Is there a Formula for Formulaic Language?

Author 1: Richard S. Forsyth (Independent researcher),

Author 2: Łukasz Grabowski (Opole University)

Abstract: This paper focuses on detecting and measuring traces of 'formulaic language'. For this purpose, we test a number of computational formulae that quantify the degree to which a text type incorporates inflexible sequences of words. We assess these candidate indices using a number of reference corpora representing a wide variety of text types, both routine and creative. We adopt the concept of 'phrase-frame' proposed by Fletcher (2002-2007) as a means of exploring phraseological pattern variability. To date, there have been few studies explicitly addressing this issue, with the exception of Roemer (2010). We examine 10 productivity indices, including Roemer's VPR, the Herfindahl-Hirschman index, Simpson's diversity index and relative Shannon entropy. We report that a novel measure, which we term Hapaxity, best meets our criteria, and show how this index of micro-productivity (in phrase-frames) may be used to assess macro-productivity (in text registers), thus quantifying an important aspect of a register's reliance on formulaic subsequences.

Keywords: formulaic language, phraseology, phrase-frames, corpus-driven research, linguistic statistics

1. Introduction

Linguists of various schools have noted the "unlimited creative potential" (Eggins 1994: 117) of human language. Chomsky, in particular, has emphasized linguistic creativity:

> "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." (Chomsky 1972: 118).

Yet when we receive a 'form letter' from a government department or cast our eyes inattentively over an example of legalistic corporate 'boilerplate' at the foot of an email message we are forced to acknowledge that some everyday uses of language fall well short of this ideal of unbounded creativity. Thus we observe a polarity in linguistic expression -- from regurgitated boilerplate on one side to creative innovation on the other. The term 'formulaic language' denotes language nearer the left pole than the right, less rigid than simple cut-&-paste but nevertheless allowing only a restricted range of expressive options.

This polarity arises because, despite the creative potential of language attested by Chomsky and others, opposing tendencies also operate in both speech and writing that lead to the production of prefabricated phrases. One such tendency has been dubbed by Sinclair (1991) "the idiom principle". Also, there is ample empirical evidence available now to linguists showing that "language users never choose words randomly, and language is essentially non-random" (Kilgarriff 2005: 263-264) simply because we tend to speak or write with specific purposes in mind. This means that very often linguists study texts exemplifying routinized patterns of linguistic behaviour. Thus, the "deadly repetitiousness of language" (Bolinger 1965: 570) is what stares us "in the face from the text" and makes many uses of language restricted and formulaic (Firth 1968 [1957]). As Halliday puts it (2014), "repeated patterns require less brain power both to produce and to understand."

However 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 study is therefore to evaluate a number of computable indices of linguistic flexibility/inflexibility which could serve as indicators of the degree to which a corpus 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 (Wray 2002: 4), one that has not been answered in a comprehensive manner. According to some often-cited estimates (e.g. Erman & Warren 2000) formulas constitute up to 50% of spoken and written texts; there are also studies where this proportion is estimated to be as high as 80% (Altenberg 1998) or as low as 32% (Foster 2001). For example, Altenberg (1998: 101) claims that “a rough estimation indicates that over 80 per cent of the words in the corpus form part of a recurrent

word-combination in one way or another", where a recurrent word combination is "any continuous string of words occurring more than once in identical form" (ibid.).[1]

According to Wray (2002: 28 & 2009: 36), such claims as to the proportion of language viewed as formulaic are hardly compatible with each other. The main reasons for this incompatibility are the different ways in which researchers operationalize the concept of formulaic language and the different methodologies used so far to measure proportion of formulaic language in texts. More specifically, Wray (2009: 36) notes that researchers apply different frequency thresholds, explore strings of word forms of different lengths, use either inductive or deductive approaches to selection and further analysis of alleged formulas, to name but a few factors. This is no surprise since the term 'formulaic sequence' is "intentionally all-encompassing, covering a wide range of phraseology": this situation makes it "difficult to identify absolute criteria which define formulaic sequences" (Schmitt & Carter 2004: 3).

With such discrepancies in view, there appears to be no agreed method of taking a group of corpora and ranking them from the most to the least formulaic on an objective basis. We cannot expect a single method to cover all the different aspects of such a complex phenomenon. However, in the following section we consider some of the chief attributes of formulaic language, which need to be taken into account in any operationalization of the concept.

2. Formulaic language: clearing the undergrowth

Generally speaking, formulaicity in texts refers to the frequent use of a high number of formulaic sequences (or, in short, formulas) characterized as "various types of wordstrings"

[1] However, having applied additional criteria, namely the length of at least three words and the frequency of occurrence in the corpus equal or higher than 10, as well as deletion of unintentional repetitions and stuttering, the proportion fell to 3 % of tokens and 1 % of types of the initial total sample of recurrent word-combinations in spoken texts (Altenberg 1998: 101-104).

which appear 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" (Wray & Perkins 2000: 1; Wray 2002: 9). This phenomenon is explained by, among others, the tendency of certain words to have "an especially strong relationship with each other in creating their meaning" (Wray 2008: 9). Furthermore, Wray (2008: 19) argues that her conceptualization of formulaic sequences, focusing on storage of formulas and their retrieval from memory, is meant to be as inclusive as possible so that no linguistic material for which one can put forward any kind of argument for being treated as formulaic is dismissed.

In fact, researchers of various disciplines have proposed the whole range of types of units that may represent formulaic language in speech or writing. Also referred to as clichés, formulas – conceptualized as ready-made reproducible signs – have been explored in folklore studies as a unit of description of folklore texts, e.g. in the 'general cliché theory' (*общая теория клише*) (Permyakov 1970). In this view, clichés encompass individual phrases, proverbs, sayings, jokes, riddles, stories, fairy tales or anecdotes, and their primary function is that of facilitating the construction and reception of utterances (Chlebda 2003: 25). From the ethnolinguistic perspective, formulas are tantamount to stereotypes and defined as stable and reproducible combinations of words, culturally conveyed, socially conditioned, stored in the collective memory of language users and retrieved therefrom as semantic wholes (Bartmiński 2007: 71). Lancioni (2009: 222) argues that "formulas do matter in written languages based upon an oral-formulaic tradition" and that in studies on oral tradition the prosodic and metric constraints underlying the use of formulas have been emphasized at the expense of the formulas' collocational requirements. In a similar vein, Pawley (2009: 14) claims that speech formulas have been of particular importance in, among others, literary research on epic sung poetry, anthropology, folklore studies.

In more linguistically-oriented studies[2], formulas are seen as various types of linguistic constructions and phraseologies with restricted forms and variability as well as with restricted distributions (e.g. proverbs, sayings, clichés, catchphrases, idioms, aphorisms, n-grams, lexical bundles, clusters or collocations); they form chunks of routinized linguistic behaviour associated with frequently performed routines, retrieved from memory as wholes and processed as such.[3] Schmitt and Carter (2004: 1-2) argue that formulaic sequences are ubiquitous in language use and occur in "so many forms that it is presently difficult to develop a comprehensive definition of the phenomenon". In fact, many different types of sequences deserving the epithet formulaic are studied by researchers specializing in language acquisition, psycholinguistics, neurolinguistics, lexicography, computational phraseology, natural language processing, to name but a few fields, and each approach provides a different perspective on formulaic language.[4]

It is, however, widely accepted that formulaic language exhibits less free variation than non-formulaic language. On this basis, linguists have proposed a number of measures of linguistic variety (or lack of variety) that could reasonably be regarded as negatively (or positively) correlated with formulaic language.[5] For example, with respect to n-grams, that is, contiguous sequences of n words, or phrase frames, that is, variants of n-grams identical

[2] A selection of studies on grammatical and semantic features of formulas as well as on their evolution and stylistic distribution across languages and text types can be found in Corrigan, Moravcsik, Ouali and Wheatley (2009), where formulaic language is explored from the perspective of syntax, pragmatics, discourse structure or language acquisition, among others. Another interesting collection of studies (a more pedagogically-oriented one) on formulaic language can be found in Schmitt (2004) or Wood (2010a, 2010b). Also, Wray (2002) presents an overview of studies on formulaic sequences from a wide variety of perspectives, such as corpus linguistics, language acquisition, lexicography and phraseology, to name but a few.

[3] For example, using eye-tracking technology Underwood, Schmitt and Galpin (2004) found that we tend to read words faster when they are parts of formulaic sequences as compared with their being used or embedded in non-formulaic text. There are also studies (e.g. Kuiper 1996) that emphasize the formulas' productive advantages for language users.

[4] For example, psycholinguists focus their attention on determining whether formulas of various types are stored as single items in the language users' mental lexicons, notably in the context of language acquisition (Schmitt and Carter (2004: 2). Research in natural language processing, notably targeted at machine translation, focuses on the identification of apparently similar (syntactically, semantically, functionally) multi-word formulas in parallel or comparable corpora (e.g. Ren et al. 2009), etc.

[5] A more detailed overview of measures applied, before the year 2002, to formulaic language can be found in Wray (2002: 28-31).

except for one word, the slot-filler (Fletcher 2002-2007), such measures include relative pervasiveness of frequent n-grams (Ellis et al. 2009) or the variant-to-phrase-frame-ratio (Roemer 2010: 105) respectively. The latter metric (VPR, in short) is meant to determine the degree of variation of a given phrase frame through dividing the number of distinct slot-fillers by the number of phrase frame tokens (ibid.). It thus represents a context-bound version of the type-token ratio. There are also indices from other fields, such as Simpson's diversity index, used in ecology to assess biodiversity, which could be adapted for this purpose and which have not been applied in corpus linguistic research on formulaic language conducted so far.

Among the desiderata for any such index proposed to measure the pervasiveness of formulaic language in a text or corpus are:

(1) that it should enable ranking of corpora from least to most formulaic;

(2) that it should help to identify the sequences or substructures that contribute to this ranking;

(3) that it should allow a comparison of these individual substructures on a common basis;

(4) that it should be relatively unaffected by corpus size.

Such an index, which would be of great help to researchers exploring formulaic language, has not been developed so far: this study constitutes a step towards finding one. In what follows, the scope of this paper as well as the research material and methodology are presented in greater detail.

3. Scope of the study, research material and methodology

In this study, we approach formulaic language from the textual rather than psycholinguistic perspective, although both approaches are not mutually exclusive. In the textual perspective, typical of corpus linguistic phraseological research, such aspects as fixedness, frequency of

occurrence, or pattern variability are among the main criteria used to define and/or detect formulaic language (Schmitt & Carter 2004; Schmitt 2005; Wray 2002, 2008, 2009). Hence, the search for traces of formulaic language in texts requires that a convenient unit of analysis be selected, preferably one that captures the essential features of the phenomenon under scrutiny and, at the same time, has not been explored in the comprehensive manner so far.

3.1 Phrase frames as the unit of analysis

In view of the above, in the present investigation we have followed Roemer (2010) in taking the phrase-frame (from now on 'p-frame') as a unit of analysis. Originally introduced by Fletcher (2002-2007), p-frames are defined as sets of variants of an n-gram identical except for one word in either the initial, medial or final position, e.g. *in the * of, if you have any *, if you * any* etc. Other scholars refer to such patterns as "formulaic frames" (Biber 2009) or “lexical frames” (Gray & Biber 2013), which typically consist of invariable function words with an intervening variable slot for content words (although Biber (2009) also allows two variable slots in formulaic frames). However, it is also possible that an intervening variable slot can be filled by a function word. Recently, several studies have employed p-frames as a means of discovering and generalizing phraseological patterns used across different registers or text varieties, or of illustrating register variation (e.g. Roemer 2009, 2010; Gerbig 2010; Fuster-Marquez 2014).

We believe that p-frames provide a convenient means of tapping into an important aspect of formulaic language since -- as explained above -- the frequency of recurrent sequences of words, as well as their fixedness and/or pattern variability jointly determine the degree to which a text is formulaic. For example, Schmitt (2005: 13) argues that “formulaic language is usually conceptualized as being basically fixed”, yet also adds that its various types or textual manifestations, e.g. recurrent n-grams or idioms, reveal a varying degree of

pattern variability. We consider that a high degree of reliance on invariant or nearly-fixed formulas will manifest itself as a relative lack of variety in commonly used p-frames. In summary, our position is that unproductive p-frames tend to indicate formulaic texts; productive p-frames tend to indicate the reverse.

In our study, we analyze only those p-frames that consist of four adjacent words with a variable slot (e.g. *in the * of*, *the end of **). This is motivated by the fact that, first, such p-frames have been explored in earlier studies (e.g. Roemer 2009, 2010; Fuster-Marquez 2014) and, second, that research on 4-word lexical bundles revealed that such units have a more readily recognizable range of structures and functions (Hyland 2008: 8; Chen & Baker 2010: 32).

To illustrate the sort of data being analyzed, four p-frames of size four from a corpus of UN Security Council Resolutions (see Appendix 1) are listed below.

*as well * the*	*93*	*1*
as well as the		*93*
*to * in the*	*89*	*20*
to serve in the		*14*
to cooperate in the		*14*
to assist in the		*14*
to participate in the		*8*
to stability in the		*5*
to play in the		*5*
to include in the		*5*
to vote in the		*4*
to date in the		*4*
to consider in the		*3*
to submit in the		*2*
to interfere in the		*2*
to contribute in the		*2*
to states in the		*1*
to interview in the		*1*
to efforts in the		*1*
to deploy in the		*1*
to countries in the		*1*
to conflicts in the		*1*
to assisting in the		*1*
*the * agreement and*	*85*	*7*

the peace agreement and		*54*
the ceasefire agreement and		*9*
the bonn agreement and		*9*
the arusha agreement and		*7*
the framework agreement and		*4*
the luanda agreement and		*1*
the algiers agreement and		*1*
*remain seized of **	*85*	*2*
remain seized of the		*82*
remain seized of this		*3*

The first of these (*as well * the*) occurs 93 times and has just a single slot-filler, the conjunction *as*. It is a fixed phrase. The second example (*to * in the*) is much more productive. It occurs 89 times with 20 different slot-fillers. The three most common of these (verbs *serve*, *cooperate* and *assist*) all occur 14 times. There are also seven fillers that occur only once each. We term these once-occurring items 'hapax legomena', or hapaxes, and thus we would say that this p-frame had 7 hapaxes among its 20 slot-fillers. (Note that our software, by default, converts all letters to lower case and omits punctuation marks.) The third p-frame (*the * agreement and*) is intermediate in terms of productivity: it occurs 85 times with 7 different variants. Its most common slot-filler, *peace*, dominates, with 54 of the 85 occurrences. It also has 2 hapaxes. The last example (*remain seized of **) is virtually fixed: it has only 2 variants. Of its 85 occurrences, 82 have the definite article as slot-filler and only 3 have the demonstrative *this*.

3.2 Investigated indicators

Using p-frames as the unit of analysis, we examine 10 quantitative indicators of productivity as inverse correlates of formulaic language, all based on the distributions of slot-fillers in p-frames, as illustrated above. They all point in the same direction, in the sense that higher scores indicate greater productivity or variety among the slot-fillers. Some of these indices are not well-known within corpus linguistics. The Herfindahl-Hirschman index (Hirschman 1964)

is used in the field of economics as a measure of the concentration of a market. It sums the squares of the proportional market shares of each firm: the more the market is dominated by the biggest firm(s), the higher the index. We use the adjusted version of this index which allows it to range from 0 to 1 (unlike the unadjusted index which ranges from 1/N to 1) and subtract it from 1 to give higher values to less concentrated, i.e. more productive, distributions. Simpson's index (Simpson 1949) is used in ecology as an index of biodiversity. A number of variations on this formula go by the name of Simpson's index in the literature. We have used the version given in Upton & Cook (2006), which is actually equivalent to the unadjusted Herfindahl-Hirschman index. Shannon entropy (Shannon 1948), a fundamental metric in information theory, is widely used in many fields, including linguistics. Balance, Hapaxity, HV and Nonfocus were devised for this investigation. All the indices explored in this paper are described in greater detail in Table 1. Where no citation is given in the table, the index is, to our knowledge, a novelty, devised for the present study. The symbols used in these formulae are explained later in Table 2.

Table 1. Formulae tested as measures of p-frame productivity

Short Name	**Formula**	**Brief Description**
Balance	1 - (L-R) / N	Score from zero to +1 measuring the flatness of the frequency distribution of slot-fillers when ranked in descending order of frequency. (1 when all slot-fillers occur equally often.)
Hapaxity	(Haps - Fmax) / N	Number of hapax legomena (once-occurring slot-fillers) less number of occurrences of the most frequent slot-filler divided by the number of occurrences of the given p-frame. Range: -1 to +1.
Haprate	Haps / N	Number of hapax legomena (once-occurring slot-fillers) as a proportion of the number of occurrences of the p-frame; proposed by Baayen (1992) in another context as an index of morphological productivity.
HC	ln(V) / ln(N)	Bilogarithmic Type-Token ratio, i.e. Herdan's C (Herdan 1964).
HH	1 - (H - 1/N) / (1 - 1/N)	Herfindahl-Hirschman index, adjusted to range fully from 0 to 1, and inverted to have higher values with lower concentrations (Hirschman 1964).

HV	$(1 + Haps + \sqrt{Haps}) / (2 + V + \sqrt{V})$	Hapax legomena divided by vocabulary size, with quasi-Bayesian attenuation factor proportional to square-root of frequency.
Nonfocus	1 - Fmax / N	Proportion of p-frame occurrences devoted to the most common slot-filler subtracted from 1. (Zero if only a single type.)
Rent	E / log2(V)	Relative entropy of the distribution of slot-fillers in a p-frame, E being average entropy (Shannon, 1948) of the distribution (see Table 2).
Simpidx	$1 - \sum p_i\text{^}2$	Simpson's index of diversity (effectively an unscaled version of HH), p_i being the occurrences of slot-filler i as a proportion of the total (Simpson 1949).
TTPC	100 × V / N	Type-Token Percentage, equivalent to VPR (Roemer 2010)[6] but with no minimum-frequency cutoff.

Table 2. Components of productivity formulae

Symbol	**Description**
E	$E = - \sum p_i \times \log2(p_i)$, p_i being the proportional frequency of the ith slot-filler
Fmax	number of occurrences of the most frequent slot-filler in a given p-frame
H	$H = \sum p_i\text{^}2$, sum of squared proportional frequencies of each slot-filler
Haps	number of once-occurring slot-fillers in a given p-frame
L	summed occurrences of all slot-fillers in the left (more frequent) half of the list of slot-fillers when ranked in descending order of frequency
N	number of occurrences of a given p-frame
R	summed occurrences of all slot-fillers in the right (less frequent) half of the list of slot-fillers when ranked in descending order of frequency
V	vocabulary size, i.e. number of different slot-fillers in a given p-frame (variants)

[6] We have used the term TTPC in this paper to distinguish it from VPR as proposed by Roemer (2010). This might seem to introduce unnecessary terminological overlap, since both measures are based on dividing the number of distinct slot-fillers by the total frequency of the p-frame concerned, i.e. both represent the type-token ratio in different guises. However, the figures quoted by Roemer (ibid.) seem likely to have been computed using an option of the kfNgram software which means that slot-fillers occurring fewer than three times are simply ignored. With this setting, the occurrence totals are also rescaled; in other words, the p-frame frequency is given as the sum of the frequencies of all slot-fillers occurring at least three times, which is usually less than the actual occurrence frequency of that p-frame. This has a drastic effect on all productivity indices computed from such data, including TTPC/VPR. At the same time the number of p-frames occurring with a given threshold frequency in any corpus decreases, thus reducing the amount of data available for comparison, which we regard as undesirable. Moreover, the choice of minimum qualifying frequency (should it be 5, 4, 3 or some other number?) introduces potential for miscommunication between researchers about what a particular VPR score signifies. These considerations, together with the significance of the concept of *hapax legomena* in linguistic studies generally, lead us to believe that it is better to deal with all the slot-fillers of p-frames, down to singletons, rather than imposing an arbitrary cutoff point. We have performed the tests reported above on that basis.

3.3 Research material

Evaluation of the indices described above involved a bootstrapping process. First, we gathered corpora representing a number of different text types exhibiting varying degrees of formulaic language.[7] Initially we had no solid grounds for rating or ranking them on a common scale from the most to the least formulaic. However, among them there were certain pair-wise comparisons where we could be confident that one corpus was indisputably more formulaic in its language usage than another.

In order to determine the most discriminating p-frame productivity index, we used ten corpora divided into 2 groups. Summary details of each group are given in Tables 3 and 4. All corpora consist of English-language texts. Word counts are as computed by a purpose-designed Python3 tokenizer used in the present study.

Table 3. Relatively more formulaic samples

Short name	**Documents**	**Total word-tokens**	**Median document length**	**Description**
ACAD	51	422109	6781	26 research articles and 25 book chapters on pharmacology
LEAF	461	482373	946	Patient information leaflets describing 461 pharmaceutical products (extracted from the Patient Information Leaflet Corpus 2.0)
PROT	240	529174	2052	Clinical Trial Protocols from the European Medicines Agency
SUMP	136	654572	4704	Summaries of Product Characteristics from OPUS website (Tiedemann, 2009)
UGAR	672	882364	955	UN General Assembly Resolutions, 2000-2003
USCR	274	246417	633	UN Security Council Resolutions, 2000-2004

[7] The corpora used in this study have been collected for personal non-commercial research, and hence are not publicly available in their totality. However, as indicated in Appendix 1, most of the source corpora are available freely online (under the links provided).

Table 4. Relatively less formulaic samples

Short name	**Documents**	**Total word-tokens**	**Median document length**	**Description**
EW	44	365158	8228	44 short stories by Edith Wharton
TEDS	1555	3371424	2264	TED talk transcripts (English) obtained from collection held at WIT3 website
WC	49	145492	2704	45 speeches by Winston Churchill, plus 2 chapters and 2 prefaces from his four-volume biography of Marlborough
LOBCORP	500	1020188	2034	Lancaster-Oslo-Bergen corpus (Hofland & Johansson 1982)

The first four corpora listed in Table 3 come from the pharmaceutical domain. They were gathered as part of an ongoing investigation into lexical and phraseological patterns and their discourse functions across pharmaceutical text types (e.g. Grabowski 2015). The last 2 corpora in Table 3 were extracted from the publications of the United Nations, collated by the German Research Centre for Artificial Intelligence (www.dfki.de). These UN corpora consist of resolutions by the General Assembly and the Security Council. The USCR texts cover a slightly longer period since fewer Security Council resolutions are issued each year.

The corpora listed in Table 4 are intended by comparison to exemplify more creative discourse. The first consists of short stories by the novelist Edith Wharton. Because they are imaginative literary compositions and written over a period of 46 years, they offer scope for stylistic variety. TEDS is a collection of English-language transcripts of talks given as part of the TED programme, hosted at www.ted.com. They consist of partly scripted oral presentations by more than a thousand different speakers ranging over a great number of subject areas and thus, although arguably belonging to a single genre, should exhibit considerable linguistic diversity. The texts by Churchill cover a period of over 60 years and thus deal with a wide range of topics. As he was a Nobel laureate in literature we assume his command of English was more creative than average. (In fact, his Nobel-prize acceptance

speech is included in the sample.) Finally, the LOB Corpus (Hofland & Johansson 1982) is a generic corpus covering several registers and thus can safely be presumed to include more linguistic variety than any of the narrowly specialized corpora in Table 3.

More details, including sources, of all the corpora presented in Table 3 and 4, as well as information about three additional corpora used as holdout samples, can be found in Appendix 1.

Importantly, we consider all text types in Table 3 relatively high in usage of formulaic language. This is a subjective judgement, yet even a cursory inspection reveals that these corpora contain a large number of recurrent sequences of words. Some examples are listed in Table 5. Such sequences are characteristic of the registers concerned but would be extremely rare in a general corpus merely on the basis of the relative frequency of their individual words. They are therefore tell-tale signs of formulaic language.

Table 5. Examples of overused fixed token sequences

Fixed token-sequence	**Source corpus**	**Percentage of source-corpus documents in which it occurs**
if you have any questions or are not sure about anything ask	LEAF	29.87
4.5 interaction with other medicinal products and other forms of interaction	SUMP	99.26
skin and subcutaneous tissue disorders	SUMP	77.94
review by the competent authority or ethics committee in the country concerned	PROT	100.00
resolution adopted by the general assembly on the report	UGAR	70.26
adopted by the security council at its	USCR	100.00
decides to remain actively seized of the matter	USCR	49.64

3.4 Overall Outline of the study

Using p-frames as the unit of analysis, the study is an attempt at identification of an effective index that can be used to detect traces of formulaic language in texts. To that end, the study consists of a number of stages.

Firstly, we perform index calibration: in order to assess the relative effectiveness of the 10 different indices described earlier in this paper for the purpose in hand, we apply each index to 12 comparisons between a pair of corpora. At this stage 12 calibration comparisons are conducted where a relatively formulaic corpus is compared with a less formulaic reference corpus. To evaluate the indices, in each comparison each index is scored according to how strongly it discriminates (using Student's paired t-test) between the 2 members of the contrasting pair. Then the ranks for each index are aggregated to give a 'winner', i.e. a measure that on this benchmark data appears to be most effective in separating text types showing high versus low levels of p-frame productivity which, because formulaic phrasing is inherently repetitive, must correlate with low versus high levels of reliance on formulaic language.

Secondly, we rank text types: the 'winning' (most discriminating) index is then used to compare the p-frame productivity of each of our text types against a reference corpus, the LOB Corpus (Hofland & Johansson 1982). This yields an ordering in terms of p-frame productivity, which allows us to rank our corpora on a common scale, from the most productive (in terms of pattern variability of p-frames) to the least productive, or – by implication – from the least to the most formulaic. Consequently, at this stage our aim is to obtain an insight into macro-productivity of the p-frames found in each corpus.

Finally, after developing a ranking of text types in terms of p-frame productivity, we examine which p-frames contribute the most to the ranking determined by aggregate macro-productivity scores obtained in the previous phase. Considering the comparison with the largest divergence score illustrates how each p-frame contributes to the scoring and thus how the behaviour of individual p-frames may be used to elucidate the details of the contrast between text types.

The study concludes with a discussion of the results, including assessment of its methodology, and conclusions with suggestions on how this research could be pursued further in the future.

4. Empirical results

4.1. Calibration Comparisons

In order to decide which of the 10 productivity indices (Table 1) is best at detecting traces of formulaic language, we set up 12 paired comparisons among our corpora. In each comparison a relatively formulaic corpus was compared with a reference corpus that we could be confident was less formulaic. The 12 comparisons made are listed in Table 6.

Table 6. Corpus comparisons

Formulaic Test Corpus	**Less Formulaic Reference Corpora**
ACAD	EW, LOBCORP
LEAF	ACAD, LOBCORP
PROT	ACAD, LOBCORP
SUMP	ACAD, LOBCORP
UGAR	TEDS, LOBCORP
USCR	WC, LOBCORP

Thus each corpus in the left-hand column was compared against 2 reference corpora. It will be seen that each text corpus was contrasted with the LOB corpus as well as another reference corpus that may be described as hand-picked, providing a specific comparison corpus as well as a more generic comparison corpus, representing a range of text types.

The role of ACAD, which appears on both sides, should perhaps be further elaborated. Work on identifying formulas in academic speech and writing (Simpson-Vlach & Ellis 2010) has revealed that academic discourse tends to be more formulaic than, for instance, fiction; but it is quite clear that, among our pharmaceutical text types, the texts in ACAD exhibit considerably more linguistic variety than LEAF, PROT or SUMP. Thus ACAD is a

borderline case in our collection. Since it is broadly in the same topic area as the other pharmaceutical corpora, it seemed the most suitable reference corpus for them.

Because differences in corpus size could be a potential confounding factor in these comparisons, random subsamples of approximately equal size were made from each of the comparison corpora other than LOBCORP (which acts as a common benchmark). The selection procedure employed was to place the documents of the corpus in a random order and then append documents one at a time to the growing subcorpus until the size of the subsample exceeded 144,000 tokens. The figure of 144,000 was chosen to approximate the size of the smallest corpus involved, namely WC, which thus was used in full.

In this randomized selection process, 2 disjoint subsamples were extracted (i.e. A and B), so that one could be held out for a post-testing phase, except in the cases of WC and USCR. The former had to be used in full, while the latter was not large enough to provide two completely disjoint samples. In the case of USCR, two partially overlapping samples were created.

Having established 12 calibration comparisons, our testing procedure was as follows. For each subcorpus, all p-frames based on contiguous sequences of four words that satisfied the following two conditions were generated.

(1) It must occur at least once per 10,000 tokens.

(2) It must occur at least 10 times in total.

For reference, Table 7 gives the actual sample sizes resulting from this procedure, for both A and B samples (and WC) along with the number of qualifying p-frames extracted in each case.

Table 7. Sizes of selected subcorpora

Subcorpus	Documents	Tokens	Median document size	P-frames extracted
ACAD144a	17	147077	7766	643
ACAD144b	19	148185	5676	417
EW144a	20	150666	6178	183
EW144b	17	146068	9172	213
LEAF144a	142	144675	939	2415
LEAF144b	139	144913	946	2609
PROT144a	66	145065	2063	5624
PROT144b	65	146802	2045	5670
SUMP144a	31	145658	4798	2763
SUMP144b	31	145664	4640	2462
TEDS144a	66	145742	2254	317
TEDS144b	74	144452	1861	338
UGAR144a	143	144023	800	3090
UGAR144b	112	144058	1072	3135
USCR144a	161	144432	631	2890
USCR144b	152	144062	627	2788
WC	49	145492	2704	491

In the next phase, all 12 pairwise corpus comparisons were conducted using the A samples. In each case, only those p-frames that occurred in both corpora were considered. This tended to exclude idiosyncratic p-frames such as *remain * seized of* but retain p-frames such as *the * of the* which might be said to be characteristic of the English language in general rather than of a particular genre or register. Thus we are not so much counting completely prefabricated phrases as contrasting creativity within commonly used constructions.[8]

Having selected the pool of p-frames, each of the 10 productivity indices was computed for each p-frame in both corpora. This produced, for each index, two numeric vectors -- the productivity scores of each p-frame in both corpora. As these scores were matched, a paired t-test could be computed from this data. The value of the t-statistic was used as a measure of the degree of differentiation achieved by the index concerned on the pairing under test. That produced 10 t-scores, which were then ranked. The rank of each

[8] Because the pool of p-frames selected in this manner may differ between different comparisons, it is possible to envisage cases where this procedure could lead to misleading results, especially if the comparison and reference corpora are stylistically very different and the pool of common p-frames is small. However, our results so far tend to indicate that the p-frames actually used are standard 'building blocks' common to most varieties of English.

productivity index provided an indication of how well it had performed in differentiating between the more and the less formulaic member of the pair. Then the ranks were simply aggregated over the 12 comparisons.

Table 8 shows the mean rank of each productivity index resulting from this procedure. Higher ranks indicate more successful discrimination. Highest and lowest scores in each column are in bold type.

Table 8. Mean rankings of 10 productivity indices in discriminating more from less formulaic corpora in 12 comparisons

Productivity Index	**Mean Rank with all slots included (1 to 4)**	**Mean Rank with left-most slot omitted (2 to 4)**	**Mean Rank with both end-slots omitted (2 to 3)**
Balance	4.00	3.67	4.42
Hapaxity	**8.25**	**8.33**	**8.17**
Haprate	7.75	7.67	7.33
HC	7.08	6.75	7.17
HH	**2.17**	**2.42**	**2.50**
HV	7.00	6.08	5.67
Nonfocus	5.50	6.08	5.92
Rent	2.50	2.92	2.83
Simpidx	4.33	4.33	4.50
TTPC	6.42	6.75	6.50

An initial finding is that whether or not either or both end-slots were excluded makes little difference to the ordering of the indices. A more substantive finding is that the measures brought in from outside linguistics, such as HH, Simpson's Index and Relative Entropy fare rather poorly. In the light of these results, it would seem unlikely that they are suitable for this particular purpose. HH and Simpson's Index are monotonically related: both performed poorly in this trial. Both the two indices that performed best in this trial employ *hapax legomena* in some form. We believe that either of these, Hapaxity or Haprate, could be viable indices for the present purpose (and perhaps also Herdan's C and TTPC). In the following sections we concentrate on the highest-ranked index, Hapaxity.

4.2 Macro-productivity of p-frames

Having selected a productivity measure, namely Hapaxity, that performs well in paired comparisons where all we could be sure of was that one member of the pair of test corpora was relatively more pervaded by formulaic language than the other, we can now take a further step and use that index to rank all our subcorpora -- from the most productive (in terms of p-frame contexts) to the least productive, hence by implication from the least to the most formulaic.

With this aim in mind, we took the LOB corpus, the largest, and most deliberately general multi-register text sample in our collection, as a benchmark reference corpus and compared every other corpus against it using Hapaxity on the pool of four-item p-frames which were found in both corpora. In this case we excluded p-frames with the empty slot at the left. Thus, for example, if both * *one of the* and *one of the* * were present, only the latter would be considered in the analyses. This corresponds to considering the results in the middle column of Table 8 as definitive.

Furthermore, we did not use the paired t-statistic as a macro-productivity index, because this takes into account the number of comparisons whereas we required a score that would be comparable across corpus-comparisons with different numbers of p-frames in common. Instead we computed the differences between the Hapaxity scores for each p-frame and divided each of these differences by the standard deviation of the Hapaxity scores of the reference corpus, i.e. by a measure of how much variation might be expected among the productivity scores of the p-frames in the reference corpus. The final score was then the mean of these scaled differences, that is, an average z-score.

In addition, we applied this procedure to the B samples in order to test how consistent it is. We also applied it to three holdout subcorpora whose details can be found in Appendix 1: chapters by Agatha Christie (AC); news items from CORDIS, the EU Commission Community Research & Development Information Service (CORD); and legal regulations

from the JRC Acquis corpus (JRCA). From each of these corpora we took a random subsample of texts totalling approximately 145,000 words selected as described earlier, in section 3.1. Table 9 presents the scores obtained as a result of comparing these test corpora to LOBCORP, a reference corpus, in terms of p-frame productivity as quantified by Hapaxity.

Table 9. Relative p-frame productivity of 10 test corpora

Test Corpus	**Mean scaled difference in Hapaxity**	**Number of p-frames in common**
EW144b	0.7038	54
EW144a	0.6157	52
[AC]	0.2136	45
WC	0.2123	70
TEDS144b	0.1281	54
[CORD]	0.0792	58
TEDS144a	0.0366	53
ACAD144b	-0.0535	54
ACAD144a	-0.8225	50
LEAF144a	-1.3751	38
LEAF144b	-1.4109	41
[JRCA]	-1.4140	51
UGAR144b	-1.7062	49
UGAR144a	-1.7905	48
USCR144b	-1.9917	48
USCR144a	-2.1735	46
SUMP144b	-2.4703	37
SUMP144a	-2.5356	36
PROT144b	-3.4628	18
PROT144a	-4.0737	18

Table 9 lists the corpora in descending order of relative p-frame productivity. The point to note about this procedure is that, apart from the intrusion of the holdout subcorpora which were excluded from the index calibration, it would make no difference to the ordering of the text types whether A or B subsamples were used. In other words, the relative order of the source corpora in terms of p-frame productivity is preserved over both random subsamples. This gives some reassurance that the procedure is stable. The largest difference between A and B subsamples occurred with the ACAD corpus, which could be regarded as containing two different text types, research articles and textbook chapters (see Appendix 1). Even in this case, the corpus ordering was unaffected.

As far as the holdout samples are concerned, the order is what would be expected, given the nature of the registers involved. Fiction by an eminent novelist (AC) is the most productive, while legalistic EU regulations (JRCA) are the least productive -- with the news reports of CORD somewhere in between.

It is perhaps also worth noting that the ranking presented above is very similar to that which would be obtained by ranking the corpora inversely by number of p-frames found therein (Table 7) (Spearman's rho = -0.9053, n=20). In other words, the gross number of p-frames exceeding a threshold frequency in a subcorpus of fixed size is also a reasonable indicator of formulaic language, which is consistent with what previous researchers have assumed. For example, Wray (2008: 102-103) argues that "even for those who question whether frequency determines formulaicity, it would take a strong case to successfully argue that the frequent examples of formulaic sequences are not a good place to start". In a similar vein, Schmitt and Carter (2004: 2) view frequency of sequences of words in texts as a criterion that is particularly often-cited in corpus linguistic research on formulaic language, apart from fixedness and variability (Schmitt 2005: 13).

It may also be interesting to note that, according to Hapaxity, the fiction of Agatha Christie is less productive, at least in the sample from her writings under scrutiny, than that of Edith Wharton, though marginally more productive than the speeches by Winston Churchill. In reality, their scores are close, so the relative positions of these authors in Table 9 might not be preserved if another reference corpus were used. However, the fact that the creative authors are all clearly rated to use p-frame contexts more productively than the five most rigid text samples (both types of UN resolutions, the Summaries of Product Characteristics, the Patient Information Leaflets and the Clinical Trials Protocols) does tend to support this approach to the detection of formulaic language.

The most striking contrast observed relative to the benchmark corpus involved the Clinical Trial Protocols (PROT). More specifically, one could say that the p-frames shared by PROT and LOBCORP corpus were between 3.46 and 4.07 standard deviations less productive in the former than in the latter corpus. This statement is elaborated in the next section, aimed at identification of those p-frames that contributed the most to this contrast.

4.3 Micro-productivity of p-frames

As mentioned earlier, it is natural to wonder which linguistic elements (i.e. which particular p-frames) are contributing to the ratings summarized in Table 9. As an example of how this issue might be elucidated, we consider the comparison with the largest divergence versus LOBCORP, which also involves the fewest p-frames, thus is easiest to summarize.

Table 10 lists the 18 p-frames common to PROT and LOBCORP, ranked from the smallest to the largest scaled difference between their Hapaxities. It can be seen that even in the context of phraseological patterns of English which are commonplace rather than topic-specific, the Clinical Trial Protocols show much less variety than does the generic multi-register LOB corpus. In fact, the least productive of the shared p-frames in LOBCORP (*the end of* *) is more productive than all but four of the shared p-frames in PROT (*in the * of, .the * of the, for the * of, with the * of*).

Table 10. Example Corpus Comparison between PROT and LOBCORP, showing which p-frames contribute how much to the aggregate productivity score

Rank	**P-frame**	**Hapaxity in PROT**	**Hapaxity in LOBCORP**	**Difference**	**Scaled Difference (SD = 0.2807456)**
1	*the end of *	-0.91111	-0.40000	-0.51111	-1.820545
2	*the * of the*	-0.28980	0.25953	-0.54933	-1.956683
3	*in the * of*	-0.39394	0.22798	-0.62192	-2.215244
4	*for the * of*	-0.28571	0.47250	-0.75821	-2.700701
5	*with the * of*	-0.25000	0.53801	-0.78801	-2.806847
6	*of the * of*	-0.41096	0.49810	-0.90906	-3.238021
7	*of * of the*	-0.68182	0.31897	-1.00079	-3.564758

8	*with a * of*	-0.52941	0.57983	-1.10924	-3.951051
9	*it is not **	-0.94118	0.25153	-1.19271	-4.248366
10	*to be * in*	-0.74675	0.55319	-1.29994	-4.630313
11	*be * in the*	-0.74434	0.55932	-1.30366	-4.643564
12	*the * in the*	-0.73077	0.66102	-1.39179	-4.957478
13	*of the * is*	-0.70588	0.70213	-1.40801	-5.015253
14	*of the * to*	-0.64607	0.78808	-1.43415	-5.108362
15	*of the * and*	-0.74359	0.69395	-1.43754	-5.120437
16	*end of the **	-0.93000	0.53788	-1.46788	-5.228506
17	*was * in the*	-1.00000	0.69106	-1.69106	-6.023461
18	*of the * the*	-0.95000	0.76172	-1.71172	-6.097051

The p-frame with the sharpest contrast is *of the * the*, which has a high variety of slot-fillers in the LOB Corpus, but virtually none in the PROT corpus. The difference is illustrated below by listing the variants of this p-frame from the A subsample of PROT with those of a subsample of comparable size from LOBCORP.

PROT (random subsample of 66 texts, 145065 tokens) :

```
of the * the      40    2
of the trial the  38
of the investigator the 2
```

LOBCORP (random subsample of 71 texts, 144546 tokens) :

```
of the * the      33    31
of the peace the  2
of the bible the  2
of the wind the   1
of the white the  1
of the timetable the    1
of the river the  1
of the region the 1
of the redeemer the     1
of the range the  1
of the path the   1
of the part the   1
of the neck the   1
of the mind the   1
of the men the    1
of the kgb the    1
of the island the 1
of the image the  1
of the home the   1
of the hair the   1
of the girls the  1
of the face the   1
of the evangelists the  1
of the emotions the     1
```

```
of the discontent the    1
of the decisions the     1
of the costumes the      1
of the confusion the     1
of the city the    1
of the century the       1
of the cases the   1
of the age the     1
```

Such listings provide some relevant information concerning the contrast, but tell us very little about the contexts in which this p-frame is found. For this reason, we have written a program in Python3 to produce PFIC listings (p-frames in context). This program accepts as input a number of p-frames and shows their local contexts.

The p-frame contexts below illustrate how the p-frame *of the * the* is dominated in the PROT subcorpus by just two slot-fillers, one of which, the noun *trial*, accounts for 38 of the 40 cases.

```
doctype : protocol (PROT)
of_the_*_the

ctps033.txt
0004040: the trial e 2.1 main objective of the trial the trial investigates the bene
0004655: nts e 2.2 secondary objectives of the trial the main secondary outcome meas
ctps034.txt
0006227: the trial e 2.1 main objective of the trial the main objective of the trial
ctps036.txt
0000422:  2005-004508-35 a 3 full title of the trial the use of pet ct scanning to a
ctps042.txt
0015547: the trial e 2.1 main objective of the trial the objective of this study is
ctps046.txt
0004010: the trial e 2.1 main objective of the trial the objective of this clinical
ctps056.txt
0000427:  2006-000082-10 a 3 full title of the trial the efficacy of oral transmucos
ctps058.txt
0003834: the trial e 2.1 main objective of the trial the purpose of focus is to eval
ctps066.txt
0003727: the trial e 2.1 main objective of the trial the primary endpoint will be th
ctps067.txt
0003703: the trial e 2.1 main objective of the trial the aim of this trial is to det
0003926: nal e 2.2 secondary objectives of the trial the trial also aims to collect
ctps085.txt
0000442:  2007-000940-28 a 3 full title of the trial the rimonabant in treatment of
0004140: the trial e 2.1 main objective of the trial the primary outcome will be the
ctps090.txt
0000434:  2007-002293-54 a 3 full title of the trial the apex trial effects of allop
0003802: the trial e 2.1 main objective of the trial the primary objective of this s
0004025: e x e 2.2 secondary objectives of the trial the secondary objectives of thi
ctps093.txt
0004101: the trial e 2.1 main objective of the trial the primary objective will be t
ctps106.txt
0004883: the trial e 2.1 main objective of the trial the primary aim of this study i
ctps114.txt
0006230: the trial e 2.1 main objective of the trial the primary objective of this s
0006800: sed e 2.2 secondary objectives of the trial the secondary objective of this
ctps135.txt
0004106: the trial e 2.1 main objective of the trial the principal question this pil
ctps140.txt
0004031: lly e 2.2 secondary objectives of the trial the visible drain blood loss fi
ctps146.txt
```

```
0008706: ine drug screen in the opinion of the investigator the investigator may fol
0011068:  of 4 weeks or in the judgment of the investigator the subject meets criter
0004073: the trial e 2.1 main objective of the trial the primary objective of this s
0004565: nse e 2.2 secondary objectives of the trial the secondary objectives of the
ctps147.txt
0000436:  2009-013226-17 a 3 full title of the trial the enigma ii trial nitrous oxi
ctps155.txt
0000427:  2009-015903-22 a 3 full title of the trial the effect of eicosapentaenoic
0004070: the trial e 2.1 main objective of the trial the study aims to show the firs
ctps167.txt
0003842: the trial e 2.1 main objective of the trial the primary objective of this s
ctps171.txt
0000425:  2007-005534-36 a 3 full title of the trial the effect of exenatide on sati
ctps172.txt
0004806: the trial e 2.1 main objective of the trial the objective of this study is
ctps186.txt
0000421:  2008-002336-15 a 3 full title of the trial the effect of prednisolone vers
0011749: one e 2.2 secondary objectives of the trial the secondary objectives of thi
ctps192.txt
0008949: the trial e 2.1 main objective of the trial the primary objective of this s
0009182: ebo e 2.2 secondary objectives of the trial the secondary objectives are to
ctps197.txt
0006141: the trial e 2.1 main objective of the trial the trial is conducted in two p
ctps218.txt
0004052: the trial e 2.1 main objective of the trial the objectives of this study ar
ctps229.txt
0004187: the trial e 2.1 main objective of the trial the primary aim of the scratch
ctps240.txt
0008134: the trial e 2.1 main objective of the trial the objective of this trial is
```

By contrast, results from the LOBCORP subsample are shown below. Only 2 of the 31 slot-fillers are repeated more than once.

```
doctype : lobcorp (LOBCORP)
of_the_*_the

LA02.txt
0002959:  akoradi  for  the  past  week  root  of  the  discontent  the  austerity  budget  inclu
0006373:  eace  and  for  inciting  a  breach  of  the  peace  the  summonses  say  they  are  like
0011008:  prisoned  for  inciting  a  breach  of  the  peace  the  committee's  president  89-  y
LA23.txt
0002011:  st  11  lb  3  lb  below  the  middle  of  the  range  the  handicapper  has  certainly  t
LA24.txt
0003799:  ee  how  it  can  be  all  the  fault  of  the  girls  the  secretary  of  great  yarmouth
LA27.txt
0007796:  ains  was  extended  to  the  whole  of  the  island  the  measures  had  previously  ap
LB06.txt
0005825:  nal  issue  by  misrepresentation  of  the  decisions  the  introduction  of  a  pseud
LC04.txt
0009864:  was  a  splendid  interpretation  of  the  part  the  rest  of  the  cast  were  well  c
LC12.txt
0007809:  emes  reflect  the  controversies  of  the  age  the  quebec  act  with  its  threat  of
LD06.txt
0009316:  o  friday  were  days  one  to  five  of  the  timetable  the  following  monday  was  da
LD15.txt
0002448:  among  the  english  translations  of  the  bible  the  design  draws  attention  to  s
0003851:  among  the  english  translations  of  the  bible  the  commemoration  editions  will
0000884:  r's  guilt  by  the  atoning  blood  of  the  redeemer  the  lord's  anointed  mercy  is
LE09.txt
0009692:  first  cave  rushes  a  large  part  of  the  river  the  second  penetrates  under  the
LE10.txt
0002929:  hange  in  the  wind  into  the  eye  of  the  wind  the  fantail  consists  of  what  is
LE34.txt
0010404:  bly  according  to  the  condition  of  the  hair  the  most  common  change  of  textur
LF13.txt
0008350:  most  sensational  murder  trials  of  the  century  the  defence  had  picked  lawren
LG05.txt
0008642:  ision  of  aim  at  the  very  heart  of  the  confusion  the  resolute  but  unbroken  g
LG13.txt
0006692:  war  were  in  his  opinion  things  of  the  mind  the  real  task  was  to  get  better
LG27.txt
0000267:  in  her  stead  undisputed  queen  of  the  home  the  children  and  all  official  so
```

```
LG43.txt
0004431:  als and the cutting and making of the costumes the opera producer is called
LG55.txt
0003743:  worked upon now comes the turn of the emotions the object of study is now t
0011588:  te in their operation the goal of the kgb the present designation for the r
LG61.txt
0004710:  be achieved by a modification of the image the simplified picture which go
LJ03.txt
0005837:  soil formation characteristic of the region the two groups of soils exempl
LJ17.txt
0003312:  pper chest wall and right side of the neck the following day he complained
LJ64.txt
0007107:  ge and was blinded in a street of the city the two sisters who had little l
0011068:  seems coarser in the portrayal of the evangelists the bodies tend to disint
0008767:  s the broadening of the planes of the face the empress and her consort are
LK14.txt
0005819:  m and she sat down on the side of the white the rumanian introduced her to
LL23.txt
0003317:  ee how this would help in most of the cases the case up that afternoon had
0009776:  ige carpet like a continuation of the path the hall stand held one umbrella
LN17.txt
0009991:  the registered stuff said one of the men the man with the keys jerked his
```

It is clear from the above listings that the p-frame in question (i.e. *of the * the*) is used in very different ways in the two corpora. Purists may worry that the great majority of examples of this particular p-frame cross over a syntactic boundary, but it nevertheless defines a context where one subcorpus exhibits much less flexibility than the other. After all, even in the context of a sentence ending with "*of the* <noun>.", followed by the definite article, there are many ways to instantiate the noun in question, as shown by the LOBCORP examples.

It is perhaps worth noting here that in this study we ignore the problem of homonymy or polysemy of p-frames. This is consistent with our main objective of quantifying overall indicators of formulaic language rather than in-depth linguistic analyses of particular patterns. Although the slot-fillers are usually semantically-related, it is also possible for them to carry various meanings and hence have different discourse functions in texts. For example, in Clinical Trial Protocols the p-frame *the * of the* is actualized as *the start/end/duration/date of the* [*trial*] or *the safety/effect/efficacy of the* [*trial*], among others. In fact, *start* and *end* are temporal markers while *safety* or *efficacy* refer to abstract properties of clinical trials. Yet in the ranking of corpora in terms of the macro-productivity of p-frames, our point of departure

is the p-frames' orthographic form rather than their meaning or discourse function in a larger context or co-text.

5. Discussion and conclusions

Focusing on ways of detecting traces of formulaic language and quantifying its prevalence in texts, this paper aimed to evaluate a number of computable indices of linguistic flexibility applied to p-frames (Fletcher 2002-2007) based on contiguous sequences of four words. In this paper, p-frames have been considered as a means of tapping into important aspects of formulaic language, that is, the frequency of occurrence as well the degree of fixedness and pattern variability (Schmitt & Carter 2004; Schmitt 2005; Wray 2009).

The study was conducted in a number of stages. Firstly, 12 calibration comparisons were conducted to identify an optimum index, out of 10 analyzed, for differentiating between more and less formulaic p-frames in individual pairs of corpora under scrutiny. The procedure revealed a 'winning' index, namely Hapaxity. Having established this, Hapaxity was further used to rank all corpora, from the most productive in terms of p-frame contexts (literary compositions by Edith Wharton) to the least productive (Clinical Trial Protocols), using the LOB Corpus (Hofland & Johansson 1982) as a benchmark -- hence, by implication, ranking the corpora from the least to the most formulaic. Finally, in order to identify particular p-frames that contribute the most to the ranking (and, by implication, to the formulaic nature of particular corpora), we compared the differential Hapaxity scores of shared p-frames in individual pairs of corpora under scrutiny. As a case in point, we ranked p-frames found in both the LOB corpus and Clinical Trial Protocols (PROT) from the smallest to the largest scaled difference between their Hapaxities. Thus, at this stage of the study, we focused on pattern variability of individual p-frames, which we call micro-productivity.

To conclude, a study involving thirteen corpora in a single language, English, can only be regarded as provisional. Nevertheless, we have extended the range of application of the p-frame concept and acquired valuable empirical information concerning which indices of p-frame productivity are likely to prove useful. We have also shown that, starting with information merely that one of a pair of corpora is more formulaic than another, we can use an index of individual p-frame productivity to give scores on a common scale to different text types and thus arrive at a relative ranking of several corpora, representing a range of different text registers from creative to routinized, in terms of p-frame productivity.

Moreover, despite the fact that formulaicity is a nebulous concept, with no rigorous definition agreed by linguistic researchers, this study has introduced and tested a novel index, Hapaxity, which our results suggest could provide a relatively consistent means of quantifying an important aspect of formulaic language. More, of course, remains to be done in evaluating its usefulness, for instance by applying it to languages other than English. Moreover, the procedure of cutting all corpora down to the length of the smallest to eliminate size effects, as used in this study, is inconvenient and could potentially lead to information loss; therefore an investigation to determine precisely how Hapaxity scores vary with corpus size would be a natural next stage of research. This could also consider a few of the other indices, such as Haprate, that also performed well in our calibration trial (Table 8). Nevertheless, we believe that p-frame Hapaxity has already shown its worth as an addition to the toolkit of researchers exploring formulaic language. More generally, it could, like the other indices, be applied in contexts other than those defined by p-frames; for example, in cases where corpora have been parsed or tagged to indicate more linguistically grounded phraseological units.

From a technical perspective, it would also be possible, in principle, to reengineer our software to utilize information on frequency of occurrence of slot-fillers used in contexts outside particular p-frames, a feature currently offered neither by kfNgram (Fletcher 2002-

2007) nor AntConc (Anthony 2014), thus allowing statistics related to strength of association between the slot-fillers and invariant components of the p-frames to be computed. Another avenue for future development would be to use this methodology to compare individual documents with a base corpus, though whether this particular approach would generalize to texts of only a few thousands or even hundreds of words is an open question.

Regarding other future avenues, the methods used in this study as well as its results open up a wide range of applications in linguistics, notably in research on phraseology, stylistics, translation or English language teaching, to name but a few. For example, our methods may help one explore specific phraseological differences between different registers. Notably, identification of a corpus with the most productive p-frames or identification of the most (or the least) productive p-frames within a corpus represents a promising starting point for exploration of phraseological variety of various text types, e.g. legal, medical, academic or literary. This may also offer an opportunity for stylistic studies, notably in terms of comparisons of phraseologies used by different authors or deemed typical of particular literary genres. Furthermore, this approach could be applied to research on translation universals, notably T-universals (Chesterman 2004): a comparison of micro-productivity and contexts of use of shared p-frames across non-translational and translational texts produced in the same language seems to be particularly attractive for testing the simplification or the levelling-out hypotheses (Baker 1995, 1996). Finally, as regards English language teaching, in particular teaching ESP, our methods may help select from a wide variety of phraseological patterns those p-frames which potentially contribute the most to formulaic nature of particular text types, registers or genres. Such information may be valuable for pedagogical purposes.

All in all, it is hoped that this paper will help other researchers interested in exploring variability of phraseological patterns and/or formulaic language in texts, notably when using p-frames as the unit of analysis.

Acknowledgements

We wish to thank two anonymous reviewers for helpful comments on an earlier draft. We are also indebted to several different corpus compilers, as indicated in Appendix 1.

References

Altenberg, B. 1998. "On the phraseology of spoken English: The evidence of recurrent word combinations". In A. Cowie (ed.), *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press. 101–122.

Anthony, L. 2014. *AntConc (ver. 3.4.1)*. Available at: http://www.antlab.sci.waseda.ac.jp/software/antconc3.2.4w.exe (accessed February 2014).

Baayen, H. 1992. "Quantitative aspects of morphological productivity". In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1991*. Berlin: Springer. 109–149.

Baker, M. 1995. "Corpora in translation studies: An overview and some suggestions for future research". *Target*, 7 (2), 223–243.

Baker, M. 1996. "Corpus-based translation studies: The challenges that lie ahead". In H. Somers (Ed.), *Terminology, LSP and Translation: Studies in Language Engineering. In Honour of Juan C. Sager*. Amsterdam: John Benjamins, 175–186.

Bartmiński, J. 2007. "Stereotyp jako przedmiot lingwistyki". In: J. Bartmiński (ed.), *Stereotypy mieszkają w języku*. Lublin: Wydawnictwo UMCS. 53–71.

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 (3), 275–311.

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

Bouayad-Agha, N. and Kilgarriff, A. 1999. “Duplication in Corpora” In *Proceedings of the 2nd CLUK Colloquium.* Colchester, Essex, 11-12 Jan 1999. Available at: http://www.kilgarriff.co.uk/Publications/1999-BouayadAghaKilg-CLUK.pdf (accessed June 2012).

Bouayad-Agha, N. 2006. The Patient Information Leaflet (PIL) 2.0 corpus. Available at: http://mcs.open.ac.uk/nlg/old_projects/pills/corpus/PIL/ (accessed May 2012).

Chen, Y.-H. and Baker, P. 2010. “Lexical bundles in L1 and L2 academic writing”. *Language Learning and Technology*, 14 (2), 30–49.

Chesterman, A. 2004. “Hypothesis about translation universals”. In G. Hansen, K. Malmkjaer & D. Gile (eds.), *Claims, Changes and Challenges in Translation Studies*. Amsterdam: John Benjamins. 1-13.

Chlebda, W. 2003. *Elementy frazematyki: wprowadzenie do frazeologii nadawcy*. Łask: Leksem.

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

Corrigan, R., Moravcsik, E., Ouali, H. and Wheatley, K. (eds.) 2009. *Formulaic Language*. *Vol. 1. Distribution and historical change*. Amsterdam: John Benjamins.

Eggins. S. 1994. *An Introduction to Systemic Functional Linguistics*. London: Pinter.

Ellis, N.C., Roemer, U., Brook O’Donnell, M., Gries, S. & Wulff, S. 2009. “Measuring the formulaicity of language”. Paper presented at colloquium *SLA and the inseparability o*f *vocabulary and syntax*. Denver, Colorado, 21-24 Mar 2009. Available at: http://researchers.tistory.com/attachment/cfile7.uf@154A94334F197CE02E3025.pdf (accessed January 2012).

Erman, B. and Warren, B. 2000. “The idiom principle and the open choice principle”. *Text*, 20 (1), 29–62 (cited in Schmitt & Carter 2004: 1).

Firth, J.R. 1968. “Linguistics and Translation”. In: F. Palmer (ed.), *Selected Papers of J. R. Firth 1952-1959*, London: Longman, 84-95 (cited in Toolan 1996: 161–162 and Roemer 2010: 96).

Fletcher, W. 2002-2007. *KfNgram*. Annapolis, MD: USNA. Available at: http://www.kwicfinder.com/kfNgram/kfNgramHelp.html (accessed November 2011).

Foster, P. 2001. “Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers”. In: M. Bygate, P. Skehan, & M. Swain (eds.), *Researching pedagogic tasks: Second language learning, teaching, and testing*. Harlow: Longman. 75–93. (cited in Schmitt 2005: 14).

Fuster-Marquez, M. 2014. “Lexical bundles and phrase frames in the language of hotel websites”. *English Text Construction*, 7 (1), 84–121.

Gerbig, A. 2011. “Key words and key phrases in a corpus of travel writing”. In: M. Bondi & M. Scott (eds.), *Keyness in Texts*. Amsterdam: John Benjamins. 147–168.

Grabowski, Ł. 2015. “Keywords and lexical bundles within English pharmaceutical discourse: a corpus-driven description”. *English for Specific Purposes*, 38: 23–33.

Grabowski, Ł. 2015b. *Phraseology in English pharmaceutical discourse: A corpus driven study of register variation*. Opole: Wydawnictwo Uniwersytetu Opolskiego.

Gray, B. and Biber, D. 2013. “Lexical frames in academic prose and conversation”. *International Journal of Corpus Linguistics*, 18 (1), 109–135.

Halliday, M.A.K. 2014. That “certain cut”: towards a characterology of Mandarin Chinese. *Functional Linguistics*, 1(2), doi:10.1186/2196-419X-1-2 .

Herdan, G. 1964. *Quantitative Linguistics*. London: Butterworth.

Hirschman, A. 1964. “The Paternity of an Index”. *The American Economic Review* (American Economic Association), 54 (5), 761.

Hofland, K. and Johansson, S. 1982. *Word frequencies in British and American English.* Bergen: Norwegian Computing Centre for the Humanities/London: Longman.

Hyland, K. 2008. “As can be seen: Lexical bundles and disciplinary variation”. *English for Specific Purposes*, 27, 4–21.

Kilgarriff, A. 2005. “Language is never ever ever random”. *Corpus Linguistics and Linguistic Theory*, 1 (2), 263–276.

Kuiper, K. 1996. *Smooth Talkers*: *The Linguistic Performance of Auctioneers and Sportscasters*. New York: Erlbaum (also cited in Wray 2002: 17).

Lancioni, G. 2009. “Formulaic models and formulaicity in Classical and Modern Standard Arabic”. In: R. Corrigan, E. Moravcsik, H. Ouali & K. Wheatley (eds.), *Formulaic Language*. *Vol. 1. Distribution and historical change*. Amsterdam: John Benjamins. 219–238.

Pawley, A. 2009. “Grammarians’ languages versus humanists’ languages and the place of speech act formula in models of linguistic competence.” In: R. Corrigan, E. Moravcsik, H. Ouali & K. Wheatley (eds.), *Formulaic Language*. *Vol. 1. Distribution and historical change*. Amsterdam: John Benjamins. 3–26.

Permyakov, G. 1970. *От поговорки до сказки (Заметки по общей теории клише)* [From Sayings to Fairytales. Notes on General Cliché Theory]. Moscow: Nauka (cited in Chlebda 2003: 25).

Ren, Z, Lu,Y., Cao, J., Liu, Q. & Huang, Y. 2009. “Improving Statistical Machine Translation Using Domain Bilingual Multiword Expressions.” *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications. MWE’ 09*, 47–54. Stroudsburg: Association for Computational Linguistics. Available at: http://www.aclweb.org/anthology/W09-2907 (accessed November 2014).

Roemer, U. 2009. “English in Academia: Does Nativeness Matter?”. *Anglistik: International Journal of English Studies*, 20 (2), 89–100.

Roemer, U. 2010. “Establishing the phraseological profile of a text type. The construction of meaning in academic book reviews”. *English Text Construction*, 3 (1), 95–119.

Schmitt, N. 2005. “Formulaic language: fixed and varied. Estudios de linguistica aplicada (ELIA), 6, 13–39. Available at: http://institucional.us.es/revistas/elia/6/art.2.pdf (accessed November 2014).

Schmitt, N. and Carter, R. 2004. “Formulaic sequences in action: An introduction”. In N. Schmitt (ed.), *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins, 1–22.

Shannon, Claude E. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3): 379–405.

Simpson, E.H. 1949. Measurement of diversity. *Nature*, 163, 688.

Simpson-Vlach, R. and Ellis, N.C. 2010. An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, 31 (4): 487–512.

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

Tiedemann, J. 2009. “News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces”. In: N. Nicolov, K. Bontcheva, G. Angelova and R. Mitkov (eds.), *Recent Advances in Natural Language Processing*, 5. Amsterdam/Philadelphia: John Benjamins. 237–248.

Underwood, G., Schmitt, N. and Galpin, A. 2004. “The eyes have it: An eye-movement study into the processing of formulaic sequences”. In: N. Schmitt (ed.), *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins. 153–172.

Upton, G. & Cook, I. 2006. *Oxford Dictionary of Statistics*. Oxford: Oxford University Press.

Wood, D. (ed.) 2010a. *Perspectives on Formulaic Language: Acquisition and Communication*. London: Continuum.

Wood, D. (ed.) 2010b. *Formulaic Language and Second Language Speech Fluency. Background, Evidence and Classroom Applications*. London: Continuum.

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

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

Wray, A. 2009. "Identifying formulaic language. Persistent challenges and new opportunities". In: R. Corrigan, E. Moravcsik, H. Ouali & K. Wheatley (eds.). *Formulaic Language. Vol. 1. Distribution and historical change*. Amsterdam: John Benjamins. 27–51.

Wray, A. & Perkins, M. 2000. The functions of formulaic language: an integrated model. *Language & Communication*, 20, 1–28.

Appendix 1. Details of corpora used in the study, supplementary to data in Tables 3 & 4.

Relatively formulaic corpora

ACAD.
Twenty-six of the documents in the ACAD Corpus consisted of research articles. More specifically, 4 texts were extracted from Volume 3 Issue 2 (dated February 2012) of the *International Journal of Pharmaceutical Sciences and Research* (IJPSR, available at www.ijpsr.com), 3 texts were extracted from Volume 2 Issue 1 (1 text), dated January-March 2012, and Volume 1 Issue 4 (2 texts), dated October-December 2011, of *The International Journal of Pharmacy and Biological Sciences* (IJPBS, available at www.ijpbs.com), and 15 texts were extracted from the following edited volume: Piscitelli, S, Rodvold, K. (Eds.) (2005). *Drug Interactions in Infectious Diseases*: 2nd Edition. Totowa, NJ: Humana Press. The twenty-five book chapters were extracted from the following textbooks: Bauer, L. (2008). *Applied Clinical Pharmacokinetics*. 2nd Edition. New York: McGraw-Hill Medical (5 chapters from Part I and Part II); Hollinger, M. (2003). *Introduction to Pharmacology*. 2nd Edition. London/New York: Taylor & Francis (13 chapters); Craig, Ch. & Stitzel, R. (2004). *Modern Pharmacology with Clinical Applications*. 6th Edition. Lippincott: Williams & Wilkins (7 chapters).

LEAF.
The patient information leaflets (LEAF) were extracted from the Patient Information Leaflet Corpus 2.0, originally compiled at the Natural Language Technology Group at the University of Brighton and discussed in greater detail by Buoayad-Agha and Kilgarriff (1999) and Buoayad-Agha (2006). This corpus is available at http://www.mcs.open.ac.uk/nlg/old_projects/pills/corpus . It contains 465 texts but four were excluded from this study as near-duplicates.

PROT.
The Clinical Trial Protocols (PROT) were downloaded from the Clinical Trials Register (CTR) database of the European Union. This database is hosted by the European Medicines Agency (EMA) and readily available at https://www.clinicaltrialsregister.eu/index.html .

SUMP.
The Summaries of Product Characteristics (SUMP corpus), were downloaded from the Open Source Parallel Corpus (OPUS) Project website at http://opus.lingfil.uu.se/EMEA.php (Tiedemann 2009).

UGAR.
These 672 documents are resolutions of the United Nations General Assembly sessions 54 to 57, four dated 1999 and the rest from the years 2000-2003. They were extracted from the publications of the United Nations, collated by DFKI GmbH (Deutsches Forschungszentrum für Künstliche Intelligenz), available from www.euromatrixplus.eu/multi-UN/. A total of 720 documents were extracted but 38 excluded as duplicates or near-duplicates. (The General Assembly reissues near-identical versions of earlier resolutions from time to time.) A further 10 documents of less than 100 words in length were also excluded.

USCR.
This corpus contains 274 resolutions passed by the United Nations Security Council in the period 2000-2004. The texts were extracted from the publications of the United Nations, collated by DFKI GmbH (Deutsches Forschungszentrum für Künstliche Intelligenz) www.dfki.de , available from www.euromatrixplus.eu/multi-UN/. With USCR, 3 texts were excluded as duplicates as well as 11 for being less than 100 words in length. The USCR texts cover a slightly longer period since fewer Security Council resolutions are issued each year.

Relatively creative corpora

EW.
These 44 stories have been collected by first author over a number of years and are not currently publicly available, although most of the tales by Edith Wharton can be found at http://public.wsu.edu/~campbelld/wharton/shortstories.htm and/or at http://www.gutenberg.org/browse/authors/w (project Gutenberg).

TEDS.
1555 transcripts in English of talks given as part of the TED initiative (www.ted.com). Obtained from collection held at WIT3 website https://wit3.fbk.eu .

WC.
The texts by Churchill (WC) cover a period of over 60 years and thus deal with a wide range of topics. They all have a political or historical focus, but as Churchill was a Nobel laureate in literature we assume his command of English was more creative than average. His Nobel-prize acceptance speech is also included in the sample. The texts comprise 45 speeches by Winston Churchill, sourced from http://www.winstonchurchill.org/learn/speeches/speeches-of-winston-churchill as well as 2 individual chapters and 2 prefaces from his four-volume biography of Marlborough.

LOBCORP.
The LOB Corpus (Hofland & Johansson 1982) is a generic corpus covering several registers and thus can safely be presumed to include more linguistic variety than any of the narrowly specialized corpora.

Holdout corpora

AC.
65 chapters from 53 novels by Agatha Christie, collected by the first author, comprising 176,975 tokens in total. A random subsample of 51 chapters totalling 146,068 words was used in these experiments. Not publicly available.

CORD.
5,756 documents totalling 2,649,108 tokens taken from the EU Commission Community Research & Development Information Service (CORDIS). Obtained from
http://psi.amu.edu.pl/en/index.php?title=Parallel_Corpora
by selecting all texts of 200 words or more from the English collection. A random subsample of 311 documents comprising 144,308 words was used in the experiments described.

JRCA.
JRC-Acquis is a collection of legislative text of the European Union and currently comprises selected texts written between the 1950s and now. Original source:
https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis
The Acquis Communautaire (AC) is the total body of European Union (EU) law applicable in the EU Member States. The portion used here was the English-language section of the Greek-English JRC-Acquis parallel corpus, obtainable from
http://opus.lingfil.uu.se/JRC-Acquis.php
A random subsample of 58 agreements, conventions, directives and other legal texts from this corpus, comprising 145,241 words, was used in the experiment reported.