Lemmatization of Multi-word Common Noun Phrases and Named Entities in Polish

Michał Marcińczuk
G4.19 Research Group
Department of Computational Intelligence
Faculty of Computer Science and Management
Wrocław University of Technology, Wrocław, Poland
michal.marcinczuk@pwr.edu.pl

Abstract

In the paper we present a tool for lemmatization of multi-word common noun phrases and named entities for Polish called PoLem[1]. The tool is based on a set of manually crafted rules and heuristics utilizing a set of dictionaries (including morphological, named entities and inflection patterns). The accuracy of lemmatization obtained by the tool reached 97.99% on a dataset with multi-word common noun phrases and 86.17% for case-sensitive evaluation on a dataset with named entities.

1 Introduction

In the article we cover the problem of multi-word common noun phrase and named entity lemmatization for Polish. The task relies on generating a nominative form of an expression[2]. For example the following named entities — *Janem Nowakiem* (a person name in an instrumental case) and *Jana Nowaka* (a genitive case of the same person name) — should be lemmatized to *Jan Nowak*. Both, lemmatization of multi-word common noun phrases and named entities are challenging because Polish is a highly inflectional language and a single expression can have several inflected forms.

The complexity of multi-word common noun phrase lemmatization is caused by the fact that the expected lemma is not a simple concatenation of base forms for each word in the phrase. In most cases only the head of the phrase is changed to a nominative form and the remaining tokens, which are the modifiers of the head, should remain in a specific case. For example in the phrase *piwnicy domu* (Eng. *house basement*) only the first word should be changed to their nominative form while the second word should remain in the genitive form, i.e. *piwnica domu*. A simple concatenation of tokens' base forms would produce a phrase *piwnica dom* which is not correct.

In the case of named entities the task is much more complex due to the following reasons:

1. Named entities consist of many words which are not present in the morphological dictionaries. For such words it is impossible to generate the desired form using only a morphological dictionary. A more generic method is required.
2. Some foreign proper names subject to inflection and some not.
3. The desired lemma of a named entity depends on the named entity category. For example *Słowackiego* (a person last name in genitive or accusative) should be lemmatized to *Słowacki* in case of person name and to *Słowackiego* in case of street name.
4. Capitalization do matter. For example a country name *Polska* (Eng. *Poland*) should be lemmatized to *Polska* but not to *polska*.

We took the following assumptions:

Assumption 1 *The lemma should have the same number and gender as the input expression.*

Assumption 2 *The multi-word common noun phrase is neither a proper name nor contains a proper name.*

Assumption 3 *The named entity can consist of any number of tokens, i.e. one or more.*

We also require that:

[1] http://nlp.pwr.wroc.pl/polem

[2] By *expression* we understand a multi-word common noun phrase or a named entity.

Proceedings of Recent Advances in Natural Language Processing, pages 483–491,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_064

Requirement 1 *The expression is described with a disambiguated morphological information.*

Requirement 2 *In case of named entities the semantic category of the named entity is known.*

Our paper is divided into three main parts. In the first part we present the related literature overview, evaluation datasets and baselines. In the second part we present the development of a set of rules for multi-word common noun phrase lemmatization. In the last part we extend the set of rules with some heuristic and new rules in order to increase the coverage and accuracy of named entity lemmatization.

2 Related Works

According to our best knowledge there are several researches on phrase lemmatization for Polish which include both multi-word phrases and named entities. One of them is a rule-base approach in which the lemmatization rules were combined with a grammar for recognition of noun phrases (Degórski, 2012). The lemmatization rules were manually created as an extension to a grammar for Polish (Głowińska, 2008) — the lemmas are generated for the recognized phrases. This method does not produce the final lemmas as it requires a form generator to obtain the desired forms for the generated morphological tags and base forms. The method was evaluated on a set of 336 phrases. 158 of them were correctly recognized and the accuracy for them was 82.9%.

Another approach was presented by Radziszewski (2013b). The method was based on an automatic generation of lemmatization rules using Conditional Random Fields for noun phrases. The authors obtained the accuracy of 80.7% on a set of 564 noun phrases (containing single- and multi-word phrases).

The last approach was also based on an automatic generation of lemmatization rules from a corpus (Małyszko et al., 2015). The method obtained the accuracy of 82.1% on a set of 888 phrases (only 83% of 1063 tested phrases were marked as processable).

3 Evaluation Datasets

The dataset of multi-word common noun phrases was created by extracting occurrences of keywords from the KPWr corpus (Broda et al., 2012)[3]. The corpus contains 1628 documents annotated with keywords. The documents are tagged with the WCRFT tagger (Radziszewski, 2013a). We have extracted 3965 occurrences of keywords from the documents' content. Then we selected those phrases which conform the Assumptions 2 — 1728 in total. Then the set was divided into two random subsets — a train set with 1329 instances and a test set with 399 instances. The train set was used to develop a set of lemmatization rules and the test set was used for the final evaluation. As the keywords were extracted from documents tagged with a morphological tagger the dataset conforms the Requirement 1.

The dataset of named entities was created by extracting named entities from the same corpus. The corpus contains 1349 documents annotated with lemmatized named entities. We have extracted 21 449 occurrences of named entities from the documents' content. The set was also divided into two random subsets — a train set with 14 104 named entities and a test set with 7 345 named entities. The train set was used to extend the basic set of lemmatization rules and the test set was used for the final evaluation.

4 Baseline Results

To establish the baseline we measured the accuracy for three lemmatization methods. For multi-words common noun phrases we used only case-insensitive evaluation. For named entities we used both case-sensitive (CS) and case-insensitive (CI) evaluations. The results are presented in Table 1. The baseline methods are:

1. Concatenation of text forms — the text form is a form which appears in the document content.

2. Concatenation of base forms — base forms were assigned by the morphological tagger for each token.

3. Lemmatization grammar (Degórski, 2012) for the Spejd tool (Przepiórkowski, 2008).

Figure 1 presents a sample phrases with all the mentioned token attributes.

[3] https://clarin-pl.eu/dspace/handle/11321/270

Lemmatization method	Multi-word phrases		Named entities	
	Train	Test	Train	Test
Number of named entities	*1329*	*388*	*14 104*	*7 345*
Concatenation of text forms CI	41.82%	35.84%	56.62%	57.17%
Concatenation of text forms CS	-	-	56.06%	56.68%
Concatenation of base forms CI	23.80%	22.31%	75.46%	73.08%
Concatenation of base forms CS	-	-	44.02%	43.97%
Spejd (only recognized phrases) CI	79.31%	80.49%	83.42%	82.08%
Spejd (only recognized phrases) CS	-	-	42.83%	42.55%
Spejd (with text forms) CI	69.45%	67.42%	67.37%	67.36%
Spejd (with text forms) CS	-	-	44.01%	44.04%

Table 1: Baseline accuracy of lemmatization for different methods on the train and test sets.

Token	**1 (head)**	**2**	**3**
Text form	organu	pierwszej	instancji
Base form	organ	pierwszy	instancja
Morphological tag	subst:sg:gen:m3	adj:sg:gen:f:pos	subst:sg:gen:f
Expected lemma	organ	pierwszej	instancji
Translation	[3] authority	[1] first	[2] instance

Figure 1: A sample expression with its text form, base form, morphological tags and the expected lemma.

4.1 Baseline for Multi-word Common Noun Phrases

Using the heuristic-based approach we were able to generate lemmas for every phrase in the dataset. However, as expected, the accuracy was very low — 41.82% for the text forms and 23.80% for the base forms on the train set and 35.84% and 22.31% for the test set.

Using the Spejd lemmatization grammar (Degórski, 2012) we were not able to obtain lemmas for every phrase in the dataset. For the train set we were able to generate lemmas for 952 phrases out of 1329 (72% coverage) and for the test set for 287 out of 399 (also 72% coverage). The reason is that the lemmas are generated for specific phrases recognized by the grammar and in some cases our phrases do not overlap with the phrases matched by the grammar. For the recognized phrases the accuracy was 79.31% and 80.49% for the train and the test sets respectively. To overcome the problem of missing lemmas, for the phrases for which Spejd did not generate any lemma we took the text form as a lemma. The final accuracy for the complete dataset was 69.45% and 67.42% for the train and the test sets respectively.

4.2 Baseline for Named Entities

Using text forms as lemmas we obtained accuracy above 56%. The case-sensitive evaluation for the text forms drops the accuracy by less than 1 pp. This indicates that almost all named entities appears in the text in their expected casing, i.e. camel case, all upper, all lower, mix case, etc. In turn, using concatenation of base forms we obtained accuracy near 75%. This means that for 25% of named entities some of the tokens requires a transformation other than the change to their singular masculine nominative form. The case-sensitive evaluation for base form concatenation drops the accuracy to 44% what shows than the token lemmas do not hold the expected capitalization. Even when we apply the casing as for the text forms to the concatenation of base forms we will not increase the accuracy above 75% — by the analogy to the difference between case-insensitive and case-sensitive evaluation for the concatenation of text forms. This shows that in order to handle the remaining 25% of named entities we need a more sophisticated method of lemmatization.

The Spejd lemmatization grammar (Degórski, 2012) obtained the accuracy of 83% but only for near 56% of all named entities. The reason of the low coverage is the same as for multi-word

phrases. To handle such cases we combined the Spejd lemmatization grammar with text forms for the named entities which do not match the recognized phrases. The accuracy for case-insensitive evaluation dropped to 67%. For case-sensitive evaluation the accuracy drops even more to 44%.

5 Lemmatization of Multi-word Phrases

5.1 Rule Development

Lemmatization of multi-word common noun phrases mostly relies on finding the correct combination of forms for each word in the phrase. Most of the words and their inflected forms are present in the morphological dictionary. The difficulty is to find the correct form for each token based on the phrase structure. To achieve this goal we developed a set of lemmatization rules.

Each rule consists of two elements — a set of constraints and a set of transformations. When the constrains are satisfied for a given phrase then the transformations are used to obtain the expected word form for each token in the phrase. Sample rules are presented on Figure 2. The constraints test tokens' morphological attributes to check whether they have specific values (text form, base form, case, gender and/or number). The constraints can also check if there is an agreement between specific words in the phrase. To encode the constraints we used the WCCL formalism (Radziszewski et al., 2011). In order to reduce the number of required rules we identified such patterns, where only the first words need some kind of transformation while the remaining words are unchanged. The rules are called the *tail rules*. A sample *tail rule* is presented on Figure 2b.

The transformations (`transformations` tag) are used to generate specific forms of the words matched in the phrase. The transformation can change case or gender of the word. The attribute `index` identifies the word in the phrase and the remaining attributes indicate the expected value of the morphological attributes. For instance, `cas="nom"` means that the word should be in nominative. If the value of a morphological attribute is unchanged then the attribute is omitted.

The initial set of rules did not cover all phrases from the train set because some of them were tagged incorrectly. To overcome the tagger errors we have added some rules with relaxed constraints (for instance we ignored the words agreement). The rules with relaxed constraints are distinguished from the remaining set of rules by the *Fix* suffix in their name. The final set of rules consists of 27 rules.

```
<rule name="SubstAdj_Agr">
   <wccl match="complete">
      and(
         inter(class[1],{adj}),
         inter(class[0],{subst,ger,depr}),
         agrpp(0,1,{nmb,gnd,cas})
      )
   </wccl>
   <transformations>
      <set index="0" cas="nom"/>
      <set index="1" cas="nom"/>
   </transformations>
</rule>
```

(a) A sample standard lemmatization rule.

```
<rule name="AdjSubstTail">
   <wccl match="prefix">
      and(
        inter(class[0],{adj,ppas,pact}),
        inter(class[1],{subst,ger,depr}),
        agr(0,1,{nmb,gnd,cas})
      )
   </wccl>
   <transformations>
      <set index="0" cas="nom"/>
      <set index="1" cas="nom"/>
   </transformations>
</rule>
```

(b) A sample *tail* lemmatization rule.

```
<rule name="SubstAdvHyphenAdj_FixGndM1">
   <wccl match="complete">
      and(
         inter(class[0],{subst,ger,depr}),
         inter(gnd[0],{m1}),
         inter(class[1],{adv}),
         regex(orth[2],"-"),
         inter(class[3],{adj,ppas,pact}),
      )
   </wccl>
   <transformations>
      <set index="0" cas="nom"/>
      <set index="3" cas="nom" gnd="m1"/>
   </transformations>
</rule>
```

(c) A sample *fix* lemmatization rule.

Figure 2: Sample lemmatization rules.

The lemmatization rules are executed in a specific order and the first rule for which the constraints are satisfied for given phrase is used to generate the lemma. At first the set of standard rules is executed. If none of the rules is matched, then the set of *fix rules* is used and, at the end, the set of *tail rules* is used. If the constraints are satisfied, then the transformations for the rule are applied. If a rule does not contain any transformation for a word then the unmodified text form is taken. In other case, the input base form and the morphological tag are taken and the transformation is applied, i.e. the specified attribute values are substituted. Then the modified values are used to generate a new word form using a morphologi-

cal analyzer called Morfeusz (Woliński, 2006).

5.2 Evaluation

Table 2 contains the results of evaluation on both sets presented in Section 3. The set of 27 rules was enough to cover all multi-word phrases in the train set and it obtained the accuracy of 99.10%. Also a high accuracy of 97.99% was obtained on the test set which was not used during rule development. In both cases the accuracy was higher than any baseline method presented in Section 4.

Evaluation	Train	Test
Multi-word phrases		
PoLem' CI	99.10%	97.99%
Named Entities		
PoLem' CI	85.56%	84.64%
PoLem' CS	81.96%	80.66%

Table 2: Accuracy of the initial lemmatization rules.

We analyzed the incorrectly generated lemmas for the train set in order to find the sources of errors. We found out that there is no simple solution to handle those cases without any additional resources. We identified the following types of problems:

Tagger errors:

- incorrect number — one of the tokens has incorrect number (singular or plural). Some of the *fix rules* force the correct number for the first or the second token. However, in some cases this leads to an error, because the rule changes the number for the correctly disambiguated word. For such cases the rule should determine for which the disambiguation was incorrect. This is possible for those phrases for which one of the tokens has only plural for singular interpretations. For example for phrase *pytania prawnego* (Eng. *legal question*) the tagger assigned the following interpretations: (1) *subst:pl:nom:n* (2) *adj:sg:gen:m3:pos*, while for the first token it should be *subst:sg:gen:n*. The second token can have only singular interpretation what is a sufficient indicator that the whole phrase should be singular, not plural.

- incorrect part of speech — one of the tokens has incorrect part of speech. For example for phrase *zmienne środowiskowe* (End. *environmental variables*) the tagger assigned the following part of speeches: *adj adj* for the subsequent words. The first token should be recognized as a noun instead of an adjective. In this case we also should check other possible interpretations to find out the possibly correct tags.

Sense disambiguation — polysemous words might have different schemes of inflection. For example word *pasza* means: (1) fodder or (2) pasha. The word has different plural form for both meanings: *pasz* for (1) and *paszowie* for (2). To handle this problem it might be necessary to check the collocations for those forms. For example, for phrase *pasze lecznicze* (Eng. *healing fodders*) word *healing* will more likely co-occur with *fodder* than *pasha*.

More than one possible form — there are words which have more than one possible form. For example word *koszt* (Eng. *cost*) has two possible plural forms: *koszty* and *koszta*. The first form is more common than the other one. For such cases a frequency list might be helpful to determine the more frequent form.

The initial set of rules also obtained a high accuracy on the dataset of named entities — 85.65% for the train set and 84.64% for the test set. However, for case-sensitive evaluation (marked as CS in the table) the accuracy dropped to 81.96% and 80.66% respectively. The results are also higher than for any presented baseline method. At this stage the dataset of named entities was not used in the development of lemmatization rules. In the next section we present the extension of the initial set of rules based on the analysis of the train set of the dataset of named entities.

6 Lemmatization of Named Entities

In this section we present several extensions of the initial set of lemmatization rules developed for multi-word common noun phrases which improved the accuracy of named entity lemmatization. The following subsections describe in details each of them.

6.1 Generic Lexicons

We used two large lexicons of proper names which are applied before lemmatization rules. The first

one was extracted from a morphological dictionary called Morfeusz SGJP (Woliński, 2006). We have selected 39 084 entries marked as geographical names. As the morphological dictionary contains only single words, the lexicon is used to lemmatize single-word named entities.

The second lexicon contains a list of inflected named entities extracted from Polish Wikipedia. The list is a part of NELexicon2[4]. The list of inflected forms was created by extracting internal links from Polish Wikipedia. Each pair consists of a link text and a title of Wikipedia page to which the link directs. The list was filtered by selecting those pairs which have the same number of elements (in the link text and the page title) and the consecutive words have the same base form or have the same prefix of a certain length. The list of pairs was filtered with a list of known proper names. The list consists of 110 178 pairs (single- and multi-word proper names of various categories).

6.2 Category-based Lexicons

For person names we created a separate lexicon which contains solely inflected forms of person names with their lemmas. The lexicon consists of names from NELexicon2 marked as a person name and the lists of first names and last names from Morfeusz SGJP. The lemmatization procedure using this lexicon is based on a rule that for a person name containing only first and last names each part of the name is changed to their respective nominative form. For each person name we divide the name into single words. Then for each word we lookup its' base form in the lexicon. If for every word we can determine the base form then the final lemma is a concatenation of the found base forms. We also defined a list of words which are never inflected, i.e. *św.* (Eng. *Saint*), *von* (and other similar words which appear in foreign last names). If at least one of the words cannot be lemmatized this way we do not generate any lemma.

6.3 Inflection Rules

The dictionaries of person names and geographical names misses many inflected forms of the names. To increase the coverage we have generated a frequency list of suffixes changes based on the morphological dictionary Morfeusz SGJP. Figure 3 presents the most frequent inflection rules for person names. The list consists of lines in the following form: *subst:sg:gen:m1 iego i 1100 0.98*. This means that 98% of names tagged as *subst:sg:gen:m1* which are ended with *iego* have a base form ended with *i*. For instance, the name *Grzybowskiego* tagged as *subst:sg:gen:m1* should be lemmatized to *Grzybowski*. We have created two inflection rule lists, separately for person names and geographical names. In the first run we try to find a set of inflection rules for every single word in the name which leads to a form that is present in the NELexicon2. If we fail to find such a set then we find a set of inflection rules with the highest confidence. On this step we ignore inflections which are less frequent than 50 occurrences. The generated form is treated as a possible lemma.

```
subst:sg:loc:m1   im    i   1112 0.99
subst:sg:inst:m1  im    i   1112 0.99
subst:pl:loc:m1   ich   i   1112 0.99
subst:pl:inst:m1  imi   i   1112 0.99
subst:pl:gen:m1   ich   i   1112 0.99
subst:pl:dat:m1   im    i   1112 0.99
subst:pl:acc:m1   ich   i   1112 0.99
subst:sg:gen:m1   iego  i   1100 0.98
subst:sg:dat:m1   iemu  i   1100 0.98
subst:sg:acc:m1   iego  i   1100 0.98
depr:pl:voc:m2    ie    i   1100 0.93
depr:pl:nom:m2    ie    i   1100 0.93
subst:sg:inst:f   ą     a   1099 1.00
subst:sg:acc:f    ą     a   1061 1.00
subst:sg:loc:m1   kim   ki  1030 0.99
subst:sg:inst:m1  kim   ki  1030 0.99
subst:sg:gen:m1   kiego ki  1030 0.99
subst:sg:dat:m1   kiemu ki  1030 0.99
subst:sg:acc:m1   kiego ki  1030 0.99
subst:pl:loc:m1   kich  ki  1030 0.99
(...)
```

Figure 3: The most common suffix changes for person last names.

6.4 Category-based Rules

The last extension is a set of category-specific lemmatization rules which override the initial set of lemmaitzation rules. The rules reflect the nature of lemmatization of specific proper name categories.

6.4.1 Road Names

The lemma for a road name that is an adjective should be in a genitive case and feminine gender. For instance, *ulicy Białej* (Eng. White street; a locative form) should be lemmatized to *ulica Biała* instead of *biały* which is the base form in the morphological dictionary of the common word.

[4] https://clarin-pl.eu/dspace/handle/11321/247

6.4.2 Voivodeship Names

Similar rule applies for Polish names of voivodeships. The lemmatized form must be in a nominative case and neutral gender instead of masculine which is the default base form in the morphological dictionary. For instance *województwie kieleckim* (Eng. *kieleckie voivodeship*; a locative form) should be lemmatized to *województwo kieleckie* instead of *kielecki*.

6.4.3 Person names

The majority of Polish names consists of a first name and a last name or two first names and a last name, i.e. a sequence of nouns. The generic rule for a sequence of nouns assumes that the first noun is the head of the phrase and the remaining nouns a the head's modifiers. The rule changes the case of the head to the nominative case and keep the case of the modifiers. In case of person names all elements must be changed to their nominative forms. The rule overrides the generic rule by changing all the words to their nominative forms.

6.5 Evaluation

Table 3 contains results for the extended version of PoLem on the dataset of named entities. The final version of PoLem obtained the accuracy of 89.80% for the case-insensitive (CI) evaluation and 87.35% for the case-sensitive (CS) evaluation on the train set. Comparing with the initial set of lemmatization rules we obtained an improvement of near 4–6 pp. Similar improvement was obtained on the test set which was not seen during the development. In the Appendix A we presented the lemmatization accuracy for each named entity category separately. The evaluation shows that there are still some major problems with lemmatization for some categories of named entities.

Evaluation	Train	Test
Named entities		
PoLem" CI	89.80%	88.45%
PoLem" CS	87.35%	86.17%

Table 3: Accuracy of the initial lemmatization rules.

The largest number of incorrect lemmas was obtained for *people names* (`nam_liv_person*`). There are several reasons for this relatively large number of errors. One of them is the gender ambiguity. There are many name forms which can be a male or a female name. For example *Antonia* can be a female name in nominative or a male name in genitive. The dictionaries of person names we used do not contain information about the name gander so we could not utilize the information about the gender assigned by the tagger. On the other hand, the tagger tends to treat most of male names in genitive as female names in nominative. To handle this type of problem some kind of post-processing with an access to the source document would be required. Similar problem applies to person last names. Different last names have the same inflected form so it is impossible to determine the correct nominative form without considering all variants of the same last name which appeared in the same document.

The second category with a high number of incorrect lemmas was *city name* (`nam_loc_gpe_city`). For this category the majority of errors were caused by the fact that there are many names which are also common words (nouns and adjectives). The names were assigned a nominative form of the common word while the expected lemma has a different form.

7 Summary

In the paper we deal with the problem of multi-word phrases and named entity lemmatization for Polish. We presented several baseline methods which do not provide satisfactory results. We showed that a small set of 27 rules was enough to cover all phrases with high accuracy. Latter, the set of initial rules was extended with a set of heuristic utilizing different types of lexicons, inflection rules and several new category-specific lemmatization rules to improve the lemmatization of named entities. The extended version of PoLem improved the accuracy of lemmatization by more than 4 percentage points for the case-insensitive evaluation on the train set and by 6 percentage points for the case-sensitive evaluation. Similar improvement was obtained on the test set which was not used in the development process.

The PoLem tool will be made available in a form of a web-service as a part of the CLARIN-PL infrastructure. The announcement will be published on `http://nlp.pwr.wroc.pl/polem`.

Acknowledgments

Work financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

A Detailed results for the test set

True	False	Accuracy	Method	Coverage
120	36	76.92%	nam_adj	1.11%
105	18	85.37%	nam_adj_city	0.87%
430	10	97.73%	nam_adj_country	3.12%
10	13	43.48%	nam_adj_person	0.16%
15	22	40.54%	nam_eve	0.26%
141	36	79.66%	nam_eve_human	1.25%
2	1	66.67%	nam_eve_human_aniversary	0.02%
30	1	96.77%	nam_eve_human_cultural	0.22%
20	6	76.92%	nam_eve_human_holiday	0.18%
103	6	94.50%	nam_eve_human_sport	0.77%
1	0	100.00%	nam_eve_natural_phenomenom	0.01%
5	0	100.00%	nam_fac	0.04%
7	4	63.64%	nam_fac_bridge	0.08%
0	2	0.00%	nam_fac_crossroad	0.01%
182	36	83.49%	nam_fac_goe	1.55%
26	13	66.67%	nam_fac_goe_stop	0.28%
8	3	72.73%	nam_fac_park	0.08%
290	35	89.23%	nam_fac_road	2.30%
27	14	65.85%	nam_fac_square	0.29%
15	14	51.72%	nam_fac_system	0.21%
18	1	94.74%	nam_liv_animal	0.13%
8	6	57.14%	nam_liv_character	0.10%
110	26	80.88%	nam_liv_god	0.96%
38	4	90.48%	nam_liv_habitant	0.30%
1945	**314**	86.10%	**nam_liv_person**	16.02%
96	32	75.00%	nam_liv_person_add	0.91%
1366	**122**	91.80%	**nam_liv_person_first**	10.55%
1421	**184**	88.54%	**nam_liv_person_last**	11.38%
7	0	100.00%	nam_liv_plant	0.05%
24	8	75.00%	nam_loc	0.23%
38	10	79.17%	nam_loc_astronomical	0.34%
34	25	57.63%	nam_loc_country_region	0.42%
72	33	68.57%	nam_loc_gpe_admin1	0.74%
21	8	72.41%	nam_loc_gpe_admin2	0.21%
66	4	94.29%	nam_loc_gpe_admin3	0.50%
1121	**140**	88.90%	**nam_loc_gpe_city**	8.94%
11	0	100.00%	nam_loc_gpe_conurbation	0.08%
785	22	97.27%	nam_loc_gpe_country	5.72%
60	10	85.71%	nam_loc_gpe_district	0.50%
42	10	80.77%	nam_loc_gpe_subdivision	0.37%
36	5	87.80%	nam_loc_historical_region	0.29%
1	0	100.00%	nam_loc_hydronym	0.01%
1	0	100.00%	nam_loc_hydronym_bay	0.01%
1	0	100.00%	nam_loc_hydronym_lagoon	0.01%
7	2	77.78%	nam_loc_hydronym_lake	0.06%
3	0	100.00%	nam_loc_hydronym_ocean	0.02%
42	8	84.00%	nam_loc_hydronym_river	0.35%
6	0	100.00%	nam_loc_hydronym_sea	0.04%
4	0	100.00%	nam_loc_land	0.03%
60	0	100.00%	nam_loc_land_continent	0.43%
1	0	100.00%	nam_loc_land_desert	0.01%
23	4	85.19%	nam_loc_land_island	0.19%
48	7	87.27%	nam_loc_land_mountain	0.39%
5	0	100.00%	nam_loc_land_peak	0.04%
7	0	100.00%	nam_loc_land_peninsula	0.05%
4	0	100.00%	nam_loc_land_protected_area	0.03%
19	5	79.17%	nam_loc_land_region	0.17%
1	0	100.00%	nam_num	0.01%
1	0	100.00%	nam_num_flat	0.01%
25	0	100.00%	nam_num_house	0.18%
13	0	100.00%	nam_num_phone	0.09%
2	0	100.00%	nam_num_postal_code	0.01%
6	0	100.00%	nam_org	0.04%
333	43	88.56%	nam_org_company	2.67%
25	23	52.08%	nam_org_group	0.34%
49	10	83.05%	nam_org_group_band	0.42%
227	31	87.98%	nam_org_group_team	1.83%
521	34	93.87%	nam_org_institution	3.93%
17	1	94.44%	nam_org_institution_full	0.13%
153	27	85.00%	nam_org_nation	1.28%
420	30	93.33%	nam_org_organization	3.19%
5	2	71.43%	nam_org_organization_sub	0.05%
141	9	94.00%	nam_org_political_party	1.06%
50	4	92.59%	nam_oth	0.38%
1	0	100.00%	nam_oth_address_street	0.01%
52	18	74.29%	nam_oth_currency	0.50%
7	1	87.50%	nam_oth_data_format	0.06%
1	0	100.00%	nam_oth_ip	0.01%
33	4	89.19%	nam_oth_license	0.26%
1	0	100.00%	nam_oth_mail	0.01%
15	2	88.24%	nam_oth_position	0.12%
118	59	66.67%	nam_oth_tech	1.25%
11	0	100.00%	nam_oth_www	0.08%
2	2	50.00%	nam_pro	0.03%
19	1	95.00%	nam_pro_award	0.14%
135	39	77.59%	nam_pro_brand	1.23%
4	0	100.00%	nam_pro_media	0.03%
194	16	92.38%	nam_pro_media_periodic	1.49%
11	2	84.62%	nam_pro_media_radio	0.09%
31	3	91.18%	nam_pro_media_tv	0.24%
146	48	75.26%	nam_pro_media_web	1.38%
61	33	64.89%	nam_pro_model_car	0.67%
1	0	100.00%	nam_pro_model_phone	0.01%
6	0	100.00%	nam_pro_model_plane	0.04%
57	29	66.28%	nam_pro_software	0.61%
22	1	95.65%	nam_pro_software_game	0.16%
2	2	50.00%	nam_pro_software_os	0.03%
1	0	100.00%	nam_pro_software_version	0.01%
135	19	87.66%	nam_pro_title	1.09%
19	2	90.48%	nam_pro_title_album	0.15%
8	1	88.89%	nam_pro_title_article	0.06%
3	0	100.00%	nam_pro_title_boardgame	0.02%
20	5	80.00%	nam_pro_title_book	0.18%
45	26	63.38%	nam_pro_title_document	0.50%
5	0	100.00%	nam_pro_title_painting	0.04%
2	0	100.00%	nam_pro_title_radio	0.01%
8	3	72.73%	nam_pro_title_song	0.08%
12	2	85.71%	nam_pro_title_treaty	0.10%
28	3	90.32%	nam_pro_title_tv	0.22%
14	2	87.50%	nam_pro_vehicle	0.11%
12307	1797	87.26%	Total	100.00%

References

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012. KPWr: Towards a Free Corpus of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of LREC'12*. ELRA, Istanbul, Turkey.

Łukasz Degórski. 2012. Towards the lemmatisation of Polish nominal syntactic groups using a shallow grammar. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7053 LNCS(250467):370–378.

Katarzyna Głowińska. 2008. Anotacja składniowa NKJP. In (Przepiórkowski, 2008), pages 107–127. https://books.google.pl/books?id=V076OgAACAAJ.

Jacek Małyszko, Witold Abramowicz, Aagata Filipowska, and Tomasz Wagner. 2015. Lemmatization of Multi-Word Entity Names for Polish Language Using Rules Automatically Generated Based on the Corpus Analysis. *Human Language Technologies as a Challenge for Computer Science and Linguistics* pages 540–544.

Adam Przepiórkowski. 2008. *Powierzchniowe przetwarzanie języka polskiego*. Problemy współczesnej nauki, teoria i zastosowania: Inżynieria lingwistyczna. Akademicka Oficyna Wydawnicza "Exit". https://books.google.pl/books?id=V076OgAACAAJ.

Adam Radziszewski. 2013a. A tiered CRF tagger for Polish. In R. Bembenik, Ł. Skonieczny, H. Rybiński, M. Kryszkiewicz, and M. Niezgódka, editors, *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*, Springer Verlag.

Adam Radziszewski. 2013b. Learning to lemmatise Polish noun phrases. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, {ACL} 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*. The Association for Computer Linguistics, pages 701–709. http://aclweb.org/anthology/P/P13/P13-1069.pdf.

Adam Radziszewski, Adam Wardyński, and Tomasz Śniatowski. 2011. WCCL: A Morpho-syntactic Feature Toolkit. In Ivan Habernal and Václav Matousek, editors, *Proceedings of Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic*. Springer, Pilsen, volume 6836 of *Lecture Notes in Computer Science*, pages 434—-441.

Marcin Woliński. 2006. *Morfeusz — a Practical Tool for the Morphological Analysis of Polish*, Springer Berlin Heidelberg, Berlin, Heidelberg, pages 511–520.