Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1842–1848
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Interannotator Agreement for Lexico-Semantic Annotation of a Corpus

Elżbieta Hajnicz
Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland
hajnicz@ipipan.waw.pl

Abstract

This paper examines the procedure for lexico-semantic annotation of the *Basic Corpus of Polish Metaphors* that is the first step for annotating metaphoric expressions occurring in it. The procedure involves correcting the morphosyntactic annotation of part of the corpus that is automatically annotated on the morphosyntactic level. The main procedure concerns annotation of adjectives, adverbs, nouns and verbs (including gerunds and participles), including abbreviations of the words that belong to the above classes. It is composed of three steps: deciding whether a particular occurrence of a word is asemantic (e.g. anaphoric or strictly grammatical), whether we are dealing with a multi-word expression, reciprocal usages of the *się* marker and pluralia tantum, which may involve annotation with two lexical units (having two different lemmas) for a single token. We propose an interannotator agreement statistics adequate for this procedure. Finally, we discuss the preliminary results of annotation of a fragment of the corpus.

Keywords: corpus, lexico-semantic annotation, interannotator agreement, Polish

1. Introduction

In this paper we want to describe the procedure for lexico-semantic annotation of the *Basic Corpus of Polish Metaphor* (BCPM), which is a first step for annotation of metaphoric expressions. This task is part of the *Cognitive and socio-cultural analysis of metaphoric expressions in Polish texts* project, aimed at, among other things, automatic detection of metaphoric expressions in Polish texts.

Annotated corpora form a basis for natural language processing. Several corpora annotated on various levels of linguistic information exist: morphosyntactic, syntactic (shallow or deep), multi-word expression (MWEs), and finally semantic, including lexico-semantic, and for various languages.

The usual procedure is to manually annotate a small corpus, use it to train NLP tools and annotate the whole corpus (a substantially larger one) by means of those tools. The quality of the manual annotation is crucial here. To ensure this happens, at least two linguists usually annotate each text sample in the corpus and the conflicts are resolved by a superannotator. Furthermore, a so-called interannotator agreement is calculated to show the difficulty of the task and the quality of its performance.

Several statistics are used to calculate the interannotator agreement. The simplest calculate the percentage of identical annotation. The most popular coefficients are Cohen's (Cohen, 1960) κ, Scott's (Scott, 1955) π and Bennett's S (Bennett et al., 1954) which take into account the possibility of chance agreement. All the statistics satisfy the equation (1), but they differ in calculating $P(E)$. S is the simplest in this respect, as it considers only the cardinality of the set of classes, whereas the two other coefficients take into account the distribution of decisions of annotators in the corpus $\hat{P}_{a_1}(k), \hat{P}_{a_1}(k)$, where a_1, a_2 are annotators.

$$\kappa, \pi, S = \frac{P(A) - P(E)}{1 - P(E)}, \quad (1)$$

$$P(A) = \frac{\sum_{t \in T} p_t^a}{\mathbf{t}}, \quad (2)$$

$$P(E) = \sum_{k \in K} \hat{P}_{a_1}(k) \cdot \hat{P}_{a_2}(k), \quad (3)$$

where $t \in T$ is a particular token and $\mathbf{k}, \mathbf{t}$ are cardinalities of the sets K, T, respectively. For S coefficient we have

$$P^S(E) = \sum_{k \in K} \frac{1}{\mathbf{k}} \cdot \frac{1}{\mathbf{k}} = \frac{1}{\mathbf{k}}, \quad (4)$$

which can be alternatively calculated as:

$$P^S(E) = \frac{\sum_{t \in T} p_t^e}{\mathbf{t}}, \quad (5)$$

where $p_t^e = \frac{1}{\mathbf{k}}$.

κ considers annotators' choices independently, i.e. $\hat{P}^{\kappa}_{a_i}(k) = \frac{\mathbf{n}^i_k}{\mathbf{t}}$, where $\mathbf{n}^i_k$ indicates how often an annotator i chooses category k. Contrarily, π averages these value, i.e. $\hat{P}^{\pi}_{a_1}(k) = \hat{P}^{\pi}_{a_2}(k) = \hat{P}^{\pi}(k) = \frac{\mathbf{n}^1_k + \mathbf{n}^2_k}{\mathbf{2t}} = \frac{\mathbf{n}_k}{\mathbf{2t}}$. When each sample is annotated by two annotators, but there are several of them involved in the whole procedure, $\hat{P}^{\kappa}_{a_1}(k)$ and $\hat{P}^{\kappa}_{a_2}(k)$ values depend on the way particular annotators are assigned as the first or the second. $\hat{P}^{\pi}_{a_i}(k)$ is not sensitive for such partitions, hence it is more suitable in such cases.

In contrast to Bennett's S, reformulating $P^{\pi}(E)$ in a way calculating it for each token separately is not straightforward. However, putting

$$p_e^t = \frac{\hat{P}^\pi(k_1^t) + \hat{P}^\pi(k_2^t)}{2} = \frac{\frac{\mathbf{n}_{k_1^t}}{2\mathbf{t}} + \frac{\mathbf{n}_{k_2^t}}{2\mathbf{t}}}{2} = \frac{\sum_{k \in K} \mathbb{1}_k(k_1^t) \cdot \mathbf{n}_{k_1^t} + \mathbb{1}_k(k_2^t) \cdot \mathbf{n}_{k_2^t}}{4\mathbf{t}}$$

we obtain

$$P^\pi(E) = \frac{\sum_{t \in T} \sum_{k \in K} \mathbb{1}_k(k_1^t) \cdot \mathbf{n}_{k_1^t} + \mathbb{1}_k(k_2^t) \cdot \mathbf{n}_{k_2^t}}{4\mathbf{t}^2} = \frac{\sum_{k \in K} \sum_{t \in T} \mathbb{1}_k(k_1^t) \cdot \mathbf{n}_{k_1^t} + \mathbb{1}_k(k_2^t) \cdot \mathbf{n}_{k_2^t}}{4\mathbf{t}^2} = \frac{\sum_{k \in K} \mathbf{n}_k^1 \cdot \mathbf{n}_k + \mathbf{n}_k^2 \cdot \mathbf{n}_k}{4\mathbf{t}^2} = \frac{\sum_{k \in K} (\mathbf{n}_k)^2}{4\mathbf{t}^2},$$

as $\sum_{t \in T} \mathbb{1}_k(k_i^t) = \mathbf{n}_k^i$ for $i = 1, 2$ annotators identifiers.

The formulae are calculated with the assumption that the set of classes K is the same for all tokens. However, this is not the case in practical applications, hence $\hat{P}_a(k)$ should be calculated w.r.t. tokens for which a category k could be chosen. Unfortunately, in the case of lexico-semantic annotations, the set of categories (LUs) is different for each lexeme. Since both distributions of lexemes and their senses are Zipfian, most LUs are chosen by an annotator once of twice. Therefore, Bennett's S seems to be the only reliable coefficient, with $p_t^e = \frac{1}{\mathbf{k}_t}$, where $\mathbf{k}_t$ is the cardinality of the set of classes K_t appropriate for a token t.

All the statistics are based on the assumption that annotators choose the value from predefined lists (potentially different for different tokens). In this paper we want to show, analysing a particular procedure applied for lexico-semantic annotation of BCPM, that in practice such a single choice may consist of a chain of interdependent decisions, and annotators can agree or disagree at every step. Thus, we have to value each such decision separately and then combine the result[1].

In what follows, we present other lexico-semantically annotated corpora (cf. section 2.). Section 3. includes the main information about the corpus being lexico-semantically annotated and about the Polish wordnet used in this annotation. The entire procedure for annotation is discussed in section 4., whereas a method of adapting standard interannotator agreement statistics to this particular procedure is proposed in section 5.. Finally, in section 6. we discuss the preliminary interannotator agreement results of two phases of annotation calculated for a small fragment of the corpus that is already annotated by two linguists.

2. Related works

The most famous semantically annotated corpus is SemCor (Miller et al., 1993), a subcorpus of the Brown Corpus (Francis and Kucera, 1964 revised and amplified 1979) containing 250 000 words semantically annotated by means of Princeton WordNet[2] (PWN) (Miller et al., 1990; Fellbaum, 1998; Miller and Fellbaum, 2007) synset identifiers. Annotation was performed by means of a dedicated interface called ConText (Leacock, 1993). The corpus was preprocessed in order to find proper names and collocations (the ones present in PWN). The collocations were joined into single units by concatenating them with underscores (e.g., *took_place*). ConText performs a corpus word by word (only open-class words). Annotators choose an appropriate sense from a list. They also have the possibility to add comments, when no available sense is considered appropriate.

A 1.7 mln. subcorpus of the British National Corpus was semantically annotated manually as a part of the Hector lexicographic project (Atkins, 1991). All occurrences of 300 word types that have between 300 and 1000 occurrences in this subcorpus were tagged, resulting in 220 000 tagged tokens.

As for Slavic languages, most words in the balanced subcorpus of the Russian National Corpus (RNC) (Grishina and Rakhilina, 2005) were semantically annotated. The semantic annotation (Apresjan et al., 2006; Lashevskaja, 2006; Kustova et al., 2007) is based on a hierarchical taxonomic classification of a Russian lexicon *Lexicograph*[3] (Filipenko et al., 1992). The texts were semantically tagged with the *Semmarkup* program (created by A. Polyakov).

For Polish, lexico-semantic annotation was performed for the sake of experiments in word sense disambiguation (WSD), and was limited to small sets of highly polysemous words, e.g. Broda and Piasecki (2011) annotated 13 nouns with the number of PlWordnet senses varying from 3 to 15 (only 72 of the total number of differentiated senses were encountered in the resulting corpus). All occurrences of 106 selected lexemes (50 nouns, 48 verbs and 8 adjectives) in NKJP 1M were annotated with coarse senses, cf. ch. 7 of (Przepiórkowski et al., 2012). PlWordnet-based[4] annotation of all open-class tokens was performed for the *Składnica* part of the NKJP 1M corpus (Hajnicz, 2014b; Hajnicz, 2014a). Unfortunately, only parsed sentences were considered.

3. Resources

BCPM is composed of two parts:

- 700 samples of the Polish Coreference Corpus (PCC), randomly selected in a way that balances various registers of texts accordingly to NKJP assumptions, cf. ch. 3 of (Przepiórkowski et al., 2012),
- 2000 samples of a fragment of NKJP 1M considered in the *Składnica* treebank, selected in way that maximises its size and the number of sentences that have parses in *Składnica*, but preserving the balance of registers.

NKJP 1M is a subcorpus of the Polish National Corpus (Polish acronym NKJP) manually annotated on the morphosyntactic level, cf. ch. 6 of (Przepiórkowski et al., 2012). The PCC (Ogrodniczuk et al., 2015), in turn, is randomly selected from the whole NKJP corpus. Therefore, BCPM as a

[1] Theoretically, we can combine these decisions into one set of classes. However, its elements will be tupples that represent chains of possible decisions.

[2] http://wordnet.princeton.edu/

[3] www.lexicograph.ru

[4] It's 2.0 version.

whole is part of NKJP. The size of the whole BCPM corpus is 344,118 tokens. It is worth noting that the PCC annotation has not been manually corrected on the morphosyntactic level.

In contrast to NKJP, we decided to lexico-semantically annotate tokens with very fine-grained semantic types represented by wordnet lexical units. We use PLWORDNET (Piasecki et al., 2009), in particular its 4.0 version (Dziob and Piasecki, 2018; Piasecki et al., 2016). It includes 288243 lexical units for 190648 lemmas, 54791 (52709) of them being multi-word[5].

PLWORDNET is a network of lexico-semantic relations, an electronic thesaurus with a structure modelled on that of the Princeton WordNet and those constructed in the EuroWordNet project. Polish WordNet describes the meaning of a lexical unit by placing it in a network representing relations such as synonymy, hypernymy, meronymy, etc.

A lexical unit (LU) is a string which has its morphosyntactic characteristics and a meaning as a whole. Therefore, it may be an idiom or even a collocation, but not a productive syntactic structure (Derwojedowa et al., 2008). An LU is represented as a pair ⟨lemma, meaning⟩, the last being a natural number. Technically, any LU also has its unique numeric identifier. Each lexical unit belongs to a synset, which is a set of synonyms.

4. The procedure for annotation

The annotation is performed independently by two linguists, and conflicts are resolved by a third. The whole procedure, together with the annotation of metaphoric expressions, is performed by means of the *WebAnno* tool (de Castilho et al., 2016) by means of a web browser. In what follows we present the lexico-semantic part of this procedure.

The lexico-semantic annotation is based on the morphosyntactic level of annotation. Since the PCC is automatically annotated on that level, annotators have to deal with erroneously annotated tokens. To make the lexico-semantic annotation more comprehensive, we ask annotators to correct errors on the basic level, part of speech (POS) and lemma, required for the lexico-semantic annotation. This correction includes spelling errors, segmentation errors (*tech nicz ny* instead of *techniczny* 'technical'), hyphened tokens (*naprawdę* 'really' for *n-a-p-r-a-w-d-ę*), resolving abbreviations (*w.* is used for *wiek* 'century', *wiersz* 'line', *wieś* 'village', *wyspa* 'island' and *wielki* 'large'), etc. There are 6 possible error codes (`lemma_error`, `pos_error`, `tag_error`, `spell_error`, `hyphen` and `case` being a subcase of `lemma_error` limited to the differences in case) and 14 POSs[6].

Next, an annotator has to decide whether a particular token should undergo annotation. Only adjectives, adverbs, nouns and verbs (including gerunds and participles) are annotated as they appear in PLWORDNET. Typically, the annotation consists in choosing the corresponding PLWORDNET lexical unit or stating that no such LU exists. Nevertheless, there are some asemantic usages of words that need not to be annotated. They include:

- grammatical usage of a word, mainly the verb BYĆ 'to be' in future and passive constructions or a correlate TO 'this',
- interrogative or anaphoric usage of a pronoun,
- the nominal element of a compound preposition, e.g. *na temat* lit. 'on the subject', 'about',
- rhetoric usage of a word,
- neologisms.

Personal pronouns are not represented in PLWORDNET, with one exception: JA 'I' meaning 'ego'. The case of interrogatives is a bit more complicated. There are strict instructions in the annotation manual how to deal with pronouns. What is important here is that they cannot be annotated fully automatically.

A somewhat different situation appears for named entities (NEs). Several are included in PLWORDNET, mainly geographical names, but most of them are not. Therefore, a detailed annotation of such tokens, besides the tag `name`, is optional.

Unfortunately, our annotation rules are still more complicated. PLWORDNET contains several multi-words expressions. The simplest are composed of a verb and the reflexive marker *się*, e.g. BAĆ SIĘ 'to fear'. Usually such annotations exclude each other, e.g. UCIEKAĆ 'to run away' and UCIEKAĆ SIĘ 'to resort'. However, PLWORDNET includes reciprocal usages of the *się* marker, e.g. ATAKOWAĆ SIĘ 'attack each other' for ATAKOWAĆ 'attack', which makes both meanings adequate.

Typical MWEs may be compositional (e.g. *dawka śmiertelna* 'deadly dose' is a 'dose') or not (e.g. there is not a meaning for CENTRUM 'centre' for *centrum handlowe* 'shopping centre', 'mall'). Therefore, we decided to allow linguists to optionally annotate elements of MWEs. This will be especially important, if the corpus is used to train WSD methods, as usually they do not consider MWEs. Technically, we assume that a corresponding LU for a MWE is assigned to its head while annotating its other elements turns to be optional.

The procedure for pluralia tantum is similar. Most of them are contemporary used only in the plural (e.g. SKRUPUŁY 'scruple') or have another meaning in plural (e.g. ZABIEGI 'efforts' vs. ZABIEG 'treatment'). However, some are distinguished solely for conventional, cultural reasons, e.g. *święta* 'holidays' (e.g. Christmas, Easter, not 'vacations') is connected with ŚWIĘTO 'holiday', and we want to preserve this connection in annotation.

Last but not least, meanings in PLWORDNET are distinguished in a very detailed, fine-grained way. Hence, sometimes it is hard to decide which sense is adequate in a particular context. We decided to demand assigning a single sense for every annotated item. Nevertheless, annotators are allowed to point out a list of senses that are very close to a chosen one and seem adequate as well. These senses are supposed to be used in the procedure of updating the annotation to the new versions of PLWORDNET, if a chosen LU is deleted.

[5]Besides 9333 LUs of 4998 verb lemmas with inherent *się* marker and 6325 LUs for 4002 corresponding gerund lemmas.

[6]NKJP distinguishes 35 POSs, cf. section 6.3 of (Przepiórkowski et al., 2012).

5. Quality of annotation and interannotator agreement

In the previous section we have shown that the procedure for lexico-semantic annotation of BCPM (especially its PCC part) is a complicated, hierarchical process. As a consequence, evaluation of its quality and calculating the interannotator agreement in particular are very important tasks.

5.1. Interannotator agreement on correction of morphosyntactic annotation

The first annotator's decision – stating whether the lemma and POS of a particular token is correct and correcting them if needed – influences the whole annotation procedure. Lexico-semantic annotations of two different ⟨lemma, POS⟩ pairs cannot be compared. On the other hand, the correction itself is not a genuine part of the lexico-semantic annotation *per se*. Therefore, we decided to evaluate this step separately. There are four possibilities:

1. Only one annotator makes corrections,
2. Both annotators make them, but their corrections differ,
3. Both annotators agree on their corrections,
4. Both annotators accept the original annotation.

We decided to calculate two various statistics: taking into account all tokens (i.e. including the case 4.) or considering only potentially improper tokens, i.e. the ones for which at least one annotator intervene. This means considering a set $T_C \subseteq T$ of tokens with morphosyntactic annotation changed.

$$p_t^a = \begin{cases} 1 & \text{if annotators fully agree,} \\ 0.8 & \text{if annotators agree on lemma and POS,} \\ 0.6 & \text{if annotators agree on error code and lemma or POS,} \\ 0.4 & \text{if annotators agree on lemma or POS,} \\ 0.1 & \text{if annotators agree on error code only,} \\ 0 & \text{if annotator fully disagree.} \end{cases} \quad (6)$$

In (6) values of p_t^a w.r.t. particular annotators' decisions are proposed, which enables us to calculate $P(A)$ in (1). $P(E)$ is calculated accordingly to Scott's π assumptions.

5.2. Interannotator agreement on lexico-semantic annotation

The lexico-semantic annotation itself consists of three steps:

S1. deciding, whether a token should undergo annotation,

S2. deciding, whether it is a case of *pluralia tantum*, the reciprocal *się* marker or MWE,

S3. performing actual annotation, potentially on two levels, including a decision, whether there are senses "close" to the chosen one adequate in the context.

There are 6 tags (`grammatical`, `anaphora`, `interrogative`, `prep_element`, `brev:phrase` and `rhetoric`) used for an asemantic occurrence of a word[7].

For simplicity, elements of MWEs or NEs that are not annotated are treated as asemantic as well. However, they are used for different POSs and cases, in a completely different context, so this is very unlikely to confuse them. Therefore, we decided to treat this decision as a binary one.

The next decision concerns how "additional" annotation (pluralia, reciprocal *się*, MWEs) should be treated. The simplest way is to treat them as independent annotations, a sort of "added tokens". However, this is not the case. Annotating MWEs and pluralia tantum is more important than "basic" single-token annotation as a more precise one (hence its weight is 0.6), whereas the role of LUs with reciprocal *się*is auxiliary (hence its weight is 0.4). Instead, we have made a simplifying assumption that a particular token can be annotated only in two ways. It is a bit controversial: a MWE or a NE can include a plurale tantum or a reciprocal verb phrase. The above weight will be referred to in the overall formulae as ω_b, ω_a for "basic" and "additional" annotation, respectively. To simplify the evaluation procedure, we decided that performing "additional" annotation only is equivalent to performing two levels of annotation with the `lack` value assigned for "basic" annotation.

Nevertheless, asemantic annotations, on the one hand, and "additional" annotation, on the other hand, are proportionally rare: the most typical case is a "basic" annotation. Since this is a case of a close set of classes, the same for all tokens, we can calculate interannotator agreement for these two phases in a spirit of Scott's π. We will refer to it in the overall formula as β_l, l = 2+0, 2+1, 2+2, 1+0, 1+1, depending on whether annotators choose 0 (asemantic), one (only basic) or two levels of annotation of a token. The chance of agreeing or not on the type of "additional" annotation will be denoted as α_0, α_1.

Furthermore, even if the annotators agree on the number and types of annotation, they can choose a different LU (including a `lack` decision).

The agreement of "close" values is not calculated. Instead, they are used to evaluate the degree of disagreement of annotation of a particular token, i.e.

C1. choices of both annotators are the same (100% of agreement despite "close" values);

C2. choices of both annotators are included in the partners "close" lists (60% of agreement);

C3. a choice of one annotator is included in the partner "close" list (20% of agreement);

C4. neither choice is included in the partner "close" list (0% of agreement).

This weight will be referred to in the overall formulae as γ. The combinations of these decisions results in values of the annotator's agreement p_t^a and expected agreement p_t^e for a token t gathered in table 1.

6. Preliminary results

The annotation of the BCPM corpus is an ongoing task. Till now, only 162 samples of the PCC part of the corpus composed of 46,350 tokens was annotated twice, which enables us to calculate the interannotator agreement[8]. The fre-

[7] A `neologism` is actually a kind of `lack` tag, indicating that a word is not supposed to be considered in PLWORDNET.

[8] 12 annotators are involved in the procedure.

Table 1: The level of agreement of two annotators depending of types of annotation

Annotators' choice	p_t^a	p_t^e
One annotator assigns two LUs and the second none	0	β_{2+0}
One annotator assigns two LUs and the second one	$\omega_b \cdot \gamma_b$	$\beta_{2+1} \cdot \frac{1}{\mathbf{k}_t^b}$
Both annotators assign two LUs with a different type of the "additional" annotation	$\omega_b \cdot \gamma_b$	$\beta_{2+2} \cdot \frac{1}{\mathbf{k}_t^b} \cdot \alpha_0$
Both annotators assign two LUs with the same type of the "additional" annotation	$\omega_b \cdot \gamma_b + \omega_a \cdot \gamma_a$	$\beta_{2+2} \cdot \frac{1}{\mathbf{k}_t^b} \cdot \frac{1}{\mathbf{k}_t^a} \cdot \alpha_1$
One annotator assigns one LU and the second none	0	β_{1+0}
Both annotators assign one LU	$\omega_b \cdot \gamma_b$	$\beta_{1+1} \cdot \frac{1}{\mathbf{k}_t^a}$
Both annotators assign none LU of a different type	0	β_{0+0}
Both annotators assign none LU of the same type	1	β_{0+0}

Table 2: The number of tokens annotated on the particular level of annotation

Type of tokens	number	percent	
considered	46,350	13,47%	
annotated	27,191	58.66%	
corrected	1,549	3.34%	
corrected twice	234	0.50%	(15%)
semantic	25,778	55.62%	
strange	41	0.09%	

quency of tokens annotated in particular ways are presented in table 2. The percentage of considered tokens is calculated w.r.t. the size of the whole corpus, the other are calculated w.r.t. the frequency of considered tokens.

6.1. Correction of morphosyntactic annotation

Correction of errors of the automatic morphosyntactic annotation turned to be marginal, it concerns 3.34% of tokens. What is much more surprising, only 15% of them is corrected by two annotators. The reason is that some annotators ignored this step and focused on choosing senses accordingly to the text level. On the other hand, some other were too thorough, correcting conjunctions, prepositions etc. not supposed to be annotated. Because of that, we decided to calculate agreement for three sets of tokens: all annotated tokens, tokens corrected by at least one annotator and tokens corrected by two annotators. In all cases, we present basic agreement $P(A)$, expected agreement $P(E)$ and resulting agreement φ. The numbers of particular types of corrections used to calculate $P(E)$ are shown in table 3.
The results presented in table 4 show that the most influential is decision whether to correct morphosyntactic annotation of a token or not. Nevertheless, its impact on the whole procedure is weak and can be ignored.

6.2. Actual semantic annotation

Semantic annotation was performed for more than 55% of tokens. The numbers do not include tokens for which the corrected lemma or POS differ.
The frequency of levels of annotation used to calculate $P(E)$ as in table 1 is presented in table 5. This is not surprising that more than 80% of semantically annotated tokens have single, "basic" interpretation given by both annotators. Asemantic usages range between 6% and 15%, but annotators do not agree on that. Two level annotation is marginal. As one may expect, if both annotators decide to assign an "additional" annotation, they agree on its type, only 4 (0.98%) of 189 such annotations are inconsistent w.r.t. the type.

Table 3: The number of particular types of corrected morphosyntactic errors

error type	number	percent
case	107	3.45%
pos_error	228	7.36%
lemma_error	468	15.11%
tag_error	537	17.33%
spell_error	20	0.65%
hyphen	1	0.00%
no_annot	422	13.62%
none	1315	42.45%

Table 4: The interannotator agreement for the correction of morphosyntax

Type of tokens	$P(A)$	$P(E)$	φ
annotated	0.969	0.223	0.961
corrected	0.463	0.258	0.277
corrected twice	0.786	0.340	0.676

Table 5: The frequency of particular numbers of annotation levels

β_{0+0}	1769	0.0686	β_{2+0}	31	0.0012
β_{1+0}	2413	0.0936	β_{2+1}	548	0.0213
β_{1+1}	20828	0.8080	β_{2+2}	189	0.0073

The results, presented in table 6, are poor. The reason is that the task is complicated and some annotators seem to misunderstand the instructions. The best results are obtained for the most frequent class, namely nouns. Nevertheless, the

Table 6: The interannotator agreement for the actual semantic annotation

Type of tokens	freq.	$P(A)$	$P(E)$	φ
all semantic	25778	0.551	0.211	0.456
nouns	12171	0.637	0.201	0.546
verbs	6272	0.519	0.209	0.391
adjectives	4507	0.527	0.194	0.413
adverbs	1606	0.601	0.281	0.445

differences between classes are not important.

7. Conclusions

In this paper, we have examined the procedure for lexico-semantic annotation of a particular corpus, namely the *Basic Corpus of Polish Metaphors*, by means of a particular repository of senses, namely PLWORDNET. Furthermore, we have shown that it is hard to apply any of the standard interannotator agreement statistics directly and propose a method for adapting them to this very procedure.

Most authors declare using Cohen's κ for this. However, they usually do not analyse why they chose this particular statistics and whether it is appropriate for their decision. According to Artstein and Poesio (2008), there are several terminological inconsistencies concerning interannotator agreement statistics in the literature.

The tools used for corpora annotation offer to calculate some interannotator agreement statistics. In particular, the *WebAnno* we are using, makes it possible to calculate Kohen's Kappa, Fleiss' Kappa and Krippendorff's Alpha. However, in order to do it properly, a tool needs to know all interdependencies among categories of tags. This is not the case in *WebAnno*. de Castilho et al. (2016) report possibility of constrains concerning applicability of one category (or its set of values) w.r.t. another category only for categories with close set of values (selected from a list). What is more, inserting a sense number "manually" or choosing it from the list is merely a technical difference, but it can influence the calculation, the first being a choice from the opened set of values, the second being a choice from a close set of values. The first can be interpreted as binary classification (agree/disagree) or limiting the set of values to the introduced ones (close word assumption). None of these interpretations is correct, as such a set of numbers that represent senses are understood as uniform for all tokens.

To sum up, the most important conclusion is that the choice of an interannotator agreement statistics that is appropriate for a particular annotation task is not obvious and it needs reasonable consideration every time.

As for the results of annotation, large number of spelling errors etc. shows that traditional linguists do not understood and disregard computational requirements of the annotation procedure. Further analysis reveals that some annotators have misunderstood the instructions. We are aware that the procedure for annotation is complicated and some its cases turned to be controversial. Nevertheless, this emphasises the sense of evaluating the results of annotation in such a preliminary stage. It is a good moment to establish weak point of the procedure (the great role of superannotators), change the instructions and train the team of annotators.

The improved version of the guidelines of the annotation procedure (in Polish) is available at `http://zil.ipipan.waw.pl/CORMETAN?action=AttachFile&do=view&target=instrukcja_sem-web.pdf`.

8. Bibliographical References

Apresjan, J., Boguslavsky, I., Iomdin, L., Iomdin, B., Sannikov, A., and Sizov, V. (2006). A syntactically and semantically tagged corpus of Russian: State of the art and prospects. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, pages 1378–1381, Genoa, Italy.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Atkins, S. (1991). Tools for computer-aided corpus lexicography: The Hector project. *Acta Linguistica Hungarica*, 41:5–72.

Bennett, E. M., Alpert, R., and Goldstein, A. C. (1954). Communications through limited-response questioning. *Public Opinion Quarterly*, 18(3):303–308.

Broda, B. and Piasecki, M. (2011). Evaluating lexicographer controlled semi-automatic word sense disambiguation method in a large scale experiment. *Control and Cybernetics*, 40(2):419–436.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

de Castilho, R. E., Éva Mújdricza-Maydt, Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., and Biemann, C. (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanitites (LT4DH)*, pages 76–84, Osaka, Japan.

Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawisławska, M., and Broda, B. (2008). Words, concepts and relations in the construction of Polish WordNet. In Attila Tanacs, et al., editors, *Proceedings of the Global WordNet Conference*, pages 162–177, Seged, Hungary.

Dziob, A. and Piasecki, M. (2018). Implementation of the verb model in plWordNet 4.0. In Francis Bond, et al., editors, *Proceedings of the 9th International WordNet Conference (GWC 2018)*, pages 114–123, , Singapore. Global Wordnet Association.

Christiane Fellbaum, editor. (1998). *WordNet — An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Filipenko, M., Paducheva, E., and Rakhilina, E. (1992). Semantic dictionary viewed as a lexical database. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-1992)*, pages 1295–1299, Nantes, France.

Francis, W. N. and Kucera, H. (1964, revised and amplified 1979). Brown corpus manual. Internet.

Grishina, E. and Rakhilina, E. (2005). Russian National

Corpus (RNC): an overview and perspectives. In *Proceedings of the AATSEEL 2005*.

Hajnicz, E. (2014a). Lexico-semantic annotation of *składnica* treebank by means of PLWN lexical units. In Heili Orav, et al., editors, *Proceedings of the 7th International WordNet Conference (GWC 2014)*, pages 23–31, Tartu, Estonia. University of Tartu.

Hajnicz, E. (2014b). Procedure of the lexico-semantic annotation of *składnica* treebank. In Nicoletta Calzolari, et al., editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, pages 2290–2297, Reykjavík, Iceland. ELRA.

Kustova, G., Lashevskaja, O., Rakhilina, E., and Paducheva, E. (2007). On taxonomy in cognitive semantics and corpus linguistics: Parts of body. In *Proceedings of the 10th International Cognitive Conference*, Cracow, Poland.

Lashevskaja, O. (2006). Corpus-aided construction grammar: Semantic tools in the Russian National Corpus. In *Proceedings of the 2th International Meeting of the German Cognitive Linguistic Association*, Munich, Germany.

Leacock, C. (1993). Context: A toot for semantic tagging of text: User's guide. Technical Report 54, Cognitive Science Laboratory, Princeton University, Princeton, NJ.

Miller, G. A. and Fellbaum, C. (2007). WordNet then and now. *Language Resources and Evaluation*, 41:209–214.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An online lexical database. *International Journal of Lexicography*, 3(4):235–244.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. (1993). A semantic concordance. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 303–308, Plainsboro, NJ.

Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A., and Zawisławska, M. (2015). *Coreference in Polish: Annotation, Resolution and Evaluation*. Walter De Gruyter.

Piasecki, M., Szpakowicz, S., and Broda, B. (2009). *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, Poland.

Piasecki, M., Szpakowicz, S., Maziarz, M., and Rudnicka, E. (2016). PlWordNet 3.0 – almost there. In Verginica Barbu Mititelu, et al., editors, *Proceedings of the 8th International WordNet Conference (GWC 2016)*, pages 290–299, Bucharest, Romania. Global Wordnet Association.

Adam Przepiórkowski, et al., editors. (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw, Poland.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325.

9. Language Resource References

NKJP Consortium. (2012). *Manually annotated subcorpus of National Corpus of Polish (NKJP 1M)*. NKJP Consortium, http://clip.ipipan.waw.pl/NationalCorpusOfPolish, 1.0.

G4.19 Group at Department of Artificial Intelligence, Wroclaw University of Technology. (2015). *Polish wordnet plWordNet*. Department of Artificial Intelligence, Wroclaw University of Technology, http://plwordnet4.clarin-pl.eu/, 2.1.

Maciej Ogrodniczuk. (2019). *Polish Coreference Corpus*. Institute of Computer Science, Polish Academy of Sciences, http://zil.ipipan.waw.pl/PolishCoreferenceCorpus, 1.5.

Acknowledgements

These research was financed by the Polish National Science Centre, within the project **2014/15/B/ST6/05186** *Compositional distributional semantic models for identification, discrimination and disambiguation of senses in Polish texts* and the project **2018/29/B/HS1/01773** *Cognitive and sociocultural analysis of metaphorical expressions in Polish texts* (CORMETAN).