A Procedural Definition of Multi-word Lexical Units

Marek Maziarz
Wrocław University of Technology, Wrocław, Poland
mawroc@gmail.com

Stan Szpakowicz
University of Ottawa, Ottawa, Ontario, Canada
szpak@eecs.uottawa.ca

Maciej Piasecki
Wrocław University of Technology, Wrocław, Poland
maciej.piasecki@pwr.wroc.pl

Abstract

Multi-word expressions evade a closed definition. Linguists and computational linguists rely on intuition or build lists of MWE types; while practical, that is scientifically and aesthetically unsatisfying. Without presuming to solve a daunting theoretical problem, we propose a decision procedure which steers a lexicographer toward acceptance or rejection of an N-gram as a lexical unit: a decision tree classifies N-grams as MWE or not MWE. It will succeed if it agrees with the native speakers' judgment. We need a small, linguistically credible set of features, to contend with the multiplicity of adequate trees. Decision tree induction works with a fixed set of annotated classification examples, but the lexical material for MWE recognition is too large to make annotation feasible. We rely on small-scale statistically significant sampling, and on intuition. Of a few decision trees produced by informed trial and error, we select one we consider best in our circumstances. That tree, deployed in a large-scale wordnet construction project, allowed us to gather dependable statistics on its usefulness in lexicographers' work. Our goal: systematic expansion of a wordnet by tens of thousands of MWEs in a manner as free of personal biases as possible.

1 Motivation

Multi-word expressions (MWEs) are present in almost every lexical resource. Their recognition can facilitate many natural language engineering tasks: information extraction, automated indexing, question answering and machine translation, to name a few. The unwavering interest in MWEs contends with the vagueness of the notion itself. There are too many, and too divergent, descriptions of just what an MWE is. Computational linguists have sought – with mixed success – a clear, "closed-formula" definition. It turns out that not only is the term "multi-word expression" not visible in linguistic literature, but that there also is no consensus on fixed phraseological expressions, non-compositional expressions, idiomatic expressions, lexicalised expressions, collocations *etc*. Most sources in traditional and computational linguistics alike seem to make do with a list of types of lexical connections in lieu of a definition. That may be practical, but it is neither scientifically nor aesthetically satisfying.

Piasecki et al. (2009) and Maziarz et al. (2013) present plWordNet, a very large wordnet and a comprehensive lexical resource for Polish. It describes most of Polish single-word lexical units and many multi-word expressions, but the coverage of the latter must increase significantly. Before that has happened, one needs to decide what *are* MWEs which merit inclusion in plWordNet, and how to make a group of lexicographers apply the definition consistently when they work on wordnet expansion.

We aim to develop a decision procedure which steers a lexicographer toward unequivocal acceptance or rejection of an N-gram as a unit in the lexical system of the language at hand. Just like a formal grammar sets precise boundaries to include things intuitively ungrammatical and exclude things intuitively grammatical, an MWE decision procedure cannot be perfect. It will be a success if it agrees to a high degree with the

Proceedings of Recent Advances in Natural Language Processing, pages 427–435,
Hissar, Bulgaria, Sep 7–9 2015.

native speakers' judgment. We do not presume to offer a solution to a theoretical problem of a clearly daunting magnitude, but we do propose a kind of practical solution which appears to work in the development of plWordNet, and which can be adapted to other languages and resources.[1]

2 An Intuitive Definition of Multi-word Lexical Unit

There is no commonly accepted definition of multi-word lexical units (MWLUs) (Granger and Paquot, 2008, p. 31). Many characteristics have been proposed as distinguishing MWLUs from regular, productive expressions in natural languages (Zgusta, 1971). Let us note two interwoven perspectives: lexicalisation and restrictedness. The former fits well the goal of building a dictionary (wordnets *are* dictionaries, among other things). An MWLU is a unit of the lexical system, stored in what is often called a *mental lexicon* (Nooteboom, 2011, p. 3).[2]

The restrictedness perspective emphasises restrictions on an MWLU's syntactic structure, meaning and use. A variety of restrictedness criteria have been proposed (Zgusta, 1971). The most frequently invoked one is semantic non-compositionality (Malmkjær, 1991, p. 291).[3] Idioms are par excellence non-compositional, and semantically the most restricted, but there are many other less pronounced cases. Restrictedness can be only considered on a continuous scale, where natural characteristic points – breaks between classes – are hard to come by.[4]

We aim to define MWLU in the spirit of lexicalisation by using means (linguistic tests) developed according to restrictedness. We will favour criteria more constrictive, and easier to work with, than complex notion of semantic non-compositionality (Svensson, 2008).

From our point of view, then, a multi-word lexical unit is fundamentally

> an expression built from more than one word, associated with a definite meaning somehow stored in one's mental lexicon and immediately retrieved from memory as a whole.

As a result, an MWLU is intuitively perceived by lexicographers as worth including in a dictionary.

Such an intuitive definition is hopelessly impractical, so, for the needs of the lexical resource building practice, we have also formulated an operational definition which takes the form of a combination of criteria implemented as linguistic tests. The criteria are meant to be applied to a MWLU candidate in appropriate, pre-defined sequences. The following sections will introduce both the criteria and the way of combining them.

3 Evaluating the Quality of the Intuitive Definition

We gave the intuitive definition (ID) from section 2, and a set of 129 monosemous word combinations, to 14 linguists who work on plWordNet. The composition of the set, collected by hand, was motivated by a few sources: Polish phraseology publications, *e.g.,* (Lewicki, 2003; Nowakowska, 2005; Müldner-Nieckowski, 2007), and a general dictionary (Dubisz, 2006). We balanced it so as to represent various stages of lexicality. The subjects answered *Yes*, *Don't know* or *No* when asked if a given *stimulus* – word combination – was a lexical unit. The 14 answers for each word combination were mapped into numbers (*Yes* $\rightarrow$ 1, *Don't know* $\rightarrow$ 0, *No* $\rightarrow$ -1) and then summed up.

Figure 1 shows that some word combinations achieved the maximum score of 14 (*e.g., szkoła podstawowa* 'primary school'), and some the minimum of -14 (*e.g., pies Marka* 'Mark's dog'). The intermediate possibilities include *mały ekran* 'the small screen = TV', *samolot transportowy* 'a freight plane' and *zjawisko językowe* 'linguistic phenomenon', all scoring at 0. The distribution is clearly not normal.

Consider a linguist's choices as they arise from probability distribution. The Central Limit Theorem says that if independent random variables X_i, $i = 1, 2, \ldots$ have the same distribution, finite mean and variance, then the sum $X_1 + X_2 +$

[1] We believe that it is unique to rely on machine learning algorithms and to use Cohen's κ to test the whole procedure on many subjects. That is why there is no related-work section in this paper. Müldner-Nieckowski (2003) arranged certain criteria into one procedure, but his proposal differed in several ways: (1) he based the procedure only on his knowledge and intuition, (2) he applied points to a candidate MWE, then a score was calculated and the final decision made according to an arbitrary threshold, (3) he proposed no tests of decision consistency between many people.

[2] «The basic prerequisite for according lemma status to a multi-word items is that it has undergone some kind of lexicalisation, *i.e.,* that it has been stored in our mental lexicon as a unit.'» (Svensén, 2009, pp. 102-3).

[3] The meaning of non-compositional MWLU cannot be reproduced from the meaning of its parts (Granger and Paquot, 2008, p. 31).

[4] «It is impossible to establish a sharp boundary between free combinations and set ones. It can be shown that there are different degrees of 'setness'.» (Zgusta, 1971, p. 154).

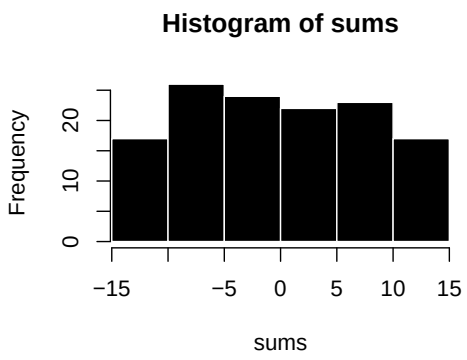

Figure 1: A histogram of summed decisions of 14 linguists for 129 word combinations, in groups of five. The Shapiro-Wilk test shows that the distribution differs from the normal distribution (W = 0.9605, p-value = 0.0008472).

... converges to normal distribution N (Meester, 2008, p. 179). The empirical distribution of $X_1 + \ldots + X_{14}$ shown in Figure 1 obviously is not normal, which leads us to the finding that, although independently, linguists reacted similarly to the same word combination stimuli.

It seems that linguists have an *intuition* on the lexicality of multi-word combinations. What many of the subjects share is perhaps the same, or quite similar, mental lexicon. But is this intuition consistent from one subject to another?

The histogram in Figure 1 suggests that many word combinations are judged inconsistently (about 1/3 of them score close to 0). There is rather low inter-annotator agreement between pairs of subjects on this set of 129 word combinations. We assume that both -1 and 0 signal non-lexical multi-word combinations, while +1 means a lexical unit. On the overall list of 129 items, the average Cohen's κ = 0.317, a "fair" value according to Landis and Koch (1971, p. 165). In computational linguistics such a result could be rejected, because a common assumption puts the κ value which guarantees reliable results at no less than 0.8, with κ above 0.67 deemed only tolerable.[5] For phraseologists, however, a value a little over 0.3 is not surprising at all, because everyone has their own mental lexicon or intuition on lexical items (Müldner-Nieckowski, 2003). The question

[5]Reidsma and Carletta (2008) show that this rule of thumb does not always work. Sometimes lower κ makes the results reliable, sometimes even $\kappa \geq 0.8$ does not suffice. The authors recommend checking whether differences between annotators are systematic or random.

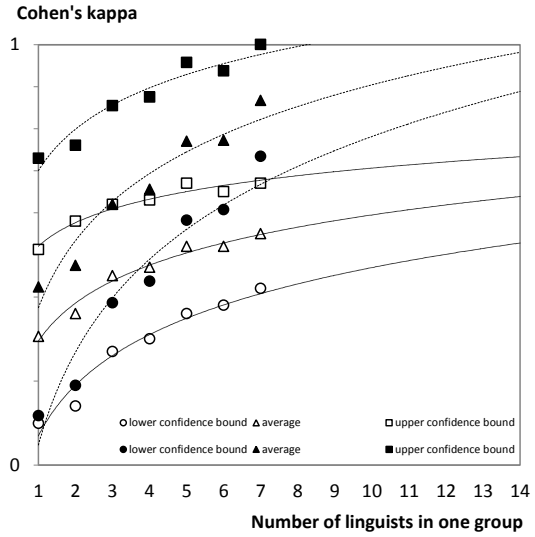

Figure 2: Cohen's κ between the summed decisions of two groups of $k = 1..7$ linguists. Confidence intervals at $\alpha = 0.05$. We start the diagram from $k = 1$, the ordinary two-subject inter-annotator agreement (one linguist per group). White figures stand for all word combinations, black figures – for word combinations with certain status: $\left|\sum_{i=1}^{14} X_i\right| > 3$.

now arises whether these lexicons are comparable, and how to achieve better κ values.

It is an open question whether averaging the independent judgments of two or more subjects increases κ. To seek an answer, we gathered the answers of 14 linguists and averaged their decisions within two independent groups. Given a matrix of judgments with 129 rows (word combinations) and 14 columns (linguists), we sampled $2k$ columns without repetition, k going into group A and k into group B, k = 1..7. Next, we sampled 129 rows with repetition for all $2k$ linguists, summed up group A and group B separately. A positive sum made the word combination an LU, a non-negative sum – not an LU. Cohen's κ was calculated for groups A and B as if they were individual annotators. This sampling was repeated 10,000 times. For a 95% simple percentile confidence interval, we took the values #250 and #9751 (DiCiccio and Efron, 1996; DiCiccio and Romano, 1988; Artstein and Poesio, 2008). The lower and upper confidence bounds appear in Figure 2 (white figures, dotted lines).

It is noticeable that increasing k increases κ. For 7-subject virtual teams, we have a confidence interval of 0.42 to 0.67, much better, but still be-

low the level of agreement desirable in computational linguistics. The extrapolation of confidence bounds into higher values of k via logarithmic functions ($R^2 \geq 0.95$) gives even higher κ values, CI = (0.53, 0.74) for k = 14. If the logarithmic extrapolation works for k > 14, we can say that we finally reach acceptable κ.

If we remove the least stable word combinations,[6] Cohen's κ increases a good deal, as Figure 2 (black figures, dashed lines) shows. As we can see, we get very good κ values even for teams of 5-7 people.

It is worth remarking that such increasing curves for κ would never emerge by chance. This proves that our definition mirrors linguists' intuition quite well. If we gathered many linguists, gave them the definition and asked to agree on the status of each multi-word combination, we would end up with a fairly appropriate dictionary of multi-word lexical units.

We have studied the quality of decision procedures by comparing their results with averaged decisions of 14 ($L1$-varia) or 5 ($L2$-plWN, $L3$-NAcoll) linguists (after the removal of word combinations of the least certain status, *i.e.,* $\left|\sum_{i=1}^{14} X_i\right| \leq 3$ and $\left|\sum_{i=1}^{5} X_i\right| \leq 1$ respectively.

4 Using the Intuitive Definition

Grouping linguists into teams of 14 (or more) significantly reduces inconsistencies of their decisions. Despite this advantage, it must be said that such a procedure is unacceptably labour-intensive. If we want to build a large dictionary of multi-word items, we must seek another solution.

We have decided to use the intuitive definition only to calibrate a special procedure of applying the status of lexicality or non-lexicality to virtually every given word combination. We posited four requirements for such a procedure.

- It must reflect common intuition of a team of linguists adjusting their decisions on what is and what is not lexical. (This is measured by precision, recall and F_1-score.)
- It must guarantee that linguists who follow it will work consistently. (We check this point using Cohen's κ.)
- It must agree with linguistic and phraseological knowledge about lexicality. (This condition was met up-front: our procedures build on criteria taken from phraseology literature.)
- It must not be too complicated. (That is why we tended to prefer simpler models over more complex ones. This criterion relies on the procedure designer's intuition.)

The calibration was performed on three sets of word combinations:

1. $L1$-varia – the already discussed set of 129 monosemous word combinations taken from various sources, annotated by 14 people (section 3). This set is the most universal, because of the largest annotator group but also various multi-word combination types (idioms, terms, compounds, collocations and loose word combinations).
2. $L2$-plWN – the set of 200 multi-word items randomly taken from plWordNet, annotated by 5 people. This representative sample set contains mainly multi-word lexical units but also some non-lexical ones, inherited from the "pre-theoretic" early stages of the development of plWordNet.
3. $L3$-NAcoll – the set of 200 Noun+Adjective collocations, drawn randomly from a set of 10,000 best Noun+Adjective pairs according to a point-wise mutual information algorithm (Bouma, 2009). The set was also annotated by 5 linguists. This type is the most common in plWordNet (almost 50% of multi-word lexical unit instances – see Figure 3).

After having many subjects annotate the lists, we chose a few people from each group of annotators (3 from the $L1$ group, 2 from $L2$ and 4 from $L3$) and gave them several linguistic criteria out of which we wanted to construct the final procedure. Here are the criteria, operationalised in the form of substitution tests.

- The specialist character of a word combination – specialist register and its terminological character (Zgusta, 1971, p. 144).
- Non-compositionality of a word combination – its metaphoric character (Müldner-Nieckowski, 2007, 117), hyponymy between a word combination and its syntactic head, ability to be paraphrased (Zgusta, 1971, p. 144).
- Syntactic criteria of non-separability and fixed word order (Bright, 1992, pp. 286-8),

[6] Those with the sum of 14 votes oscillating around 0. We did throw out word combinations with $\left|\sum_{i=1}^{14} X_i\right| \leq 3$.

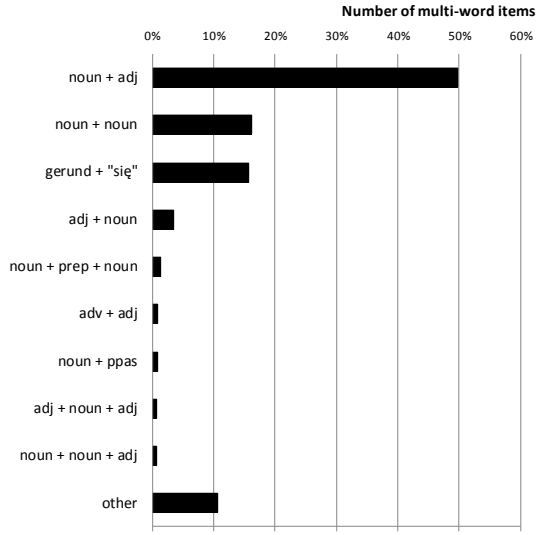

Figure 3: Multi-word items in plWordNet by syntactic pattern.

(Pike, 1967), (Müldner-Nieckowski, 2007, p. 100).

- A derivational criterion of the possibility of forming a one-word derivative from a multi-word lexical unit base (Svensén, 2009, pp. 102-103).
- An ontological criterion of a multi-word combination being a sign for a unique object type (Szober, 1967, p. 113), (Svensén, 2009, pp. 102-103).

We excluded from tests those criteria which seemed unproductive, *e.g.,* having one-word counterpart in another language.[7] Equipped with this criterion repertoire, the linguists described every multi-word combination.

The three matrices of linguists' choices now consist of independent variables (linguistic criteria) and a predicted variable (the level of lexicality measured by the sum of linguists' choices). These matrices were given to machine learning algorithms which tried to perform the best classification from linguistic criteria into the lexicality score. We worked in the Weka environment (Hall et al., 2009), and found that decision tree induction gave the best results.

5 Planting Trees

The decision trees, the embodiment of our procedure, were evaluated in accordance with the four requirements listed in section 4. We applied Cohen's κ measure for inter-annotator agreement and standard model efficiency measures. Figure 4 presents one of those trees, made by Weka's J48 decision tree induction for the set $L2$-plWN.

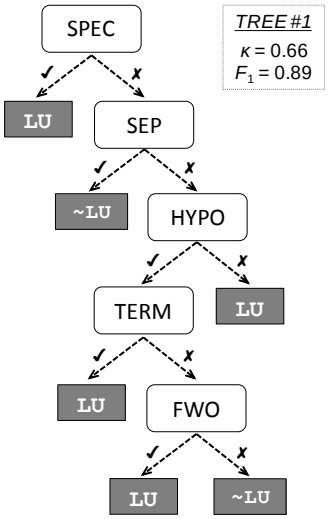

Figure 4: Tree #1 for the set $L2$-plWN. Legend: SPEC – specialist register, SEP – separability, HYPO – hyponymy between a word combination and its syntactic head, TERM – terminology, FWO – fixed word order, LU – MWLU, ∼LU – loose combination.

The results were initially promising: averaged $F_1 = 89\%$, $P_{LU} = 96\%$, $R_{LU} = 84\%$. It turned out fast, however, that trees adequate for LUs already in plWordNet were disappointing when applied to the more general set of $L1$-varia word combinations. Apart from acceptable Cohen's κ, we had inferior procedure performance, with F_1 = 52%. Table 1 shows more details.

Tree #2 behaved similarly: good κ and good model performance for LUs in plWordNet did not turn into a good F_1-score for the $L1$-varia set. Cohen's κ was reasonable. See Table 1 again.

We then started from a more general set $L1$, but good trees were hard to obtain. The best was tree #3, but the results were inconclusive (Table 2): very good behaviour, but κ still low. What is more important, the tree was very complicated, so it would be difficult to improve κ.

At the end of the day, we found ourselves with a couple of trees made for the $L2$ set which worked poorly on $L1$, and one tree for $L1$ which also worked on $L2$ but with moderate values of κ.

That is why we have decided to construct a de-

[7] What is lexical in one language need not be lexical in another. Otherwise, there would be no lexical gaps.

Procedure	tree #1			tree #2		
	$L2$-plWN, κ=0.66			$L2$-plWN, κ=0.67		
	P	R	F_1	P	R	F_1
LU	96%	84%	90%	90%	78%	83%
∼LU	81%	96%	88%	80%	91%	85%
Averaged	88%	90%	**89%**	85%	84%	**84%**
	$L1$-varia, κ=0.58			$L1$-varia, κ=0.59		
	P	R	F_1	P	R	F_1
LU	65%	34%	45%	86%	21%	34%
∼LU	47%	80%	59%	47%	96%	63%
Averaged	56%	57%	**52%**	66%	58%	**48%**

Table 1: Precision, recall and F_1-measure of trees #1 and #2 for the sets $L2$-plWN and $L1$-varia.

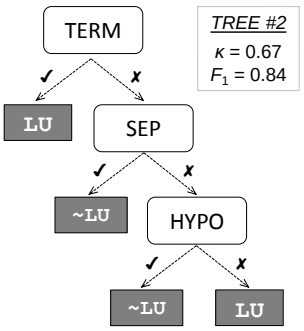

Figure 5: Tree #2 for the set $L2$-plWN. Legend: TERM - terminology, SEP – separability, HYPO – hyponymy between a word combination and its syntactic head, LU – MWLU, ∼LU – loose combination.

cision tree for the most frequent structural type Noun+Adjective. Since the performance of the tree on the set $L3$-NAcoll was very good, we generalised it onto other structural types and checked it on the most general set $L1$-varia. We also inspected the κ values for the $L2$-plWN set. A brief description appears in section 6.

6 Collocations of the Noun+Adjective Type

The ability to have a paraphrase is a criterion aimed at detecting non-compositionality, similarly to criteria for a word combination being a hyponym of its own syntactic head (HYPO in Figures 4 and 5) and metaphoricity (MET in Figure 6). A word being specialist or being a term are also very close criteria. The combination of non-compositionality and terminology can be traced in all of the trees we present here.

Among many other trees, Weka gave us a lex-

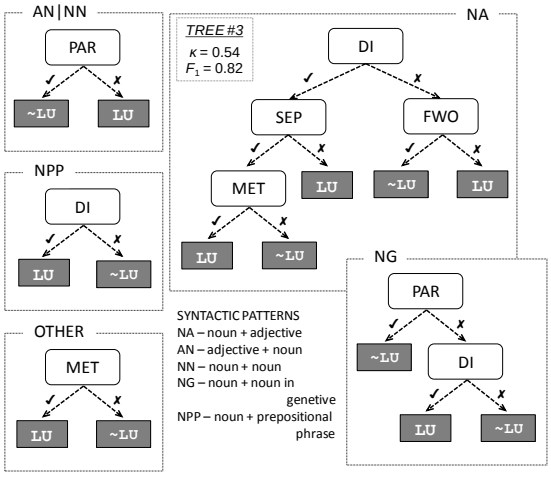

Figure 6: Tree #3 for the set $L1$-varia. Legend: DI – intuitive definition, SEP – separability, HYPO – hyponymy between a word combination and its syntactic head, MET – metaphoricity, PAR – paraphraseability, LU – MWLU, ∼LU – loose combination.

tree #3			
$L1$-varia, κ=0.54	P	R	F_1
LU	93%	74%	82%
∼LU	74%	93%	82%
Average	82%	82%	82%
$L2$-plWN, κ=0.49	P	R	F_1
LU	100%	92%	96%
∼LU	67%	100%	81%
Average	84%	96%	88%

Table 2: Precision, recall and F_1-measure of tree #3 for the sets $L2$-plWN and $L1$-varia.

icographically intriguing tree #4 (Figure 7). We have finally decided to use the criteria TERM and PAR in a very simple decision procedure called TP. Further experiments were run on the TP tree and noun+adjective word combinations from the $L3$-NAcoll set,[8] on all structural types from the most general set $L1$, and from plWordNet ($L2$-plWN set).

In order to improve the recall of LU recognition, we added the criteria of separability and fixedness based on the IPI PAN Corpus (IPIC) counts to the criteria for tree #5 (Figure 8); we call it TP_{IPIC}.[9] The tree is a hybrid, since it binds human-driven decision paths with the semi-automatic verification of syntactic irregularities.

[8] Recall that Noun+Adj combinations are the most frequent in plWordNet: 50%.

[9] http://korpus.pl/

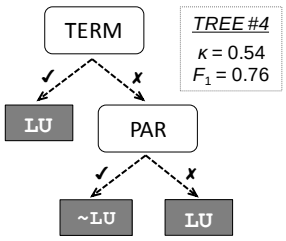

Figure 7: Tree #4 for the set $L3$-NAcoll. Legend: TERM – terminology, PAR – paraphraseability, LU – MWLU, ∼LU – loose combination.

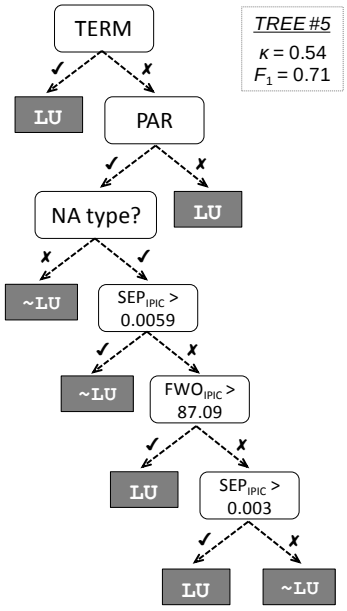

Figure 8: Tree #5 for the set $L3$-NAcoll. Legend: TERM – terminology, PAR – paraphraseability, NA type? – noun and a postposed adjective, SEP$_{IPIC}$ – separability according to corpus statistics, FWO$_{IPIC}$ – fixed word error according to corpus statistics, LU – MWLU, ∼LU – loose combination.

We checked the performance of both trees on the $L3$-NAcoll set (thoroughly) and on the $L1$-varia set. For the plWordNet set ($L2$-plWN), we only have checked κ. We looked if the trees achieved high precision and recall in recognising LUs and F-score,[10] as well as sufficient Cohen's κ, and compared them to decision procedure based only on our intuitive decision (ID). In a series of experiments with 2 to 6 linguists, we found that precision and F-score of simple ID procedure was

[10]We aim to construct a wordnet, so we focus mainly on LUs, not on word combinations rejected by linguists.

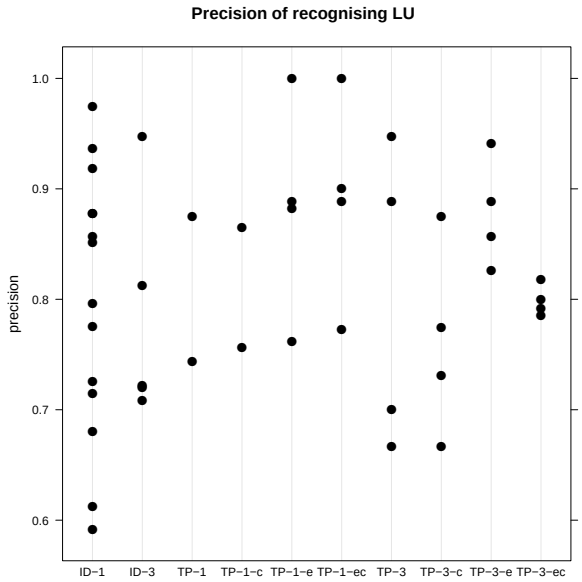

Figure 9: Precision of recognising a LU by procedures TP, TP$_{IPIC}$ and ID. Experiments for Noun+Adj word combinations on the $L3$-NAcoll set, and for all structural types on the $L1$-varia. Legend: ID – intuitive definition, TP – tree #4 procedure, 'c' – signals tree #5 procedure, digits 3 and 1 denote the $L3$ and $L1$ set, 'e' marks procedures run by experienced annotators. Precision values in experiments 'TP-1' & 'TP-3', 'TP-3-c' & 'TP-1-c', 'TP-3-e' & 'T-1-e', and 'TP-3-ec' & 'T-1-ec', are indistinguishable.

comparable to TP and TP$_{IPIC}$ (Figures 9 and 10), but the inter-annotator agreement of plWordNet editors pairs was the best for TP$_{IPIC}$ (Figure 11).

It is interesting that the κ values improved visibly when we compared experienced linguists, who have worked with the procedure for several months (scores marked with 'e') with inexperienced linguists. Please inspect in particular the similarly rising κ values in the sequence of tests 'TP1' < 'TP1c' < 'TP1e' < 'TP1ec' and 'TP3' < 'TP3c' < 'TP3e' < 'TP3ec' in Figure 11.

Good performance of our procedures on different word combination sets (NA in $L3$ set, all structural types in $L1$ and $L2$ sets), tested by many subjects (2-6, experienced and inexperienced), shows that these procedures are useful. Thus, from low-to-moderate κ, the procedures lift us to the area of acceptable κ.

7 Final Thoughts

Using procedures TP and TP$_{IPIC}$ boosts κ. Using them for a few months results in an even steeper

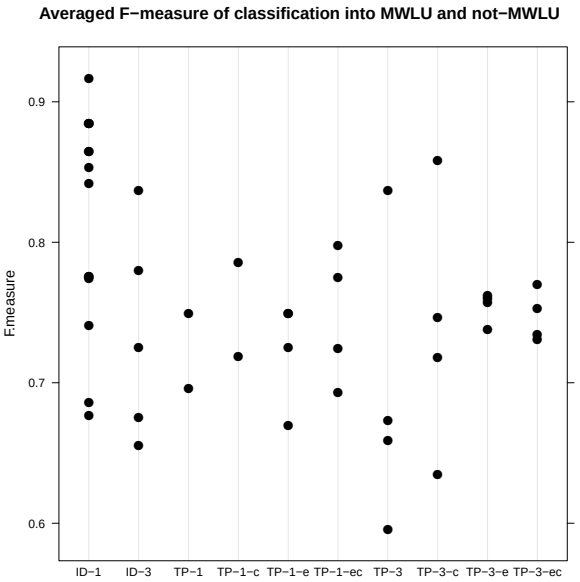

Figure 10: An averaged F-score for three procedures (TP, TP$_{IPIC}$ and ID). Experiments performed for noun+adjective word combinations on the $L3$-NAcoll set and for all structural types on the $L1$-varia. Marks as in previous legend. Be aware of small variance in the case of experienced linguists ('e').

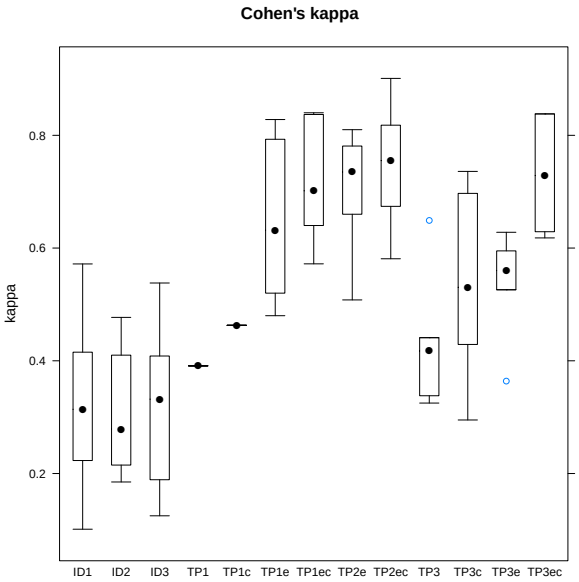

Figure 11: Boxplots of Cohen's κ for three procedures: (i) simple ID procedure performed on the sets $L3$-NAcoll ('ID3', 5 linguists), $L2$-plWN ('ID2', 4 linguists) and $L1$-varia ('ID1', 14 linguists), (ii) TP procedure ('TP1' on $L1$, 'TP3' on $L3$), (ii) TP$_{IPIC}$ procedure, *i.e.,* TP with extensions for syntactic irregularities measured on the IPI PAN Corpus ('TP3c' ran on $L3$ and 'TP1c' checked on $L1$), e – signals TP and TP$_{IPIC}$ performed by experienced plWordNet editors. The t-test performed for the $L3$ set found the means of ID procedure, the TP procedure used by experienced linguists ('TP3e') and TP$_{IPIC}$ (used both by experienced and inexperienced linguists: 'TP3c', 'TP3ec') statistically different with the p-value ≈ 0.005. For sets $L1$-varia and $L2$-plWN we can prove statistical differences between 'ID1' and 'ID2' procedures and for experienced linguists applying procedures TP and TP$_{IPIC}$ ('TP1e', 'TP2e', 'TP1ec' & 'TP2ec', respectively). Note a perfect fit of the boxplots of κ for the ID procedure ran on sets $L1$, $L2$, $L3$.

κ rise ('TPne' and 'TP$n$ec' in Figure 11). In the end, we get a workable procedural definition of a multi-word lexical unit.

Taking the perspective of a large wordnet as a comprehensive reference lexico-semantic resource, we divided MLUs into three classes:

- *terms* – multi-word terminological units,
- *idioms* – semantically non-compositional units,
- *compounds* – units which manifest syntactic irregularity.

The latter class is represented in plWordNet by noun+adjective pairs which show syntactic irregularity, *e.g.,* non-separability and fixed word order.

Using this procedure, nearly 30,000 word combinations were annotated in plWordNet.

Here is a simple conclusion from the work presented in this paper: one *can* leverage vague linguistics intuitions about multi-word lexical units into a constructive classification procedure. Another conclusion: while it is not the ultimate goal to use machine learning to build up a wordnet, machine learning can still be a lot of help.

Acknowledgments

Work supported by the Polish Ministry of Education and Science, Project CLARIN-PL, the European Innovative Economy Programme project POIG.01.01.02-14-013/09, and by the EU's 7FP under grant agreement No. 316097 [ENGINE].

References

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Gerlof Bouma. 2009. Normalized (Pointwise) Mutual Information in Collocation Extraction. In *Proceedings of the Biennial GSCL Conference*.

William Bright, editor. 1992. *International Encyclopaedia of Linguistics*, volume II. Oxford University Press, Oxford.

Thomas J. DiCiccio and Bradley Efron. 1996. Bootstrap Confidence Intervals. *Statistical Science*, 11(3):189–212.

Thomas J. DiCiccio and Joseph P. Romano. 1988. A Review of Bootstrap Confidence Intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(3):338–354.

Stanisław Dubisz. 2006. Wstęp [Introduction]. In Stanisław Dubisz, editor, *Uniwersalny słownik języka polskiego PWN. Wersja 3.0 (elektroniczna na CD) [A universal dictionary of Polish. Version 3.0 (electronic on CD)]*. Polish Scientific Publishers PWN.

Sylviane Granger and Magali Paquot. 2008. Disentangling the phraseological web. In Sylviane Granger and Fanny Meunier, editors, *Phraseology: An interdisciplinary perspective*. John Benjamins Publishing Company.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18.

J. Richard Landis and Gary G. Koch. 1971. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Andrzej Maria Lewicki. 2003. *Studia z teorii frazeologii [Research in the theory of phraseology]*. Leksem.

Kirsten Malmkjær, editor. 1991. *The Linguistics Encyclopedia*. Routledge, London & New York.

Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. 2013. The chicken-and-egg problem in wordnet design: synonymy, synsets and constitutive relations. *Language Resources and Evaluation*, 47(3):769–796.

Ronald Meester. 2008. *A Natural Introduction to Probability Theory*. Birkhäuser, 2nd edition.

Piotr Müldner-Nieckowski. 2003. *Wielki słownik frazeologiczny języka polskiego [Grand phraseological Polish dictionary]*. Świat Książki.

Piotr Müldner-Nieckowski. 2007. *Frazeologia poszerzona: Studium leksykograficzne [Extended phraseology: Lexicographic study]*. Oficyna Wydawnicza Volumen, Warszawa.

Sieb Nooteboom. 2011. Self-monitoring for speech errors in novel phrases and phrasal lexical items. In *Yearbook of Phraseology*. de Gruyter.

Alicja Nowakowska. 2005. *Świat roślin w polskiej frazeologii [The plant kingdom in Polish phraseology]*. Acta Universitatis Wratislaviensis 2755. Wydawnictwo Uniwersytetu Wrocławskiego.

Maciej Piasecki, Stanisław Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Wrocław University of Technology Press.

Kenneth Lee Pike. 1967. *Language in Relation to a Unified Theory of the Structure of Human Behaviour*. Mouton, The Hague.

Dennis Reidsma and Jean Carletta. 2008. Squibs: Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

Maria Helena Svensson. 2008. A very complex criterion of fixedness: Non-compositionality. In *Phraseology: An interdisciplinary perspective*. John Benjamins Publishing Company.

Bo Svensén. 2009. *A handbook of lexicography: the theory and practice of dictionary-making*. Cambridge University Press.

Stanisław Szober. 1967. *Gramatyka języka polskiego [A grammar of Polish]*. Polish Scientific Publishers PWN.

Ladislav Zgusta. 1971. *Manual of Lexicography*. Janua Linguarum. Series Maior. de Gruyter.