# The Online Version of *Grammatical Dictionary of Polish*

**Marcin Woliński, Witold Kieraś**
Institute of Computer Science, Polish Academy of Sciences
Jana Kazimierza 5, 01-248 Warszawa, Poland
wolinski@ipipan.waw.pl

**Abstract**

We present the new online edition of a dictionary of Polish inflection – the *Grammatical Dictionary of Polish* (`http://sgjp.pl`). The dictionary is interesting for several reasons: it is comprehensive (over 330,000 lexemes corresponding to almost 4,300,000 different textual words; 1116 handcrafted inflectional patterns), the inflection is presented in an explicit manner in the form of carefully designed tables, the user interface facilitates advanced queries by several features (lemmas, forms, applicable grammatical categories, types of inflection). Moreover, the data of the dictionary is used in morphological analysers, including our product Morfeusz (`http://sgjp.pl/morfeusz`). From the very beginning, the dictionary was meant to be convenient for human reader as well as to be ready for use in NLP applications. In the paper we briefly discuss both aspects of the resource.

**Keywords:** inflectional dictionary, Polish, morphological analysis

## 1. About the Dictionary

The idea of the dictionary was conceived by its primary author Zygmunt Saloni some 35 years ago under the influence of a grammatical dictionary of Russian (Zalizniak, 1977). The project evolved slowly, first by analysing grammatical information in the largest dictionary of Polish printed on paper in 11 volumes (Doroszewski, 1958–1969). Important milestones were the new systematisation of Polish declension (Gruszczyński, 1989), the schematic reverse index of Polish word forms (Tokarski, 1993), the dictionary of Polish conjugation (Saloni, 2001). The mentioned works were published as paper books, but they were based on data prepared using databases. In particular, the verbal database became the seed for description of other parts of speech. The first and second edition of the *Grammatical Dictionary of Polish* (*Słownik gramatyczny języka polskiego*, SGJP) appeared in the form of a computer program (Saloni et al., 2007b; Saloni et al., 2012). At this stage, the team included Zygmunt Saloni, Marcin Woliński, Robert Wołosz, Włodzimierz Gruszczyński, and Danuta Skowrońska as the authors, with many formal and informal co-workers (listed at the project's web page). For the third edition we decided to change the form to a web application available `http://sgjp.pl` (Saloni et al., 2015).

## 2. The Scope of SGJP

SGJP covers the whole list of entries of the mentioned dictionary of Polish (Doroszewski, 1958–1969), including a whole range of archaic, obsolete, dialectal and otherwise stylistically marked words, since its extensive lexical basis goes back to even last decades of 18th century. On the other hand, numerous new words were added, including a significant number of proper names as well as some modern vocabulary collected during various linguistic investigations. As a result, the original lexical basis of Doroszewski's dictionary, estimated at ca. 130,000 lexemes, was significantly extended.

The number of lexemes of various grammatical classes is illustrated in Table 1.

The primary scope of interest in SGJP is inflection. Unlike in most other dictionaries, inflectional paradigms are presented explicitly in the form of tables containing all possible surface forms of the given lexeme. It is worth noting that Polish is a heavily inflected language – the paradigm for nouns consists of 14–16 forms, for adjectives the table comprises 23–25 cells, while a typical verb has 99 forms (including analytic ones) with some possible variations.

The dictionary provides no definitions. Short glosses suggesting meanings are used in case of homonymous or less known entries. Lexemes are defined by the identity of a paradigm, regardless of meanings. For example, only one lexeme PARA is considered since both its meanings ('vapour' and 'pair') have exactly the same inflectional forms. On the other hand, SGJP has three lexemes with the lemma PŁYWAK, since its 3 meanings ('swimmer', 'great diving beetle', and 'float') result in paradigms differing in the accusative.

Repertoire of grammatical categories and their values is based on the tradition of Polish grammar, but it also uses solutions proposed earlier by the members of the group (Saloni et al., 2007a). In particular, the dictionary uses the detailed system of 9 genders proposed by (Saloni, 1976) and some non-traditional grammatical categories including accomodability for numeral forms (Saloni, 1977) and depreciativity for masculine personal nouns (Bień and Saloni, 1982; Saloni, 1988).

An important feature of the system is that the description is two-level. At the first level, surface forms are described, i.e. all orthographic words are enumerated. For example, a typical adjective can appear in texts as 11 different words. Only at the second level grammatical features are attached to the forms. This provides for flexibility. In particular, we use a different second level when presenting the data to the reader of the dictionary and when generating data for morphological analysis. When the inflectional tables for the human reader are created, 11 forms of the adjective are distributed among 23 cells of the table (cf. Figure 1). On the other hand, in the data for morphological analysis the same forms are coupled with 106 distinct tags (resulting from combining 9 genders, 7 cases and 2 numbers).

| | 1st edition | | 3rd online edition | |
|---|---|---|---|---|
| | **entries** | **patterns** | **entries** | **patterns** |
| **total** | **244,669** | **1,095** | **334,845** | **1,116** |
| prefixes | 81 | 2 | 112 | 2 |
| lexemes | 244,588 | 1,095 | 334,733 | 1,116 |
| nominal inflection | 135,529 | 762 | 172,178 | 767 |
| personal pronouns | 6 | 6 | 6 | 6 |
| gerunds | 29,590 | 2 | 29,526 | 2 |
| deadjectival | 28,980 | 1 | 62,445 | 1 |
| regular nouns | 76,953 | | 80,201 | |
| proper names | 8,782 | | 10,710 | |
| common | 68,171 | | 69,491 | |
| adjectival inflection | 65,671 | 71 | 103,761 | 77 |
| participles | 34,301 | | 36,304 | |
| active | 13,931 | 1 | 13,877 | 1 |
| passive | 20,370 | | 22,427 | 4 |
| regular adjectives | 31,370 | 71 | 67,457 | 77 |
| comparative | 950 | 1 | 1065 | 1 |
| positive | 30,420 | 70 | 66,392 | 77 |
| deadjectival adverbs | 11,146 | 1 | 26,577 | 1 |
| comparative | 1,106 | 1 | 1,243 | 1 |
| positive | 10,040 | 1 | 25,512 | 1 |
| numerals | 98 | 45 | 117 | 47 |
| verbs | 29,532 | 215 | 29,955 | 222 |
| predicatives | 35 | 1 | 30 | 1 |
| conjugated | 29,497 | 214 | 29,925 | 221 |
| other | 2,612 | 2 | 1,967 | 2 |
| other adverbs | 491 | 1 | 502 | 1 |
| particles | 193 | 1 | 201 | 1 |
| prepositions | 113 | 2 | 118 | 2 |
| conjunctions & complementizers | 121 | 1 | 121 | 1 |
| interjections | 458 | 1 | 490 | 1 |
| abbreviations | 1,117 | 1 | 392 | 17 |
| other | 119 | 1 | 143 | 1 |

Table 1: Numbers of entries of various grammatical classes in SGJP. The most noticeable change between 1st and 3rd edition concerns significant increase in the number of deadjectival nouns and positive adjectives and adverbs, which was caused by automatic generation of their negated forms by adding prefix *nie-*.

Obviously, the dictionary does not contain all Polish lexemes. However, we hope that almost all inflectional patterns for Polish have already been identified. This claim can be backed by the work on Polimorf (Woliński et al., 2012), where the data of SGJP was merged with a community-built dictionary of similar size (`sjp.pl`). In the process, lexemes of the other dictionary were matched against SGJP's inflectional patterns. This process required adding less than 10 patterns to the system (mainly for proper names).

Besides purely inflectional features, the dictionary notes case government of prepositions; it includes gender for nouns; categorises numeral forms with respect to their relation with nouns (agreement or government); for verbs it provides information on the aspect (perfective/imperfective), transitivity, co-occurrence with the reflexive marker *się* (obligatory or optional). Case government of verbs is not generally noted – we delegate this feature to a specific valency dictionary, e.g. (Przepiórkowski et al., 2014). SGJP includes information on selected highly regular derivational

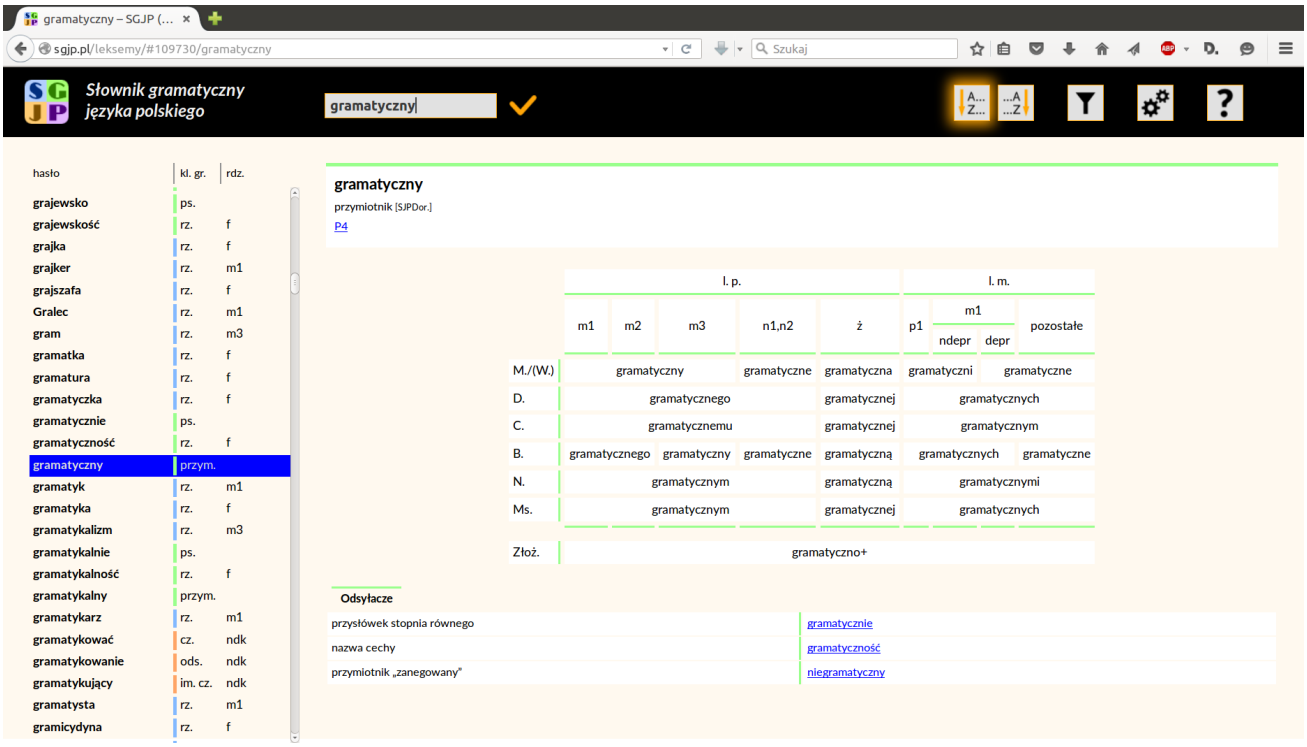


Figure 1: Interface of the dictionary showing inflection of the adjective GRAMATYCZNY *'grammatical'*. The left panel shows lemmas with parts of speech ('rz.' for noun, 'przym.' for adjective, 'cz.' for verb, etc.) and selected features (gender for nouns, aspect for verbs). The right panel displays the given lexeme. This headword is characterised as an adjective ('przymiotnik') inflecting according to pattern P4. The table is organised by features characteristic for the given part of speech. In the case of an adjective these are: grammatical case (rows), number and gender (columns). Some cells span multiple columns to increase readability. The table includes also a special form used in compounds ('Złoż.'). The section below includes derivational links ('Odsyłacze') to the adverb, noun naming the feature, and antonym.

relations, in particular between verbs and their nominal and adjectival derivatives (gerunds and participles), between adjectives and adverbs or nominal names of qualities, between positive and comparative adjectives (superlative adjectives are derived implicitly from comparative ones).

## 3. Online Edition of SGJP

The interface of the online edition of SGJP has been realized as a JavaScript internet application (see Figure 1) backed by a database on the server side. When designing the application we strove to provide a user experience close to that of a desktop application.

In particular, to display the list of entries (left part of the screen in Figure 1) we use the SlickGrid JavaScript component that allows to create the illusion that the full list of entries is displayed directly in the browser. In reality, the component queries the database selectively and only loads these pieces of the list that are needed for the portion visible on the screen. This interface was meant for the advanced users of the dictionary, who are expected to compare several lexemes of interest. In such applications pagination is not a comfortable option because page boundaries often interfere with what the user wants to see.

Two sorting orders are available – the usual alphabetical order of headwords (cf. Figure 1), as well as the order of reversed headwords (cf. Figure 2). The latter is not often seen in electronic dictionaries, although there existed reversed indices for some printed dictionaries (for example the index for Doroszewski's dictionary (Grzegorczykowa and Puzynina, 1973)). This order causes entries that end in a similar way to appear together, allowing to observe similarities in their inflection.

Entries can be searched for by any form belonging to their paradigm. We think this feature can be very useful for foreigners learning Polish, since Polish inflection is often non-obvious with (rare) extremities such as forms not having any prefix common with the lemma (e.g., the noun DECH *'breath'* has *tchem* as one of its forms). The forms matching query are highlighted in the resulting inflectional table.

The new feature of the online edition is filtering. Filtering criteria can reference the following features: the headword, any inflected form, part of speech, name of inflectional pattern used, labels, types of proper names, gender, aspect, reflexivity. It is also possible to filter by the number of different patterns or genders assigned to a lexeme. Conditions concerning headwords, forms, and pattern names can reference their parts or be specified using regular expressions. For example it is possible to find all lexemes having inflected forms with particular endings.

A new systematic classification of inflectional patterns has been prepared for the third edition, resulting in particular in renaming all patterns. New names have internal structure: the beginning specifies a rough inflectional group, further characterised by the next parts of the name. This can be

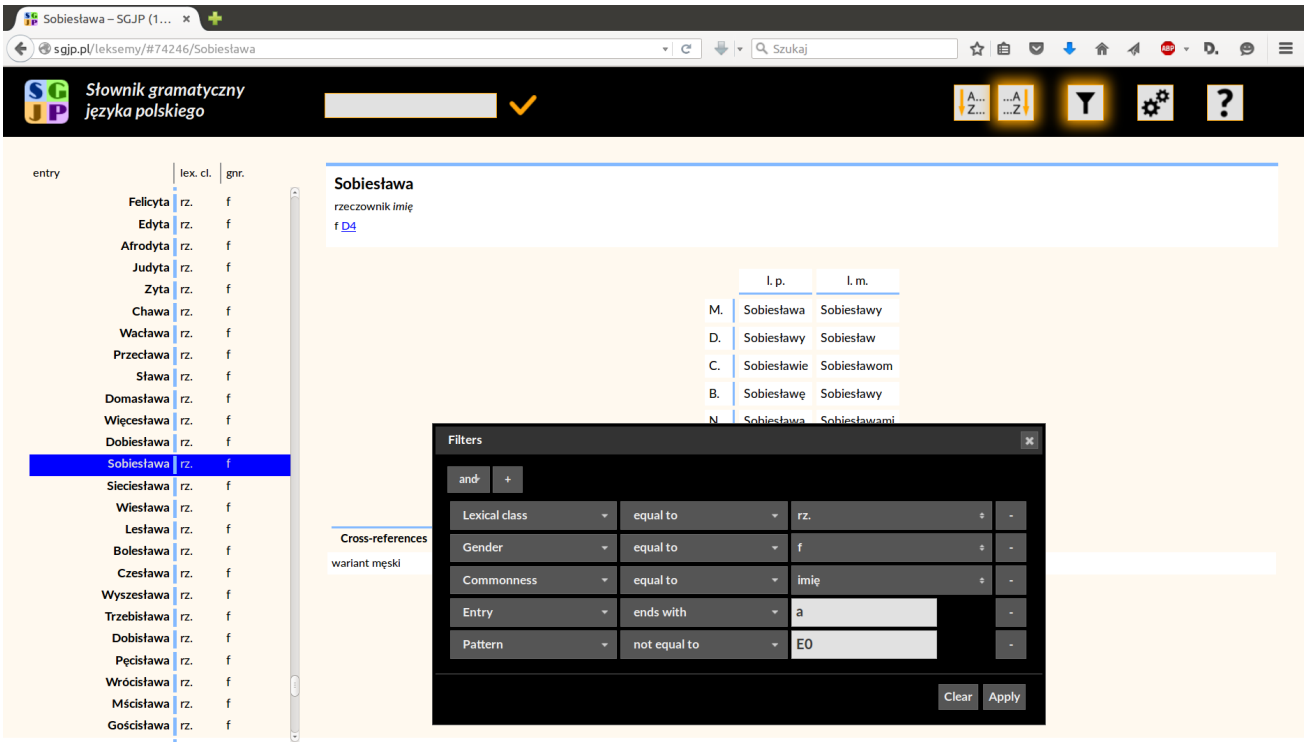


Figure 2: The result of filtering feminine first names shown sorted by reversed headwords. Using filters it is easy to verify, e.g., that Polish feminine first names inflect if and only if they end in an *-a*.

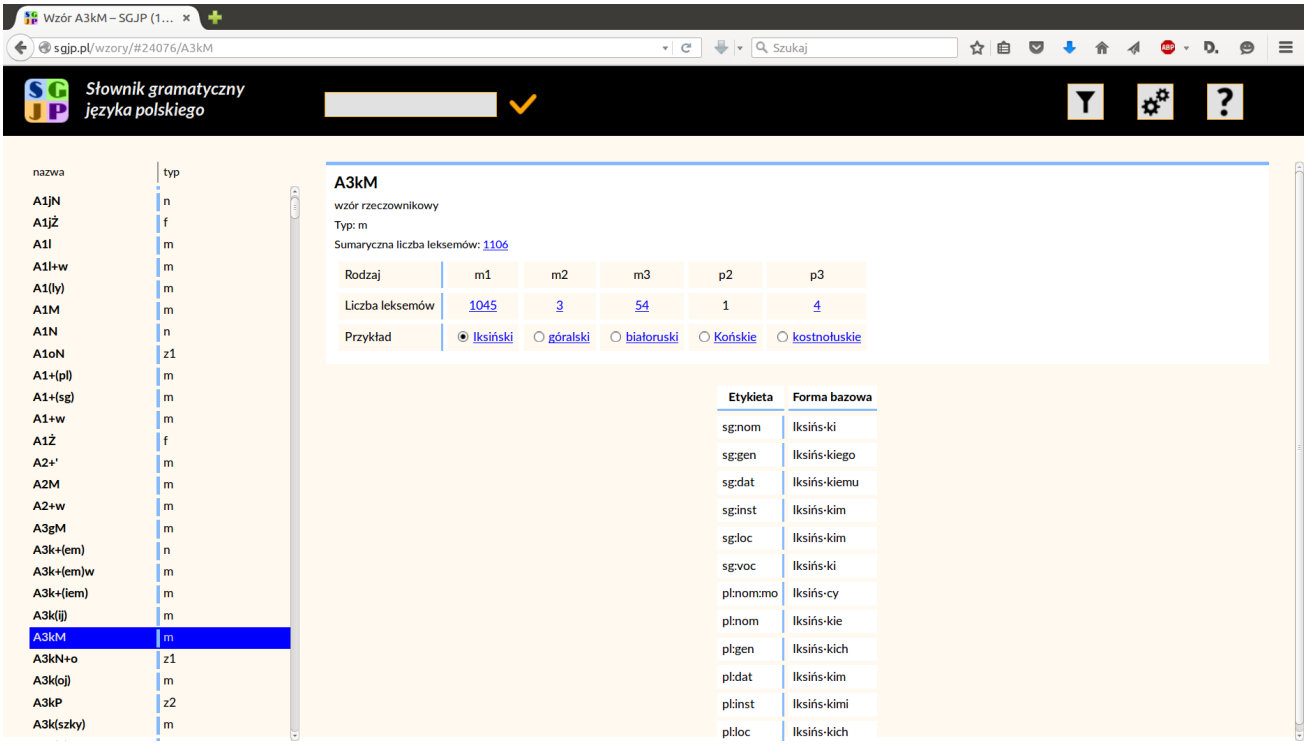


Figure 3: The view of inflectional patterns in SGJP showing the pattern named A3kM. The left panel lists pattern names with their inflectional types (for nominal patterns the type can be: uninflected (0), masculine (m), feminine (f), neuter (n), pronominal (z1, z2)). The right panel shows the given pattern in detail. The A3kM pattern gets characterised as a nominal pattern ('wzór rzeczownikowy') of masculine inflectional type. Total number of lexemes using this pattern is 1106. Subtotals are provided for each gender (the masculine inflectional type does not limit the pattern for use only with masculine genders). Here example lexemes of 5 genders are shown. The forms generated by the pattern are shown below.

used in filtering to compare lexemes using similar patterns or study variation within rough groups.
Patterns are visualised in a dedicated view (cf. Figure 3), which is new in this edition of the dictionary. The view shows how a pattern works using an example and how many lexemes inflect according to the given pattern. For nouns the count is provided separately for each gender of nouns using this pattern. Links provide a way to easily filter the lexemes using the given pattern. The view of patterns can also be filtered.
Inflectional patterns in SGJP describe forms in terms of a stem common to all forms and endings differentiating the forms. In Figure 3 these elements are separated with a dot: *Iksińs·ki*. Due to the extensive irregularity in Polish inflection the number of patterns used in the dictionary is high – see Table 1.

## 4. Dictionary Editor's Interface

The most important change is not visible to the end users. Up to version 2. the data of SGJP was maintained as a set of Microsoft Access database files. Each of the files contained entries of one part of speech and was manipulated (at any given time) by only one of the authors. This organisation of work crystallised in the 1990s, when the authors did not yet have constant access to the Internet.
When creating the online version we merged and converted all data to a common format (Woliński, 2009). Currently the data is hosted on a server and editors of the dictionary use a web-based interface. This has the obvious advantage of allowing editors to work simultaneously and simplifies maintenance of data considerably. Moreover, accepted changes in the data are automatically and immediately visible in the online version.
The editor's interface includes several features targeted at guiding the editors. The most important of them is a tool that suggests inflectional patterns for a new lexeme based on the lexemes already present in the dictionary. The tool displays a list of matching patterns sorted by similarity of lexemes that use them to the one in question (see Figure 4). Usually, since the dictionary is already very rich, the first suggestion is correct.
For example, Figure 4 shows suggestions for the lexeme PREKARIAT 'precariat'. The longest common suffix with a noun of the same gender (m3) present in the dictionary is *kariat* for the lexeme WIKARIAT 'curacy' using the pattern B4t+u. The result of applying this pattern to the lexeme `prekariat` is shown on the right, which allows to verify that these are indeed correct inflectional forms. The next rows of the table show other matching patterns. The pattern B4ta+u of the second row would result in the form *prekariecie* for the locative and the vocative. The pattern B4ta+(u) would result in *prekariata* as the genitive and so on.

## 5. Morfeusz SGJP

SGJP is used as a source of data for the inflectional analyser Morfeusz (Woliński, 2014; Woliński, 2006), a tool commonly used by the Polish NLP community.
The list of forms needed for Morfeusz gets periodically generated by the server hosting SGJP. Then a binary compiled

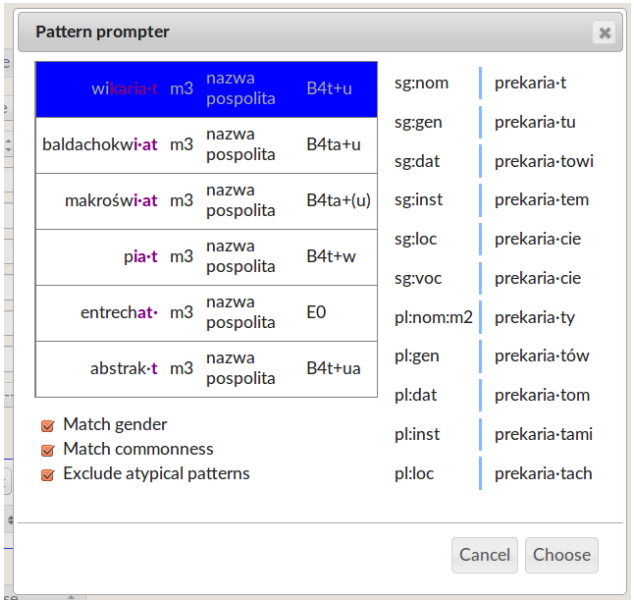


Figure 4: Tool suggesting inflectional patterns for new entries

dictionary is generated and ready to install packages for several operating systems are built. The process is completely automatic and is triggered each week, provided any changes were introduced in the data. This way up-to-date versions of the analyser are easily available and the time between introducing a change in the dictionary and its visibility in the analyser is very short.
The tool developed for the maintenance of SGJP facilitates work on multiple dictionaries, which allows to create domain-specific dictionaries that are separate from SGJP proper. Since SGJP is a general dictionary of Polish, we do not intend to extend it with, e.g., medical terminology. However, it is easy to create a separate 'medical' dictionary that will get exported as data for Morfeusz together with the basic dictionary.
Morfeusz and its SGJP-based dictionary are distributed under the liberal two clause BSD license. In particular, the clear text form of the list of all inflected words with Morfeusz tags is made available. The licence allows for any use of the resources, commercial or otherwise, provided the authorship of the resource is acknowledged.

## 6. Conclusions and future work

We have reported on changes in the dictionary that is used as an important resource for many NLP applications involving Polish. The resource – both lexical data and the model of Polish inflection – is being used in several other applications. The new *Great Dictionary of Polish* (*Wielki słownik języka polskiego*, see: `http://www.wsjp.pl`), that is currently under preparation, imports its grammatical information directly from SGJP. Moreover, a dictionary of XIX century Polish is being developed based on the research paradigm and tools created for SGJP (Derwojedowa et al., 2014). The whole inflectional model of SGJP is relatively easy to adapt for other applications concerning Polish inflection, both historical and contemporary. For example, a possible extensive inflectional dictionary of proper names could

be developed in the same manner and with the same tools, which would significantly enhance morphological analysis used in many NLP projects for Polish.

The new version of the dictionary has a modern web-based interface and interesting new features for advanced users. But, more importantly, the data of the dictionary has been reorganised and a dedicated tool has been implemented to ease corrections in the dictionary and ensure smooth further development. Last but not least, transforming SGJP from a desktop electronic dictionary to a full featured web-based application also increased the number of its possible users, including students and language enthusiasts outside of academia.

## 7. Acknowledgements

This work has been financed by the Polish National Science Centre grants 2011/01/B/HS2/04695 and 2014/15/B/HS2/03119.

## 8. Bibliographical References

Bień, J. S. and Saloni, Z. (1982). Pojęcie wyrazu morfologicznego i jego zastosowanie do opisu fleksji polskiej (wersja wstępna). *Prace Filologiczne*, XXXI:31–45.

Nicoletta Calzolari, et al., editors. (2014). *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, Reykjavík, Iceland. ELRA.

Derwojedowa, M., Kieraś, W., Skowrońska, D., and Wołosz, R. (2014). Współczesne narzędzia leksykograficzne a analiza tekstów dawniejszych. *Polonica*, XXXIV:21–27.

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

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

Renata Grzegorczykowa et al., editors. (1973). *Indeks a tergo do Słownika języka polskiego pod redakcją Witolda Doroszewskiego*. PWN, Warszawa.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., Woliński, M., Skwarski, F., and Świdziński, M. (2014). Walenty: Towards a comprehensive valence dictionary of Polish. In Calzolari et al. (Calzolari et al., 2014), pages 2785–2792.

Saloni, Z., Gruszczyński, W., Woliński, M., and Wołosz, R. (2007a). Grammatical dictionary of Polish. presentation by the authors. *Studies in Polish Linguistics*, 4:5–25.

Saloni, Z., Gruszczyński, W., Woliński, M., and Wołosz, R. (2007b). *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., Woliński, M., Wołosz, R., Gruszczyński, W., and Skowrońska, D. (2015). *Słownik gramatyczny języka polskiego*. `http://sgjp.pl`, 3 edition.

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

Saloni, Z. (1977). Kategorie gramatyczne liczebników we współczesnym języku polskim. In *Studia gramatyczne I*, pages 145–173. Wrocław.

Saloni, Z. (1988). O tzw. formach nieosobowych [rzeczowników] męskoosobowych we współczesnej polszczyźnie. *Biuletyn Polskiego Towarzystwa Językoznawczego*, XLI:155–166.

Saloni, Z. (2001). *Czasownik polski. Odmiana, słownik*. Wiedza Powszechna, Warszawa.

Tokarski, J. (1993). *Schematyczny indeks a tergo polskich form wyrazowych*, red. Zygmunt Saloni. Wydawnictwo Naukowe PWN, Warszawa.

Woliński, M., Miłkowski, M., Ogrodniczuk, M., Przepiórkowski, A., and Szałkiewicz, 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. (2006). Morfeusz — a practical tool for the morphological analysis of Polish. In Mieczysław Kłopotek, et al., editors, *Intelligent Information Processing and Web Mining, IIS:IIPWM'06 Proceedings*, pages 503–512. Springer.

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

Woliński, M. (2014). Morfeusz reloaded. In Calzolari et al. (Calzolari et al., 2014), pages 1106–1111.

Zalizniak, A. (1977). *Grammaticheskij slovar' russkogo yazyka*. Russkij yazyk, Moscow, 1 edition.

## 9. Language Resource References

ZIL IPI PAN. (2014). *Morfeusz 2*. `http://sgjp.pl/morfeusz`, version 2.0.

Zygmunt Saloni and Marcin Woliński and Robert Wołosz and Włodzimierz Gruszczyński and Danuta Skowrońska. (2015). *Grammatical Dictionary of Polish*. `http://sgjp.pl`, version 3.