Part of Speech Tagging for Polish: State of the Art and Future Perspectives

Łukasz Kobyliński and Witold Kieraś

Institute of Computer Science, Polish Academy of Sciences
Jana Kazimierza 5, 01-248 Warszawa, Poland
lkobylinski@ipipan.waw.pl, wkieras@uw.edu.pl

Abstract. In this paper we discuss the intricacies of Polish language part of speech tagging, present the current state of the art by comparing available taggers in detail and show the main obstacles that are a limiting factor in achieving an accuracy of Polish POS tagging higher than 91% of correctly tagged word segments. As this result is not only lower than in the case of English taggers, but also below those for other highly inflective languages, such as Czech and Slovene, we try to identify the main weaknesses of the taggers, their underlying algorithms, the training data, or difficulties inherent to the language to explain this difference. For this purpose we analyze the errors made individually by each of the available Polish POS taggers, an ensemble of the taggers and also by a publicly available well-known OpenNLP tagger, adapted to Polish tagset. Finally, we propose further steps that should be taken to narrow down the gap between Polish and English POS tagging performance.

1 Introduction

There is an ongoing discussion whether the problem of part of speech tagging is already solved, at least for English (see e.g. [1]), by reaching the tagging error rates similar or lower than the human inter-annotator agreement, which is ca. 97%. In the case of languages with rich morphology, such as Polish, there is however no doubt that the accuracies of around 91% delivered by taggers leave much to be desired and more work is needed to proclaim this task as solved.

The problem is that while the work on taggers for Polish continues for more than a decade now, the progress is very slow and even reaching the goal of 97% seems very distant. This is in spite of using the latest achievements in machine learning, increasing the size of training data and perfecting the available morphosyntactic dictionaries. The less than perfect quality of automatic POS tagging impacts other NLP tools and the accuracy of other layers of syntactic and semantic annotation generated by these tools.

In this paper, we try to answer the question what is the underlying difficulty in getting closer to the error rates presented by taggers for other languages. For one thing, the task of POS tagging in the case of Polish is much more difficult than in the case of English because of the morphology of the language: the set of all possible tags consists of more than 4 000 choices (ca. 1 500 appear

in a manually tagged corpus) versus 30–200 for English. This doesn't answer the question however, because taggers for languages with tagsets of similar size, such as Czech and Slovene, have proved to achieve higher accuracies. We discuss Polish tagset and morphological dictionary in Section 2.

If not the language itself, maybe the difficulty lies in the available language resources, or the chosen approach to use them for training the taggers? We briefly present the structure of the National Corpus of Polish, used as the training material for ML methods in Section 2.3, discuss the previously proposed taggers in Section 3 and show the difficulty in adapting existing methods to Polish in Section 4.

We then elaborate on the process of evaluating the quality of taggers in Section 5.1, follow with experimental data concerning individual taggers in Section 5.2 and combining them into an ensemble of classifiers in Section 5.3. In Section 6 we discuss the results, the most common mistakes made by the taggers and the problems inherent to the language. Finally, we close with conclusions and perspectives in Section 7.

2 Available Resources

2.1 Polish Tagset

There have been several attempts to define a tagset for Polish, usually connected with the development of a reference text corpus, a morphological dictionary, or a tagger. The first formulation of a tagset, which has been used for tagging the IPI PAN corpus,[1] has been proposed in [2]. In this paper we use the more current version of the tagset, proposed for annotating the National Corpus of Polish and described in [3].

2.2 Morfeusz – Morphosyntactic Dictionary

Morfeusz is the most commonly used morphological analyzer for Polish. Although it was recently reimplemented from scratch and significantly enhanced [4], due to technical reasons we were forced to use its previous version [5] as neither the taggers nor the training corpus were adapted to use the newer version of the analyzer.

Morfeusz uses a lexical input obtained from the Grammatical Dictionary of Polish [6], the largest database of Polish inflectional paradigms.[2] The dictionary consists of over 330 000 lexical entries and nearly 7 million wordforms representing over 1100 different inflectional patterns. Its extensive lexical basis goes back to even last decades of 18th century vocabulary which on one hand makes it a desirable resource for morphological analysis, but on the other hand compels to deal with large amounts of archaic, obsolete, dialectal and otherwise stylistically marked lexical entries.

[1] IPI PAN corpus was the first large, POS-tagged reference corpus of Polish, now superseded by the National Corpus of Polish.

[2] Now available also on-line at `http://sgjp.pl`

2.3 National Corpus of Polish

For training and testing purposes we have used a manually annotated corpus of about 1.2 million words created for National Corpus of Polish project [7]. Contrary to many other resources used for training statistical POS taggers such as Penn Treebank or Prague Dependency Treebank, our training corpus consists not only of newspaper samples but also of fiction and non-fiction, scientific and educational texts, Internet (blogs, fora, Usenet, Wikipedia and other webpages) and oral text samples (media and conversations). Newspapers and magazines constitute only 49% of the corpus. This diversity of data gives us a wider representation of language registers and genres, but may presumably affect both training and evaluation processes (see Section 6.4 for a discussion on that topic).

3 Previous Work – POS Taggers for Polish

The first tagger for Polish, proposed by [8], has never been publicly released and is not included in further discussion. TaKIPI tagger, described in [9], assumes a heterogeneous approach to tagging, combining hand-crafted rules with decision trees. TaKIPI is tied to the original, now obsolete IPI PAN corpus tagset and is also excluded from further experiments.

Currently available taggers, using the latest version of the tagset, include: Pantera [10] (an adaptation of the Brill's algorithm to morphologically rich languages), WMBT [11] (a memory based tagger), WCRFT [12] (a tagger based on Conditional Random Fields) and Concraft [13] (another approach to adaptation of CRFs to the problem of POS tagging). Evaluation of performance of a combination of these taggers has been presented in [14].

4 OpenNLP – A Case Study in Adapting a Known Tagger to Polish

One of the questions that may be asked is whether we really need another implementation of a POS tagger, given that so many have already been proposed and in fact several are open sourced and freely available. Such implementations have an unquestionable advantage of being easy to use, supplemented with well-developed user interfaces and (possibly) well tested by multiple users and developers. Unfortunately, such generic tools are not easily adapted to specific languages and associated language resources, or the implemented algorithms do not perform well in case of highly inflective languages, with large tagsets.

An example of such an existing implementation of NLP algorithms, including a POS tagger, is the Apache OpenNLP library.[3] It is a Java-based toolkit, supporting such NLP tasks as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution. Two main machine learning methods are used for solving these tasks: maximum

[3] `http://opennlp.apache.org/`

entropy and perceptron. The selection of available ML methods presents another problem when using an existing toolkit. Although maximum entropy has been successfully implemented in many English taggers, several more recent alternatives have been proposed to date (such as CRFs).

We were able to successfully train both a maximum entropy and perceptron-based POS tagger using the OpenNLP toolkit; the accuracy of these models is presented in Table 1 and Figure 3.[4] As expected, the tagger performs worse than the approaches proposed specifically for Polish. We elaborate on some of our observations below.

Tiered vs one-pass classification Considering the large number of POS tags in Polish tagset, most of the tested taggers divide the task of selecting the correct tag for a word token into several stages, or tiers. For example, the grammatical class is first disambiguated by the WCRFT tagger (1 out of 35 possibilities in the NCP corpus[5]) and after that decision the number of possible combinations of more specific tag parts (e.g. grammatical number, case) is greatly reduced. In the case of OpenNLP, the tags are selected directly, in one pass, which amounts to a problem of selecting 1 out of 1000 possible combinations.

Use of training data The task of tagging is understood as the task of morphosyntactic disambiguation in the case of most of Polish taggers. As such, the models in these taggers are trained to eliminate incorrect possible tags (produced by the morphosyntactic dictionary) for each of the analyzed word tokens. Therefore, during training, not only the correct tag for a particular word is taken into account, but also the set of all possible selections. This helps to make the correct decision in a similar context during tagging. In the case of OpenNLP, only the correct tag is used for training the model.

Use of morphosyntactic dictionary As stated above, previously proposed taggers for Polish rely heavily on morphosyntactic dictionaries and are in fact trained to disambiguate between one of several possibilities generated by the dictionary and not to produce the tags themselves. OpenNLP on the other hand uses the dictionary only for speeding up the beam search algorithm and the model is trained to select one of all possible (previously seen during training) tags. As can be seen from the results in Table 1, this is an advantage in case of tagging unknown words (unknown to the dictionary).

5 Evaluating and Combining the Taggers

5.1 Evaluation Methodology

It is undoubtedly difficult to compare the performance of various approaches to tagging between different languages. That is because of the differences inherent

[4] As the difference in accuracy between these two approaches turned out not to be statistically significant, we have limited further experiments to maximum entropy models. Trained models available at: `http://zil.ipipan.waw.pl/OpenNLP`

[5] In fact, this number is further reduced by the morphosyntactic analyser.

to languages, which were mentioned earlier (e.g. the size of the tagset), or differences in the structure and character of the training and testing material. The problem of evaluating taggers is however much broader and may lead to misconceptions about their real-world performance even when looking at methods proposed for the same language.

There are some obvious conditions that have to remain unchanged to warrant an unbiased comparison: training and testing corpora, additional dictionaries, or the statistical method used to calculate the tagger accuracy. The more subtle decision, often not explicitly stated, is the choice of exact part of the processing pipeline at which the tagger accuracy is measured.

We have decided to continue the line of thought proposed in [15] and evaluate the performance of taggers given plain text as input and measure the accuracy of correct tag assignments to correctly segmented word tokens. This mimics the real-world application of taggers, but hides several stages of processing into one accuracy result (token and sentence segmentation, as well as the morphosyntactic disambiguation itself). Consequently, we also use the accuracy measure proposed in [15], namely the *accuracy lower bound* (Acc_{lower}), which treats all segmentation mistakes as tagger errors. We also distinguish errors made on tokens which are known to morphosyntactic dictionary (Acc^K_{lower}) and on tokens for which no morphosyntactic interpretation is provided by the dictionary (Acc^U_{lower}).

For the details of data preprocessing we followed the procedure described in [15] and [14].

5.2 Performance of Individual Taggers

We have firstly re-evaluated all available Polish taggers, using the evaluation methodology described above and language resources described in Section 2. Accuracy measures have been calculated by performing ten-fold cross validation of the available training data. The results (presented in Table 1) are on-par with previously published data, but this time we have also included the results for OpenNLP tagger, evaluated using the same methodology.

Table 1: Performance of Individual Taggers.

n	Tager	Acc_{lower}	Acc^K_{lower}	Acc^U_{lower}	training time	tagging time
1	Pantera	88.95%	91.22%	15.19%	2 624 s	186 s
2	WMBT	90.33%	91.26%	60.25%	548 s	4 338 s
3	WCRFT	90.76%	91.92%	53.18%	27 242 s	420 s
4	Concraft	91.07%	92.06%	58.81%	26 675 s	403 s
5	OpenNLP	87.24%	88.02%	62.05%	11 095 s	362 s

We have also compared tagger efficiency by measuring training and tagging times of each of the methods on the same machine. We used 1.1M tokens both

for training and tagging stages and measured the total processing time, including model loading/saving time and other I/O operations (e.g. reading/writing the tokens).

It is worth noting that while the overall accuracy of the OpenNLP tagger is significantly lower than for any other tested approach, it performs better in the case of words unknown to the morphological dictionary. This might suggest that there is room for improvement in the implementations of the best performing methods in the case of such unknown tokens. Usually, a different tagging strategy has to be employed, as the task is different than the usual morphological disambiguation, as in the case of known words.

5.3 Ensemble of Taggers

Next, we tested the hypothesis that the accuracy of an ensemble of taggers increases with each added component tagger, even if its accuracy is lower than the average accuracy of the group. This is because wrong decisions are usually different between taggers and they do not negatively influence the overall accuracy of a voted ensemble. We have indeed observed a slight positive impact of including the OpenNLP tagger into an ensemble of all the tested taggers (see Figure 1).

In this experiment we have used the setup and strategies described in [14]. The accuracy of an 'oracle' tagger is a hypothetical result of a perfect ensemble voting strategy, in which the correct choice is always made among the tags produced by individual taggers. 'Simple' approach is majority voting, 'weighted' gives advantage to better taggers in case of a draw, while 'per-class' gives advantage to taggers which are known to perform better for a particular grammatical class.

6 Why Are the Taggers Wrong?

6.1 The Most Common Errors

We use the term "part of speech tagging" referring to the process of choosing a proper morphosyntactic interpretation for a given token, which means that it is not restricted to choosing a correct part of speech label, but also a proper lemma and proper values of all grammatical categories of the wordform. In fact, when it comes to literal meaning of the term (i.e. choosing a correct POS label), the problem is rather simple and all tested taggers obtain relatively good results in this task. For nouns (subst), adjectives (adj), numerals (num), past tense verb forms (praet), passive adjectival participle (ppas), active adjectival participle (pact) and pronouns (ppron12 and ppron3) taggers tend to assign a correct part of speech rather than an incorrect one. Significant problems in assigning a correct part of speech label can be observed mostly in the area of grammatically not inflected words such as prepositions (prep), particles (qub), adverbs (adv) and

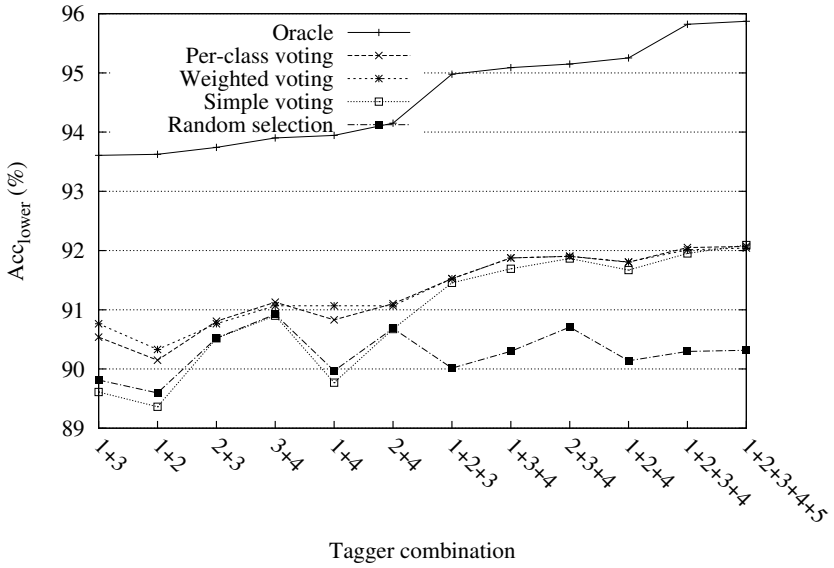

Fig. 1: Accuracy of an ensemble of taggers.

conjunctions (conj).[6] However, error rate grows dramatically if one expects the tagger to also choose the correct lemma and all grammatical category values such as: gender, number, case, person, tense, aspect etc. See Figure 2 for a comparison between the percentage of errors made by selecting an incorrect grammatical class vs. errors in tagging other grammatical categories.

6.2 Homonymy

The need for disambiguation of the morphological analyzer's output arises from homonymy. In general, the more homonymy in a certain language, the more difficult it is to disambiguate. Polish is definitely on the harder side as its average homonymy rate reaches 47%, which means that nearly every second word in a text is morphologically ambiguous.

Problems with homonymous words might be of different nature. One of the homonymy types the most difficult to deal with is syncretism, which is also the most common one. By syncretism we understand homonymy restricted to the inflectional paradigm of a single lexeme.[7] In other words, it means that some tokens can be analyzed as different wordforms of the same lexeme. For example,

[6] One exception from this general observation are gerunds (ger), which are however systematically homonymous with nouns and thus are extremely difficult to disambiguate not only for taggers, but also for the human annotator.

[7] This phenomenon is typical of fusional languages such as Polish and other Slavonic languages.

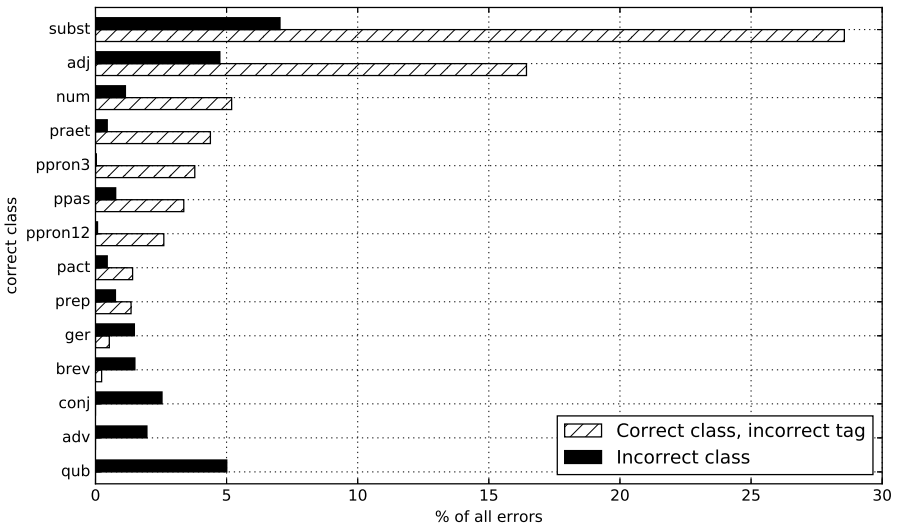

Fig. 2: Tagger errors in choosing the correct grammatical class vs selecting correct tag in a correctly identified class.

the noun PIES 'dog' declines by two grammatical numbers (singular and plural) and seven grammatical cases, so its full paradigm consists of fourteen wordforms, but only ten orthographically distinct strings, seven wordforms being syncretic. In the case of PIES the syncretic wordforms are: *psa* (genitive or accusative singular), *psie* (locative or vocative singular) and *psy* (nominative, accusative or vocative plural). In fact, every Polish noun and adjective shares this feature to some extent. Most of the syncretisms are systemic (i.e. typical to certain declension classes), but some might be also accidental.

While analyzing the most frequent errors produced by the taggers, one can observe that syncretism is responsible for a significant part of them. Table 2a contains a list of most commonly mismatched part of speech tags, which reflects frequent homonyms between grammatical classes like nouns (subst) and adjectives (adj) or conjunctions (conj) and particles (qub). On the other hand, table 2b presents the most commonly mismatched full tags restricted to the noun class, which reflects the most typical syncretisms in noun paradigms characteristic to certain grammatical genders (feminine, neuter and three masculine subgenders named m1, m2, m3). It is easy to observe that these mismatches are usually restricted only to case and sometimes to number, but never to gender (tags in column 1 and 2 are always labeled with the same gender). The first four rows in the table represent two most common syncretisms between singular nominative and singular accusative of masculine inanimate and neuter nouns. Since nominative and accusative are cases typically assigned to subject and object of a sentence, achieving 100% accuracy of morphological disambiguation of those

Table 2: The most common errors made by the taggers.

(a) Errors in grammatical class selection.

tagger	reference	% of errors
adj	subst	1.9422
conj	qub	1.7373
subst	adj	1.5811
adv	qub	1.5518
subst	ger	1.4835
qub	conj	1.4737
subst	brev	1.3664
ger	subst	1.2493
num	adj	1.0541
ppas	adj	1.0541
qub	adv	0.8882
adj	ppas	0.6930

(b) Errors in specific tagging of *subst* class.

tagger	reference	% of errors
sg:nom:m3	sg:acc:m3	1.8641
sg:acc:m3	sg:nom:m3	1.4933
sg:acc:n	sg:nom:n	1.3078
sg:nom:n	sg:acc:n	1.1419
pl:nom:m3	pl:acc:m3	0.8686
pl:acc:f	pl:nom:f	0.7613
pl:nom:f	pl:acc:f	0.6930
pl:acc:m3	pl:nom:m3	0.6637
sg:gen:m1	sg:acc:m1	0.6539
sg:acc:m1	sg:gen:m1	0.5173
pl:nom:n	pl:acc:n	0.4685
sg:gen:f	pl:gen:f	0.4587

tokens would require at least partial syntactic analysis. This actually shows that the most frequent mistakes in assigning tags to nouns are also those that are most difficult to avoid. All the entries in the table represent syncretisms typical of large classes of Polish nouns.

Syncretism is not restricted to nouns, but involves also adjectives as well as past forms of verbs. All of these are the sources of common errors of taggers. This proves that syncretism is a serious problem in morphological disambiguation, far more difficult than simple part of speech tagging.

But ambiguous analyses do not necessarily involve only syncretism. Another source of the problem is proper homonymy of inflectional forms of different lexemes – either systemic, serial and motivated by derivational processes or accidental, connecting words derivationally and semantically unrelated. The former was extensively researched by [16] and could be exemplified by masculine nouns differing only in subgender (personal, animate or inanimate) and thus sharing most of the forms in their paradigms, e.g. ADMIRAŁ 'admiral of the fleet' (m1, masculine personal) or 'red admiral butterfly' (m2, masculine animate). The same applies to three homonymous nouns BOKSER 'boxer' which reflect the polysemy of the English word BOXER ('sportsman', 'dog' and 'engine') but is marked on the morphological level by a slightly different inflection. Another example are productive series of pairs such as FIZYK 'physicist' (masculine) and FIZYKA 'physics' (feminine) which systematically share three homonymous forms in their paradigms. Accidental homonymy could be on the other hand illustrated by nouns PALETA (feminine) 'palette/pallet', PALET (masculine) 'payment warrant; archaic legal term', PALETO (neuter) 'coat; archaic loan word'. Each represents a different grammatical gender and thus different inflectional type, but

they share some forms, in particular, all three have the same singular locative *palecie*.

An interesting example of this phenomenon is its occasional conjugation with the ambiguity of prepositions' case government. Usually prepositional context helps in disambiguating a noun that is governed by the preposition by ruling out interpretations inconsistent with the case government. However, in some cases the ambiguity of a noun "responds" to the ambiguity of a preposition. Consider a phrase *w krypcie* consisting of two tokens: a preposition W that requires a nominal phrase either in accusative or locative, and a word *krypcie* that could be analyzed either as a locative singular form of the noun KRYPTA 'vault', or an accusative plural of the noun KRYPEĆ 'primitive wooden shoe'. The latter interpretation of the noun is highly unlikely since the word KRYPEĆ is both obsolete and dialectal but it cannot be ruled out on the basis of prepositional government and it calls for other solutions.

Possible linguistic solution to problems illustrated above could be extensive use of non-inflectional information about lexical units extracted from the dictionary, especially about all kinds of stylistic markedness of words (archaisms, dialectalisms, slang etc.) and scope restrictions of their usage (scientific jargon, medical terminology etc.). Also any kind of systematic statistical information about frequencies of word occurrences might improve disambiguation results.

6.3 Lemmatization vs Disambiguation

Another issue that was illustrated by some examples above is the problem of lemmatization. All the tested taggers were aimed at disambiguating morphosyntactic tags, while none of them treated lemmatization as a separate task. In practice the taggers simply ignore the lemmas and take only tags into account. This strategy is reasonable since in most cases it should lead to choosing the correct lemma as well, but sometimes it could result in choosing a completely unlikely lemma before a more probable one. This applies to the example of *w krypcie* shown above – some taggers choose the archaic and rare word KRYPEĆ before stylistically unmarked KRYPTA. The same applies to the nouns OGRÓD 'garden' and OGRODA 'fence' which share the singular locative form *ogrodzie*. Some taggers choose the archaic OGRODA before stylistically unmarked OGRÓD in locative context, which means that in solving such cases they do not take any other significant factor into account and if tags are identical or different but equally justified (as in the case of KRYPEĆ), determining a lemma is more or less a matter of random choice. Avoiding such situations requires an approach in which lemmatization and disambiguation are separate tasks of a tagger as it was suggested in [17]. The basis on which a tagger should resolve a certain lemma is itself a separate issue, but at least two sources of information may turn out useful: text frequency of words and stylistical markedness provided by a dictionary.

6.4 Training Data

A comparison of tagger evaluations between Polish and other languages reveals that there is a significant difference in the structure of National Corpus of Polish, the training material for all presented experiments, and corpora used for training and testing taggers in other languages. For example, English taggers are trained and tested on the part of Penn Treebank which consists exclusively of newspaper articles from the Wall Street Journal. As stated in Section 2.3, the NCP consists of a variety of sources, including newspaper articles, but also books, spoken dialogues and data collected from discussion groups on the Internet.

National Corpus of Polish is also smaller than some of the corpora in other languages (1.2M tokens). As such, we wanted to test the hypothesis that 1) the structure of the corpus might influence (negatively) the accuracy of POS tagging and 2) extending the training corpus with more reference data is another possible approach to increasing tagger performance, besides work on the methods themselves.

In order to test the first hypothesis we have evaluated the performance of the OpenNLP tagger on several subcorpora of the NCP, removing data from the sources such as discussion forums, spoken dialogue and books. Each of the subcorpora was chosen to contain roughly the same number of tokens (ca. 54% of the whole corpus), to eliminate the influence of training data size. Based on the experimental data presented in Table 3 we may indeed observe that there is a relationship between the degree of homogeneity of the data and tagging accuracy. This supports the argument that tagging results for Polish are not directly comparable to other languages, for which evaluations are commonly performed on corpora consisting exclusively of newspaper articles.

Table 3: Accuracy of tagging vs source of the training and testing data (ten-fold cross-validation).

Train/test data	Acc_{lower}	Acc^{K}_{lower}	Acc^{U}_{lower}	$Avg_{unknown}$
all	85.45%	86.25%	60.01%	3.04%
without: *internet*	85.54%	86.37%	60.68%	2.90%
w/o: *internet*, *spoken*	85.71%	86.47%	60.76%	2.96%
w/o: *internet*, *spoken*, *books*	86.21%	87.10%	61.83%	3.50%

The influence of the training data size on each of the tested methods has been presented in Figure 3. In this experiment, we have trained the taggers with randomly drawn subsets of the available training data, increasing data size from 10 000 tokens to 1M tokens, and tested their accuracy on a 100 000 tokens data set. For the best performing taggers, doubling the training data size results in ca. 1 percentage point increase in accuracy.

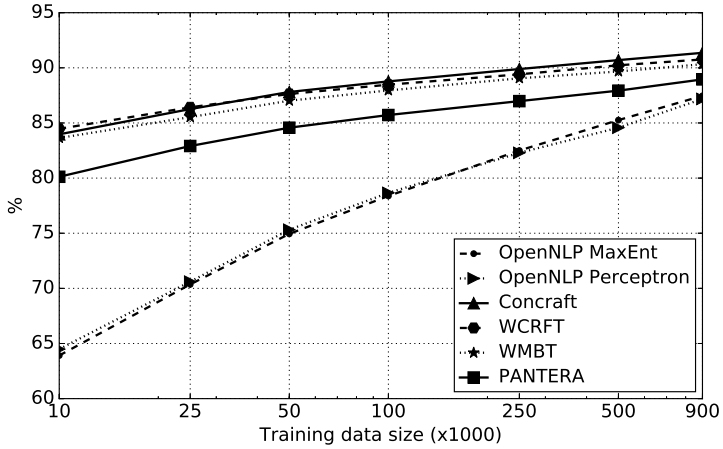

Fig. 3: Learning curve for the tested taggers. Test data size: 100 000 tokens.

7 Conclusions and Future Work

In conclusion, we believe that more work concerning Polish language resources is needed to overcome the problem of limited accuracy of POS taggers. Tagging accuracy is directly related to the complexity of this task and Polish is one of the languages with largest tagsets. The accuracy is also directly related to the size of available training data and morphological dictionaries. Some work in this area is already in progress, as the new version of Morfeusz analyzer is under development.

We have also shown that the specific data, which is usually used to evaluate the performance of Polish taggers may negatively impact their results, in comparison with evaluations done for other languages. The linguistic quality and consistency of newspaper articles is usually much higher than that of a text acquired from the Internet, or transcribed dialogues.

In our opinion, future work on POS taggers for Polish should focus on utilizing more of the information available in external language resources (such as stylistic marks of words in the morphological dictionary), tackle the problem of unknown words in more efficient way and also address the problem of lemmatization, which was left out in taggers to date.

Acknowledgment

Work financed by the Polish Ministry of Science and Higher Education, a program in support of scientific units involved in the development of a European research infrastructure for the humanities and social sciences in the scope of the CLARIN ERIC consortium.

References

1. Manning, C.D.: Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In: Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I. CICLing'11, Berlin, Heidelberg, Springer-Verlag (2011) 171–189
2. Przepiórkowski, A., Woliński, M.: The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish. In: Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), EACL 2003. (2003) 109–116
3. Przepiórkowski, A.: A comparison of two morphosyntactic tagsets of Polish. In Koseska-Toszewa, V., Dimitrova, L., Roszko, R., eds.: Representing Semantics in Digital Lexicography: Proceedings of MONDILEX Fourth Open Workshop, Warsaw (2009) 138–144
4. Woliński, M.: Morfeusz reloaded. [18] 1106–1111
5. Woliński, M.: Morfeusz — a practical tool for the morphological analysis of Polish. In Kłopotek, M.A., Wierzchoń, S.T., Trojanowski, K., eds.: Intelligent Information Processing and Web Mining. Advances in Soft Computing. Springer-Verlag, Berlin (2006) 503–512
6. Saloni, Z., Woliński, M., Wołosz, R., Gruszczyński, W., Skowrońska, D.: Słownik gramatyczny języka polskiego. 2. edn. Warszawa (2012)
7. Przepiórkowski, A., Bańko, M., Górski, R., Lewandowska-Tomaszczyk, B., eds.: Narodowy Korpus Języka Polskiego. Warszawa (2012)
8. Dębowski, Ł.: Trigram morphosyntactic tagger for polish. In: In Proceedings of the International IIS:IIPWM'04 Conference, Springer-Verlag (2004) 409–413
9. Piasecki, M.: Polish tagger TaKIPI: Rule based construction and optimisation. Task Quarterly **11** (2007) 151–167
10. Acedański, S.: A morphosyntactic Brill tagger for inflectional languages. In: Advances in Natural Language Processing. (2010) 3–14
11. Radziszewski, A., Śniatowski, T.: A Memory-Based Tagger for Polish. In: Proceedings of the LTC 2011. (2011)
12. Radziszewski, A.: A tiered CRF tagger for Polish. In Bembenik, R., Skonieczny, Ł., Rybiński, H., Kryszkiewicz, M., Niezgódka, M., eds.: Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions. Springer Verlag (2013) 215–230
13. Waszczuk, J.: Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In: Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), Mumbai, India (2012) 2789–2804
14. Kobyliński, Ł.: PoliTa: A multitagger for Polish. [18] 2949–2954
15. Radziszewski, A., Acedański, S.: Taggers gonna tag: an argument against evaluating disambiguation capacities of morphosyntactic taggers. In: Proceedings of TSD 2012. LNCS, Springer-Verlag (2012) 81–87
16. Awramiuk, E.: Systemowość polskiej homonimii międzyparadygmatycznej. Białystok (1999)
17. Radziszewski, A.: Evaluation of lemmatisation accuracy of four polish taggers. In: Proceedings of the LTC 2013. (2013)
18. Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavík, Iceland, ELRA (2014)