Multi-word Lexical Units Recognition in WordNet

Marek Maziarz*, Ewa Rudnicka*, Łukasz Grabowski○

Wroclaw University of Science and Technology*, University of Opole○

Wybrzeże Wyspiańskiego 27, 50-370 Wrocław*****, pl. Kopernika 11A, 45-040 Opole○

marek.maziarz@pwr.edu.pl, ewa.rudnicka@pwr.edu.pl, lukasz@uni.opole.pl

Abstract

WordNet is a state-of-the-art lexical resource used in many tasks in Natural Language Processing, also in multi-word expression (MWE) recognition. However, not all MWEs recorded in WordNet could be indisputably called lexicalised. Some of them are semantically compositional and show no signs of idiosyncrasy. This state of affairs affects all evaluation measures that use the list of all WordNet MWEs as a gold standard. We propose a method of distinguishing between lexicalised and non-lexicalised word combinations in WordNet, taking into account lexicality features, such as semantic compositionality, MWE length and translational criterion. Both a rule-based approach and a ridge logistic regression are applied, beating a random baseline in precision of singling out lexicalised MWEs, as well as in recall of ruling out cases of non-lexicalised MWEs.

Keywords: multi-word expressions, Princeton WordNet, lexicality, lexicography, semantic compositionality, sentence embeddings

1. Introduction

The paper takes as its focus the lexicality status of English multi-word expressions (henceforth MWEs) found in Princeton WordNet (PWN, Fellbaum 1998) as well as in its extension enWordNet (enWN), built by a team of bilingual linguists working within the plWordNet group (Rudnicka et al. 2015). Our goal is to devise a method that can be used to distinguish between lexicalized and non-lexicalized multi-word expressions.

The term *multi-word expression* needs clarification. Multi-word expressions are neither ordinary words, nor ordinary syntactic structures, they lie somewhere in-between (Siyanova-Chanturia et al., 2017). Sag et al. (2002) define multi-word expressions as "idiosyncratic interpretations that cross word boundaries (or spaces)". Apart from the frequently mentioned idiosyncrasy or idiomaticity, researchers also emphasise other aspects of MWE nature (Constant et. al., 2017). These include its statistically non-trivial co-occurrence patterns, their status of vocabulary units similar to single words (lexicology, semantics and grammarians' point of view), as well as their particularly baffling behaviour traces, such as syntactic discontiguity, semantic non-compositionality, form variability, and form ambiguity (Calzolari et al., 2002).

MWEs are a real challenge for Natural Language Processing, since their idiosyncratic properties may lead statistical approaches astray (Sag et al., 2002). Constant et al (2017) underline the need for manually validated MWE lexicons, of higher quality than automatically extracted lists. In this paper, we take a closer look at Princeton WordNet, one of the crucial lexical sources of MWEs for English. From a lexicographic perspective, there are a number of fully compositional MWEs in WordNet that can hardly be ascribed the status of a vocabulary unit or found in any existing dictionary. These are exemplified by *rich people* 'people who have possessions and wealth (considered as a group)' or *psychology department* 'the academic department responsible for teaching and research in psychology'. In the newest initiative on the expansion and correction of Princeton WordNet called the Open English WordNet, McCrae et al. (2020: 3) postulate not to add such fully compositional MWEs in the English WordNet. As an example, they give the MWE *French Army* whose meaning can be fully deducted from the meanings of its component words *French* and *army*, both already present in the WordNet.

A different category of synsets with compositional MWEs are the ones exemplified by *biological group* 'a group of plants or animals' or *animal group* 'a group of animals' which are more units of (language) taxonomy than of language itself. They help to organise wordnet structure building top level hierarchy, yet since their lexicality status is very much different from 'ordinary' synsets, they might be tagged as 'artificial synsets', as it is done, for instance, in GermaNet (Hamp & Feldweg, 1997) and plWordNet projects (Piasecki et al., 2009).

Still another group form synsets built of MWEs of the pattern *piece/article of*, such as *piece/article of furniture* (appearing in the same synset as *furniture*) with the gloss 'furnishings that make a room or other area ready for occupancy'. The lexicality status of such MWEs is also varied. Since WordNet is considered the gold standard for NLP-oriented lexicography (MCrae et al., 2019) and the list of WordNet MWEs is used in the MWE recognition task (Riedl & Biemann, 2016; Schneider et al., 2014), as well as for evaluative purposes in the MWE extraction process (Pearce, 2001; Farahmand et al., 2014), assessing the lexicality status of MWEs in WordNet is a task worth researching.

In this paper, we use the term MWE in a broader sense as a cover term for both free and set word combinations (Zgusta, 1971). For word combinations within language vocabulary we reserve the term *multi-word lexical units* (MWLUs). We assume that MWEs that function similarly to single-word lexical units should be called MWLUs and as such recorded in dictionaries or lexical databases. The remaining MWEs should be treated as non-lexicalised ones (non-MWLUs). In other words, we argue that all MWLUs are MWEs, yet the opposite is not always true, the approach taken by Maziarz et al. (in print).

Proceedings of the 18th Workshop on Multiword Expressions (MWE 2022) n@LREC2022, pages 49–54

In short, we compare a sample of PWN and enWN MWEs with several general-purpose English dictionaries. Shedding the burden of proof on dictionary editors, we consider MWLU those MWEs that appear in at least one of the dictionaries. Employing additional linguistic and lexicographic criteria, such as MWE length, its semantic compositionality or translational equivalent criterion (for details, see Sec. 2), we scrutinise their usefulness in the task of automatic recognition of lexicalized MWEs (MWLUs) (Sec. 3 and 4). Since the selected dictionaries contain general usage vocabulary, we believe that such a list of core lexicalised MWEs can be used in NLP for evaluative purposes in the MWE extraction task.[1]

2. Data set

In order to build a data set for our experiments, we applied a rule-based procedure aimed at extracting MWEs from PWN and enWN. For these purposes, MWEs were operationally defined as sequences of graphic words (Sag et al. 2002), separated by at least one space. To extract them, we first drew all PWN and enWN synsets containing such MWE lemmas and next built a dataset of MWE lexical units (LUs, that is lemma, POS, sense number triplets). An inspection of the obtained dataset has shown a number of proper names and specialist terms. We decided to remove them from the MWE dataset, since we focus on common nouns and general-usage vocabulary. Proper names were identified by the internal *Instance* relation and by the inter-lingual I-instance relation to plWordNet synsets. Biological taxonomy and chemistry terms were singled out on the basis of hyponymy relation to the following top synsets: {organism 1}, {biological group 1}, {chemical element 1} and {chemical 1}. After the filtering, we were left with 39,406 MWEs. Their part of speech (POS) statistics are given in Table 1.

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

Table 1: POS statistics for the MWE dataset

Table 1 shows that most MWEs in the dataset are nouns (86%). Verbs make up for 11%, while adjectives and adverbs are scarce, with (2%) and (1%), respectively.

Now, to verify the lexicality of MWEs in our dataset we decided to consult general-purpose English dictionaries such as Collins[2], Longman[3], Oxford Lexico[4] and Merriam-Webster[5]. For these purposes, we drew a random sample of 200 MWEs from our dataset and looked them up in the reference dictionaries. Crucially, we paid close attention to MWE senses checking if the sense of an MWE in a dictionary matches its sense from PWN or enWN. MWEs that were recorded in at least one dictionary were considered MWLUs. The ones absent from the dictionaries were treated as non-lexicalised multi-word expressions and called non-MWLUs. The results of manual annotation are given in Table 2.

class	nouns	verbs	adject.	adverbs	sum
MWLU	114	9	0	1	124
nonMWLU	68	6	0	1	76
sum	183	15	0	2	200
%	92%	8%	0%	1%	100%

Table 2: POS and lexicality status statistics for a random 200 MWE sample

Table 2 shows that the distribution of POS in the 200 sample mirrors its distribution in the whole MWE dataset (cf. Tab. 1). As for the lexicality status, MWLUs overgrow non-MWLUs by almost a half.

Our ultimate goal was to come up with algorithms which would allow us to recognise MWLUs and non-MWLU in our dataset of 39k MWEs. To this end we applied both a rule-based approach and a statistical one. The 200 MWE sample was used to train a logistic classifier and to evaluate both approaches (Sections 3 and 4). We decided to use several features and automatically annotated with them both the small sample and the whole MWE dataset. The features were as follows:

- the presence of an inter-lingual *I*-synonymy link (Rudnicka et al., 2012) between a pair of Polish (plWordNet) - English (PWN or enWN) synsets (with the English one containing an MWE);
- the presence of an MWE lemma in a Polish-English conglomerate 'cascade' dictionary (Kędzia et al., 2013);
- the length of an MWE in terms of the number of its characters (excluding spaces) and spaces between component words;
- the cosine of an angle between an MPNet (Masked and Permuted Pre-trained Neural Network, Song et al., 2020) sentence embedding 768D vectors calculated separately for an MWE lemma itself and its WordNet gloss;
- the number of an MWE sense.

All of the above *lexicality* features were used in logistic regression, while the I-synonymy and cascade dictionary criteria were solely used in a rule-based approach.

The inter-lingual synonymy relation entails close correspondence in meanings and relation structures (Rudnicka et al., 2012). Our hypothesis is that English MWEs from synsets holding this relation are more likely to be lexicalised. The idea behind using a cascade dictionary is similar. The cascade dictionary is a collection of 12 Polish-English dictionaries and lexical resources arranged in a cascade with the most reliable on the top (Kędzia et al., 2013). It was sufficient to find an MWE in at least one of its 12 sub-dictionaries; we also used each separate dictionary as a predictor for regression (see. Sec. 4 below). Another feature - the length of an MWE is correlated with MWE's frequency in corpora.[6]

[1] Still, corpora do not contain all fixed expressions found in dictionaries and the frequency of such MWEs varies to a great extent (Svensén, 2009: 191).

[2] https://www.collinsdictionary.com/

[3] https://www.ldoceonline.com/

[4] https://www.lexico.com/

[5] https://www.merriam-webster.com/

[6] Indeed, taking SemCor 3.0 corpus (Mihalcea & Moldovan, 2001) and calculating frequency counts *f* for all MWE lexical

We assumed that longer MWEs are much rarer in usage than shorter ones. Semantic similarity between MWE's WordNet definition (gloss) and its lemma approximated semantic compositionality of the MWE (measured in the vector space using word embeddings and cosine similarity).[7] Semantically non-compositional MWEs were supposed to be defined with the use of words semantically more distant from MWE elements. The number of senses is correlated with the relative frequency of a given sense (when compared to other senses; more frequent senses are equipped with higher ranks).[8]

In Figure 1, we present descriptive statistics for five lexicality features (MWE length, MPNet cosine, I-synonymy criterion, cascade dictionary equivalent test and WordNet sense number) across our MWE sample. As shown, MWEs from dictionaries are shorter (Welch's test[9]: $t(123.14) = -4.66$, $p < .001$) and less semantically similar to their definitions (Welch's $t(160.85) = -2.20$, $p = .029$) than MWEs not found in the four reference dictionaries. MWLUs are also more frequently found in the cascade dictionary ("cascDict") than non-lexicalised MWEs (Pearson's chi-squared test: $\chi^2(1, N = 200) = 10.81$, $p = .001$), while the *I*-synonymy ("synonymy") criterion does not clearly determine class boundaries: 58% of nonMWLUs and 51% of MWLUs had their inter-lingual equivalent in plWN ($\chi^2 = 0.688$, $p = .407$).[10] Also, higher sense numbers ("senseNo") are characteristic for non-lexicalised MWEs: $t(123) = 2.31$, $p = .023$.

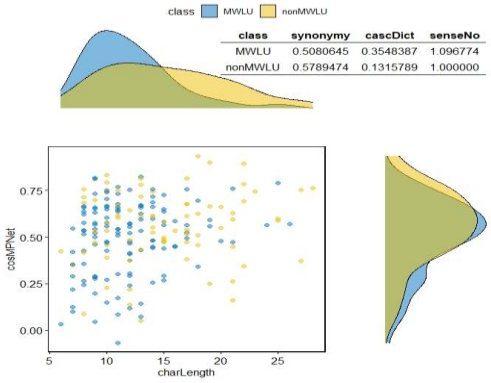

Figure 1: Dictionary-based MWE classes related to lemma length (charLength ~ MWE rarity) and similarity between an MWE definition and its lemma (cosMPNet ~ MWE semantic compositionality). The proportion of *I*-synonymy and cascade dictionary cases per class are shown in the top-right table. Abbreviations: "charLength" - the length of an MWE in characters; "cosMPNet" - cosine similarity of MPNet vectors calculated for MWE lemma and its enWN definition; "synonymy" - *I*-synonymy case; "cascDict" - MWE found in at least one of 12 cascade sub-dictionaries, "senseNo" - sense rank, i.e. the ordinal number of wordnet sense (aka *variant*).

3. Rule-based approach

We attempted to verify the validity of the inter-lingual equivalent criterion in distinguishing between MWLUs and non-MWLUs. We assumed that lexicalised MWEs should have *I*-synonymy and should be found in at least one out of 12 cascade sub-dictionaries. The rule-based procedure allowed to determine the MWLU class with high precision and low recall. In a one-tailed bootstrap percentile test (Tibshirani & Efron, 1993) for greater than zero difference the approach gained higher MWLU class precision than the uniform random baseline ($p = .027$). Also the non-MWLU class was successfully ruled out with 87% recall, clearly above the baseline ($p < .001$). On the other hand, the recall of the MWLU class was lower than the random baseline ($p < .001$). We also could not prove any difference between the rule-based model and the random baseline in terms of the non-MWLU class precision ($p = .17$).

200 MWE sample		real class		efficacy	
		MWLU	Non MWLU	P	R
rule-based	MWLU	32	10	**76%**	26%
	Non MWLU	92	66	42%	**87%**
random baseline		real class		efficacy	
		MWLU	Non MWLU	P	R
random class	MWLU	62	38	62%	**50%**
	Non MWLU	62	38	38%	50%

Table 3: Confusion matrix and efficacy of the rule-based approach. Symbols: P - precision; R - recall. Bolded values are significant at .95 confidence level in bootstrap testing

Figure 2 presents classes established using manual rules with regard to five lexicality features for the 200-MWE sample. In contrast to real classes (Fig. 1), the rules seem not to take into account neither MWEs' length (~ frequency) nor their semantic compositionality. Instead, they rely solely (quite successfully) on translational criteria.

All in all, from the initial sample of 39,406 English MWEs, we obtained 6,390 potential MWLUs with the estimated precision of 76%. To validate the result, we randomly selected 18 MWEs out of the prediction class of MWLUs. Only 2 MWEs were not found in the four reference English dictionaries, which yields a 90%

units we obtained the following values of MWE mean lengths $mL(f)$ (in characters): $mL(f = 1) = 11.1$, $mL(2) = 10.7$, $mL(3) = 10.4$, $mL(4) = 10.0$, $mL(f = 5+) = 9.8$. A similar law for simplex words is known as Zipf's law (Piantadosi et al., 2011).

[7] Pickard (2020), for instance, used the cosine similarity for the comparison between MWE lemma embedding vectors and their constituent word embeddings.

[8] For SemCor 3.0 and all lexical units (not only for MWEs) we obtained following values of mean frequency F per a given sense ordinal number #*n*: $F(\#1) = 8.0$, $F(\#2) = 7.2$, $F(\#3) = 5.8$, $F(\#4+) = 5.2$. If we take into account only MWEs, we would get something unintuitive: $F_{MWE}(\#1) = 2.7$ and $F_{MWE}(\#2+) = 2.8$ (the difference is significant at 5% significance level in the U Mann-Whitney test).

[9] Welch's *t*-test for two samples (Delacre et al., 2017).

[10] This unexpected tendency can be due to the fact that I-synonymy is a synset-level relation and not a lexical unit one. Synsets are built of one, two or even more lexical units and the degree of interlingual correspondence between specific pairs of Polish-English LUs within a pair of Polish-English synsets linked by this relation can differ.

confidence interval for the precision in the 67-97% range.[11]

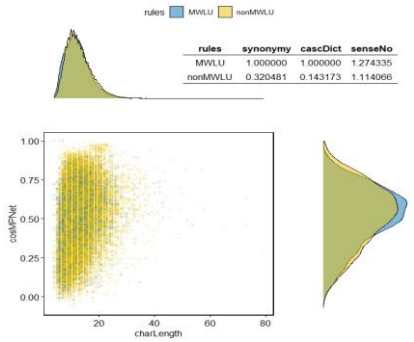

Figure 2: Classes obtained using manual rules with regard to lemma length (charLength ~ MWE rarity) and similarity between an MWE's definition and its lemma (cosMPNet ~ MWE semantic compositionality). Proportion of *I*-synonymy and cascade dictionary cases per class are shown in the top-right table. Abbreviations as in Figure 1.

4. Statistical approach

We used the same 200 MWE sample and took into consideration all calculated lexicality features. Then, ridge logistic regression was applied and precision and recall statistics were calculated in non-parametric .632 bootstrap cross-validation, with 1,000 repetitions (Efron, 1983; Efron & Tibshirani, 1997). Each time the training data set had to be balanced by resampling the smaller class of non-lexicalised MWEs with replacement. Table 4 presents the mean efficacy of the logistic regression approach. The confusion matrix was averaged from probabilities of each cell in 1,000 iterations.[12] The random baseline was obtained by assuming equal probabilities of both classes. In one-tailed test[13] the logistic model turned out to be better than the uniform distribution random baseline with regard to the precision of the MWLU class ($p < .001$), as well as the precision and the recall of the nonMWLU class ($p = .001$ and $p = .002$, respectively). The difference between the logistic model and the random baseline was insignificant, when we compared the recall of the MWLU class ($yp = .702$ in the test). The recall for the MWLU class was approximately twice as high as in the rule-based approach.

200 MWE sample		real class		efficacy	
		MWLU	Non MWLU	P	R
RLR class	MWLU	55.9	12.8	**83%**	45%
	Non MWLU	68.2	63.1	**49%**	**83%**
random baseline		real class		efficacy	
		MWLU	Non MWLU	P	R
rand. class	MWLU	62	38	62%	50%
	Non MWLU	62	38	38%	50%

Table 4: Mean confusion matrix and mean efficacy of the statistical approach. Symbols: P - precision, R - recall, RLR - ridge logistic regression, rand. - random. Bolded values are significant at .95 confidence level in bootstrap testing

We re-taught the model on the whole 200 MWE sample and used it to assess the lexicality of all MWEs in WordNet.[14] The sign of parameters of the regression function for the non-MWLU class is clearly visible in Figure 3. *I*-synonymy ("synonymy"), MWE length ("charLength"), MWE complexity ("noOfSpaces") and MPNet vectors cosine ("cosMPNet") are positively correlated with the confidence level of logistic regression and help increase the probability of ascribing non-lexicality status to an MWE, while sense variant number ("senseNo") and cascade dictionary criterion ("cascDict") decrease the probability (blue bars, "coeff"). MWE length in characters, cascade dictionary criterion and cosine similarity could be treated as the most prominent lexicality features in the regression model (Spearman's rho > 0.3).

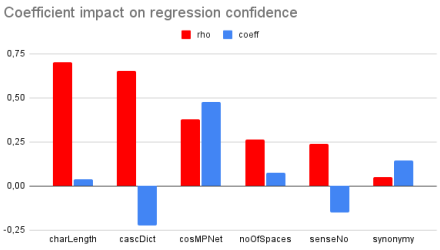

Figure 3: Relative impact of lexicality features on the logistic function for the "non-MWLU" prediction class. With blue bars we mark regression coefficients (both cascDict and senseNo are negatively correlated with the confidence measure). Lexicality features are ordered according to the value of their *absolute* correlation with the regression confidence (red bars).

Finally, 18,971 MWEs were labelled "MWLU" by the logistic model (48% of all 39,406 MWEs). Figure 4 shows descriptive statistics for five lexicality features including MWE length, its semantic compositionality, I-synonymy, cascade dictionary criterion and sense variant. The prediction class "MWLU", as compared to "non-MWLU" class, contained more frequent and less semantically

[11] These include 16 MWEs found in dictionaries: *acid precipitation*, *alkaline battery*, *computational linguistics*, *cross section*, *dialectical materialism*, *electronic paper*, *field-effect transistor*, *fire ship*, *knock over*, *lapis lazuli*, *melanocyte-stimulating hormone*, *safe sex*, *white paper*, *wind farm*, *wisdom tooth*, *yolk sac*; while two MWEs were not included in any of our reference dictionaries, namely *diplomatic mission* and *masonry heater*.

[12] From the equation

$$[1] \quad Pr_i(j) = \sum_{i=1}^{n}[0.632 \cdot Pr_i^{test}(j) + 0.368 \cdot Pr_i^{subst}(j)],$$

where $Pr_i(j)$ signifies the probability of the *j*-th cell in *i*-th repetition, the superscript ^{test} denotes the bootstrap out-of-bag testing sample, while ^{subst} refers to the bootstrap (unbalanced) training set substituted to the model taught on the very same (though balanced) sample (Efron, 1983).

[13] We calculated percentile intervals for differences between 0.632 bootstrap CV logistic regression results and random baseline estimates.

[14] More precisely speaking, those MWEs that were neither proper names, nor chemistry/biology terms.

compositional MWEs, which was intuitively expected. Also, cases of cascade sub-dictionaries were much more frequent in the "MWLU" class than in the class of "non-MWLU". *I*-synonymy, however, was more frequently found in the "non-MWLU" class. Prominent senses (with higher ranks/ lower sense ordinal numbers) occurred more frequently across the non-lexicalised class of MWEs.

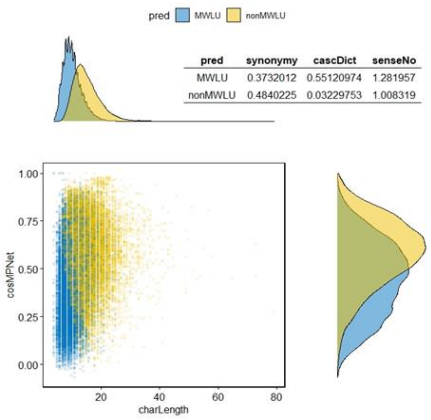

Figure 4: Logistic regression prediction classes with regard to lemma length (charLength ~ MWE rarity) and similarity between an MWE's definition and its lemma (cosMPNet ~ MWE semantic compositionality). Proportion of *I*-synonymy and cascade dictionary cases per class is shown in the top-right table. Abbreviations as in Figure 1

To validate the precision of the "MWLU" class assignments, we randomly drew 50 MWEs from the 19k "MWLU" prediction class. Only 9 word combinations were not found in any of the four reference English dictionaries, which means a 95% confidence interval for the precision in the range of 71-90%.[15] This result is in accordance with the .632 bootstrap CV precision estimate (P = 83%).

We publish datasets used in this research under the CC BY-SA 4.0 licence on GitHub (https://github.com/MarekMaziarz/Multi-word-lexical-units).

[15]These include 41 MWEs found in dictionaries: *acid precipitation, alkaline battery, anonymous ftp, Ashcan school, Babinski sign, bell ringing, blank out, cerebral peduncle, chin wag, cloven foot, come to life, cross section, double up, dust mop, easy chair, electronic paper, field-effect transistor, fire ship, fish cake, food allergy, frig around, go on, Gram stain, ice tongs, knock over, lapis lazuli, light up, on one hand, OTC stock, peel off, post exchange, procrustean bed, rat cheese, safe sex, squad room, tank suit, wet suit, white paper, wind farm, wisdom tooth, yolk sac*; 9 MWEs were not spotted in any of our four dictionaries, i.e. *butt against, chip at, dummy up, iron trap, masonry heater, pack of cards, soaking up, vena pylorica* and *vulvar slit.* Please note that validation samples drawn for both rule-based and statistical approaches partially overlapped (cf. footnote 11).

5. Conclusions

In this study, we undertook an attempt at extracting the subset of multi-word lexical units (MWLUs) from PWN and its extension, enWordNet, by using two different approaches. In a rule-based approach and logistic regression, we were able to filter out many non-lexicalised MWEs with high precision (> 70%). The completeness of both approaches differed though. The rule-based approach yielded approximately 25% of all MWLUs, while the statistical approach captured nearly 50% of the existing MWLUs. In absolute figures, we obtained 6,4k MWLUs and 19k MWLUs from WordNet, respectively. Importantly, both approaches made use of different lexicality features. As regards the rule-based approach, the features such as *I*-synonymy and cascade dictionary equivalent were used, while the statistical approach additionally capitalised on other automatically measured features: 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 also on the existence of equivalents in each constituent cascade dictionary.

The proposed procedures and methods, which were designed to extract multi-word lexical units (MWLUs) from PWN and enWN, can be applied to NLP as a gold standard list of lexicalised MWEs. For example, they can be used to evaluate MWEs extracted from corpora. Moreover, additional research is still required to develop more precise guidelines for the inclusion of MWLUs into lexical databases such as wordnets.

Our approach, though successful, for sure could be improved. The greatest need now is to broaden the recall of the MWLU class. We plan to do this by applying new features to the statistical approach. Also the translational criterion of I-synonymy could be administered more properly on the level of lexical units (not synsets).

6. Acknowledgements

This research has been funded by the Polish National Science Centre under the grant agreement No UMO-2019/33/B/HS2/02814.

7. Bibliographical References

Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., & Zampolli, A. (2002, May). Towards best practice for multiword expressions in computational lexicons. In *LREC Vol. 2*, pp. 1934-1940.

Constant, M., Eryiğit, G., Monti, J., Van Der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, 43(4): 837-892.

Delacre, M., Lakens, D., & Leys, C. (2017). "Why psychologists should by default use Welch's t-test instead of student's t-test." *International Review of Social Psychology*, 30(1): 92-101. https://doi.org/10.5334/irsp.82

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78(382): 316-331.

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

Farahmand, M., & Martins, R. T. (2014). A supervised model for extraction of multiword expressions, based on statistical context features. In *Proceedings of the 10th workshop on multiword expressions (MWE)* pp. 10-16.

Fellbaum, Ch. (Ed.) (1998). WordNet: An electronic lexical database, Cambridge, MA: MIT Press, 1998.

Hamp, B. and H. Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications.*

Kędzia, P., Piasecki, M., Rudnicka, E. and K. Przybycień. (2013). Automatic prompt system in the process of mapping plWordNet on Princeton WordNet." *Cognitive Studies*, 13: 123-141.

Maziarz, M., Grabowski, Ł, Piotrowski, T., Rudnicka, E. & Piasecki, M. (in print). Lexicalisation of Polish and English word combinations: an empirical study. *Poznań Studies in Contemporary Linguistics*.

McCrae, J., Rademaker, A. Bond. F., Rudnicka, E., and Ch. Fellbaum. (2019). English WordNet 2019 – an open-source wordNet for English. *Proceedings of Global Wordnet Conference 2019.* Oficyna Wydawnicza Politechniki Wrocławskiej: Wrocław.

McCrae, J., Rademaker, A., Rudnicka, E., and F. Bond. (2020). English wordNet 2020: improving and extending a wordnet for English using an open-source methodology. *Proceedings of LREC 2020.*

Mihalcea, R., & Moldovan, D. (2001, July). Pattern learning and active feature selection for word sense disambiguation. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems,* pp. 127-130.

Pearce, D. (2001). Synonymy in collocation extraction. In *Proceedings of the workshop on WordNet and other lexical resources, Second meeting of the North American chapter of the Association for Computational Linguistics,* pp. 41-46.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimised for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9), 3526-3529.

Piasecki M., Szpakowicz S., Broda B. (2009). A Wordnet from the ground up. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej.

Pickard, T. (2020). Comparing word2vec and glove for automatic measurement of MWE compositionality. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pp. 95-100.

Riedl, M., & Biemann, Ch. (2016). Impact of MWE resources on multiword recognition. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 107–111, Berlin, Germany. Association for Computational Linguistics.

Rudnicka, E., Maziarz, M., Piasecki, M. and S. Szpakowicz. (2012). A strategy of mapping Polish WordNet onto Princeton WordNet. In Kay, M. and Boitet, Ch. (eds.), *Proceedings of COLING 2012: Posters*. Mumbai, India, pp. 1039-1048. www.aclweb.org/anthology/C12-2101

Rudnicka, E. Witkowski, W, and M. Kaliński. (2015). Towards the methodology for extending Princeton WordNet. *Cognitive Studies 15.*

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *International conference on intelligent text processing and computational linguistics*, pp. 1-15. Springer, Berlin, Heidelberg.

Siyanova-Chanturia, A., Conklin, K., Caffarra, S., Kaan, E., & van Heuven, W. J. (2017). Representation and processing of multi-word expressions in the brain. *Brain and language*, 175: 111-122.

Schneider, N., Danchik, E., Dyer, Ch., & Smith, N. A. (2014). Discriminative lexical semantic segmentation with gaps: running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2020). MPNet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33: 16857-16867.

Svensén, B. (2009). *A handbook of lexicography* (Vol. 1235). Cambridge: Cambridge University Press.

Tibshirani, R. J., & Efron, B. (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57: 1-436.

Zgusta, L. (1971). *Manual of lexicography*. Academia, Publishing House of Czechoslovak Academy of Sciences, Prague.