Andreas Buerki

(How) is Formulaic Language Universal? Insights from Korean, German and English

Abstract: Items of formulaic language, also referred to as phraseological units or common turns of phrase, are in evidence in a very large number of languages. However, the extent to which languages feature such formulaic material is unclear. Similarly, how formulaicity may be understood across typologically different languages and whether indeed there is a concept of formulaic language that applies across languages, are questions which have not generally been discussed. Using a novel data set consisting of topically matched corpora in three typologically different languages (Korean, German and English), this study proposes an empirically founded universal concept for formulaic language and discusses what the shape of this concept implies for the theoretical understanding of formulaic language going forward. In particular, it is argued that the nexus of the concept of formulaic language cannot be fixed at any particular structural level (such as the phrase or the level of polylexicality) and incorporates elements specified at varying levels of schematicity. This means that a cross-linguistic concept of formulaic language fits in well with a constructionist view of linguistic structure.

1 Introduction

In this chapter, I set out to assess whether formulaic language (FL) can be regarded as universal in a comprehensive sense, and if so, what such a universal concept of FL looks like. To make this assessment possible, data from Korean, German and English are used – between them, these languages cover the spectrum of morphological typology, which is arguably the most pertinent typological classification when it comes to FL.

One way of characterising FL is to say that it represents habitual turns of phrase in a speech community (cf. Burger et al. 1982: 1; Coulmas 1979; Erman and Warren 2000; Fillmore et al. 1988; Howarth 1998: 25; Langacker 2008: 84; Pawley 2001). Such typical ways of putting things may include conversational formulae (e.g. *Thank you very much – not at all),* collocations (like *face a challenge*, or *utter*

https://doi.org/10.1515/9783110669824-005

disgrace), multi-word terms (*open letter*, *contempt of court*) as well as other habitual sequences (*half an hour*, *no chance of X*, *behind closed doors*) and, to the extent to which they are in recurrent use within a community, idioms (like *get one's knickers in a twist*) and even proverbs (*garbage in, garbage out*).

FL is held to be of central importance to the functioning of language in a number of key ways. For example, besides making up a sizable portion of language in use (Altenberg 1998; Butler 2005: 223), knowledge of FL is thought a prerequisite for full proficiency in a language, register, dialect or sociolect. This is because habitual turns of phrase are crucially only a subset of all expressions that might be judged grammatical (e.g. Bally 1909: 73; Pawley and Syder 1983: 191; O'Keeffe et al. 2007: 60) and so knowledge of the boundaries of grammaticality alone is insufficient. FL is also thought to ease processing load during language production and thus it is nothing less than a key enabler of fluency in language (Nattinger and DeCarrico 1992, Pawley and Syder 1983, Wray and Perkins 2000). Further, research suggests that FL is key to successful mutual understanding in communication because items of FL activate a range of social, situational and cultural contextual cues (Erman 2007: 26; Feilke 1994, 2003: 213, Wray 2008: 20–21). Hence even in lingua franca communication among L2 speakers, communities move fast to establish a stock of FL to aid mutual understanding, as shown by Seidlhofer (2009). In short, much in language depends on FL.

It is likely that items of FL are found in languages universally (Colson 2008: 191). Previous phraseological research has established the existence of FL phenomena in very many different languages, including all major European languages and less widely spoken European languages and dialects (cf. overview in Burger et al. 2007: part XIV and the survey of 74 European and 17 non-European languages in Piirainen 2012) as well as Arabic (Abdou 2011), Catalan (Bladas 2012), Chinese (Shei and Hsieh 2012), Hebrew (Al-Haj et al. 2014), Hindi (Shama 2017), Japanese (Namba 2010), Korean (Kim et al. 2001), English as a Lingua Franca (Kecskes 2007; Seidlhofer 2009) and indeed artificial languages like Esperanto, Interlingua and Ido (cf. Fiedler 2007), to name only a few of the more recently investigated varieties (See also major comparative works including the recent Idström and Piirainen 2012; Benigni et al. 2015 and the large number of monolingual and multilingual phrasebooks and idiom dictionaries, e.g. anon. 2010; Cownie 2001).[1] Consequently, there is every reason to expect that languages

[1] Arguably even programming languages feature items of FL that represent habitual ways of coding tasks in a programming language (cf. *programming idioms*, e.g. in Maruch and Maruch 2011: ch. 21).

that have not yet had their phraseology documented will nevertheless be shown to feature FL.

Crucially, however, the points made above regarding the importance of FL to the functioning of language in general require that FL is not only found in all languages but found in comparable measure in all languages (universality in the comprehensive sense): it would be difficult to maintain that some languages feature a greater density of habitual ways of expression than others (all else being equal) or that fluency and mutual understanding is better or more easily achieved in some languages than others by virtue of their higher rate of FL occurrence. To date, no quantitative cross-linguistic studies have confirmed whether FL is indeed found in similar measure across different languages or whether the degree of reliance on FL in fact varies between languages and language varieties, though results of some studies appear to point to non-universality of FL in the comprehensive sense (e.g. Kim 2009).

This is a matter of very considerable consequence for the study of FL: if it were found that different languages rely on FL to very differing degrees, widely-accepted theoretical claims about the importance and role of FL (such as those outlined above) would require a fundamental re-examination – there would be a strong possibility that FL may in fact be a mere epiphenomenon, a language-specific reflex of a more general, yet to be formulated principle that manifests itself differently in different languages, rather than a phenomenon of theoretical interest in itself. If, on the other hand, a coherent concept of FL can be formulated that is equally valid across typologically diverse languages, it would reaffirm the significance of FL in linguistic theory and contribute substantially to an understanding of FL that is able to sustain the continued expansion of phraseological research into new domains and its application to new data.

In the following, I will first outline some of the main ways in which FL has been understood. Then previous research relevant to the question of comprehensive universality will be reviewed, along with the relevant concepts of linguistic typology. The *data and procedure* section subsequently outlines how a trilingual, topic-matched corpus of around 80 million words of Korean, German and English was put together and how it was used to test the universality of the concept of FL. In the final two sections, results of this analysis are presented and their significance discussed.

2 Background

2.1 Formulaic Language

There is a range of current understandings of and approaches to FL and phraseological phenomena. This complicates any statements made about FL in general because it begs the question to which understanding of FL those generalisations apply. But the plurality of understandings is also a sign of the multi-faceted nature of the phenomenon at hand which invites a diversity of approaches and conceptualisations and it is an index of the vitality of research into FL which attracts scholars from diverse fields, and with diverse interests and research agendas.

At the risk of a degree of oversimplification, it is nevertheless possible to identify main strands of thinking on FL which I will do by reviewing three main approaches: the traditional phraseological, the psycholinguistic and the corpus linguistic. Traditional phraseology considers the criterion-triplet of polylexicality (i.e. items involving more than one word), idiomaticity (semantic and/or syntactic irregularity) and fixedness (or stability) of key importance in conceptualising FL (cf. Burger et al. 1982). While the criteria of polylexicality and fixedness are common to most concepts of FL, the prominence of the criterion of idiomaticity has meant that idioms and proverbial expressions, although shown to be comparatively infrequent in language use (Moon 1995: V, 1998: 81; Colson 2007), have tended to be a particular (and occasionally exclusive) focal point within this strand of thinking. In the second strand, here dubbed psycholinguistic, the aspect of mental processing features particularly prominently. Sinclair described relevant entities as “phrases that constitute single choices” (1991: 110), and the idea of prefabrication is prominent in this strand, as for example in Wray’s definition of a formulaic sequence 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.
> (Wray 2002: 9)

Since processing occurs in individuals’ heads, formulaicity in this view might be understood primarily as a feature of idiolect rather than of the shared language system. The final line of thinking focuses on the aspect of conventionality in relation to speech communities, as manifested in language use. This can be summed up in the characterisation of FL as expressions that represent habitual ways of putting things in a community. Early formulations referred to “combinations sanctioned by usage” (Bally 1909: 73, my translation), while more recent work in this line of thinking has described FL as conventional or institutionalised phrases (Pawley 2001: 122; Bybee 2010: 35, respectively; cf. also Howarth 1998: 25; Brunner and Steyer 2007: 2). In specifically corpuslinguistic work, conventionality is typically measured via variously modulated measures of frequency of occurrence, as in the pioneering study by Altenberg and Eeg-Olofsson (1990) that made clear that idiomatic sequences are vastly outnumbered by conventional, non-idiomatic sequences that should nevertheless be considered instances of FL.

While conceptions of FL and their associated terminologies are therefore diverse, they also coincide in key characteristics, such as their tendency to involve units larger than words that display stability of form across instances of use. Views diverge on the importance of idiomaticity and on whether the mental processing of individuals or the shared conventions of a language community are the most relevant aspects of FL. Although the approach followed in this study is an inclusive, corpus-linguistic approach of the third strand, the commonalities between strands ensure that conclusions drawn are relevant to FL in general.[2]

2.2 Universality and Typological Difference

Above it was pointed out that previous research has established the existence of items of FL in a diverse range of languages. It was also argued that if the well-established theoretical claims about FL are to be maintained, the mere existence of tokens of FL in the languages of the world provides insufficient support for the universality of the full concept of FL. Only evidence of comprehensive universality (i.e. of comparable levels of recourse to items of FL across languages) would confirm the central importance of FL to the functioning of language in general. In the following, therefore, the focus will be on previous research that throws light on aspects of this comprehensive type of universality.

Although generally the concept of FL is most often treated as cross-linguistically unproblematic in FL literature, a number of authors have overtly commented on aspects of *comprehensive universality* and cross-linguistic concepts of FL. Wray (2002), for example, offers comments about the influence of flexible word-order on the nature of FL and highlights the fundamental nature in which typological differences can affect FL:

[2] Although, due to the rarity of narrowly idiomatic expressions among all items of FL, results will be less relevant to a conception of phraseology that is concerned exclusively with items displaying semantic and/or syntactic irregularity.

> While an English phrase might be fully fixed except for, say, the verb morphology, its German equivalent might need to contain two slots for the verb, with one or the other being filled according to the syntactic environment.
>
> (Wray 2002: 269; similarly Heid 2012 and others)

Like most theoretical works, however, Wray otherwise presents findings in terms of properties of language in general while drawing primarily on a single language. Although some theoretical treatments are more circumspect when suggesting generalisations across languages (e.g. Fellbaum 2007: 2), the discussion of cross-linguistic aspects is largely left to one side, leading to Colson's perceptive comment that "[o]n the basis of European syntax, we may have a slightly biased view of what phraseology looks like in other [i.e. non-European] languages" (Colson 2008: 193).

Specifically cross-linguistic studies have overwhelmingly focussed on strongly idiomatic items of FL (for overviews and discussion of contrastive phraseology see esp. Colson 2008; also Burger et al. 2007: part XIII; Földes 1997) and have uncovered findings particularly relating to the figurative semantics of idioms and their possible implications for universal tendencies in human cognition (e.g. Dobrovol'skij and Piirainen 2005) and how widely similar idioms are shared between languages (e.g. Piirainen 2012). Other studies discussing cross-linguistic aspects (e.g. Butler 1997, 2005; Cortes 2008; Granger 2014) have presented insightful comparisons of form and function in items of FL, typically across pairs of languages. Though none of these studies directly address the question of comprehensive universality or propose adjustments to the concept of FL based on their comparisons, Granger observes in relation to lexical bundles in French and English that "the overall number of n-grams may differ across languages" (2014: 61) and that

> a lexical bundle approach [to FL] is likely to generate more interesting results if the languages compared are sufficiently close morphologically, lexically and syntactically.
>
> (Granger 2014: 61)

However, Granger views this and similar issues caused by "typological differences between languages" (60) as methodological problems to which solutions need to be found rather than matters of theoretical importance. Similarly, Kim (2009) in a comparison of Korean lexical bundles of three-word length in conversation and academic texts, finds that "in Korean, [...] lexical bundles are generally rare overall due to the wide range and variety of word endings" and "[t]he findings of the current study [...] suggest that typological differences are obviously central to any explanation of these differences" (2009: 157). The fundamental questions this raises regarding the nature of FL are not discussed.

On the other hand, Durrant's (2013) study of formulaicity in Turkish offers important insights regarding the nature of FL in relation to language typology: based on extensive corpus evidence, he demonstrates formulaicity at the morpheme sequence level and suggests that in agglutinating languages, this may pick up the shortfall in the number of recurring multi-word items of FL. Durrant maintains that

> [s]ince individual word forms are rare, so too are high-frequency word combinations. [...] it may be that collocation is better described as relationships between lemmas, or between specifiable subsets of a lemma, or even between suffix combinations, abstracted from lexical roots.
>
> (Durrant 2013: 34)

Similar insight may be gleaned from treatments of FL in languages that do not mark word boundaries orthographically, such as Chinese. In Chinese orthography, characters represent single "syllables associated with a morpheme" (Sun 2006: 102) and are not grouped orthographically into words. Since morphemes are furthermore "more indeterminate with respect to their bound [or] free status" (Sun 2006: 46) the word "is neither a particularly intuitive concept nor easily defined" (Sun 2006: 46–49), creating immediate problems for the FL-criterion of polylexicality. Hence Shei and Hsieh, when describing items of FL in Chinese place the locus of formulaicity at the morphological level: they point out that "there are traditionally a huge number of four-morpheme units called *cheng2yu3* ([...] "established language", "idiom") [...] used to show erudition or simply for succinct meaning making" (2012: 327), but that the "issue of large habitually formed morpheme groups [...] is not so well investigated to date" (2012: 328). They then proceed to outline a "method which can separate idiomatic expression from ad hoc polysyllabic [i.e. polymorphemic] strings" (2012: 328), operating, again, at the morpheme level.

In summary, discussions of cross-linguistic aspects of FL, where they have occurred at all, have rarely engaged with the question of *comprehensive universality* or the concept of FL that might underlie it. The studies that have compared semantic, functional and structural aspects of items of FL have not, in general, commented on the effects of differing morphological behaviour among languages on the concept of FL or on quantitative aspects, leading to the apparent assumption that existing understandings of FL are unproblematically universal. Kim (2009), Granger (2014) and Durrant (2013) have shown, however, that this cannot be assumed and that models of FL as recurrent strings of word-forms, for example, are unlikely to be universal in the comprehensive sense. Consequently, the question of how recurrent complex units should be conceived of, across very different languages, has so far not been investigated at anything approaching the depth which would be necessary to support the ambitious research programme that is currently pursued in the area of FL, or indeed to safeguard the theoretical importance currently attached to the concept of FL. The next section lays out how the question of whether, and if so in what way, FL is comprehensively universal was assessed in this study.

3 Data and Procedure

How does one work out whether and how FL is universal in the comprehensive sense? The approach taken in this study is a quantitative, corpus-linguistic one involving three basic steps: in a first step, a novel genre, topic, structure and size-matched trilingual corpus of languages representative of the breadth of diversity found across morphological typology was compiled. Next, comprehensive automatic extractions of items of FL from each of the languages represented in the corpus were carried out, employing various candidate universal concepts of FL and measuring their effects. In the third and final step, the concept of FL that succeeded in yielding a closely comparable number of extracted items of FL across the three language sections of the corpus (thus simulating a comprehensively universal concept of FL) was assessed in terms of whether it is a theoretically viable concept of FL or one that does not form a plausible basis for the shape of a universal concept of FL. In the former case, the relevant simulated universal FL concept would furnish the basis for an explanation of how FL is universal; the latter case would suggest that FL is not universal in the comprehensive sense.

3.1 Corpus Compilation

In compiling a corpus for present purposes, a range of features needed to be considered to obtain valid results. The most fundamental of these was the choice of languages compared. Known factors likely to influence FL-density, including genre, topic and corpus size, also needed controlling across the different language sub-corpora.

The languages chosen for the comparison were Korean, German and English. Languages can be classified in various ways according to a multitude of features. Some of the more common linguistic typologies have classified languages according to word order, vocabulary or morphological type. While all of these criteria will influence FL to some extent, in this study, morphological classification was used as the basis for source data selection as this type of variation is clearly pertinent to FL (cf. below as well as Durrant 2013; Granger 2014).

The discussion of morphological typology in this section essentially follows Whaley (1997: 127–148). Morphological typology can be understood as a classification of the morphological behaviour of a language on two semi-independent continua. One is the continuum of synthesis (or morphemes per word ratio), with isolating languages (few morphemes per word) at one extreme and synthetic ones (many morphemes per word) at the other. The other continuum is that of fusion, with agglutinating languages (where individual morphemes remain recognisable as they are combined) at one end and fusional or (in)flectional languages where morphemes typically merge with one another, at the other extreme. Languages are placed at different points on the continua according to their tendencies which are, however, not necessarily uniform (Song 2001: 43). Korean, German and English take up different positions on the continua and therefore represent the breadth of diversity found across morphological typology: English is the most isolating language of the three, whereas German is more synthetic and also more fusional. Korean is yet more synthetic, though unlike German it is agglutinating (Sohn 2001: chapter 8). The situation is roughly sketched in figure 1.

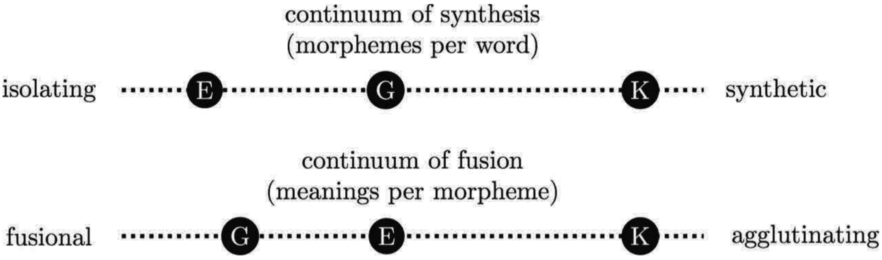

Fig. 1: Continua of synthesis and fusion. Note: E=English, G=German, K=Korean; placements are approximate

It is well known that genre and register influence the types of FL found, but crucially here, genres are also known to differ in the degree to which they rely on FL (Ädel and Erman 2012: 81; Biber 2006, 2009; Biber and Barbieri 2007; Biber et al. 2003; Kuiper 2009; Lenk and Stein 2011; Stein 2007). Topic may well have similar effects and it was therefore decided to control for topic as well as genre. To exclude possible effects of both, while avoiding the complications of translated texts, the sub-corpora for each language were drawn from Wikipedia articles, with 75% of articles in each language being on shared topics and 25% of articles on topics not covered by the respective other languages (see table 1 for an overview).[3] A 100% match of topics would not have been feasible as article topics in some languages correspond to sections of more general articles in others and vice versa. It was also thought important to capture a proportion of language used to discuss indigenous topics, as it were, because shared topics would inevitably be more globalised in nature. Since several shorter texts may not be equivalent to a single text of the same total size in important respects, corpus structure in terms of number of documents was also matched across language sub-corpora. Where necessary, articles had random paragraphs removed in order to match sub-corpora in both overall size and in the number of documents included.

Tab. 1: Corpus composition. Note: SID = syllable information density; shared docs = documents with shared topics across languages

	total docs	shared docs	syll. count	word count	SID
Korean	63,075	40,545	67,164,785	25,021,576	1
German	63,075	40,349	55,840,652	28,636,204	1.203
English	63,075	40,501	48,004,421	29,077,310	1.4

Perhaps the most obvious factor to be controlled was sub-corpus size. Traditionally in corpus linguistics, size is measured in number of words. However, words are not cross-typologically stable units. As laid out in the above discussion on morphological typology, isolating languages tend to split morphemes into many words while synthetic languages pack many morphemes into single words resulting in situations where whole phrases in isolating languages like English are equivalent to single words in highly synthetic languages like Korean with obvious implications for measurements of corpus size.[4] A measure of corpus size independent of the concept of 'word' was therefore required and a measure based on syllabic information density (SID) was chosen instead. SID (Pellegrino et al.

[3] Translation across Wikipedia pages in various languages does occur, but "articles in the different versions are often written directly in the respective target-language" (Mc Donough Dolmaya 2015: 16). Warncke-Wang et al. (2012) found that of the 1,253,523 articles of the German Wikipedia, only 0.306% were as translations, and only 0.267% of the English Language Wikipedia. In any case, however, due to article creation and editing being collaborative and continuous, even articles with translation activity at a certain stage in their history are not likely to be translated texts in any conventional sense.

[4] The concept of a word is problematic from a theoretical point of view, both within and even more so across languages (cf. Dixon and Aikhenvald 2002).

2011; Oh et al. 2013) measures the amount of information packed into a syllable and then allows for corpus size to be specified on the basis of a density-adjusted number of syllables, rather than words, leading to a balanced amount of language across sub-corpora.

To determine equivalent sub-corpus sizes based on SID, densities were first obtained for each language. This was done on the basis of a set of 825 sentences of Korean, German and English that were translation equivalents of each other and of mixed translation direction. The sentences were obtained from the Tatoeba database of sentence translations (Ho 2009). Information density was then calculated as the ratio of the total number of syllables found in the Korean sentences (baseline) to the number of syllables of German and English respectively. This resulted in the quotients given in the final column of table 1. These indicate that Korean has the lowest SID, followed by German and then English, which packs the most information into a single syllable. These figures were cross-validated against those obtained by Oh using different data (Oh et al. 2014; Oh, personal communication) and proved closely similar. Densities were then used to calculate the target number of syllables needed for each sub-corpus by dividing the baseline (Korean) syllable count by the SID for each of the other languages. The resulting figures are again shown in table 1. As the word counts of table 1 indicate, although Korean features the lowest SID (therefore requiring the highest number of syllables), Korean words contain the most syllables on average and so when measured in words, the Korean sub-corpus is the smallest, followed by the German and then the English language sub-corpus. The amount of language compared, however, is equivalent.

In terms of the actual process of corpus construction, the full Wikipedia dumps for all articles in Korean, German and English (as per February 2013) were downloaded, divided into one document per article and then cleaned and stripped of Wikipedia's XML and non-textual information using WikiExtractor (Attardi and Fuschetto 2012). The relevant documents as per table 1 were then compiled into a trilingual corpus, observing the target syllable and document counts as outlined above. Random paragraphs of some documents were left out in order to achieve the target syllable count within the necessary number of documents. To facilitate the subsequent analyses, a morphological annotation layer was added. For German and English, TreeTagger (Schmid 1994) was used to add part-of-speech, lemmas and morphological parsing; HanNanum (Park 2011) was used to add the same to the Korean sub-corpus, additionally annotating morpheme boundaries.

3.2 Identification of FL

This section describes the procedure employed in the identification of items of FL in corpus data and the options available within the procedure to simulate various underlying FL-concepts. Above, items of FL were characterised as expressions representing habitual ways of putting things in a speech community. The idea of conventional ways of putting things implies that there are both units of meaning (i.e. things to be 'put'), and linguistic forms conventionally associated with those meanings (i.e. ways of putting them). For the purposes of automatic identification and extraction, therefore, the operationalization in (1) was used:

(1) Frequent sequences of linguistic elements forming a semantic unit

Linguistic elements were taken to be word forms in the first instance (more specifically, white space delimited orthographic words) with the option to also consider lemmas (i.e. words abstracted away from features like case marking), morphemes (i.e. sub-lexical units of meaning) and combinations of these. Sequences of 2 to 9 elements in length were considered. Following the corpus-linguistic strand of thinking on FL, conventionalisation was measured via frequency of occurrence in corpus material; *frequent* was taken as minimally occurring twice per million words. A *semantic unit* was deemed a word sequence possessing the sort of semantic unity typical of words and structurally complete phrases. Semantic unity was also attributed to sequences that, while lacking this unity, can acquire it through the addition of a single, semantically or formally restricted variable element at either edge of the sequence (such as when *in search of* does not form a full semantic unit unless a variable element on the right is added, i.e. *in search of X* where *X* is restricted semantically to something prized that is being pursued). For reasons of practicality, the phenomenon of sequence-internal variable slots (such as *at the [young/early/average/premature] age of X*) was not specifically catered for as only continuous sequences of elements were extracted. There is no indication that this decision affected the three tested languages unequally, and the most frequent fillers of variable slots will be extracted in-situ as an additional sequence type (i.e. *at the age of X*, *at the early age of X* and *at the young age of X* as separate types). For a more detailed discussion of internal variability, see Buerki (2016).

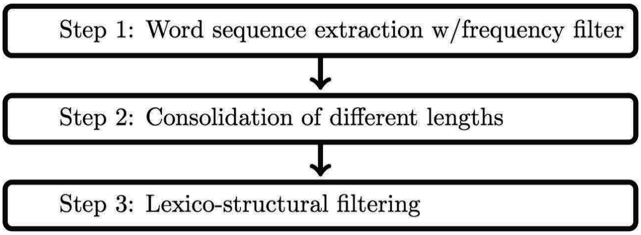
Step 1: Word sequence extraction w/frequency filter
Step 2: Consolidation of different lengths
Step 3: Lexico-structural filtering

Fig. 2: Main steps of the identification procedure

The actual identification of items of FL from each sub-corpus was conducted in three steps (cf. figure 2). The extraction from each sub-corpus of all word sequences occurring at least twice per million words (step 1) was carried out using the N-Gram Processor (Buerki 2013). To aid accuracy, sequences across sentence and sentence-equivalent boundaries were blocked and an additive stop list was used. The stop list contained the 200 most frequent word forms of the respective language according to the Leipzig Corpus Portal (anon. 2001) and served to eliminate exclusively sequences that are made up entirely of stop-listed (i.e. very high-frequency) words.[5] In step 2, the various lengths of identified sequences had their frequencies consolidated and were combined into a single list using SubString (Buerki 2017). At step 3, lexico-structural filters were applied to the lists of sequences to remove sequences that were likely to lack semantic unity. One entry of the lexico-structural filter for English, for example, bars sequences ending in the word ‘and’ as most such sequences would fail to show semantic unity. A detailed discussion of the extraction procedure (applied to a different data set) is found in Buerki (2012).

Extraction accuracy was established as follows. A random sample (n = 300 types) of automatically identified sequences in each language was rated for compliance with the operationalisation in (1) by the author and independently by an L1 speaker of the respective language acting as a research assistant. Extraction accuracy at the baseline (i.e. using sequences of orthographic word forms exclusively) varied between languages and raters in the range of 72% to 75% of sequence types rated as operationalisation compliant. Recall (the comprehensiveness of an extraction) is difficult to assess in this scenario, but is typically inversely related to accuracy, that is, higher accuracy leads to lower recall and vice-versa (Manning and Schütze 1999: chapter 5). The accuracy figures achieved

[5] For German, a stop list based on the top 150 (rather than 200) most frequent words proved sufficient to yield comparable extraction accuracy to the other languages.

were therefore regarded as suited to present purposes. Notably, the achievement of a narrow range of variation in extraction accuracy between the three languages was critical because it means that comparability between extractions across languages was successfully maintained. A higher accuracy for one language, for example, would almost certainly have caused a lower number of sequences to be extracted for that language, thus introducing a bias. Thus a robust identification procedure was applied to enable the subsequent quantitative comparison of FL across the three languages studied.

3.3 Simulation of FL Concepts

As noted above, the first (baseline) FL-concept tested for universality employed orthographic word form sequences as the basic building blocks. This represents a traditional FL-concept in that it accepts the multi-word level as the relevant level at which formulaicity is manifested and it is also very conservative in terms of fixedness – it takes the view that all elements of a habitual turn of phrase are fully fixed such that, for example, the sequences in (2) are deemed separate types of sequences, each needing to satisfy FL-status on its own, rather than being tokens of one sequence.

(2) *consists of X*
consisting of X
consisted of X
consist of X

Two exceptions to full fixedness applied even at the baseline level (in addition to allowing variable slots at either edge): all numbers (whether in figures or words) were replaced by the label NUM, and occurrences of the names for months of the year were replaced by the label NMONTH. This allowed the identification of sequences like those in (3) as a single type.

(3) *NUM days later (two/ten/21 days later)*
in NMOUNTH of that year (in April/July/August of that year)
in the early NUMth century (in the early twentieth/17th century)

Although adequate for many cases, previous studies have shown that as a general requirement, (almost) complete fixedness is not realistic as items of FL are subject to a substantial amount of variation (Wray 2002: chapter 14; Sinclair 2004: 161; Langlotz 2006; Dutton 2009). An exception here is the idea of lexical bundles

(Biber et al. 1999; Biber and Conrad 1999) which uniquely requires complete fixedness. Since there is no definition of lexical bundles independent of their operationalization (resulting in a conflation of theory and method) it remains unclear whether full fixedness is of theoretical importance to the idea of lexical bundles or simply a methodological expediency.

As reported in the next section and expected on the basis of previous research (Granger 2014; Durrant 2013; Kim 2009), the baseline concept of FL failed to produce comparable FL-densities in the three languages. Consequently, progressive changes were made to the FL-concept tested until approximate parity in FL-densities across the three languages was reached. In this iterative process, modifications to the FL-concept were progressively stepped up through aspects of fixedness to more fundamental alterations concerning the level at which formulaicity applies. While at each stage, modifications to simulated FL-concepts were made incrementally and with a view to maintaining plausibility as far as possible, it is important to recall that the goal was to take the simulation to whatever level necessary to produce approximate parity in FL-density across languages, and subsequently to assess whether the resulting comprehensively universal FL-concept is a plausible one or not. Thus it was never in doubt whether parity could be achieved (this is a relatively simple exercise), but rather what modifications would be necessary, to what extent alterations would be needed and whether the resulting concept was plausible. The results of this process are detailed in the next section.

4 Results

As a baseline for comparisons, the results of a FL-concept of (almost) complete fixedness and taking the orthographic (white space separated) word sequence as the level at which formulaicity is manifested, are presented in figure 3 and table 2. Several key observations result: first, the number of items of FL identified across the languages is vastly different (both in terms of types as well as tokens) and therefore the underlying concept of FL is clearly not universal in the comprehensive sense. It is evident, therefore, that an understanding of FL similar to the baseline concept used here has to be regarded as a language-specific phenomenon in that density of occurrence varies greatly between languages. Perhaps the most prominent such concept is the idea of lexical bundles, which is even more fixed and depends to a much greater extent on (ultra-high) frequency of occurrence as a defining characteristic than the baseline concept used here. A second immediate observation is that the number of items identified as FL in each language parallels the placement of the respective language on the continuum of synthesis (cf. figure 1). This confirms the dependence of the baseline concept of FL on typology – something distinctly undesirable for a concept of importance to language in general rather than certain languages only.

Tab. 2: Items of FL under the baseline FL concept

	FL-types	FL-tokens
Korean	10,617	1,480,862
German	19,114	2,677,999
English	25,712	3,727,071

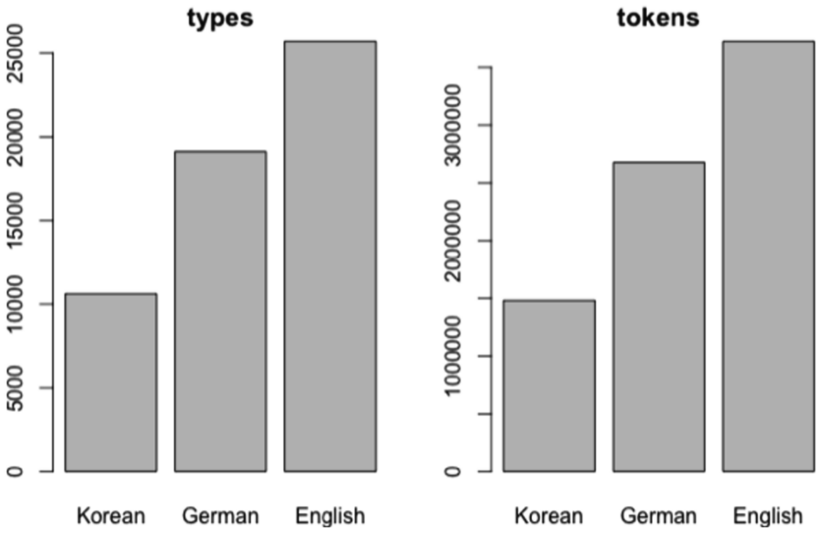

Fig. 3: Items of FL under the baseline FL

By contrast, the figures obtained by employing a simulated universal concept of FL are presented in figure 4 and table 3. These figures show that it is entirely possible to automatically identify a comparable number of items as formulaic in each of the languages. The question to consider is whether the underlying concept of FL is a plausible, coherent and sensible concept within the context of what is known about FL. To make this assessment, the changes to identification parameters implemented to move from the baseline concept of FL to the simulated universal concept are set out below, and subsequently assessed.

Tab. 3: Items of FL under the universal FL concept

	FL-types	FL-tokens
Korean	24,345	3,619,171
German	26,577	3,807,337
English	25,712	3,727,071

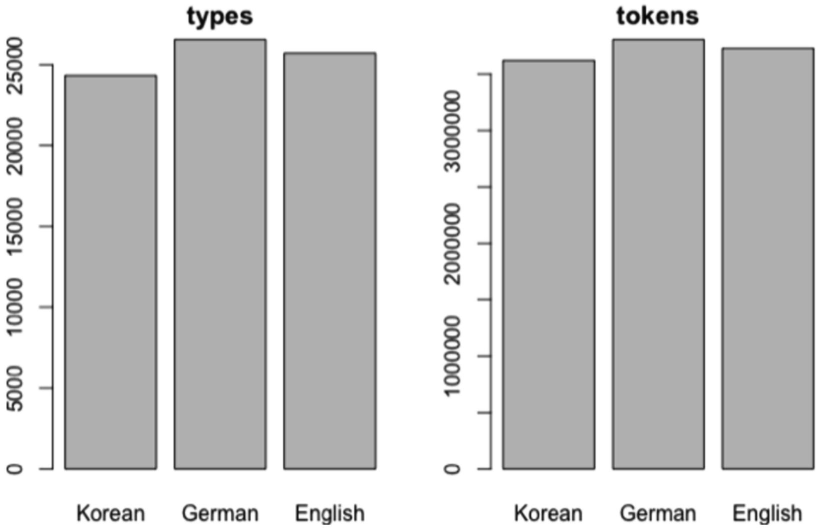

Fig. 4: Items of FL under the universal FL concept

4.1 Adjustments

The adjustments indicated below were implemented by adapting a version of the source corpus and then re-running the FL identification procedure with commensurate adjustments to stop lists and filters where necessary.

4.1.1 Fixedness

The first set of adjustments was made to the degree of fixedness: as pointed out above, phraseological research has long maintained that many items of FL require certain types of flexibility. Such flexible items could be said to be underspecified to a degree, or specified at a more schematic level than the word form sequence – and they require adjustments to fit contexts of use. One type of flexibility is the occurrence of variable slots as seen above. Others are alternations in word order and inflectional morphology. The effects of morphological typology seen in the results reported above suggest that inflectional morphology is pertinent to the differences observed in the data and so the first set of adjustments to the concept of FL was made to reduce fixedness in areas of inflectional morphology.

In Korean, this flexibility was simulated by removing case markers of subject (-*이/가 [i/ga]*) object (-*을/를 [eul/reul]*) and topic (-*은/는 [eun/neun]*) as realised by the bound morphemes indicated, as well as all plural markers (-*들[deul]*).[6] Notably, the absence of these markers does not necessarily result in ungrammaticality as they are "frequently omittable" (Sohn 2001: 231). Korean also possesses an elaborate system of verbal (and in some cases adjectival) inflection to mark politeness levels (Sohn 2001: 231–241), though other aspects, such as grammatical person, are not marked morphologically. The formal style used in texts like Wikipedia articles, however, means that only a very narrow range of these inflections is manifest, rendering intervention superfluous. To exemplify effects of adjustments made, items in (4) can be seen united under a single sequence type (5) as a consequence of the adjustments.

(4)	*버스 정류장을*	*[beoseu jeongriujangeul]*	*bus stop-OBJ*
	버스 정류장은	*[beoseu jeongriujangeun]*	*bus stop-TOPIC*
	버스 정류장이	*[beoseu jeongriujangi]*	*bus stop-SUBJ*
	버스 정류장	*[beoseu jeongriujang]*	*bus stop*
(5)	*버스 정류장*	*[beoseu jeongriujang]*	*bus stop*

Morphology to mark tense/aspect, mode and modality was left unadjusted – as in other languages (including German and English), these are expressed partly

[6] There is some disagreement over whether these markers are more suffix-like (as assumed here) or more word-like (cf. Sohn 2001: 231). As current orthography does not typically afford these markers the status of orthographic word, they are taken as bound morphology here.

by inflectional morphology and partly periphrastically. The relaxation of fixedness would therefore not have contributed to addressing typological differences between the languages compared.

Methodologically, the adjustments mentioned were implemented by producing a version of the source corpus that had all items deleted that were marked by the morphological parser as instances of the Korean subject, object, topic and plural markers. The FL-identification procedure was then re-run to produce a new list of items of FL.

In the German sub-corpus, an equivalent reduction in fixedness was targeted by masking all verbal inflections for grammatical person (but tense/aspect, mode and modality was again retained as this is marked in all the languages under investigation and would therefore not target differences).[7] Further, all case and gender inflection was masked on definite and indefinite articles, adjectives and nouns (but number distinctions were retained as they occur in the English sub-corpus as well and were deemed an overly harsh generalisation for these languages). Again, to illustrate the effect of some of these adjustments, sequences in (6) appear united under (7) after the adjustments.

(6)	*die Bundesrepublik Deutschland*	*the Federal Republic of Germany*	[nominative]
	der Bundesrepublik Deutschland	*the Federal Republic of Germany*	[dative/genitive]
(7)	ARTDEF Bundesrepublik Deutschland		

Similarly, after adjustments the nine attested sequence types in (8) appear under (9) as four generalized types.

(8)	*zur Verfügung stehen*	*be available*	[1st/3rd pers. pl, pres. tense]
	zur Verfügung steht	*be available*	[3rd pers. sg/2nd pers. pl, pres. tense]
	zur Verfügung stehe	*be available*	[3rd pers. sg, subjunctive I]
	zur Verfügung stünden	*be available*	[1st/3rd pers. pl, subjunctive II]
	zur Verfügung stand	*be available*	[1st/3rd pers. sg, past tense]
	zur Verfügung standen	*be available*	[1st/3rd pers. pl, past tense]
	zur Verfügung stehende	*available*	[adjectival, case/number marked]
	zur Verfügung stehenden	*available*	[adjectival, case/number marked]
	zur Verfügung stehender	*available*	[adjectival, case/number marked]
(9)	*zur Verfügung stehen_IndPres*	[indicative, present tense]	
	zur Verfügung stehen_IndPast	[indicative, past tense]	
	zur Verfügung stehen_Subj	[subjunctive]	
	zur Verfügung stehend	[adjectival]	

[7] This was done by replacing finite verbs with lemmas marked for tense and mode.

As shown, distinctions in tense and mode are retained, but flexibility is introduced with regard to grammatical person (for verbs) and case/number marking on adjectival expressions. Notably, not all available forms of the respective inflectional paradigms are attested in the corpus (*zur Verfügung stehen* [be available] does not occur in the data with inflection for first or second person singular, for example) and some forms occur only a few times. This is partly due to the particularities of the corpus, of course, but is also a manifestation of a degree of fixedness of the expression. Therefore, even when extensive flexibility in terms of inflection is introduced, this does not necessarily lead to the identification of as many more items of FL on the basis of heightened recurrence as might be expected. The examples also show that the range of inflectional morphemes is further limited by the fusion of morphemes – the last three forms in (8) represent all possible combinations of case and number marking.

Methodologically, these adjustments were again achieved by modifying a copy of the source corpus in which all German verb forms, adjectival forms, forms of the definite and indefinite article and noun forms were replaced with the respective lemma (plus the added information on mode, tense, number, etc. that was to be retained) as seen in (9). The FL-identification was then re-run.

In English, an equivalent level of flexibility is inherent due to the absence of some of the equivalent inflectional morphology on the one hand and the isolating morphology on the other. The effect of the latter is seen in (6), where English would require the addition of the free morpheme *of* for genitive case marking in the second line, but this would still leave the recurring 5-element sequence *the Federal Republic of Germany* intact (and easily identifiable) in both lines of (6).

Although further flexibility could have been introduced to the simulated FL concept, this was not deemed judicious because the adjustments introduced already cover the aspects of flexibility that are pertinent to the typological differences in morphology present in the data set: there would have been little gain, for example, in such sweeping adjustments as a complete generalisation over tense marking because morphological tense marking is not a feature on which the three languages differ categorically.[8] Despite little further room for sensible reductions in fixedness, checks at this stage of the simulation indicated that a

[8] While the focus of this study is on morphological differences, it is likely that a simulated generalisation over aspects of word order would reduce differences in this regard between Korean and German as languages with more word order variation on the one hand and English with less word order variation on the other (although, of course, there is some word-order variation in English as well; cf. Heid 2012). It has to be left to future studies to ascertain the magnitude of the impact of these differences.

comprehensively universal concept of FL was not yet achieved. The simulation was therefore stepped up to comprise another area to which previous research has drawn attention: the level(s) at which formulaicity operates.

4.1.2 Levels of Focus

The final step to the universal concept of FL that produced the figures of table 3 required adjustments to the levels at which constituent elements of FL are recognised, to include certain units at the morpheme level. In Korean, 14 common bound morphemes occurring word-finally (the translation equivalents of which are generally free morphemes in German and English), were separated from their hosts so that they became eligible for recognition as independent constituents of formulaic sequences. The morphemes concerned are: -*의* ([ui] of); -*에서* ([eseo] from); -*에* ([e] at); -*로*/*으로* ([ro/euro] towards); -*과*/*와* ([gwa/wa] and); -*하고* ([hago] and); -*고* ([go] and); -*에게* ([ege] to); -*도* ([do] too), -*부터* ([buteo] from); -*까지* ([kkaji] until); -*만* ([man] only); -*마다* ([mada] every); -*지* ([ji] not). This was implemented by identifying all instances of the named morphemes in a copy of the source corpus and isolating them from their hosts through the insertion of white space characters. Identification occurred with the help of part-of-speech tags supplied by the morphological parser, as many of the forms involved, being single or double syllables, also occur as constituents of other lexical items, or as homographs).

In German, compounds consisting of common words were separated so their constituents become eligible for recognition as independent constituents. Although both Korean and English feature compounds as well, German is particularly noted for its use of compounding and the length of its compounds (cf. Russ 1994: 221–225), making this an important area where formulaicity remains unrecognized in one language but picked up in others due only to differences in morphological typology. In addition, German compounds are typically single orthographic words where many English and some Korean compounds consist of multiple orthographic words (hyphens were treated as separate words in all languages, resulting in hyphenated compounds like *open-minded* being treated as 3-element expressions). In example (7) above, the cross-linguistic effect of German compounds is drawn into focus as the German expression consists of three elements, whereas the English gloss consists of five. After compound-separation, (7) appeared as in (10) featuring four elements.[9]

[9] *Deutschland* might also be split but was left whole by the splitting software (s. below).

(10) ARTDEF Bundes republik Deutschland

jWordSplitter (Naber 2015) was used to divide German compounds in a copy of the source text and the FL-identification procedure was repeated. jWordSplitter divides noun compounds and some verbal and adjectival compounds. Due to necessarily limited coverage of the morphological dictionaries used, jWordSplitter is in practice most effective splitting compounds that consist of common word forms, giving it a fairly light touch that turns out to be well suited for the level of adjustments required in the present case.

As before, no equivalent adjustments to the English language data was necessary as the adjusted Korean bound morphemes are already independent in English (as well as German) and the adjustment to German compounds now approximated the state of compounds in English (and Korean). A sample of identified items of FL, including items identified only after adjustments (marked with an asterisk), is shown in table 4.

Tab. 4: Sample of identified FL. Notes: * items resulting from an adjusted FL-concept; X = variable slot; NUM = numbers; ARTDEF = definite article

	items of FL	**gloss**
Korean	어느 정도 [eoneu jeongdo]	roughly, to some extent
	영어 로* [ieongeo ro]	in English
	지금 도* [jigeum do]	even now
	X 과 함께* [X gwa hamkke]	together with X
	박사 학위[를] 취득하었다 [baksa hakwi[reul] chwideukhaeotta]	received [their] PhD
	유럽 연합 의* [yureop yeonhap ui]	of the European Union
	X 후 곧바로 [X hu gotbaro]	right after X
	X 때 마다* [X ttae mada]	always when X
	그 다음 에* [geu daeum e]	after that
	X 에 따라 달라진다* [X e ttara dallajinda]	differ depending on X
	X 있는 것으로 알려져 있다 [X inneun geoseuro allyeojyeo itta]	it has become known that there is X
	첼로 협주곡 [chello hyeobjugok]	cello concerto
	X (으)로 인하여* [(eu)ro inhayeo]	because of X
	X 와 같이* [X wa gachi]	with X
	오래 된 [orae doen]	old (lit. long been)
	X 에 대한 지원* [X e daehan jiwon]	support for X
	NUM 살 의 나이 로* [NUM sal ui nai ro]	at the age of NUM
German	aus diesem Grund	for this reason
	zu diesem Zeit punkt*	at this point in time

	items of FL	gloss
	ARTDEF Stadt zentrum*	the city centre
	in diesem Sinne	in this way/sense
	anhand von X	by means of X
	Tage buch*	diary (day book)
	auch sonst	at any rate / anyway
	immer größer	bigger and bigger
	dazu führen, dass X*	lead to the outcome that X
	siehe unten	see below
	miteinander verbunden	connected to each other
	wie zum Beispiel X	as for example X
	DEFART so genannte/r/n X*	the so-called X
	bis zu seinem Tod NUM	until his death in NUM
	hinzu kommen/kommt, dass X*	added to this, X
	nach dem Krieg	after the war
	bereits im NUMten Jahrhundert	going back to the NUMth century
	in Frage gestellt	questioned
	stehen/steht unter Denkmal schutz*	be listed (i.e. be a listed building)
English	X was released in NUM	
	until his death in NUM	
	large amounts of X	
	natural resources	
	mainland China	
	Member of Parliament	
	internal combustion	
	consistent with X	
	open to the public	
	the Olympic Games	
	on several occasions	
	science and technology	
	in the US state of X	
	by the early NUMs	
	special effects	
	it is thought that X	
	incompatible with X	
	on the grounds of X	
	in a NUM – NUM victory over X	
	due to the fact that X	

4.2 Assessing the Simulated Concept of FL

Having reviewed the underpinnings of the simulated universal concept of FL that underlies the figures presented in table 4, we can now outline its main features: with regard to fixedness, the universal concept of FL allows flexibility minimally in the areas of inflectional morphology to do with case marking, marking of agreement and if necessary marking of number to allow items of FL that specify these aspects at a more schematic level than the word form. Additionally, the universal concept of FL used is flexible with regard to the locus of formulaicity and recognizes formulaicity at the morpheme level.[10]

It is important to note that in this context *flexibility* does not mean that in all cases items of FL must be pitched at the most schematic level: the results discussed suggest that many items may be pitched at that level, but others are not and there will be different elements of the same item at differing levels of schematicity. For example, the German item of FL *eines Tages* (some day, at some point in time; lit. *of a day*) is fixed in the genitive case, but the more schematic *ARTINDEF Tag*[11] (a day) is still a common turn of phrase forming a semantic unit regardless of case marking. Similarly, the Korean phrase *예를 들면 [iereul deulmeon]* (*for example*, lit. *if [we] take an example*) invariably specifies the object case marker *–를 [-reul],* including in all 1,214 occurrences of the expression in the corpus, despite case marking in general often being omitted, as discussed above). Similar examples are mentioned by Granger (2014: 60) (see also Tognini-Bonelli (2001) for a defense of the word form as relevant unit). In this sense, the identification of items of FL carried out above was a *simulation* of a flexible FL concept; an actual identification based on a flexible FL concept would identify items at

[10] It may be argued that instead of the flexibility claimed to be necessary, it may be sufficient (or at least partially sufficient) simply to adjust the minimum frequency level for less isolating languages as part of the identification procedure (as Granger 2014 suggests), or that, in effect, the need for flexibility is created artificially by using frequency as part of the operationalization of FL. But this argument would be problematic: frequencies would have to be lowered very substantially from an already low threshold to get a similar effect because unlike in certain other procedures, frequency is only one element of the operationalization of FL used. A substantial lowering of threshold frequency would result in a much lower accuracy of identification (unless replacement filtering devices are used), meaning that the additional items would be unlikely to be bona fide items of FL in the sense used in this study. More fundamentally, frequency bears theoretical significance as it is used to operationalize conventionality and so is a fundamental, rather than accidental, aspect of FL according to the understanding of FL put forward. Consequently, adjustments to take account of this are justifiable.

[11] ARTINDEF is the label used for a lemma of the indefinite article.

their most relevant level(s) of schematicity, which might well differ for each constituent element.

It can now be considered whether the universal concept of FL described is plausible. There are two main considerations that strongly suggest that this universal concept of FL, besides succeeding empirically, also forms a coherent and sensible concept from the point of view of theory. First, the features of increased flexibility in levels of schematicity and locus of formulaicity are not novel features, but have been suggested, albeit more tentatively, by previous studies as outlined above. This analysis has principally added an indication of their scope and necessity. Second, specification at various and mixed levels of schematicity (with some elements highly fixed in all aspects and others much less so) and the loss of the significance of the distinction between the word and morpheme levels are features that are not unusual: if we turn to constructionist approaches to grammar (also known as Construction Grammar and noted for their tight interfacing with phraseological theory and data, cf. Van Lancker Sidtis 2015; Buerki 2016), these features are not only accommodated but predicted as features for linguistic structures across language (Hilpert 2014; Hoffmann and Trousdale 2013). In constructionist theory, all of language consists of constructions that are specified at the full range of levels of schematicity from fully substantive (lexically fixed) to fully schematic. For example, the fully schematic ditransitive construction in (11) is as much a bona-fide construction as the partially lexically substantive construction in (12) or the fully substantive construction in (13).

(11) <Subj V Obj1 Obj2> (e.g. I handed her the book)
(12) <the Xer the Yer> (e.g. the bigger the better)
(13) <blue jeans>

Further, in constructionist theory, constructions exist from the level of single morpheme or morpheme group to that of phrase without a theoretically significant distinction between word and morpheme level constructions (cf. table 1.1 in Goldberg 2006: 5). Consequently, constructions like *<prebook>* or *<over-V>* (as in *overeat, oversleep*, etc.) are as much constructions as *<blue jeans>* or phrase-level constructions (11) and (12). From a constructionist viewpoint it therefore comes as no surprise that a universal concept of FL should admit items that are specified at various and mixed levels of schematicity, such as specification of the exact word form for all elements as in (14), specification at word form level for all but one element in (15) where the second element is specified at a more schematic level that allows case marking flexibility, specification at a fairly abstract level as in (16), which only contains two fully substantive elements, or indeed (17) which is formulaic at the morpheme sequence level.

(14) eines Tages (one day)
(15) 박사 학의(를) 취득하였다 *[baksa hakwi(reul) chuieukhaeotta]* (received [their] PhD)
(16) X [was/is to be/will be/is due to be] released in NUM
(17) Tagebuch (diary, lit. book of days)

The comprehensively universal concept of FL outlined above therefore not only succeeds in demonstrating its comprehensive universality across the three languages in our data, but also presents itself as a plausible concept of FL, taking previous phraseological research and insights from constructionist theory into account.

5 Discussion and Outlook

The results of this study fall into three general areas of significance. The first concerns the concept of FL and in what sense it is applicable universally to different types of languages. Here results show that it is possible to construct a concept of FL that applies in equal measure to isolating languages such as English with a low morpheme-per-word ratio, languages like German that feature a vast array of case, gender and agreement morphology, as well as polysynthetic, agglutinating languages like Korean, where individual words are often equivalents of whole phrases in more isolating languages. This is significant, because although the existence of FL is documented in a wide range of languages, previously FL was not subject to large-scale cross-linguistic comparison of quantitative aspects and such comparisons as have been conducted have yielded stark cross-linguistic differences in the number of items of FL identified. This posed fundamental challenges to the adequacy of the theoretical claims outlined above, most principally to the central importance of FL to the functioning of language in general, but these claims have now been safeguarded by the presentation of a plausible universal concept of FL.

Second, results crucially also reveal that this cross-linguistically viable concept of FL must incorporate two key aspects that have hitherto not been prominently discussed or applied: on the one hand, the inclusion within the concept of FL of more flexible, more schematic forms that require fine-tuning at time of use (as well as fully substantive forms that do not) is a requirement for a plausibly universal concept of FL, not an optional or marginal feature. While some items of FL are best identified as fully substantive forms that allow their use in context without any further adjustments, more schematic forms that require morphological fine-tuning must equally be recognised as FL. In the data, this fine-tuning typically consists of adjustments for case, number, or person, but may include other aspects. The point here is that without the ability to stipulate schematic forms, many individual, fully substantive forms will on their own be too rare to be reasonably considered common turns of phrase in their own right and this shortfall has vastly more serious consequences for languages that use, for example, case marking than for languages that do not, resulting in vastly different amounts of FL being detected between such languages. On the other hand, an adjustment of a more radical nature is required if the notion of FL is to be a universal one: the traditional fixation on FL as sequences of words needs to be relativized and sequences of sub-word-level linguistic items need to be eligible for recognition as legitimate items of FL. Again, the data indicate that this is necessary for the notion of FL become universally applicable. Thus results indicate that a universal concept of FL is viable but absolutely requires the admission of sequences that need a degree of fine-tuning at the time of use, and further requires a discounting of the importance of the word level that has hitherto been a prominent feature in conceptualisations of FL.

Third, results also suggest adjustments to the place of FL in an overall theory of language. In terms of theories of linguistic structure (i.e. syntax and morphology), notionally, FL can be integrated into various frameworks (cf. Wray 2008: chapter 7) or it may be envisaged as a completely separate module or "subsystem" (Dobrovol'skij 1992: 279) of the grammar. However, the requirement for a universal concept of FL to discount the significance of the word level, and the inclusion of sequences at differing levels of schematicity, strongly support and integrate with constructivist approaches to grammar. These approaches place linguistic constructions (from fully substantive phrases to fully schematic constructions), rather than words and rules of combination, at the centre of theoretical thinking. Items of FL function in this view as constructions of a particular, namely a predominantly substantive, type. Therefore, a universal concept of FL suggests a natural integration with constructivist theories of language where FL is able to take up an important place, commensurate with its importance in accounting for how language operates.

There are of course also a number of limitations to consider: only some, though arguably the most pertinent, aspects of how languages vary have been considered in this study. Detailed consideration of other aspects, such as the effects of freer word orders in some languages, and other features of languages not investigated in this study will no doubt add further important detail to a universal concept of FL. In its outline however, the concept put forward is unlikely to change dramatically.

Overall, results obtained offer strong evidence for a cross-linguistically robust notion of FL and how it fits into a larger theoretical context. This advances the field of research into FL by placing it on a firmer footing and by affirming its importance in accounting for how language works. This firmer footing can sustain current interest in the phenomenon and contribute to stimulating further research into theoretical as well as applied aspects of FL.[12]

References

Abdou, Ashraf (2011): *Arabic idioms: A corpus based study*. London: Routledge.

Al-Haj, Hassan, Alon Itai & Shuly Wintner (2014): Lexical Representation of Multiword Expressions in Morphologically-complex Languages. *International Journal of Lexicography* 27 (2), 130–170.

Altenberg, Bengt (1998): On the phraseology of spoken English: The evidence of recurrent word-combinations. In Anthony Paul Cowie (ed.), *Phraseology: Theory, analysis and applications*, 101–122. Oxford: Clarendon Press.

Altenberg, Bengt & Mats Eeg-Olofsson (1990): Phraseology in Spoken English: Presentation of a Project. In Jan Aarts & Willem Meijs (eds.), *Theory and Practice in Corpus Linguistics*, 1–26. Amsterdam: Rodopi.

anon. (2001): *Word lists*. Retrieved from http://wortschatz.uni-leipzig.de/html/wliste.html, accessed April 30, 2015.

anon. (2010): *Get the Hang of it. 3000 Redewendungen in fünf Sprachen*. Köln: Anaconda.

Attardi, Giuseppe & Antonio Fuschetto (2012): *WikiExtractor 2.2* [computer programme].

Ädel, Annelie & Britt Erman (2012): Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes* 31 (2), 81–92.

Bally, Charles (1909): *Traité de stylistique française, premier volume,* Paris: Librairie C. Klincksieck.

Benigni, Valentina, Paola Cotta Ramusino, Fabio Mollica & Elmar Schafroth (2015): How to apply CxG to phraseology: a multilingual research project. *Journal of Social Sciences* 11 (3), 275–288.

Biber, Douglas (2006): *University language: a corpus-based study of spoken and written registers*. Amsterdam, Philadelphia: John Benjamins.

Biber, Douglas (2009): A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14 (3), 275–311.

Biber, Douglas & Federica Barbieri (2007): Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26 (3), 263–286.

Biber, Douglas & Susan Conrad (1999): Lexical bundles in conversation and academic prose. *Language and Computers* 26, 181–190.

[12] The work was supported in part by the Swiss National Science Foundation under Grant P2BSP1_148623.

Biber, Douglas, Susan Conrad & Viviana Cortes (2003): Lexical bundles in speech and writing: an initial taxonomy. In Andrew Wilson, Paul Rayson & Tony McEnery (eds.), *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*, 71–92. Frankfurt am Main: Peter Lang.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan (1999): *Longman grammar of spoken and written English*. Harlow: Pearson.

Bladas, Òscar (2012): Conversational routines, formulaic language and subjectification. *Journal of Pragmatics* 44 (8), 929–957.

Brunner, Annelen & Kathrin Steyer (2007): Corpus-driven study of multi-word expressions based on collocations from a very large corpus. *Proceedings of CL2007, University of Birmingham, UK, 27–30 July 2007*. Retrieved from http://corpus.bham.ac.uk/corplingproceedings07/paper/182_Paper.pdf, accessed September 21, 2012.

Buerki, Andreas (2012): Korpusgeleitete Extraktion von Mehrwortsequenzen aus (diachronen) Korpora. In Natalia Filatkina, Ane Kleine-Engel, Marcel Dräger & Harald Burger (Hrsg.), *Aspekte der historischen Phraseologie und Phraseographie*, 263–292. Heidelberg: Universitätsverlag Winter.

Buerki, Andreas (2013): *N-Gram Processor 0.4* [computer programme].

Buerki, Andreas (2016): Formulaic sequences: a drop in the ocean of constructions or something more significant? *European Journal of English Studies* 20 (1), 15–34.

Buerki, Andreas (2017): Frequency consolidation among word n-grams: A practical procedure. In Ruslan Mitkov (ed.), *Computational and corpus-based phraseology*, 432–446. Cham: Springer.

Burger, Harald, Annelies Buhofer & Ambros Sialm (1982): *Handbuch der Phraseologie.* Berlin, New York: De Gruyter.

Burger, Harald, Dmitrij Dobrovol'skij, Peter Kühn & Neal R. Norrick (eds.) (2007): *Phraseology: an international handbook of contemporary research*. Berlin, New York: De Gruyter.

Butler, Christopher S. (1997): Repeated word combinations in spoken and written text: some implications for functional grammar. In Christopher Butler, John H Connolly, Richard A Gatward & Roel M. Vismans (eds.), *A fund of ideas: recent developments in functional grammar*, 60–77. Amsterdam: IFOTT.

Butler, Christopher S. (2005): Formulaic language: an overview with particular reference to the cross-linguistic perspective. *Pragmatics and beyond. New series* 140, 221–242.

Bybee, Joan L. (2010): *Language, usage and cognition,* Cambridge: Cambridge University Press.

Colson, Jean-Pierre (2007): The World Wide Web as a corpus for set phrases. In Harald Burger, Dmitrij Dobrovol'skij, Peter Kühn & Neal R. Norrick (eds.), *Phraseology: an international handbook of contemporary* research, 1071–1077. Berlin, New York: De Gruyter.

Colson, Jean-Pierre (2008): Cross-linguistic phraseological studies. In Sylviane Granger & Fanny Meunier (eds.), *Phraseology: an interdisciplinary perspective*, 192–206. Amsterdam: John Benjamins.

Cortes, Viviana (2008): A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora* 3 (1), 43–57.

Coulmas, Florian (1979): On the sociolinguistic relevance of routine formulae. *Journal of Pragmatics* 3 (3/4), 239–266.

Cownie, Alun R. (2001): *Dictionary of Welsh and English idiomatic phrases*. Cardiff: University of Wales Press.

Dobrovol‘skij, Dmitrij (1988): *Phraseologie als Objekt der Universalienlinguistik*. Leipzig: Enzyklopädie.

Dobrovol’skij, Dmitrij (1992): Phraseological universals: theoretical and applied aspects. In Michel Kefer & Johann van der Auwera (eds.), *Meaning and grammar: cross-linguistic perspectives*, 279–301. Berlin, New York: De Gruyter.

Dobrovol’skij, Dmitrij (2000): Contrastive idiom analysis: Russian and German idioms in theory and in the bilingual dictionary. *International Journal of Lexicography* 13 (3), 169–186.

Dobrovol’skij, Dmitrij & Elisabeth Piirainen (2005): *Figurative Language: Cross-cultural and cross-linguistic Perspectives*. Oxford: Elsevier.

Durrant, Philip (2013): Formulaicity in an agglutinating language: the case of Turkish. *Corpus Linguistics and Linguistic Theory* 9 (1), 1–38.

Dutton, Kelly (2009): *Exploring the boundaries of formulaic sequences: a corpus-based study of lexical substitution and insertion in contemporary British English*. Saarbrücken: VDM.

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

Erman, Britt & Beatrice Warren (2000): The idiom principle and the open choice principle. *Text* 20 (1), 29–62.

Feilke, Helmuth (1994): *Common sense-Kompetenz: Überlegungen zu einer Theorie des „sympathischen“ und „natürlichen“ Meinens und Verstehens*. Frankfurt am Main: Suhrkamp.

Feilke, Helmuth (2003): Textroutine, Textsemantik und sprachliches Wissen. In Angelika Linke, Hanspeter Ortner & Paul R. Portmann (Hrsg.), *Sprache und mehr. Ansichten einer Linguistik der sprachlichen Praxis*, 209–230. Tübingen: Niemeyer.

Fellbaum, Christiane (2007): Introduction. In Christiane Fellbaum (ed.), *Idioms and collocations: corpus-based linguistic and lexicographic studies*, 1–22. London: Continuum.

Fillmore, Charles, Paul Kay & Mary O’Connor (1988): Regularity and idiomaticity in grammatical constructions: the case of *let alone*. *Language* 64 (3), 501–538.

Granger, Sylviane (2014): A lexical bundle approach to comparing languages: Stems in English and French. *Languages in Contrast* 14 (1), 58–72.

Goldberg, Adele E. (2006): *Constructions at work: the nature of generalization in language*. Oxford: Oxford University Press.

Heid, Ulrich (2012): German noun+verb collocations in the sentence context: morphosyntactic properties contributing to idiomaticity. In Thomas Herbst, Susen Faulhaber & Peter Uhrig (eds.), *The phraseological view of language: A tribute to John Sinclair*, 283–311. Berlin, New York: De Gruyter.

Hilpert, Martin (2014): *Construction grammar and its application to English*. Edinburgh: Edinburgh University Press.

Ho, Trang (2009): *Tatoeba Project*. Retrieved from http://tatoeba.org, accessed December 1, 2014.

Hoffmann, Thomas & Graeme Trousdale (eds.) (2013): *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press.

Howarth, Peter (1998): Phraseology and second language proficiency. *Applied Linguistics* 19 (1), 24–44.

Idström, Anna & Elisabeth Piirainen (eds.) (2012): *Endangered metaphors*. Amsterdam: John Benjamins.

Kecskes, Istvan (2007): Formulaic language in English Lingua Franca. In Istvan Kecskes & Laurence R. Horn (eds.), *Explorations in pragmatics: Linguistic, cognitive and intercultural aspects* (Volume 1), 191–218. Berlin, Boston: De Gruyter.

Kim, Seonho, Juntae Yoon & Mansuk Song (2001): Automatic extraction of collocations from Korean text. *Computers and the Humanities* 35 (3), 273–297.

Kim, You-Jin (2009): Korean lexical bundles in conversation and academic texts. *Corpora* 4 (2), 135–165.

Kuiper, Koenrad (2009): *Formulaic Genres*. Houndmills: Palgrave Macmillan.

Langacker, Ronald W. (2008): Cognitive Grammar as a Basis for Language Instruction. In Peter Robinson & Nick C. Ellis (eds.), *Handbook of cognitive linguistics and second language acquisition*, 66–88. Abingdon: Routledge.

Langlotz, Andreas (2006): *Idiomatic creativity: a cognitive-linguistic model of idiom-representation and idiom-variation in English*. Amsterdam: John Benjamins.

Lenk, Hartmut E. H. & Stephan Stein (Hrsg.) (2011): *Phraseologismen in Textsorten*. Hildesheim: Olms.

McDonough Dolmaya, Julie (2015): Revision history: Translation trends in Wikipedia. *Translation Studies* 8 (1), 16–34.

Manning, Christopher D. & Hinrich Schütze (1999): *Foundations of statistical natural language processing*. Cambridge, Mass.: Massachusetts Institute of Technology Press.

Maruch, Stef & Aahz Maruch (2011): *Python for dummies*. Chichester: Wiley.

Moon, Rosamund (1995): Introduction. In John McH Sinclair (ed.), *Collins COBUILD dictionary of idioms,* iv–vii. London: HarperCollins.

Moon, Rosamund (1998): Frequencies and Forms of Phrasal Lexemes in English. In Anthony Paul Cowie (ed.), *Phraseology: theory, analysis and applications*, 79–100. Oxford: Clarendon Press.

Naber, Daniel (2015): *jWordSplitter* [computer programme].

Namba, Kazuhiko (2010): Formulaicity in code-switching: Criteria for identifying formulaic sequences. In David Wood (ed.), *Perspectives on formulaic language: Acquisition and communication*, 129–150. London: Continuum.

Nattinger, James R. & Jeanette S. DeCarrico (1992): *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Oh, Yoon Mi, François Pellegrino & Edigio Marsico (2014): La complexité des langues du monde. *Pour la Science* 82, 66–71.

Oh, Yoon Mi, François Pellegrino, Egidio Marsico & Christophe Coupé (2013): A Quantitative and Typological Approach to Correlating Linguistic Complexity. *Proceedings from the 5th Conference on Quantitative Investigations in Theoretical Linguistics* (*QITL) Leuven*, 12-14 September 2013, 71–75. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=8B48B870A6ECAF02510C5D8B364DAC33?doi=10.1.1.398.7882&rep=rep1&type=pdf, accessed November 22, 2014.

O'Keeffe, Anne, Michael McCarthy & Roland Carter (2007): *From corpus to classroom: language use and language teaching*. Cambridge: Cambridge University Press.

Park, Sangwon (2011): *HanNanum* [computer programme].

Pawley, Andrew (2001): Phraseology, linguistics and the dictionary. *International Journal of Lexicography* 14 (2), 122–134.

Pawley, Andrew & Frances Syder (1983): Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In Jack C. Richards & Richard W. Schmidt (eds.), *Language and communication*, 191–226. Harlow: Longman.

Pellegrino, François, Christophe Coupé & Egidio Marsico (2011): Across-Language Perspective on Speech Information Rate. *Language* 87 (3), 539–558.

Piirainen, Elisabeth (2012): *Widespread idioms in Europe and beyond: toward a lexicon of common figurative units*. New York: Peter Lang.

Russ, Charles (1994): *The German language today: a linguistic introduction*. London: Routledge.

Schmid, Helmut (1994): *Probablistic Part-of-Speech Tagging Using Decision Trees. Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK.* Retrieved from http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf, accessed April 20, 2015.

Sharma, Sunil (2017): Happiness and metaphors: a perspective from Hindi phraseology. *Yearbook of Phraseology* 8, 161–180.

Shei, Chris & Hsun-Ping Hsieh (2012): Linkit: a call system for learning Chinese characters, words, and phrases. *Computer Assisted Language Learning* 25 (4), 319–338.

Seidlhofer, Barbara (2009): Accommodation and the idiom principle in English as a Lingua Franca. *Intercultural Pragmatics* 6 (2), 195–215.

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

Sinclair, John McHardy (2004): *Trust the text*. London: Routledge.

Sohn, Ho-Min (2001): *The Korean Language*. Cambridge: Cambridge University Press.

Song, Jae Jung (2001): *Linguistic typology: Morphology and syntax*. Harlow: Pearson Longman.

Stein, Stephan (2007): Mündlichkeit und Schriftlichkeit aus phraseologischer Perspektive. In Harald Burger, Dmitrij Dobrovol'skij, Peter Kühn & Neal R. Norrick (eds.), *Phraseology: an international handbook of contemporary research*, 220–236. Berlin, New York: De Gruyter.

Sun, Chaofen (2006): *Chinese: A linguistic introduction*. Cambridge: Cambridge University Press.

Tognini-Bonelli, Elena (2001): *Corpus linguistics at work*. Amsterdam: John Benjamins.

Van Lancker Sidtis, Diana (2015): Formulaic language in an emergentist framework. In Brian MacWhinney & William O'Grady (eds.), *The handbook of language emergence*, 578–599. Chichester: Wiley-Blackwell.

Warncke-Wang, Morten, Anuradha Uduwage, Zhenhua Dong & John Riedl (2012): In search of the Ur-Wikipedia: universality, similarity, and translation in the Wikipedia inter-language link network. *Proceedings of the eighth annual international symposium on wikis and open collaboration, Linz (A)*. Retrieved from https://dl.acm.org/citation.cfm?id=2462959, accessed August 07, 2018.

Whaley, Lindsay J. (1997): *Introduction to typology: the unity and diversity of language*. London: SAGE.

Wray, Alison (2002): *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.

Wray, Alison (2008): *Formulaic language: pushing the boundaries.* Oxford: Oxford University Press.

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