

A relational model of Polish inflection

Marcin Woliński

Instytut Podstaw Informatyki PAN
ul. Ordona 21, 01-237 Warszawa, Poland
wolinski@ipipan.waw.pl

Abstract

The subject of this article is a description of Polish inflection in the form of a relational database. The description has been developed for a grammatical dictionary of Polish that aims at complete inflectional characterisation of all Polish lexemes. We show that all subtleties of Polish inflection can be accounted for with a uniform and relatively compact relational model.

Introduction

The grammatical dictionary of Polish (Saloni *et al.* (in print), henceforth: SGJP) aims at providing a description of inflection as complete as possible. Although the dictionary is large (about 180,000 lexemes, 3,600,000 orthographic words), it does not of course include all Polish words. It is hoped, however, that the dictionary includes all possible inflectional patterns for all inflecting lexemes of Polish.

The idea of SGJP was conceived by Saloni under the influence of Zalizniak's grammatical dictionary of Russian (Zalizniak, 1977). SGJP is based on numerous earlier works: Tokarski's and Saloni's work on inflection (Tokarski, 1993), Saloni's description of Polish verbs (Saloni, 2001), Gruszczyński's description of Polish nouns (Gruszczyński, 1989), Wołosz's morphological data (Wołosz, 2005) and some other.

The scope of the dictionary is complete morphological characterisation and basic syntactic characterisation of Polish words. This means that for each lexeme all its inflected forms are given with values of all morphological categories (categories for which given lexeme inflects). Moreover values of some syntactic features are provided: gender for nouns, aspect for verbs, (required) case for prepositions. The dictionary also contains some links between lexemes, e.g., between elements of aspectual pairs for verbs, between a verb and its nominal derivatives (gerund and participles), between adjectives and adverbs derived from them, between positive and comparative adjectives.

In SGJP the inflection is described according to theoretical decisions set in the above-mentioned works. Due to the large amount of data involved SGJP is being worked on using relational database machinery. A question arises whether it is possible to model the inflection according to the set rules within the relational model. Such a possibility would mean the whole work on the dictionary could be done with database tools alone. Otherwise the database would be merely a means of storage while some other facilities would be needed, e.g., to generate all inflected forms from dictionary data.

In the following, we give an affirmative answer to this question and present a relational model of SGJP (we assume some basic knowledge of relational modelling from the reader).

1. The model

For the user, SGJP has the form of a dedicated interface programme (cf. fig. 1). The data in the backend is represented as a relational database—the programme communicates with the database in SQL. In this section we will briefly present the schema of this database.

The central entity is `Lexeme` (cf. fig. 2). It is the basic unit of description in the dictionary.¹ Its attributes include a numerical `ID` (which is the primary key), the `lemma` (base form) and a homonym number `hom` (these two attributes form an alternate key). Other attributes are the grammatical class (part of speech) `pos` and several attributes needed to present the entry to the user: labels, glosses, comments, source, and so on. Although the attributes are quite numerous, the most important for the present article are the `ID` and `pos`.

In the works of Tokarski, Saloni, and Gruszczyński inflection is described by means of inflectional patterns. Consider all inflected forms of a Polish lexeme. In almost all cases, all the forms share a common first part (and in the remaining cases we assume this part to be empty). We shall call this common part a *stem* and the rest of each form an *ending*.² To describe inflection of a lexeme we have to state its stem and the inflectional pattern containing all respective endings.

Inflectional patterns are modelled here with entities `Pattern` and `Ending`. Patterns are identified with `patternIDs` and classified with the attribute `pat_type`, whose role will be explained in section 3. For each instance of `Pattern` we have several instances of `Ending` with the respective endings as the value of attribute `ending`.

For the sake of compactness `Endings` describe only what we call *basic inflected forms*. Other inflected forms are generated from the basic ones in a way which does not depend on the lexeme's inflectional pattern, but depends on the grammatical class, and so will be discussed for each class separately in the following sections. The mapping between basic inflected forms and the remaining ones is provided by the entity `Form`.

¹A `Lexeme` may be represented by several entries in the list of entries of the dictionary. In one of the user-selectable views of the interface all inflected forms are used as entries.

²Quite commonly these parts are not what could be called a stem or an ending from the morphological point of view.

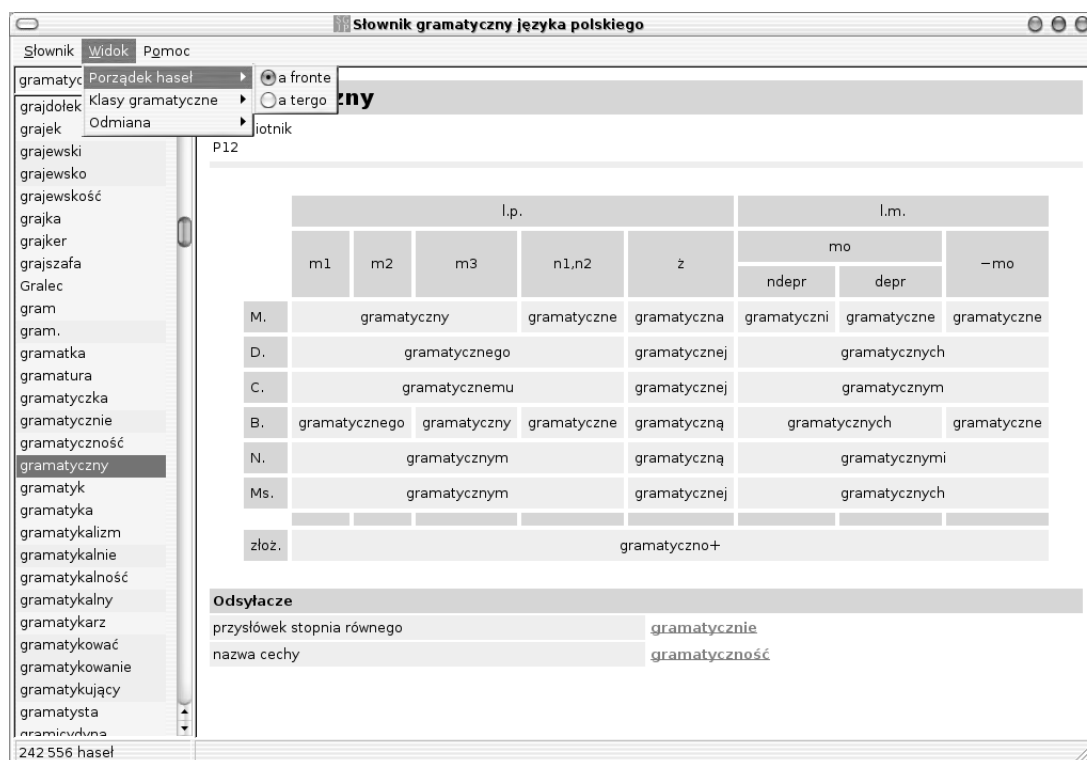


Figure 1: The interface of the grammatical dictionary

Sometimes it is necessary to assign more than one Pattern to a Lexeme. For that reason `patternID` is not an attribute of `Lexeme`. A separate entity `Inflection` models the many-to-many relationship between `Lexemes` and `Patterns`.

Entity `Link` is used to represent the links between `Lexemes` mentioned in the introduction. The attribute `link_type` describes type of the relation modelled by the given `Link`.

2. Adjectives

We start with presentation of adjectives since they provide the simplest example of distinction between basic inflected forms and inflected forms.

Adjectives in SGJP are described according to the principles set in the work of Tokarski (1993). A typical Polish adjective can be realised (represented) in texts by any of 11 shapes (orthographic words). However, if we try to attach the values of case, number, and gender to these shapes we end up with $7 \times 2 \times 9 = 126$ combinations.³

To make this plethora of forms manageable, inflectional patterns for adjectives describe only 11 basic inflected forms whose basic form tag `bafotag` is a number from 1 to 11. So for each adjectival `Pattern` the entity `Ending` has 11 instances.⁴

³A rather detailed system of 9 genders for Polish is used. It includes masculine personal m1 (e.g., *profesor*), animal m2 (*pies*), inanimate m3 (*stół*); neuter n1 (*dziecko*), n2 (*okno*); feminine f (*wanna*); plurale tantum p1 (*państwo*), p2 (*drzwi*), and p3 (*spodnie*) (Mańczak, 1956; Saloni, 1976).

⁴In fact up to 4 more basic forms are used to account for some additional forms present only for some adjectives.

These basic forms are mapped to actual inflected forms by the instances of the entity `Form`. For example, for any adjectival pattern the basic form with `bafotag` of 2 (e.g., *białego*) can be interpreted as genitive singular of any masculine or neuter gender or accusative singular of masculine personal or masculine animal genders:

pos	bafotag	tag
adj	2	sg:gen:m1
adj	2	sg:gen:m2
adj	2	sg:gen:m3
adj	2	sg:gen:n1
adj	2	sg:gen:n2
adj	2	sg:acc:m1
adj	2	sg:acc:m2

This mapping is universal to all adjectival patterns.

Some additional flexibility is present in this scheme thanks to the use of the mapping attribute of `Form`. This attribute selects which of the various presentations of forms is delivered to the user. This applies to lexemes of all grammatical classes. We use three mappings in SGJP. The one presented above is used to show all inflected forms to the user. In fact, the tags used in this mapping are different than those shown in the table above since, e.g., the inflection tables for adjectives are presented in a compacted form and do not have 126 cells. The second mapping includes only basic inflected forms, so it shows the real representation of inflectional patterns to the user. The third one is used to generate all forms necessary for searching in the dictionary. This includes some forms which are not normally displayed, but to which the program should react when typed in the search window (e.g., negated gerunds). One more mapping is used to convert the dictionary to

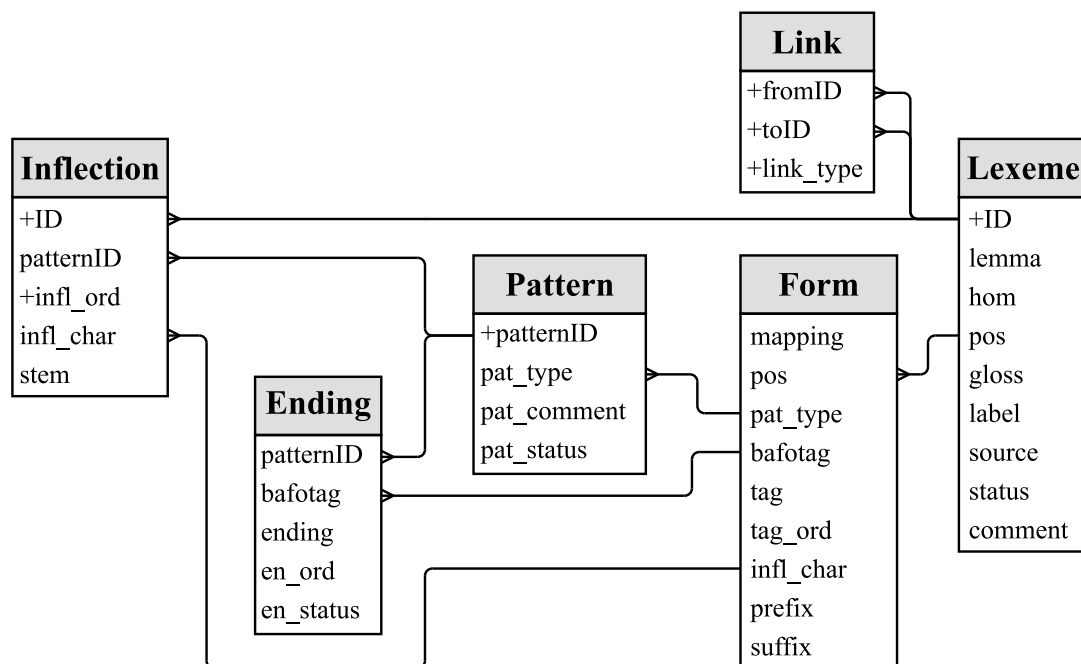


Figure 2: The schema of the dictionary database (slightly simplified)

the form used by the morphological analyser *Morfeusz* (Woliński, 2006).

The key point of these remarks is that providing another view of data is rather trivial since the interface of SGJP is driven by the contents of the table `Form`.

3. Nouns

Inflection of Polish nouns is described in SGJP in a more complicated way. Inflectional patterns for nouns constructed according to the simple stem-ending rule would be very numerous. However, some of them differ in a very regular manner. For example we can find triples of nouns of all masculine genders which differ only in forms of the accusative case and presence of a special form of nominative plural for masculine personal nouns. This applies for example to nouns *pletwonurek* m1, *skowronek* m2, and *naglówek* m3. The following rule works for all masculine Polish nouns: accusative singular is equal to genitive for gender m1 and m2, and to nominative for m3; accusative plural is equal to genitive for gender m1 and to nominative for m2 and m3. It makes sense to have only one inflectional pattern for these lexemes (Gruszczyński, 2001; Gruszczyński and Saloni, 2006). For that reason accusative forms are not included in the set of basic inflected forms for nouns. The right form of accusative is created depending on the value of `infl_char` attribute, which for nouns carries the value of gender.

The next complication is introduced by masculine personal nouns, which have two possible forms of nominative plural differentiated with the value of depreciativity (Saloni, 1988). For example nominative plural of the noun *pletwonurek* has a neutral variant *pletwonurkowie* and a stylistically marked (depreciative) variant *pletwonurki*. The only possible form for gender m2 and m3 has the same ending as the depreciative form for m1. For that reason we

have two basic inflected forms for nominative plural. Both are used for masculine personal nouns, and only the second for the remaining masculine genders.

Unfortunately the above remarks do not apply to feminine nouns. Those have a specific form of singular accusative which has to be noted explicitly. Moreover, in Polish we have some masculine nouns which inflect in a way so similar to feminine nouns that it makes sense to have a common pattern for the two (e.g., *poeta* m1 ‘poet’ inflects almost exactly in the same way as *kobieta* f ‘woman’). Moreover for some feminine nouns we need to account for two regular variants of genitive plural (e.g., *funkcyj/funkcyj*).

Yet another set of basic inflected forms is needed for neuter nouns, since for them accusative and vocative is always equal to nominative in both numbers. Also there exist masculine (personal) nouns which inflect similarly to neuter nouns (e.g., the p2 *plurale tantum* noun *mistrzostwa* has the same forms as *dynamo* n2 in plural).

To account for these phenomena we introduce types of inflectional patterns differentiated with the attribute `pat_type` of `Pattern`. This attribute together with the gender contained in the `infl_char` of a given `Inflection` selects the right instance of `Form`. Three pattern types have been introduced for masculine, feminine, and neuter type of inflection (not gender). One more type is used for non-inflecting nouns, which have just one shape used for all grammatical forms. The following table lists instances of `Form` for singular accusative forms of various pattern types:

pos	pat_type	infl_char	bafotag	tag
subst	m	m1	sg:gen	sg:acc
subst	m	m2	sg:gen	sg:acc
subst	m	m3	sg:nom	sg:acc
subst	f	any	sg:acc	sg:acc
subst	n	m1	sg:gen	sg:acc
subst	n	m2	sg:gen	sg:acc
subst	n	m3	sg:nom	sg:acc
subst	n	n1	sg:nom	sg:acc
subst	n	n2	sg:nom	sg:acc
subst	0	any	lemma	sg:acc

The complete list of bafotags used for various types of inflection pat_type is as follows:

m	f	n
sg:nom	sg:nom	sg:nom
sg:gen	sg:gen	sg:gen
sg:dat	sg:dat	sg:dat
	sg:acc	
sg:inst	sg:inst	sg:inst
sg:loc		sg:loc
sg:voc	sg:voc	
pl:nom:m1	pl:nom:m1	pl:nom
pl:nom:m2	pl:nom:m2	
pl:gen	pl:gen:funi	pl:gen:m
	pl:gen:fnuni	pl:gen:n
	pl:gen:m	
pl:dat	pl:dat	pl:dat
pl:inst	pl:inst	pl:inst
pl:loc	pl:loc	pl:loc

A word of explanation is due as for why `infl_char` is an attribute of `Inflection` and not of `Lexeme`. There are nouns in Polish whose gender is not stable. For example the noun *człowieczyisko* can be reasonably included both in the m1 and n2 class. Similarly *cabernet* can be m2, m3, or n2. In this case we choose to have one `Lexeme` with multiple `Inflections` differing in gender. Of course for regular homonyms (e.g., *bokser* m1 ‘boxer (athlete)’, m2 ‘bulldog’, and m3 ‘type of engine’) `SGJP` has separate `Lexemes`.

4. Verbs

The main feature which determines the set of forms of a typical Polish verb is its aspect. Present tense in indicative mood, adverbial simultaneous participle, and adjectival active participle are specific to imperfective verbs. Perfective verbs form simple future tense and adverbial anterior participle. For that reason in our model aspect is kept in the `infl_char` attribute for verbs.

Verbal forms are very numerous (and, not like in the case of adjectives, this means numerous different orthographic words). Fortunately they can be easily derived from twelve basic inflected forms (Saloni, 2000). For example the basic form denoted with `bafotag` of 10 is used to create the impersonal past form (e.g., *wiedzion-o*) and

all forms of the passive adjectival participle except for m1 nominative plural (e.g., *wiedzion-y*, *wiedzion-e*, *wiedzion-a*, ..., *wiedzion-ych*). We construct verbal forms from the stem specific for a `Lexeme`, ending specific for the basic inflected form, and `suffix`⁵ characteristic for a `Form`. For the verb *wieść* used above the stem is *wi-*, the ending 10 is *-edzion-*, and the suffixes are marked in the above examples.

Moreover some forms, namely the negated gerunds, are formed by prepending the prefix *nie* to the affirmative forms. For that purpose we use the attribute `prefix` of `Form`. The following table lists `Forms` derived from basic inflected form 10:

pos	infl_char	bafotag	tag	prefix	suffix
v	any	10	imps		<i>o</i>
v	any	10	ppas:sg:nom:m1:aff		<i>y</i>
v	any	10	ppas:sg:nom:m2:aff		<i>y</i>
v	any	10	ppas:sg:nom:m3:aff		<i>y</i>
v	any	10	ppas:sg:nom:n1:aff		<i>e</i>
v	any	10	ppas:sg:nom:n2:aff		<i>e</i>
v	any	10	ppas:sg:nom:f:aff		<i>a</i>
			...		
v	any	10	ppas:pl:loc:p3:aff		<i>ych</i>
v	any	10	ppas:sg:nom:m1:neg	<i>nie</i>	<i>y</i>
v	any	10	ppas:sg:nom:m2:neg	<i>nie</i>	<i>y</i>
v	any	10	ppas:sg:nom:m3:neg	<i>nie</i>	<i>y</i>
v	any	10	ppas:sg:nom:n1:neg	<i>nie</i>	<i>e</i>
v	any	10	ppas:sg:nom:n2:neg	<i>nie</i>	<i>e</i>
v	any	10	ppas:sg:nom:f:neg	<i>nie</i>	<i>a</i>
			...		
v	any	10	ppas:pl:loc:p3:neg	<i>nie</i>	<i>ych</i>

The `prefix` and `suffix` is empty for other classes except for superlative degree of adjectives which is formed with `prefix naj`.

The class of verbs includes some ones with very non-typical inflection. These include verbs like *powinien* which has very limited set of forms as well as pseudo-verbs which do not inflect for person (*braknie*, *warto*). For these groups separate pattern types have been introduced.

5. Other grammatical classes

The class of numerals is very small, only 92 `Lexemes` in `SGJP`, but very irregular. It includes numerals which do not inflect at all (e.g., *pół*), those which inflect only for gender (e.g., *półtora*), those inflecting for gender and case, and finally those which inflect for gender, case, and the category of accomodability which specifies whether given form agrees with the noun (e.g., in the phrase *dwaj chłopcy*) or requires the noun to be in genitive (*dwóch chłopców*) (Saloni, 1977).

Inflectional patterns for numerals of each of these four groups belong to a separate pattern type in our description.

⁵The terms *prefix* and *suffix* are used here in the technical meaning of an arbitrary first (respectively last) part of a string.

Lexemes traditionally described as nominal, adjectival, adverbial, and numeral pronouns are treated in SGJP as regular nouns, adjectives, adverbs, and numerals (Saloni, 1974). The class of pronouns is limited to personal pronouns (including the reflexive pronoun *się*). These lexemes cannot be treated as nouns since they have no value of gender assigned. Moreover some of them have special forms depending on whether they appear on an accented position in the sentence (e.g., *ciebie* vs. *cię*) or whether they appear after a preposition (e.g., *jego* vs. *niego*). We need three separate pattern types to describe Polish personal pronouns.

The dictionary lists as well non-inflecting lexemes, which is rather trivial. An interesting point is however that some prepositions have two forms depending on phonological features of the following word (e.g., *w* and *we*). We obviously use a dedicated inflectional pattern for these lexemes.

6. Conclusion

It may seem that with the given simple construction of inflectional patterns the description of morphology is almost trivial. However, numerous issues which need to be taken into the account show that this is not the case.

In general an inflected form consists of four parts: prefix, stem, ending, and suffix controlled by several entities of the model. Each of these parts can be empty.

The following table lists numbers of Patterns for various grammatical classes in the dictionary:

adjectives	71
nouns	744
verbs	214
numerals	45
pronouns	6

Due to irregularities in Polish inflection there exist Patterns which are needed for only one Lexeme.

The presented model covers all inflectional phenomena accounted for in SGJP. The model features a rather dense net of relations but still is rather compact and manageable. In particular it provides a unified method of generating forms of a lexeme of any grammatical class although the inflectional patterns for particular classes are constructed in a substantially different way.

7. References

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

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

Gruszczyński, Włodzimierz, 2001. Rzeczowniki w słowniku gramatycznym współczesnego języka polskiego. In W. Gruszczyński, U. Andrejewicz, M. Bańko, and D. Kopcińska (eds.), *Nie bez znaczenia... Prace ofiarowane Profesorowi Zygmuntowi Saloniu z okazji jubileuszu 15000 dni pracy naukowej*. Białystok, pages 99–116.

Gruszczyński, Włodzimierz and Zygmunt Saloni, 1978. Składnia grup liczebnikowych we współczesnym języku polskim. In *Studia gramatyczne II*. Wrocław, pages 17–42.

Gruszczyński, Włodzimierz and Zygmunt Saloni, 2006. Notowanie informacji o odmianie rzeczowników w projektowanym *Słowniku gramatycznym języka polskiego*. In I. Bobrowski and K. Kowalik (eds.), *Od fonemu do zdania. Prace dedykowane Profesorowi Romanowi Laskowskiemu*. Kraków, pages 203–213.

Mańczak, W., 1956. Ile rodzajów jest w polskim? *Język Polski*, XXXVI(2):116–121.

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

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

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

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

Saloni, Zygmunt, 2000. *Wstęp do koniugacji polskiej*. Olsztyn: Wydawnictwo Uniwersytetu Warmińsko-Mazurskiego.

Saloni, Zygmunt, 2001. *Czasownik polski. Odmiana, słownik*. Warszawa: Wiedza Powszechna (3rd ed. — 2007).

Saloni, Zygmunt, Włodzimierz Gruszczyński, Marcin Woliński, and Robert Wołosz, (in print). *Słownik gramatyczny języka polskiego*. Warszawa: Wiedza Powszechna.

Saloni, Zygmunt and Marcin Woliński, 2003. A computerized description of Polish conjugation. In Peter Kosta, Joanna Błaszczak, Jens Frasek, Ljudmila Geist, and Marzena Żygis (eds.), *Investigations into Formal Slavic Linguistics (Contributions of the Fourth European Conference on Formal Description on Slavic Languages)*.

Tokarski, Jan, 1973. *Fleksja polska*. Wydawnictwo Naukowe PWN (2nd ed. — 2002).

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

Woliński, Marcin, 2006. Morfeusz — a practical tool for the morphological analysis of Polish. In Mieczysław Kłopotek, Sławomir Wierzchoń, and Krzysztof Trojanowski (eds.), *Intelligent Information Processing and Web Mining, IIS:IIPWM'06 Proceedings*. Springer.

Wołosz, Robert, 2005. *Efektywna metoda analizy i syntezy morfologicznej w języku polskim*. Akademicka Oficyna Wydawnicza EXIT.

Zaluzniak, Andrei, 1977. *Grammaticheskij slovar' ruskogo yazyka*. Moscow: Russkij yazyk, 1st ed. (4th ed. — 2003).