic. Nevertheless, and despite the evidently important insights gained from the analysis of selected examples picked through close reading of texts as demonstrated by existing work on the language of Brexit, a more rigorous approach, it would seem, is useful for at least two reasons pertaining to present aims: first, despite the many obvious patterns, there may well be other features of the discourse that are more hidden and less accessible to conscious selection than might be assumed. Therefore an approach that is data-led in that, as much as possible, the structure of the data itself suggests items for analysis, ensures that less obvious patterns are not as easily missed. Second, a researcher's judgement in selecting items for analysis is likely influenced by their own experience, their own perception of the important features of that discourse, or their own experience with language, thereby inevitably introducing a selection bias that may inadvertently mask certain features of the discourse before they are recognised. Therefore, a data-led and comprehensive approach such as the one outlined below seeks to defer the vital expert judgements involved in the selection of features to as late a stage in the analysis as feasible. In the present case, this means that the identification of phraseological features is based on a comprehensive automatic extraction of phraseology from corpus materials, and the identification of Brexit phraseology is achieved through a careful contrasting of texts on Brexit with contemporary texts on other topics.

Data

The data used for the investigation comprised 80 million words of media texts published between the 20 **February2016** (when the date of the Brexit referendum was announced) and 31 **October2019** (when the UK missed the second deadline

[1]The discourse on Brexit (or Brexit discourse) is here used variously in the singular or plural; the singular encompasses all sub-discourses (plural) that might exist within the total societal discourse on Brexit.

for leaving). Texts were obtained from a content provider's database and comprise texts from UK publications, sampled on random dates of each month of the period under investigation. Texts are included from national newspapers (including their online outlets), regional and Sunday newspapers as well as sources such as the Press Association Newswire, magazines like MoneyWeek and the Spectator, trade journals like Banking and Credit News and the Metal Bulletin, web publications like independent.co.uk and a sprinkling of transcripts of radio and television broadcasts (the latter making up 3.6% of documents). This very broad range of publications included means that the corpus data cover not only the discourse on national news media, but arguably approach the full range of public discussion, including specialist and popular discourse.[2]

On each sampled date, texts producing a closely similar number of total words were sampled from texts containing the word *Brexit* on the one hand and from texts excluding the word *Brexit* on the other hand. This resulted in 40 million words each of Brexit-related texts and non-Brexit-related texts. Only complete texts were included. In addition, to allow comparisons between different phases of Brexit, word counts were balanced across the four periods shown in Table 7.1, with each period containing 10 million words of text on Brexit and 10 million words of text on non-Brexit topics. Doubtlessly, periods different from those chosen could have been used – the significance of the transition dates on which divisions are based, however, makes this a reasonable representation of the phases of Brexit, even though it certainly is not the only possible representation. The chosen transition dates are the day the referendum was announced (20 **February2016**), the day of the referendum itself (23 **June2019**), the day the UK officially notified the EU of its intention to leave (29 **March2017**), the end of the 2-year negotiation period (29 **March2019**) and the end of the first extension period to negotiations, 31 **October2019**. The unequal lengths of periods mean that shorter periods were sampled more densely in terms of sampling dates than longer periods to obtain the same number of words for each period.

The resulting structure along the two dimensions of topic (Brexit vs. non-Brexit texts) and time period (periods 1 to 4) is shown in Figure 8.1. Across the eight resulting sub-corpora, 161,850 texts of an average length of just under 500 words, published on 156 different days, across 747 different media titles are represented in the corpus. A breakdown by media type is shown in Figure 7.2.

Method

The identification of relevant Brexit phraseology was carried out in three main steps. First, phraseological expressions were automatically extracted from each

[2]It does not, of course, sample discussion in semi-private and private discourse which will be influenced by, but could differ from public discourse.

Table 6.1: Phases of Brexit represented in the corpus

periods	dates		length	wordcount
1: pre-referendum	20/02/2016	–	4 months	20 million
	23/06/2016		9 months	20 million
2: post-referendum	24/06/2016	–	24 months	20 million
	28/03/2017		7 months	20 million
to	29/03/2017	–		
invocation of art. 50	29/03/2019			
	30/03/2019	–		
3: initial 2-year exit negotiation period	31/10/ 2019			
4: extension period				

Figure 6.1: Corpus structure. Note: each sub-corpus has a size of 10 million words

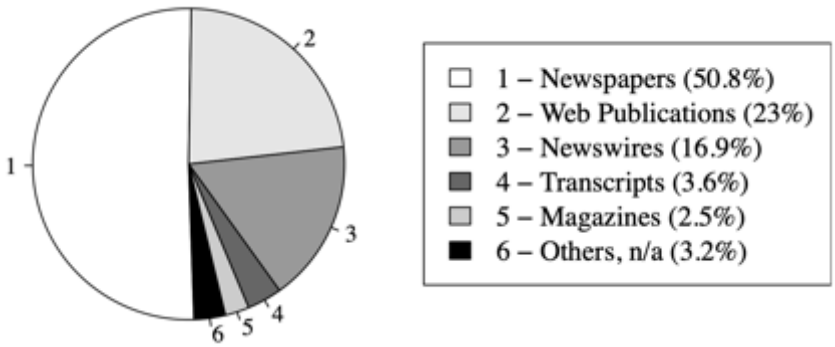


Figure 6.2: Breakdown of document sources. Note: Newspapers include UK-wide titles as well as regional and local titles and free papers like the Metro. Web Publications include online portals of print media as well as online-only publications such as The Independent; Newswires include specialist Newswires such as Sport News or Global Banking News; Transcripts are of radio and TV programmes; Magazines include titles like The Spectator and the THES (100% = 161,850 documents).

of the eight sub-corpora of 10 million words. Second, Brexit-related phraseological expressions were identified by contrasting expressions extracted from the Brexit sub-corpora with those extracted from non-Brexit sub-corpora. Finally, Brexit phraseology of the different phases of Brexit were compared with each other to find out to what extent these differed.

To facilitate the first step, an operationalization of the concept of phraseological patterns as consecutive word sequences of two to seven words in length, occurring at least three times per million words in the corpus and forming a semantic unit was chosen (cf. also Buerki, 2020: ch. 3). This operationalises the earlier definition of phrases (= word sequences) that represent common ways (occurring >= 3/M) of putting things (semantic units) in a speech community. The particular speech community here are English speakers in the UK (co-referential with the discourse community whose discourses we seek to explore).

Measuring conventionality via frequency has definite drawbacks (as Wray, 2002:31, and others have pointed out, there are highly conventionalized phrases that are rare in general language use). However, it is felt that for present purposes, this drawback is sufficiently mitigated by the chosen low frequency threshold, which is very low compared to the range typically employed in corpus-linguistic research on formulaic language (e.g. 4/M in McCarthy and Carter, 2002:12; 10/M in Biber et al., 1999) and by not using frequency as the sole criterion. Further, although looking exclusively at consecutive word sequences biases the search towards the more fully lexically specific end of the phraseological spectrum, internal variable elements (e.g. *a [nice/nasty/complete] surprise*) tend to follow a Zipfian distribution (**Ellis2012**), meaning that the few most frequent of those elements account for most of the variation and can often be extracted *in situ* if a low minimum frequency is chosen as is the case here. Nevertheless, it remains important when looking at results, to explore patterns around automatically identified sequences to establish a fuller picture of phraseological patterns. Finally, the idea that sequences should form semantic units (similar to the way single words or structurally complete phrases form semantic units) is an important additional constraint by which sequences are excluded that happen to be frequent purely because their constituent words are highly frequent (e.g. *of the, and he*). Some sequences naturally only form semantic units once we allow for the implicit variable element (e.g. *at the age of X*); these sequences were also deemed to possess semantic unity.

Figure 6.3: Summary of extraction procedure

With this operationalization in place, a full, automatic extraction of operationalization-compliant sequences out of the eight sub-corpora was carried out. As mentioned

above, a full automatic extraction has the advantage of allowing a maximally inter-subjective identification of relevant expressions as well as being able to deal with the large amounts of data available, whereas a manual identification of expressions would necessarily need to be highly selective with respect to the data that can be processed. The procedure described in **Buerki2016**, summarized in Figure 7.3, was employed for the extraction of phraseological expressions. This procedure has previously achieved good accuracy (Buerki, 2016:23), a high recall due to a low frequency cut-off, and remains to date the only fully documented procedure that extracts consolidated word-sequences of different lengths, making it a tool well suited for present purposes. The eight lists so derived contained between 25,182 and 30,202 types of phraseological expressions. The number of word tokens contained in extracted expressions varied from 4.4 million words (4-Non-Brexit list) to 5.6 million words (5-Brexit list), out of the 10 million words contained in each of the eight sub-corpora.

To identify Brexit-specific phraseology, the lists of expressions obtained from the four non-Brexit sub-corpora (1-Non-Brexit to 4-Non-Brexit in Figure 8.1) were first combined into a single master list of non-Brexit expressions containing all expression types from the four time periods (51,760 expression types in total). Second, each of the four lists derived from the Brexit sub-corpora (1-Brexit to 4-Brexit in Figure 8.1) were compared to the non-Brexit master list.

As the goal was to identify Brexit-specific phraseology, the emphasis was, in the first instance, on phraseological patterns exclusive to the Brexit sub-corpora.[3] Consequently, each of the four Brexit-sub-corpus derived lists were compared to the non-Brexit master list and any shared expressions were eliminated from the Brexit lists, leaving only expressions unique to the Brexit lists. Given the size of the corpora underlying the non-Brexit master list (four times 10 million words), it was felt that the risk of identifying an expression as Brexit phraseology in error because of a chance absence from the master list was low enough to be acceptable. For analysis, a Brexit expression master list was created, containing all expression types of the four periods, with frequencies across the periods summed. This list was ordered in decreasing order of frequency and the focus was on mid- to high frequency expressions across the periods.

A secondary analysis was prepared of relative, rather than absolute Brexit expressions, that is, expressions that differ between Brexit and non-Brexit sub-corpora only in frequency (rather than in being present in one and absent in the other). For this purpose, expressions were identified that were far more frequent

[3]Some of the exclusive expressions do occur in the non-Brexit sub-corpora, but below the frequency at which they are here considered phraseological expressions.

in the Brexit sub-corpora than in the non-Brexit sub-corpora by using the keyness formula suggested by **Kilgarriff2009**, with a +N value of 100 added to each frequency count to prioritise mid- to high-frequency expressions (s. Kilgarriff, 2009). Consequently, an ordered list of absolute Brexit expressions, as well as one of relative Brexit expressions was created for analysis.

Finally, to investigate the development of Brexit phraseology across the four periods of investigation, two further lists were prepared: a list of expressions exclusive to a single Brexit period, not found among the expressions of other periods, and a second list of expressions with frequencies that varied most across the four Brexit periods.[4]

As the number of expressions on each of these lists was far too large for individual analysis, the lists were ordered according to likely typicality for each list's focus, providing a rigorous and maximally inter-subjective basis for a judged selection and analysis of individual expressions and groups of expressions. The results of these selections and their analyses are presented in the next section; their implications are discussed in the final section of this chapter.

4 Results

This section presents results that allow us to draw conclusions regarding the questions set out above. We start by considering the data on the question of whether the Brexit discourse is more phraseologically rich than comparable discourses, that is, whether it has a higher density of phraseological expressions – something that, while not previously formally analysed, may be suggested by popular comments about the discourse being reduced to slogans or by the finding that there is a large amount of metaphor applied to the Brexit discourse, much of it via set expressions and phrases (cf. Charteris-Black, 2019:2). Table 7.2 shows the proportion of word tokens that are part of automatically extracted phraseological expressions in Brexit and non-Brexit texts across the four periods of study. Non-Brexit texts consistently show a notably lower density of phraseological expressions.

Before we can conclude, however, that Brexit discourse is more phraseologically dense, it is worth considering whether the difference in density might be influenced by the difference between a topically more uniform set of texts (Brexit texts all contain the word *Brexit* and might therefore be in a sense *about* Brexit)

[4]The top 8,000 Brexit expressions, together with their frequencies across all four periods, are searchable at brexitphrases.herokuapp.com.

Table 6.2: Phraseological density of Brexit and non-Brexit texts. Note: figures represent the ratio proportion of total word tokens that are part of phraseological expressions. For key to periods, see Table 7.1.

	Brexit texts	non-Brexit texts
Period 1	0.63	0.46
Period 2	0.53	0.46
Period 3	0.49	0.44
Period 4	0.56	0.46
mean	0.55	0.46

and a topically diverse set of texts (non-Brexit texts cover all other topics). If topical diversity might be linked to a greater diversity of phraseological expression, it could mean that it is more difficult to identify conventional turns of phrase in topically diverse texts (leading to fewer expressions being identified) because there is less repetition of the same phraseological patterns. It seems, however, that although there may well be topic-related influences, the conclusion of clearly higher phraseological density in Brexit texts is justified: first, although Brexit texts are at least partially about Brexit because they contain the word *Brexit*, the phenomenon of Brexit is characteristically one that touches all areas of life. That is to say, there may well be discussion of Brexit in texts about sport, cooking or money matters as well as politics and news, whereas similarly, non-Brexit texts may cover any of these topics. In this sense, Brexit texts still contain remarkable topical diversity. Second, to see if indeed non-Brexit phraseology was more diverse, two indicators show no clear support for notably more diversity in non-Brexit phraseology: in terms of type-token ratios of expressions, Brexit texts show an average type to token ratio of 1:66 among phraseological expressions, for non-Brexit texts of 1:70, indicating only a somewhat elevated diversity in expressions rather than a dramatic difference. In terms of how internally diverse Brexit expressions and non-Brexit expressions are, one might look at what proportion of expression types and tokens are shared between pairs of periods across Brexit and non-Brexit texts: out of a mean of 30,539 expression types identified per period in Brexit texts (range: 28,040 – 34,188), on average 47.9% are shared with each of the other three periods. In the case of non-Brexit texts, the figure is notably higher at 57.5% (mean of types 25,693; range 25,182 – 26,600), indicating less diversity within the phraseology of non-Brexit texts, rather than more diversity. For expression tokens, the equivalent figures are 63% for Brexit texts and 71% for non-Brexit texts. The comparatively elevated phraseological density

of Brexit texts therefore seems to be a genuine feature of the Brexit discourse rather than attributable to differences in topical diversity.

Turning now to examples of Brexit expressions, it is clear that these are multi-faceted and rich. They range from multi-word terms to slogans to collocations and other usual sequences, many of which are embedded in new phraseological patterns that have become conventional ways of talking about Brexit-related issues. Some of the expressions are new coinages, some are used in new senses or previously existed only in specialist discourses. We start by looking at multi-word terms. Among the large number of multi-word terms are examples (??) to (??).[5]

- the single market and customs union

allow/end free movement / freedom of movement

trigger article 50

avoid/return to a hard [Irish] border

cut/reduce net migration

- a second referendum
- the cliff edge
- a Brexit deal
- the remain / leave camp
- hard Brexiteers
- hard / soft Brexit
- the will of the [British] people
- project fear
- the British people
- red line[s]
- personal attack[s]

[5]Square brackets, [], indicate optional variable elements of an expression, a slash indicates alternatives.

- Brexit uncertainty

These have become well-established and widely used across the different phases of Brexit. (??) to (??) are technical terms that have been catapulted into widespread use in the Brexit discourse. One remarkable feature is how, around these expressions, extended phraseological patterns and very clear collocational preferences have arisen that are evidently the product of rapid conventionalization: for example, frequent patterns involving either the whole phrase in (??), or one of the co-ordinated expressions, include *[in/outside/access to/stay in/remain in/leave] the [European/EU's/EU] single market.* Similarly, (??), (??), (??) and (??) have their preferred verbal collocates as indicated. Expressions that are topically similar to (??) include *the Irish border issue* and *the Good Friday Agreement,* the latter an example of a relative Brexit expression (occurring in non-Brexit texts, but at a much lower frequency). This cluster highlights the magnitude of difficulties created by Brexit on the island of Ireland. Perhaps the most interesting case of collocational preference is (??) which has developed a virtually exclusive preference for forms of the verb *to trigger*. The alternative *invoke article 50* makes a brief appearance in phase two of Brexit, but is unable to establish itself. The metaphorical triggering seems to suggest the setting in motion of an unstoppable train of events (whereas a revocation can follow an invocation), forcing a particular conceptualisation of events. There are other examples where there seems to be a competition between alternative phrasings, with the existence of a dominant pattern that is (or becomes) the conventional way of expression: (??) is by far the most usual way of referring to the idea of another referendum on the UK's membership of the EU to be held after the Brexit referendum of 2016. However, in period 4, there are other expressions for the same concept (*a people's vote*; *a confirmatory referendum*; *a final say referendum*), none of which reach the frequency of *a second referendum*. This appears to show that once a phraseological expression is established as the usual way to express a meaning, it is very difficult to challenge it, even if the conventional expression arguably carries with it a certain (possibly undesirable) conceptualisation of the world: the alternatives to *a second referendum* seek to avoid the implication of a re-run of the 2016 referendum, for example. Similarly, (??) is a conventional expression for referring to a sudden, disorderly exit from the EU,[6] but arguably carries with it the vivid picture of catastrophe that many Brexit-supporting language users might wish to dispel. Similar, arguably forced, conceptualisations built into phraseological expressions

[6] *disorderly exit* itself only just makes the cut for being a phraseological expression in its own right in period 4, but in no other period.

are evident in (??) reflecting the language of deal-making (cf. Achilleos-Sarll and Martill, 2019)[7], as well as in (??) to (??). (??) to (??) are expressions that shine a light on concepts evidently prominent in the discourse – the concept of a nation, of tough negotiation stances, the bruising arguments, and of the effects of Brexit. A shift in convention is observable in the data in relation to labels applied to people supporting or opposing Brexit: *the remain camp* and *the leave camp* (notably militaristic expressions, cf. Achilleos-Sarll and Martill, 2019) have fallen out of use as phraseological expressions by period 4, being replaced by snappier single word expressions (e.g. *remainers / leavers*, also (??) and similar). *leavers and remainers*, as per (??), is a binomial attested with increasing frequency from period 2 onwards, but there are also attestations of the form *remainers and leavers* in the data, showing that this expression has not yet reached the status of *irreversible* binomial. A similar observation holds for (??), which is far more frequent in the cited form, but also appears as *the customs union and [the] single market* and so has not solidified to an irreversible binomial either, showing that these expressions are still in the process of being fully established.

- leavers and remainers
- our children and grandchildren (periods 1& 3)
- parents and grandparents (period 1)
- strong and stable (periods 3& 4)
- our [European] friends and partners (period 4)

The binomials shown in (??) to (??) are more clearly irreversible in their order. Particularly fascinating are (??) and (??) – not only are they relative Brexit expressions (they occur at low frequency in non-Brexit texts and predate the period of observation as phraseological expressions), they are also specific to period 1, the pre-referendum period. (??) speaks to the epochal gravity and enormity of the Brexit decision (demonstrated by typical usages such as 'with our children and grandchildren 's future at stake […]') and to an extent, they both also stand for the two sides of the argument: a proportion of contexts for (??) suggests that 'our children and grandchildren' would wish to remain, whereas the referents of (??) tend in most contexts to be thought of as natural leavers (e.g. 'Cameron has pleaded with parents and grandparents to vote to stay').

[7]The more neutral *withdrawal agreement* only becomes a phraseological expression in period 4.

There are, however, irreversible binomials that were formed within the Brexit discourse itself: (??) and (??) are examples of (deliberate) coinages that are also closely tied to specific periods of Brexit: (??) appears in period 3 and peaks in period 4 (the run-up to the **June2017** general election), and (??) is exclusive to period 4 where it serves as a label for the EU. Whereas (??) and (??), below, conceptualise the EU as (ex-) family (cf. also Islentyeva, 2019), (??) appears friendlier, but with the sting of a far more clearly distanced relationship. The nature of (??) as a deliberate coining that appears in official communications and pronouncements in period 4, opens the possibility that the distancing might be another example of attempted forced conceptualisation.

Further, the category of phrasal verbs and other verbal patterns is exemplified by expressions (??) to (??).[8]

- pressed on X
- lashed out at X
- press ahead with X
- argued for X
- free up [money/£350 million a week/cash/…]
- leave on WTO terms
- revised down

lumped/fell/tumbled by as much as X

- raised the prospect of X
- have a negative impact on X

Expressions (??), as in 'Mr Barclay was pressed on what would happen if …', to (??) by virtue of appearing (at above-threshold frequency) exclusively in Brexit-related texts appear to reveal something of the way in which social actors behave in public discourse: they might be evasive (??), brash in tone (??), stubborn (??), and engaged in permanent arguments (??), in addition to indulging in *personal attacks* (16). (??) is found in contexts where it is argued that leaving the EU will result in improved state finances, (??) could be seen as an alternative rendering

[8]In these examples as elsewhere, *X* stands for a non-optional variable element.

of the situation referred to in (??) as *the cliff edge* and (??) to (??) chime with (??) above in indicating recurrent patterns related to the impact of Brexit.

A further category of expression are slogans, where deliberate coinages are most clearly in evidence. It is noteworthy that such slogans show up in the data (which after all do not include political speeches as such, nor parliamentary proceedings) – it seems the originators of these expressions are influential enough to be able to achieve wide dissemination. These expressions show a tendency to simplify issues and are ideologically highly charged. (??), (??), (??) and (??), discussed earlier, share with these expressions their likely or certain status of being deliberate coinages.

- the best possible deal
- Brexit means Brexit
- no deal is better than a bad deal
- take back control [of [immigration / our laws, borders, money and trade / ...]]
- the best deal [for [families and businesses / Britain / every part of the UK / ...]]

Examples of other usual sequences and phraseological patterns are shown in (??) to (??) and include epoch references as in (??) to (??); the latter is a relative Brexit expression (occurring at far lower frequencies in non-Brexit texts) and suggests that Brexit is being perceived as an event of a magnitude that invites comparison to World War II. Various patterns referencing uncertainty (including the difficulty of predictions as in (??)) and repercussions of Brexit, (??) to (??), are also evident. Further clues to the societal climate (??) and the seemingly ever-present possibility of short-term dramatic shifts in situations (??), are also apparent.

- since the referendum / since the Brexit vote / following Brexit
- in the post-Brexit [era/world] / [in] post-Brexit Britain
- since the Second World War
- fall in the value of the pound / the weak[er] pound

ter/more/stronger than expected

- the uncertainty surrounding [Brexit/the status of EU nationals/the UK's future relationship with the EU/...]
- because of Brexit / as a result of [Brexit/leaving [the EU]]
- what Brexit will mean for X

concerned about the impact/effect of Brexit [on X]

- the potential impact of X
- X [has] warned [of] Y / warning that [Brexit could] X

what Brexit means [for [business/Wales/the future of X/...]]

- the [pressing/critical/biggest/major/real/...] issues facing X
- the anger [of X / at X / among X / ...]
- at the time of writing

The final category, proverbs and idioms, is exemplified in (??) to (??). Their comparatively high frequency in the discourse is highly notable because pure idioms and proverbs are rare in normal language use (Moon, 1998). (??) is an allusion to the proverb *You can't have your cake and eat it* and occurs in texts in relation to aspects of the Brexit negotiating position of the UK government, as noted in previous analyses (Charteris-Black, 2019; Musolff, 2019). In this sense, it could be seen as challenging the assertive *red line[s]* (??) in the negotiations that refer to positions that will be preserved under all circumstances. (??) is similarly used as a criticism of negotiating positions taken by the UK government, and finds its counterpoint perhaps in the accusation of (??). (??) is an expression specific to period 4; contexts suggest it is an evaluation of the series of exit date extensions of that period.

have cake and eat[ing] it

- cherry pick[ing]
- kick[ing] the can down the road (period 4)

Before concluding the review of examples of Brexit phraseology, it is important to highlight examples of discontinuity in the phraseology of Brexit across

the four time periods covered. Above, we reviewed figures showing that, on average, each period shares 47.9% of its Brexit expression types and 63% of expression tokens with each of the other periods. Clearly, there is therefore continuity in Brexit expressions across various periods and most expressions so far reviewed occur in most periods.[9] But there are also significant shifts: we have already encountered shifts in designations of supporters and detractors of Brexit from (??) to (??). Similarly, examples (??) to (??), which occur above the phraseological threshold (3/M) in only one or two periods, show that there is a clear sense in which different periods have their own phraseology.

- referendum on Britain's membership of the EU (period 1)
- a European army (period 1)
- economic migrants (period 1)
- ever closer union (period 1)
- concerns about immigration (period 1)
- invoke article 50 (period 2)
- divorce proceedings (period 2)
- in the aftermath of the Brexit vote (period 2)
- the fallout from the Brexit vote (period 2)
- regulatory divergence (period 3)
- speech in Florence (period 3)

a/the [second] meaningful vote (period 3)

- a constitutional crisis (periods 3 and 4)

Brexit divorce bill (periods 3 and 4)

- fourth meaningful vote (period 4)
- the Benn act (period 4)

[9] It is likely also that many of these expressions will remain part of the language for the long term – (??), for example, might well become a staple phrase similar to (??).

- reaching an agreement is still possible (period 4)
- more and more difficult (period 4)
- proroguing parliament (period 4)
- abject surrender (period 4)
- if Britain leaves the EU (absent in period 2)
- recognition that X (absent in period 4)

In some cases, period-specific phraseological expressions have been replaced by other expressions in subsequent periods – (??) was replaced by *the referendum* or *the Brexit referendum*, likely for reasons of economy of expression. (??) to (??), as well as (??) and (??) as noted earlier, appear to express concerns largely of the pre-referendum period. (??) and (??) are among expressions reflecting the shock of the immediate post-referendum period and the processes that had to be initiated, e.g. (??) and (??). Periods 3 and 4 seem to have some overlap in their concerns and expressions of these periods document wranglings over agreements between the EU and the UK, or over the lack of agreements, as in (??), (??) to (??).

Other expressions appear (and disappear) with the relevance of their denotations: *meaningful vote* in (??) and (??) – a technical term for a vote in parliament (repeated several times) on the withdrawal agreement negotiated by former Prime Minister May – became irrelevant shortly after the events, similarly (??).

By contrast to expressions that only appear in one or two periods, (??) and (??) are examples of expressions that are (conspicuously) absent from only one of the Brexit periods. (??), for example, appears in contexts where social actors concede a point with an amount of humility – its fall from a top frequency of over 7/M in period 1 to disappearance from the discourse in period 4 may be a further indicator of polarisation and a worsening of the tone of the discourse.

There are also notable shifts in frequencies of expressions that appear across periods: (??), for example, while frequent in all four periods, shows a generally increasing frequency development, whereas mentions of (??) follow a generally decreasing trend. Former Prime Minister May's slogan in (??) comes in at over 8 times per million words in period 2, but its frequency has halved by period 4 while her slogan in (??), similarly, is very frequent in periods 2 and 3, but halves in frequency for period 4.

The appearance and disappearance of expressions in specific periods shows that where needed, conventionalised expressions can be brought into use in a

speech community over very short intervals, indeed. This seems clear evidence for the existence of pro-tem phraseological expressions (to borrow Seidlhofer's term), not just at the micro-level of small group communication (cf. Seidlhofer, 2009), but at the level of a complete speech community such as the speakers of British English. From the point of view of phraseological theory, it is stunning to see widely used and circulated pro-tem phraseology documented in the data – as well as seeing such a rapid creation of new phraseological expressions and to observe expressions as they are formed, such as the binomials that are not fully irreversible. Both in terms of the significance of pro-tem phraseological expressions at the level of a speech community as well as the rapidity of phraseological development, these observations open up areas of phraseological study that require more investigation and reflection, but they are in agreement with recent findings on the rapidity of phraseological change (cf. Buerki, 2019).

Having quantified aspects of the phraseology of Brexit and reviewed examples showing the richness and diversity as well as the continuity and discontinuity of Brexit phraseology, we are now in a position to discuss, in the final section, answers to the questions posed at the outset.

5 Discussion

Results presented in the preceding section now allow us to attempt a reading of the discourses of Brexit through their phraseology. Three areas seem worth particular mention:

First, the domains and meanings encoded phraseologically, in other words, those meanings sufficiently frequently communicated to develop usual turns of phrase used in the Brexit discourse, allow us to read some of the concerns of the Brexit discourse, what is happening within society and what the issues of public concern are. At a necessarily general level, these include:

- An impression of the complexity of Brexit (the specialist multi-word terms that have entered common use in the discourse)
- Conversely, efforts to present Brexit in simplistic terms, cf. the slogans in (??) to (??)
- Indicators of the roughness of much of the discourse: (??), (??), (??) to (??), (??)
- Topical preoccupations on borders and immigration, e.g. (??), (??), (??), (??) to (??), cf. also **Mair2019**

- Polarisation, with many terms appearing in opposing sets, cf. further discussion below
- A very deep sense of uncertainty, expressed in (??), (??) to (??), (??) to (??), (??), (??)
- The epoch-defining status of Brexit as in (??) to (??)
- A realization of deepening crisis within the discourse itself with the emergence of (??), perhaps emblematically (??) and the collocation (??) brought back from the obscurity of history.

The data here show that popularly recognised high-profile expressions, e.g. (??) or (??), do reflect the phraseology of Brexit, but there are many more subtle patterns that have slipped under the radar while just as much part of Brexit phraseology. The bottom-up, comprehensive procedure employed in this study was able to bring these to the surface as well. Some of them are less noticed patterns that are able to add extra insight that is important: the patterns around uncertainty and consequences of Brexit, for example, are more extensive and prominent than perhaps generally acknowledged and verbal patterns revealing aspects of the tone of the discourse are equally highly revealing.

Second, diachronic aspects of the analysis suggest that the discourse itself is fast-paced and ever changing. On the one hand, this can be read as speaking to the ever new challenges and consequences of Brexit emerging and requiring discussion (e.g. in (??) and (??)), and thus further to the sense of instability and insecurity evident within the discourse. There appear to be expressions that are tied to particular phases of Brexit in a way that makes them appear dated or out of place in other periods. In this respect, there are phraseologies (plural) of Brexit, as well as a common phraseological bedrock of Brexit expressions. This points similarly to a plurality of discourses within the overall Brexit discourse.

On the other hand, variation and fast change also reveal intense conflicts over alternative conceptualisations of key aspects of the Brexit narrative that are played out to a notable extent in phraseology: opposing expressions, some deliberately coined, some more naturally occurring, are vying for the status of the usual way in which their meaning is expressed (and with it the usual way in which that domain is conceptualized). Cases in point are the expressions (??) vs. (??) as well as (??) vs. (??) and (??) vs. (??). These are by the end of the period of observation still very much contested. As observed, no one side or tendency has been successful in getting all 'their' conceptualisations accepted in the

community – there remains a diversity of ideologically incompatible conceptualisations in current use, pointing to an ongoing and vast array of domains that remain contested in the discourse. There are some aspects that have been settled – *people's vote* vs. *second referendum,* are shown to have been settled in favour of the second conceptualisation, for example, but these are relatively few. Notably, (??), *if Britain leaves the EU*, shows a high frequency in the pre-referendum period, but disappears as a common turn of phrase in period 2 (the immediate post-referendum period) indicating that the most fundamental question (whether or not Brexit will happen) appears settled. However, the data show that it re-emerges as a common turn of phrase in periods 3 and 4, showing that by the end of the period of observation, the most fundamental question regresses into the category of what is contested within society.

Third, beyond struggles in the diachronic development of phraseological expressions, Brexit phraseology in general very often seems ideologically charged, as shown in examples (??), (??), (??), (??), (??) and (??) to (??), in particular. This indicates the limited possibility of neutrality in this discourse: participants in the discourse cannot but take sides in one way or another if they wish to speak. Particularly in cases where expressions force a certain conceptualisation of events, often through conceptual metaphors (Lakoff and Johnson, 1980), deliberate coinages feature strongly. This attempted forced conceptualisation of a domain, particulary in political discourse, is sometimes labelled 'framing' (Lakoff, 2010) and has been extensively documented by Lakoff. The finding of a concentration of attempted forced conceptualisation in phraseological expressions indicates that phraseology itself appears to be instrumentalised (not to say *weaponised*) by actors in the discourse.[10] However, slogans and other deliberate coinages are joined by more naturally conventionalised expressions and many of the coined terms are themselves embedded in usual phrasings, while other attempted coinages are very short-lived or so evidently counter-factual that they can now only be used with irony as is the case with (??) and so these linguistic power struggles are not artificial in nature. Rather, they can reasonably be read to reflect societal struggles, some of which have recently been labelled *culture wars* (Sobolewska and Ford, 2020).

A notable additional feature uncovered is that the Brexit discourse is more peppered with phraseology than comparable discourses, perhaps reflecting an entrenchment of often polarized views among those participating in the discourse.

[10] **Wintour2020** reports that 'Foreign Office staff have been banned from using certain words and phrases in discussing Brexit – including "implementation period", "no deal", "special partnership" and even Brexit itself […]'

Likely parallels can also be drawn to what **Szerszunowicz2015** termed a *periodic growth of phrasemes*, an intensive increase in the number of phraseological expressions 'triggered by an important event in the history of a particular culture' (2015:103). Szerszunowicz demonstrates the phenomenon using the 1989 change of system in Poland which 'influenced greatly all spheres of life in Poland, such as politics, economy, culture' (2015:103) and led to the creation of a great number of new phraseological expressions. Although Szerszunowicz primarily documents an increase in types (rather than specifically an increase in the phraseological density of texts, i.e. of tokens), she notes a general colloquialisation of public discourse which included an increased use of idioms and sayings as well as that 'the ability to include many [scientific] terms [and expressions] into public speeches was an important element' (2015:108). The observation that the Brexit discourse is more phraseologically dense than discourses on other topics (as well as the parallels regarding the creation of many new expression types) could therefore also point to the magnitude of change (in this case in all spheres of life in Britain) brought on by Brexit – a finding that contributes intriguing facets of insight to the emerging field of Brexit studies.

To conclude, I would like to propose three theses, based on discussed findings. These should serve to move the current state of research forward, firstly by underpinning and supporting findings of existing work that has sought to bring phraseology to bear on discourse analytical questions. Secondly by making it more attractive to overtly declare phraseological work in discourse analysis as such and in so doing benefit from the support of phraseological theory, and finally by encouraging the use of phraseological means of reading discourses to a far fuller extent in the interest of further advances in discourse analysis and in the interest of the stretching, testing and mapping out of the boundaries of the validity of these theses:

- Phraseological patterns allow deep insight into what is happening within society, what the issues of public concern are, what is contested or settled (including how this changes over time) – in short, they allow us to read the discourses of a community. The likely precondition to this is a robust, data-led identification of relevant patterns. This first thesis is, so the hope, borne out by the results presented, but also leads to the realisation of thesis 2.

- Discourses crystallise (to a remarkable extent) in phraseology. This is regardless of whether the items of phraseology under scrutiny are attempts at deliberate coinages (where successful, these develop their own embeddings and extended patterns through natural conventionalisation) or not

and whether they are pro-tem items or patterns that are part of the language over longer periods of time. Indeed, where diachronic aspects can be assessed, this necessarily adds important additional angles (cf. Mair, 2019) even over short periods of time, such as the 44-month period investigated in this study.

- Phraseological theory explains why all this should be the case: 1 and 2 above are not merely empirical curiosities but follow from the essence of phraseology as common turns of phrase that represent conventional, usual ways of putting things in a speech community. As such, items of phraseology that are the result of communal discursive practice and negotiation (part of the sediment of social practice, to speak with Bourdieu, 1977), are by virtue of their nature salient and contain not only more propositional meaning than other items of language but carry pre-understandings and conceptualisations of reality (Feilke, 1994, Lakoff, 2010). That is why they allow deeply penetrating access to the discourses of a community.

References

Boguslavsky, Igor. 2003. On the passive and discontinuous valency slots. In *MTT conference proceedings, 16–18 June 2003, Paris*.

Burger, Harald. 2003. *Phraseologie: Eine Einführung am Beispiel des Deutschen* (Grundlagen der Germanistik GrG 36). Berlin: Erich Schmidt Verlag.

Cowie, Anthony Paul. 1998. *Phraseology: Theory, analysis, and applications*. Oxford: Oxford University Press.

Cowie, Anthony Paul & Ronald Mackin (eds.). 1975. *Oxford dictionary of current idiomatic English*. London: Oxford University Press.

Cowie, Anthony Paul, Ronald Mackin & I. R McCaig. 1993. *Oxford dictionary of English idioms*. Oxford; New York: Oxford University Press.

Crowther, J., S. Dignen & D. Lea. 2003. *Oxford collocations dictionary: For students of English*. Oxford: Oxford University Press.

Evert, S. 2005. *The statistics of word cooccurrences*. Word Pairs and Collocations. University of Stuttgart. (Doctoral dissertation).

Hausmann, Franz Josef. 2004. Was sind eigentlich kollokationen. In K. Steyer (ed.), *Wortverbindungen: Mehr oder weniger fest*, 309–334. Berlin: De Gruyter.

Heid, Ulrich & Rufus H Gouws. 2006. A model for a multifunctional dictionary of collocations. 979–988.

Moon, R. 1998. *Fixed expressions and idioms in English : A corpus-based approach*. Oxford: Oxford University Press.

Moroz, Andrzej. 2013. Zależność a kookurencja: Dwa różne sposoby wiązania wyrażeń. *Studia Językoznawcze* 12. 121–132.
Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Straka, Milan & Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. Vancouver, Canada*, 88–99. Association for Computational Linguistics.

Chapter 7

Grammar patterns as an exploratory tool for studying formulaicity in English-to-Polish translation: A corpus-based study

Łukasz Grabowski
University of Opole

Nicholas Groom
University of Birmingham

In this chapter, we explore the use – and argue for the usefulness – of the concept of grammar patterns (**FrancisEtAl1996**6, 1998; **HunstonFrancis1999**) in descriptive research on the English-to-Polish translation of formulaic language. Specifically, we use the *Paralela* English-Polish parallel corpus (Pęzik 2016) to explore – largely in terms of frequency distributions – the use of the Polish equivalents of selected English multi-word items, which are textual manifestations of grammar patterns. As a proof-of-concept, we will focus on a pre-selected grammar pattern ('*it* v-link ADJ *to*-inf'), where a given word (e.g. the adjective *possible*) may convey different senses depending on the pattern in which it occurs (**Groom2005**). We aim to check whether and to what extent somewhat similar lexico-grammatical patterns emerge from the Polish language data (i.e. from the Polish translations under scrutiny). Using this method, we are also able to investigate whether the Polish equivalents of that English grammar pattern (i.e. from its textual manifestations) are realized with the same level of regularity in translation. Early findings revealed the English pattern '*it* v-link ADJ *to*-inf', when filled by adjectives conveying the sense of 'difficulty', corresponds to a set of Polish syntagmatic patterns, such as 'ADJ v-link, *aby*', 'ADV v-link' or 'ADV' (the last two ones followed by verbs in the infinitive form), which contribute to the formulaicity of Polish texts. It is argued not only that the study's findings indicate that grammar patterns are a useful starting point for the exploration of formulaicity in translation, but also that they help us explain some

more general differences in terms of semantics, pragmatics and usage in source texts and their translations.

1 Introduction

1.1 General remarks on Pattern Grammar

In the last three decades, corpus linguists have proposed a number of novel research methods and procedures for capturing and exploring recurrent patterns of language use. One of them is Pattern Grammar (**HunstonFrancis2000**), a corpus-based grammar of the English language that describes the lexico-syntactic environments of individual lexical items. These 'grammar patterns' (henceforth GPs) are defined as "a phraseology frequently associated with (a sense of) a word, particularly in terms of the prepositions, groups, and clauses that follow the word" (**HunstonFrancis2000**: 3) or as "all the words and structures which are regularly associated with the word and which contribute to its meaning" (ibid. 37). As argued by **Roemer2009**, GPs are neither single words nor empty grammatical structures; they constitute abstract representations of frequent lexico-grammatical patterns. For example, Hunston and **Francis2000** claim that in the patterns ***it*** **v-link ADJ that** (e.g. *it is interesting/clear that*) and ***it*** **v-link ADJ *to*-inf** (e.g. *it is sensible/possible to*), the adjectives belong to similar meaningfully-related groups, e.g. expressing a range of different concepts such as likelihood, importance, desirability and obviousness.

Whilst Pattern Grammar is principally an empirical approach to linguistic description, it also puts forward two major theoretical claims. The first of these is that the different meanings (i.e. senses) of polysemous words can be distinguished on the basis of their typical occurrence in different patterns. Consider, by way of example, the following two (invented) sentences:

- ***It's*** **possible that** *she sent a message.*
- ***It's*** **possible to** *send a message.*

It should be immediately clear that, although the GPs highlighted in bold in these sentences both feature the same adjective, ***possible***, the meanings that they make are very different. In example (**??**), the highlighted words can be replaced with the single adverb ***maybe***, whereas in example (**??**) the highlighted part can be replaced with the two words ***you*** **can**. From a traditional perspective, it might

be said that this shows that the adjective ***possible*** has two different meanings: an epistemic modal meaning in example (**??**), and a dynamic modal meaning in example (**??**). From the point of view of Pattern Grammar, however, the epistemic and dynamic modal meanings made by these two sentences reside not in any of the individual words, nor in the morphosyntactic configurations into which these words fall, but in the interaction between them. That is, the epistemic 'maybe' meaning belongs to the whole sequence '*it is possible* + that-clause', and the dynamic 'you can' meaning is made by the whole sequence '*it is possible* + to-infinitive clause'. This insight is of major significance because, if it is true for a language as a whole, it means that semantic ambiguity is virtually non-existent in naturally-occurring language. Each of the different meanings of a polysemous word will be associated with a different structural pattern, and language users will interpret which meaning is being made in each case on the basis of this patterning. In other words, Pattern Grammar proposes that disambiguation is far more a matter of attending to linguistic co-text than it is a matter of attending to extralinguistic context, as is generally assumed in traditional theories of semantics and pragmatics.

The second major theoretical claim put forward by Pattern Grammar is that "words which share a given pattern tend also to share an aspect of meaning" (**HunstonFrancis2000**: 3). To return to examples (**??**) and (**??**) above, ***possible*** is just one member of a class of adjectives that have a broadly epistemic modal meaning when they occur in the pattern ***it*** **v-link ADJ** ***that*** (e.g. *certain, definite, demonstrable, doubtful, feasible, implausible, incredible, irrefutable, likely, true, uncertain, unthinkable,* etc.), and just one of many adjectives that have a broadly dynamic modal meaning when they occur in the pattern ***it*** **v-link ADJ** ***to*****-inf** (e.g. *difficult, easy, feasible, hard, impossible, impractical., simple, tough, tricky*, etc.). In Pattern Grammar, these classes of words that share a particular pattern/meaning association are referred to as 'meaning groups'. It is important to note that each grammatical pattern is not restricted to just one meaning group; on the contrary, many if not most patterns have several different meaning groups associated with them. For example, **FrancisEtAl1998** identify eight distinct meaning groups for the pattern ***it*** **v-link ADJ** ***that***, and nine different meaning groups for the pattern ***it*** **v-link ADJ to-*inf***.

Although Pattern Grammar is unquestionably a corpus-based approach to grammatical description, it is nevertheless highly reliant on the manual qualitative analysis of concordance lines, and thus on the application of human judgement in the identifying of patterns and the meaning groups associated with them. This is because the identification of GPs in attested language corpora essentially involves perceiving a similarity between linguistic items which may not be iden-

tical on the surface, but which have some underlying regularity of form and meaning, which at the current time can only be reliably identified by a human analyst.[1] For example, the sequence of words *it is ironic that* is a representative of a general pattern beginning with it, ending with a *that*-clause (with optional *that*) and containing a linking verb (e.g. *be, seem, look*), followed by an adjective expressing evaluation of a given situation, e.g. *it is not surprising that, it looks very unlikely that, it seems very peculiar that* (**HunstonFrancis2000**: 154). Such judgements are often highly nuanced and subtle, and it is sometimes problematic even for human researchers, let alone computer software, to distinguish between sequences of words which are formally the same yet differ in terms of their patterns, e.g. 'annoyance + relative clause' vs. 'annoyance + appositive that-clause' (**HunstonFrancis2000**: 67).

Despite its labour intensive nature, Pattern Grammar has been widely and successfully applied to the description and analysis of English over the last two decades or so. Perhaps the best known fruits of this research are the two monumental *COBUILD Grammar Patterns* reference works (**FrancisEtAl1996**; **19981996**), one covering 700 patterns across 9,000 different verb senses, and the other covering 100 further patterns across 10,000 nouns and adjectives. In addition to these general reference works, corpus-based studies have also provided empirical evidence that GPs differ in systematic ways across different discourses and genres, with academic English being a particular focus of interest (e.g. **Groom2005**; **Charles2006**; **20072006**; **Larsson2016**; **SuHunston2019**). The COBUILD Grammar Patterns reference works have also been used as an empirical basis for psycholinguistic investigations into first and second language speaker knowledge of verb argument constructions (e.g. **EllisEtAl2014**; **RömerEtAl2014**; **20152014**).

These latter inter-varietal research perspectives suggest that Pattern Grammar might also be used to carry out empirical research in cross-linguistic contexts. Surprisingly, however, no work along such lines has been done to date; in fact, although there is some comparative cross-linguistic research on grammar patterns in the specialized genre of judicial decisions (**PontrandolfoGozdz-Roszkowski2014**) focusing on the discoursal function of evaluation in English and Italian, PG has as yet never been systematically applied to any other language than English (Hunston, personal communication).

[1]The automatic identification of GPs has been demonstrated as feasible in principle (**MasonHunston2004**), but has not as yet been achieved at scale on open text.

1.2 Towards Pattern Grammar for languages other than English

In fact, there has been no usage-based description similar to Pattern Grammar capturing recurrent lexico-grammatical patterns of language use in Polish, i.e. based on the relationship between meanings or discourse functions on the one hand and structural patterns on the other. However, similar to valency dictionaries developed for English (e.g. *Erlangen Valency Patternbank*[2] or *Pattern Dictionary of English Verbs*[3]), there are dictionaries of Polish that combine syntactic and semantic information on particular lexical items in ways which bear some comparison with the Pattern Grammar approach. One of them is *Walenty*[4], a comprehensive valency dictionary of Polish (**PrzepiórkowskiEtAl2017a**; **2017b2017a**) that specifies arguments of predicates for verbs (and – to a lesser extent – for nouns, adjectives and adverbs), which consist of comprehensive syntactic and semantic formalisms, the latter ones not fully implemented yet (**PrzepiórkowskiEtAl2017a**: 10). Although utilized by two parsers of Polish (*Świgra* and POLFIE), the dictionary has been designed primarily for computer processing of natural language texts as well as for researchers (linguists) and lexicographers (ibid.: 9). For example (and unlike in the case of Pattern Grammar), verbs recorded in *Walenty* are described using the valency frames, each with a set of argument specifications (ibid.), e.g. the verb *otruć* ('to poison') has three syntactic frames and one semantic frame, all accompanied by examples of use (Figure 8.1).

Although *Walenty* is similar to Pattern Grammar in that it presents syntactic environments of lexical items, it does not allow the researcher to browse through larger lexico-grammatical patterns (e.g. verbs, nouns or adjectives and the words or syntagmatic frames that follow or precede the words) or to classify the slot-fillers in those patterns into specific meaning or functional groups, which is precisely what the Pattern Grammar approach does. For example, the so-called 'introductory *it*' pattern followed by a link-verb, adjective and to-infinitive clause (***it* v-link ADJ *to*-inf**)[5] reveals a number of meaning groups for adjectives filling in the pattern (Figure 7.2) as well as specific examples of their use (Figure 7.3).

[2]EVP is available at: http://www.patternbank.uni-erlangen.de/cgi-bin/patternbank.cgi; it provides a list of valency patterns for 511 verbs, 544 adjectives and 274 nouns. See **HerbstEtAl2004** for more details.

[3]PDEV is available at: http://pdev.org.uk/#about_cpa; it provides systematic description of meaning and use of verb patterns of 1451 verbs (as of 9 **July2019**). See Hanks (2008a, 2013) for more details.

[4]Primarily a dictionary of Polish subcategorization frames, Walenty can be directly compared to FrameNet (http://framenet. icsi.berkeley.edu) grounded in the theory of frame semantics (**Fillmore1982**).

[5]The example is available at: https://grammar.collinsdictionary.com/grammar-pattern/it-v-link-adj-to-inf_1

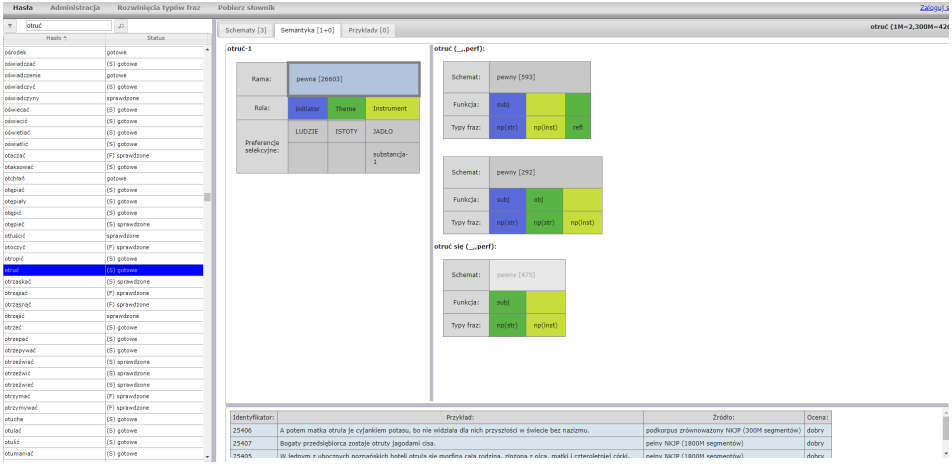


Figure 7.1: . The verb otruć ‘to poison’ in Walenty (retrieved from: http://walenty.clarin-pl.eu)

Grammar Patterns › Adjectives

it v-link ADJ to-inf

it v-link ADJ to-inf

The adjective group follows introductory *it* and a **link verb**, and is followed by a **to-infinitive clause**.

it	verb group	Adjective group	to-infinitive clause
It	is	easy	to see what he means.
It	is	essential	to pay in advance.
It	's	hard	to believe he just forgot where he was.

Adjectives with this pattern belong to the following meaning groups:

The 'accurate' and 'illogical' group
The 'easy' and 'difficult' group
The 'selfish' and 'dangerous' group
The 'exciting' and 'safe' group
The 'surprising' and 'interesting' group
The 'important' group
The 'legal' group
The 'funny' group
The 'enough' group

Figure 7.2: . The pattern ‘it v-link ADJ to-inf’

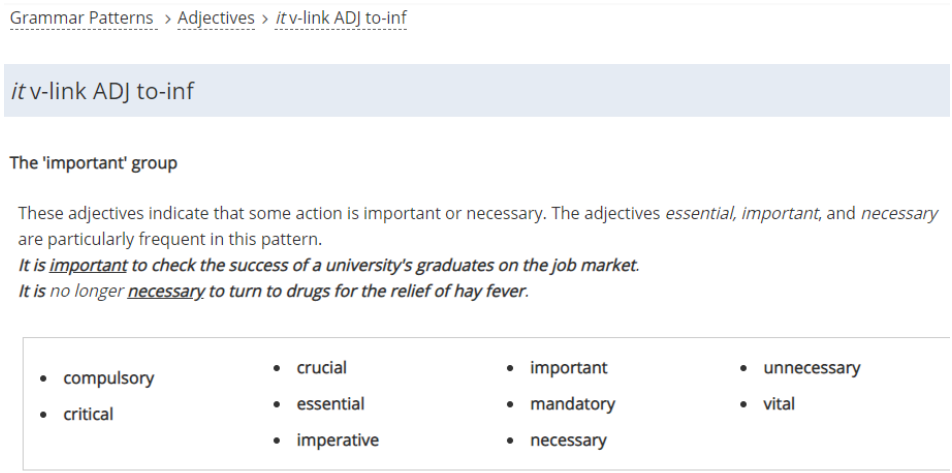


Figure 7.3: . The pattern 'it v-link ADJ to-inf' filled with adjectives conveying the sense of importance.

As can be seen in Figure 7.3, the sense of importance (or attitudinal stance) of the following proposition is conveyed by the entire grammar pattern ***it* v-link ADJ *to*-inf** rather than by individual adjectives filling in the pattern. Consequently, Pattern Grammar offers an inventory of lexico-grammatical constructions which constitute pairings of form with semantic or discoursal function.[6] Such a resource describing lexico-grammatical patterns extracted from corpora in a bottom-up fashion has not been developed for the Polish language so far.

The main goal of this study is therefore to explore whether GPs may be employed as a useful exploratory tool for cross-linguistic studies. In other words, we aim to identify and describe lexico-grammatical patterns that emerge from the English-to-Polish translations – extracted from Paralela corpus (Pęzik 2016) – of a pre-selected English GP performing specific discoursal functions. More precisely, we will focus on one English 'introductory it' pattern with *to*-clause complementation (***it* *v-link* ADJ to-inf**), which will constitute the starting point for our analysis.[7] However, the *tertium comparationis* are the functional cate-

[6] In practice, identification of GPs is possible by analyzing concordance lines, which later involves grouping the patterns into notional categories (e.g. topical or functional ones) on the basis of different types of meanings conveyed in contexts. Although individual words can help in determination of those groupings, a qualitative analysis of a wider context of their occurrence is necessary to form appropriate groups and thus identify the patterns (**HunstonFrancis2000**: 162).

[7] Both patterns were studied by **Groom2005** in terms of their variability across different academic registers.

gories or discoursal functions (necessity, importance, obviousness etc.) contingent on individual words filling in the patterns or, in other words, worked out on the basis of intuitive understandings of words in the English GPs. This will allow us to investigate whether we may find corresponding generalizable lexicogrammatical patterns in Polish that perform the same discoursal functions as the English GP under scrutiny. The underlying assumption is that in theory the Polish translations should convey the same information as their English source texts. We will also attempt to verify whether GPs are useful as a discovery tool for detecting translation patterns. We believe that a study like this one may pave the way to a wider application of the Pattern Grammar approach for the description of languages other than English (in this case, Polish) and/or for cross-linguistic comparisons or translation-oriented research, which have not been attempted hitherto.

The chapter is structured as follows. In §**??**, the research material (i.e. the study corpus), units of analysis and methodology will be described. Next, we describe and discuss the empirical results exemplifying translation patterns emerging from textual realizations of GPs in the sample of English source-texts and their Polish translations. The concluding section discusses the study limitations and offers suggestions on how this research may be developed further in the future.

2 Methodology

2.1 Research material

For the purposes of the current study, we will use the English-Polish parallel corpus *Paralela* (Pęzik 2016), which is available at: http://paralela.clarin-pl.eu. The parallel corpus under scrutiny includes a little more than 262 million word tokens in 10,877,000 translation segments from various text types and genres, including written, spoken and to-be spoken texts, e.g. legal documents from various European Union institutions (legislation, transcripts of proceedings of the European Parliament etc.), press releases, medical texts, film subtitles, popular science texts, literary classics, transcripts of European Parliament proceedings etc. As a rule, a parallel corpus contains source texts aligned with their translations in the target language. In this study, we will analyze a pre-selected GP ('*it* v-link ADJ *to*-inf') found in a single genre of English source texts, namely European Parliament proceedings (henceforth EPP), and we will try to align its textual realizations with their Polish equivalents as found in translation segments in the corresponding Polish sub-corpus (with more than 13 million word tokens in almost 700,000

translation segments)[8] Then, we will attempt to quantitatively and qualitatively analyze the target language equivalents and see whether any regular patterns, i.e. lexico-syntactic associations similar to GPs, emerge from the Polish language data.

2.2 Units of analysis

At least in theory, parallel corpora include texts (originals and translations) that express the same meanings and perform the same discoursal functions, which allows one to search for correspondences between linguistic items in source and target texts (**Johansson2007**: 9, cited in **Marco2019**: 43). As mentioned earlier, in this study we will focus on a pre-selected GP, which means that the lexical items under scrutiny will not be extracted from texts in a bottom-up approach. This is mainly because we do not have access to full texts collected in *Paralela*. Hence, the use of the study corpora will be limited to the analysis of bilingual concordances illustrating particular translation patterns, i.e. frequency and distribution of the Polish equivalents of the English GP, which provide a starting point for our investigation.

We will capitalize on the results of the study conducted by **Groom2005**, who analyzed, among others, two GPs, namely ***it* v-link ADJ that** and ***it* v-link ADJ to-inf**. **Groom2005** noticed that adjectives which convey the meaning (i.e. sense) that can be generalized as "validity" (e.g. *clear, inconceivable, obvious*) tend to fall into the pattern ***it* v-link *ADJ* that** while the adjectives conveying the sense of "difficulty" (e.g. *difficult, easy, hard*) tend to fall into the pattern ***it* v-link ADJ to-inf**[9]. Also, it was found that depending on the pattern, one and the same word (e.g. the adjective *possible*) may convey different senses, that is either "difficulty" or "validity" (**Groom2005**). In summary, the findings of the study conducted by **Groom2005** provide strong evidence of the relationship between particular sense conveyed by particular words and the structural patterns in which those words tend to occur.

We will search for a single pre-selected pattern using the SlopeQ query syntax implemented in *Paralela* (Pęzik 2016) as well as morphosyntactic tags (e.g. *<tag=j.*>* stands for adjectives). Thus, the pattern ***it* v-link ADJ to-inf** will be searched for using the following query:

[8]The transcripts were originally extracted from Europarl corpus (**Koehn2005**) and included in Paralela (Pęzik 2016). The debates were recorded on 11–12th and 23rd **October2006**, and translated from English into Polish.

[9]However, **Groom2005** found that distributions and more fine-grained rhetorical functions of the adjectives in those grammar patterns vary across corpora representing different language varieties.

- *it <tag=v.*> <tag=j.*> to* (34,047 occurrences in Paralela; 3,067 occurrences in EPP).

In view of a high number of occurrences, it is necessary to filter out the results to facilitate the qualitative analysis of concordance lines. To this end, we will apply systematic sampling by recording translation pairs of every 30th concordance. This means that we will ultimately focus on a sample of 100 translation pairs (i.e. bilingual concordances illustrating Polish translations of specific textual realizations of the English GP) extracted from the EPP sub-corpus of *Paralela*.

2.3 Research questions and hypotheses

The main problem addressed in this study concerns whether the corresponding items (translation equivalents) found in target texts can also be described in terms of recurrent lexico-syntactic associations similar to the GP identified in the English source texts. Hence, in this exploratory paper we aim to provide answers to the following research questions:

1. What are the Polish equivalents of the English source-language multi-word items emerging from GPs?
2. Can the Polish equivalents be generalized into a more abstract set of lexico-grammatical patterns similar to GPs?
3. Can the Pattern Grammar approach be used for a description of Polish?
4. Can GPs be applied as a unit of analysis in cross-linguistic contexts?

2.4 Research questions and hypotheses

This study will be conducted in a number of stages. First, we will preselect one GP (described earlier) and develop an inventory of their textual realizations in English source-texts, i.e. in the EPP sub-corpus of *Paralela*. Then, we will generalize the results by means of grouping overlapping textual realizations into a list of n-grams performing specific discourse functions. After identification of the Polish equivalent, or translation variant (be it a single-word or a multi-word unit), we will investigate whether the observed patterns of Polish translations can be generalized to a more abstract phraseological, syntagmatic or lexico-grammatical units, similar to GPs.

3 Preliminary results: grammar patterns in contrast

In this exploratory study, we used the GP ***it* v-link ADJ to-inf** as a unit of analysis and a tool for discovering potential translation patterns. All in all, the said GP occurred in the EPP sub-corpus of Paralela (Pęzik 2016) 3,067 times. In order to limit the amount of data for manual analysis of bilingual concordances, we used systematic sampling and selected every 30th concordance, which resulted in the set of 100 English-Polish translation segments to be analyzed. Depending on the adjectives filing in the slot, we used the procedure put forward by **Groom2005** and – by conducting manual analysis of English source-text fragments – we classified the textual variants of the GP found in the sample into semantic/functional categories (discoursal functions) corresponding to the senses conveyed by the adjectives. This way, the linguistic data were classified into the following categories: 'importance', 'validity', 'desirability' and 'difficulty'.

First, we present the results of the analysis of the textual instantiations of the GP ***it* v-link ADJ to-inf** as filled with adjectives conveying the sense of 'difficulty' in the English-original texts in the EPP sub-corpus (8 occurrences) and its equivalents in the Polish translations.

Table 7.1: Textual realizations of the GP

Textual realizations in the English-original	**Discoursal function**	**Polish equivalents**	**Generalized pattern of Polish translations**
it is ***difficult*** *to* (??)	DIFFICULTY	***trudno*** *jest* (??)	ADV v-link v-inf
it is ***hard*** *to* (??)	DIFFICULTY	***trudno***, *aby* (??)	ADV, *aby*
		trudno (??)	ADV v-inf
it is ***easier*** *to* (??)	DIFFICULTY	***łatwiej*** (??)	ADV v-inf

***it* v-link ADJ to-inf** *in English source-texts and their Polish translations: discoursal function of difficulty*

The findings (Table 7.1) show that in the sample under scrutiny there are three different instantiations of the English GP '*it* v-link ADJ *to*-inf' when filled with adjectives conveying the sense of 'difficulty', namely *it is difficult to*, *it is hard to* and *it is easier to*, with the total frequency of 8. Those multi-word items have the following Polish equivalents in Paralela corpus (Pęzik 2016), such as *trudno jest* (used 5 times as an equivalent of *it is difficult to*), *trudno, aby* and *trudno* (used 1 each as equivalents of *it is hard to*) and *łatwiej* (used once as an equivalent of *it is*

easier to). Apart from insights into certain translational choices, it has been possible to reconstruct abstract lexico-grammatical patterns (or syntagmatic frames) based on the Polish equivalents and conveying the sense of difficulty with respect to the following proposition, e.g. 'ADV v-link v-inf' (*trudno jest*) or 'ADV v-inf', both followed by a verb in infinitive form (*trudno jest udowodnić* 'it is difficult to prove', *trudno uwierzyć* 'it is difficult to believe') or 'ADV, *aby*' (*trudno, aby* etc.) followed by a complement clause, e.g.:,

(1) ***It* is hard to** *believe when reading it.*
Trudno *uwierzyć w te słowa, kiedy się je czyta.* [EVOeRj]
This is because ***it* is difficult to** *prove that the service rendered was of poor quality.*
Dzieje się tak dlatego, że ***trudno* jest** *udowodnić, że świadczone usługi były złej jakości.* [qoavWA]
Simplification of the common agricultural policy is a beautiful idea, and ***it* is hard to** *imagine that someone would oppose it.*
Uproszczenie wspólnej polityki rolnej to piękna idea i ***trudno,* aby** *ktoś był jej przeciwny.* [ea2Eoq]

As can be seen, the English GP '*it* v-link ADJ *to*-inf', when filled by adjectives conveying the sense of 'difficulty', corresponds to a set of Polish syntagmatic patterns, such as 'ADV, *aby* + complement clause', 'ADV v-link v-infinitive' or 'ADV v-infinitive'.

The following example under scrutiny refers to the GP '*it* v-link ADJ *to*-inf' as filled with adjectives (*possible, clear, impossible, true*) conveying the sense of 'validity' in the English-source texts (Table 7.2).

The findings show that there is more variety among Polish lexico-grammatical patterns that convey the sense of 'validity' in the Polish translations as compared with the sense of 'difficulty'. The most frequent one (4 occurrences) is the pattern 'ADJ v-link' (with positional variation) and 'ADJ v-link, *że*', which is realized with the following Polish equivalents, namely *możliwe* (*będzie*); *możliwe jest;* (…) *jasne jest, że*; *jest* (…) *jasne, że.* Other patterns are centred around nouns ('NN', 'NN v-link' and 'NN v-link, *że*'), which include the following words and phrases: *możliwość, niemożliwością jest* and *prawdą jest, że* respectively, e.g.:

(2) *I would like to say that it is becoming clear in this discussion that* ***it* is possible to** *have a 'two-speed ' Europe.*

Table 7.2: Textual realizations of the GP 'it v-link ADJ to-inf' in English source-texts and their Polish translations: discoursal function of validity

Textual realizations in the English-original	**Discoursal function**	**Polish equivalents**	**Generalized pattern of Polish translations**
(will) it be ***possible*** *to* (??)	VALIDITY	***możliwe*** *(będzie)* (??)	ADJ v-link
it is ***clear*** *to (us/me)* (??)	VALIDITY	*dla (nas)* ***jasne*** *jest, że* (??)	ADJ v-link, *że* 'that'
		jest dla (mnie) ***jasne****, że* (??)	v-link ADJ, że 'that'
it is ***impossible*** *to* (??)	VALIDITY	***niemożliwością*** *jest* (??) 'impossibility is'	NN v-link NEGprt + MOD v
		nie można (??) '	
it is ***possible*** *to* (??)	VALIDITY	***możliwe*** *jest* (??) 'possible that'	ADJ v-link MODv v-inf
		można (??) 'may/-can'	TRANSFORMATION
		możliwość[10] (??) 'possibility'	
it is ***true*** *to* (??) say	VALIDITY	***prawdą*** *jest, że* (??) 'truth is that'	NN v-link, *że* 'that'

(EN) *I would like to say that* ***it*** **is becoming clear** *in this discussion that it is possible to have a 'two-speed' Europe.* (PL) *Chciałbym powiedzieć, że w tej dyskusji* ***oczywista*** **staje się możliwość** *istnienia Europy "dwóch szybkości".* [Drag45]

Chciałbym powiedzieć, że w tej dyskusji oczywista staje się ***możliwość*** *istnienia Europy "dwóch szybkości*[11]*".*[Drag45]

There are also two patterns centred on the modal verb *można* 'can/may' ('MODv', 'NEG PRT MODv'), namely *można* and *nie można*, followed by the infinitive form of the verb, e.g.:

(3) (…) *but the one thing I have learned is that* ***it*** **is impossible to** *book a ticket on the Eurostar when you are travelling.*
(…) *lecz nauczyłem się jednego, a mianowicie, że* ***nie*** **można** *zarezerwować biletu na pociąg.* [BNzRG4]

A large group of multi-word items in the English originals conveys a sense of 'importance' (e.g. *it is important to, it is crucial to, it is essential that*) and there is a high variety among their Polish equivalents, which can be grouped into a number of lexico-grammatical patterns. The most prominent ones include the pattern 'ADJ v-link' and 'ADJ v-link, *aby/by*' (e.g. *istotne jest*; *ważne jest, aby*; *niezbędne jest; konieczne jest*), with adjectives occasionally modified by adverbs (e.g. *bardzo ważne jest*, *niezwykle istotne jest*). It is particularly noticeable that the pattern 'ADJ v-link' is followed by a verbal noun functioning as a direct object, while the pattern 'ADJ v-link, *aby/*by' or 'ADJ, *by*') is followed by the verb in the infinitive form, e.g.:

(4) *Finally, I would like to stress that* ***it*** **is important to** *have full transparency regarding the founding of the initiative and sources of financial support for the organisers.*
Wreszcie, pragnę podkreślić, że ***ważne*** **jest** *zapewnienie pełnej przejrzystości w odniesieniu do finansowania inicjatywy oraz źródeł wsparcia finansowego dla jej organizatorów* [RDwmvo]
It **is important to** *overcome the current problems that characterise the sector: the lack of competition, the regulatory over-dependence on ratings, and the low reliability of notes.*
Ważne **jest**, **aby** *rozwiązać problemy, które obecnie spotyka się w tej branży, a mianowicie przesadne zawierzanie ratingom i małą wiarygodność ocen.* [APaN8r]

[11]The phrase *Europa dwóch szybkości* 'two-speed Europe' has not been adopted in the Polish press discourse. Instead, the phrase *Europa dwóch prędkości* has become commonly used.

Other frequent patterns include 'MODv v-inf' (*trzeba, musimy/musi, należy* followed by verbs in the infinitive form) or 'NN ADJ v-link' (e.g. *sprawą podstawową jest*; *ważną rzeczą jest/są*) or 'V ADJ NN' (e.g. *ma fundamentalne znaczenie*), cf. Table 7.3 below.

Finally, similar correspondences can be observed in the case of structures conveying the sense of 'desirability' in the English originals and their Polish equivalents (Table 7.4). The most frequent patterns – with positional variation due to free word order – in the Polish translations were found to be 'ADJ v-link' (e.g. *konieczne jest, jest niestosowne, nieodpowiedzialne jest, słuszne jest*), 'ADV v-link' (e.g. *dobrze jest*), 'MODv v-inf' (e.g. *warto, należy*). The comparison of certain GPs provides interesting insights into cross-linguistic correspondences on the syntactic level. For example, the English pattern '*it* v-link ADJ v-inf' may correspond, among others, to the Polish GP 'MODv v-inf', where the position of direct and indirect object with respect to the patterns under scrutiny changes, e.g.

***It* is essential to** *give people* [Indirect object] *access* [Direct object] *to healthcare, drinking water and sanitation.*

Ludziom [Indirect object] ***należy*** *zapewnić dostęp* [Direct object][13] *do opieki zdrowotnej, wody pitnej i kanalizacji.* [DrZG0o]

We also identified ready-made formulas corresponding to specific speech acts, (e.g. the polite phrase used after a greeting *Cieszymy się, że jest Pan tu z name* versus *it is nice to have you with us*) or transformations resulting from stylistic changes in the translation as compared to the original. Both fixed formulas and transformations do not fall into GPs that we attempted to identify among the Polish equivalents.

4 Discussion and conclusions

The findings of our exploratory study indicate that the Pattern Grammar approach holds unexplored potential for cross-linguistic research, of both text-oriented and system-oriented kinds. The approach showcased in our study used a set of GPs as a starting point to identify recurrent English multi-word items with specific meanings (senses) and discoursal functions, which were then aligned with their Polish equivalents. We obtained further empirical evidence against a widely held misconception about translation whereby one has to translate a given fixed phrase from L1 (English) into a corresponding fixed phrase in L2 (Polish). Using this methodology, we were able to obtain a set of lexico-grammatical frames emerging from the English-to-Polish translation patterns in the corpus under

[13] Lit. 'people should be given access'

Table 7.3: Textual realizations of the GP 'it v-link ADJ to-inf' in English source-texts and their Polish translations: discoursal function of importance

Textual realizations in the English-original	Discoursal function	Polish equivalents	Generalized pattern in Polish translations
it is ***important*** *to* (??)	IMPORTANCE	***ważne*** *jest, aby* (??)	ADJ v-link, *aby*
it is ***crucial*** *to* (??)	IMPORTANCE	***istotne*** *jest* (??)	ADJ v-link
		kluczowe *znaczenie ma* (??)	ADJ NN V
it is ***essential*** *to* (??)	IMPORTANCE	***konieczne*** *jest* (??)	ADJ v-link
			ADJ NN V
		kluczowe *znaczenie ma* (??)	ADJ v-link MODv v-inf
		niezbędne *jest* (??)	V ADJ NN
		trzeba (??)	(ADV) ADJ v-link
		ma ***istotne*** **znaczenie** (1)	MODv v-inf
		niezwykle istotne *jest* (??)	
		należy (??)	
it is ***fundamental*** *to* (??)	IMPORTANCE	*ma* ***fundamentalne*** **znaczenie** (??)	V ADJ NN
it is ***imperative*** *to* (??)	IMPORTANCE	***bardzo*** **ważne** *jest* (??)	(ADV) ADJ v-link
it is ***important*** *to* (??)	IMPORTANCE	*należy* (??)	MODv v-inf
		ważne *jest* (??)	ADJ v-link
		musimy (??)	MODv v-inf
		należy	MODv ADV v-inf
		koniecznie (??)	ADJ v-link, *aby*
		ważne *jest, aby* (??)	MODv v-inf
			ADJ v-link
		trzeba (??)	ADJ v-link
		bardzo **ważne** *jest* (??)	ADJ N v-link ADJ, *by*
		konieczne *jest* (??)	ADJ v-link
		ważną **rzeczą** *jest/są* (??)	
		ważne*, by* (??)	

Table 7.4: Textual realizations of the GP 'it v-link ADJ to-inf' in English source-texts and their Polish translations: discoursal function of desirability

Textual realizations in the English-original	Discoursal function	Polish equivalents	Generalized pattern in Polish translations
*it is **necessary** to* (??)	DESIRABILITY	***konieczne** jest* (??)	ADJ v-link
*it is **advisable** to* (??)	DESIRABILITY	OMISSION[14] (??)	OMISSION
*it is **appropriate** to* (??)	DESIRABILITY	***warto*** (??) 'it is worth'	MODv v-inf
*it is **better** to* (??)	DESIRABILITY	***lepiej*** (??) 'better'	ADV V
*it is **fair** to* (??) (*say that…*)	DESIRABILITY	***uczciwie*** (1) 'fairly'	TRANSFORMATION[15]
*it is **good** to* (1) know	DESIRABILITY	***dobrze** jest* (??)	ADV v-link
*it is **inappropriate** to* (??)	DESIRABILITY	*jest **niestosowne*** (??)	v-link ADJ
*it is **irresponsible** to* (??)	DESIRABILITY	***nieodpowiedzialne** jest* (??)	ADJ v-link
*it is **nice** to* (??) (*have you with us*)	DESIRABILITY	*Cieszymy się, że* (??) *jest Pan tu z nami* 'we are happy to have you here with us'	FIXED FORMULA
*it is **pointless** to* (??)	DESIRABILITY	*nie ma sensu* (??)	NEGprt V NN
*it is **profitable** to* (??)	DESIRABILITY	*przynosi zysk*	V NN
*it is **right** to* (??)	DESIRABILITY	*należy* (??) *jest **słuszne*** (??) ***słuszne** jest* (??)	MODv v-inf v-link ADJ ADJ v-link
*it is **sensible** to* (??)	DESIRABILITY	*(jest)* ***rozsądne*** (1)	v-link ADJ
*it is **unacceptable** to* (??)	DESIRABILITY	*nie do przyjęcia jest* (??)	NEGprt PREP NN v-link
*it is **uplifting** to* (1)	DESIRABILITY	*świadomość* (…) *podnosi na duchu* (??) 'awareness of (…) raises sb's spirits'	TRANSFORMATION

scrutiny (EPP). As well as yielding insights into language in use in English-to-Polish translation in the EPP corpus, our methodology revealed a number of novel and valuable cross-linguistic correspondences between formulaic structures in English and Polish. As our study was based on a specific text type, it remains to be seen to what extent those Polish lexico-grammatical patterns can be generalized to other text types or genres (if we adopt a textual perspective), and to the Polish language system (if we adopt a cross-linguistic systemic perspective). Nevertheless, our preliminary results are certainly encouraging.

The findings of the current study also show that the Pattern Grammar approach holds unexplored potential not only for cross-linguistic studies, but also for the description of recurrent lexico-syntactic constructions found in other languages, such as Polish. Clearly, such a research agenda needs to be pursued with caution in view of the many typological differences between English and Polish (cf. **Fisiak1978**; Willim & Mańczak-**Wohlfeld1997**). In practice, what this means is that a "Polish version" of Pattern Grammar will need to be mindful of formal descriptions of Polish, and that we should expect to find non-formally corresponding grammar patterns that convey the same meanings or discourse functions in English and Polish, as was the case with the English GP '*it* v-link ADJ *to*-inf' and the lexico-syntactic patterns that emerged from the Polish translations, rather than exact equivalents for each given English GP. Indeed, we may find that there are one-to-one, one-to-many, many-to-many and many-to-one relationships between GPs conveying the same meaning or discourse function in English and Polish. Since adjective complementation patterns have received considerable attention (e.g. **SuHunston2019**), it would be useful for future research to conduct analyses that involve other patterns (e.g. *there* v-link *sth/anything/nothing* ADJ *about/in* NP). Since this study concerns translation, it would be also interesting for future research to examine whether the observed translation patterns are triggered by English (due to interference), and if so then to what extent, or whether they are natural and equally formulaic in native Polish texts. Since this preliminary study is exploratory in nature, no frequency threshold was set to verify when an equivalent is regarded as a pattern; however, setting such a threshold will be essential in the future.

Finally, Pattern Grammar-based descriptions of Polish may also feed into the future development of more comprehensive lexicographic resources, the so-called 'dictionaries of constructions' or 'constructicons' which are presently being compiled for several European languages, including German (**BoasZiem2018**), Swedish (**LyngfeltEtAl2018b**) and Russian (**JandaEtAl2018**) [16] , to name but a few, and

[16] According to **JandaEtAl2018**, the Russian constructicon prioritizes multi-word constructions.

which are designed to model entire languages as inventories of constructions at all levels, i.e. from morpheme to discourse. A promising example of this synergy of approaches is provided by **PerekPatten2019**, who are currently building an English constructicon by semi-automatically combining GPs (as represented by the COBUILD reference works) with the semantic frames and valency relations found in the FrameNet database. Since constructicons may serve lexicographers, language learners as well as NLP applications, it is postulated that such a resource, which could utilize information already available in valency dictionaries (e.g. *Walenty*), should be also developed in the future for the Polish language[17].

4.1 Acknowledgements

This research has been funded by the Polish National Agency for Academic Exchange (NAWA) under the agreement no: PPN/BEK/2018/1/00081/, and conducted during a research stay of Łukasz Grabowski at the University of Birmingham (UK).

References

Bardovi-Harlig, Kathleen. 2019. Formulaic language in second language pragmatics research. In Anna Siyanova-Chanturia & Ana Pellicer-Sanchez (eds.), *Understanding formulaic language: A second language perspective*, 97–114. Oxford, UK: Routledge.

Boers, Frank & Seth Lindstromberg. 2012. Experimental and intervention studies on formulaic sequences in a second language. *Annual Review of Applied Linguistics* 32. 83–110.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*. 2nd edn. Hillsdale, NJ: Erlbaum.

Craik, Fergus I. M. 2002. Levels of processing: Past, present and future? *Memory* 10(5/6). 305–318.

Dabrowska, Ewa & Elena Lieven. 2005. Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics* 16(3). 437–474.

Dahlmann, Irina. 2009. *Towards a multi-word unit inventory of spoken discourse*. Nottingham, UK: University of Nottingham. (Doctoral dissertation).

Ding, Yanren. 2007. Text memorization and imitation: The practices of successful Chinese learners of English. *System* 35(2). 271–280.

[17]The idea of developing a construction grammar of Polish has also been briefly mentioned recently by **Wierzbicka-Piotrowska2019**.

Ecke, Peter & Christopher J. Hall. 2013. Tracking tip-of-the-tongue states in a multilingual speaker: Evidence of attrition or instability in lexical systems? *International Journal of Bilingualism* 17(6). 734–751.

Erman, Britt. 2007. Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics* 12(1). 25–53.

ETS. 2019. *TOEIC listening and reading test.* https://www.etsglobal.org/fr/en/test-type-family/toeic-listening-and-reading-test) (5 October, 2019).

Fitzpatrick, Tess & Alison Wray. 2006. Breaking up is not so hard to do: Individual differences in L2 memorisation. *Canadian Modern Language Review* 63(1). 35–57.

Fletcher, William. 2011. *Phrases in English website.* http://phrasesinenglish.org (14 January, 2016).

Granger, Sylviane. 2019. Formulaic sequences in learner corpora: Collocations and lexical bundles. In Anna Siyanova-Chanturia & Ana Pellicer-Sanchez (eds.), 228–247. Oxford, UK: Routledge.

Jiang, Nan. 2000. Lexical representation and development in a second language. *Applied Linguistics* 21(1). 47–77.

Kornell, Nate & Kalif E. Vaughn. 2016. How retrieval attempts affect learning: A review and synthesis. *Psychology of Learning and Motivation* 65. 1–30.

Meunier, Fanny. 2012. Formulaic language and language teaching. *Annual Review of Applied Linguistics* 32. 111–129.

Myles, Florence & Caroline Cordier. 2017. Formulaic sequences (FS) cannot be an umbrella term in SLA. *Studies in Second Language Acquisition* 39. 3–28.

Nelson, Thomas O. 1977. Repetition and depth of processing. *Journal of Verbal Learning and Verbal Behavior* 16. 151–171.

Noice, Helga & Tony Noice. 2006. What studies of actors and acting can tell us about memory and cognitive functioning. *Current directions in psychological science* 15(1). 14–18.

Paquot, Magali & Sylviane Granger. 2012. Formulaic language in learner corpora. *Annual Review of Applied Linguistics* 32. 130–149.

Pawley, Andrew & Frances H. Syder. 1983. Two puzzles for linguistic theory: Native-like selection and native-like fluency. In Jack C. Richards & R.W. Schmidt (eds.), *Language and communication*, 191–226. London/New York: Routledge.

Schmitt, Norbert & Ronald Carter. 2004. Formulaic sequences in action: An introduction. In Norbert Schmitt (ed.), *Formulaic sequences: Acquisition, processing and use*, 1–22. Amsterdam, Netherlands: John Benjamins.

Segalowitz, Norman. 2010. *Cognitive bases of second language fluency.* Abingdon: Routledge.

Siyanova-Chanturia, Anna. 2015. On the 'holistic' nature of formulaic language. *Corpus Linguistics and Linguistic Theory* 11(2). 285–301.

Tremblay, Antoine & Harald Baayen. 2010. Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In David Wood (ed.), *Perspectives on formulaic language: Acquisition and communication*, 151–173. London: The Continuum International Publishing Group.

Wray, A. 2002. *Formulaic language and the lexicon*. Cambridge: CUP.

Wray, A. 2008. *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.

Wray, A. & M. R. Perkins. 2000. The functions of formulaic language: An integrated model. *Language and Communication* 20(1). 1–28.

Wray, Alison. 2004. 'Here's one I prepared earlier': Formulaic language learning on television. In Norbert Schmitt (ed.), *Formulaic sequences: Acquisition, processing and use*, 249–268. Amsterdam: John Benjamins.

Wray, Alison. 2019. Concluding question: Why don't second language learners more proactively target formulaic sequences? In Anna Siyanova-Chanturia & Ana Pellicer-Sanchez (eds.), *Understanding formulaic language: A second language perspective*, 248–269. Oxford, UK: Routledge.

Wray, Alison & Tess Fitzpatrick. 2008. Why can't you just leave it alone? Deviations from memorised language as a guide to nativelike competence. In Fanny Meunier & Sylvie Granger (eds.), *Phraseology in foreign language learning and teaching*, 123–148. Philadelphia, PA: John Benjamins.

Yamashita, Junko & Nan Jiang. 2010. L1 influence on the acquisition of l2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *TESOL Quarterly* 44. 647–668.

Yoshimura, Yuki & Brian MacWhinney. 2007. Proceedings of the SLaTE Workshop on Speech and Language Technology in Education. International Speech Communication Association (ISCA). The effect of oral repetition on L2 speech fluency: An experimental tool & language tutor.

Chapter 8

God, the Devil, and Christ: A corpus study of Russian syntactic idioms and their English and Finnish translation correspondences

Mikhail Mikhailov
Tampere University (Finland)

In any language, phrases like *holy Christ/God/cow* can be found. They are sometimes called syntactic idioms, because they are identified in the first place by their syntactic structure and only in the second place by their variable lexical elements. Such expressions are difficult to present in dictionaries, and for this reason they are problematic for language learners. In this paper, the structure, meanings, and use of the Russian construction *Nominative + s 'with' + Instrumental* (*bog s toboj* 'god with you', *čërt s nim* 'the devil with him', etc.) as well as its equivalents in other languages are studied. The construction has four main meanings: 'blessing', 'disagreement', 'permission', and 'acceptance with disapproval'. These meanings are determined by context, and in many cases the expressions are ambiguous. A large web corpus of Russian, ruTenTen11, was used for studying the composition of the construction, its obligatory and optional components, and its functioning in speech. To study the English and Finnish correspondences of the construction, data from parallel corpora of literary texts were used. Parallel concordances demonstrated the absence of direct equivalents for the construction in both English and Finnish. The data also show that this construction is often misunderstood by translators. This phenomenon is obviously connected to insufficient information supplied by monolingual and bilingual dictionaries. The use of the CxG methodology helps to make syntactic idioms more visible and provide better descriptions for them.

1 Introduction

Only the core part of a language consists of free sequences of elements that are combined according to the language's basic rules. The remaining – quite substantial – part consists of so-called "exceptions", for which no clear-cut rules can be suggested. While some rules can be worked out, they are so complicated that it is extremely difficult to use them. This is one of the reasons why learning languages is difficult.

Among those non-free sequences of components are expressions that cannot be interpreted from their constituent parts because of a certain added meaning. Such units are called idiomatic expressions. Many of them are registered at the end of dictionary entries after the basic meanings of the main lexical element are explained. For example, the expression *to kick the bucket* would be probably found at the end of the entry on the noun *bucket*, or, less likely, at the end of the entry on the verb *to kick*.

Some idiomatic expressions pretend to be free expressions. Is the English phrase *How do you do?* idiomatic? Evidently, it is, although it does look like a normal English phrase. A person who says this phrase is not really interested in the health or personal problems of the addressee, neither is the phrase a question. The phrase should be uttered exactly in this form in the situation of being introduced to a person and has the same meaning and function as *Nice to meet you*. Any changes to the phrase (*How are you? How are you doing? ?How did you do*? etc.) may lead to a communicative failure. Hence, there are good reasons to treat the speech formula *How do you do?* as an idiom.

Other borderline cases are combinations of a noun or a verb with a preposition or an adverb, and a good example of this would be English phrasal verbs like *put on*, *show off*, *cut in*, *run out*, etc. Some of these phrases can be registered in dictionaries as idioms, while some are believed to be free expressions. In any case, it is clear that all of them are difficult for non-native speakers and they often cause misinterpretation.

A good example of such a mistake caused by the misunderstanding of an idiomatic expression is a passage from the adaptation of John Wilson's tragedy The City of the Plague (1816) by the Russian poet Alexander Pushkin (*Pir vo vremja čumy* [The feast in the time of the plague], 1832).

Here is the quotation from the original English text:

(1) Priest. O impious table! Spread by impious hands!
Mocking with feast and song and revelry
The silent air of death that hangs above it,

<...>
I could have thought that hell's exulting fiends
With shouts of devilish laughter dragged away
Some harden'd atheist's soul unto perdition.
Several voices. How well he talks of hell! Go on, old boy!

The Russian translation of the last line looks like this:

(2) Несколько голосов. Он мастерски об аде
several. voice..GEN.PL he..3.NOM skillfully.ADV about. hell..LOC
говорит! Ступай, старик! ступай своей дорогой!
talk..3SG. GOIMP 'old_man'..NOM GOIMP own.. way..SG
'Several voices. He skilfully talks about hell. Go away, old man!'

In the original text of the play, the audience mockingly encourages the priest to continue his speech. Pushkin evidently understood *go on* as 'continue on your way' and the reaction of the priest's audience in the Russian translation is the opposite.[1] Pushkin read in the original and translated many English authors – Shakespeare, Byron, Milton – and his translations show a very good understanding of the source text. The error in the translation of Wilson is most likely caused by a lack of knowledge of the spoken language[2] and the possible scarceness of information in the dictionaries of that time.

The modern world is more open, there are more language manuals and dictionaries, the methods of learning languages have improved, and people speak foreign languages much better than in Pushkin's time. Besides, online dictionaries, text corpora, and encyclopaedias make it possible to make very complicated queries. Does this mean that idiomatic expressions do not present problems for learners and translators nowadays?

In any language, there can be found idiomatic expressions that have idiomaticity programmed into their syntactic structure; they are a kind of frame into which variable lexical components can be inserted. For example, there is an English tautological expression *N-Pl will be N-Pl*, which is most often realized as *boys will be boys* (enTenTen15: 878 occurrences, 0.05 ipm[3]), but one can coin other phrases based on that pattern: *men will be men* (enTenTen15: 29 occurrences), *women will*

[1]The matter was discussed on Russian social media in 2019 with many Russian scholars participating, Yakov Testelets and Dmitri Sitchinava among them.

[2]The opinions on Pushkin's command of English are very contradictory; some researchers believe he spoke the language fluently, while others think he could barely read, see **Zaharov2008** for more information.

[3]ipm = instances per million words.

be women (enTenTen15: 3 occurrences), *students will be students* (enTenTen15: 7 occurrences), etc. Such expressions are sometimes called syntactic idioms or phraseoschemes (see, e.g. Baranov & Dobrovol'skij 2008: 16), and usually they are not registered in dictionaries of idioms, partly due to technical issues (e.g. where to place the entry) and partly because of their very complicated semantics. However, such idioms often become topics for linguistic publications, for example **Wierzbicka1987** on *boys will be boys*.

In this paper, I will study the Russian syntactic idiom *N-Nom s 'with'-N/Pron-Instr* (hereafter, I will use a shorter version *N-s-N*, although it is less precise), which can be realized in expressions like *bog s toboj* 'god with you', *čërt s rabotoj* 'devil with work', etc. I will study the idiom with the help of corpus data and describe its structure and meaning using the formalisms of Construction Grammar (CxG, see Fried & Östman 2004). I will check parallel corpora for possible correspondences of this idiom in other languages and ascertain whether translators understand it correctly.

My main sources of data will be ruTenTen11, Russian-English and English-Russian parallel corpora at the Russian National Corpus (RNC) and the Russian-Finnish and Finnish-Russian parallel corpora ParRus and ParFin compiled at Tampere University (**MikhailovHärme2015**, **HärmeMikhailov2016**).

2 The construction *N-s-N*: An overview

Let us start with usage examples from ruTenTen11, a Russian language corpus hosted at SketchEngine (sketchengine.eu).

(3) a. Пожалел псарь хорошенькую девочку и сказал: «Ну и ступай.
pity.SG
Бог с тобой, бедная девочка!»
dog-trainer..NOMSG pretty.ADJ.ACC girl..ACCSG and. say.Past

Participle and.PTCP GOIMP god..NOMSG with. you..2 poor.ADJ.NOMFSG

girl..NOMSG
'The dog-trainer took pity on the pretty girl and he said: Go. God be with you, poor girl!'

b. Путин работает. И Бог с ним!
Putin..NOM work..3SG and.PTCP god..NOM with. he..
'Putin is working, and let him be!'

c. - Наверное, вы правы. Бог с ней, с политикой.
maybe.ADV you..2.NOM.PL right.ADVerbial god..NOM

with. she..3F.SG with. politics.F.SG
'Maybe you are right and we should not discuss politics.'

d. Однако все дела закончить не удастся. И бог с ними!
however.ADV all.

thing..PL.ACC finish. not.PTCP succeed.REFL.FUT.3SG and.PTCP

god..NOM with. it..3.PL.
'However, we won't be able to finish all the jobs, and we need not do them!'

Example (??) is different from the remaining three. In spite of the word *bog* 'god' being the headword of the expression, *bog s X* 'god with X', only this example really relates to piety. The meaning can be described as wishing somebody success, prosperity, or other achievements – later on, I group these meanings together as 'blessing'. In the remaining three examples, the expression has the meaning of acceptance of an inevitable state of affairs the speaker (probably) does not approve of – later on, I refer to this meaning as 'acceptance'. The findings of this chapter will demonstrate that the meaning of 'acceptance' is much more frequent in present-day Russian than the meaning of 'blessing'.

The expression *bog s X* 'god with X' is not unique, and many expressions can be found of a similar structure with different nouns in the initial position.

(4) a. Честное слово, чёрт с ней, с такой работой!
honest.ADJ..NOMSG word...NOMSG devil..NOMSG with. she.F.SG

with. such.. job.F..SG
'I am serious, to hell with such a job!'

b. Правда, я застрял на Ибице, ну и хрен с ним <...>
however.ADV I. stick.SG on.

Ibiza..LOC so.PTCP and.PTCP HORSERADISH..NOM with. it..SG
'Although I am stuck in Ibiza, it is no big deal…'

c. Ладно. Пёс с ними, с высокими идеями.
allright.ADV dog..NOMSG with. it..PL. with.

high.ADJ.PL. idea.F.PL.
'OK, I do not care about these high ideas.'

The construction is flexible, and it has two variable components. The first component should be a noun in the Nominative case, the second is the preposition *s* 'with', and the third can be a noun or a pronoun in the Instrumental case. Additionally, the construction has optional elements. It can be introduced with particles or particle combinations *da*, *i*, *da i*, *nu*, *nu i*, and *nu da i*. If the third component is a pronoun, it can be explicitated (i.e. be made more explicit) with a propositional group headed with preposition *s* 'with' and a noun, sometimes with an attribute, like in examples (**??**) and (**??**).

The expressions are very typical in spoken Russian and rather misleading for non-native speakers, as many colloquialisms are. The meaning often depends on the context and the intonation. The construction is used in the written language as well, and many examples can be found in fiction, mass media, and letter exchange.

The two most frequent of them, *bog s X* and *čërt s X*, are occasionally registered in dictionaries of the Russian language. The Ozhegov-Shvedova Dictionary (**OzhegovShvedova1992**) has both (and even the *pës s X* 'dog with X'), while the Concise Academic Dictionary of Russian (**MAS1984**) has only *čërt s X*. The third popular dictionary of the Russian language, Efremova's Dictionary, does not register any of these idioms (neither does it seem to register syntactic idioms at all).

Russian phraseological dictionaries, even the latest and the most complete Academic Dictionary of Russian phraseology (Baranov & Dobrovol'skij 2015) register only *bog s X* and ignore *čërt s X*.

The bilingual Russian-English Phraseological Dictionary by Sophia **Lubensky1995** is the most accurate with this group of idioms: it registers both *bog s X* and *čërt s X* and mentions that the first element can be replaced by other words, *bog* 'god' > *gospod'* 'Lord', *Hristos* 'Christ', *čërt* 'devil' > *shut* 'clown', *pës* 'dog', *prah* 'ashes', and *hren* 'cock, vulg.'.

In Constructicon for Russian, a repository of Russian constructions (https://spraakbanken.gu
rus), only the construction *čërt s X* is registered with the following definition: "This construction expresses consent with [a participant or situation] Theme imposed on the speaker. The speaker negatively evaluates the participant or situation, and contrary to their will, accepts these conditions". According to Constructicon, the X component can only be a pronoun (which is not true, cf. e.g. a quite

acceptable phrase *čërt s karantinom* 'to devil with the quarantine'). The article does not mention the possibilities of changing *čërt* 'devil' to other nouns, and *bog s X* was not registered, at least at the time this paper was written.

However, in spite of the fact that an average Russian native speaker is very likely to connect the expressions *bog s X*, *čërt s X*, *hren s X*, etc., these relations are not shown in Russian monolingual dictionaries and only partly registered in Lubensky's Russian-English Dictionary.

In linguistic literature, the construction *N-s-N* has not yet been a subject of special study, although Dobrovol'skij et al. (2019: 12) mention in their paper that this construction is productive and deserves a separate study.

Thus, neither dictionaries and lexical databases nor current linguistic research provide a thorough analysis of this syntactic idiom and give a concise picture of its structure, meanings, and functioning in speech. In this publication, I will therefore try to fill this gap.

3 Obtaining the corpus data

As it has already been mentioned, the construction *N-s-N* belongs to language for general purposes, and it is not likely to be found in specialist discourse. It can be used in posts on social media and other informal messages, in mass media texts, and in fiction. Thus, to collect data on this construction, we need a corpus of language for general purposes, and this corpus must be very large, because frequencies of multiword expressions are much lower than frequencies of single words. We also need a concordancing tool with the capacity to make complicated queries to look up syntactic constructions. Currently, the most suitable resource is SketchEngine, which uses its own ruTenTen11, currently the largest corpus of the Russian language available (18.2 G running words). The service permits the download of search results in convenient formats, which was very important for the current study. Therefore, the choice to use SketchEngine and ruTenTen11 was easy.

The service supports the CQL query language, and this makes it possible to run very complicated search queries. However, in this particular case, it was problematic to obtain the data in one step. The problem is that the sequence *Nominative + s + Instrumental* is very common in the Russian language, and searching for it directly would produce an immense amount of noise like *kofe s molokom* 'coffee with milk', *obed s drugom* 'lunch with friend', *kniga s kartinkami* 'book with pictures', etc. Of course, one can always search for particular words in particular forms, but it was necessary to find out first what lexemes serve as the

first component of the construction, the noun in the Nominative case. For this reason, I decided to start with a search on the sequences *Noun.Nominative + s + Personal_pronoun.Instrumental* forming a sentence, i.e. delimited with end-of-sentence punctuation marks. Of course, such a search would not yield all the relevant data, and there might still be some noise in the results (e.g. *obed so mnoj*[4] 'lunch with me' or the above-mentioned *kofe s molokom* 'coffee with milk' as separate sentences). Still, the task of this particular query was not to find all the data with 100% precision, but to get a list of candidates for the headword of our construction.

The first query therefore had the following form:

(5) Query 1.

```
[word="\." | word="!" | word="\?"][wordну=""]?[wordи=""]?[tag="
    N..sn.*"]
[wordco="?"][tag="P....i.*"][word="\." | word="!" | word="\?" |
     word=";"]}
```

I will give here only a very brief explanation of the query: for more details, see the manual of the CQL language on the website of SketchEngine (https://www.sketchengine.eu basics/). Each token of the search phrase is put in square brackets. A full stop means any character. A question mark after any element (token, character) means that it is optional.[5] An asterisk means that the preceding element can be repeated from zero to an indefinite number of times. The tag "word" is used for querying by tokens (running words and punctuation marks), "lemma" by dictionary form, and "tag" by grammatical features. Different tags of the same token can be combined by the logical operators | ("OR"), & ("AND") and ! ("NOT"). So, Query 1 can be read as follows: "a full stop, an exclamation mark, or a question mark – optional particle *nu* – optional particle *i* – a noun in the Nominative singular (the codes for grammatical forms are explained in the tagsets for each language; the Russian tagset can be found here: https://www.sketchengine.eu/russian-tagset/) – a preposition *s* 'with' or its phonetic variant *so* – a pronoun in the Instrumental case – a full stop, an exclamation mark, a question mark, or a semicolon".

[4]The Russian preposition *s* 'with' has a phonetic variant *so* that is used if the next word starts with a combination of consonants, e.g. *so mnoj*, *so stakanom*, *so zvonom*, etc. This variant is included in the search queries of this study, for example in Query 1 below.

[5]To include in a query "real" full stops, question marks, asterisks, and other characters with special meaning, they should be preceded with a backslash ("\.", "\?", "*", etc.).

This query running on a Gigacorpus would have produced a vast concordance that I did not need, so I ordered 10,000 random examples. After loading the concordance into R, separating the first noun into a separate column and creating a frequency list of these nouns, I obtained a table with 1,280 lines. To be on the safe side, I decided to check the whole frequency list, even the single occurrences. As it has been already mentioned, the combination N.NOM+S+Pron. is very common in Russian, and even after restricting the sequence to a separate sentence, many nouns on the list had nothing to do with the construction in question. After removing the noise, the list was dramatically reduced to what can be seen in Table 7.1.

Having the list of headword candidates, it was easy to run the queries to collect all usage examples for the construction *N-s-N* with the words from the list. All the queries run on the second stage of the search were formed like Query 2 below. This particular query looks up the constructions with *bog* 'god' as the headword. The construction does not have to be a separate sentence (commas added to initial and final tokens), and the third element can be a noun or a pronoun.

(6) Query 2.

```
[word="\." | word="!" | word="\?" | word=","][lemmaну=""]?[
    lemmaи=""]?[lemmaбог=""][wordc="" | wordco=""][tag="P....i
    .*" | tag="N...i.*"][word="\." | word="!" | word="\?" |
    word=";" | word=","]
```

The queries for all headwords from the list in Table 7.1 were done by replacing the headword in Query 1 with relevant words: [lemma="бог"] → [lemma="хрен"], [lemma="леший"], etc. In some cases for the words that have different spellings or could have been lemmatized incorrectly, matching with regular expressions was used or a **lemma** tag was replaced with a **word** tag, for example, [lemma="ч[е|о|ё]рт"], [lemma="х.й"], [word="[г|Г]осподь"].

The second search produced a more exact picture, because this time all examples and not a random sample were collected, and all the variants of the construction were looked up.

To check the precision of the search, a random sample of 1,000 examples was generated from the concordance and manually checked. Only 9 examples were wrong and the precision was therefore

(7) 991/1000 x 100= 99.1%.

Table 8.1: The frequency list of the headwords.

Head			Freq
бог	bog	‘god’	2,622
черт	čërt	‘devil’	1,802
хрен	hren	‘horseradish’	907
господь	gospodʹ	‘Lord’	520
фиг	fig	‘fig’	391
шут	šut	‘fool’	105
христос	hristos	‘Christ’	100
хер	her	‘prick’	92
пес	pës	‘dog’	85
аллах	allah	‘Allah’	30
леший	lešij	‘forest imp’	16
хуй	huj	‘prick’	15
дьявол	dʹâvol	‘devil, satan’	13
бес	bes	‘devil’	8
шайтан	šajtan	‘devil for muslims’	5
иисус	iisus	‘Jesus’	4
демон	demon	‘demon’	3
сатана	satana	‘satan’	3
будда	budda	‘buddha’	2
холера	holera	‘cholera’	2
госдеп	gosdep	‘Department of State’	1
зевс	zevs	‘zeus’	1
перун	perun	‘Perun, Slavic god of thunder’	1
фюрер	fjurer	‘führer, Hitler’	1
член	člen	‘organ’	1

Evaluating the recall is more difficult due to the size of the corpus. The search was more or less accurate concerning high-frequency nouns detected with the help of Query 1. However, there might have been a number of low-frequency words used in the construction, and they may not have been detected with the query. Let us assume that there were around 500 examples with the names of people, gods, and mythological creatures that had a low frequency and passed unnoticed. Besides, there are always misspelled words and typos. With an error rate of 5%, about 1,600 headwords or other important components of the construction could have contained a typo, and these contexts would not have been found with the query. Another issue is the parsing accuracy. According to **NivreFang2017**, the accuracy of Russian Universal Dependency parsers is currently on the level of 79.79%. An accuracy of 79.79% for 30,000 examples means about 6,500 examples might have been incorrectly annotated and not found by the query. Thus, the recall of the search would be

(8) (1-(500+1600+6500)/30000)*100 = 71.3%.

The estimation is rough, but it is clear that one cannot expect a high recall rate in a very large and noisy corpus.

The results of the second search are presented in Table 7.2. The words after *šajtan* had very low frequencies and were removed from the table. The absolute (F) and relative (ipm) frequencies are given for each headword, as along with the log-likelihood index (LL) (**Dunning1993**, **Xiao2015**: 111, **Levshina2015**: 223–239). The values of LL are significant for all headwords at the p>0.0001 level.

Grammatical constructions are certain lexemes that occur in speech in certain forms and in a certain order, and therefore collocation searches can provide additional information on the composition and use of constructions. Collocation searches on large Russian language corpora can be performed with the online collocator CoCoCo (http://cococo.cosyco.ru/, **KopotevEtAl2016**, **Kormacheva2020**). Unfortunately, grammatical searches are not available in the current version,[6] and therefore one can only submit queries on the concrete lexical realizations of constructions. I tested the headwords from Table 7.2 in combination with the word *s* 'with' (*bog + s*, *fig + s*, etc.) and no third collocation could be found for the words *gospod'*, *šut*, *hren*, *allah*, *lešij*, *bes*, and *šajtan*. For the remaining headwords, only pronoun collocations were found. The search for the noun preceding the phrases *s nim*, *s nej*, and *s nimi* yielded the collocates *bog*, *fig*, *hren*, *čërt*, and *shut*. The CoCoCo service performs searches on three Russian corpora: Taiga,

[6]To be more precise, it is possible to compose a query with a grammatical form and no lexeme, but the search does not work.

Table 8.2: Headwords of the construction N-s-N: ruTenTen11.

Word	Translit	Meaning	F	ipm	Connotation	LL
бог	bog	'god'	10706	0.586	pos	7160.01
черт	čërt	'devil'	6981	0.382	neg	5694.86
хрен	hren	'horse-radish'	4054	0.222	neg	6173.16
фиг	fig	'fig'	2585	0.141	neg	3158.73
господь	gospodʹ	'Lord'	164	0.09	pos	8029.45
шут	šut	'fool'	1087	0.059	neg	1278.75
хуй	huj	'prick'	923	0.05	neg	511.64
пес	pës	'dog'	907	0.05	neg	76.94
хер	her	'prick'	473	0.026	neg	456.73
христос	hristos	'Christ'	304	0.017	pos	8190.28
аллах	allah	'Allah'	123	0.007	neg	2826.67
леший	lešij	'forest imp'	85	0.005	neg	58.79
дьявол	d'âvol	'devil, satan'	81	0.004	neg	1459.43
бес	bes	'devil'	37	0.002	neg	1011.68
шайтан	šajtan	'devil for muslims'	33	0.002	neg	26.24

the Russian National Corpus, and I-ru. Taiga (**ShavrinaShapovalova2017**) is the largest corpus and the only one suitable for our searches (and even there was not enough data for some words). This shows that for studying syntactic idioms, one needs very large data sets, and the existing manually collected corpora are too small. Evidently, this is the reason why only part of my findings was confirmed with CoCoCo. Sadly, webcorpora like ruTenTen11 are also problematic in terms of data quality.

The total number of examples collected with Query 2 was 30,019 and the relative frequency of the construction was 1.64 ipm. This is a low frequency, e.g. the Frequency Dictionary of Russian by **LjashevskajaSharov2009** includes 20,000 words with a relative frequency higher than 2.6 ipm. An additional problem is that unlike lexemes, constructions cannot be detected by means of tokenization and lemmatization.

After studying Table 7.2, we can define the semantic restrictions for the head of the *N-s-N* construction: 'divine force', 'dark force', or 'masculine sexual organ' (obviously serving as a euphemism for dark forces, other swear words will not

work in this construction).

In most cases, the collected examples have only obligatory components without optional particles at the beginning and the nominal group at the end. Still, 11,645 examples have emphasizing particles in the initial position of the construction: *da* (2,634), *da i* (1,948), *da nu i* (**??**), *i* (2,936), *nu* (**??**), and *nu i* (3,582). About 23% of the examples (6,897) have the optional nominal group with explicitation of the pronoun of the construction's third obligatory element.

It is impossible to analyse in detail the usage examples that were collected, as the size of the concordance was more than 30,000 items. Still, some of the most typical meanings could be found by studying random examples and collocations. The number of meanings has grown from the two meanings detected at the beginning of section 2 of this paper to the following four meanings:

- Blessing: X gives Y a blessing to perform Z (headwords: *bog* 'god', *gospod'* 'Lord', *Hristos* 'Christ')
- Disagreement, disbelief, surprise: X disagrees with Y / does not believe Y / is surprised with what Y says (headword: *bog* 'god', *gospod'* 'Lord', *Hristos* 'Christ')
- Permission: X allows Y to perform Z (headwords: *bog* 'god', *gospod'* 'Lord', *Hristos* 'Christ')
- Acceptance with disapproval: X is reluctant that Y is planning Z, but cannot prevent it (headwords: *fig* 'fig', *hren* 'horseradish', *her* 'prick')

The meanings can be connected: on the one hand, a positive attitude to Z (blessing, permission, see example (**??**)) can gradually turn into a negative (acceptance with disapproval, see example (**??**)). On the other hand, a blessing can transform into a disagreement (see example (**??**)).

One might think that blessing and disagreement have nothing in common. However, disagreement can be expressed by blessing the existence of another point of view and in this way express that the speaker's point of view differs from that of the interlocutor. The nouns in the first position should be *bog/gospod'/Hristos*. The "evil forces" are not fit for expressing respect that is necessary for this meaning. The third component must be the second-person pronoun *ty/vy*. An exclamation mark is very typical for such contexts.

(9) a. — Ну, какая коррупция в Англии, да Бог с вами!
well.PTCP what.

corruption.F.NOM in. England..LOC and.PTCP god..NOM with. you..2..PL
'What corruption in England are you talking of? I don't believe you.'

b. Что вы, господь с вами! это не он.
what. you. Lord..NOM with. you...PL this. not.PTCP he..3.NOM
'What are you talking about, you are wrong! It is not he!'

c. — Что ты!
what. you..2SG
Христос с тобой! — воскликнул я, несколько испуганный.
Christ..NOM with.

you..2.SG shout.SG I..NOM slightly.ADV frighten.PTCP.PASSSG.NOM
'What are you talking about! You are wrong!' I shouted, a little frightened.'

The connections of the meanings are shown schematically in Figure 8.1.

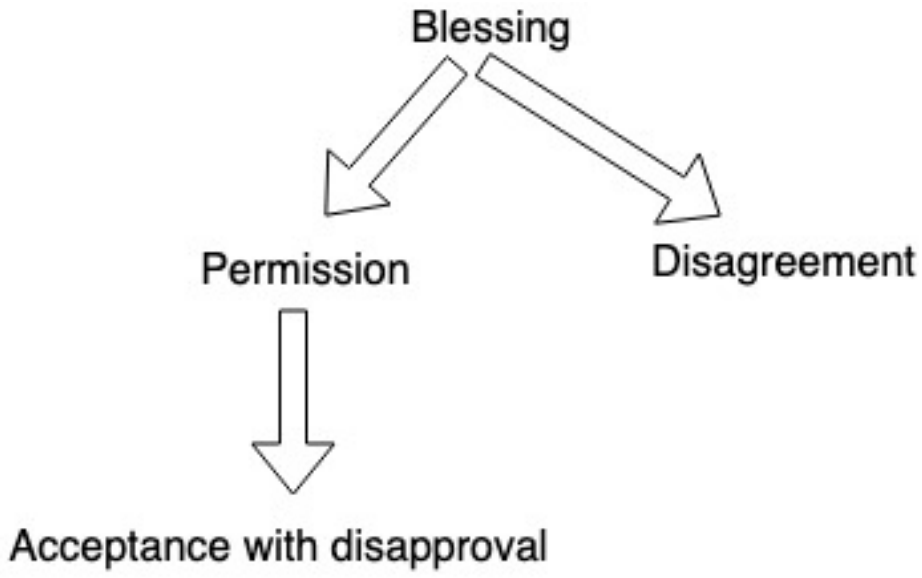


Figure 8.1: The meanings of the construction N-s-N.

Although *bog* ‘god’ occupies the first row of Table 7.2 and is almost twice as frequent as *čërt* ‘devil’, the words with negative connotations clearly dominate the list, and the sums of frequencies of the expressions with the negatively coloured headwords outmatch the positively coloured. The result is 17,369 versus 12,650, which makes 58% against 42%. Even in cases where the headword is positively coloured, there might be contexts with negative connotations (e.g. example (**??**)).

To sum up, the meanings of the construction are interrelated and have many borderline cases. For this reason, it is practical not to treat them as separate homonymous constructions, but rather as a single construction.

4 Constructing the construction

To present the *N-s-N* construction as a whole, I will take advantage of the box notation used in Construction Grammar (CxG). This notation is “a convenient way of organizing all the information needed to give an adequate account of linguistic structure” (Fried & Östman 2004: 13). The result of summing up the findings from the concordances is presented in the following box diagrams (Fig. 2–4).

In section 3, the existence of semantic and structural variation in the research data was demonstrated. The easiest way to handle this heterogeneity is to define three variants of the construction *N-s-N*. Still, it is better to treat them as variants of the same construction rather than as independent constructions. The first variant (*N-s-N_a*, Fig. 2) covers the meaning ‘blessing’; the second one (*N-s-N_b*, Fig. 3) has the meaning ‘disagreement’, ‘surprise’, or ‘disbelief’; and the last one (*N-s-N_c*, Fig. 4) handles the remaining meanings.

The construction *N-s-N_a* (Fig. 2) is the simplest. The choice of the first noun is limited to three: *bog* ‘god’, *gospod’* ‘Lord’, and *Hristos* ‘Christ’. The second nominal component is always a second-person pronoun. This pronoun can be explicitated by the optional noun phrase with the noun in the Nominative (or, rather, Vocative) case. The beneficiary of this construction is always a person (or a personified animal/artefact, etc).

(10) Здоровья Вам и Вашим близким,
health..GEN you..2.DAT.PL and. you..POSS.DAT.PL relative..DAT.PL,
терпения! Бог с Вами!
patience..GEN! God..NOM with. you.2..PL
‘I wish you and your relatives health and patience. God be with you!’

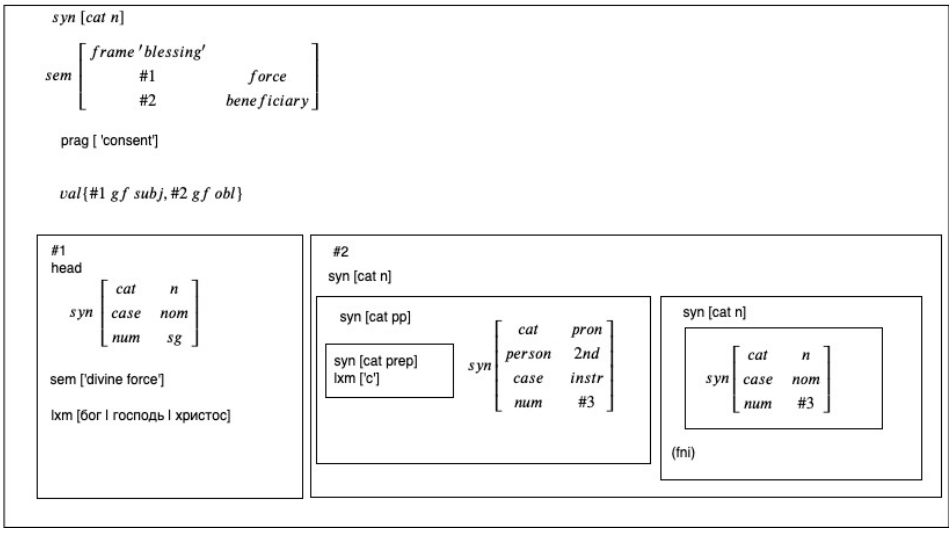


Figure 8.2: The representation of the N-s-N_a construction

(11) Прощай, Оля, господь с тобой.
farewell.ADV Olja..NOM, Lord..NOM with. you..2.SG
'Farewell Olja, the Lord be with you.'

(12) Прощайте! Христос с вами!
farewell.ADV! Christ..NOM with. you..2..PL
'Farewell! Christ be with you!'

The variant *N-s-N_b* (Fig. 3) looks very similar to the previous one and can be easily confused with it, as we will see in section 5 of this chapter. When spoken, the intonation of this variant is different from *N-s-N_a*, with a phrasal stress on the first nominal element; graphically it may be expressed with the exclamation mark. Besides, there is a structural difference: an optional particle *nu* or *da* in the beginning.

(13) a. i. xnumiv. The construction *N-s-N_c* (Fig. 4) gives more freedom to choose the first nominal element. Any noun from the list of Table 7.2 can be used, including the three nouns used in the *N-s-N_a* and *N-s-N_b* variants, and the list is open and other nouns with the semantics of 'superhuman force' can be used (see section 3 for details). The second nominal element can be

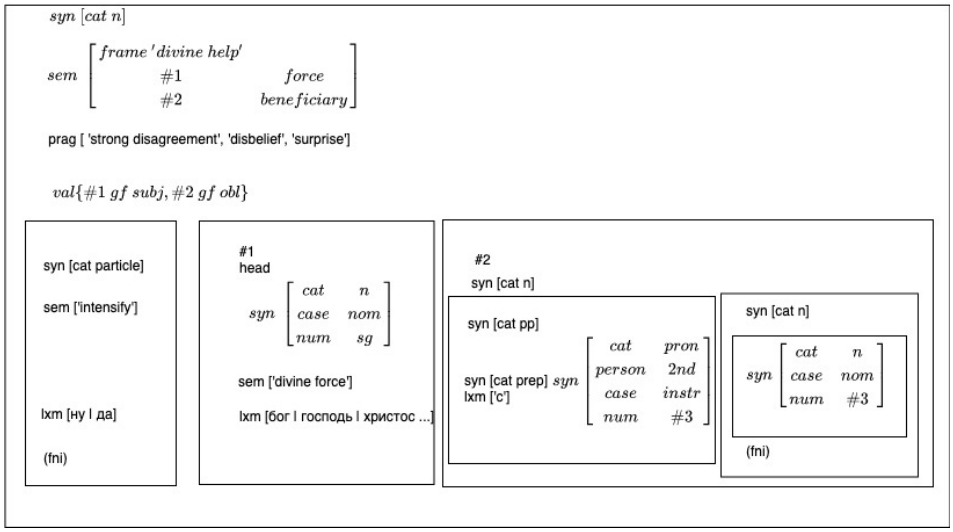


Figure 8.3: The representation of the N-s-N_b construction.

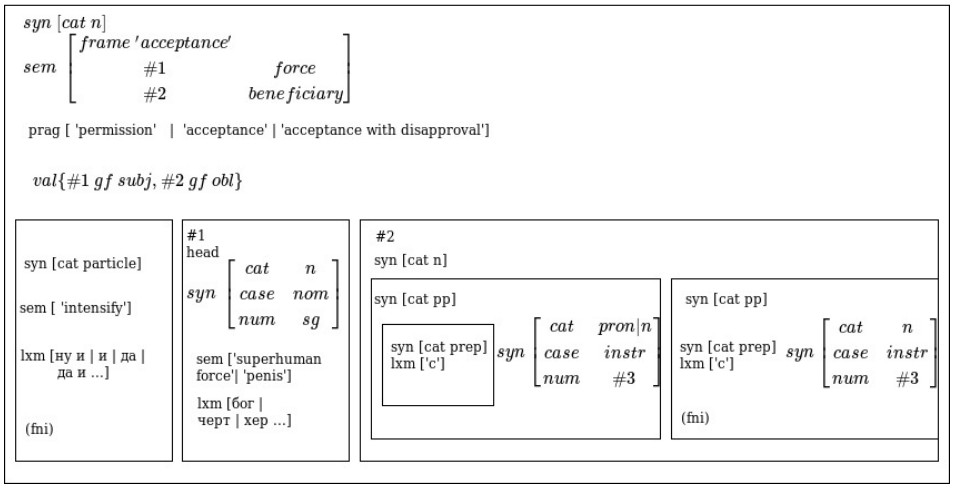


Figure 8.4: fig:mikhailov:4 The representation of the N-s-N_c construction.

a noun or a pronoun, and there are no semantic restrictions: it can be a person, a thing, an activity, a situation, etc.

This variant can have an optional initial element: a particle or a combination of particles that work as an intensifier. The palette is richer than in *N-s-N_b*, which has only two options. At least the following combinations are used quite frequently: *da, i*, *da i*, *nu i*, and *nu da i*. The most frequently used is the combination *nu i* (3,582 examples in the concordance). The expression *nu i* is used in other contexts as well (e.g. *Nu i durak* 'what a fool you are'), and Dobrovol'skij et al. 2019 claim that it is a separate construction.

xnumiv. Диска с ПО, естественно,
disk..GENSG with. software._Abr, "of_course".ADV,
тоже никакого нет, ну и
also.PTCP none..GENSG no.Pred, well.PTCP and.PTCP
аллах с ним.
Allah..NOM with. he..3.SG
'Of course there is no software included, well, I don't care.'

xnumiv. Да бес с ними, с
and.PTCP devil..NOM..NOMSG with. he...3.PL with.
британцами.
brit...PL
'I don't care about the brits.'

capitalize 'Brit' (2x)?

xnumiv. Радуйтесь, и Господь с вами.
enjoyIMP.PL and.PTCP Lord..NOM with. you..2..PL
'Be happy and the Lord be with you.'

Another optional component is the prepositional phrase, which can be used for the explicitation of the second nominal element if the latter is a pronoun. Unlike the construction N-s-N_a, this element is not in the Nominative case, but it repeats the structure of the second element: preposition s 'with' + noun phrase in the Instrumental case.

This prepositional phrase can be a combination of a preposition with a single noun (**??**), but it can have quite a complicated structure (9b and 9c).

xnumiv. Черт с ним, с народом!
devil..NOM with. he..SG, with. people..SG
'I don't care about the people!'

xnumiv. Бог с ней, с этой конкретной
god..NOM with. she..3F.SG with. this..SG concrete.ADJ.SG
передачей.
programme.F.SG
'I don't care about this particular programme.'

xnumiv. Впрочем, аллах с ними, с
anyway.ADV, Allah..NOM with. it...3.PL with.
нашими смешными читательскими проблемами!
our..POSS..PL funny.ADJ..PL reader.ADJ..PL problem.F..PL
'Anyway I do not care about our funny readers' problems!'

In rare cases, the use of two pronouns is possible, like in (??), where the speaker is the beneficiary.

xnumiv. Бог с ним, со мной.
god..NOM with. he..3.SG with. I..1.SG
'I do not care about myself.'

If the context is limited, the construction may become ambiguous, as in example (??), which can be interpreted as a blessing, as an acceptance, and even as disagreement.

xnumiv. – Друже Чумак, <…>
friend..VOC Chumak..NOM I..NOMSG no.PTCP
я не требую никаких
demand..1SG no..NEG.GEN.PL explanation..GEN.PL. god..NOM
объяснений. Бог с вами.
with. you...PL
'Friend Chumak, I am not demanding any explanations. God be with you / I do not care / Not at all.'

The most frequently used is the construction N-s-N_c. A check of the same random sample of 1,000 examples that was used for calculating the precision of the search (see section 2) confirmed this: among the 991 correct examples, only 49 (4.9%) belonged to N-s-N_a. The construction N-s-N_b occurred about the same number of times, in 47 examples (4.7%), and all the remaining 895 (90.3%) examples belonged to N-s-N_c.

5 The challenges of parallel concordancing

To check the equivalents that are used when translating contexts containing a certain construction, one needs many examples from parallel texts. This becomes a problem when studying multiword expressions, because their frequencies are low and therefore large amounts of text are needed to get enough examples. As it has already been mentioned, the best source of data for studying idiomatic expressions are fiction and mass media texts. Such texts are available in parallel corpora, but the sizes of parallel corpora of literary texts are quite modest compared to gigaword monolingual corpora. The data I used for this study were as follows:

1. Parallel corpora at the Russian National Corpus (RNC)
- Russian-English subcorpus (6.5 M running words)
- English-Russian subcorpus (18 M running words)

1. Parallel corpora at Tampere University
- ParRus, the Russian-Finnish corpus of fiction texts (6 M running words)
- ParFin, the Finnish-Russian corpus of fiction texts (3 M running words)

It is obvious that the amounts of data from these parallel corpora are microscopic in comparison with ruTenTen11. Besides, the Russian-English subcorpus of the RNC is not well-balanced: works by Vladimir Nabokov clearly dominate over all other authors and periods. However, there were no other data available. Parallel corpora at SketchEngine are larger, but their composition is unclear, and it is impossible to filter out indirect translations and pseudotranslations. Hence, our data will be suitable only for detecting general tendencies for some of the expressions.

It is easy to observe in Table 7.3 that the normalized frequencies of headwords are much higher than in ruTenTen11, although not all expressions were found (only seven of fifteen). This can be explained by the structure of ruTenTen11, which contains many genres in which the construction *N-s-N* is never used. The causes of the differences in frequencies between parallel corpora are the corpora's imbalance and their construction from whole texts, so a couple of very long texts could skew the whole collection.

Table 8.3: Frequencies of the headwords N-s-N construction in the parallel corpora.

Word			RuEnF	ipm	EnRuF	ipm	RuFiF	ipm	FiRuF	ipm
бог	bog	‘god’	65	9.86	23	1.27	73	23.09	9	5.03
господь	gospod	‘Lord’	8	1.21	13	0.72	0	0	0	0
пес	pes	‘dog’	1	0.15	0	0	0	0	0	0
Христос	Hristos	‘Christ’	10	1.52	0	0	0	0	0	0
черт	čert	‘devil’	42	6.37	28	1.55	52	16.44	6	3.35
шут	šut	‘clown’	2	0.30	1	0.06	0	0	0	0
хрен	hren	‘horseradish’	0	9	0.50	0	0	0	0	

The comparison of the frequencies of *N-s-N* in ruTenTen11 and the parallel corpora demonstrates that the frequencies of expressions are much less stable than that of single words, and it is problematic to obtain reliable statistics from the observations. For example, the frequency of the expression *bog s X* is 9.8 ipm in Russian-English RNC and 23.1 ipm in ParRus, although both corpora are collections of Russian fiction texts.

Regardless, one important observation can be made from the frequencies: the construction *N-s-N* is much more frequent in the original Russian texts than in the translations from English and Finnish into Russian. This is the sign of the evident absence of matching constructions in both English and Finnish. The findings are also in line with **Tirkkonen-Condit2004**’s (**Tirkkonen-Condit2004**) hypothesis about the underrepresentation of unique items of the source language in translated language.

The statistics from the parallel concordances give the impression that something is not right. As it was shown in the previous sections, the construction *N-s-N* is polysemous, and the actual meaning depends on the context. The most misleading is the construction with *bog* ‘god’ as a headword: it can be used in all three variants of the construction described in section 4 of this paper. The variant *N-s-N_a* is not very frequent: I demonstrated this by the study of random examples. Still, in the

Russian-English data, 28 contexts out of 65 were translated into English with expressions containing the word *god*. In the Russian-Finnish data, there are 73 contexts with *bog* 'god', and 48 of them are translated with the expressions containing *jumala* 'god', *herra* 'Lord', or *luoja* 'Creator'. From the above-mentioned study of random examples, I would have expected that only about 7% of contexts of *bog s X* would belong to the *N-s-N_a* variant, while the statistics from the parallel corpora show a much higher rate in both the Russian-English and Russian-Finnish data.

It is true that the data are not balanced, and it is true that the frequencies of the expressions in our data vary greatly. It is therefore quite possible that the data from the parallel corpora might contain far more *N-s-N_a* contexts than the ruTenTen11 data. For this reason, it is necessary to check the actual contexts to confirm the statistical observations.

The checking of the Russian-English concordance with *bog* 'god' on the Russian side and *god* on the English side confirmed my suspicions: 19 cases out of 28 show an obvious misunderstanding of the source text.

xnumiv. «
"well.PTCP
Ну, бог с тобой, оставайся уж», — решила в тоске Грушенька, состра
god..NOM

with. you..2.SG stayIMP well.PTCP", - decide.FSG in.

melancholy..LOCSG Grushenka..NOM, compassionately.ADV

he..DATSG smile.

'OK, I don't care, you can stay, decided Grushenka in her melancholy and smiled at him compassionately.'

"Well, God bless you, you'd better stay, then," Grushenka decided in her grief, smiling compassionately at him.

F. Dostoevsky. Brat'â Karamazovy [The Brothers Karamazov] (1878, transl. C. Garnett, 1912)

In example (**??**), the speaker reluctantly gives the interlocutor her permission to stay, while the translator obviously understood the

expression as a blessing or at least as a demonstration of piety (which is strange for Grushenka, who, as we know, was not a very pious person).

In the Russian-Finnish data, 44 contexts with an obvious misunderstanding were found. An additional factor for misinterpreting is Russian-Finnish dictionaries, some of which register the phrase *bog s X* only with the meaning of blessing (see, e.g. **KuusinenOllikainen1984**).

xnumiv. Господин Разумихин не то-с, да
mister..NOMSG Razumihin..NOM not.PTCP that., and.PTCP
и человек посторонний, прибежал
and.Patricle person..NOMSG stranger.ADJ.NOMSG, run.SG
ко мне весь такой бледный ...
to. I..1.DAT all..NOM such.ADJ.NOMSG pale.ADJ.NOMSG. but.PTCP
Ну да бог с ним, что его
and.PTCP god..NOM with. he..3.SG, what. he..3.ACCSG here.ADV
сюда мешать.
involve.

'Mister Razumihin is a stranger, but he ran to me so pale. Never mind, why shall we involve him in this.'

xnumiv. Herra Razumihinhan on vallan
mister..NOM Razumihin..NOM be. power..GEN
toista maata, sivullinen ihminen,
another.ADJ. country.. stranger.ADJ.NOM person..NOM
vaikka hän juoksi silloin kasvot kalpeina
although. he..NOM run. then.ADV face..NOM white.ADJ.
luokseni ... Luoja hänen kanssaan, eihän
"to me".ADV. Lord..NOM he..GEN with. not.PTCP
hänellä ole tässä osaa eikä arpaa.
he..3.ALL be. here.ADV. part.. not.PTCP lot..SG

'Mister Razumihin is like from another country, a stranger, still he ran to me with a white face. God be with him, he has nothing to do in this business.'

F. Dostoevsky. Prestuplenie i nakazanie [Crime and Punishment] (1866, transl. J. Konkka, 1970)

The expression *čërt s X* 'devil with X' also contains a trap: it can be interpreted as swearing and blasphemy, although in most

cases it does not and fits quite well into the *N-s-N_c* construction.

xnumiv. Об чем? Ну, да черт с
about. what..LOC well.PTCP and.PTCP devil..NOM with.
тобой, пожалуй, не сказывай.
you..SG maybe.ADV not.PTCP tellIMPSG
'What about? Well, do not tell, I don't mind.'

xnumiv. What about? Confound you, don't tell me then.
F. Dostoevsky. Prestuplenie i nakazanie [Crime and Punishment] (1866, transl. C. Garnett, 1914)

One might think that such things take place only in very old translations of even older source texts. However, this is not so: in (??) is an example of a relatively recently published translation from Russian into Finnish.

xnumiv. Черт с ним! – сердито
devil..NOM with. he..SG angrily.ADV think.FSG
подумала Вероника.
Veronika..NOM.
'I don't care, thought Veronika angrily.'

xnumiv. "Hitto", Veronika mietti vihaisena.
devil..NOM, Veronika..NOM think..3SG angry.ADJ.SG
'Devil, thought Veronika angrily.'

xnumiv. A. Marinina. Za vse nado platit′ [You have to pay for everything] (1995, transl. O. Kuukasjärvi, 2005)

It should be mentioned that the parallel concordance also provided enough examples with interesting solutions for this construction. I will give here only two examples from the Russian-English data. In (??) an English expression *all right* is used, while in (??) the meaning of expression is explicitated (*I will take it*).

xnumiv. Ну что ж, бог с вами,
well.PTCP what. but.PTCP, god..NOM with. you..2..PL,
пусть пять рублей будет.
let.PTCP five.um.NOM rouble..GEN.PL be.V.FUTure.3SG.
Только деньги попрошу вперед. "OK, let it
only.ADV money..ACC.PL ask.V.FUT.1SG forward.ADV

be five roubles, but I would like to have the money in advance."
"Well, all right, make it five roubles. Only I want the money in advance, please."

Ilya Ilf, Evgeny Petrov. Двенадцать стульев (Dvenadcat′ stul′ev) [The Twelve Chairs] (1927, transl. John Richardson, 1961).

xnumiv. Ну, бог с вами, — сказал
well.PTCP god..NOM with. you..2..PL say. Mahin..NOM,
Махин, кладя на витрину купон. "'OK, I
put. on. counter..ACCSG coupon..ACCSG
agree,' said Mahin putting the coupon on the counter".
"Well, I will take it," said Mahin, and put the coupon on the counter.

Leo Tolstoy. Fal′šivyj kupon [The Forged Coupon] (1889–1904, transl. Louise and Aylmer Maude, 1911)

To sum up the findings from the parallel concordances, the main problem of the data obtained from translations from Russian into other languages is the possibility of misunderstanding the source texts by translators. Hence, translations from other languages into Russian quite unexpectedly become a very useful source of reference data. Translators into Russian write in their native language and their work is addressed to other native speakers of Russian. As a result, the expression that served as a stimulus for the Russian expression may be with a few reservations used as an equivalent for translating in the opposite direction. Of course, in this case there is an issue of the correct understanding of the source text in language X.

The RNC's English-Russian subcorpus is larger and richer than the Russian-English one. In spite of this, the construction *N-s-N* features in it much less frequently (see Table 7.3). Still, the parallel concordance produces some interesting solutions that seem suitable for translating from Russian into English as well.

xnumiv. I've been told I ought to have a salon, whatever that may be. Never mind. Go on, Badger.
Мне частенько говорили, что мне надо
I..1.DATSG often.ADV say..PL that. I..1.DATSG need.Pred

бы завести салон, что бы это
would.PTCP start. salon..ACCSG, what. would.PTCP this.
там ни значило. Ну, бог с
there.ADV not.PTCP mean.. well.PTCP god..NOMSG with.
ним! Продолжай, Барсук.
he..3.SG. continueIMP, Badger..NOM.
'They often said to me that I should start a salon, whatever it may mean. Continue, Badger.'

xnumiv. Kenneth Grahame. The Wind in the Willows, (1908, transl. I. Tokmakova, 1988)

xnumiv. "You still have half your balls there." "I don't care. This will set my game back a month."
– У тебя еще осталась половина мячей. –
by. you..2.GENSG still.PTCP
И черт с ним. Это отбросит мою технику на месяц назад.
remain.FSG

half..NOMSG ball..GEN.PL. and.PTCP devil..NOM with. he..3.SG.

It. throw.V.FUTure.3SG my..POSSF.ACCSG technique..ACCSG on.

month..ACCSG back.ADV
'You still have half of the balls. I don't care. It will throw my technique a month back.'

xnumiv. Michael Connelly. City Of Bones (2002, transl. D. Vozniakevitch, 2006)

The same can be observed in the Finnish-Russian parallel concordance obtained from the ParFin corpus.

xnumiv. Lukeneilla ihmisillä sellanen
study.PTCP.ALL.PL person..ALL.PL such.ADJ.NOMSG
on ja hyvä niin.
be.V.3SG and. good.ADJ.NOMSG so.ADV
'Educated people have this and this is good.'

xnumiv. У тех, кто учился, есть, и
by. he..GEN.PL, who..NOM study.SG, be..3, and.PTCP

бог с ними.
god..NOM with. he...PL.
'Those who studied have it and let it be'
Hotakainen, Kari, Juoksuhaudantie (transl. I. Uretski)

Strangely, although the English stimuli *never mind* and *I don't care*, as well as the Finnish stimulus *hyvä niin* 'OK' can be considered as very good variants for conveying the meaning of the Russian construction *N-s-N_c*, they are not very typical for translations from Russian. The expression *never mind* occurs only 7 times in the Russian-English concordance, and the verb *care* only three times. In the Russian-Finnish parallel concordance, there is not a single example of *hyvä niin* used as an equivalent for *N-s-N*.

6 Discussion

The case study performed in this paper demonstrates the usefulness of monolingual and parallel corpora for studying constructions. Corpora provide information on the variability of constructions and statistics. Monolingual concordancing is helpful in the study of the components of the construction, the lexemes used for its realization, and even semantic issues. The analysis reveals that the construction *N-s-N* can be implemented in the form of ready-made phrases (like *bog s nim, čërt s nim*, etc.) that are used very frequently, as well as in the form of *hapax legomena* constructed with the same template. As a result, some phrases may be registered in dictionaries, while occasionalisms remain both outside dictionaries due to their rarity and outside grammar descriptions due to their specificity. Evidently, the best way of describing and storing such units would be databases like FrameNet or Constructicon.

To study the links of the construction with other languages, parallel corpora were used. However, the usability of this resource was limited. Parallel corpora did not help so much in looking up translation equivalents as one might have expected. The first reason was that the search did not return enough usage examples; one would have needed much larger data collections to

obtain a parallel concordance at least comparable with the monolingual concordance from ruTenTen11. The data that were available were sufficient only for demonstrating the fact that the *N-s-N* construction in Russian does not have corresponding constructions in English or Finnish, and that this absence causes difficulties for translators.

The second reason was the rather high rate of errors in the translations. Of course, one might expect errors in any language data – this is quite natural – but in this case the errors were repeated, and their main cause was misinterpretation of the source text. On the one hand, this is a challenge to modern statistical and neural machine translation technologies, which are based on parallel corpora and use human translations for modelling MT. The developers of MT presume that there might be errors and mistakes in the data, but are they ready for errors on such a scale? On the other hand, this is a challenge to the belief that the translation of a literary work into another language is the *same* story told in other words. The real data show that literary translators sometimes do not understand the source text well enough.

Why does this happen? The first priority of a literary translator is to produce a good target text, one that meets the standards of a literary text. The correspondence of the translation to the source text comes second, and it is not likely that every passage of the translation is compared to the original text. Of course, the translation should not be very different, but how correct should it be? There is also some evidence that the literary translators' command of the source language is not as advanced as one might expect. For example, Nikolai Chukovski, one of the leading Russian literary translators working from the 1920s to the 1960s, was very critical of his own proficiency in English (**ChukovskiChukovski2004**), and there existed writers (and especially poets) who "translated" by editing earlier translations or literal translations produced by other people (see, e.g. **Kamovnikova2019**).

These issues make the use of parallel corpora of literary texts a specific resource. They cannot be, for example, the main source of data for bilingual dictionaries, but rather reference data for rechecking translation equivalents. Parallel corpora also

demonstrate that even nowadays, proficiency in non-native languages is limited and needs to be improved. The data from parallel corpora might be of great help in finding such weak points.

References

Biber, D. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14. 275–311.

Buerki, A. 2016. *Formulaic sequences: a drop in the ocean of constructions or something more significant? European Journal of English Studies* 20(1). 15–34.

Caines, A. & P. Buttery. 2017. The effect of task and topic on opportunity of use in learner corpora. In V. Brezina & L. Flowerdew (eds.), *Learner corpus research: New perspectives and applications*, 5–27.

Callies, M. 2015. Learner corpus methodology. In S. Granger, G. Gilquin & F. Meunier (eds.), *The Cambridge handbook of learner corpus research*, 35–55.

Chen, Y. H. & P. Baker. 2010. Lexical bundles in l1 and l2 academic writing. *Language Learning and Technology* 14(2). 30–49.

Colson, J. P. 2016. Set phrases around GLOBALIZATION: An experiment in corpus-based computational phraseology. In F. Alonso Almeida, I. Ortega Barrera, E. Quintana Toledo & M. E. Sánchez Cuervo (eds.), *Input a word, analyze the world. Selected approaches to corpus linguistics, pp*, 141–152. Newcastle: Cambridge Scholars Publishing.

Cortes, V. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes* 23. 397–423.

Durrant, P. 2017. Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics* 38(2). 165–193.

Gilquin, G. & M. Paquot. 2008. Too chatty: Learner academic writing and register variation. *English Text Construction* 1(1). 41–61.

Gledhill, C. 2011. The "lexicogrammar" approach to analysing phraseology and collocation in ESP texts. *ASP. la revue du GERAS* 59. 5–23.

Granger, S. & P. Rayson. 1998. Automatic profiling of learner texts. In S. Granger (ed.), *Learner English on computer*, 119–131. London: Longman.

Halliday, M. A.K. 2014. *Halliday's introduction to functional grammar (4th ed.)* Revised by C. M.I. M. Matthiessen. Oxen: Routledge.

Herbst, T. 2015. Why construction grammar catches the worm and corpus data can drive you crazy: Accounting for idiomatic and non-idiomatic idiomaticity. *Journal of Social Sciences* 11(3). 91–110.

Hyland, K. 2016. Academic publishing and the myth of linguistic injustice. *Journal of Second Language Writing* 31. 58–69.

Hyland, K. 2008a. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27. 4–21.

Kecskes, I. 2016. Deliberate creativity and formulaic language use. In K. Allan, A. Capone & I. Kecskes (eds.), *Pragmemes and theories of language use, perspectives in pragmatics, philosophy & psychology 9, pp*, 3–20. Cham, Switzerland: Springer International Publishing.

Martinez, R. & N. Schmitt. 2012. A phrasal expression list. *Applied Linguistics* 33(3). 299–320.

Nesi, H. & S. Gardner. 2012. *Genres across the disciplines: Student writing in higher education*. Cambridge: Cambridge University Press.

Pérez-Llantada, C. 2014. Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes* 14. 84–94.

Römer, U. 2019. A corpus perspective on the development of verb constructions in second language learners. *International Journal of Corpus Linguistics* 24(3). 268–290.

Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Siyanova-Chanturia, A. 2013. Eye-tracking and ERPs in multi-word expression research: A state-of-the-art review of the method and findings. *The Mental Lexicon* 8(2). 245–268.

Soler, J. & Y. Wang. 2019. What gets published in predatory journals: A corpus-based comparison of two journals in political science. *Learned Publishing* 32. 259–269.
Wang, Y. 2018. As hill seems to suggest: Variability in formulaic sequences with interpersonal functions in l1 novice and expert academic writing. *Journal of English for Academic Purposes* 33. 12–23.
Wang, Y. 2019. A functional analysis of text-oriented formulaic expressions in written academic discourse: Multiword sequences vs single words. *English for Specific Purposes* 54. 50–61.
Wray, A. 2002. *Formulaic language and the lexicon.* Cambridge: CUP.
Wray, A. 2008. *Formulaic language: Pushing the boundaries.* Oxford: Oxford University Press.
Wray, A. & M. R. Perkins. 2000. The functions of formulaic language: An integrated model. *Language and Communication* 20(1). 1–28.