International Journal of Lexicography, 2020, 1–26
doi: 10.1093/ijl/ecaa008
Article

Article

IDENTIFYING DICTIONARY-RELEVANT FORMULAIC SEQUENCES IN WRITTEN AND SPOKEN CORPORA

Kaja Dobrovoljc
Faculty of Arts, University of Ljubljana; Jozef Stefan Institute, Ljubljana, Slovenia (kaja.dobrovoljc@ijs.si)

Abstract

In view of the pervasiveness of formulaic language in human communication and the growing awareness of its relevance to modern lexicography, this study presents a corpus-driven identification, analysis and comparison of dictionary-relevant formulaic sequences in reference corpora of written and spoken Slovenian. The sequences were identified using a semi-automatic approach, whereby the most frequently recurring word combinations in each corpus were ranked according to their statistical salience and manually inspected for formulaic expressions with lexicographic relevance. Despite its semantic heterogeneity, the resulting list illustrates the distinct characteristics of formulaic multi-word expressions, such as high frequency of usage, prevalent inclusion of grammatical words and common non-propositional meaning, especially in speech, where research revealed numerous understudied formulaic expressions related to interaction management and mitigation. The final evaluation of measures used in the identification process demonstrates their relative suitability for corpus-driven identification of dictionary-relevant formulaic expressions, with their precision varying in relation to corpus size and length of sequences under investigation.

Key words: phraseology, lexical bundles, corpus analysis, lexical frequency

1. Introduction

Following the pioneering theoretical discussion on the prominence of lexical patterning (Firth 1957, Bolinger 1976), the last three decades have seen an extensive body of research on the formulaic nature of language use, exposing the multitude of multi-word combinations which language users seem to store and retrieve as single vocabulary units (Sinclair 1991:114, Wray 2002:9). In addition to the most commonly studied groups of multi-word expressions, such as idioms (e.g. *break a leg*), proverbs (e.g. *barking dogs*

seldom bite) or collocations (e.g. *heavy rain*), defined by their distinct semantic or syntactic characteristics (Cowie 1994, Atkins and Rundell 2008), a number of corpus-driven (Biber et al. 1999, Erman and Warren 2000, Biber et al. 2004), psycholinguistic (Conklin and Schmitt 2008, Tremblay et al. 2011) and phonological (Lin 2010) investigations have shown that special formulaic status can also be attributed to the most frequently recurring sequences of words in a language—variously termed formulaic sequences, lexical bundles, prefabricated expressions, routine formulae, etc.—which are not necessarily structurally complete and/or semantically non-compositional (e.g. *this means that*, *have a nice day*, *let me guess*).

Despite their pervasiveness in language use and the empirical evidence of their holistic processing, formulaic sequences have largely been disregarded in current lexicographic practice (Paquot 2015), due to their less conspicuous nature in comparison to phraseological units with stronger cognitive salience (Hanks 2013). In keeping with general findings that formulaic sequences represent one of the key indicators of native-like linguistic performance and fluency (Granger 1998, Wood 2010), this trend, however, is slowly changing under the influence of learner and bilingual lexicography (Siepmann 2008, Granger and Lefer 2016), which argues for greater coverage, accessibility and systematic description of this type of expressions.

Nevertheless, lexicography-oriented studies mostly remain limited to theoretical and practical considerations related to the general relevance of this category and its representation in dictionaries, drawing on examples of predefined subsets of formulaic expressions, such as sequences with metadiscursive function (Siepmann 2005, Granger and Lefer 2016). Much less work has been dedicated to the methodological aspects of an exhaustive bottom-up identification and description of this statistics-driven class of expressions as a whole. This is an especially pertinent issue given the lack of consensus on the optimal method for measuring formulaicity in general (Granger and Paquot 2008, Biber 2009, Gries 2012), a fact often reflected in past work on corpus-driven formulaic sequence extraction, where lexical sequences occurring above a certain threshold have either been considered relevant due to their frequency alone (Biber 2009) or filtered by additional measures of lexical association (Simpson-Vlach and Ellis 2010) and/or various qualitative formal and semantic criteria (Wray 2008, Martinez and Schmitt 2012).

To offer new insights into which frequently recurring sequences of words in a language are actually relevant in terms of lexicography and how to best identify them in language corpora, this study presents the identification and analysis of formulaic sequences with potential dictionary relevance in reference corpora of Slovenian. Specifically, it aims to (i) identify a representative inventory of dictionary-relevant formulaic sequences; (ii) analyse their formal and semantic characteristics; and (iii) determine the optimal method for their corpus-driven identification. To give an exhaustive overview of formulaic multiword expressions in the language, we perform our analysis on the reference corpora of both written and spoken Slovenian especially in view of the fact that the field of speech-specific lexis in general has so far remained under-researched (Siepmann 2015, Verdonik and Maučec 2016).

The remainder of this article is structured as follows. The two corpora are presented in Section 2, which is followed by the description of formulaic sequence extraction, selection and annotation in Section 3 and an in-depth discussion of methodological issues related to such subjective categorization task in Section 4. A detailed analysis and comparison of the

identified dictionary-relevant formulaic sequences in both corpora is given in Section 5, followed by the final evaluation of various statistical measures used for their identification in Section 6. Finally, we summarize our findings in Section 7 and discuss their implications for future lexicographic investigations of formulaic expressions in general.

2. Data

2.1. Written corpus Gigafida

Gigafida is a reference corpus of written Slovenian containing more than 1.3 billion words from newspapers (47.8%), magazines (16.5%), web texts (28.0%), fiction (3.5%), non-fiction (3.8%) and other works (0.3%) published from 1990 to 2018. Gigafida thus provides a representative sample of modern standard written Slovenian and is intended for usage-based descriptive linguistic studies and the compiling and development of corpus-based dictionaries (Gantar et al. 2016, Kosem et al. 2018, Arhar Holdt et al. 2018), grammars, teaching materials and language technologies for Slovenian. The present study is based on the recently released Gigafida 2.0 version of the corpus (Krek et al. 2019), which has been improved in comparison to the first release (Logar et al. 2012) by adding texts published after the year 2011, removing non-standard online communication and excluding duplicates. The corpus is freely accessible as part of the CJVT language resources portal,[1] as well as SketchEngine,[2] noSketchEngine[3] and KonText[4] corpus querying tools.

2.2. Spoken corpus GOS

GOS is a reference corpus of spoken Slovenian containing approximately 120 hours (1 million words) of transcribed spontaneous speech from different everyday situations, balanced to be representative of speaker demographics (sex, age, region, education) and channel (TV, radio, telephone, personal contact), as well as type of spoken communication settings. These include both public and non-public speech events, categorized into public informative and educational speech (35%), such as television and radio shows, interviews, discussions, school lessons and academic lectures; public entertainment speech (22%), such as talk shows, morning radio shows and sports broadcasting; non-public non-private speech (15%), such as work meetings, consultations and services; and non-public private speech (28%), such as conversations between family and friends. This study used the freely available GOS version 1.0 (Zwitter Vitez et al. 2013), which can also be accessed through an official audio-supported concordancer,[5] as well as noSketchEngine[6] and KonText[7] corpus querying tools. There are two levels of transcription available (Verdonik et al. 2013); the present work is based on normalized GOS transcriptions, that is the transcriptions with standardized spelling which neutralizes any variation in pronunciation.

3. Method

3.1. Extraction of formulaic sequences

The initial list of the most frequently recurring sequences of words in both corpora was extracted using the n-gram extraction module of LIST (Krsnik et al. 2019), an open-source tool for statistical analysis of large-scale corpora. Specifically, we extracted sequences of

two to five contiguous tokens (occurring within sentence/utterance boundaries) with a minimal relative frequency of twenty occurrences per million, in line with standard practice in formulaic language research, where frequency thresholds usually range from five to forty occurrences per million.[8] Given the orthographic distinctions between the two corpora, sequences of normalized and lowercased word forms were extracted from the spoken and written corpus, respectively, excluding punctuation. In addition to maintaining direct comparability between the two lists, this also enabled the consolidation of numerous capitalization and punctuation-based variants occurring in written corpora, such as a joint count of the variants *Kljub temu da* / *kljub temu da* / *Kljub temu, da* / *kljub temu, da* of the formulaic expression *kljub temu da* 'despite the fact that'. Using this extraction procedure, 2,687 and 4,895 formulaic sequences have been identified in the Gigafida written and GOS spoken corpus, respectively, confirming the presence of a large body of formulaic language in written and especially spoken Slovenian.

3.2. Ranking of formulaic sequences

In the second step of the formulaic sequence identification procedure, the resulting lists of all n-grams fulfilling the above criteria were ranked according to their statistical salience. Given the on-going discussion on the optimal method for measuring formulaicity mentioned in the Introduction, the list of extracted sequences in both corpora was ranked according to six statistical measures producing six distinct recommendations of the most pertinent formulaic sequences in each corpus. These methods include the absolute frequency count and five commonly used association measures (Evert 2009, Pecina 2010) that either measure the effect-size (i.e. the strength of statistical attraction between words), such as the Dice coefficient, point-wise mutual information and cubic mutual information, or statistical significance (i.e. the amount of evidence for a positive statistical association between words), such as t-score and simple log-likelihood measures. Figure 1 gives the exact equation for each association measure using the nomenclature by Ramisch et al. (2010), who also take into account the applicability of measures to n-grams longer than two words, that is to sequences of words w_1 to w_n, with observed marginal word frequencies $c(w_1)\ldots c(w_n)$ and the observed n-gram frequency $c(w_1 \ldots w_n)$ in a corpus of N words. The expected frequency of words co-occurring by chance is calculated as $E(w_1 \ldots w_n) \approx \frac{c(w_1)\ldots c(w_n)}{N^{n-1}}$.[9]

The subsequent analysis of formulaic sequences in each corpus looked at the top ranked candidates identified by each of the six measures. Given the difference in the number of formulaic sequences occurring in each corpus, different thresholds were selected, that is the top 500 candidates for the Gigafida written (amounting to 1,215 distinct sequences) and the top 1,000 candidates for the GOS spoken corpus (2,374 distinct sequences), with both cut-off points representing approximately the top 20% of the most salient candidates among all the extracted sequences in each corpus (i.e. 18.6% and 20.4% for Gigafida and GOS, respectively). Table 1 gives the quantitative summary of the extraction and ranking procedures in both corpora.

3.3. Annotation of dictionary-relevant formulaic sequences

In the final stage, the resulting lists of 1,215 and 2,374 top-ranked formulaic sequences in Gigafida and GOS, respectively, were manually inspected to identify sequences with lexicographic relevance. This was accomplished by means of two separate annotation campaigns

$$\textbf{MI} = \log_2 \frac{c(w_1 \ldots w_n)}{E(w_1 \ldots w_n)} \qquad \textbf{MI}^3 = \log_2 \frac{c(w_1 \ldots w_n)^3}{E(w_1 \ldots w_n)}$$

$$\textbf{Dice} = \frac{n \times c(w_1 \ldots w_n)}{\sum_{i=1}^{n} c(w_i)} \qquad \textbf{t-score} = \frac{c(w_1 \ldots w_n) - E(w_1 \ldots w_n)}{\sqrt{c(w_1 \ldots w_n)}}$$

$$\textbf{simple-LL} = 2 \times (c(w_1 \ldots w_n) \times \log \frac{c(w_1 \ldots w_n)}{E(w_1 \ldots w_n)} - \big(c(w_1 \ldots w_n) - E(w_1 \ldots w_n)\big))$$

Figure 1: Association measures used in the study: pointwise mutual information (MI), cubic mutual information (MI^3), Dice coefficient (Dice), t-score, simple log-likelihood (simple-LL).

Table 1: The number of all extracted and top-ranking formulaic sequences in Gigafida and GOS by sequence length.

	Gigafida (written)		GOS (spoken)	
no. of words	all sequences	top-ranked	all sequences	top-ranked
2	2,281	809	3,999	1,808
3	393	393	834	504
4	10	10	53	53
5	3	3	9	9
Total	**2,687**	**1,215**	**4,895**	**2,374**

(one for each corpus) with a broad range of categorization tasks (Dobrovoljc 2019) carried out by four pre-trained native-Slovenian students of linguistics, who were asked to categorize the sequences as either relevant or irrelevant, based on the instructions given in the annotation guidelines.

Since one of our main objectives was to investigate the concept of dictionary-relevance itself, the guidelines were intentionally kept short, simple and neutral in terms of existing dictionary-making theory and practice. The annotators were thus asked to identify formulaic sequences they would expect to find in a general monolingual dictionary with a wide spectrum of potential users (both native and non-native speakers) in mind, either as independent dictionary entries or in any other section of the entry microstructure used for highlighting the headword's typical multi-word units. In line with the related state-of-the-art approaches to Slovene multi-word expressions (Gantar et al. 2016, Gantar et al. 2019a), they were primarily asked to identify any sequence functioning as a multi-word unit with a recognizable independent meaning or function, ranging from semantically transparent collocations (e.g. *prehodno stanje* 'transition state', *na internetu* 'on the Internet') and syntactic expressions (e.g. *zaradi tega ker* 'due to the fact that', *bolj ali manj* 'more or less') to non-compositional fixed expressions (e.g. *javni sektor* 'public sector', *sto osemdeset stopinj* 'a hundred and eighty degrees') and phraseological units with metaphorical, expressive or pragmatic meaning (e.g. *dame in gospodje* 'ladies and gentlemen', *to je to* 'that's it', *na zdravje* 'cheers')—with the exception of a few illustrative examples, however, no specific definitions of these categories were provided. The annotators were instructed to label all other sequences, perceived as free word combinations, as irrelevant, such as structurally

incomplete sentence fragments (e.g. *da gre za* 'that it is about', *predlagam da* 'I suggest that'), the most common type of frequently recurring sequences in general (Biber et al. 2004, Dobrovoljc 2019).

Each sequence was annotated by two independent annotators, with only one decision allowed per sequence. In case of ambiguity, the annotators were advised to inspect a random sample of concordances in the corpus (with the links provided as part of the annotation spreadsheet). They were instructed to label a sequence as relevant regardless of the frequency of dictionary-relevant usage, for instance in the case of sequence *to je* 'that is', which can either occur as an irrelevant sentence fragment, as seen in example (1), or as a fixed expression with a discourse-organizing function, as demonstrated in example (2).

1. *Ah, to je normalno, pojdi domov.*
   'Ah, that is normal, go home.'
2. *Če imamo radi okrasne koprive, pravočasno, to je ob koncu poletja in na začetku jeseni, poskrbimo za podmladek in naredimo potaknjence.*
   'If we like painted nettles, we have to take care of the offspring and propagate the cuttings in time, that is at the end of the summer or in the beginning of the fall.'

4. Inter-annotator agreement

The annotators agreed on the (ir)relevance of 997 out of 1,215 formulaic sequences in the Gigafida written corpus, amounting to an 82.1% absolute agreement between the two annotators and a Cohen's kappa coefficient of 0.54. A somewhat lower agreement was observed for formulaic sequences in the GOS spoken corpus, where the annotators agreed on 1,840 out of 2,374 (77.5%) sequences with a Cohen's Kappa coefficient of 0.43. Although it is difficult to give an exact interpretation of these scores—with most Kappa interpretation scales describing it as moderate (Viera and Garrett 2005)—they confirm a satisfactory degree of inter-rater reliability, comparable with the results observed in related work on semantically non-compositional multi-word expressions in Slovenian (Gantar et al. 2019a), especially given the distinct nature of the annotation task. This is specific not only in terms of linguistic categorization (subjective interpretation of a relatively abstract concept), but also in terms of items under investigation (ambiguous sequences with competing interpretations) and the annotation setting itself (lack of immediate context, simple guidelines, non-expert annotators). With spoken language sequences in particular, the annotators faced an additional challenge of having to judge the relevance of typically spoken lexical phenomena, with which they had no previous experience in traditional Slovenian dictionaries (Verdonik and Maučec 2016).

To illustrate these issues, we give a detailed analysis of the most frequent inter-annotator disagreement in the sections below. Based on this analysis, an updated version of the guidelines was created and applied in the final adjudication of the competing decisions by an expert third annotator (author of the guidelines and this study). Given the equivocal nature of such categorization, however, information on individual annotator's decisions has been preserved in the released version of the annotated lists, in order to enable future work based on new or refined categorization criteria.[10]

4.1. Borderline sequences in the Gigafida written corpus

As with any annotation task, some degree of disagreement on the relevance of formulaic sequences in Gigafida can simply be attributed to misinterpretation of the original

guidelines, such as annotating proper names (e.g. *Evropska komisija* 'European Commission'), sentence fragments (e.g. *je na primer* 'is for example') or compositional discourse marker co-occurrences (*in tako* 'and so') as dictionary relevant despite explicit mention of these exclusion categories in the guidelines. Nevertheless, a large proportion of remaining disagreement reveals very specific sets of formulaic expressions causing ambiguity.

In Gigafida, by far the most disagreement occurs with sequences involving prepositions. In addition to prepositional phrases with adverbial or modal function (e.g. *brez težav*, lit. 'without issues', i.e. easily) and multi-word prepositions (e.g. *v zvezi z* 'in relation to') with a relatively straightforward multi-word unit status due to their syntactic fixedness and semantic idiomaticity, other types of prepositional phrases have also emerged as ambiguous, such as adjuncts (e.g. *iz tujine* 'from abroad', *na fotografiji* 'in the photo'); case-marking prepositional phrases introducing a nominal (e.g. *na vrhu* 'on top [of something]'); modified prepositions (*daleč od* 'far from', *takoj po* 'right after'); and combinations of content words and prepositions indicating typical valency (*čas za* 'time for', *govorimo o* 'talk about', *odvisno od* 'depending on'). Essentially, all these examples point to an underspecified category of (grammatical) collocations, as well as cases of formulaic propositional collocations involving numerals (*ob 17. uri* 'at 5 o'clock') and semantically underspecified modifiers (*zelo pomembno* 'very important', *nekaj dni* 'a few days', *naslednje leto* 'next year'), which had an equal degree of inter-annotator disagreement. Given the ongoing discussion on the interweaving statistical, syntactic and semantic criteria for collocation delimitation, a burning issue in Slovenian lexicography (Kosem et al. 2018) as well as general lexicography, we decided to keep to the original, inclusive approach in this exploratory stage; all of these types of collocations were therefore considered to be relevant, especially in view of their outstanding frequency of usage and undisputed contribution to the illustration of semantic and colligational tendencies of the words they contain.

The second source of frequent disagreement in Gigafida involves discourse-structuring devices, such as multi-word discourse connectives (*zato ker* 'because', *v tem primeru* 'in this case'), modified connectives (*bolj kot* 'more than', *tudi če* 'even if'), connectives with semantically bleach contrastive particle *pa* (*sicer pa* 'otherwise') and discourse-organizing sentence stems (*kar pomeni da* 'which means that'). Given their frequency, fixedness and idiomatic function these expressions have all been labelled as relevant. This same decision was also applied in cases of disagreement involving modified adverbs (*kar nekaj* 'quite a lot', *vse več* 'more and more') and inherently reflexive verbs (*se da* 'it is possible').

4.2. Borderline sequences in the GOS spoken corpus

Points of disagreement similar to those in Gigafida also emerged in the list of formulaic sequences identified in the spoken GOS corpus, such as complex predicates (*bi moral* 'should have', *ne ve* 'does not know'), prepositional phrases involving personal pronouns (*k meni* 'to me'), clause beginnings (*dobro da* 'it's good that') and fragments of longer multi-word expressions (*[na] drugi strani* '[on] the other side')—adjudicated as irrelevant—as well as collocations involving numerals (*petnajst minut* 'fifteen minutes'), semantically bleach collocates (*nekaj drugega* 'something else', *tule gor* 'up here') and typical valency prepositions (*govorimo o* 'talk about', *hvala za* 'thanks for')—adjudicated as relevant.

Apart from that, the inter-annotator disagreement in GOS also revealed several new types of borderline formulaic sequences, related to typically spoken phenomena. Firstly, there is much broader disagreement related to discourse-structuring devices that not only include connectives, such as *glede na to da* 'given the fact that', *tudi če* 'even if', *takrat ko* 'exactly when', but also expressions related to interaction management, such as discourse particles (*bi rekel* 'say'), interjections (*a ja* 'oh really', *daj no* 'come on') and general extenders (*ali kaj* 'or what'). Such a discrepancy hardly comes as a surprise, as native speakers often disregard such formulaic expressions as self-explanatory and semantically vague; however, they were all labelled as relevant in the final round of annotation because of their essential role in second language acquisition and understanding.

A similar explanation can be given for the second large area of inter-annotator disagreement, the sequences marking formulaic replies and inquiries, such as *daj nehaj* 'stop it', *glej glej* 'well well', *pa kaj* 'so what', *ni nujno* 'not necessarily', *to pa res* 'that's true' or *bo šlo* 'all good', *kaj mislite* 'what do you think', *kako si* 'how are you', *še kaj* 'anything else'. Here, the guidelines were updated, defining such sequences as relevant if they are relatively fixed, occur with high frequency and have an identifiable independent pragmatic function. Less fixed units with a high degree of syntactic and functional compositionality, such as *aha ja* 'ok yes', *ampak ja* 'but yes', *ne morem* 'I can't', *ja saj vem* 'yes I know', *se spomniš* 'do you remember', *kje si* 'where are you', were marked as irrelevant in this particular study.

5. Dictionary-relevant formulaic sequences in spoken and written Slovenian

As summarized in Table 2, the extraction, annotation and adjudication process described in Sections 3 and 4 above resulted in the final lists of 420 dictionary-relevant formulaic sequences in the Gigafida written corpus and 604 dictionary-relevant formulaic sequences in the spoken GOS corpus. For both modes of communication, this number represents a substantial share of top-ranked formulaic sequences (i.e. 35% and 25.4% of all annotated candidates in Gigafida and GOS, respectively), confirming the general usefulness of the identification method(s) selected. Interestingly, despite the larger number of formulaic sequences in speech in comparison to writing (Table 1), the percentage of dictionary-relevant sequences is greater in writing than in speech. In addition to potential differences in sequence recall due to corpus size—discussed in Section 6—this might also be due to differences in the formulaic language of both modes in general, such as a larger number of

Table 2: The number and frequency of dictionary-relevant formulaic sequences in Gigafida and GOS by sequence length.

	Gigafida (written)		GOS (spoken)	
no of. words	types	avg. frequency	types	avg. frequency
2	335	70.2	489	128.8
3	81	36.7	101	62.3
4	4	37.8	14	36.5
Total	**420**	**63.4**	**604**	**115.5**

structurally incomplete sequences in speech (Biber et al. 2004, Dobrovoljc 2019), including sentence fragments involving discourse particles and fillers (e.g. *eee to je* 'uhm that is', *a ne in* 'right and').

5.1. General overview

In both lists, most dictionary-relevant sequences, hereinafter also referred to as formulaic multi-word expressions, consist of two words, followed by three- and four-word sequences; no longer multi-word expressions have been identified.[11] In terms of frequency of usage, the average dictionary-relevant formulaic sequence occurs with a much higher normalized frequency than the original threshold of 20 occurrences per million, namely 63.4 occurrences per million in the Gigafida written corpus (median: 37.4/M) and 115.5 occurrences in GOS spoken corpus (median: 42.5/M). Keeping in mind the above observation on the smaller percentage of dictionary-relevant sequences in speech, the observed difference in the average frequency of usage of formulaic multi-word expressions in both modes suggests that speakers use a more limited set of formulaic multi-word expressions, but do so much more often.

In addition to high frequency of usage, formulaic multi-word expressions identified in both corpora also exhibit a distinct lexical structure with respect to content (i.e. nouns, verbs, adjectives, adverbs, numerals and abbreviations) and function words (i.e. prepositions, conjunctions, particles, interjections, pronouns and auxiliary verbs), as shown in Figure 2. In fact, more than 62% of formulaic sequences in Gigafida and 79% in GOS include at least one function word, with as much as 11% and 20% of all formulaic sequences in Gigafida and GOS, respectively, consisting of function words only (e.g. *kljub temu da*, lit. 'despite this that', i.e. despite the fact that). This is an important feature that distinguishes formulaic multi-word expressions from mainstream corpus-based phraseology research, which focuses primarily on content word combinations.

A brief overview of the semantic characteristics of the formulaic multi-word expressions under investigation groups them into sequences with a propositional function, such as referential expressions with nominal, verbal, adjectival and adverbial meaning (e.g. *v zadnjem*

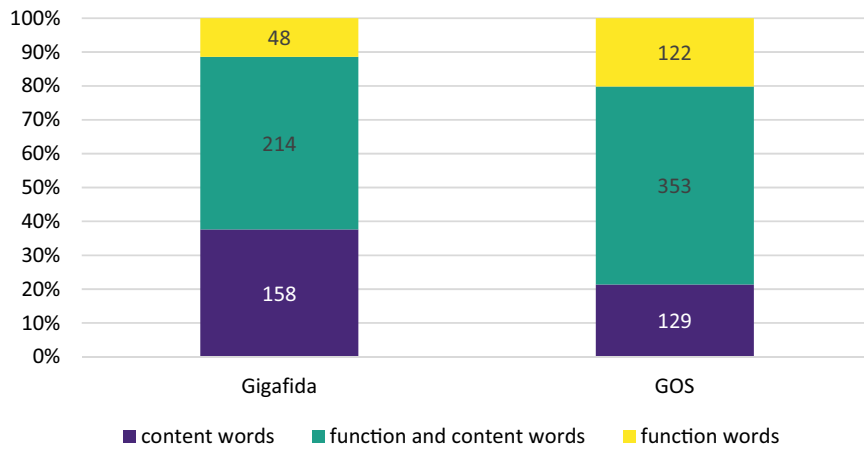

Figure 2: Dictionary-relevant formulaic sequences according to lexical structure.

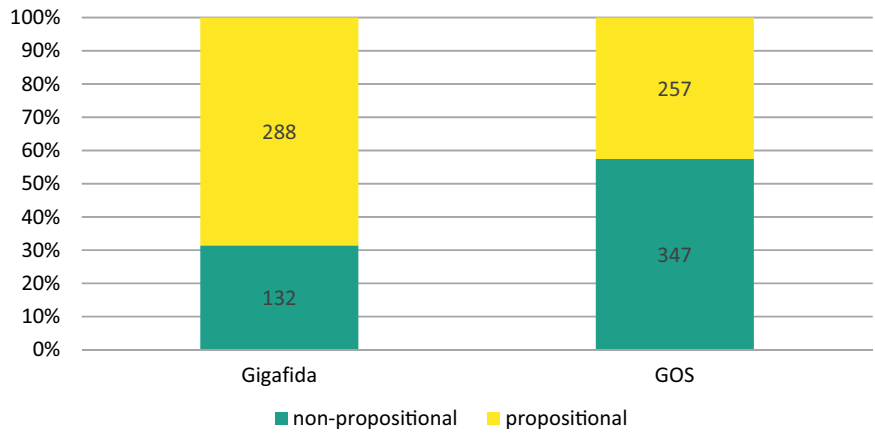

Figure 3: Dictionary-relevant formulaic sequences by semantic function.

času 'lately', *tiskovna agencija* 'press agency', *glava družine* 'head of the family'), and sequences with a non-propositional function, such as expressions used for conveying stance marking (e.g. *v bistvu* 'actually, in fact', *po besedah* 'according to', *bolj ali manj* 'more or less) and discourse/interaction organization (e.g. *na primer* 'for example, *za začetek* 'to start with', *dobro jutro* 'good morning')—a categorisation very similar to the notion of metadiscursive lexical bundles (Granger and Lefer 2016) derived from the functional taxonomy proposed by Biber et al. (2004). Figure 3 shows that a substantial amount of non-propositional formulaic multi-word expressions has been identified in both corpora; this is especially true for the GOS spoken corpus, where there is predominance of non-propositional multi-word expressions (31.4% in Gigafida and 57.5% in GOS). We discuss the multi-word expressions belonging to each group in more detail in the following sections.

5.2. Comparison of formulaic sequences in writing and speech

Although the two corpora return a similar-sized inventory of dictionary-relevant formulaic language in Slovenian with some common distributional, lexical and semantic characteristics discussed above, they exhibit a rather limited overlap, with only 149 sequences occurring in both lists (Figure 4). Thus, up to 75.3% of sequences in the GOS spoken corpus and 64.5% of sequences in the Gigafida written corpus are unique to each mode of communication, reaffirming the need to observe this lexical phenomenon in a broad and exhaustive spectrum of language use.

As expected, formulaic sequences occurring in both corpora mainly include mode-neutral, commonly used multi-word expressions, such as commonly used discourse connectives (*tako da* 'so that', *potem pa* 'then', *zato ker* 'because', *tako kot* 'just like'), multi-word prepositions (*v zvezi z* 'in relation to', *eden od* 'one of') and stance expressions (*se mi zdi* 'I think', *zelo dobro* 'very good, *tako rekoč* 'so to speak'), as well as different types of commonly used multi-word expressions with referential meaning, such as adverbials denoting time (*še enkrat* 'once more', *do konca* 'until the end', *v soboto* 'on Saturday', *dve leti* 'two years', *čim prej* 'as soon as possible') and quantity (*kar nekaj* 'quite a lot', *še bolj* 'even more', *zelo malo* 'very little'), multi-word predicates (*gre za* 'it is about', *prišlo je do* 'it

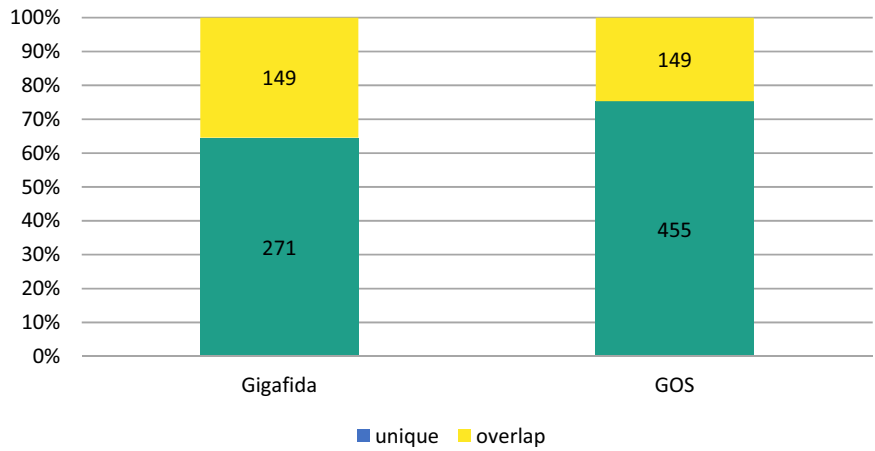

Figure 4: The number of overlapping and unique dictionary-relevant formulaic sequences in each corpus.

came to', *se je zgodilo* 'it happened') and some common nominal phrases (*rojstni dan* 'birthday', *tisoč evrov* 'thousand euros', *predsednik vlade* 'prime minister', *spletni strani* 'website').

Indeed, some of these categories also occur in the list of formulaic sequences unique to the Gigafida list, such as comparative multi-word prepositions (*za razliko od* 'in opposition to', *v nasprotju z* 'in contrast to', *v primerjavi z* 'in comparison to') and various discourse connectives (*še posebno* 'in particular', *vendar pa* 'however', *v času ko*, lit. 'in the time when', i.e. when, *tudi zato ker* 'also because', *kljub vsemu* 'despite everything'), suggesting either a greater presence of these functions in writing in comparison to speech or the development of mode-specific vocabulary to express them. Most Gigafida-unique sequences, however, reflect the distinct vocabulary related to the corpus text type distribution (Section 2). This mostly includes lexical collocations and fixed expressions connected to media (*tiskovna agencija* 'press agency', *novinarski konferenci* 'press conference'), sports (*svetovno prvenstvo* 'world championship', *v skupnem seštevku* 'in the overall standings',), legislation (*državnega zbora* 'national assembly', *človekovih pravic* 'human rights', *kaznivih dejanj* 'criminal acts'), politics (*zunanji minister* 'foreign secretary', *za notranje zadeve* 'internal affairs') and economy (*predsednik uprave* 'board manager', *električne energije* 'electricity'), but also non-propositional, metadiscursive phraseology, such as evidentials (*po navedbah* 'according to', *po njegovem mnenju* 'in his opinion'). Some unique sequences are also due to orthographic differences between the two corpora, such as collocations with numerals (e.g. *leta 2000* 'in the year 2000', *ob 17. uri* 'at 5 o'clock') and multi-word abbreviations (e.g. *prof. dr.*), which are transcribed in their unabbreviated full form in GOS.

Perhaps the most interesting set of formulaic multi-word expressions—at least in terms of redefining the focus of traditional lexicography—emerges when looking at the list of GOS-unique sequences. These include a heterogeneous set of non-propositional multi-word expressions, specific to interpersonal interaction, such as formulaic replies and questions (*kaj še* 'what else', *točno to* 'exactly', *kaj pa* 'what about, *pa ja* 'sure', *tako je* 'that's right',

a res 'really'), discourse particles and interjections (*a ne* 'right', *a veš* 'you know', *bi rekel* 'say', *boš videl* 'you'll see', *glej glej* 'well well', *daj daj* 'come on'), expressions of politeness (*hvala lepa* 'thank you very much', *dobro jutro* 'good morning', *na zdravje* 'cheers', *se opravičujem* 'excuse me'), as well as hedging devices (*ne vem* 'I don't know', *v bistvu* 'in fact', *ali pa nekaj takega* 'or something like that', *na neki način* 'in a way', *kaj pa jaz vem* 'what do I know', *če hočeš* 'if you will') and other stance-marking expressions (*po mojem* 'in my opinion', *na žalost* 'unfortunately', *več ali manj* 'more or less').

Other speech-specific lexical features include speech-specific discourse-structuring devices (*se pravi* 'that is to say', *zaradi tega ker* 'because', *drugače pa* 'otherwise', *konec koncev* 'after all', *razen če* 'unless', *če pogledamo* 'if we look at'), colloquial expressions and constructions (*bla bla* 'blah blah', *na hitro* 'quickly', *ful dobro* 'awesome', *zelo zelo* 'very very', *ne bo šlo* 'not gonna happen', *zna biti* 'it might be', *ni važno* 'it doesn't matter', *za jesti* 'to eat', *vse sorte* 'all sorts'), as well as other expressions related to topics, specific to individual speech settings and events, such as talk shows (*en aplavz* 'a round of applause', *po oglasih* 'after the commercials'), radio shows (*vse najboljše* 'happy birthday', *na cestah* 'on the road', *v studiu* 'in the studio'), political debates (*nacionalni interes* 'national interest', *prehodno stanje* 'transition state', *v javnem sektorju* 'in public sector', *vaše mnenje* 'your opinion'), sports broadcasting (*v letošnji sezoni* 'this season', *v veleslalomu* 'in giant slalom'), and everyday conversations between family and friends (*v šolo* 'to school', *na morje* 'to the seaside', *moj oče* 'my dad', *na faksu* 'in college', *v trgovino* 'to the store).

6. Evaluation of statistical measures for identification of dictionary-relevant sequences

Given the distinct distributional, lexical and semantic features of formulaic sequences in comparison to other types of multi-word expressions, on the one hand, and the differences between the two corpora, on the other, the last stage of our research involved a final comparison of the six statistical measures used (Section 3.2) in terms of their precision in identifying dictionary-relevant formulaic sequences. Similarly to related work on discourse-structuring multi-word expressions in Slovenian (Dobrovoljc 2017), we present our results in the form of a precision plot (Evert 2009), which shows the percentage of dictionary-relevant formulaic sequences among the *n* best-ranked multi-word expression candidates in each of the six lists. In contrast to comparing methods at more or less randomly selected cut-off point, such as top 100 candidates, a precision plot gives a more stable comparison across different sets of candidates—a more realistic scenario in corpus-based lexicography, where longer n-best inspections usually take place.[12]

6.1. General comparison

The precision plot for the Gigafida written corpus, shown in Figure 5, gives relatively straightforward results, as the Dice coefficient exhibits a substantially higher precision in extracting dictionary-relevant formulaic sequences in comparison to other statistical measures applied. This measure is also very consistent across n-best intervals, with precision ranging from 0.56 to 0.64—meaning that more than half of the sequences ranked using the Dice coefficient are expected to be dictionary-relevant. Other measures seem much less useful for detecting this particular type of multi-word expressions in the Gigafida corpus of written Slovenian, with none of the measures surpassing 0.34 precision. In particular,

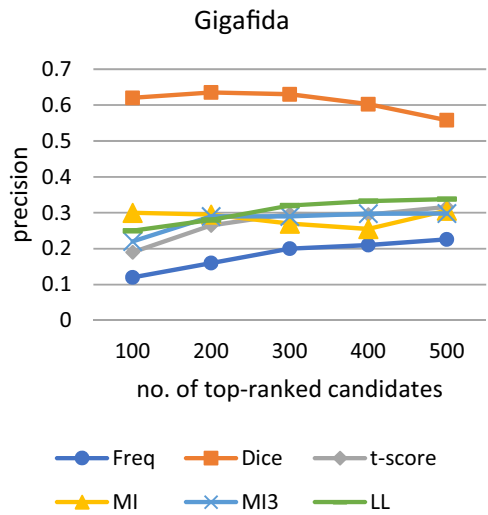

Figure 5: Precision plot for identifying dictionary-relevant formulaic sequences in the Gigafida written corpus.

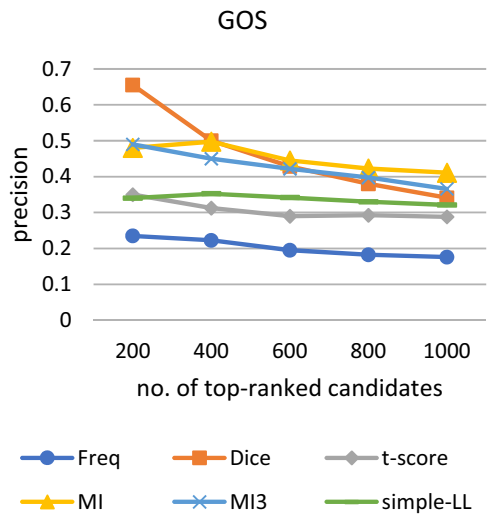

Figure 6: Precision plot for identifying dictionary-relevant formulaic sequences in the GOS spoken corpus.

similar results are observed in the MI, MI^3, LL and t-score association measures (with precision rates and relationships changing across intervals), while the worst performance in identifying dictionary relevant formulaic sequences is observed when relying on frequency alone.

However, it is much more difficult to identify a single best-performing measure for the GOS spoken corpus, as the differences between them are much smaller (Figure 6). While Dice exhibits the highest precision when looking at the top-fifth interval alone, with a fully

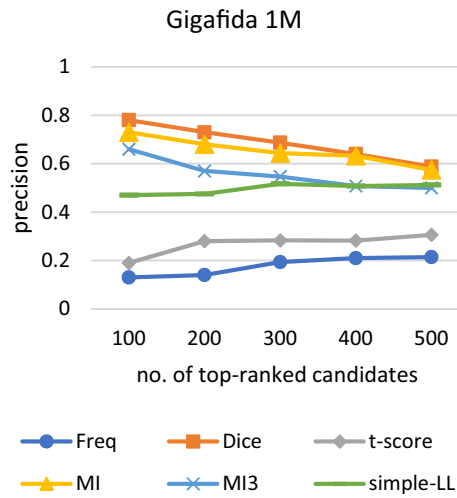

Figure 7: Precision plot for identifying dictionary-relevant formulaic sequences in the sampled Gigafida written corpus.

comparable precision rate to that of Gigafida (0.66), its precision declines when a larger number of candidates is inspected, becoming comparable to that of MI and MI^3. Slightly worse results are observed for the LL and t-score measures, while frequency remains the worst-performing measure in GOS as well. This confirms that despite the distinct formal and distributional characteristics of formulaic expressions, which have often been seen as detrimental to association-based measures, association measures are generally more appropriate for identifying dictionary-relevant multi-word expressions than ranking by frequency alone, albeit with different effect.

6.2. Controlling for corpus size and sequence length

Although the results in Figures 5 and 6 give a useful comparison of the appropriateness of individual measures for identifying this specific set of formulaic multi-word expressions in the two specific corpora, they do not imply the differences in identifying formulaic multi-word expressions in speech and writing in general. Numerous factors influence the performance of specific measures, starting with corpus size, as some measures are known to be sensitive to population size (Evert 2009, Gries 2012). In order to neutralize the effect of corpus size, we therefore performed the same evaluation on a Gigafida sample with a comparable number of tokens to that of GOS. Specifically, we sampled random Gigafida paragraphs, amounting to 1,000,035 tokens in total, in which 443 dictionary-relevant formulaic sequences were identified using an identical procedure to that of the original Gigafida corpus (Section 3).[13] The resulting precision plot for the Gigafida 1M sample in Figure 7 above confirms that some of the measures also depend on corpus size. While Dice, t-score and frequency ranking exhibit a similar precision in both the original and sampled written corpus, the MI, MI^3 and LL measures perform much better in the small Gigafida 1M sample, making the results much more comparable to that of the GOS spoken corpus (Figure 6).

Table 3: The number of extracted, top-ranked and dictionary-relevant sequences longer than two words in each corpus.

	Gigafida	Gigafida 1M	GOS
all extracted sequences (rel. freq. $\geq$ 20/M)	2,687	2,846	4,895
extracted sequences with 3 to 5 words	406	412	896
top-ranked sequences with 3 to 5 words	406	270	566
dictionary-relevant sequences with 3 to 5 words	86	81	115

However, it would be premature to conclude that the performance of these measures depends solely on the corpus size, since this only holds true if we consider formulaic sequences of different lengths (that is, sequences containing differing numbers of words) as a uniform set. This is an assumption that has attracted the attention of many researchers, who examine both the associated frequency thresholds (Cortes 2015, Chen and Baker 2016, Bestgen 2018) and the usefulness of different association measures for sequences longer than two words (Biber 2009, Simpson-Vlach and Ellis 2010, Gries 2013, Gries 2015). As discussed in Note 9, the latter is a particularly challenging issue, given there is no common consensus on the optimal method of extending association scores to sequences longer than two words.

These observations are also confirmed by our study, as the original comparison between the three corpora changes significantly when we limit our evaluation to two-word formulaic expressions only. Most strikingly, the original differences in specific measure performance between the large and the sampled Gigafida written corpus (cf. Figures 5 and 7) disappear completely, given the much better precision of the MI, MI^3 and LL measures in the large Gigafida corpus when disregarding sequences longer than two words. This supports our original assumption that it is not only the corpus size that affects the performance of specific measures, but rather the sequence length in combination with the corpus size.

To illustrate this more clearly, Table 3 presents the number of extracted, top-ranking and dictionary-relevant sequences longer than two words in each corpus: while a very similar percentage of 3- to 5-word sequences occurs above the given threshold in all three corpora (approx. 15-18% of all extracted sequences), their recall among the top-ranking candidates is uneven, with all the extracted longer sequences selected in the large Gigafida corpus and only two thirds of the longer sequences selected in the small one-million-word corpora (GOS and Gigafida 1M).

The results for specific association measures in Table 4 show that the three measures demonstrating the greatest improvement in performance when we control for corpus size and sequence length (i.e. MI, MI^3 and LL) are also the measures with the highest recall of (irrelevant) sequences longer than two words in the large, one-billion-word Gigafida written corpus. In fact, sequences longer than two words represent the vast majority of top-ranking sequences for all three measures in the large Gigafida corpus, while their percentage is much smaller in the one-million-word corpora (for example, 406 vs. 236 sequences longer than two words among the top-500 sequences ranked by MI in the Gigafida and Gigafida 1M corpus, respectively).

The question whether this bias towards sequences longer than two words in large corpora is a characteristic of the measures themselves or a consequence of the selected method

Table 4: The number of top-ranked and dictionary-relevant sequences longer than two words identified by each statistical measure in each corpus.

	Gigafida		Gigafida 1M		GOS	
measure	top-500	relevant	top-500	relevant	top-1,000	relevant
Freq	43	2	41	3	96	29
Dice	26	9	10	8	46	18
t-score	78	16	76	16	242	58
MI	406	85	236	79	481	110
MI^3	373	82	180	72	369	101
simple-LL	313	73	182	73	325	95

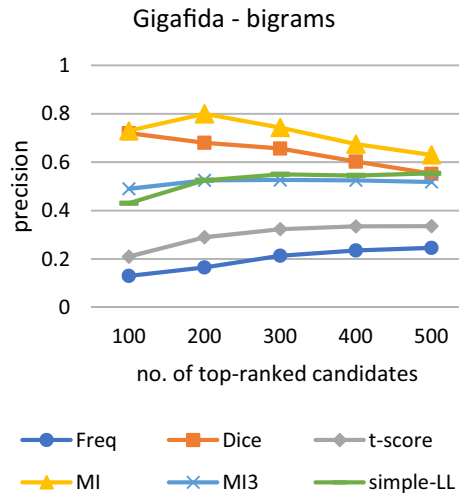

Figure 8: Precision plot for identifying two-word dictionary-relevant formulaic sequences in the Gigafida written corpus.

for their extension to sequences longer than two words (see Note 9) lies beyond the scope of this study; however, our results clearly contribute to the related discussion on the significance of sequence length in corpus-driven formulaic language identification. Future work on the topic of longer formulaic expressions should also consider lowering the frequency threshold selected in this study, with Bestgen (2018), for example, recommending a threshold above 10 occurrences per million for sequences longer than two words in corpora larger than 500,000 words.

Going back to the original question on the differences in measure performance in the spoken and written corpora, the controlled comparison in Figures 8, 9 and 10 confirms that the measures exhibit a relatively similar performance regardless of the mode, with the MI score and frequency-based ranking exhibiting the best and worst performance, respectively. However, in contrast to the GOS spoken corpus (Figure 9), where all five association measures tend to converge with a growing number of candidates, the Gigafida 1M written corpus (Figure 10) shows a more stable gap between t-score and other association

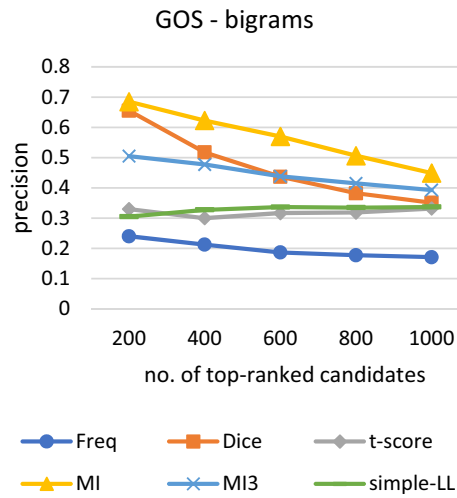

Figure 9: Precision plot for identifying two-word dictionary-relevant formulaic sequences in the GOS spoken corpus.

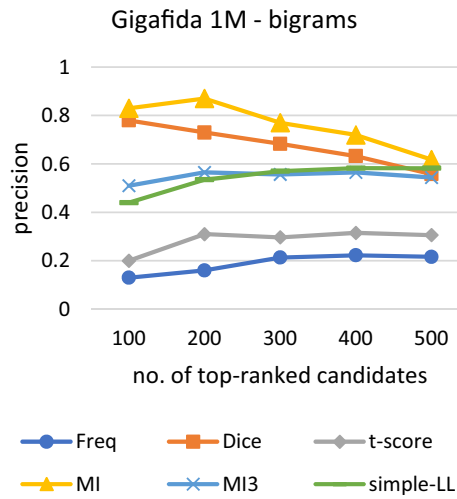

Figure 10: Precision plot for identifying two-word dictionary-relevant formulaic sequences in the sampled Gigafida written corpus.

measures. This observation is in line with the established differences in lexical and distributional features of formulaic language in each mode, namely a higher overall frequency of usage and a larger number of sequences containing high-frequency grammatical words in speech (Section 5.1), which correspond with the ability of t-score to identify word combinations that are frequent and non-exclusive (Evert 2009, Gablasova et al. 2017, Brezina 2018).

Nevertheless, relating these differences to differences in semantic characteristics of the sequences identified in both corpora, would be an oversimplification: there is no

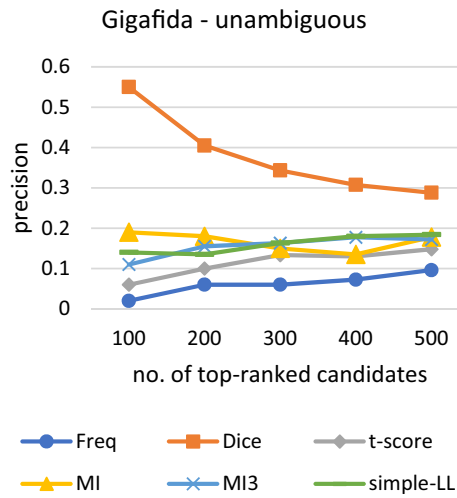

Figure 11: Precision plot for identifying unambiguous dictionary-relevant formulaic sequences in the Gigafida written corpus.

straightforward mapping between semantics, on the one hand, and lexical/distributional features on the other. Propositional formulaic expressions, which are more frequent in writing, often also contain frequent function words (with prepositional phrases being an obvious example); conversely, non-propositional expressions, more frequent in speech, may also consist of content words alone (e.g. *dobro jutro* 'good morning', *čakaj malo* 'wait a minute').

6.3. Controlling for definition

The precision evaluation for the subset of unambiguous formulaic multi-word expressions, that is the most salient multi-word units that all three annotators considered relevant, illustrates that the above results for all sequences (Section 6.1) and two-word sequences only (Section 6.2) do not depend on the final delimitation of the lexical items belonging to this category (Section 4). As shown in Figures 11–16, very similar results for all three corpora are observed for the subset of sequences, unanimously agreed on by the annotators. The only noticeable exception is the better performance of the MI score in the GOS spoken corpus (cf. Figures 15 and 9), which is expected given its characteristic capacity to identify rare, idiosyncratic word combinations, which are more likely to be unequivocally recognized as fixed multi-word combinations.

7. Discussion and conclusions

The large number of frequently recurring word sequences in reference corpora of written and spoken Slovenian identified by present study affirms previous observations on the formulaic nature of human communication in general, and spoken language in particular; it demonstrates that this is also true for Slovenian, a free word order language with complex morphology. The work presented above resulted in an extensive open-access inventory of formulaic sequences in Slovenian with an additional delimitation of dictionary-relevant formulaic sequences, consisting of multi-word expressions of various types. As such, the two

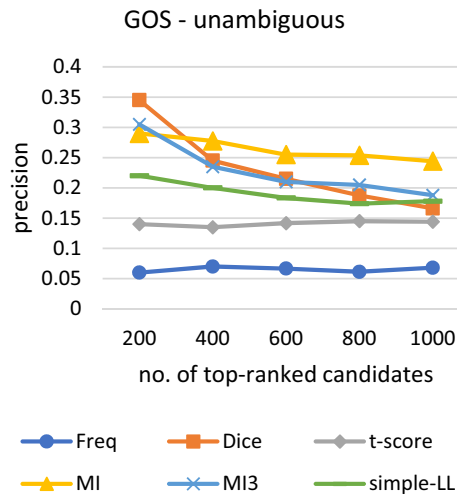

Figure 12: Precision plot for identifying unambiguous dictionary-relevant formulaic sequences in the GOS spoken corpus.

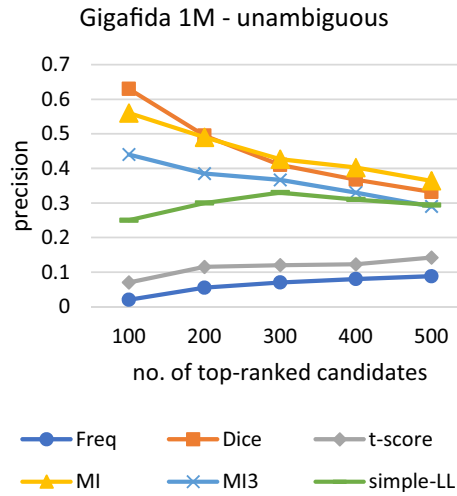

Figure 13: Precision plot for identifying unambiguous dictionary-relevant formulaic sequences in the sampled Gigafida written corpus.

lexicons present an invaluable lexical resource not only for Slovenian lexicography, but also for other disciplines interested in the study and teaching of formulaic language, such as applied linguistics, computational linguistics, neurolinguistics and pragmatics. This is especially true given the notable number of identified multi-word expressions with discourse-structuring, stance-marking and other non-propositional functions, which have often been overlooked in the existing corpus-based collections of multi-word units in Slovenian (Ljubešić et al. 2015, Gantar et al. 2016, Kosem et al. 2018).

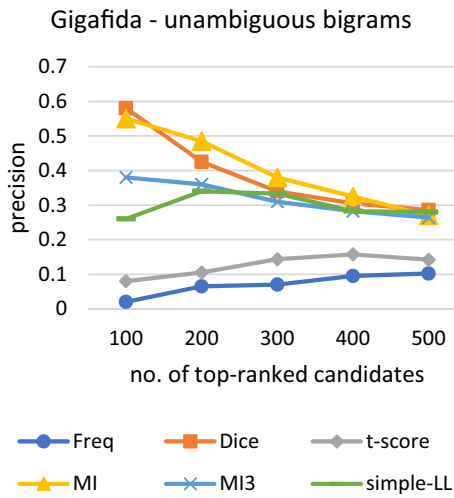

Figure 14: Precision plot for identifying unambiguous two-word dictionary-relevant formulaic sequences in the Gigafida written corpus.

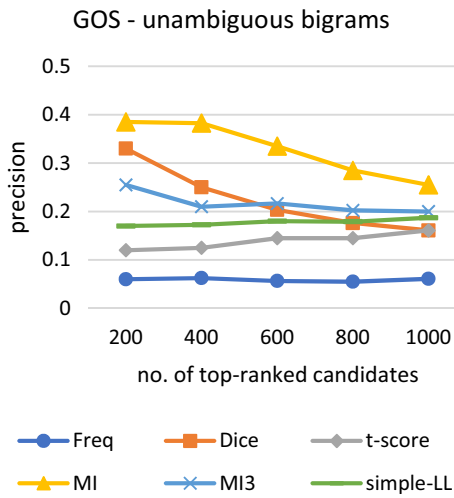

Figure 15: Precision plot for identifying unambiguous two-word dictionary-relevant formulaic sequences in the GOS spoken corpus.

In addition to language-specific contributions, however, the study presents several important findings related to the intersection of formulaic language research and lexicography in general. Firstly, the relatively low agreement between annotators on the dictionary-relevance of specific formulaic sequences empirically confirms the relativity of the concept of dictionary relevance itself, that is the kind of multi-word expressions one should expect to find in a general dictionary. This long-standing issue in lexicography (Atkins and Rundell 2008, Granger and Paquot 2008) has led to the design of detailed dictionary making protocols and classifications regarding multi-word expressions, which are based on

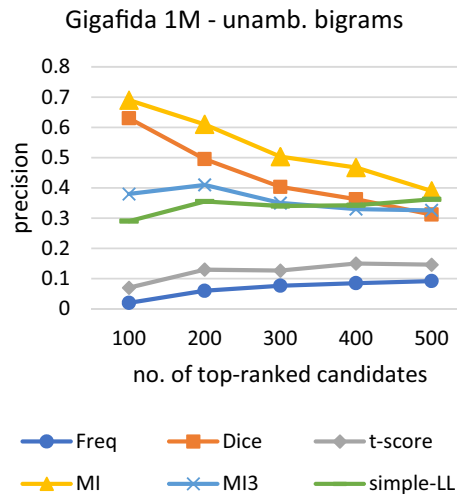

Figure 16: Precision plot for identifying unambiguous two-word dictionary-relevant formulaic sequences in the sampled Gigafida written corpus.

different linguistic and user-oriented criteria (Bergenholtz and Gouws 2013, Gantar et al. 2019b) and often part of the data collection process itself, for example with predefined headwords, syntactic structures, frequency thresholds and salience scores (Gantar et al. 2016).

Numerous formulaic sequences from this work would no doubt also emerge in such mainstream phraseology extraction procedures. The inductive, bottom-up approach taken in this study reveals, however, that statistically most salient formulaic expressions often defy cognitive-driven criteria, such as semantic non-compositionality, lexicalization, structural completeness, syntactic fixedness and headword-oriented analysis. On the other hand, they exhibit an extraordinary frequency of usage, which makes them difficult to ignore. This is best illustrated by the heterogeneous set of ambiguous formulaic expressions presented in Section 4, from grammatical collocations and colligations in writing to formulaic lexical patterns in speech. The heterogeneity of the expressions identified demonstrates that the notion of formulaicity and the challenges it poses to the lexis-grammar interface extend to a much broader set of multi-word expressions than the metadiscursive formulas alone (Granger and Lefer 2016). Regardless of the actual delimitation and subcategorization of multi-word expressions in specific future dictionaries, it is thus important that formulaic phenomena are systematically addressed in the corresponding guidelines, as well.

Secondly, our results highlight the importance of spoken data in dictionary making process and linguistic research in general, showing that—in line with the observations on the nature of formulaic language in general (Biber et al. 1999, Erman and Warren 2000, Biber et al. 2004)—formulaic *multi-word expressions* are not only more prominent in speech, but also different to those in writing. This confirms previous lexicographic observations that written corpora, irrespective of their size or structure, cannot provide sufficient insight in spoken language lexis (Siepmann 2015, Verdonik and Maučec 2016). In the case of formulaic multi-word expressions in particular, this does not only include various types of speech-specific referential multi-word expressions, but also a number of non-propositional

expressions related to discourse structuring, interaction management and speaker-hearer mitigation, which confirms Altenberg's (1998) view on formulaic expressions as convenient building blocks in spontaneous language production that pervade all levels of linguistic organization—lexical, grammatical and pragmatic. While these conventionalized expressions often seem redundant to native speakers and lexicographers alike due to their semantic compositionality, they are essential to language learners and dictionaries aimed at both receptive and productive language use (Siepmann 2008, Granger and Lefer 2016).

Finally, we contribute to this effort by evaluating the usefulness of different measures for identifying dictionary-relevant sequences in Slovenian, highlighting the importance of their careful consideration in formulaic sequence research design, given their distinct distributional and lexical characteristics, such as high frequency of usage and prevalent inclusion of grammatical words. However, with the exception of the general observation that association measures outperform the frequency-based sequence ranking in all experimental settings, the conclusions on best-performing association measures are not straightforward. While the Dice coefficient has been identified as the single best-performing measure for detecting this particular set of relevant sequences in the Gigafida written corpus—reaffirming the predominant use of this measure and its derivatives in related work on Slovenian (Gantar et al. 2016, Kosem et al. 2018)—the differences between measures are much less distinct in the GOS spoken corpus. This observation, however, can only partially be related to the distinct nature of formulaic language in each mode; rather, it is caused by the complex interplay between mathematical characteristics of specific association measures on the one hand, and a combination of corpus size and sequence length on the other. When we consider the extraction of two-word sequences only, the performance of the association measures becomes very similar for both modes, with MI—a measure with the longest-standing tradition in lexicography (Church and Hanks 1990)—exhibiting the best performance for both corpora used in this study, regardless of size. Future evaluation is thus needed to generalize these findings to different corpora, languages and types of formulaic expressions, with our evidence supporting Pecina's (2010) observation that the development of ensemble methods should be given priority over identifying a single best-performing measure, not to mention the multitude of newly emerging methods for multi-word expression identification based on richly annotated corpora and/or distributional semantics (Markantonatou et al. 2018, Cordeiro et al. 2019, Gantar et al. 2019b).

Acknowledgment

We would like to thank the IJL editors and two anonymous reviewers for their insightful and constructive comments. This study was supported by the Slovenian Research Agency through the research core funding no. P6-0411 (Language resources and technologies for Slovene language) and the research project no. J6-8256 (New grammar of contemporary standard Slovene: sources and methods).

Notes

1. https://viri.cjvt.si/gigafida/
2. https://www.sketchengine.eu/corpora-and-languages/corpus-list/
3. https://www.clarin.si/noske/run.cgi/corp_info?corpname=gfida20_dedup&struct_attr_stats=1
4. https://www.clarin.si/kontext/first_form?corpname=gfida20_dedup
5. http://www.korpus-gos.net/

6. https://www.clarin.si/noske/run.cgi/corp_info?corpname=gos&struct_attr_stats=1
7. https://www.clarin.si/kontext/first_form?corpname=gos
8. Although normalising raw frequencies to the number of occurrences per million words remains a popular approach in corpus-driven formulaic sequence identification (Bestgen 2018), several studies have emphasized the potentially problematic selection of the same frequency threshold for corpora of different sizes, arguing that more sequences are usually selected in smaller corpora than in larger ones (Cortes 2015, Gray 2016, Bestgen 2019). Given that the aim of this study is not to identify an exhaustive list of formulaic expressions in either of the corpora, but to illustrate their pervasiveness and relevance to lexicography, we adopt the predominant approach of selecting a common normalized frequency threshold for both corpora; we address this issue, however, by introducing a one-million-word sample of the Gigafida written corpus in Section 6.
9. As many researchers have pointed out, the question of extending association measures (generally developed for two-word association calculation) to sequences longer than two words is far from being a straightforward methodological decision (see Gries 2015, for example); this is why many different approaches have been proposed so far (e.g. da Silva and Lopes 1999, Van de Cruys 2011, Kilgarriff et al. 2012, Gries 2013, to name just a few), usually developed and evaluated for specific association measures. This study applies the extension method featured in the popular `mwetoolkit` tool (Ramisch 2015), due to its universal applicability to association measures of various types and its computational efficiency, as it only relies on computing the frequencies of the n-gram and the individual words it contains. We discuss the potential effect of this methodological decision in Section 6.2.
10. The annotation guidelines and the resulting lexicons for both corpora are publicly available at CLARIN.SI online repository under an open-source license (Dobrovoljc et al. 2020a, Dobrovoljc et al. 2020b).
11. The significant number of multi-word expressions longer than two words highlights the importance of looking beyond two-word combinations, usually the focus of data-driven phraseology identification. At the same time, the prevailing amount of two-word sequences carries equally important implications for the complementary field of formulaic language research, which has usually restricted itself to sequences longer than two words, mostly due to the methodological convenience of more manageable amounts of data.
12. Although precision plots aim to capture the differences in performance between individual measures, it should be noted that the successfully identified candidates are not necessarily unique to each measure. Specifically, 145 (34.5%) and 137 (22.7%) multi-word expressions have only been identified by a single association measure in Gigafida and GOS, respectively, while the majority of candidates has been identified by at least two measures. Interestingly, only 55 expressions have been identified by all six measures in GOS, while only two expressions occurred in all six top-ranking lists in Gigafida (i.e. *kljub vsemu* 'despite everything', *ves čas* 'all the time').
13. The original and the sampled Gigafida written corpora exhibit a similar number of formulaic sequences in general (2,687 vs. 2,846 formulaic sequences in the original and sampled Gigafida, respectively), as well as a similar number of top-ranking candidates (1,215 vs. 1,123) and dictionary-relevant sequences in particular (420 vs. 443).

References

Altenberg, B. 1998. 'On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations'. In Cowie, A. P. (ed), *Phraseology. Theory, Analysis, and Applications*. Oxford: Oxford University Press, 101–124.

Arhar Holdt, Š., J. Čibej, K. Dobrovoljc, P. Gantar, V. Gorjanc, B. Klemenc, I. Kosem, S. Krek, C. Laskowski and M. Robnik Šikonja. 2018. 'Thesaurus of Modern Slovene: By the Community for the Community'. In Krek, S., J. Čibej, V. Gorjanc and I. Kosem (eds), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, 401–410.

Atkins, B. T. S. and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press.

Bergenholtz, H. and R. Gouws. 2013. 'A Lexicographical Perspective on the Classification of Multiword Combinations'. *International Journal of Lexicography* 27.1: 1–24.

Bestgen, Y. 2018. 'Evaluating the Frequency Threshold for Selecting Lexical Bundles by Means of an Extension of the Fisher's Exact Test'. *Corpora* 13.2: 205–228.

Bestgen, Y. 2019. 'Comparing Lexical Bundles across Corpora of Different Sizes: The Zipfian Problem'. *Journal of Quantitative Linguistics:* 1–19.

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.

Biber, D., S. Conrad and V. Cortes. 2004. 'If You Look at . . .: Lexical Bundles in University Teaching and Textbooks'. *Applied Linguistics* 25.3: 371–405.

Biber, D., S. Johansson, G. Leech and S. Conrad. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education Limited.

Bolinger, D. 1976. 'Meaning and Memory'. *Forum Linguisticum* 1.1: 1–14.

Brezina, V. 2018. *Statistics in Corpus Linguistics*. Cambridge University Press.

Chen, Y. H. and P. Baker. 2016. 'Investigating Criterial Discourse Features across Second Language Development: Lexical Bundles in Rated Learner Essays, CEFR B1, B2 and C1'. *Applied Linguistics* 37.6: 849–880.

Church, K. W. and P. Hanks. 1990. 'Word Association Norms, Mutual Information, and Lexicography'. *Computational Linguistics* 16.1: 22–29.

Conklin, K. and N. Schmitt. 2008. 'Formulaic Sequences: Are They Processed More Quickly than Nonformulaic Language by Native and Nonnative Speakers?' *Applied Linguistics* 29.1: 72–89.

Cordeiro, S., A. Villavicencio, M. Idiart and C. Ramisch. 2019. 'Unsupervised Compositionality Prediction of Nominal Compounds'. *Computational Linguistics* 45.1:1–57

Cortes, V. 2015. 'Situating Lexical Bundles in the Formulaic Language Spectrum: Origins and Functional Analysis Developments'. In Cortes, V. and E. Csomay (eds), *Corpus-based Research in Applied Linguistics: Studies in Honor of Doug Biber*. John Benjamins, 197–216.

Cowie, A. P. 1994. 'Phraseology'. In Asher, R.E. (ed), *The Encyclopedia of Language and Linguistics*, Oxford: Pergamon Press, 3168–3171.

Dobrovoljc, K. 2017. 'Multi-Word Discourse Markers and Their Corpus-Driven Identification'. *International Journal of Corpus Linguistics* 22.4: 551–582.

Dobrovoljc, K. 2019. 'Annotating Formulaic Sequences in Spoken Slovenian: Structure, Function and Relevance'. In A. Friederich, D. Zeyrek and J. Hoek (eds), *Proceedings of the 13th Linguistic Annotation Workshop*, Association for Computational Linguistics, 108-112.

Dobrovoljc, K., R. Roblek, C. Vianello, A. Diaci and Z. Vuga. 2020*a*. 'List of Formulaic Sequences in Spoken Slovenian'. *Slovenian Language Resource Repository CLARIN.SI*.

Dobrovoljc, K., R. Roblek, C. Vianello, A. Diaci and Z. Vuga. 2020*b*. 'List of Formulaic Sequences in Standard Written Slovenian'. *Slovenian Language Resource Repository CLARIN.SI*.

Erman, B. and B. Warren. 2000. 'The Idiom Principle and the Open Choice Principle'. *Text - Interdisciplinary Journal for the Study of Discourse* 20.1: 29–62.

Evert, S. 2009. 'Corpora and Collocations'. In Lüdeling, A. and M. Kytö (eds), *Corpus Linguistics. An International Handbook*. Berlin/New York: Mouton de Gruyter, 1212–1248.

Firth, J. 1957. *'Papers in Linguistics, 1934-1951'*. London: Oxford University Press.

Gablasova, D., V. Brezina and T. McEnery. 2017. 'Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence'. *Language Learning* 67.S1: 155–179.

Gantar, P., J. Čibej and M. Bon. 2019*a*. 'Slovene Multi-Word Units: Identification, Categorization, and Representation'. *Proceedings of the Europhras 2019 Conference*. In press.

Gantar, P., L. Colman, C. Parra Escartín and H. Martínez Alonso. 2019*b*. 'Multiword Expressions: Between Lexicography and NLP'. *International Journal of Lexicography* 32.2: 138–162.

Gantar, P., I. Kosem and S. Krek. 2016. 'Discovering Automated Lexicography: The Case of the Slovene Lexical Database'. *International Journal of Lexicography* 29.2: 200–225.

Granger, S. 1998. 'Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae'. In Cowie A.P. (ed), *Phraseology: Theory, Analysis and Applications*. Oxford University Press, 145–160.

Granger, S. and M.-A. Lefer. 2016. 'From General to Learners' Bilingual Dictionaries: Towards a More Effective Fulfilment of Advanced Learners' Phraseological Needs'. *International Journal of Lexicography* 29.3: 279–295.

Granger, S. and M. Paquot. 2008. 'Disentangling the Phraseological Web'. Granger, S. and F. Meunier (eds), *Phraseology: An Interdisciplinary Perspective*. Amsterdam: John Benjamins Publishing Company, 27–49.

Gray, B. 2016. 'Lexical Bundles'. In Baker, P. and J. Egbert (eds), *Triangulating Methodological Approaches in Corpus Linguistic Research*. Routledge, 33–55.

Gries, S. T. 2012. 'Frequencies, Probabilities, and Association Measures in Usage-/Exemplar-Based Linguistics: Some Necessary Clarification'. *Studies in Language* 11.3: 477–510.

Gries, S. T. 2013. '50-Something Years of Work on Collocations: What Is or Should Be Next'. *International Journal of Corpus Linguistics* 18.1: 137–166.

Gries, S. T. 2015. 'Some Current Quantitative Problems in Corpus Linguistics and a Sketch of Some Solutions'. *Language and Linguistics* 16.1: 93–117.

Hanks, P. 2013. *Lexical Analysis: Norms and Exploitations*. The MIT Press.

Kilgarriff, A., P. Rychly, V. Kovar and V. Baisa. 2012. 'Finding Multiwords of More Than Two Words'. In Fjeld, V.F, and J. M. Torjusen (eds), *Proceedings of the 15th EURALEX International Congress*. Department of Linguistics and Scandinavian Studies, University of Oslo, 693–700

Kosem, I., S. Krek, P. Gantar, Š. Arhar Holdt, J. Čibej and C. Laskowski. 2018. 'Collocations Dictionary of Modern Slovene'. In Krek, S., J. Čibej, V. Gorjanc and I. Kosem (eds), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Znanstvena založba Filozofske fakultete Univerze v Ljubljani, 989–997.

Krek, S., Š. Arhar Holdt, J. Čibej, A. Repar and N. Ljubešić. 2019. *Gigafida 2.0 Corpus Compilation: Specifications*. Accessed on 2 September 2019. https://www.cjvt.si/gigafida/wp-content/uploads/sites/10/2019/06/Gigafida2.0_specifikacije.pdf.

Krsnik, L., Š. Arhar Holdt, J. Čibej, K. Dobrovoljc, A. Ključevšek, S. Krek and M. Robnik-Šikonja. 2019. 'Corpus Extraction Tool LIST 1.0'. *Slovenian Language Resource Repository CLARIN.SI*.

Lin, P. M. S. 2010. 'The Phonology of Formulaic Sequences: A Review'. In Wood, D. (ed), *Perspectives on Formulaic Language: Acquisition and Communication*. London: Continuum, 174–193.

Ljubešić, N., K. Dobrovoljc and D. Fišer. 2015. '*MWELex – MWE Lexica of Croatian, Slovene and Serbian Extracted from Parsed Corpora'. *Informatica* 39.3: 293–300.

Logar, N., M. Grčar, M. Brakus, T. Erjavec, Š. Arhar Holdt and S. Krek. 2012. *Korpusi Slovenskega Jezika Gigafida, KRES, CcGigafida in CcKRES: Gradnja, Vsebina, Uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Markantonatou, S., C. Ramisch, A. Savary and V. Vincze. 2018. *Multiword Expressions at Length and in Depth: Extended papers from the MWE 2017 workshop*. Berlin: Language Science Press.

Martinez, R. and N. Schmitt. 2012. 'A Phrasal Expressions List'. *Applied Linguistics* 33.3: 299–320.

Paquot, M. 2015. 'Lexicography and Phraseology'. In Biber, D. and R. Reppen (eds), *The Cambridge Handbook of English Corpus Linguistics*. Cambridge University Press, 460–477.

Pecina, P. 2010. 'Lexical Association Measures and Collocation Extraction'. *Language Resources and Evaluation* 44.1–2: 137–158.

Ramisch, C. 2015. *Multiword Expressions Acquisition: A Generic and Open Framework*. Cham: Springer.

Ramisch, C., A. Villavicencio and C. Boitet. 2010. 'Multiword expressions in the wild? The mwetoolkit comes in handy'. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010): Demonstrations*, 57–60.

Siepmann, D. 2005. *Discourse Markers Across Languages: A Contrastive Study of Second-Level Discourse Markers in Native and Non-Native Text with Implications for General and Pedagogic Lexicography*. London & New York: Routledge.

Siepmann, D. 2008. 'Phraseology in Learners' Dictionaries: What, Where and How?' *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins Publishing Company, 185–202.

Siepmann, D. 2015. 'Dictionaries and Spoken Language: A Corpus-Based Review of French Dictionaries'. *International Journal of Lexicography* 28.2: 139–168.

da Silva, J., F. da and G. P. Lopes. 1999. 'A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multi-Word Units from Corpora'. In Rogers J. (ed), *Proceedings of the 6th Meeting on Mathematics of Language*: 369–381.

Simpson-Vlach, R. and N. C. Ellis. 2010. 'An Academic Formulas List: New Methods in Phraseology Research'. *Applied Linguistics* 31.4: 487–512.

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

Tremblay, A., B. Derwing, G. Libben and C. Westbury. 2011. 'Processing Advantages of Lexical Bundles: Evidence From Self-Paced Reading and Sentence Recall Tasks'. *Language Learning* 61.2: 569–613.

Van de Cruys, T. 2011. 'Two Multivariate Generalizations of Pointwise Mutual Information'. *Proceedings of the Workshop on Distributional Semantics and Compositionality*, Association for Computational Linguistics, 16–20.

Verdonik, D., I. Kosem, A. Z. Vitez, S. Krek and M. Stabej. 2013. 'Compilation, Transcription and Usage of a Reference Speech Corpus: The Case of the Slovene Corpus GOS'. *Language Resources and Evaluation* 47.4: 1031–1048.

Verdonik, D. and M. S. Maučec. 2016. 'A Speech Corpus as a Source of Lexical Information'. *International Journal of Lexicography* 30.2: 143–166.

Viera, A. J. and J. M. Garrett. 2005. 'Understanding Interobserver Agreement: The Kappa Statistic'. *Family Medicine* 37.5: 360–363.

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

Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press.

Wray, A. 2008. *Formulaic Language: Pushing the Boundaries*. Oxford University Press.

Zwitter Vitez, A., J. Zemljarič Miklavčič, S. Krek, M. Stabej and T. Erjavec. 2013. 'Spoken Corpus Gos 1.0'. *Slovenian Language Resource Repository CLARIN.SI*.