Annotating formulaic sequences in spoken Slovenian: structure, function and relevance

Kaja Dobrovoljc
Jozef Stefan Institute, Ljubljana, Slovenia
University of Ljubljana, Slovenia
kaja.dobrovoljc@ijs.si

Abstract

This paper presents the identification of formulaic sequences in the reference corpus of spoken Slovenian and their annotation in terms of syntactic structure, pragmatic function and lexicographic relevance. The annotation campaign, specific in terms of setting, subjectivity and the multifunctionality of items under investigation, resulted in a preliminary lexicon of formulaic sequences in spoken Slovenian with immediate potential for future explorations in formulaic language research. This is especially relevant for the notable number of identified multi-word expressions with discourse-structuring and stance-marking functions, which have often been overlooked by traditional phraseology research.

1 Introduction

There has been an extensive body of research on the formulaic nature of language in the last three decades (Wray, 2013) exposing the large number of multi-word combinations that speakers seem to process as single vocabulary units (Sinclair, 1991; Wray). In addition to the most commonly studied groups of multi-word expressions, such as idioms (e.g. *break a leg*) and collocations (e.g. *heavy rain*), corpus-driven research (Biber, 2009; Conklin and Schmitt, 2012) has shown that formulaic status can also be attributed to frequently recurring sequences of words (variously termed formulaic sequences or lexical bundles), which are not necessarily structurally or semantically complete (e.g. *this means that*).

Although there is a general consensus on the need to systematically identify and formalize formulaic sequences, both for native and non-native speakers of a language (Simpson-Vlach and Ellis, 2010; Brooke et al., 2015), there has been less discussion on the optimal approach to their linguistic description and (sub)categorization. In addition, few studies that do involve some kind of quantification of formulaic sequences by syntactic, semantic or other properties, rarely report on the methodological issues related to the categorization itself.

To provide insight on the nature of formulaic language in (spoken) Slovenian, and the methodological aspects related to its linguistic categorization in general, this paper presents the annotation of formulaic sequences in the reference corpus of spoken Slovenian in terms of syntactic structure, pragmatic function and semantic relevance. After a short presentation of the corpus (Section 2) and the formulaic sequence extraction (Section 3), we present the annotation workflow and the guidelines in Section 4. Given several distinct aspects of this annotation campaign, a detailed analysis of inter-annotator disagreements is given in Section 5, followed by the presentation and discussion of the resulting list of annotated sequences in Section 6.

2 GOS corpus

GOS is the reference corpus of spoken Slovenian including approximately 120 hours (1 million tokens) of spontaneous speech in different everyday situations in public (radio and TV shows, school lessons and lectures) and non-public settings (meetings, consultations, services, private conversations).

The recordings, balanced for communication channels, situations and speaker demographics, have been manually transcribed in both pronunciation-based and standardized spelling (Verdonik et al., 2013). In this research, version 1.0 of the GOS corpus was used, freely available for download from the CLARIN.SI repository (Zwitter Vitez et al., 2013).[1]

[1]For GOS corpus browsing and listening see also the on-

Proceedings of the 13th Linguistic Annotation Workshop, pages 108–112
Florence, Italy, August 1, 2019.

3 Identification of formulaic sequences

3.1 N-gram extraction

To generate the list of formulaic sequences in GOS corpus, the LIST extraction tool (Krsnik et al., 2019) was used to extract all n-grams of length 2-5 tokens (words with normalized spelling) occurring above the frequency threshold of 20 occurrences per million. In addition to frequency counts, the tool also calculates the strength of association between words in a given n-gram, using three effect-size measures (Dice coeeficient, point-wise mutual information, and cubic mutual information) and two significance measures (t-score, simple log-likelihood), extended for multi-word combinations (Ramisch et al., 2010), as well.

3.2 N-gram ranking

There is no uniform consensus on the optimal method for measuring formulaicity in a language, with methods ranging from raw frequency counts to specific association measures (Biber, 2009; Gries, 2012), producing only partially overlapping recommendations of the most salient multi-word units in a language (Evert, 2009), including Slovenian (Dobrovoljc, 2017). Instead of opting for a single method, we narrowed the initial list of frequently recurring n-grams to the union of top-1,000 candidates ranked by each of the six methods (frequency, Dice, t-score, LL, MI, MI^3). This amounted to the final list of 2,374 formulaic sequences for subsequent annotation (Table 1).

Length	No. of types	Example
2-gram	1,808	*ja ja*
3-gram	504	*se mi zdi*
4-gram	53	*glede na to da*
5-gram	9	*osem nič osem nič nič*
Total	2,374	

Table 1: Number of identified formulaic sequences in GOS by length. (English translations from top to bottom: "yes yes", "it seems", "given the fact that", "eight zero eight zero zero".)

4 Annotation of formulaic sequences

The list of formulaic sequences has been split into multicolumn spreadsheets containing the sequences, slots for predefined labels and the hyperlinks to the corresponding concordances in GOS.

line concordancer at www.korpus-gos.net.

Each spreadsheet was manually annotated by two independent annotators (trained native speakers) based on the guidelines summarized below, with disagreements adjudicated by an expert third annotator.

4.1 Syntactic structure

In terms of syntactic structure, the sequences have been categorized into structurally complete and incomplete sequences. Structurally complete are the sequences that can be attributed a specific syntactic role in a utterance. This includes complete utterances or phrases (e.g. *to je res* "that's true", *no no* "well well"), sentence elements, such as predicates (*boš videl* "you-will see"), predicate arguments (*glava družine* "head of the family") and adjuncts (*pol ure* "half an hour"), as well as modifiers (*bolj ali manj* "more or less"), multi-word conjunctions (*zaradi tega ker* "given the fact that"), and connectives (*tako da* "so that").

Incomplete sequences, on the other hand, include fragments of the above constructions (*da bi se* "that they", *minute čez* "minutes past"), including speech-specific sequences involving fillers (*eee in eee* "uhm and uhm"), discourse markers (*ja tako da* "yes so") and repetitions (*kaj kaj* "what what").

4.2 Pragmatic function

In terms of pragmatic function, the guidelines followed previous influential functional taxonomies (Simpson-Vlach and Ellis, 2010; Biber et al., 2004), in which formulaic sequences are divided into referential expressions that reference physical or abstract entities and their properties (e.g. *to je bilo* "that was", *v skladu z* "in line with", *uradni list št.* "official gazette no.'), stance expressions that express attitudes or assessments of certainty (e.g. *na nek način* "in a way", *se mi zdi* "I think", *naj bi bil* "is supposed to", *ja ne vem* "well I don't know"), and discourse organizers that contribute to textual and interactional coherence (e.g. *kar pomeni da* "which means that", *to se pravi* "that is to say", *tako da je* "so that is", *ja ja ja* "yes yes yes").

4.3 Lexicographic relevance

In order to determine which formulaic sequences are potentially relevant for inclusion in future dictionaries and similar lexical resources for Slovenian, the annotators were asked to label the sequence in terms of its semantic relevance, i.e.

whether the sequence is a multi-word expression they would expect to find in a general dictionary intended for both native and non-native speakers of Slovenian. Specifically, they were instructed to identify multi-word expressions as opposed to free word combinations, ranging from collocations (*na internetu* "on the Internet") to fixed multi-word units with denominative (*javni sektor* "public sector"), syntactic (*kljub temu da* "despite the fact that"), or pragmatic functions (*tako rekoč* "so to speak", *dame in gospodje* "ladies and gentlemen"), regardless of semantic transparency.

4.4 Disambiguation

Only one label was allowed per category. In case of ambiguity, the annotators were advised to inspect a random sample of the concordances provided and decide for the most frequently occurring structural or functional interpretation, i.e. a primary interpretation for the given string. For semantic relevance, on the other hand, the annotators were instructed to label a sequence as relevant regardless of the frequency of this particular usage.

5 Inter-annotator agreement

On average, the two annotators agreed on 81.6% of categorization decisions, with disagreements distributed similarly across different n-gram lengths. This confirms the relatively high level of subjectivity involved in this annotation task, specific not just in terms of categories (intuitive interpretations of abstract classes), but also in terms of items under investigations (highly ambiguous and multifunctional), and the annotation setting itself (lack of immediate context, simple guidelines).

As expected, best inter-annotator agreement was observed for syntactic structure (86% absolute agreement, Cohen's Kappa 0.66), where annotators mostly disagreed on the structure of sequences occurring as both syntactically complete and incomplete units with similar frequency distribution (e.g. *veš kaj* "you know what"). Other frequent groups with structure disagreement include predicates with transitive verbs (*bom rekel* "I-will say"), numerals (*deset tisoč* "ten thousand"), repetitions (*dobro dobro* "good good"), fragments of prepositional phrases (*današnji dan* "(on) this day"), as well as strings of discourse connectives (*in s tem* "and thus"), and clause stems (*kar pomeni* "which means").

For pragmatic function, the moderate inter-annotator agreement (81% agreement, Cohen's Kappa 0.54) was mostly due to disagreement on the referential or discourse-organizing role of specific groups of sequences, such as sentence fragments containing discourse particles and connectives (*zato je* "so is", *eee mi* "uhm us"), anaphors (*na ta način* "in this way"), and words with metadiscursive meaning (*govorimo o* "we-are-talking about", *v nadaljevanju* "in the continuation"). Similarly, expressions with competing referential and stance-marking interpretations include sequences with modal verbs and adverbs (*morati* "have to", *lahko* "can"), verbs of reasoning (*vedeti* "know", *misliti* "think") and the conditional auxiliary *bi* "would".

The lowest agreement was observed for semantic relevance (78% agreement, Cohen's Kappa 0.43), where the annotators disagreed on the relevance of semantically bleach multi-word units, such as discourse particles (*bi rekel* "say"), interjections (*a ja* "oh really", *daj no* "come on") and general extenders (*ali kaj* "or what"); modified connectives (*tudi če* "even if", *takrat ko* "exactly when"); institutionalized matrix clauses (*kar pomeni da* "which means that", *predlagam da* "I suggest that"), as well as collocations involving numerals (*petnajst minut* "fifteen minutes"), deictics (*vse to* "all this", *z drugimi* "with others") and auxiliary verbs (*bomo naredili* "we-will do").

For all three categories, the competing annotations were resolved by an expert third annotator. However, given the high level of ambiguity and subjectivity inherent to the annotation task, the information on the degree of inter-annotator agreement for each decision has been preserved in the final data release.[2]

6 List of annotated sequences

In general, the distribution of specific annotation labels in the resulting list of formulaic sequences (summarized in Table 2) confirms previous empirical observations that formulaic sequences mostly consist of structurally incomplete n-grams (72.2%) with referential function (72.0%) that do not correspond to traditional dictionary-relevant multi-word expressions (74.6%). Specifically, 50.6% of sequences (1,201) have been labelled with this exact combination of characteris-

[2]The resulting list and annotation guidelines will be freely available for download through the CLARIN.SI repository in accordance with the project deliverable timeline. Project website: http://slovnica.ijs.si/

tics, among which sentence fragments (*da je* "that is", *je to* "is this", *ki je v* "which is in") prevail.

Category	Label	N
structure	complete	661
	incomplete	1,713
function	referential	1,709
	stance	306
	discourse	359
relevance	yes	604
	no	1,770
Total		2,374

Table 2: Number of annotated formulaic sequences in GOS by type.

Nevertheless, the annotated list reveals several other groups of formulaic language in spoken Slovenian with potential relevance for further linguistic inquiries and applications. From the point of syntactic structure, the structurally complete sequences (27.8%) include a diverse set of constructions, ranging from sentence elements, such as predicates (*smo rekli* "we-have said"), and adjuncts (*v Sloveniji* "in Slovenia", *dve leti* "two years"), to various types of modifiers (*še en* "another") and sentence-peripheral multi-word expressions. This last group also corresponds to the function-related findings that show a notable share of formulaic sequences with discourse-organizing (15.1%, e.g. *tako da* "so that", *na primer* "for example", *a ne* "right", *dobro jutro* "good morning") and stance-marking functions (12.9%, e.g. *se mi zdi* "it seems", *mislim da* "I think", *po svoje* "in a way"), confirming the importance of discourse structuring, interaction management and speaker mitigation in speech.

In line with the observations above, the subset of sequences recognized as dictionary-relevant (25.4%) includes a heterogeneous set of speech-specific multi-word expressions, such as formulaic replies and questions (*kaj še* "what else" *točno to* "exactly", *kaj pa jaz vem* "what do I know"), expressions of politeness (*hvala lepa* "thank you very much"), temporal expressions (*na začetku* "in the beginning", *še zmeraj* "still", *do zdaj* "until now"), intensifiers (*zelo zelo* "very very", *še bolj* "even more"), discourse-structuring devices (*pri tem* "in doing so", *prav tako* "as well"), hedging expressions (*ne vem* "I don't know", *v bistvu* "actually"), colloquial expressions (*na hitro* "quickly", *ful dobro* "awesome"), as well as other expressions related to event-specific topics (*na televiziji* "on TV", *predsednik vlade* "prime minister", *v letošnji sezoni* "this season"). Although the large majority of dictionary-relevant sequences consists of syntactically complete units, some incomplete structures have also been marked as relevant, such as multi-word prepositions (*ne glede na* "regardless of"), verbs and phrases with typical prepositions (*govorimo o* "talk about", *pride do* "come to", *hvala lepa za* "thanks for", *priložnost za* "a chance to") and discourse-structuring sentence stems (*to pomeni da* "this means that", *če pogledamo* "if we look at").

7 Conclusion

This paper presented the identification of the most frequent and statistically prominent word n-grams in the reference spoken corpus of Slovenian and their annotation in terms of syntactic structure, pragmatic function and lexicographic relevance. The annotation campaign resulted in a preliminary lexicon of formulaic sequences in (spoken) Slovenian with a high potential for future explorations in both theoretical and applied formulaic language research.

In particular in relation to the latter, our research represents an important addition to existing corpus-based collections of multi-word units in Slovenian (Gantar et al., 2016; Kosem et al., 2018; Ljubešić et al., 2015), which predominantly focus on units with propositional meaning. The large number of formulaic expressions with discourse-organizing and stance-marking functions identified in this research, however, confirms the need for future investigations of non-propositional multi-word expressions, as well.

In doing so, we plan to extend our work to the identification and annotation of formulaic sequences in written texts, drawing on the findings and observations presented above. In addition to the immediate benefits to lexicography, language teaching and natural language processing, an exhaustive inventory of formulaic sequences in Slovenian will also enable further research on methods for their identification and categorization. This also includes a comparison with manual formulaic sequence identification in corpora, bringing insight to issues related to instance-level annotation, as well.

Acknowledgments

The author acknowledges the financial support from 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*).

References

Douglas Biber. 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.

Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. If you look at ...: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics*, 25(3):371–405.

Julian Brooke, Adam Hammond, David Jacob, Vivian Tsang, Graeme Hirst, and Fraser Shein. 2015. Building a lexicon of formulaic language for language learners. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 96–104, Denver, Colorado. Association for Computational Linguistics.

Kathy Conklin and Norbert Schmitt. 2012. The Processing of Formulaic Language. *Annual Review of Applied Linguistics*, 32:45–61.

Kaja Dobrovoljc. 2017. Multi-word discourse markers and their corpus-driven identification. *International journal of corpus linguistics*, 22(4):551–582.

Stefan Evert. 2009. Corpora and collocations. In *Corpus Linguistics. An International Handbook*, volume 1, pages 1212–1248.

Polona Gantar, Iztok Kosem, and Simon Krek. 2016. Discovering Automated Lexicography: The Case of the Slovene Lexical Database. *International Journal of Lexicography*, 29(2):200–225.

Stefan Th Gries. 2012. Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics: some necessary clarification. *Studies in Language*, 11(3):477–510.

Iztok Kosem, Simon Krek, Polona Gantar, Špela Arhar Holdt, Jaka Čibej, and Cyprian Laskowski. 2018. Collocations Dictionary of Modern Slovene. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pages 989–997.

Luka Krsnik, Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Aleksander Ključevšek, Simon Krek, and Marko Robnik-Šikonja. 2019. Corpus extraction tool LIST 1.0. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Kaja Dobrovoljc, and Darja Fišer. 2015. *MWELex – MWE Lexica of Croatian, Slovene and Serbian Extracted from Parsed Corpora. *Informatica*, 39(3):293–300.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. Multiword expressions in the wild?: The mwetoolkit comes in handy. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, COLING '10, pages 57–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

John Sinclair. 1991. *Corpus, concordance, collocation*. Oxford University Press, Oxford.

Darinka Verdonik, Iztok Kosem, Ana Zwitter Vitez, Simon Krek, and Marko 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.

Alison Wray. *Formulaic Language and the Lexicon*. Cambridge University Press.

Alison Wray. 2013. Formulaic Language. *Language Teaching*, 46(3):316–334.

Ana Zwitter Vitez, Jana Zemljarič Miklavčič, Simon Krek, Marko Stabej, and Tomaž Erjavec. 2013. Spoken corpus Gos 1.0. Slovenian language resource repository CLARIN.SI.