PoliMorf: a (not so) new open morphological dictionary for Polish

Marcin Woliński, Marcin Miłkowski,
Maciej Ogrodniczuk, Adam Przepiórkowski, Łukasz Szałkiewicz

Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, Warsaw, Poland

marcin.wolinski@ipipan.waw.pl, marcin.milkowski@ifispan.waw.pl,
maciej.ogrodniczuk@ipipan.waw.pl, adam.przepiorkowski@ipipan.waw.pl

Abstract

This paper presents preliminary results of an effort aiming at the creation of a morphological dictionary of Polish, PoliMorf, available under a very liberal BSD-style license. The dictionary is a result of a merger of two existing resources, SGJP and Morfologik and was prepared within the CESAR/META-NET initiative. The work completed so far includes re-licensing of the two dictionaries and filling the new resource with the morphological data semi-automatically unified from both sources.

The merging process is controlled by the collaborative dictionary development web application Kuźnia, also implemented within the project. The tool involves several advanced features such as using SGJP inflectional patterns for form generation, possibility of attaching dictionary labels and classification schemes to lexemes, dictionary source record and change tracking.

Since SGJP and Morfologik are already used in a significant number of Natural Language Processing projects in Poland, we expect PoliMorf to become the Polish morphological dictionary of choice for many years to come.

Keywords: morphological dictionaries, Polish, Morfeusz SGJP, Morfologik, PoliMorf

1. Introduction

The CESAR project (part of the META-NET; cf. `http://www.meta-net.eu/projects/cesar`), running from February 2011 to January 2013, intends to address the issue of long-term sustainability, interoperability and re-usability of language resources and tools (LRTs) for Central and East European languages by enhancing, upgrading, standardising and cross-linking them, thus contributing to the open linguistic infrastructure. In Poland, CESAR triggered a series of improvements in the field, starting with standardisation and liberalisation of licensing models used for LRTs.

The licensing of morphological resources is crucial, as it affects all further stages of processing: during the process of morphological analysis, fragments of an inflectional dictionary are copied to the analysed text. In effect, according to lawyers, any text with added morphological information becomes a derivative work of the dictionary. In case the licensing terms of the dictionary are in contradiction with the rights to the text itself, the text becomes non-distributable. The only way out of this conundrum seems to be the use of an inflectional dictionary with an extremely liberal licence.

This paper presents some preliminary results of an effort — carried out within CESAR — aiming at the creation of such a dictionary. Given the number of morphological dictionaries of Polish (section 2.), creating another one from scratch would be a waste of time and resources. Instead, the owners of two of the most popular and comprehensive dictionaries — produced independently and expected to be partially complementary — agreed to license them on the very liberal 2-clause BSD licence and join efforts in developing the ultimate morphological resource for Polish (section 3.): PoliMorf.

2. Polish Morphological Dictionaries

Although a report published over 10 years ago, Hajnicz and Kupść 2001, already mentions 12 morphological dictionaries or analysers for Polish, most of them are not publicly available or are not free even for non-commercial scientific purposes. Until recently only a few such resources of a reasonable size and quality were freely available for research, most notably:

- UAM Text Tools (`http://utt.amu.edu.pl/`; Vetulani and Obrębski 1997, Obrębski and Stolarski 2006), with the underlying dictionaries now licensed under both Creative Commons (CC) Attribution Non-Commercial Share Alike (`by-nc-sa`) and GNU General Public License (GPL),
- Morfeusz, until recently free for non-commercial use, but not open source, and
- Morfologik, until recently available on GNU Lesser General Public Licence (LGPL) and CC `sa`.

The last two resources, forming the basis of the (not so) new morphological dictionary PoliMorf, are described in more detail below.

2.1. SGJP and Morfeusz SGJP

Morfeusz SGJP (`http://sgjp.pl/morfeusz`, Woliński 2006) is a morphological analyser for Polish whose inflectional data (dictionary) comes from SGJP — *Grammatical Dictionary of Polish* (Saloni et al. 2007a).

SGJP is the result of several years of work of an informal group led by Prof. Saloni. The work started in the 1980s by digitising the list of headwords of the 11-volume Doroszewski's dictionary of Polish (Doroszewski

	Lexemes	Patterns
total	323,946	1095
nouns	168,929	762
common	68,682	
proper	9,919	
gerunds	29,851	
deadjectival (*-ość*)	60,477	
adjectives	98,705	71
"regular"	64,033	
participles	34,672	
numerals	116	45
verbs	29,804	215
non-inflecting	26,392	2
adverbs	25,344	
prepositions	115	
other	933	

Table 1: Number of entries representing various grammatical classes in SGJP

1958–1969). The grammatical description in SGJP is based on new concepts proposed in the 2nd half of the 20th century (cf. Saloni et al. 2007b), with many detailed solutions proposed by the members of the team (e.g. Tokarski 1993, Gruszczyński 1989, Saloni 2001). PoliMorf will use data from the second edition of SGJP presented in numbers in Table 1. The stated number of lexemes corresponds to 4,223,981 word forms (counting syncretic forms of the same lexeme as one unit).

Inflection in SGJP is represented with inflectional patterns, which describe forms in terms of a stem common to all forms and endings differentiating the forms. The model of inflection is in fact more complicated (cf. Woliński 2009), but the high level of irregularity in Polish inflection still leads to numerous inflectional patterns — over a thousand.

An important feature of SGJP is that it is to some extent tagset-agnostic. Formation of inflected words is described separately from labelling them with grammatical features. Morfeusz SGJP uses the IPI PAN Tagset (Przepiórkowski and Woliński 2003), but adoption of a radically different tagset would be equally easy.

The license of Morfeusz SGJP restricted its use in the past, even if it was already fairly liberal. Re-licensing paved the way for the integration with the other large morphological resource for Polish.

2.2. Morfologik

Morfologik is probably the first truly open source morphological dictionary of Polish. It is accompanied with an analyser library, Morfologik-stemming. It contains 216,992 lexemes and 3,475,809 word forms.

The dictionary was created by enriching the Polish ispell/hunspell dictionary with morphological information, which was possible thanks to the structure of the original dictionary that retained important grammatical distinctions (Miłkowski 2010). The process of conversion relied on a series of scripts, and the resulting dictionary was later augmented with manually entered information. Unfortunately, the original source dictionary did not contain sufficient structure to allow reliable detection of some information, such as the exact subgender of the masculine for substantives. This information was added manually and using heuristic methods, however its reliability is low. Considering the fact that the substantives are about one third of the dictionary content (and almost half of them are masculine), this limitation is severe.

The tagset of the dictionary is inspired by the IPI PAN Tagset (Przepiórkowski and Woliński 2003). However, Morfologik diverges from that tagset and from Morfeusz, as it never splits orthographic ("space-to-space") words into smaller dictionary words (i.e., so-called *agglutination* is not considered). Moreover, due to the lack of information in the ispell dictionary, some forms are not completely annotated, and are marked as irregular. There is, however, some additional markup added to reflexive verbs, which is not present in the original IPI PAN Tagset. This was introduced for the purposes of the grammar checker LanguageTool that used the dictionary extensively. In contrast to SGJP, Morfologik was closely linked with a variant of the IPI PAN Tagset and adoption of a radically different tagset was not practical because of the flat textual representation of morphological data.

2.3. Differences in Lexicographical Approaches

SGJP and Morfologik differed in lexicographic approaches to Polish morphology. The differences in how information in both of them is structured are therefore not merely a result of some historical coincidences, and the fact that Morfologik, in particular, was not a scientific project but an effort to produce a resource for practical purposes.

One of the most important differences is the analysis of certain forms of Polish verbs, as noted already in the previous subsection. While in SGJP the verb form such as *poszedłbym* 'I would go', lit. 'go-would-I', is analysed into three component parts, namely *poszedł* 'go', *by* 'would', and *m* 'I', in Morfologik it remains a single unit. Because of that, some ambiguous words have to be represented by the Morfeusz SGJP analyser as graphs, while the Morfologik-stemming represents them as flat lists. For example, the word *miałem* is both a past masculine form of *mieć* 'to have' and also a singular instrumental form of *miał* 'dust'. It is quite obvious that, for this reason, the two resources could not be simply concatenated, and they needed to be brought under a common data representation scheme.

It is worth stressing that, for simplicity, some researchers preferred a flat-list representation of ambiguous forms, and that accounted for some popularity of Morfologik and its morphological analyser library, Morfologik-stemming, even if the quality of the dictionary was lower than of SGJP. Morfeusz SGJP, however, was harder to interface with those NLP tools which do not support graph representations of results of morphological analysis.

Morfologik is based on an open source effort to expand the Polish ispell/hunspell dictionary for the needs of the gaming site, `http://kurnik.pl/`. This dictionary is de-

veloped by an online community (especially fans of word games) and is currently available at `http://www.sjp.pl/` (where SJP stands for *Słownik Języka Polskiego*, 'Dictionary of Polish Language'). Initially, only the words that were already acknowledged to exist in previously published modern dictionaries of Polish were allowed in the dictionary. This helped to avoid the charges of arbitrariness in allowing or disallowing certain word forms in word games.

Only after some time, it was realized that the huge size of the dictionary might be detrimental for spell-checking purposes: some common typos share the same form with rare words (like *sie*, which is a typo used instead of *się*, but is analysed as a plural form of *si*, an archaic adjective used currently only in its masculine genitive singular form *siego* in *Do siego roku!* 'Happy New Year!'). Consequently, some words or word forms were removed from the spelling-check version of the SJP dictionary; they were also removed from the Morfologik resource. Nonetheless, it might be said that Morfologik is ecumenically pluralistic in its approach to morphology: whatever form was present in modern dictionaries of Polish, it was admitted by the administrators of SJP. This stands in stark contrast to SGJP, where the decision to introduce or remove a word was based on systematic grammatical principles that were explicitly formulated. On the other hand, the grammatical rules of the inflection in the SJP have been implemented on the base of Saloni's works, which is a link between the two dictionaries discussed in the present paper.

Another difference between SGJP and Morfologik concerns the plural form of proper names. While proper names were added in SJP, the basis of Morfologik, quite late (they were not allowed to be used in the word game), once they were added, it was on a massive scale. However, only some plural forms (according to special rules) were added. In SGJP, on the other hand, all proper names are associated with their plural form, whether it is actually used this way, or not.

3. Collaborative Dictionary Development

A web-based tool for collaborative dictionary development, Kuźnia (Polish for 'forge'), has been implemented within CESAR. The system manages a database of lexemes and makes it possible to edit their descriptions, first of all to characterise their inflectional paradigms (cf. Figure 1).

The database is modelled after SGJP, in particular its inflectional patterns are used directly. The system makes it possible to attach various labels to lexemes. Besides typical dictionary labels like *informal* or *dated*, special labels are used for excluding some forms from spell-checking dictionaries. This way a special variant of the dictionary can be generated which does not contain certain theoretically correct but extremely infrequent words (i.e., potential false negatives in spell-checking).

Moreover, the system makes it possible to specify a classification scheme (or several classification schemes), which the lexemes are to follow. This mechanism is currently used to classify lexemes into common and proper names (with some subclasses).

A feature of Kuźnia crucial in the context of this paper is that it is able to work with multiple dictionaries and retain the information about the source dictionary of each lexeme. But, even more importantly, this mechanism facilitates work on domain dictionaries. These are not a subject of the CESAR project itself, but the need seems obvious: for example, when processing medical texts, a dictionary of medical terms would be needed and, at the same time, such specialist terms should not enter the general dictionary.

Kuźnia makes it possible to generate various derivative forms of its dictionaries. The process is configurable, so that it is possible to generate inflectional dictionaries assuming various tagsets and even various rules of segmenting the text (e.g., whether analytical forms of verbs are included in the output). Moreover, it is possible to generate, e.g., forms included in the medical dictionary plus forms from the general dictionary but without those marked as *dated*.

The database of Kuźnia was initialised with SGJP data and currently Morfologik data is being added and manually verified using the system's facilities (see the next section). After importing all available resources, the database will probably be extended with frequent — as witnessed by the National Corpus of Polish (Przepiórkowski et al. 2010; `http://nkjp.pl`) — forms absent from PoliMorf. At that stage of dictionary development it will be important to be able to easily assign inflectional patterns to the lexemes introduced into the system. Fortunately, Kuźnia includes a helpful tool, which has two modes of operation. In both modes the user is first asked for the basic information about the lexeme: its lemma, part of speech, gender for nouns and aspect for verbs, and whether the lexeme is a proper name.

Then, in the first mode, the tool suggests patterns used by lexemes with the base form having a similar ending (close in lexicographical ordering of reversed headwords) and the same characteristics (e.g., a noun of the same gender and the same status as a proper name/common word). Since the dictionary is already large, this mode almost always proposes the right pattern(s).

In the second mode, the user is asked to give a few inflected forms of the lexeme. The system asks for forms which give the most discriminative information with respect to selecting the right inflectional pattern. Usually it is possible to select the right pattern using only a few forms.

If no matching inflectional pattern is found, the lexeme is passed to the editor specialised in inflectional patterns. This is not a common situation, but it happens especially with uncommon proper names from foreign languages.

4. Merging the Dictionaries

As mentioned in section 3., the merging of the two resources within PoliMorf was initiated by the import of the SGJP data into Kuźnia, followed by the import of Morfologik. An attempt was made to automatically match entries of Morfologik to the structure and inflectional paradigms of SGJP.

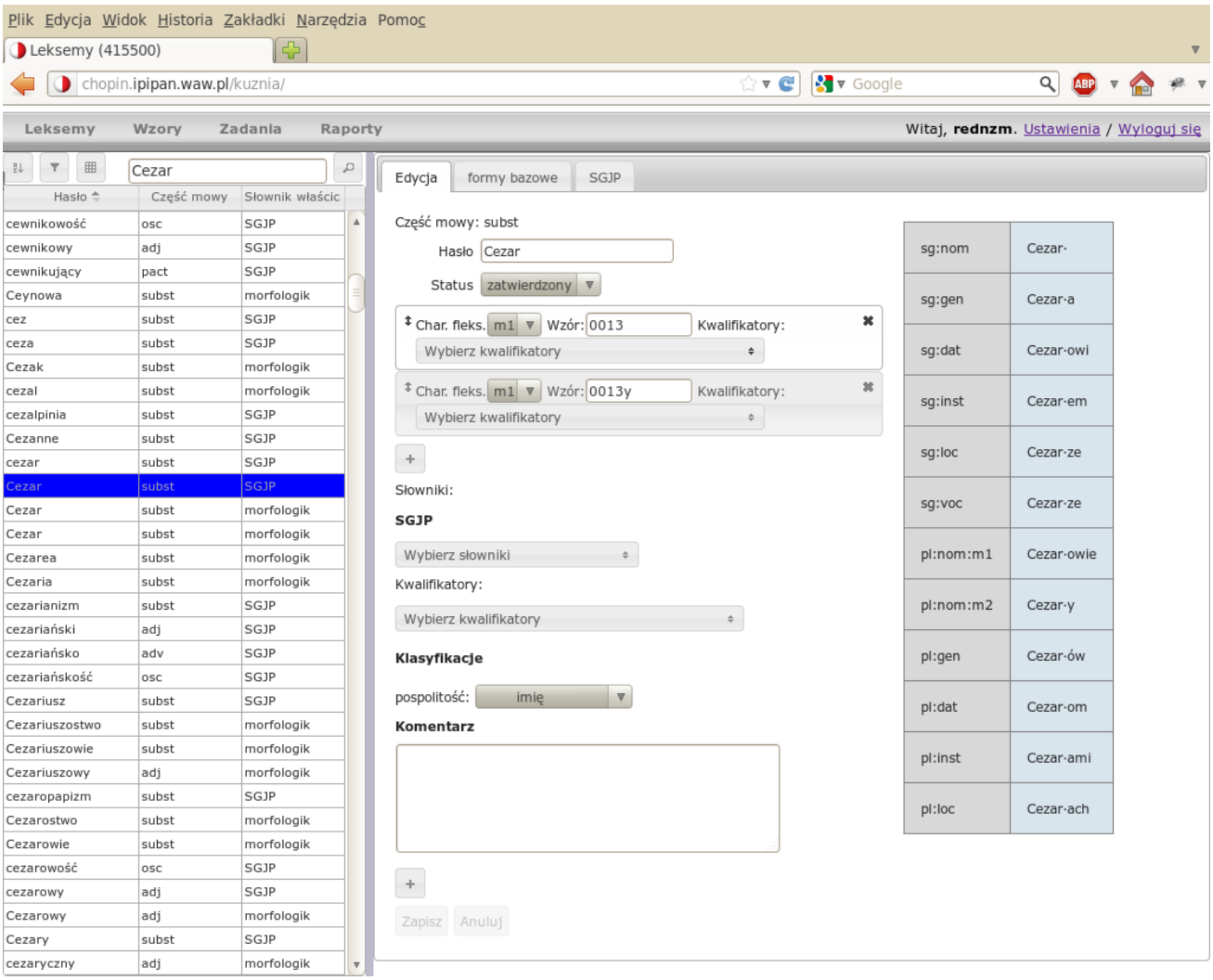

Figure 1: The main screen of Kuźnia showing the description of the noun *Cezar*. The highlighted entry shows inflection according to SGJP with two assigned inflectional patterns differing in nominative plural (*Cezarowie* or *Cezarzy*). The other two entries for *Cezar* were imported from Morfologik. These 3 entries will be merged in PoliMorf.

	common	SGJP only	Morfologik only	total
nouns	74342	94587	50922	219851
adjectives	66434	32271	33224	131929
verbs	17442	12362	4023	33827
non-inflecting	9111	17281	2304	28696
total	**167329**	**156501**	**90473**	**414303**

Table 2: Numbers of lexemes introduced to PoliMorf by each of the source dictionaries

In order to evaluate the quality of the automatic merging procedure and to plan the necessary manual work, 100 lexemes were randomly sampled from each of the following groups:

1. lexemes present in both dictionaries with the same set of inflectional forms (such lexeme pairs are collapsed into one in PoliMorf),
2. new lexemes from Morfologik, to which inflectional characteristics have been unequivocally assigned,
3. new lexemes from Morfologik, which could not be automatically assigned:
   (a) parts of speech,
   (b) inflectional properties (e.g., nouns' gender),
   (c) inflectional paradigm;
4. lexemes present in both dictionaries which could not be collapsed due to:
   (a) grammatical differences between descriptions in the two dictionaries,
   (b) erroneous data import,
   (c) wrong description in one of the dictionaries.

Group 1 includes over 100,000 lexemes. No errors were

detected in its sample. In the sample of group 2 no errors were detected, either. The total number of lexemes of this type is about 40,000. New lexemes from Morfologik represent several parts of speech and come from various semantic fields. A large proportion of these lexemes is constituted by proper names, foreign as well as Polish, including a few thousand Polish names (in particular, in PoliMorf, *Woliński* already present in SGJP, was joined by *Miłkowski* and *Przepiórkowski*).

Group 3 consists mainly of uninflected lexemes and masculine nouns. In Polish, the masculine comes in several variants but it is virtually impossible to detect the variant based only on the surface form of the word ending, which was used in semi-automatic process to import Morfologik's data. Moreover, Morfologik did not contain correct annotation for these nouns in the first place, as the automatic conversion of the original SJP, again for the same reason, did not allow it.

Finally, the perusal of a sample of 100 lexemes of group 4 suggests what kind of manual work is necessary to be performed in order to merge the two lexicons.[1] Apart from differences concerning inflectional paradigms, particularly many differences between the two resources were related to the presence or absence of passive forms in these dictionaries.

5. Conclusion

At the time of submitting this paper, completed work includes re-licensing of the two morphological dictionaries, creating the collaborative dictionary development tool Kuźnia, filling it with the SGJP data and – to some extent – merging the data with Morfologik.

Since Morfeusz and Morfologik are already used in a significant number of Natural Language Processing projects in Poland, we expect PoliMorf to become the Polish morphological dictionary of choice for many years to come.

6. Acknowledgements

Research funded in 2011–2013 within CESAR (CEntral and South-east europeAn Resources), a European (CIP ICT-PSP) project (grant agreement 271022), part of META-NET.

7. References

Witold Doroszewski, editor. *Słownik języka polskiego PAN*. Wiedza Powszechna – PWN, 1958–1969.

Włodzimierz Gruszczyński. *Fleksja rzeczowników pospolitych we współczesnej polszczyźnie pisanej*, volume 122 of *Prace językoznawcze*. Zakład Narodowy im. Ossolińskich, Wrocław, 1989.

Elżbieta Hajnicz and Anna Kupść. Przegląd analizatorów morfologicznych dla języka polskiego. IPI PAN Research Report 937, Institute of Computer Science, Polish Academy of Sciences, Warsaw, 2001.

Marcin Miłkowski. Developing an open-source, rule-based proofreading tool. *Software: Practice and Experience*, 40(7):543–566, 2010.

Tomasz Obrębski and Michał Stolarski. UAM text tools — a flexible NLP architecture. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006*, pages 2259–2262, Genoa, 2006. ELRA.

Adam Przepiórkowski and Marcin Woliński. 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, 2003.

Adam Przepiórkowski, Rafał L. Górski, Marek Łaziński, and Piotr Pęzik. Recent developments in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta, 2010. ELRA.

Zygmunt Saloni. *Czasownik polski. Odmiana, słownik*. Wiedza Powszechna, Warszawa, 2001.

Zygmunt Saloni, Włodzimierz Gruszczyński, Marcin Woliński, and Robert Wołosz. *Słownik gramatyczny języka polskiego*. Wiedza Powszechna, Warsaw, 2007a.

Zygmunt Saloni, Włodzimierz Gruszczyński, Marcin Woliński, and Robert Wołosz. Grammatical dictionary of Polish. Presentation by the authors. *Studies in Polish Linguistics*, 4:5–25, 2007b.

Jan Tokarski. *Schematyczny indeks a tergo polskich form wyrazowych*, edited by Zygmunt Saloni. Wydawnictwo Naukowe PWN, Warszawa, 1993.

Zygmunt Vetulani and Tomasz Obrębski. Morphological tagging of texts using the lemmatizer of the 'POLEX' electronic dictionary. In Barbara Lewandowska-Tomaszczyk and Patrick James Melia, editors, *PALC'97: Practical Applications in Language Corpora*, pages 496–505, Łódź, 1997. Łódź University Press.

Marcin Woliński. Morfeusz — a practical tool for the morphological analysis of Polish. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 503–512. Springer-Verlag, Berlin, 2006.

Marcin Woliński. A relational model of Polish inflection in *Grammatical Dictionary of Polish*. In Zygmunt Vetulani and Hans Uszkoreit, editors, *Human Language Technology. Challenges of the Information Society. Third Language and Technology Conference, LTC 2007. Revised Selected Papers*, volume 5603 of *LNAI*, pages 96–106. Springer-Verlag, 2009.

[1] Initial work was concerned with the differentiation of surnames with adjectival inflection (e.g., *Jurecki*) from possessive adjectives (e.g., *Andersenowski* – 'belonging or related to Andersen'), which – according to Polish orthography – are capitalised.