Developing free morphological data for Polish

Adam Radziszewski

Institute of Informatics, Wrocław University of Technology, Wrocław

Marek Maziarz

Institute of Informatics, Wrocław University of Technology, Wrocław

April 4, 2011

Abstract

A limiting factor in construction of Natural Language Processing (NLP) systems is often the availability of morphological resources. This indeed happens for Polish: the freely available corpus with manual morpho-syntactic annotation (part of the IPI PAN Corpus) is not coupled with any free morphological analyser. There exists a very large morphological dictionary of Polish available under a free licence — Morfologik. Unfortunately, its tagset differs significantly from the tagset of the corpus and, what is more, its morphological description lacks desired rigour. We amend this situation by performing a massive conversion of the dictionary into the tagset compliant with the corpus. The conversion results in a free dictionary containing entries for almost 3.5 million different word forms. In this article we report on our methodology, discuss some morphological and syntactic issues related to both tagsets and present the characteristics of the resulting dictionary.

1 Background

Computer processing of text on the level of morpho-syntax requires several language resources. These usually include a manually annotated corpus and a morphological analyser, both operating on the same tagset. This is especially true for NLP tasks that employ statistical learning from annotated corpora: the corpus is rarely large enough to obtain a reliable lexical model that would account for less frequent word forms. This problem can be partially solved by providing an external, large coverage morphological dictionary. Such dictionaries are used to assign lemmas and morpho-syntactic tags to word forms, including those not present in the corpus (cf. training and usage of morpho-syntactic taggers for Polish, e.g. Piasecki (2007)).

Thus, the availability of such an *analyser–corpus* pair influences the possibility of creating NLP systems. Quite often, licensing of such resources becomes the limiting factor. First, the licence of the employed resources directly affects possible usages of the system (and its output). Second, close-sourced analysers disallow to study, extend or correct the contents of their dictionary, which seems crucial from research perspective. Thus, licensing affects not only business, but also science. The two are thought to be strongly interdependent in the area of NLP. We address the problem of lacking free resources for Polish by converting the data of a free morphological analyser into the tagset compliant with a freely available corpus, being the tagset of the IPI PAN Corpus (Przepiórkowski, 2004).

To the best of our knowledge, there exists only one freely available corpus of Polish with manual morpho-syntactic annotation, that is the new edition of the *Corpus of the Frequency dictionary of contemporary Polish* (Ogrodniczuk, 2003). The corpus contains Polish language of the 1960s and comes in two versions:

1. the original[1] version annotated with the help of the SAM-95 analyser (Bień and Szafran, 2001),
2. the re-annotated version, being a part of the *IPI PAN Corpus* project, following the tagset of the corpus.

Both versions of the corpus are available under the terms of a free licence, namely GNU GPL[2].

The aforementioned SAM-95 analyser is not free software — it is restricted to research-only purposes[3] (Bień and Szafran, 2001). The other version is annotated with the tagset of the IPI PAN corpus (henceforth the *IPIC tagset*). Unfortunately, the only available morphological analyser compliant with the tagset — Morfeusz (Woliński, 2006) — is also not free software. In fact, its licence is quite restrictive, allowing only non-commercial usage and explicitly forbidding any attempts at extracting the underlying morphological data (technically, the analyser is provided as a dynamic library with no separation between the code and the data).

Several morphological analysers have been created for Polish (Hajnicz and Kupść, 2001). To the best of our knowledge, only one of them has been released under a free licence (as of March 2011): *Morfologik*[4], a part of the *LanguageTool* open source proof-reading tool (Miłkowski, 2010). The dictionary of the analyser is quite sizeable — Morfologik 1.6 contains entries for nearly 3.5 million forms. This data is available under a dual licence: GNU LGPL or Creative Commons Share Alike (the user is free to choose; both of them are very permissive, not limiting the usage to any particular scenario). The dictionary is stored in a simple three-column text format, containing forms, lemmas and tags. Although the tagset of Morfologik resembles the IPIC tagset, there are significant differences, both in the structure of the tagset and in the rigour of description.

The corpus itself could be used as a source of morphological data. We have tried this option but the results are disappointing: the extracted morphological dictionary contains entries for less than 84 000 different word forms. This means that a morpho-syntactic tagger (or another piece of NLP

[1] This is not entirely true: the very original corpus was compiled in the years 1963–1967 and later on frequency lists based on the corpus were published (Ogrodniczuk, 2003). We refer to the first digitised version as the *original*.

[2] Actually this licence brings about an unfortunate situation: it is not clear what constitutes the *source code* and how it affects various types of language models acquired from the corpus (e.g. frequency lists or a trained tagger). Nevertheless, it is genuinely free: it guarantees that the data will remain free and, thus, may be used as a part of an open-source NLP system.

[3] This is an issue even for research projects, since it is often hard to distinguish "pure research" from "commercial" applications.

[4] http://morfologik.blogspot.com

software) equipped with this data only is likely to perform much worse that the same tagger having access to a dictionary containing entries for over three million forms. This was our motivation to bridge the gap between the free manually annotated corpus and the free morphological data. We considered two possibilities: 1) converting the corpus annotation to the tagset of Morfologik and 2) converting Morfologik to the tagset of the IPI PAN corpus. We decided on the second option. First, converting (even this large) a dictionary requires less workload than converting a reference corpus and ensuring its high quality (i.e. that contextually appropriate tags are selected). Second, we prefer to invest our labour in a general language resource, which a morphological dictionary certainly is, rather than in a corpus representing the Polish language of the 60's. The last consideration was the choice between the tagsets: while the tagset of the corpus is well-documented and designed with clear principles in mind, Morfologik's tagset is somewhat inaccurate and implicit.

In the rest of this article we briefly compare the two tagsets, introduce our auxiliary tagset, present our methodology and try to assess the quality of the converted dictionary. We hope that this paper will be useful as a companion to the data set we release under the original Morfologik licences: GNU LGPL or Creative Commons Share Alike.

2 Three tagsets

In this section we summarise both tagsets and present our *intermediate tagset* which is used to facilitate the conversion.

2.1 IPIC tagset

The IPIC tagset has been designed with extraordinary rigour. The basic assumption is that the *grammatical classes* (generalisation of the usual part-of-speech notion) are distinguished primarily on the grounds of inflection (Przepiórkowski and Woliński, 2003). In consequence, there are over 30 grammatical classes. Each class is specified for a set of *attributes* (grammatical categories) whose values must be provided. For instance nouns are specified for number, gender and case, adverbs are specified for degree. Because of these assumptions, some grammatical classes depart from traditional parts-of-speech; an utmost example is the class called *particle-adverb*, which consists of indeclinable forms that did not fit in other classes — including some non-gradable adverbs, particles and the reflexive pronoun *się*.

An important exception to this rule is the consent to *optional attributes* whose values may be omitted. For instance, some lexemes belonging to the class of prepositions are specified for *vocalicity*. This attribute has two values, denoting if the form undergoes a phonological alternation manifested as vowel addition. Some prepositions have two variants, e.g. *przed* and *przede*, while the others do not (e.g. *na*). The latter are not given any value of this attribute.

The IPIC tagset is *positional*, meaning that each tag consists of the grammatical class and a sequence of attribute values and for each grammatical class the order of its attributes is fixed. If any optional attributes are applicable, they must come after the required ones. Note, however, that this is merely a convenient assumption, since having distinct value mnemonics for all the attributes, one can easily infer the proper tag even if the order of values is altered.

Due to space limitations we do not quote the whole tagset specification. It can be found in Przepiórkowski and Woliński (2003) or Chap. 3. of Przepiórkowski (2004). Unfortunately, neither of these sources defines the optionality of attributes; this we have inferred from the corpus data and the configuration file published in Przepiórkowski (2008). The exact definition of the tagset as employed here (as well as the other tagset discussed in the next sections) can be found in the files referenced at the download site (see Section 4).

2.2 Tagset of Morfologik

The tagset of Morfologik, on the contrary, is hardly defined. The *Readme* file gives the mnemonics of the grammatical classes and the attribute values. Attributes as such are not explicitly enumerated, and, in some cases it is hard to infer which attribute some values belong to. Not all the actual classes are documented. The positionality is not respected and in the actual data the forms of the same grammatical class are likely to occur with quite a number of combinations of attributes whose values are specified. In one case the distinction between an attribute and a class is blurred (`refl` is declared as a value, although, technically, it occupies a class position in the dictionary).

Despite this inexact frame of Morfologik tagset, the actual tags closely resemble those of the IPIC tagset. This is intended and some additional remarks on the differences are given in the *Readme* file. To obtain a sketch of the *real* tagset (i.e. classes that describe the actual data) we developed a Python script that reads a morphological dictionary and outputs a list of *value usage patterns*. Each of the patterns is a subclass of one grammatical class that has a fixed number of values provided. A pattern is described by sets of values that appeared at subsequent positions (the script naively assumes that the tagset is positional). Here is a fragment of the output of the script run against data from Morfologik 1.7 RC2[5]:

```
adj/0:
adj/3:  pl.sg  acc.dat.gen.inst.loc.nom.voc  f.m.n
adj/4:  pl.sg  acc.dat.gen.inst.loc.nom.voc  f.m.m1.m2.m3.n  comp.pos.
   sup
adj/5:  pl.sg  acc.dat.gen.inst.loc.nom.voc  f.m.m1.m2.m3.n  pos  aff.
   neg

subst/1:  irreg
subst/2:  acc.gen  m1
subst/3:  pl.pltant.sg  acc.dat.gen.inst.loc.nom.voc  f.m.m1.m2.m3.n.n2
subst/4:  ger.pl.sg  acc.dat.gen.inst.loc.nom.pl.sg.voc  acc.dat.gen.
   inst.loc.m.m1.n.nom.voc  depr.m1.n.neg
```

The first line states that there are occurrences of the `adj` class with no attributes. The second one presents a pattern corresponding to three-value `adj` tags, whose first attribute can be recognised as grammatical number, the second attribute as case and third — gender.

A comparison of the grammatical classes appearing in both tagsets is presented in Table 1.

[5] This version of the data was obtained from Marcin Miłkowski on 29 October 2010. The analysis and conversion described in this paper has been carried out on this data. When the final 1.7 version was published, we updated the

IPIC	Morfologik	Name (IPIC)	Example form
adj	adj	adjective	*biały*
adja	missing	ad-adjectival adj.	*biało*(-czerwony)
adjp	adjp	post-prep. adj.	(po) *polsku*
adv	adv	adverb	*biało*
aglt	segmentation	agglut. *być*	(czytał)*em*
bedzie	verb:bedzie	future *być*	*będziesz*
conj	conj	conjunction	*lub*
depr	subst:depr	deprec. noun	*posły*
fin	verb:fin	non-past form	*czyta*
ger	subst:ger	gerund	*picie*
ign (= unknown)	ign (unreliable)	unknown	*xyz123*
imps	verb:imps	impersonal form	*czytano*
impt	verb:impt	imperative	*czytaj*
inf	verb:inf	infinitive	*czytać*
interp	punctuation	punctuation	.
	nstd (rare)		*domie*
num	num	numeral	*sześć*
numcol or num (rare)	num	collective num.	*sześcioro*
pcon, pant	pcont, pant	adv. participle	*pijąc, wypiwszy*
pact, ppas	pact, ppas	adj. participle	*pijący, pity*
ppron12, ppron3	ppron12, ppron3	personal pronoun	*ciebie, oni*
praet praet + aglt praet + by praet + by + aglt	verb:praet verb:praet verb:praet:pot verb:praet:pot	l-participle (past form) (conjunctive) (conjunctive)	*czytał* *czytałem* *czytałby* *czytałbym*
pred	pred	predicative	*widać*
prep	prep	preposition	*na*
qub	qub (unrealiable)	particle-adverb	*się, nawet*
not needed	refl		
siebie	siebie	pronoun *siebie*	*sobą*
subst	subst	noun	*mięso*
winien	winien	*winien*-like verb	*powinni*
xxs (rare)		nominal alien	*H2O*
xxx (rare)		other alien	*fantastisch*

Table 1: A comparison of IPIC and Morfologik grammatical classes. Classes marked unreliable represent Morfologik forms that are either hard to classify or should belong elsewhere than stated. Classes marked rare are rare and specific cases, whose omission is of little concern or in fact other classes are usually used to label these forms (e.g. num instead of numcol).

2.3 Segmentation issues and the intermediate tagset

One of the most important differences is related to segmentation strategies. While Morfologik follows a traditional assumption that word forms constitute separate tokens, the approach underlying the IPIC tagset is significantly different: some forms are split into several tokens. This is related to so-called *floating inflections*, treated as forms of the verb *być* (*to be*) (Woliński, 2006). For instance, *czytałem* (*I was reading*) is split into *czytał* (l-participle in singular masculine) and *em* (agglutinative *być* in first person, singular number). Similarly, *czytałbym* (the same form in conjunctive mood) is split into *czytał*, *by* (conjunctive particle) and *m*. Such a suffix may also be attached to non-verbal forms, most notably prepositions (e.g. *dlań*, being a shortened version of *dla niego*, *for him*). This splitting is linguistically well motivated — some of the suffixes are "mobile", i.e. they can appear in different places of the same sentence, being a separate word form or attached to another word, cf. (Grzegorczykowa et al., 1998, pp. 256, 262–263, 339) and (Przepiórkowski, 2004, Ch. 3).

This segmentation strategy is, however, quite atypical and practically inconvenient — virtually all tagging algorithms assume that input is a *sequence* of tokens, not a fancy graph structure. As the number of verb forms subjected to this phenomenon is quite large while the other combinations (not involving a verb) are rare, we decided to deal systematically with verbs only.

Our solution is to employ an *intermediate tagset*, being a slight and conservative modification of the original IPIC tagset with altered segmentation rules. This way, we could almost directly transfer a large number of verb forms from Morfologik dictionary, laying the burden of re-tokenisation and re-tagging on tagset conversion routine. We wrote two such routines, being configurations for the Maca system (Radziszewski and Śniatowski, 2011), implementing to-and-fro conversion. The routines have been tested extensively, including tests on the whole *Corpus of the Frequency dictionary of contemporary Polish*. We believe that, besides facilitating morphological dictionary conversion, the intermediate tagset may be useful as such for other practical purposes, e.g. it may diminish the segmentation ambiguity occurring in the IPIC tagset (at least in cases where a noun reading is ambiguous with *verb*+suffix reading, e.g. *miałem*).

The intermediate tagset, called *IKIPI*, departs from the IPIC in four aspects:

1. We extend the original `fin` class, which in IPIC denotes non-past verb forms. The new `fin` class contains both simple past and non-past forms (we keep the class name as it still meaningfully refers to *finite* verbs). The preterite forms result from joining l-participles directly followed (i.e. no spaces in-between) by agglutinative suffixes. Gender is an optional attribute, since it is morphologically manifested in preterite forms only (in this case the value of gender is taken directly from the corresponding `praet` form; Polish non-past finite verbs are morphologically not marked for gender).

2. New class is introduced for conjunctive forms: `conjt`. These forms result from joining two tokens, l-participle and *by* particle, or these two directly followed by an agglutinative suffix. In the former case, the form is marked as third person, while in the latter the value of person is taken from the suffix. Besides person, the class requires number, gender and aspect.

resulting dictionary to account for changes between these versions.

3. The new fin class is explicitly specified for tense. The forms resulting from joining are marked as prt (preterite). The tense of the forms coming from the original fin class is inferred from aspect: perfective forms are marked fut (future), while imperfective aspect entails prs (present tense). To keep it consistent, all the forms of the bedzie class (future form of *być*) are marked fut. Note that prt is a value of the introduced *tense* attribute (preterite); this should not be confused with praet, the original IPIC grammatical class (l-participle).

4. Similar joining is performed for forms of the winien class directly followed by the agglutinative suffixes. The class is not changed but only the joined forms are specified for person.

This procedure does not cover all the cases with preterite and conjunctive verbs. When the agglutinative suffix is not physically attached to the verb, the verb may still belong to a preterite or conjunctive construct — with the affix placed somewhere else (e.g. as in *bym czytał*, having the same meaning as *czytałbym — I would read*). Alternatively, the same verb form itself may be interpreted as plain third person preterite. For such forms we leave the original praet class.

We want to stress that these modifications are not meant to "fix" the original tagset. They are merely a practical means of avoiding several difficulties related to atypical segmentation strategy assumed in the IPIC tagset. A change in segmentation strategy calls for modifications in tagset, as different forms are to be accounted for. Although we tried not to violate the basic assumptions behind the original tagset, we had to sacrifice some of its simplicity. We hope that the resulting tagset and conversion routines will be helpful for other future applications where having word forms as single tokens is more important than perfectly accurate description. Besides, it allows to get slightly more traditional morpho-syntactic description of verbs, which may be an advantage in some cases.

3 Conversion procedure

In this section we present our conversion procedure. We enumerate the general problems and then specific cases.

3.1 Methodology

We have divided the data into parts — each part corresponding to one usage pattern as generated by our script (see Section 2.2). The result was 40 groups, making up all the 20 grammatical classes appearing in the Morfologik dictionary. Each part was analysed separately, although in some cases it turned out convenient to join some of them and then convert as a whole.

It was tempting to use the corpus as a source of the data — we could hope to get reasonable coverage of closed-class forms, e.g. prepositions. Unfortunately, we had to refrain from doing so as it would enforce licensing the data under GPL. In some difficult cases we did consult the corpus but to get a picture of grammatical class–attribute usage rather than to extract particular tags. Using Morfeusz was absolutely out of question due to its restrictive licence.

We relied on the papers describing the tagset, i.e. Przepiórkowski and Woliński (2003), Przepiórkowski (2004) and Przepiórkowski (2003). In some cases, we had to resort to linguistic sources: for the interpretation of grammatical genders and accommodation rules we consulted Saloni and Świdziński (1998), while the inflectional paradigms of numerals and personal pronouns were based on Grzegorczykowa et al. (1998). We relied on the prescriptive dictionary of Polish (Markowski, 2006) to define interpretations of incorrect language constructs.

The whole described conversion procedure has been carried out with the help of basic GNU/Linux command-line utils (`grep`, `sed` and the `vim` editor), some simple Python scripts and the Maca system to validate the obtained tags (Radziszewski and Śniatowski, 2011).

In the rest of this section we present particular problems and how we solved them. Some issues that are left unsolved are explicitly stated.

3.2 General problems

A technical difficulty was posed by duplicated tags in Morfologik dictionary. As the dictionary format allows for compact representations of multiple tags, these repetitions tend to be implicit, i.e. cannot be filtered by removing duplicated lines. To account for this problem, we wrote a Python script that is able to decompose compact tag representations, remove duplicates and write the dictionary back in the compact format. The script is generally useful as a tool for cleaning morphological dictionaries before compiling into transducers and has been included in the Maca system.

A general problem is that the dictionary contains some erroneous entries. During the conversion process we examined manually some small random samples of each subclass. Where the errors were too frequent, we rejected the whole subclass as unreliable (it happened to two classes: `ign` and `qub`; the latter was later re-examined and some forms recovered). Anyway, we have not assessed the scale of the problem, so the resulting dictionary may still contain a significant number of errors.

The most serious tagset-related issue was the presence of the underspecified masculine gender (`m`). The IPIC tagset contains three masculine genders (`m1`, `m2`, `m3`) motivated inflectionally. The `m` value appears across the whole dictionary, with different grammatical classes. Besides the underspecified `m`, the proper, specified masculine genders are also common. A fully reliable solution would be to inspect all these `m` values manually. Unfortunately, such tags are too frequent to perform the corrections within reasonable workload (there are at least 750 000 entries with `m` gender). We decided on a half-measure: we substituted this underspecified gender with all three values. Rudimentary inspections suggest that the underspecified masculine gender indeed occurs in places where all three values should be placed, but again, without systematic analysis this is only hypothetical. A couple of entries were marked as `m13` where `m1.m3` (notation shorthand for both variants) was probably intended.

There is some similar confusion regarding the neuter gender (although less problematic from our point of view): some forms are marked as specific `n1` or `n2` gender. Note that this distinction is documented in some of the papers on the IPIC tagset, e.g. Przepiórkowski (2003), but somehow it has not made its way into the final corpus, cf. Przepiórkowski (2004). We treat both neuter variants as plain neuter gender (`n`).

Some nouns bear the `pltant` value (*plurale tantum*); the IPIC tagset requires such forms to be plain plural (`pl`).

3.3 Particular grammatical classes

The IPIC tagset distinguishes verb types as different grammatical classes, while Morfologik puts them under generic `verb` class. Fortunately the division into classes is mirrored in attribute values. An analogous situation holds for the distinction between regular nouns, depreciative noun forms and gerunds.

The vast majority of nouns could be converted without much trouble. An exception is a subset of forms marked `subst:irreg` (almost 4000), most of which are perfectly regular nouns and should be described with respect to number, gender and case. We do not convert these forms. Unfortunately, some actual gerunds are misclassified as regular nouns; this would require extensive manual inspection and has been postponed (we leave these forms as they are).

A common problem related to verbs is the lack of aspect value for many subclasses. Fortunately, aspect is given for infinitives; this gave us the opportunity to restore the aspect for other verb forms using verb lemmas (which are infinitives). Single forms that left uncovered had to be edited manually.

Conversion of forms marked as `verb:praet` was influenced by the design of the intermediate tagset. Forms labelled `verb:praet:...:ter` had to be restored to `praet` (l-participles, not specified for person). The remaining `verb:praet` forms were converted to the new `fin` class, specified for person and the preterite tense. Similarly, `verb:praet:pot` (conjunctive) was converted to the new `conjt` class. There were only 24 forms labelled as `verb:winien`, so we could inspect them manually; the suffixed forms were left marked for person while for the rest (marked as third person) we cleared the value of person (cf. Section 2.3).

Adjectives were converted easily, with the exception of zero-attribute `adj` subclass (only 7 forms).

The class of adverbs (`adv`) and particle-adverbs (`qub`) had to be converted with special caution. First, according to the IPIC principles, zero-attribute adverbs should belong to the `qub` class. Some of these forms, however, are in fact comparative adverbs and had to be moved out of this subclass. Second, some forms marked as negated (we disregard negation in adverbs and adjectives) had incorrect lemmas assigned — we took forms as lemmas. As noted in Table 1, the forms marked `qub` in Morfologik are generally unreliable: they include lots of mistagged forms that are mostly nouns. Nevertheless, as this class is quite frequent in the corpus, we couldn't ignore the data. We employed a simple heuristics to filter out most of these nouns: we got rid of entries having different form than lemma. The vast majority of the resulting data were legitimate particle-adverbs, including non-gradable adverbs, particles and interjections. After manual removal of forms that should belong to other classes, there are 200 forms tagged as `qub`. We added one form that was removed by the heuristics, namely the reflective pronoun `się` (in Morfologik it appears with lemma *siebie*) — this is an extremely frequent Polish form.

Personal pronouns occur in Morfologik in a multitude of variants, each with different set of attributes. As these forms are not numerous but, arguably, important, we decided to take forms from Morfologik and re-analyse them manually with the help of (Grzegorczykowa et al., 1998, p. 336–339) and (Saloni and Świdziński, 1998, p. 175–177, 182).

We divided the class of numerals into collective and cardinal numerals. We revised lemmas and accommodability of collective numerals according to the principles described in (Saloni and Świdziński, 1998, p. 195–209). Some particular word forms were compared to Markowski (2006). The accommodability of cardinal numerals is currently left unchanged, although some corrections may be desired.

3.4 Problems left unsolved

As noted in Section 2.3, we leave out indeclinable forms glued with agglutinative suffixes. Some of these forms are marked as regular prepositions and conjunctions in Morfologik. We isolated such cases for treatment in future versions of the analyser.

Due to the described segmentation strategy, the forms resulting from joining indeclinable forms with agglutinative suffixes are not handled. We gathered a small list of such forms for future considerations.

Ad-adjectival adjectives (`adja`) are generally missing from Morfologik dictionary. This class reflects another IPIC-specific segmentation strategy, namely splitting of forms like *polsko-niemiecki* (*Polish-German*). Such forms are analysed as three tokens: the ad-adjectival form (indeclinable), hyphen and a regular adjective. We respect this segmentation strategy (it is actually simpler to split on hyphens by default), but give no `adja` forms in the dictionary. This can be fixed in future releases — such forms could be semi-automatically acquired from unannotated corpora.

4 Working analyser and its evaluation

We used a script to ensure that there are no duplicated tags. We produced two versions: one with all the forms and lemmas converted to lower-case and the other one kept intact for future usage. It is not trivial to use the letter case in a systematic manner, thus we use the lower-case version here.

We have compiled the lower-case dictionary into a transducer (by using the provided script) and created a Maca configuration for the analyser. The configuration refers to segmentation rules for IKIPI, assigns `interp` tags to all tokens recognised as punctuation during tokenisation and feeds the rest of tokens to the transducer (in case-insensitive mode). For convenience, we have also prepared another configuration that automatically converts the output to the IPIC tagset.

The resulting lower-case dictionary contains entries for 3 432 509 different forms and 216 986 different lemmas. To evaluate the data coverage, we employed the *Corpus of the frequency dictionary* in the IPIC tagset[6]. The corpus contains 661 839 tokens. To measure the coverage, we employed the following procedure:

1. We extracted plain text from the corpus (respecting *no-space* markers and turning sentence boundaries into double newlines).

[6] `http://korpus.pl/index.php?page=download`

2. The text was re-analysed with our analyser configuration.

3. The output corpus contained differences in tokenisation (most of them were actually created by dubious entries, mostly “DELETED_TOKEN” and similar, which actually should have been removed from a reference corpus). We employed a Python script that substituted each portion of the mistokenised input with the correct tokenisation but tags reverted to ign (unknown) and lemmas to the orthographic form. This was to treat mistokenised fragments as our failure. Actually it is a severe penalty for different tokenisation (especially given the nature of most differences), nevertheless, it was hard to compare the morphological tags in different setting.

4. We employed a tagger testing script to compare how many tokens have tags and lemmas intersecting with the reference corpus.

The achieved accuracy (understood as the number of tokens where our output intersects with the reference corpus) of tag assignment equals 88.80%, while the accuracy of lemma assignment is 95.16%.

The “source” versions of the dictionaries (i.e. text files in tab-separated format) as well as the compiled transducer and both configurations are available at the Maca download site[7].

5 Summary

We have presented our procedure of converting the morphological dictionary of Morfologik into the IPIC tagset and producing a ready-made analyser. This is the first free and open-source morphological analyser for Polish that outputs in the IPIC tagset. We believe this is a valuable language resource: its tagset has been formalised and is now compliant with the *Corpus of the frequency dictionary*, which is also freely available. What is more, the tagset seems a *de facto* standard in Polish NLP community. This can facilitate creation of language processing tools, which may be of practical benefit both for research projects and commercial applications.

However, to ensure higher quality some further work is still needed. We plan to revise the tags for the problematic grammatical classes presented in the previous section. As the dictionary is free and open, it may also benefit from other people’s contributions. The other task is to evaluate the dictionary as data source for morpho-syntactic tagging of Polish.

Acknowledgements

The project is financed by the National Centre for Research and Development (NCBiR) agreement SP/I/1/77065/10.

We would like to thank Marcin Miłkowski for his help in the described work, which included both insightful discussions and corrections of the dictionary.

[7] http://nlp.pwr.wroc.pl/redmine/projects/libpltagger/wiki

References

Bień, J. S. and Szafran, K. (2001). Analiza morfologiczna języka polskiego w praktyce. In *Biuletyn Polskiego Towarzystwa Językoznawczego XLVIII*, pages 171–184. Polskie Towarzystwo Językoznawcze.

Grzegorczykowa, R., Laskowski, R., and Wróbel, H., editors (1998). *Gramatyka współczesnego języka polskieg*, volume I. PWN.

Hajnicz, E. and Kupść, A. (2001). Przegląd analizatorów morfologicznych dla języka polskiego. Technical Report 937, IPI PAN.

Markowski, A., editor (2006). *Wielki słownik poprawnej polszczyzny*. PWN.

Miłkowski, M. (2010). Developing an open-source, rule-based proofreading tool. *Software: Practice and Experience*, 40:543–566.

Ogrodniczuk, M. (2003). Nowa edycja wzbogaconego korpusu słownika frekwencyjnego. In *Językoznawstwo w Polsce. Stan i perspektywy*, pages 181–190. Polska Akademia Nauk, Komitet Językoznawstwa, Uniwersytet Opolski, Instytut Filologii Polskiej, Opole.

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

Przepiórkowski, A. (2003). Składniowe uwarunkowania znakowania morfosyntaktycznego w korpusie IPI PAN. *Polonica*, XXII–XXIII:57–76.

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

Przepiórkowski, A. (2008). *Powierzchniowe przetwarzanie języka polskiego*. Akademicka Oficyna Wydawnicza EXIT, Warsaw. The configuration files can be obtained from http://nlp.ipipan.waw.pl/PPJP/.

Przepiórkowski, A. and Woliński, M. (2003). A flexemic tagset for Polish. In *Proceedings of Morphological Processing of Slavic Languages, EACL 2003*.

Radziszewski, A. and Śniatowski, T. (2011). Maca — a configurable tool to integrate Polish morphological data. In *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*.

Saloni, Z. and Świdziński, M. (1998). *Składnia współczesnego języka polskiego*. PWN.

Woliński, M. (2006). Morfeusz — a practical tool for the morphological analysis of Polish. In Kłopotek, M. A., Wierzchoń, S. T., and Trojanowski, K., editors, *Proceedings of IIPWM'06*, pages 511–520, Ustroń, Poland. Springer-Verlag, Berlin.