1. Introduction

Morfeusz is a morphological analyser for Polish. During its over 10 years long history it has established its position as a basic resource for morphological processing of Polish. Morfeusz was used for the annotation of The IPI PAN Corpus of Polish (Przepiórkowski, 2004) and National Corpus of Polish, NKJP, (Przepiórkowski et al., 2011), it serves as a basis for several taggers: (Dębowski, 2004), TaKIPI (Piałecki, 2007), PANTERA (Acedański, 2010), WMBT (Radziszewski and Śniatowski, 2011), Concraft-pl (Waszczuk, 2012). Moreover, Morfeusz was integrated with several parsing tools (Spejd (Przepiórkowski, 2008), Śwигра (Wołński, 2004), SproUT (Piskorski et al., 2004)) as well as with the multiword expression toolkit Multiflex (Savary, 2005). During these years we have accumulated some experience and have identified its drawbacks. Now the time has come for an overhaul of the program.

First version of Morfeusz (Woliński, 2006) was based on approximated description of Polish inflection. Soon the data was replaced with inflectional data coming from the Grammatical dictionary of Polish, SGJP (Saloni et al., 2007), which is much richer and more precise. At some point this version was released as an open source program. Then the data of SGJP was merged with the community developed dictionary of Polish (sjp.pl), resulting in the largest freely available inflectional dictionary of Polish – Polimorf (Wołński et al., 2012). Currently Morfeusz is available in both SGJP and Polimorf flavours.

The new version of the program described here includes not only an analyser but also a compatible generator. Morfeusz is getting less tightly bound to its dictionary. The present version provides an infrastructure for including domain dictionaries or replacing the basic one completely. We are also working on some optimisations of finite automata used in the program.

2. Segmentation, Morphological Analysis

As for segmentation (or tokenization), we assume that segments cannot contain blanks so each segment is contained within a word. However, we allow for words consisting of several segments. A simple example is words containing punctuation characters that have to be interpreted separately (we have a separate tag interp for punctuation).

The next level of complication involves some productive mechanisms in the language which introduce myriads of words of very low textual frequency. Polish adjectives have the ability to form compounds like zielono-niebieski meaning ‘partly green and partly blue’ and zielononiebieski meaning ‘having a color between green and blue’. This works not only for colours: ‘a box made of wood and metal’ can be drewniano-metalowe pudełko and a Polish-Czech-Hungarian summit is szczyt polsko-czesko-węgierski. Including such lexemes in the dictionary does not make much sense, since the mechanism is very regular and the meaning of a compound can be determined from its components. We have decided to split such formations into several segments. Unfortunately the hyphen is not an obvious segment boundary in Polish, since it is used in inflection of acronyms, e.g., PRL-u (genitive of PRL, the acronym for ‘People’s Republic of Poland’).

These facts lead to the conclusion that proper segmentation for Polish has to be dictionary-based. We assume that an inflectional dictionary consists of entries describing some abstract units of the language. We call these units lexemes. A lexeme can be considered to be a set of other abstract units — namely grammatical forms. Lexemes gather sets of forms which have similar relation to the reality (e.g., all denote the same physical object) and differ in some regular manner. The differences between forms are described with values of grammatical categories attributed to them. Forms are represented in texts by segments.

For identifying the lexemes we will use lemmas (base forms), which traditionally have the shape of one of the forms belonging to the lexeme but should be in fact considered arbitrary unique identifiers (see also Section 4.2.). By morphological analysis we will understand the interpretation of segments as grammatical forms. Technically that means assignment of a lemma and a tag. The lemma identifies a lexeme and the tag contains values of grammatical categories specifying the form.

In case of ambiguity, the result of morphological analysis includes all possible interpretations. We do not pay attention to the context that a word occurs in. In this setting, morphological tagging consists of morphological analysis and contextual disambiguation.

Figure 1 presents an example of morphological analysis.

Keywords: morphological analysis and generation, Polish, finite state automata.
Each row of the table includes one morphological interpretation, the lines separate groups of interpretations for respective segments. The input text was segmented into tokens (in particular the full stop was separated from the word *morfologicznej*). Corresponding lemmas were provided in the third column. The last column presents tags describing the values of grammatical categories of particular forms. The word *mam* has three interpretations: the genitive plural form of the noun *mama*, the imperative of the verb *mamieć* and the present tense form of the verb *mieć*. The word *analizy* was unambiguously associated with the lemma *analiza* but with two possible tags representing singular and plural form in different grammatical cases. The tags are positional. The first position defines the part of speech (more precisely: the flexeme, see below), the following ones stand for the values of grammatical categories of each class. For instance, the tag *subst* stands for a noun, it is followed by the values of the number, case and gender. The tags are usually abbreviated forms of Latin value names.

| 0 | 1 | Mam  | Mama [mother] | subst:pl:gen:f |
|---|---|------|----------------|----------------|
|   | 2 | próbkę | próbka [sample] | subst:sg:acc:f |
|   | 3 | analizy | analiza [analysis] | subst:sg:gen:f |
|   | 4 | morfologicznej | morfologiczny [morphological] | adj:sg:gen.dat.loc:f:pos |
|   | 5 | .  | .  | interp |

Thus, a lexeme can be considered to be a set of flexemes which are sets of forms. In total there are 13 different verbal flexemes. If they are to be considered collectively as verbal forms, the processing system has to maintain their list. The relation is purely deterministic.

We call the tagset used in Morfeusz morphosyntactic since some attributes contained in the tags are not of inflectional nature. For example we provide information on gender for nouns, although Polish nouns do not inflect for gender. Nonetheless, gender is included in the tags as an important attribute of nominal lexemes describing their syntactic features.

The tagset uses a very detailed system of 9 genders (Przepiórkowski and Wolński, 2003c) based on the works of Saloni (1976). This system was reduced to 5 genders for the NKJP (Przepiórkowski, 2009), which simplifies automatic tagging. Morfeusz, however, uses the more detailed classification, since projecting it to 5 genders is trivial but the opposite transformation is not.

The most controversial feature of the Morfeusz and NKJP tagsets concerns movable inflections. In Polish, endings of past tense forms of verbs can be detached from the verb form under some conditions. This is illustrated with the following examples:

1. *Nie wiedziałem, że to czytałeśście.*
   Not known-I that it read-you
   *I didn’t know that you have read this.*

2. *Nie wiedziałem, żeście to czytali.*
   Not known-I that/aux-you it read
   *I didn’t know that you have read this.*

The construction in the second example is probably more common in less formal texts, but with some complementizers (mainly used in the conditional) the detachment is obligatory:

3. *(Przyszedłbym, gdybyście czytała).*
   Would-have-come-I if it read-you
   *I would have come if you read this.*

4. *(Przyszedłbym, gdybyśś cie czytała.)*
   Would-have-come-I if/aux-you it read
   *I would have come if you read this.*

This means that movable inflections have to be accounted for in a linguistically adequate tagset of Polish. In the Morfeusz tagset it was decided to describe these inflections as
We prefer the library+bindings architecture to implementing the analyser within some particular NLP toolkit, since that way we are not binding users to that toolkit.

1OK, it caused problems even there: each time the size of the corpus in tokens was to be reported, it was necessary to explain how tokens of NKJP corresponded to words.

2We prefer the library+bindings architecture to implementing the analyser within some particular NLP toolkit, since that way we are not binding users to that toolkit.
Figure 2: Morphological interpretations for the sentence Coś zrobił? with ambiguous segmentation. The sentence can be read as ‘What (co) have you done (ś zrobił)?’ or ‘Did he do (zrobił) anything (coś)?’.

names, last names, pseudonyms, and patronyms. Appropriate labels have been added for the whole scope of Polimorf dictionary.

Obviously, this information is not part of morphological tagging. Nonetheless, it was included since it can be useful even when semantic processing is not done, e.g., for parsing names.

The labels include, e.g., ‘archaism’, ‘colloquialism’, ‘coarse/vulgarism’, as well as those signalling terminology of specific domains (e.g., ‘chemical’). The labels allow to filter out some interpretations when the domain of the text analysed is known. For example, when processing hospital documentation one can safely ignore archaic words and colloquialisms. It is probably also safe to ignore vocative forms of nouns (which can be homonymous with nominatives) and imperatives of verbs. Both make sense for limiting homonymy in Polish.

4.2. Generation

The generating module of Morfeusz has two flavours. The first takes a lemma and generates the full paradigm of the given lexeme (i.e. forms of all flexemes comprising the lexeme). The second takes a lemma and a tag and generates only forms matching that tag (there can be more than one). In both cases the program returns structures closely resembling results of analysis (including stylistic labels and proper name information).

In case of homonymy we use lemmas containing disambiguating elements. For example, in the case of the lexeme PIEC which in Polish can be a verb (‘to bake’) or a noun (‘an oven’) the lemmas have the form PIEC:V and PIEC:S, respectively. If there is more than one lexeme of the same grammatical class arbitrary numbers are used. For instance, the dictionary contains lexemes ZAMEK:S1 (‘a castle’, with genitive zamka) and ZAMEK:S2 (‘a lock’, with genitive zamka). In this example the difference in inflection is what forces us to introduce two separate lexemes. Even if a word has several clearly separate meanings we will consider it a single lexeme if the meanings share the whole inflectional paradigm (as in the case of the noun PARA ‘a couple’ or ‘a vapour’).

Since we lemmatise deverbal flexemes to the infinitive, we are free from some cases of systematic homonymy in Polish.

For example, the gerund MIESZKANIE derived from the verb MIESZKA/C (‘to live/inhabit’) is homonymous with a noun (‘a flat’). However, the gerund is a part of the verbal lexeme with the lemma MIESZKA/C and the lemma MIESZKANIE points unambiguously to the noun. The same goes for homonymy between adjectival participles and regular adjectives.

The SGJP dictionary contains only about 10,000 lemmas with a disambiguator. The use of arbitrary numbers is a bit unfortunate. But the analysing module always generates lemmas that can be fed back to the generator. Thus, if we analyse the word ZAMKA, we will learn that the corresponding lemma is ZAMEK:S2. Moreover, to ease this situation the generating module accepts lemmas without the disambiguating part and generates forms of all matching lexemes in response (so a call with lemma PIEC will result in both verbal and nominal forms generated).

5. Dictionaries

Previous versions of Morfeusz used to be tightly coupled with a compiled-in dictionary. In the present version we want to be able to adapt the dictionary to particular needs. The tool described in our previous paper (Woliński et al., 2012) allows to work simultaneously on several dictionaries. It is used for development of both dictionaries distributed with Morfeusz. But it can be used as well to develop domain dictionaries. Such need arises when processing, e.g., hospital documentation as some of medical terminology is too specific to be included in a general dictionary. Moreover, hospital documentation uses specific set of abbreviations which should not be considered when processing general text.

The tool allows to export a list of forms from an arbitrary set of dictionaries contained in the system. The dictionary compiling tool of Morfeusz turns such lists into a binary representation used by the Morfeusz library. Obviously lists of forms can be also prepared by other means and merged with those in Morfeusz distribution or replace them completely.

5.1. Precompiled Dictionaries

Two inflectional dictionaries are included in the program’s distribution available at http://sgjp.pl/
Table 1: Sizes of precompiled dictionaries of Morfeusz

|       | SGJP | Polimorf |
|-------|------|----------|
| lemmas| 264166 | 315055   |
| forms | 4037250 | 3844535  |

morfeusz/. The SGJP dictionary contains data from the second edition of the Grammatical dictionary of Polish (Saloni et al., 2012). The Polimorf dictionary is a merger of SGJP with Morfologik dictionary based on community developed sjp.pl dictionary (Woliński et al., 2012). Table 1 presents the sizes of these dictionaries.

5.2. Compiling Dictionaries

The core dictionary of Morfeusz maps segments to sets of possible interpretations. The dictionary is represented as a minimal deterministic finite state automaton with the transitions labelled with consecutive letters of the words and the accepting states labelled with interpretations. The automaton is generated with a variant of the algorithm presented by Daciuk et al. (2000).

Figure 3 presents the form of a dictionary that is fed to the dictionary compiler. Each row contains one grammatical form. Five columns are separated with tabulation (U+0008). Their content is as follows:

1. segment
2. lemma (including disambiguator if necessary)
3. tag
4. proper name/common classification
5. stylistic label(s) (optional)

The list used by Morfeusz contains all the inflected forms of lexemes including the special adjectival and numeral forms used in compounding. It includes as well special segments that cannot appear by themselves but can be combined with another segment to form a complete word (see below).

6. Segment Joining (Compounding)

As explained above, Morfeusz treats some orthographic words as consisting of several tokens interpreted separately. This mechanism is used, for example, to analyse compound adjectival forms like biało-czerwony ‘red and white’.

In the new version we have also taken into account less common adjectival compounds without a hyphen (ciemnoczerwony ‘dark red’) and compounds including numeral element (drwurzędowy ‘having [two rows], drugorzędowy [belonging to the] second row’). Also the rules for attaching movable inflections were extended to some inflecting lexemes, mainly pronouns (myszy ‘we/1per.pl.aux’).

The compounding mechanism is also used to guess lexemes unknown to Morfeusz that can be derived with a list of frequently used Polish prefixes. The list of prefixes is kept in the dictionary, which means users can adjust the list depending on the domain/genre of texts.

To describe allowed combinations of segments, each segment in the dictionary is associated with a segment type. These are defined in an additional file that is used by the dictionary builder together with the list of forms. Segment types can be associated with specific tags or with specific forms of specific lexemes. The latter take precedence as exceptions. Another section of the file defines possible combinations of segments in terms of regular expressions over segment types.

This mechanism allows us to experiment with segmentation rules without recompiling the program. This is useful since, e.g., the possibility of agglutinative formations occurring in the text depends on the genre of the text. The pre-compiled dictionaries of Morfeusz contain several variants of the rule set that can be selected with options at run-time.

7. Summary

Morfeusz together with its SGJP and Polimorf dictionaries is available under the very liberal Z-2-clause BSD license. This makes it accessible both for scientific and commercial uses. Morfeusz’s model of inflection has a linguistically sound base. The changes in the present version make the program more attractive for simplified practical solutions.
Fourth European Conference on Formal Description on Slavic Languages, pages 349–362.

Przepiórkowski, A., and Woliński, M. (2003c). 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, pages 109–116.

Przepiórkowski, A., Bańko, M., Górski, R. L., and Lewandowska-Tomaszczyk, B., editors. (2011). Narodowy Korpus Języka Polskiego. Wydawnictwo Naukowe PWN, Warsaw.

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

Przepiórkowski, A. (2004). Korpus IPI PAN. Wersja wstępna. Instytut Podstaw Informatyki PAN, Warszawa.

Przepiórkowski, A. (2008). Powierzchniowe przetwarzanie języka polskiego. Exit, Warsaw.

Przepiórkowski, A. (2009). A comparison of two morphosyntactic tagsets of Polish. In Koseska-Toszewa, V., Dimitrova, L., and Roszko, R., editors, Representing Semantics in Digital Lexicography: Proceedings of MONDILEX Fourth Open Workshop, pages 138–144, Warsaw.

Radoszewski, A., and Śniatowski, T. (2011). A memory-based tagger for Polish. In Vetulani, Z., editor, Proccedings of the 5th Language Technology Conference, pages 556–560, Poznań.

Saloni, Z., Gruszczyński, W., Woliński, M., and Wołosz, R. (2007). Słownik gramatyczny języka polskiego. Wiedza Powszechna, Warszawa.

Saloni, Z., Woliński, M., Wołosz, R., Gruszczyński, W., and Skowrońska, D. (2012). Słownik gramatyczny języka polskiego. Warszawa, 2 edition.

Saloni, Z. (1974). Klasyfikacja gramatyczna leksemów polskich. Język Polski, LIV:1, 3–13, z.2, 93–101.

Saloni, Z. (1976). Kategoria rodzaju we współczesnym języku polskim. In Kategorie gramatyczne grup imiennych we współczesnym języku polskim, pages 41–75. Osсолinum, Wrocław.

Savary, A. (2005). MULTIFLEX. User’s Manual and Technical Documentation. Version 1.0. Technical Report 285, LI-François Rabelais University of Tours, France.

Waszczuk, J. (2012). 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), pages 2789–2804, Mumbai, India.

Woliński, M. and Przepiórkowski, A. (2001). Projekt anotacji morfosyntaktycznej korpusu języka polskiego. Prace IPI PAN 938, Instytut Podstaw Informatyki Polskiej Akademii Nauk.

Woliński, M., Miłkowski, M., Ogrodniczuk, M., Przepiórkowski, A., and Szalkiewicz, Ł. (2012). PoliMorf: a (not so) new open morphological dictionary for Polish. In Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, pages 860–864, Istanbul, Turkey. ELRA.

Woliński, M. (2003). System znaczników morfosyntaktycznych w korpusie IPI PAN. Polonica, XXII–XXIII:39–55.

Woliński, M. (2004). Komputerowa weryfikacja gramatyki Świdzińskiego. Ph.D. thesis, Instytut Podstaw Informatyki PAN, Warszawa, December.

Woliński, M. (2005). An efficient implementation of a large grammar of Polish. In Vetulani, Z., editor, Human Language Technologies as a Challenge for Computer Science and Linguistics. 2nd Language & Technology Conference April 21–23, 2005, pages 343–347, Poznań.

Woliński, M. (2006). Morfeusz — a practical tool for the morphological analysis of Polish. In Klopotek, M., Wierchoń, S., and Trojanowski, K., editors, Intelligent Information Processing and Web Mining, IIS:IIPWM ’06 Proceedings, pages 503–512. Springer.