Claudia Lückert

The lexical profile of modern American proverbs: Detecting contextually-predictable keywords in a database of American English proverbs

Abstract: Proverbs (as *Time is money*) are conventionalized expressions that are "equivalent to a sentence" and "express generalized experiences or value judgements" (Steyer 2015: 209–210). This study aims at describing the lexical structure of the proverb inventory and at identifying 'proverbial keywords' which may be assumed to play an important role in storing proverbs in the mind (Lückert 2018). A database of American English proverbs was compiled and content words were tested for their contextual predictability (with *COCA* as 'normative corpus' using a 'goodness-of-fit' test). The results suggest that 59.4 % of the word lemmas are significantly over-represented in the proverb corpus (i.e. 'keywords'). This finding underpins the assumption that a considerable number of words are strongly associated with the proverb as a category. In the experimental part of the project it was tested whether the 'keywords' differ in how well they contribute to a strengthened memory representation of proverbs.

Keywords: contextual predictability, *Corpus of American English Proverbs* (*CAEP*), lexical profile, proverbial keywords

1 Introduction: Familiarity vs. infrequent use

Proverbs – for example *No pain, no gain* – are conventionalized, reproducible multi-word expressions that are "equivalent to a sentence", "have become part of everyday language as an expression of worldly wisdom" and "express generalized experiences or value judgements" (Steyer 2015: 209–210; also Mieder 2007). In research, proverbs have been studied from diverse perspectives ranging from pragmatic and cultural aspects (for example their use in literary works and in the media; see overview in Mieder [2009: 240–241]) to cognitive aspects (in past research, usually comprehension with a focus on metaphors, see Honeck [1997]; Lewandowska and Antos [2015]).

Claudia Lückert, Westfälische Wilhelms-Universität Münster, Englisches Seminar, Johannisstr. 12–20, 48143 Münster, Germany, c.lueckert@uni-muenster.de

The rationale of the present paper is to shed light on the lexical structure of the proverb inventory and to identify 'proverbial keywords' which may be assumed to play an important role in storing proverbs in the mind (Lückert 2018).[1] When compared to other types of fixed expressions proverbs are usually characterized by a higher degree of linguistic complexity (Mac Coinnigh 2015) and a generally low token frequency (Arnaud and Moon 1993; Moon 1998; Grzybek 2012). At the same time, however, infrequent proverbs can still be very familiar to speakers (Arnaud and Moon 1993; Grzybek 2012). This apparent discrepancy can possibly be resolved by an assumed "back-up system" in the mental lexicon (in line with usage-based theory; see Bybee [2013]) that strengthens the memory representation of proverbs (Lückert 2018).

The main aims of this paper are to demonstrate how the lexical profile of a proverb inventory of a given language can be generated using a frequency profiling paradigm (Rayson and Garside 2000) and to discuss the benefits of a rank-ordered list of proverbial keywords based on this proverb inventory. In the project this paper is part of, corpus-based analyses and psycholinguistic tasks are combined to scrutinize the effect of proverbial keywords on memory. In this paper, the focus will be on the corpus-linguistic aspects.

After an overview of the current state of linguistic research on proverbs in Section 2, the corpus-based analyses are reported on in Section 3. Section 3.1 will introduce the database (*CAEP – Corpus of American English Proverbs*) which was compiled for the project so as to make it possible to extract the most frequent words used in proverbs and those that are characterized by a high contextual predictability. The database (28,159 word tokens, 2,562 word types, 4,164 proverb variants [2,191 proverb types]) draws on published material and includes both current corpus data on the frequency of proverbs (from the *Corpus of Contemporary American English* [*COCA*], *Google Books* [American English], *Google Books Ngram Viewer* [filters 'American English' and '1990–2008'; Michel et al. 2010], and the Web using *Google Advanced Search* [filters 'English' and 'United States']) and current experimental data on familiarity (Chlosta and Grzybek 2005; Haas 2008).

In Section 3.2 the method of detecting the keywords is discussed. Section 3.3 is informed by an evaluation of the keywords. This section highlights exemplarily in which form we may benefit from a lexical profile of a given proverb inventory – e.g. against the background of 'frame semantics' (see Gawron 2011) with respect

[1] This study is part of a research project funded by the German Research Foundation (DFG): DFG LU 1942/2-1 (grant for 2015–2019). 'Proverbial keywords' in this study are content words which have been identified as strongly associated with the proverb context by corpus-linguistic analyses (see Section 3.2).

to how particular semantic frames are realized in form of lexical material which is more or less typical of the proverb as a class and from learning what is not represented in the inventory (e.g. which semantic frames are fragmentary in or absent from proverbs). Finally, in Section 4, the central findings of the present paper are summed up and aspects for future research are pointed out.

2 Proverbs in linguistic research

The study of proverbs has a long tradition reaching back to classical times (Mieder 2007: 394). Traditionally, scholars from the areas of literary and cultural studies have been involved in paremiological research. Linguistic perspectives have, however, increasingly gained importance from the twentieth century on. In the past few decades, there has been a trend towards new lines of research which stress the use of empirical, quantitative methods (Chlosta and Grzybek 2005; Grzybek 2012; Steyer 2015 etc.).

In spite of the long tradition of proverb scholarship, the problem of how to define the proverb is anything but resolved (Mieder 2007; Mac Coinnigh 2015). What is more, it has been noted that language users' intuitions about the proverb as a category do not necessarily correspond to scientific classification systems. On the contrary, language users do not seem to distinguish between proverbs and, for example, slogans or mottos – they appear to rely on "the concept *sentence as message*" (Steyer 2015: 223).

But what, then, makes a proverb a proverb? We may here consider a number of observations from influential proverb scholars. Firstly, Mieder describes the proverb as a class along the lines of a functional perspective: "Proverbs fulfill the human need to summarize experiences and observations into nuggets of wisdom that provide ready-made comments on personal relationships and social affairs. There are proverbs for every imaginable context, and they are thus as contradictory as life itself" (2007: 394).

Secondly, Steyer names a structural characteristic ("equivalent to a sentence") next to more functionally motivated features ("have become part of everyday language as an expression of worldly wisdom" and "express generalized experiences or value judgements") (2015: 209–210). And thirdly, Arora (1984) has looked into particular poetic and structural features that tend to appear in proverbs – these 'proverbial markers' include parallelism, ellipsis, and alliteration (also Mac Coinnigh 2015).

Proverbs are lexicalized, reproducible multi-word units. As pointed out in the introduction, they are, however, distinct from other types of multi-word expressions (such as conventionalized *as*-similes, collocations, or phrasal verbs etc.)

in that proverbs (a) tend to be more complex in their linguistic structure (Mac Coinnigh 2015), (b) tend to occur much more infrequently in language use (on the basis of corpus data, Arnaud and Moon 1993; Moon 1998; Grzybek 2012), and (c) can still appear very familiar to speakers even if the proverbs have a very low token frequency (Arnaud and Moon 1993; Grzybek 2012).

With regard to the characteristic of 'linguistic complexity' we may first outline generally what is understood by this term. Gries for example distinguishes between units of differing complexity (not restricted to reproducible multi-word units, though) and names items which range from "low-complexity/abstractness cases (morphemes, monomorphemic words) via polymorphemic words, *fully-fixed multi-word expressions*, and partially filled multi-word expressions to high-complexity/abstractness syntactic/argument structure constructions" (2018: 2; my emphasis). Proverbs may be grouped with the class of "fully-fixed multi-word expressions" which score comparatively high in complexity ("fully-fixed" may be questioned as proverbs tend to be variable in their structure to some degree, see below).

'Linguistic complexity' can be broken down into various types, namely 'structural complexity', 'syntactic complexity', 'clause complexity' and, for instance, 'cognitive complexity' (Szmrecsanyi 2004). The complexity of proverbs, however, *varies* across the proverb inventory. Simple word counts can be used as measures of syntactic complexity (Szmrecsanyi 2004). Proverbs in *CAEP* have a mean word count of N=6.8 words per proverb (the range is quite wide with only 2 component words as lowest limit [e.g. *Money talks*], but proverbs usually consist of a larger number of words [e.g. *The grass is always greener on the other side* with 9 words]).

Proverbs tend not only to be characterized by a higher complexity when it comes to their linguistic structure, they also tend to be variable in their form.[2] Arnaud et al. found in their study based on data from *COCA* that 77.9 % of occurrences of proverbs were not canonical (2015: 153). Some proverbs have a large 'family' of structural variants. In *CAEP* structural variants of canonical proverbs are included if they reach a minimum frequency and/or familiarity level.

The characteristic of the low frequency of occurrence is a challenge for corpus approaches in the field of phraseology. These approaches have been well established for more than two decades (Cowie 1998; Moon 1998; Steyer 2013, Steyer 2015; Gries 2015a [2013]). The study of proverbs, however, has only received

[2] 'Variability' needs to be distinguished from conscious 'modification' (Filatkina 2016: 11; also Barlow 2000).

attention from corpus-linguistic research fairly recently. Proverb use in large general language corpora has, for example, been studied by Ďurčo (2005) and Čermák (2006). This strand of research also includes diverse studies of specialized proverb corpora – for instance diachronic corpus-based research of (medieval) English proverbs (Aurich 2012) or research of the lexical frequency structure of a Slovenian proverb corpus (Grzybek 2004).

It stands to reason to study how the characteristic complexity and low frequency of proverbs can possibly be brought together with their familiarity among language users. This question may best be approached from a psycholinguistic perspective. Earlier research has mainly focused on comprehension rather than on production (Bencini 2013: 379). The storage and processing of fixed expressions has been widely discussed in psycholinguistic research (Cutting and Bock 1997; Sprenger et al. 2006; Tabossi et al. 2008; Siyanova-Chanturia et al. 2011).

Proverbs as a specific type of multi-word expression, however, have not been as widely researched to the present day. There are a number of older experimental studies on proverbs, and these usually focus on comprehension only (Gibbs and Beitel 1995; Nippold et al. 2000). Sprenger et al. (2006: 178) note that further research on various types of fixed expressions would be called for. In Lückert (2018) a novel model of 'lexical access' for proverbs is outlined in which proverbial keywords play an important role and in which the discrepancy between complexity, low frequency and familiarity of proverbs is addressed. In the following section, the corpus-linguistic methodology of detecting such keyword candidates will be dealt with.

3 The lexical profile of American English proverbs

In Section 3.1, it is discussed why a specialized proverb database was compiled for the present research in the first place. What is more, details of the corpus structure and annotations are outlined. Section 3.2 deals with the methods of detecting keywords in *CAEP*. In Section 3.3, it is argued that the keywords can be used for further, qualitative analyses.

3.1 The *Corpus of American English Proverbs* (*CAEP*)

(i) Motivation for compiling this database: Earlier quantitative studies used predefined lists of proverbs (from collections or dictionaries) either to look up occurrences in a large general language corpus (Arnaud and Moon 1993; Moon 1998;

Arnaud et al. 2015) or to scrutinize a given structure within this list treated as a corpus in itself without relating this material to other data sources (Grzybek 2001, Grzybek 2004). The number of proverb units included in the lists were often limited to several hundred items (Arnaud et al. [2015] searched for 303 items in *COCA* which actually are variants and modifications of 6 canonical proverbs; Arnaud and Moon [1993] included 250 proverbs in their study) – much larger lists were studied by Grzybek (2001, 2004).[3] The *Corpus of American English Proverbs* (*CAEP*) has a large size with 4,164 proverb variants (2,191 proverb types). Annotations for frequency in *several* large databases for all variants have been included and – if available – recent (experimental) survey data on familiarity. To the best of my knowledge, there are at present no large electronic collections of American English proverbs with such annotations.

(ii) Corpus data and annotations: The proverbs for *CAEP* have been taken from a variety of sources and have been validated as recurrent proverbs on the basis of corpus data and, in part, survey data (Steyer 2013: 79). The sources used in the present study include three proverb dictionaries/collections (Bryan and Mieder 2005; Doyle et al. 2012; Mieder 2015) and two surveys which use various experimental methods (Chlosta and Grzybek 2005; Haas 2008). All canonical forms and all variants (cited in contexts of use) were systematically extracted from the dictionaries. The two surveys were used for familiarity data for proverbs from the three collections but were also gleaned for proverbs not included in the collections.

It is a method of choice for most research in the field to extract material from existing collections and to clean up the data subsequently with the help of other methods, for instance with elicitation tests (Steyer 2013: 79). To go beyond the practice of working with predefined lists of canonical proverbs, though, further structural variants were included in *CAEP*. The additional variants were discovered mainly in two ways. Firstly, when searching hits for the proverbs from the dictionaries and surveys partial word combinations were used (so as to find hits of the target proverb with embedded strings (e.g. *they say*) and further variants – e.g. *the longest journey* aiming at *The longest journey begins with a single step* will also find *The longest journey starts with a single step*).[4] Secondly, in some instances searches were run testing some likely lexical or grammatical variations of the documented proverbs (e.g. *Almost/about/close is not good/close enough*).

[3] Some corpus-linguistic studies on phraseological units not restricted to proverbs have included considerably more items – cf. Quasthoff et al. (2010) who included roughly 6,000 phraseological units drawn from various sources.

[4] Collocates searches in *COCA* were also helpful for finding variants with embedded strings.

At times, these procedures produced hits for similar, but distinct, variants (and, in fact, proverbs) undocumented in the sources used.

These additional variants which are not included in the sources used were also integrated into the data set of *CAEP* if they reached a minimum frequency level. This practice of adding further items may not meet the requirements of systematicity as much is left to chance. Limitations of time and resources made it impossible to systematically compile lists of likely variant forms of *all* the dictionary and survey items and run searches in several large databases.[5] It is, however, still better to include relevant material which is uncovered more or less accidently than ignoring it altogether. The various collections named above can – after all – not claim to be exhaustive and fully systematic in their choice of material either (Mieder 2009; Steyer 2013: 79).

It would seem desirable to extend *CAEP* systematically in the future with more variants (including truncated forms and allusions) and to include again frequency/familiarity annotations. Common truncated forms are, at present, only included if the canonical full forms were not (sufficiently) represented in the databases used for the annotations. For example, not only the canonical proverb *Finders keepers, losers weepers* (*CAEP* F.78.1; 0.01 per mil in *COCA*) is included but also the more frequent truncated form *Finders keepers* (*CAEP* F.78.2; 0.07 per mil in *COCA*).

However, this practice is not entirely systematic. In the future, it is planned to define which truncated forms are to be included (deciding on the lower limit – namely two-word units, one-word units, or maybe even allusions without particular lexemes from the proverb alluded to) and to apply this mode to *all* items in the data set. So far, because of limitations of time and because it is not the focus of the present project, there are only very few truncated forms in *CAEP*.

Further annotations are restricted to particular components in the corpus. The sub-set of word lemmas used for the psycholinguistic experiment (word types N=251),[6] for example, has additional annotations: (a) ‘specificness score’ (number of distinct, current meanings per lemma on the basis of the *Oxford English Dictionary* as a measure of (potential) context variability [for ‘context variability’ see Criss et al. (2011)]), (b) ‘abstractness/concreteness’ (Brysbaert et al.

[5] The decision to actively look for further variants was based on search hits with partial word combinations which showed at least one lexical alternative for a constituent word (assuming that more alternatives may be likely in this slot). WordNet 2.1 was used to find semantically-related words.

[6] See Sections 1 and 3.3.

2014), and (c) 'word length' (measured in syllables; on word length see Strauss et al. [2005]).

(iii) Structure of the data set: Sub-sets of *CAEP* include the 'proverb component' ('metadata' which lists all proverb variants with sources, and, in part, first date and American English variety/origin of loan proverbs; 'raw data set' and 'cleaned data set' which both give familiarity data [Chlosta and Grzybek 2005, and Haas 2008] and frequency data), the 'word lemma component' (which includes the lemmatized lexical profile in an alphabetically ordered version and a rank-ordered one), and, lastly, the 'experimental component' (which includes the experimental material with further annotations).

The 'proverb component' of *CAEP* is divided into three sub-sets based on the frequency of occurrence and/or familiarity data (cf. the selection criteria in Table 1 below). The sub-sets comprise 'x1' proverbs (N=3,132; 75.2 %), 'x2' proverbs (N=785; 18.9 %), and the much smaller sub-set of 'x3' proverbs (N=247; 5.9 %). The present study aims only at current proverbs actually used in a language community. This is why only proverb variants of a minimum frequency of occurrence or familiarity level (category 'x1' or higher) are identified and 'dictionary proverbs' are excluded. The claim that only "current" proverbs are included is, it goes without saying, to be taken with a pinch of salt as the choice is largely based on corpus data (survey data does not cover all proverbs in *CAEP*). Corpus data cannot be entirely representative of the practices of a whole community, but *CAEP* includes at least data from *several* corpora – it is a little more balanced this way.

Proverbs tend to be infrequent in established large language corpora. *COCA*, for example, turned out to be problematic when it came to validating proverbs from existing collections (Lückert 2014). It seemed reasonable to also include data sources such as the Web (using *Google Advanced Search*) and to use a threshold of a minimum token frequency of 10 hits either on the Web (using filter options 'language' "English" and 'region' "United States" in *Google Advanced Search*) or in *Google Books* (American English).[7] Numerous searches on the Web revealed that hits below 10 usually stemmed – for the most part – from digitized versions of the collections that were used by the present study. Hits of 10 or more, on the other hand, actually comprised at least a few actual instances of use.

[7] Google frequency information is not reliable (see Colson [2007: 1072] on the limitations of the World Wide Web as a corpus).

Tab. 1: Overview of selection criteria for the proverb categories (in case of mismatches the higher category was assigned)[8]

Proverb Category	**Frequency** *Google Advanced Search* or *Google Books* (American English)	**Familiarity** (survey data) Chlosta and Grzybek 2005: completion task Haas 2008: (A) generation task (B) rating task (4-point scale)
x1	10 > 4,999 hits	10–49 % (Chlosta and Grzybek 2005) 1.00–2.49 Haas (B)
x2	5,000 > 89,999 hits	50–89 % (Chlosta and Grzybek 2005) 10–13 % Haas (A) 2.50–3.49 Haas (B)
x3	90,000$^+$ hits	90–100 % (Chlosta and Grzybek 2005) 14 %$^+$ Haas (A) 3.50–4.00 Haas (B)

Frequency and familiarity are not necessarily positively correlated (infrequent proverbs may be very familiar). If frequency/familiarity scores pointed to different proverb categories it was decided to assign the higher category to the item.[9] As a case in point, the proverb *Absence makes the heart grow fonder* (*CAEP* A.5.1; *Google Advanced Search*: 22,300 hits [accessed 22/01/2016], *Google Books* [American English]: ca. 92 hits [accessed 22/01/2016], Chlosta and Grzybek [2005]: 91.5 %, Haas [2008, B]: 3.13–3.71) may be considered. It was classified as an 'x3' proverb in spite of the comparatively low frequency data (which corresponds with 'x1' or 'x2', respectively) because of the high familiarity scores.

(iv) Methodological problems when annotating *CAEP*-data: the major problems lie in the sources used (for a general discussion of what makes counting proverbs in corpora difficult, see Steyer [2015]). The frequency data should preferably stem from spoken language corpora as proverbs are considered a predominantly oral phenomenon (Mieder 2007) – the limited size of such corpora and the overall low frequency of proverbs make this less favorable though. The corpus architecture may make proverbs "invisible" (5-gram searches only as with *Google Books* [American English] and *Google Books Ngram Viewer*); or the "40 token

[8] Various sources (for frequency and familiarity) were combined for categorization because no single source covers all proverbs. The discrepancies between frequency and familiarity result in part from survey methods (for example in case of completion tasks if several options are at hand, cf. Lückert [2018]).

[9] *CAEP* lists all scores in addition to the proverb category so that cases where there is a discrepancy may be identified.

threshold" with *Google Books* (cf. website at http://googlebooks.byu.edu/x.asp). What is more, the familiarity data has been obtained by three different methods (a completion task using the initial words, a generation task, and a rating task) – it is not entirely clear how and to what degree the results in the surveys correspond (for the major problems associated with many surveys see Ďurčo [2015]).

3.2 The detection of proverbial keywords

Keyword studies in general can be conducted along various lines. We can study which words are 'key' within one corpus and across two (or more) corpora (Rayson and Garside 2000; Köhler 2008: 305). In phraseological research, it is not new to assume that some of the linguistic material in multi-word units is characteristic in this context. It has, for example, been noted in a study on phraseological allusions that there are "image-bearing constituents" which "evoke the whole [unit] in the reader's mind" (Naciscione 2010: 70). These 'image-bearing constituents' seem similar to the keywords discussed in the present study.

Earlier quantitative research on proverbs has – so far – not addressed the question of detecting individual keywords with empirical methods. There are important quantitative studies which have shed light on diverse aspects that characterize the proverb inventory of a given language. An interesting study by Grzybek (2004), for example, has studied the lexical frequency structure of the Slovenian proverb inventory based on a historical collection of 2,429 proverb items. This study analyzed the frequencies with which individual words occur within the proverb corpus. An important finding from Grzybek's study is that the lexical repertory is "not chaotically organized" and that it shows "exactly the same regularities characterizing homogeneous texts and text corpora" (2004: 13).

It is assumed in the present study that there may be patterns in the lexical material of a given proverb inventory. The focus in the present paper will, however, not lie on frequency distribution patterns *within* the proverb corpus under consideration – it will rather lie on the frequencies of occurrence of the individual words (which are lemmatized) in the proverb corpus (as a 'target corpus') *in relation to* those in a large general language corpus (*COCA* as 'reference corpus'), differences are noted and a 'goodness-of-fit' test is used to evaluate the results.

Gries (2015b: 55) mentions a number of methods of identifying 'keywords' when comparing two corpora. With the help of Damerau's relative frequency ratio, for instance, values are computed which are interpreted in such a way that values smaller than 1 are considered 'dispreferred' and those that are larger than 1 'preferred'. A disadvantage of Damerau's relative frequency ratio, however, lies in structuring the results in groups so as to distinguish *degrees* of how typical or untypical a word is.

The problem of establishing groups of 'typicality' may be addressed by using a different paradigm. Therefore, the log-likelihood ratio (G^2) is used in the present study as a measure of 'keyness' (Dunning 1993; Rayson and Garside 2000; Gries 2016). In the present paper, G^2 is computed along the lines of Rayson and Garside (2000). The groups are set up according to the established 'critical values' that are commonly assigned to the levels of statistical significance.

In the following, it shall be explained how exactly the keywords were detected in the present study. Word lemmas may appear more often or less often than 'expected' against the background of *COCA*, that is they may differ in their 'contextual predictability'.[10] It may be argued that the two databases *CAEP* and *COCA* may be compared for the frequency patterns of word lemmas on the grounds that, by definition, proverbs are part of general language use. A collection which consists of proverbs only is then treated as a 'sub-set' of general language use and not, say, as a corpus of a 'specialized text type' (e.g. a corpus of professional medical texts) which may be problematic because of differences in sampling. In our case we aim at comparing the lexical frequency patterns of a particular sub-set (current proverbs as represented in *CAEP*) to the whole unit (general language use as represented in *COCA*). Relative frequencies may differ, which means that the lemmas are over- or under-represented in *CAEP*. It is assumed that the higher the LL values are for over-representation in *CAEP*, the more 'proverbial' are the lemmas.

Once the various values had been computed, the word frequency list was rank-ordered so that the most characteristic words are at the top of the list (Rayson and Garside 2000: 3). The words were then classified in groups according to the degree to which they are "indicative" for the lexical inventory of proverbs, namely 'sign-over' (LL+ at $p<0.05$), 'insign-over' (LL+ at $p\approx0.1$), 'insign-under' (LL- at $p\approx0.1$), 'sign-under' (LL- at $p<0.05$).

It seemed reasonable to restrict the analysis to 'content words' only which (also, and sometimes exclusively) appear in the most frequent and/or familiar

[10] The keyword calculation was run with all items of the cleaned data set rather than with the canonical variants only. Assigning the label "canonical" proverb is problematic: frequency and familiarity data often mismatches across similar variants (if one variant has a higher frequency score and a related variant a higher familiarity score – which value should take precedence?) and it is not always clear how to distinguish variants from canonical proverbs (how different does a variant need to be identified as a separate proverb?). It seemed best to focus on all items of a given frequency/familiarity irrespective of their perceived status as the resulting entrenchment is key to the present project.

proverbs of *CAEP* (347 word lemmas of sub-set 'x3').[11] So the proverb items of category 'x3' were used as a "filter", as it were, because it may be assumed that candidates for proverbial keywords will most likely turn up in this context. The lemmas of these 'x3' proverbs, however, were then considered against the background of the whole corpus (i.e. cleaned data set). The results for this sub-set suggest that 59.4 % of the lemmas (N=206) are significantly over-represented in *CAEP* (at least p<0.05), that 24.2 % of the lemmas (N=84) are insignificantly over-represented, and that 16.4 % of the lemmas (N=57) are under-represented (N=43 insignificantly, N=14 significantly). See Figure 1 below for a chart.

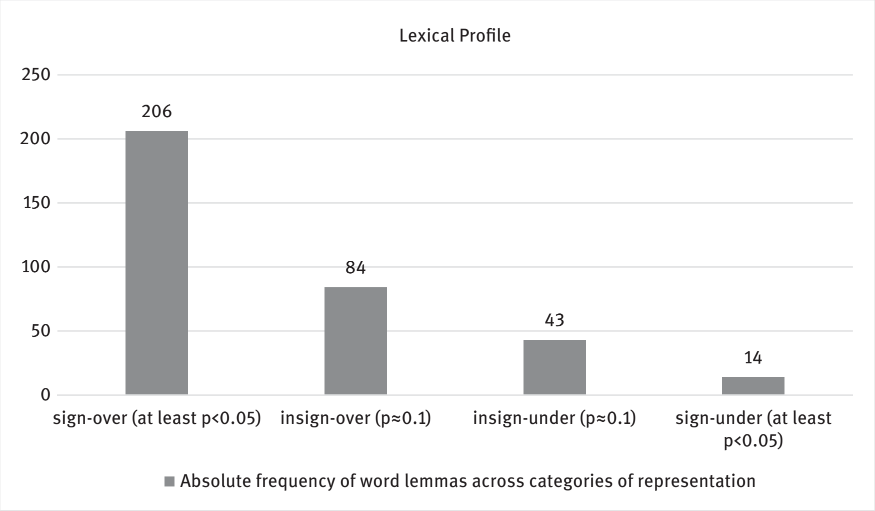

Fig. 1: Categories of representation based on the word lemmas of 'x3' proverbs

A fully automatic process of detecting keywords is not an end in itself in the present project. Instead, the keyword results can be used for further studies (see Section 3.3 below). In the experimental part of the project, the keywords were tested for their role in processing in the human mind. It should not be assumed that all, say, significantly over-represented words are equal when it comes to their 'keyness' in the real world. As a case in point, words which differ extremely in their absolute frequency

[11] X3 words were selected because we may assume that linguistic material of the most frequent and/or familiar proverbs is well entrenched in the minds of most language users. Including words which appear just in the x1 or x2 sub-set may skew results (e.g. *credulity* – in *Credulity is not a crime* (x1) – N=1 in *CAEP*, N=228 in *COCA*, "over-represented" at p<0.01, even though the proverb may be almost obsolete).

can score high in their association with the proverb inventory: compare the 'high-frequency keyword' *thing* n. (rank 6; *p*<0.001; lemma frequency in *CAEP* N=161, in *COCA* N=538,926) to the 'low-frequency keyword' *bygone* n. (rank 120; *p*<0.001; lemma frequency in *CAEP* N=2, in *COCA* N=193) which both are grouped in 'sign-over' words. It stands to reason to predict that *thing* n. and *bygone* n. will not prove to be equal when it comes to facilitating processing of proverbs (*bygone* may turn out to be more beneficial to processing on account of its strong association with the proverb context, its low absolute frequency and its low context variability etc.).

3.3 Of *God*, *dog* and what's missing

The keywords which were detected in the corpus based on the degree of their contextual association (or more precisely, the degree to which we may be certain that the results are not just chance) may be used for further studies. The overall purpose of identifying such keywords for the present project was to prepare a reaction-time experiment conducted in the Boland Lab at the University of Michigan, Ann Arbor (USA) in February 2018.[12] The keywords may, however, be of interest to a variety of other strands of research – first and foremost pragmatic, cognitive and cultural strands.

It may be assumed that a number of the keywords are tied to the cultural beliefs and norms of a given community – they are used to express views on these norms in the form of reproducible lexicalized multi-word expressions (Mieder 2007: 401–403). If the views expressed were not important to a sizable number of members of a community, such expressions would not undergo lexicalization in the first place (Schaefer 1992). If we leave aside the content word lemmas which tend to serve more structural purposes in proverbs (e.g. *like* [adj./adv.; rank 2] or *better* [n./adj./adv.; rank 5]), we can try to focus on the meaning-bearing lemmas which are strongly associated with cultural concepts. By identifying and analyzing the 'semantic frames' (Gawron 2011) of the top ranks of the keywords we may arrive at a better understanding of which concepts seem to be especially important in the proverb inventory (see Table 2 below for the top 30 ranks).

The following is meant to illustrate which 'semantic frames' may be uncovered, it is by no means meant to be a complete list.[13] We may, for example, identify

[12] Data from 97 native-speaker informants was collected. The tasks included self-paced reading, cued recall, and free recall.

[13] The semantic tagger USAS was used (http://ucrel.lancs.ac.uk/usas/tagger.html; accessed online 28/06/2018). In a few cases (*always*, *fool*, *honesty*, *waste*), however, I assigned a different category if the words were at least associatively related.

concepts such as ‘life and living things’ (cf. *life* [n.; rank 11], *live* [v.; rank 26], *die* [v.; rank 21], *kill* [v.; rank 67]), ‘time’ (cf. *time* [n.; rank 41], *always* [adv.; rank 12], *forever* [adv.; rank 37], *never* [adv.; rank 1], *tomorrow* [adv./n.; rank 28], *today* [adv./n./adj.; rank 185]; but compare *yesterday* [adv./n./adj.] which is no keyword [$p>0.1$]), ‘ability: success and failure’ (cf. *winner* [n.; rank 65], *loser* [n.; rank 39], *fool* [n./adj.; rank 16], ‘relationship/groups and affiliation’ (cf. *friend* [n.; rank 25], *alone* [adj.; rank 51], *honesty* [n.; rank 159]), ‘money’ (cf. *penny* [n.; rank 29], *waste* [v.; rank 76], *money* [n.; rank 20]), and ‘religion’ (cf. *God* [n.; rank 8], *hell* [n.; rank 45], *heaven* [n.; rank 83]).

Animal metaphors in American English proverbs seem to have a tendency to favor dog imagery (cf. *dog* [n.; rank 9] – cf. proverbs such as *Dog eat dog*, *You can’t teach an old dog new tricks*, *Every dog has his day*, *Let sleeping dogs lie*, *If you aren’t the lead dog, the scenery never changes*, *If you can’t run with the big dogs, stay on the porch*. Further animal imagery operates, for example, on cats (*cat* [n.; rank 47]) and chickens (cf. *chicken* [n.; rank 63]).

Tab. 2: Top thirty lemmas of *CAEP* keywords (sub-set of highly frequent/familiar proverbs) ranked by log likelihood

	Word Lemma List (cf. CAEP)	**POS** (cf. OED)	**OF CAEP**	**OF COCA**	**Log Likelihood Value**	**Critical Value LL (DF 1)**
1	never	adv.	205	343686	622.20	$p<0.001$
2	like	adj./adv.	91	52032	461.31	$p<0.001$
3	there	adv.	330	1463606	455.27	$p<0.001$
4	good	adj./n./adv.	165	486731	336.44	$p<0.001$
5	better	n./adj./adv.	99	144722	325.37	$p<0.001$
6	thing	n.	161	538926	294.31	$p<0.001$
7	man	n.	157	543627	278.28	$p<0.001$
8	God	n.	48	19513	275.21	$p<0.001$
9	dog	n.	65	69394	251.77	$p<0.001$
10	best	n./adj./adv.	84	198438	203.40	$p<0.001$
11	life	n.	117	436096	193.40	$p<0.001$
12	always	adv.	87	238216	188.42	$p<0.001$
13	get	v.	209	1317745	182.36	$p<0.001$
14	bad	adj./n.	59	111474	166.29	$p<0.001$
15	make	v.	173	1133929	142.05	$p<0.001$
16	fool	n./adj.	23	8748	134.87	$p<0.001$
17	every	adj./pron.	77	283478	128.81	$p<0.001$
18	love	n.	45	88541	123.55	$p<0.001$
19	count	v.	32	37191	118.84	$p<0.001$
20	money	n.	63	213246	114.01	$p<0.001$
21	die	v.	50	128836	113.14	$p<0.001$
22	acorn	n.	13	1804	101.94	$p<0.001$

Tab. 2 (continued)

	Word Lemma List (cf. CAEP)	**POS** (cf. OED)	**OF CAEP**	**OF COCA**	**Log Likelihood Value**	**Critical Value LL (DF 1)**
23	spilled	adj.	10	806	89.17	*p*<0.001
24	eat	v.	38	96504	86.94	*p*<0.001
25	friend	n.	51	190716	83.68	*p*<0.001
26	live	v.	55	231403	79.99	*p*<0.001
27	easy	adj./adv./n.	38	108370	79.38	*p*<0.001
28	tomorrow	adv./n.	25	39091	78.87	*p*<0.001
29	penny	n.	13	5012	75.77	*p*<0.001
30	evil	adj./n.	19	19061	75.67	*p*<0.001

But what is at least as interesting as what *is* represented in the ranked lemma list of the proverb corpus is what *is not* represented. Which concepts are avoided or neglected? Systematic answers to this question have to be left to future research as the present study has a different focus. What is more, finding out "what's missing" in the proverb corpus, that is which word lemmas as realizations of underlying 'semantic frames' are *absent*, would require a time-consuming procedure (automatically generated results would need to be manually checked). When we scrutinize particular 'semantic frames' which are indeed represented in *CAEP*, we may still uncover individual components which are noteworthy in their absence or low frequency of occurrence. Let us consider 'time': there seems to be a preference for expressing views about present or future times, the past appears to be neglected by comparison (cf. *today* [significantly over-represented at *p*<0.05] and *tomorrow* [significantly over-represented at *p*<0.001] compared to *yesterday* which is no keyword [insignificantly over-represented at *p*>0.1]).

Very tentatively this may be explained by functional aspects – proverbs often give advice on what and how something should be done in accordance with cultural norms, the past is often addressed only in relation to the present or future. This can be illustrated by the only proverbs (N=4) in the cleaned sub-set of *CAEP* which sport *yesterday*: first, *Today is the tomorrow that you worried about yesterday* (and its variants *Today is the tomorrow you worried about yesterday*, *Today is the tomorrow you dreamed about yesterday*), and, second, *Yesterday is gone; tomorrow has not yet come; we have only today*). The lemma *past* (adj./n.) is no keyword either (insignificantly under-represented at *p*>0.1) and appears in proverbs such as *The past does not repeat itself, but it rhymes* (the verb *repeat* indicates that the focus is – once again – on the present and/or future time), *The past does not equal the future*, and the jocular proverb *The past is not what it used to be*. A further aspect which informs us on whether particular concepts are important in the proverb inventory is the frequency and familiarity data on the individual proverb variants in *CAEP*. The four proverb variants in *CAEP* quoted above which include the lemma *yesterday* seem to be very infrequent in all general language resources used for *CAEP*-annotations.

It stands to reason that we cannot draw any general conclusions on how the concept of ‘past’ as part of the ‘semantic frame’ of ‘time’ is realized in American English proverbs based on the (non-)occurrence patterns of only two word lemmas (*yesterday*, *past*) in the proverb corpus. We could, for example, compile a list of all possible candidates of how the ‘past’ (including different word classes, different events, states, activities) may be expressed in proverbs. Then we could go ahead searching the whole proverb corpus for these candidates. This procedure would cost a lot of time and is, unfortunately, not possible in the scope of the present paper.

4 Conclusion

In the present study it has been demonstrated that a considerable number of ‘content words’ are strongly tied to the proverb as a category. Frequency patterns of content word lemmas in a specialized proverb corpus (*CAEP*) relative to a general language corpus (*COCA*) show that a large number of words are typical (to varying degrees) in the context of proverbs. The methodological problems of compiling a proverb corpus and of generating a lexical profile on its basis have been discussed. Words in the proverb corpus which are associated with the proverb as a class on a statistically significant level have been identified as ‘keywords’ that have then been included in a keyword list rank-ordered by log-likelihood values.

Creating a rank-ordered list of proverbial keywords should, however, not be an end in itself. As demonstrated in Section 3.3, such a list can be put to further, qualitative study – in a semantic framework for example. Such a keyword list can, moreover, be used for psycholinguistic experimentation as in the present project. In an experimental study at the University of Michigan it was tested whether the keywords differ in how well they activate particular proverbs and in how well they contribute to a strengthened memory representation of proverbs.

On a general plane, it is argued in the present paper that paremiological research can benefit from quantitative data. Past research has called for “the creation of new proverb collections or modern proverb information systems” which bring together various types of information (Steyer 2015: 222). *CAEP* combines information on the form of proverb variants with data on frequency and familiarity and offers a content word lemma list which informs on the strength of the association of the lemmas with the proverb context. In this way the collection may be of interest beyond the context of the present project. Researchers who work on diverse aspects (such as linguistic structuring principles in proverbs, say Panini's Principle, or the distribution of conceptual metaphors) may benefit from *CAEP* data in the future. At present, no comparable electronic resource on American English proverbs is available – it is, however, planned to make *CAEP* available for free by the end of 2018 (CLARIN Centre SFS in Tübingen, cf. http://www.sfs.uni-tuebingen.de/ascl/clarin-center/ repository.html).

References

Arnaud, Pierre J. L. & Rosamund Moon. 1993. Fréquence et emplois des proverbes anglais et français [Frequency and usage of English and French proverbs]. In Christian Plantin (ed.), *Lieux Communs, Topoï, Stéréotypes, Clichés* [Commonplaces, topoi, stereotypes, cliches], 323–341. Paris: Kimé.

Arnaud, Pierre J. L., François Maniez & Vincent Renner. 2015. Non-canonical proverbial occurrences and wordplay: A corpus investigation and an enquiry into readers' perception of humour and cleverness. In Angelika Zirker and Esme Winter-Froemel (eds.), *Wordplay and Metalinguistic/Metadiscursive Reflection: Authors, Contexts*, 135–160. Berlin & Boston: De Gruyter.

Arora, Shirley L. 1984. The perception of proverbiality. *Proverbium: Yearbook of International Proverb Scholarship* 1. 1–38.

Aurich, Claudia. 2012. *Proverb Structure in the History of English: Stability and Change. A Corpus-Based Study* (Phraseologie und Parömiologie 26). Baltmannsweiler: Schneider Verlag Hohengehren.

Barlow, Michael (2000). Usage, Blends, and Grammar. In Michael Barlow and Suzanne Kemmer (eds.), *Usage Based Models of Language*, 315–345. Stanford, CA: CSLI Publications.

Bencini, Guilia M. L. 2013. Psycholinguistics. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford Handbook of Construction Grammar*, 379–396. Oxford: Oxford University Press.

Bryan, George B. & Wolfgang Mieder. 2005. *Dictionary of Anglo-American Proverbs and Proverbial Phrases. Found in Literary Sources of the Nineteenth and Twentieth Centuries*. New York: Peter Lang.

Brysbaert, Marc, Amy Beth Warriner & Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46. 904–911.

Bybee, Joan. 2013. Usage-based theory and exemplar representation of constructions. In Thomas Hoffmann and Graeme Trousdale (eds.), *The Oxford Handbook of Construction Grammar*, 49–69. Oxford: Oxford University Press.

Čermák, František. 2006. Statistical methods for searching idioms in text corpora. In Annelies Häcki Buhofer & Harald Burger (eds.), *Phraseology in Motion I. Methoden und Kritik* [Phraseology in motion I. Methods and criticism] (Phraseologie und Parömiologie 19), 33–42. Baltmannsweiler: Hohengehren.

Chlosta, Christoph & Peter Grzybek. 2005. Varianten und Variationen anglo-amerikanischer Sprichwörter – Dokumentation einer empirischen Untersuchung [Variants and variations of American English proverbs – Documentation of an empirical study]. *ELiSe: Essener Linguistische Skripte elektronisch* 5(2). 63–145.

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.), *Phraseologie. Phraseology. Ein internationales Handbuch zeitgenössischer Forschung. An International Handbook of Contemporary Research*, Vol. 2, 1071–1077. Berlin & New York: De Gruyter.

Cowie, Anthony P. 1998. *Phraseology: Theory, Analysis, and Applications*. Oxford: Clarendon Press.

Criss, Amy H., William R. Aue & Larissa Smith. 2011. The effects of word frequency and context variability in cued recall. *Journal of Memory and Language* 64. 119–132.

Cutting, J. Cooper & Kathryn Bock. 1997. That's the way the cookie bounces: Syntactic and semantic components of experimentally elicited idiom blends. *Memory & Cognition* 25(1). 57–71.

Davies, Mark. 1990–. *Corpus of Contemporary American English* (*COCA*) (560 million words, 1990–2017). https://corpus.byu.edu/coca/old/ (accessed October 2015–January 2017).

Davies, Mark. 2011–. *Google Books* (*American English*) *Corpus* (155 billion words, 1810–2009). http://googlebooks.byu.edu/ (accessed October 2015–January 2017).

Doyle, Charles Clay, Wolfgang Mieder & Fred R. Shapiro. 2012. *The Dictionary of Modern Proverbs*. New Haven et al.: Yale University Press.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1). 61–74.

Ďurčo, Peter. 2005. *Sprichwörter in der Gegenwartssprache* [Proverbs in modern language use]. Trnava: Univerzita Sv. Cyrila a Metoda. Filozofická fakulta.

Ďurčo, Peter. 2015. Empirical research and paremiological minimum. In Hrisztalina Hrisztova-Gotthardt & Melita Aleksa Varga (eds.), *Introduction to Paremiology. A Comprehensive Guide to Proverb Studies*, 183–205. Warsaw: De Gruyter Open.

Filatkina, Natalia. 2016. Wie fest sind feste Strukturen? Beobachtungen zu Varianz in (historischen) Wörterbüchern und Texten [How fixed are fixed structures? Observations on variation in (historical) dictionaries and texts]. In Luise Borek & Andrea Rapp (eds.), *Varianz und Vielfalt interdisziplinär: Wörter und Strukturen* [Variation and variety from an interdisciplinary perspective: Words and structures], 7–27. OPAL 2.

Gawron, Jean Mark. 2011. Frame semantics. In Klaus von Heusinger, Claudia Maienborn & Paul Portner (eds.). *Semantics – An International Handbook of Natural Language Meaning* (Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science (HSK), 33/1), 664–687. eBook. Berlin: De Gruyter.

Gibbs, Raymond W. & Dinara Beitel. 1995. What proverb understanding reveals about how people think. *Psychological Bulletin* 118(1). 133–154.

Google Advanced Search. http://advancedgoogle.com/ (accessed October 2015–January 2017.)

Gries, Stefan Th. 2015a [2013]. 50-something years of work on collocations: what is or should be next… In Sebastian Hoffmann, Bettina Fischer-Starcke & Andrea Sand (eds.), *Current Issues in Phraseology*, 135–164. Amsterdam & Philadelphia: John Benjamins.

Gries, Stefan Th. 2015b. Quantitative designs and statistical techniques. In Douglas Biber & Randi Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*, 50–71. Cambridge: Cambridge University Press.

Gries, Stefan Th. 2016. *Quantitative corpus linguistics with R*, 2nd rev. and ext. edn. London & New York: Routledge, Taylor & Francis Group.

Gries, Stefan Th. 2018. Operationalizations of domain-general mechanisms cognitive linguists often rely on: A perspective from quantitative corpus linguistics. In Stefan Engelberg, Henning Lobin, Kathrin Steyer & Sascha Wolfer (eds.), *Wortschätze: Dynamik, Muster, Komplexität*, 75–90. Berlin & Boston: De Gruyter.

Grzybek, Peter. 2001. Zur Satz- und Teilsatzlänge formelhafter zweigliedriger Sprichwörter [Sentence and sentence part length of formulaic bipartite proverbs]. In Ludmila Uhlířová, Gejza Wimmer, Gabriel Altmann & Reinhard Köhler (eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs* (Festschrift für Luděk Hřebíček), 64–75. Trier: WVT.

Grzybek, Peter. 2004. A quantitative approach to lexical structure of proverbs. [Special issue: Festschrift in honour of Professor Raimund Piotrowski]. *Journal of Quantitative Linguistics* 11(1–2). 79–92.

Grzybek, Peter. 2012. Facetten des parömiologischen Rubik-Würfels. Kenntnis = Bekanntheit? [Facets of the paremiological Rubik's Cube. Knowledge = familiarity?]. In Kathrin Steyer (ed.), *Sprichwörter multilingual. Theoretische, empirische und angewandte Aspekte der modernen Parömiologie* [Proverbs from a multilingual angle. Theoretical, empirical and applied aspects of modern paremiology], 99–138. Tübingen: Gunter Narr.

Haas, Heather A. 2008. Proverb familiarity in the United States: Cross-regional comparisons of the paremiological minimum. *Journal of American Folklore* 121. 319–347.

Honeck, Richard P. 1997. *A Proverb in Mind. The Cognitive Science of Proverbial Wit and Wisdom*. Mahwah, N. J.: Lawrence Erlbaum.

Köhler, Reinhard. 2008. Properties of lexical units and systems. In Reinhard Köhler, Gabriel Altmann & Raimund Piotrowski (eds.), *Quantitative Linguistik/Quantitative Linguistics. Ein internationales Handbuch/An International Handbook* (Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science (HSK) 27), 305–312. eBook. Warsaw: De Gruyter.

Lewandowska, Anna & Gerd Antos. 2015. Cognitive aspects of proverbs. In Hrisztalina Hrisztova-Gotthardt & Melita Aleksa Varga (eds.), *Introduction to Paremiology. A Comprehensive Guide to Proverb Studies*, 162–182. Warsaw: De Gruyter Open.

Lückert, Claudia. 2014. Prosodic Aspects of Proverb Change in English: Panini's Principle. In Peter Grzybek & Vida Jesenšek (eds.), *Phraseologie im Wörterbuch und Korpus*, 181–192. Maribor: Zora.

Lückert, Claudia. 2018. A psycholinguistic approach to the conventionalisation and variation of proverb structure. In Natalia Filatkina & Sören Stumpf (eds.), *Konventionalisierung und Variation: Phraseologische und konstruktionsgrammatische Perspektiven* [Conventionalisation and variation: Perspectives from phraseology and construction grammar], 53–71. Frankfurt a. M.: Peter Lang.

Mac Coinnigh, Marcas. 2015. Structural aspects of proverbs. In Hrisztalina Hrisztova-Gotthardt & Melita Aleksa Varga (eds.), *Introduction to Paremiology. A Comprehensive Guide to Proverb Studies*, 112–132. Warsaw: De Gruyter Open.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak & Erez Lieberman Aiden. 2010. Quantitative analysis of culture using millions of digitized books. *Science* (Published online ahead of print: 16 December 2010). *Google Books Ngram Viewer*. https://books.google.com/ngrams (accessed October 2015–January 2017).

Mieder, Wolfgang. 2007. Proverbs as cultural units or items of folklore. In Harald Burger, Dmitrij Dobrovol'skij, Peter Kühn & Neal R. Norrick (eds.), *Phraseologie. Phraseology. Ein internationales Handbuch zeitgenössischer Forschung. An International Handbook of Contemporary Research*, Vol. 1, 394–414. Berlin & New York: De Gruyter.

Mieder, Wolfgang. 2009. New proverbs run deep: Prolegomena to a dictionary of modern Anglo-American proverbs. *Proverbium* 26. 237–274.

Mieder, Wolfgang. 2015. *Different Strokes for Different Folks: 1250 authentisch amerikanische Sprichwörter* [Different strokes for different folks: 1250 authentic American proverbs]. Bochum: Brockmeyer.

Moon, Rosamund. 1998. *Fixed Expressions and Idioms in English. A Corpus-Based Approach*. Oxford: Clarendon.

Naciscione, Anita. 2010. *Stylistic Use of Phraseological Units in Discourse*. Amsterdam & Philadelphia: John Benjamins.

Nippold, Marilyn A., Melissa M. Allen & Dixon I. Kirsch. 2000. How Adolescents Comprehend Unfamiliar Proverbs: The Role of Top-Down and Bottom-Up Processes. *Journal of Speech, Language, and Hearing Research* 43. 621–630.

Oxford English Dictionary (*OED*). http://www.oed.com/ (accessed October 2016–November 2017).

Quasthoff, Uwe, Fabian Schmidt & Erla Hallsteinsdottir. 2010. Häufigkeit und Struktur von Phraseologismen am Beispiel verschiedener Web-Korpora [Frequency and structure of phraseologisms on the basis of various Web-corpora]. In Stefaniya Ptashnyk, Erla Hallsteinsdottir & Noah Bubenhofer (eds.), *Korpora, Web und Datenbanken. Computergestützte Methoden in der modernen Phraseologie und Lexikographie* [Corpora, Web and databases. Computer-based methods in modern phraseology and lexicography], 37–53. Baltmansweiler: Hohengehren.

Rayson, Paul & Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000). 1–8 October 2000*, Hong Kong.

Schaefer, Ursula. 1992. *Vokalität. Altenglische Dichtung zwischen Mündlichkeit und Schriftlichkeit* [Vocality. Old English poetry between orality and literacy]. Tübingen: Narr.

Siyanova-Chanturia, Anna, Kathy Conklin & Walter van Heuven. 2011. Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37(3). 776–784.

Sprenger, Simone A., Willem J. M. Levelt & Gerard Kempen. 2006. Lexical access during the production of idiomatic phrases. *Journal of Memory and Language* 54. 161–184.

Steyer, Kathrin. 2013. *Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht* [Current word combinations. Primary patterns of language use from a corpus linguistic perspective]. Tübingen: Narr.

Steyer, Kathrin. 2015. Proverbs from a corpus-linguistic point of view. In Hrisztalina Hrisztova-Gotthardt & Melita Aleksa Varga (eds.), *Introduction to Paremiology. A Comprehensive Guide to Proverb Studies*, 206–228. Warsaw: De Gruyter Open.

Strauss, Udo, Peter Grzybek & Gabriel Altmann. 2005. Word length and word frequency. In Peter Grzybek (ed.), *Word length studies and related issues*, 255–272. Boston & Dordrecht: Kluwer.

Szmrecsanyi, Benedikt M. 2004. On Operationalizing Syntactic Complexity. In Gerald Purnelle, Cedrick Fairon & Anne Dister (eds.), *JADT 04. Le poids des mots. Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles* [Proceedings of the 7th International Conference on Textual Data Statistical Analysis], 1031–1038. Louvain: Presses Universitaires de Louvain.

Tabossi, Patrizia, Rachele Fanari & Kinou Wolf. 2008. Processing idiomatic expressions: Effects of semantic compositionality. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34(2). 313–327.