On the Evaluation of Two Polish Taggers[*]

Danuta Karwańska and **Adam Przepiórkowski**

Institute of Computer Science, Polish Academy of Sciences
and University of Warsaw

Abstract

The aim of this paper is to discuss different ways in which the performance of morphosyntactic taggers may be compared and to present the results of the comparison of two such Polish taggers. This comparison sets the stage for measuring the performance of future taggers of Polish, including the tagger being developed within the National Corpus of Polish.

Keywords: POS tagger, Polish, morphosyntactic disambiguation, TaKIPI, evaluation, IPI PAN Tagset, National Corpus of Polish.

1. Introduction

The general issue addressed in this paper is how to compare the performance of two different morphosyntactic taggers for a given language. Obviously, the problem has been addressed before, but usually in the setting in which both taggers employ exactly the same tagset, both produce exactly one tag for each token, and also the gold standard contains exactly one tag for each token. As we will see below, all these assumptions are violated to some degree in the current setting.

In cases when these three assumptions hold the standard procedure is as follows:

- the same gold standard corpus is used to evaluate both taggers;
- in the procedure known as 10-fold cross-evaluation, the gold standard is randomly split into 10 portions of the same size, and for each such portion, the tagger is trained on the remaining nine and evaluated on that portion;
- the measure known as *accuracy* (and its dual, *error*) is used for the evaluation within each of the ten experiments;

[*] This article is largely based on the unpublished report Karwańska 2009. The authors wish to thank Łukasz Dębowski and Maciej Piasecki for their comments on an earlier version of this paper. Research funded in 2007–2010 by a research and development grant from the Polish Ministry of Science and Higher Education.

- the final numerical evaluation of the tagger is the average of the ten evaluation results;
- the tagger with significantly higher *accuracy* is considered better.

Accuracy is commonly understood as the percent of *tokens* (roughly, words; also called *segments* in the following sections) on which the gold standard and the tagger agree (Manning & Schütze 1999, p. 342), although sometimes it is defined as the percent of *tags* on which they agree (Jurafsky & Martin 2009, p. 189). Conversely, *error* is the percent of tokens (or tags) on which the gold standard and the tagger *dis*agree. As we will see below, it is not clear, however, how to interpret these measures in case more than one tag is considered correct for a given token.

Why should ever more than one tag be assigned to a token, whether within a gold standard or by a tagger? As discussed, e.g., in Oliva 2001, there are perhaps rare but linguistically justified cases where it is in principle impossible to tell which of a number of interpretations is the right one, as in (1) below, taken from Przepiórkowski 2004.

(1) Pamiętam ją pijaną.
remember.1ST her.ACC drunk.ACC/INS
'I remember her drunk.'

(2) a. Pamiętam go pijanego.
remember.1ST him.ACC drunk.ACC
'I remember him drunk.'

b. Pamiętam go pijanym.
remember.1ST him.ACC drunk.INS

As shown in (2), the predicative adjective may occur either in the accusative or in the instrumental case in this construction. However, the feminine form *pijaną* in (1) is syncretic between these two grammatical cases, so both interpretations are fully correct and no context or world knowledge can provide any disambiguation clues here.

This reasoning determined the design of the IPI PAN Corpus of Polish (`http://korpus.pl/`; Przepiórkowski 2004), as well as accompanying annotation and search tools. In particular, a manually annotated subcorpus of the IPI PAN Corpus has numerous instances of multi-tagged tokens, and two taggers trained on that subcorpus, presented in § 3, occasionally disambiguate to more than one morphosyntactic interpretation.

2. Evaluation Measures

Let us assume that the segment *pijaną* may have three[1] different morphosyntactic interpretations: ppas (passive participle), adj:acc (adjective in the accusative) and adj:inst (adjective in the instrumental), and that in the gold standard (*GS*) corpus both adjectival interpretations are marked as correct.

Let us consider five different taggers, *T*1 to *T*5, making different decisions for the segment *pijaną* in (1), as in Table 1. It is clear that the tagger *T*1 fares best for this segment, as its decision is identical to the annotation in the gold standard, and *T*5 clearly performs worst here. But what is the relative value of the other three taggers?

word	**tag**	*GS*	*T*1	*T*2	*T*3	*T*4	*T*5
pijaną	ppas	0	0	0	1	1	1
	adj:acc	1	1	0	1	1	0
	adj:inst	1	1	1	1	0	0

Table 1: Different decisions of five hypothetical taggers

2.1. Accuracy-Like Measures

We will consider three interpretations of the general notion of *accuracy*:

- **correctness**: word-level accuracy understood as the percent of words for which the tagger and the gold standard agree completely;
- **weak correctness**: the percent of words for which the set of tags selected as correct by the tagger and the set of tags marked as correct in the gold standard overlap;
- **tag-level accuracy**: the percent of tags for which the tagger and the gold standard agree.

The values of these three measures for the single segment *pijaną* annotated as in Table 1 are given in Table 2.

[1] These three interpretations are an illustrative simplification of the real set of interpretations for *pijaną*; in particular, passive participles in Polish bear case, among other morphosyntactic categories.

measure	$T1$	$T2$	$T3$	$T4$	$T5$
correctness	1	0	0	0	0
weak correctness	1	1	1	1	0
tag-level accuracy	1	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	0

Table 2: Evaluation of the hypothetical taggers according to the three *accuracy*-like measures

The **tag-level accuracy** seems to be the most useful of the three measures in the tiny artificial example above, but it does not behave gracefully in case of realistic corpora of Polish, where it is not uncommon for a segment to have a dozen or two dozen interpretations, only one of which is correct. In such a case, a tagger that selects a wrong interpretation still has a very high **tag-level accuracy**, as it agrees with the gold standard on all the interpretations rejected by both. For example, in case of a segment with 10 interpretations, such an erroneous tagger would score 0.8 **tag-level accuracy**, but — more reasonably — only 0.0 **correctness** and **weak correctness**. For this reason we do not consider **tag-level accuracy** in the remainder of this paper.

Obviously, the other two measures are also not unproblematic: according to their scores, $T4$ is as good for *pijaną* as either $T2$ or $T3$, contrary to the fact that 1) both $T2$ and $T4$ got one of the two correct tags right, but for the incorrect interpretation ppas, only the former reached the right decision and, similarly, 2) the interpretation on which $T3$ and $T4$ differ is marked correctly by the former and not by the latter. Note also that **weak correctness** alone is not a sufficient measure of tagger quality: it is trivial to achieve the maximal value of this measure (by selecting all interpretations as correct). Nevertheless, as these measures have been used in the past, we will retain them also in the current evaluation exercise.

2.2. Information Retrieval Measures

It has also been proposed (van Halteren 1999) that standard Information Retrieval (IR) measures of **precision** and **recall**, as well as their harmonic mean, called **F-measure**, could be used for the evaluation of such partially disambiguating taggers:

- **precision**: the percent of tags selected by the tagger which are correct according to the gold standard;

- **recall**: the percent of correct tags that are selected by the tagger.

The values of these measures for the hypothetical taggers and the artificial 1-segment corpus *pijaną* are given in Table 3. Note that the **F-measure** reasonably evaluates $T2$ and $T3$ as better than $T4$, but it is not clear that the relative strengths between $T2$ and $T3$ are correctly reflected by this measure: which of these two taggers should be preferred is a matter of the relative importance of precision and recall for the further uses of the tagged corpus.

measure	$T1$	$T2$	$T3$	$T4$	$T5$
precision	1	1	$\frac{2}{3}$	$\frac{1}{2}$	0
recall	1	$\frac{1}{2}$	1	$\frac{1}{2}$	0
F-measure	1	$\frac{2}{3}$	$\frac{4}{5}$	$\frac{1}{2}$	0

Table 3: Evaluation of the hypothetical taggers according to the IR measures

3. Polish Taggers

There are two Polish taggers which were trained on the manually disambiguated corpus — of over 880 000 segments — created within the IPI PAN Corpus of Polish project (`http://korpus.pl/`).[2]

3.1. Dębowski's Tagger

The first of these taggers, probably the first morphosyntactic tagger of Polish[3], is a nameless tagger developed by Łukasz Dębowski at the Institute of Computer Science, Polish Academy of Sciences (ICS PAS; the Polish acronym is IPI PAN, hence the name of the IPI PAN Corpus), described in Dębowski 2004. It is a classical trigram-based statistical tagger, whose error rate was reported as 9.4%, corresponding to the **weak correctness** of

[2] Another Polish tagger, which — to the best of our knowledge — has not seen the light of day, is reported in Rudolf 2004. Recently, one more tagger has been constructed by two students of the second author at the Institute of Informatics, University of Warsaw; cf. Acedański & Gołuchowski 2009.

[3] An earlier attempt, following a different methodology, is described in Dębowski 2003.

90.6%, although the possibility of multiple correct tags is not discussed explicitly in Dębowski 2004, so it is not fully clear which notion of word-level accuracy was assumed here.

For practical reasons the tagger has never been made publicly available: it is inefficient, it consists of a rather large set of Perl scripts, and it is not easy to install it in a new environment or to maintain it. Nevertheless, this tagger was used for the morphosyntactic annotation of the first version of the 250-million-segment IPI PAN Corpus of Polish.

In the rest of the paper, we will refer to Dębowski's tagger as DT.

3.2. TaKIPI

The second edition of that corpus was tagged with TaKIPI (Piasecki 2007), a tagger originally developed by Maciej Piasecki and his colleagues within a project carried out at ICS PAS, subsequently further developed and maintained at the Wrocław University of Technology.[4] It is a hybrid system based on decision trees, unigram frequencies and manually constructed correction rules, it is rather efficient, partially due to its implementation in C, and it is publicly available as open source under the GNU General Public License.[5]

The **weak correctness** of TaKIPI reported in Piasecki & Godlewski 2006 is 92.55%, but — as the article hastens to add – *[c]omparison with 90.4% reported by [(Dębowski 2004)] is difficult, as his tagger always leaves the best one tag, and was tested on a very small part of [the manually disambiguated corpus]*. For this reason — and also in order to set the stage for the evaluation and comparison of future taggers of Polish, including the tagger to be trained within the National Corpus of Polish project (`http://nkjp.pl/`; Przepiórkowski et al. 2008, 2009) — an independent comparative evaluation of both taggers is needed, using a variety of measures introduced in the previous section. The results of such an evaluation are presented in § 4 below.

3.3. Common Tagset and Gold Standard

There are many factors which may influence the fairness of a comparison or even make the whole exercise futile. As noted in van Halteren 1999, the size of the tagset and,

[4] Currently the tagger is co-owned by the two institutions.

[5] It can be downloaded from `http://nlp.ipipan.waw.pl/TaKIPI/` and `http://www.plwordnet.pwr.wroc.pl/g419/tagger/`.

especially, the average number of interpretations per segment correlate with the difficulty of the tagging task. In case at hand, the gold standard and both taggers in principle assume the same tagset, the IPI PAN Tagset described, e.g., in Przepiórkowski & Woliński 2003a,b, with 4179 potentially possible tags, out of which up to 1642 are used in practice (Przepiórkowski 2006) and with 3.32 tags per segment (Przepiórkowski 2008, p. 44, fn. 32). However, somewhat surprisingly, it turned out that there were subtle differences between the three versions of the IPI PAN Tagset.

Each tag in this tagset has the form kl:kat$_1$:kat$_2$:...:kat$_n$, where kl is the name of a grammatical class, while each of kat$_i$ is the value of a grammatical category appropriate for that class, e.g., adj:sg:acc:f:pos for the singular accusative feminine positive degree form *pijaną* of the adjective PIJANA 'drunk'. The formal specification of the IPI PAN Tagset makes some categories for some classes optional, for example, in the case of the neutralisation of the rather ephemeral binary-valued categories of accentability and post-propositionally for some 3rd person pronouns. The gold standard and DT sometimes took advantage of this possibility to omit values of these two categories, while TaKIPI always produces maximally specific tags.

In the current experiments, in order to make the gold standard and tagger results uniform, the abbreviated tags in the gold standard and in the results of DT were expanded by adding all possible values of both binary categories; in case the abbreviated tag was marked as correct (in gold standard) or selected (by the tagger), all expended tags were so marked, and analogously for the opposite marking.

Another difference between the three tagset versions stemmed from the optionality of another ephemeral binary-valued category, accommodability, pertaining to Polish numerals. Both taggers treat this feature as if it were obligatory, while within the gold standard it only appears with human-masculine (m1) numerals in the nominative. In this case, because of the peculiarities of the category of accommodability (see the original presentations of the IPI PAN Tagset and references therein), tags were uniformly contracted, by removing the value of accommodability (and retaining the correct / selected status of the tag).

4. Evaluation Results

Both taggers were evaluated using the same version of the underlying morphological analyser, Morfeusz SIAT (Woliński 2006), namely the version for which TaKIPI was optimised, thus giving this tagger a certain competitive edge. TaKIPI has an optional module, a morphological guesser, used for words unknown to Morfeusz, but since there is no similar guesser in DT, we used TaKIPI with the guesser switched off. The evaluation was performed for full morphosyntactic tags, as well as for grammatical classes (roughly, parts of speech; POS) only. In the latter case, all interpretations sharing the same POS were contracted to one tag, marked as correct / selected exactly in cases when at least one of the original interpretations was correct / selected.

Table 4 presents the results of the evaluation. C and WC stand for **correctness** and **weak correctness**, and P, R and F — for the IR metrics of **precision**, **recall** and **F-measure**.

tagger	C	WC	P	R	F
DT	87.39%	90.59%	84.51%	83.09%	83.80
TaKIPI	86.63%	91.30%	88.83%	83.47%	86.06
DT (POS only)	96.79%	97.11%	96.75%	96.78%	96.77
TaKIPI (POS only)	94.58%	96.81%	94.67%	96.55%	95.60

Table 4: Summary of evaluation results

A number of observations may be made about these results. First, note that the **weak correctness** result of DT is slightly higher than reported in Dębowski 2004 (90.59% vs. 90.4%), and the result of TaKIPI is a little lower than reported in Piasecki & Godlewski 2006 (91.30% vs. 92.55%), thus making the two taggers more similar with respect to this measure than originally assumed. Nevertheless, the WC quality of TaKIPI is still significantly better than that of DT.

Second, in case of the strong **correctness**, DT fares better than TaKIPI. We assume that one of the reasons for that is that TaKIPI does not disambiguate between nouns and gerunds, so in case of — relatively frequent — ambiguities of this kind, TaKIPI never produces exactly the same decisions as the gold standard, which specifies either the noun interpretation or the gerund interpretation as correct. This is probably also one of the reasons why, when the tagset is constrained to grammatical classes, TaKIPI produces consistently lower results than DT, for all measures considered here; since noun and gerund are two different grammatical classes, such behaviour of TaKIPI has detrimental effect on **correctness** and **precision**.

Finally, let us note that for full morphosyntactic tagging, the evaluation results as measured by the Information Retrieval metrics are consistently higher for TaKIPI than for DT.

5. Conclusion

Contrary to earlier preliminary results, the quality of both taggers, when trained on the same corpus and using exactly the same tagset, turns out to be similar. One or the other tagger achieves better results depending on the evaluation metrics and the granularity of the tagset. For fine-grained morphosyntactic tagging TaKIPI is clearly better than DT as measured with IR metrics, and a little better in terms of weak correctness, but DT is a little better than TaKIPI in terms of (strong) correctness. For coarse-grained POS tagging, DT is uniformly better than TaKIPI; in case of correctness and precision (and, hence, also F-measure) — clearly so.

So which of the two taggers should be recommended for the morphosyntactic tagging of Polish? While the evaluation results given above do not provide an unqualified answer, practical considerations do: unlike DT, TaKIPI is not only open source and publicly available, but also reasonably efficient, so currently there is no alternative to using this tagger. But as other taggers are being developed, and existing taggers are trained on new manually annotated corpus being developed within the National Corpus of Polish, the need for the comparative evaluation of taggers will grow. We hope that the discussion in the current paper will facilitate such comparison in the future.

Bibliography

Acedański, S. & K. Gołuchowski. (2009) ‘A morphosyntactic rule-based Brill tagger for Polish’. In: Kłopotek, M. A., Przepiórkowski, A., Wierzchoń, S. T. & K. Trojanowski (Eds.), *Advances in intelligent information systems — design and applications*. Warsaw: Akademicka Oficyna Wydawnicza EXIT, 67–76.

Dębowski, Ł. (2003) ‘A reconfigurable stochastic tagger for languages with complex tag structure’. In: *Proceedings of* Morphological Processing of Slavic Languages*, EACL 2003*.

———. (2004) ‘Trigram morphosyntactic tagger for Polish’. In: Kłopotek, M. A., Wierzchoń, S. T. & K. Trojanowski (Eds.), *Intelligent Information Processing and Web Mining*. Berlin: Springer-Verlag, 409–413.

Jurafsky, D. & J. H. Martin. (2009) *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Pearson Education, Inc., 2nd ed.

Karwańska, D. (2009) ‘Porównanie tagerów dopuszczających niejednoznaczności’. Unpublished manuscript, National Corpus of Polish.

Manning, C. D. & H. Schütze. (1999) *Foundations of statistical natural language processing*. Cambridge, MA: The MIT Press.

Oliva, K. (2001) ‘On retaining ambiguity in disambiguated corpora: Programmatic reflections on why’s and how’s’. TAL (Traitement Automatique des Langues) 42(2): 487–500.

Piasecki, M. (2007) ‘Polish tagger TaKIPI: Rule based construction and optimisation’. Task Quarterly 11(1–2): 151–167.

Piasecki, M. & G. Godlewski. (2006) ‘Effective architecture of the Polish tagger’. In: Sojka, P., Kopeček, I. & K. Pala (Eds.), *Text, speech and dialogue: 9th international conference, TSD 2006, Brno, Czech Republic, September 2006*. Berlin: Springer-Verlag, 213–220. (Lecture Notes in Artificial Intelligence, 4188.)

Przepiórkowski, A. (2004) *The IPI PAN corpus: Preliminary version*. Warsaw: Institute of Computer Science, Polish Academy of Sciences.

———. (2006) ‘The potential of the IPI PAN Corpus’. Poznań studies in contemporary linguistics 41: 31–48.

———. (2008) *Powierzchniowe przetwarzanie języka polskiego*. Warsaw: Akademicka Oficyna Wydawnicza EXIT.

———. (2009) ‘A comparison of two morphosyntactic tagsets of Polish’. In: Koseska-Toszewa, V., Dimitrova, L. & R. Roszko (Eds.), *Representing semantics in digital lexicography: Proceedings of MONDILEX fourth open workshop*. Warsaw, 138–144.

Przepiórkowski, A., Górski, R. L., Lewandowska-Tomaszczyk, B. & M. Łaziński. (2008) ‘Towards the National Corpus of Polish’. In: *Proceedings of the sixth international Conference on Language Resources and Evaluation, LREC 2008*. Marrakech.

Przepiórkowski, A., Górski, R. L., Łaziński, M. & P. Pęzik. (2009) ‘Recent developments in the National Corpus of Polish’. In: Levická, J. & R. Garabík (Eds.), *Proceedings of Slovko 2009: Fifth international conference on NLP, corpus linguistics, corpus based grammar research, 25–27 November 2009, Smolenice/Bratislava, Slovakia*. Brno: Tribun.

Przepiórkowski, A. & M. Woliński. (2003a) ‘A flexemic tagset for Polish’. In: *Proceedings of Morphological processing of Slavic languages, EACL 2003*. Budapest, 33–40.

———. (2003b) ‘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*, 109–116.

Rudolf, M. (2004) *Metody automatycznej analizy korpusu tekstów polskich*. Warsaw: Uniwersytet Warszawski, Wydział Polonistyki.

van Halteren, H. (1999) ‘Performance of taggers’. In: van Halteren, H. (Ed.), *Syntactic wordclass tagging*, volume 9 of *Text, Speech and Language Technology*. Dordrecht: Kluwer, 81–94.

Woliński, M. (2006) ‘Morfeusz — a practical tool for the morphological analysis of Polish’. In: Kłopotek, M. A., Wierzchoń, S. T. & K. Trojanowski (Eds.), *Intelligent information processing and web mining*. Berlin: Springer-Verlag, 511–520.