Lexicalised and Non-lexicalized Multi-word Expressions in WordNet: a Cross-encoder Approach

Marek Maziarz[1], **Łukasz Grabowski**[2], **Tadeusz Piotrowski**[3], **Ewa Rudnicka**[1], and **Maciej Piasecki**[1]

[1]Wrocław University of Science and Technology, Wyspiańskiego 27, 50-370 Wrocław, Poland
[1]{marek.maziarz, ewa.rudnicka, maciej.piasecki}@pwr.edu.pl
[2]University of Opole, pl. Kopernika 11, 45-040 Opole, Poland
[2]lukasz@uni.opole.pl
[3]University of Wrocław, English Department, Kuźnicza 22, 50-138 Wrocław, Poland
[3]tadeusz.piotrowski@uwr.edu.pl

Abstract

Focusing on recognition of multi-word expressions (MWEs), we address the problem of recording MWEs in WordNet. In fact, not all MWEs recorded in that lexical database could with no doubt be considered as lexicalised (e.g. elements of wordnet taxonomy, quantifier phrases, certain collocations). In this paper, we use a cross-encoder approach to improve our earlier method of distinguishing between lexicalised and non-lexicalised MWEs found in WordNet using custom-designed rule-based and statistical approaches. We achieve F1-measure for the class of lexicalised word combinations close to 80%, easily beating two baselines (random and a majority class one). Language model also proves to be better than a feature-based logistic regression model.

1 Introduction

Recognition of multi-word expressions (MWEs) is one of the main tasks in the field of natural language processing (NLP) and lexicography, notably in the development of custom-designed MWE lexicons for various NLP tools or compilation of dictionaries, respectively (Gantar et al., 2018). However, MWEs are not homogeneous and there is a plethora of their definitions and operationalizations in specialized literature. For example, according to (Sag et al., 2002), the range of MWEs is very broad (including idioms, proper names, fixed phrases, compound nouns, collocations, to name but a few), as any "idiosyncratic interpretations that cross word boundaries (or spaces)" are considered to be MWEs. These idiosyncratic interpretations of a given word combination can be related to various linguistic criteria (formal, pragmatic, statistical or psycholinguistic ones), e.g, morphosyntactic patterns, constituent substitutability, semantic compositionality, frequent/recurrent use, reproducibility, collocational strength, conventionalization, pragmatic function (Woźniak, 2017; Gantar et al., 2018), and any idiosyncrasy/irregularity/non-standardness in those criteria may imply that we are potentially dealing with a MWE that is lexicalised.

We treat lexicalisation as a gradable syntax-to-lexicon process whereby a purely compositional word combination (a syntactic unit) comes to be treated as a single semantic or pragmatic unit (a lexical unit), exhibiting word-like behaviour (Lipka, 1990; Jezek, 2016; Constant et al., 2017), or – in other words – as "a conventionalized association of a contentful sense with a form at the level of the lexicon" (Van Rompaey et al., 2015, p.234). As we argue that lexicalisation is best described on a continuum, the range of multi-word expressions (MWEs) is rather wide, starting with purely compositional word combinations created *ad hoc* on one end, through collocations, to fixed phrases and idioms on the other end (Maziarz et al., 2022, 2023). However, in the theory and practice of lexicography, it is often difficult to determine which MWEs should or should not be recorded in a dictionary, i.e., treated as vocabulary/lexical units rather than mere word combinations created *ad hoc* in speech or writing. Lexicographers have to make a binary decision: either this is a *bona fide* lexical unit or not. Traditionally this status was indicated in a dictionary by place of an item in the entry and typography. More precisely, when making this binary choice, lexicographers rely on their linguistic intuition, linguistic experience and competence, contemporary and previous sources of information (dictionaries, books, corpora, etc.) to decide which MWEs to record in a dictionary, and these decisions may also differ across lexicographic traditions. For example, we looked into selected dictionaries of English and Polish and found that English lexicographers tend to record semantically compositional word combinations much more often than their Polish counterparts (Maziarz et al., 2023).

In this study, we assume that the same practical lexicographic problems apply to wordnets, as not all MWEs recorded in the Princeton WordNet can be indisputably considered to be lexicalised, e.g., such items as elements of the WordNet taxonomy (*biological group*, *animal group* etc.), quantifier phrases (*piece of furniture*, *article of furniture*), collocations (*rich people*, *psychology department*). For this reason, we need clear and operational procedures for deciding which MWEs should be included in a wordnet and which should not. By analogy to lexicography, where lexical entries in dictionaries are treated as lexical units, in this study we use the label 'multi-word lexical units' (MWLUs) for those lexicalised MWEs that should indisputably be recorded in a wordnet. Hence, our proposed procedure, combining rule-based and statistical approaches, would help us filter out MWLUs from the broad pool of MWEs recorded in WordNet (or PWN/enWN) and, in consequence, facilitate making the aforementioned dichotomous choice. The findings may help fine-tune the list of WordNet MWEs, which are often used as gold standard for NLP applications (Schneider et al., 2014; Farahmand and Martins, 2014; Riedl and Biemann, 2016). Finally, we believe that our findings will help us better understand how WordNet developers (Fellbaum, 1998) tackled the problem of recording MWEs when compiling that lexical resource.

2 Sample annotation

From Princeton WordNet and enWordNet we chose all word combinations that contained at least one space. We ruled out all proper names, as well as chemistry and biological taxonomy terms, just like we did in our previous experiment (Maziarz et al., 2022)[1]. After the filtering, we got 39,406 MWEs. Table 1 presents part of speech distribution in the dataset. 387 MWEs were randomly drawn from the remaining word combination set[2].

nouns	verbs	adjectives	adverbs
33713	4389	540	764
86%	11%	2%	1%

Table 1: POS statistics for the MWE dataset.

In order to verify the potential MWLU status of the sampled 387 word combinations, we checked how they are described in 6 dictionaries of English; we assume that if a word combination was given the headword status in the dictionaries then that indicates they are treated as multiword lexical units by native speakers of English – lexicographers – whose lexical competence surpasses that of any native speaker of English. We treat data from dictionaries thus as native speakers' response to a question: is this expression a MWLU? In other words, we believe that lexical units with headword status in dictionaries are end products of lexicalization. We are going to mention some problems with this belief below.

The dictionaries are all from established publishing houses, and will be mainly identified as such; they are: New Oxford Dictionary of English (NODE, British)[3], Merriam-Webster Collegiate (M-W, USA)[4], Collins Dictionary (CED, British)[5], New World Dictionary (N-W, USA), Collins COBUILD (COBUILD)[6], Longman Dictionary of Contemporary English (Longman)[7]. Four of those dictionaries (NODE, M-W, N-W, CED) are so-called medium, or desktop, dictionaries that are intended to be used primarily by educated native speakers of English, and two are so-called pedagogical dictionaries (COBUILD, Longman), that are intended to be used primarily by advanced learners of English or non-native speakers of English (Jackson, 2022; Cowie, 2009). We used online versions, as they are updated quite regularly in contrast to printed versions. These dictionaries were selected to ensure that we have a multi-faceted approach, and this can be shown as follows.

First, the selection was based on the needs of the intended user, as described above, but it was also based on the size and comprehensiveness of coverage. Desktop dictionaries include most of the vocabulary that educated native speakers can find

[1] We singled them out on the basis of hyponymy relation to the following top synsets: {organism 1}, {biological group 1}, {chemical element 1} and {chemical 1}.

[2] This is roughly one percent of the total 39k set. To the training set, containing 200 MWEs, used in our previous experiment (Maziarz et al., 2022), we also added 100 new MWEs as well as 50 MWEs used for final evaluation in the previous paper (already cross-checked with dictionaries). Since the 50 MWEs set represented 'MWLU' prediction class of the logistic model, we had to balance the sample to preserve the ratio of real classes. That is why additional 37 MWEs were added (recognised as non-lexicalised by the logistic classifier). We publish data sets used in this research under the CC BY 4.0 licence on GitHub (https://github.com/MarekMaziarz/MWE-recognition-in-WN).

[3] lexico.com (until August 27, 2022) and at google.com

[4] www.merriam-webster.com

[5] www.collinsdictionary.com

[6] www.collinsdictionary.com

[7] www.ldoceonline.com

in texts of English and which they may not know (that is why they reach for a dictionary), though they do not use them on their own. We used dictionaries that are meant to be used by both American or British English speakers. Pedagogical dictionaries include vocabulary of high frequency that native speakers have in their active vocabulary, the needs of a non-English user, especially from outside European culture, are not quite predictable. They have a balanced selection of British and American items, therefore we did not describe them as being British or American.

Each MWE in our sample was manually verified in terms of its occurrence as a lexical entry in any of the six dictionaries on the basis of elimination tests, starting with M-W, followed by Longman, COBUILD, CED, N-W, and concluding with NODE. In the sample we treated COBUILD, CED and N-W as one source. For example, if a MWE was recorded in M-W, then its occurrence was not checked in the remaining dictionaries, and its status as MWLU was labeled as True (T). Conversely, if the MWE was not recorded in any dictionary, then its MWLU status was labelled as False (F). We denote those non-lexicalised MWEs with the 'non-MWLU' label. Finally, in the 387 MWEs sample we obtained 144 non-lexicalised MWEs and 243 multi-word lexical units.

3 Methodology

We capitalize on and extend our earlier research (Maziarz et al., 2022), where we developed and applied a method (rule-based and statistical one using ridge logistic regression) of distinguishing between lexicalised ('MWLUs') and non-lexicalised word combinations in WordNet, taking into account selected lexicality features. In the rule-based approach, we used I-synonymy and cascade dictionary equivalents, while in the statistical approach we used MWE length measured in characters, the cosine of the angle between embedding vectors calculated for WordNet glosses and MWE lemmas, MWE sense ordering in WordNet, and the existence of equivalents in each constituent cascade dictionary. We extracted the subset of MWLUs from WordNet and its extension, enWordNet with high precision (> 70%), yet the completeness of both approaches varied. Using the rule-based approach, we obtained approximately 25% of all MWLUs, and using the statistical approach we extracted nearly 50% of the MWLUs, which translates into absolute figures as 6,4k and and 19k MWLUs respectively (ibid.). Hence, in this study we made an attempt at improving our method in order to increase the recall for extraction of the MWLU class from WordNet and enWordNet.

lemma	hypernym, definition	label
jest at	mock, subject to laughter or ridicule	0
take back	disown, take back what one has said	1

Table 2: Two examples from the sample passed to the cross-encoder. Zero means 'non-lexicalised multi-word expression', while one stands for 'multi-word lexical unit'.

This time we use a cross-encoder in the task (Reimers and Gurevych, 2019), using sentence-transformers Python library.[8] The setu4993/smaller-LaBSE model (Feng et al., 2020) rather than a large language model was used, because of a relatively small size of the manually annotated sample. We used a language-agnostic model as it could be also applied to other languages (e.g. Polish) in the future. To the cross-encoder we passed separately (i) a multi-word lemma and (ii) a synset definition (preceded by lemmas of a hypernym synset) together with (iii) the label of the sequence pair (based on entries of English dictionaries). By adding hypernymic lemmas to the semantic description (given in a definition), we attempted to provide the model with the capacity to discover semantic compositionality of a MWE, cf. (Bauer, 2019, p. 52). Two exemplar word combinations together with their semantic descriptions (i.e. a hypernym plus a definition) were presented in Table 2. We trained a classifier to automatically classify word combinations recorded in WordNet as either non-lexicalised MWEs ('non-MWLU') or multi-word lexical units ('MWLUs', that is lexicalised MWEs). Tokenizer and model inputs were truncated to 48 tokens. The number was slightly bigger than the 95th percentile of the sample definition length, cf. Fig. 1.

We fine-tuned the setu4993/smaller-LaBSE pretrained model one hundred times in a loop (with four epochs in each turn) for the need of the .632 bootstrap estimator (Efron, 1983; Jiang and Simon, 2007). In each iteration, we sampled with replacement $n_{MWLU} = 243$ examples from lexicalised MWEs and $n_{nonMWLU} = 144$ examples with replacement from the set of non-lexicalised MWEs.

[8] https://huggingface.co/sentence-transformers

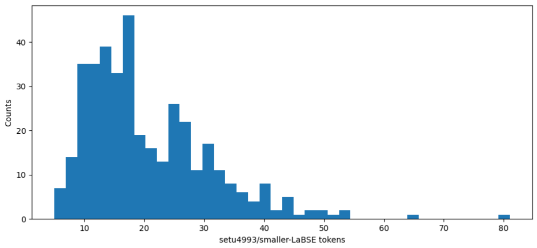

Figure 1: Histogram of lengths of sample definitions (enriched with hypernyms) in terms of LaBSE tokens. The 95th percentile for the empirical distribution equals 41, while the maximal length is 81 tokens.

In order to balance the training sample, we also additionally resampled 99 $(= 243 - 144)$ examples *without* replacement from the set of just resampled non-lexicalised MWEs. The remaining (i.e. not selected) word combinations were assigned to the evaluation (testing) data set. Thus, in the training data set both classes were balanced, while in the testing data set they were not. Within the bootstrap loop, we calculated precision (P), recall (R) and F_1 measure from confusion matrices for the language model, as well as for random and majority baselines. The results were further tested for significance with the non-parametric .632 bootstrap method (Efron, 1983; Efron and Tibshirani, 1997)[9].

The confusion matrices were obtained from Efron's .632 bootstrap rule:

$$\begin{aligned} N_i(j) = n \times Pr_i(j) = \\ n \times [0.632 \times Pr_i^{test}(j) \\ +0.368 \times Pr_i^{subst}(j)], \end{aligned} \tag{1}$$

where j $(= 1, 2, 3, 4)$ and i $(= 1, ..., B)$ denote the j-th cell of the i-th confusion matrix, $n = 387$, i.e. the whole sample size. B is the number of bootstrap iterations, in our case it is 100. Probabilities $P_i(j)$ were calculated simply as proportions of each cell counts either in testing data (out-of-bag sample, the superscript ^{test}) or in a training sample (through substitution to the model taught on the balanced sample, the symbol ^{subst}). Before calculating each cell count, the substitution sample was checked for duplicates, which were subsequently removed.

Table 3 presents the mean values of precision, recall and F_1 measure, obtained from the corresponding $\{N_i(j)\}$ matrices $(j = 1, 2, 3, 4)$. Confusion matrices presented in the table were also averaged in the following manner:

$$\begin{aligned} N(j) = \frac{\sum_{i=1}^{B} N_i(j)}{B} = \\ \frac{n \times \sum_{i=1}^{B} Pr_i(j)}{B} = \\ (n \times \sum_{i=1}^{B} [0.632 \times Pr_i^{test}(j) \\ +0.368 \times Pr_i^{subst}(j)]) \div B, \end{aligned} \tag{2}$$

where $N(j)$ stands for the j-th cell of the mean confusion matrix.

The number of epochs in each training iteration was *arbitrarily* set to 4. For BERT-like models, the number should be sufficient, although not optimal. For BERT itself, Devlin et al. (2018) recommend 2-4 epochs for fine-tuning. We selected the biggest number from that range, as we had assumed that the smaller-LaBSE model would have needed more time to optimize its weights due to a rather small annotated sample size. Our assumption was later verified with accuracy gain/loss results for each iteration (Fig. 2). The posterior evaluation revealed that setting the number of epochs to 4 almost always resulted in the highest accuracy scores. This excludes overfitting, but still our approach is prone to the problem of underfitting. We used default settings for other training parameters.

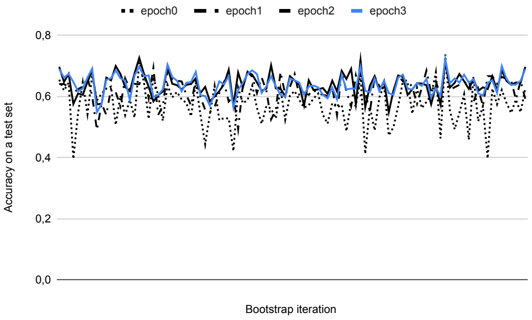

Figure 2: Accuracy gain/loss on testing sets throughout four epochs and one hundred bootstrap iterations.

4 Results

Table 3 presents the efficiency measures for the language model. In one-tailed tests the language model turned out to be better than the uniform distribution random baseline with regard to the precision, recall and F1-measure for both classes (with p-value lower than 0.01 or .025). The precision of the 'MWLU' class was also better than the majority

[9] In the same manner to (Maziarz et al., 2022).

		real		efficiency		
LaBSE model		non-MWLU	MWLU	P	R	F
prediction	non-MWLU	89.3	52.3	$\mathbf{.63}^{-}_{*}$	$\mathbf{.62}^{**}_{**}$	$\mathbf{.62}^{-}_{**}$
	MWLU	54.5	190.9	$\mathbf{.78}^{**}_{**}$	$.78_{**}$	$\mathbf{.78}_{**}$
majority baseline		non-MWLU	MWLU	P	R	F
prediction	non-MWLU	0	0	—	0	—
	MWLU	144.0	243.0	.63	$\mathbf{1}^{**}$	**.77**
random baseline		non-MWLU	MWLU	P	R	F
prediction	non-MWLU	69.7	71.9	.49	.36	.41
	MWLU	124.0	121.4	.50	.63	.55

Table 3: Confusion matrix and cross-encoder (setu4993/smaller-LaBSE) classification results for the discrimination of multi-word lexical units ("MWLUs") and non-lexicalised MWEs ("non-MWLU") in bootstrap cross-validation. Differences between the model and a random/majority baseline are statistically significant at *) <.025 or **) <.01 significance level. Comparisons with the random baseline are presented in subscript, while differences from the majority baseline are given in superscript. The presented values are averaged out over all bootstrap iteration rounds. Please note that the significance level of <0.01 was obtained when none of the bootstrap trials (out of $B = 100$ samples) found a result supporting the null hypothesis.

class baseline (with $p < 0.01$). The random baseline was obtained by sampling labels 'MWLU' and 'non-MWLU' with equal probabilities regardless real annotations, in the majority class baseline the class 'MWLU' was given to each example.

The difference between the language model and the majority class baseline was insignificant, when we compared the F1-measure for the 'MWLU' class (p = .32 in the test). The recall for the 'MWLU' class was, of course, lower than the 100% of the baseline. Comparing efficacy of smaller-LaBSE cross-encoder with a feature-based approach (Maziarz et al., 2022), we find that current F1 measure for the 'MWLU' class is much better (78% vs. 58%, p<0.01), while the measure for the 'non-MWLU' class is not worse (62% vs. 61%, $p = .31$).[10]

We retrained the model on all manual annotations and applied the fine-tuned cross-encoder to WordNet data set of 39k word combinations. Out of them, 25.5k were found to be lexicalised by the language model.

5 Conclusions

In a bootstrap cross-validation, we have found that the smaller-LaBSE cross-encoder performed very well on a manually annotated sample of nearly 400 word combinations. Both precision and recall for multi-word expressions were close to 80%, while the statistics for non-lexicalised MWEs were higher than 60%. The discrimination between lexicalised and non-lexicalised expressions worked better than two random baselines (simple uniform distribution and majority class baselines). The usage of the language model, i.e. the smaller-LaBSE cross-encoder, also improved the results obtained in (Maziarz et al., 2022) with a more traditional feature-based method. Interestingly, the cross-encoder model was given no more than bare lemmas and their synset definitions enriched only with hypernyms. No corpus frequency (a feature important in MWE recognition) was provided. We assume that the smaller-LaBSE cross-encoder (the black box *par excellance*) relied on semantic discrepancies between a word combination and its semantic description in the definition, that is, on semantic opacity/compositionality. But this assumption should be further verified in consecutive experiments in the future.

The rationale for our experiment is pivoted on lexicographic descriptions taken manually from dictionaries. A few words must be said to address possible shortcomings of this approach.

Native-speaker dictionaries are often constricted by the tradition of monolingual dictionaries in English and, what follows, by the expectations of users. This is the reservation that we voiced in Section 2: native-speaker dictionaries can include items because these items were included in some dictionaries that had been published earlier and which were quite influential. And these items are

[10] Please note that for the comparison with results from the previous experiment, we used bootstrap point estimation on mean logistic regression values, instead of paired bootstrap.

not lexical units, even though they are quite frequent in texts but the users might expect them in a dictionary. M-W and Oxford dictionaries are such influential dictionaries. In contrast, editors of pedagogical dictionaries are not constrained by tradition and one may believe that the items they include are genuine lexical items. Unfortunately, this also works in the other direction: a MWLU that is not very rare in texts may not be recorded in dictionaries because no previous dictionary recorded it. Clearly there is room for improvement both for wordnets and for "traditional" dictionaries. One obstacle for changing traditional dictionaries has been removed: they are not constrained by space, as they do not have to be printed, and may freely include MWLUs, which until recently have not been covered adequately because there was no sufficient space for them.

Acknowledgements

This research was funded by the Polish National Science Centre (NCN) under agreement no. UMO-2019/33/B/HS2/02814. Also, we would like to thank Mirosława Podhajecka for her invaluable help with data annotation.

References

Laurie Bauer. 2019. *Complex lexical units. Compounds and multi-word expressions*, chapter Compounds and multi-word expressions in English. de Gruyter.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: Multiword expression processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Anthony Cowie, editor. 2009. *The Oxford History of English Lexicography. Clarendon Press*. Clarendon Press, London.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bradley Efron. 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association*, 78(382):316–331.

Bradley Efron and Robert Tibshirani. 1997. Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.

Meghdad Farahmand and Ronaldo Teixeira Martins. 2014. A supervised model for extraction of multiword expressions, based on statistical context features. In *Proceedings of the 10th workshop on multiword expressions (MWE)*, pages 10–16.

Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. The MIT Press.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding.

Polona Gantar, Lut Colman, Carla Parra Escartín, and Héctor Martínez Alonso. 2018. Multiword expressions: Between lexicography and NLP. 32(2):138–162. _eprint: https://academic.oup.com/ijl/article-pdf/32/2/138/29012810/ecy012.pdf.

Howard Jackson. 2022. *The Bloomsbury Handbook of Lexicography. Second ed*. Bloomsbury Academic, London.

Elisabetta Jezek. 2016. *The Lexicon: An Introduction (Oxford Textbooks in Linguistics*. Oxford University Press, Oxford.

Wenyu Jiang and Richard Simon. 2007. A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. *Statistics in medicine*, 26(29):5320–5334.

Leonhard Lipka. 1990. *An Outline of English Lexicology; Lexical Structure, Word Semantics, and Word-formation*. Max Niemeyer, Tuebingen.

Marek Maziarz, Łukasz Grabowski, Tadeusz Piotrowski, Ewa Rudnicka, and Maciej Piasecki. 2023. Lexicalisation of polish and english word combinations: an empirical study. *Poznan Studies in Contemporary Linguistics*. In print.

Marek Maziarz, Ewa Rudnicka, and Łukasz Grabowski. 2022. Multi-word lexical units recognition in wordnet. In *Proceedings of the 18th Workshop on Multiword Expressions@ LREC2022*, pages 49–54.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Martin Riedl and Chris Biemann. 2016. Impact of mwe resources on multiword recognition. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 107–111.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer Berlin Heidelberg.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A Smith. 2014. Discriminative lexical semantic segmentation with gaps: running the mwe gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Tinne Van Rompaey, Kristin Davidse, and Peter Petré. 2015. Lexicalization and grammaticalization: The case of the verbo-nominal expressions be on the/one's way/road. *Functions of Language*, 22(2):232–263.

Michał Woźniak. 2017. *Jak znaleźć iglę w stogu siana? Automatyczna ekstrakcja wielosegmentowych jednostek leksykalnych z tekstu polskiego*. IJP PAN, Kraków.