A Relational Model of Polish Inflection in *Grammatical Dictionary of Polish*

Marcin Woliński

Institute of Computer Science
Polish Academy of Sciences
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 some complexities of the Polish inflectional system for various grammatical classes. Then we present a relatively compact relational model which can be used to describe Polish inflection in a uniform way.

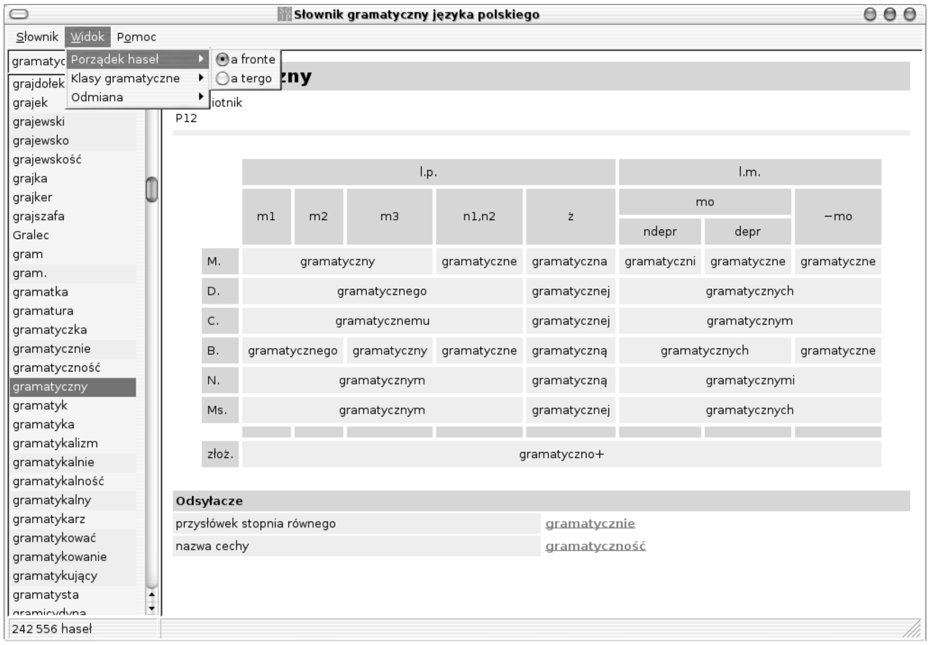
Key words: Polish morphology, inflection, relational modelling

1 Introduction

Grammatical Dictionary of Polish [1, henceforth: SGJP] aims at providing a description of Polish 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, as new words continuously enter the language. It is hoped, however, that the dictionary includes all reasonably frequent words and 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 [2]. SGJP is based on numerous earlier works: Tokarski's and Saloni's work on Polish inflectional suffixes [3], Saloni's description of Polish verbs [4, 5], Gruszczyński's description of Polish nouns [6], Wołosz's morphological data [7] and some other.

The scope of the dictionary is complete morphological characterisation and basic syntactic characterisation of Polish words. 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, comparative, and superlative adjectives.



grajdołek
grajek
grajewski
grajewsko
grajewskość
grajka
grajker
grajszafa
Gralec
gram
gram.
gramatka
gramatura
gramatyczka
gramatycznie
gramatyczność
gramatyczny
gramatyk
gramatyka
gramatykalizm
gramatykalnie
gramatykalność
gramatykalny
gramatykarz
gramatykować
gramatykowanie
gramatykujący
gramatysta
gramicydyna

242 556 haseł

P12

	l.p.					l.m.		
	m1	m2	m3	n1,n2	ż	mo		−mo
						ndepr	depr	
M.	gramatyczny			gramatyczne	gramatyczna	gramatyczni	gramatyczne	gramatyczne
D.	gramatycznego				gramatycznej	gramatycznych		
C.	gramatycznemu				gramatycznej	gramatycznym		
B.	gramatycznego		gramatyczny	gramatyczne	gramatyczną	gramatycznych		gramatyczne
N.	gramatycznym				gramatyczną	gramatycznymi		
Ms.	gramatycznym				gramatycznej	gramatycznych		
złoż.	gramatyczno+							

Odsyłacze

przysłówek stopnia równego	gramatycznie
nazwa cechy	gramatyczność

Fig. 1. The interface of *Grammatical Dictionary of Polish*

In SGJP the inflection is described according to theoretical decisions set in the above-mentioned linguistic works. Since we do not concentrate on phonological issues and our approach is paradigm-centric, the commonly used two-level morphology model [8] seems not appropriate. The model used by Wołosz [7] (which derives from Prószéky's ideas, cf. [9]) is based on similar assumptions as SGJP, but its drawback is complexity. To verify the description of any given lemma in this model one needs to understand quite complex system of constraints placed on word formation. The description used in SGJP, however, should be easy to use for involved linguists (including students, not necessarily with a formal background). One of the design goals is also to minimise the number of needed inflectional patterns. On the other hand, the description does not have to directly lead to fast tools for analysis or synthesis. The data can be converted to a finite state transducer afterwards (this is what we actually do).

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.

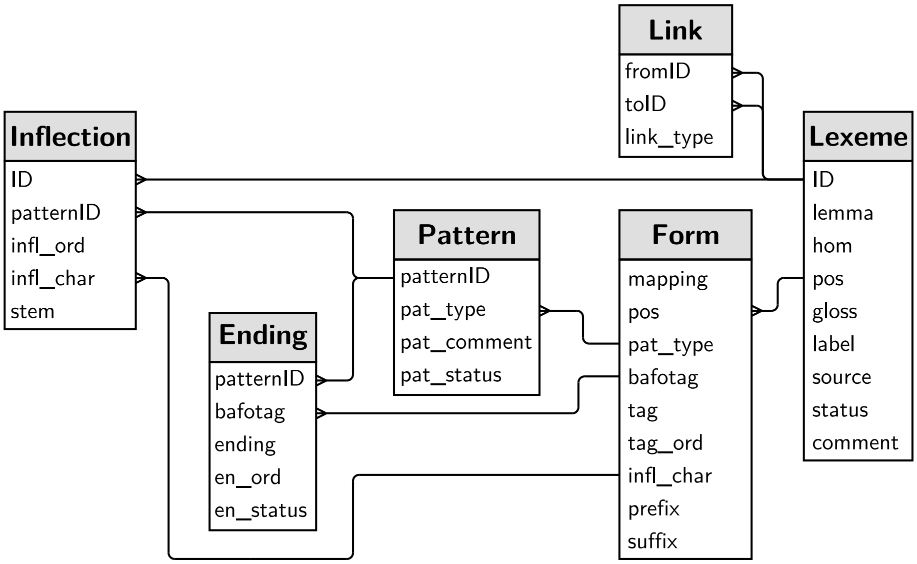


Fig. 2. The schema of the dictionary database (slightly simplified)

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).

2 The model

For the user, SGJP has the form of a dedicated interface program (cf. Fig. 1). The data in the backend is represented as a relational database—the program 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 majority of cases, all the forms share a common first part (and in the remaining

[1] Lexeme is not identical with an entry in the list of entries of the dictionary. A Lexeme may be represented by several entries. In one of the user-selectable views of the interface all inflected forms are used as entries.

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 4. For each instance of Pattern we have several instances of Ending with the respective endings as the value of the 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.

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.

3 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 [3]. 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]

These basic forms are mapped to actual inflected forms by the instances of the entity Form. For example, 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, cf. Table 1. 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 presenta-

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

[3] A rather detailed system of 9 genders for Polish is used in SGJP. It includes masculine personal m1 (e.g., *profesor*), animal m2 (*pies*), inanimate m3 (*stół*); neuter n1

Table 1. Instances of the entity Form for adjectival forms with basic form tag 2

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

tions 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 slightly compacted form and do not have 126 cells (as can be seen in Fig. 1). 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). An additional mapping is used to convert the dictionary to the form used by the morphological analyser *Morfeusz* [12].

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.

4 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 *płetwonurek* m1, *skowronek* m2, and *nagłówek* 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 [13, 14]. For that reason accusative forms are not included in the set of basic inflected forms for nouns. The right

(*dziecko*), n2 (*okno*); feminine f (*wanna*); plurale tantum p1 (*państwo*), p2 (*drzwi*), and p3 (*spodnie*); cf. [10, 11].

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

Table 2. Accusative singular Forms of nouns are generated from basic forms with different bafotags depending on pattern type (pat_type) and gender (infl_char).

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

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 [15]. For example nominative plural of the noun *płetwonurek* has a neutral variant *płetwonurkowie* and a stylistically marked (depreciative) variant *płetwonurki*. 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., *funkcji*/*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. As an example, the Table 2 lists instances of Form for singular accusative forms of various pattern types. The complete list of basic inflected forms for nouns (or more precisely their bafotags) used for various types of inflection pat_type is presented in Table 3.

Table 3. The sets of basic inflected forms (bafotags) used for various pattern types pat_type.

pat_type	m	f	n
bafotags	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 *człowieczysko* 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.

5 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 [16]. 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 to a Lexeme, ending specific to the basic inflected form, and suffix[5] characteristic for a Form. For the verb *wieść* ('to lead') used above the stem is *wi-*, the ending 10 is *-edzion-*, and the suffixes are marked in the above examples.

Some verbal forms, namely the negated gerunds and adjectival participles, are formed by prepending the prefix *nie* to the affirmative forms. For that purpose we use the attribute prefix of Form. The prefix and suffix is empty for other classes except for superlative degree of adjectives which is formed with prefix *naj*.

The Table 4 presents examples of Forms derived from basic inflected form 10.

Table 4. Verbal Forms generated from basic inflected form 10: finite impersonal past form, affirmative and negated forms of passive adjectival participle.

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

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

[5] The terms *prefix* and *suffix* are used here in the technical (and not linguistic) meaning of an arbitrary first (respectively last) part of a string.

6 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* 'two boys') or requires the noun to be in genitive (*dwóch chłopców*, 'two boys', both examples are nominative) [17].

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

Lexemes traditionally described as nominal, adjectival, adverbial, and numeral pronouns are treated in SGJP as regular nouns, adjectives, adverbs, and numerals [18]. The class of pronouns is limited to personal pronouns (including the reflective 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.

7 Conclusion

It may seem that with the given simple construction of inflectional patterns the description of inflection 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 in our model consists of four parts: prefix, stem, ending, and suffix controlled by several entities of the model. Each of these parts can be empty. However, since the mapping from Endings to Forms is universal, this complexity does not influence the process of adding new lexemes or verifying the existing description. These tasks can be successfully performed on the limited set of basic inflected forms, which are built only from a stem and an ending.

The Table 5 presents numbers of Patterns needed for various grammatical classes in the dictionary. Without the mechanism of reusing nominal patterns for various genders, we would need 1086 nominal patterns to describe all lexemes currently present in our database (46% more patterns). If verbs of different aspect were to require separate patterns, 384 verbal ones would be needed (79% more).

Due to irregularities in Polish inflection there exist numerous 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 it is rather compact and manageable. In particular it provides a unified method of generating forms

Table 5. The number of inflectional patterns needed for respective grammatical classes in SGJP.

adjectives	71
nouns	744
verbs	214
numerals	45
pronouns	6

of a lexeme of any grammatical class although the inflectional patterns for particular classes are constructed in a substantially different way.

References

1. Saloni, Z., Gruszczyński, W., Woliński, M., Wołosz, R.: Słownik gramatyczny języka polskiego. Wiedza Powszechna, Warszawa (2007)
2. Zalizniak, A.: Grammaticheskij slovar' russkogo yazyka. 1 edn. Russkij yazyk, Moscow (1977)
3. Tokarski, J.: Schematyczny indeks a tergo polskich form wyrazowych, edited by Zygmunt Saloni. Wydawnictwo Naukowe PWN, Warszawa (1993)
4. Saloni, Z.: Czasownik polski. Odmiana, słownik. Wiedza Powszechna, Warszawa (2001)
5. Saloni, Z., Woliński, M.: A computerized description of Polish conjugation. In: Kosta, P., Błaszczak, J., Frasek, J., Geist, L., Żygis, M., eds.: Investigations into Formal Slavic Linguistics (Contributions of the Fourth European Conference on Formal Description on Slavic Languages), pp. 373–384. (2003)
6. Gruszczyński, W.: Fleksja rzeczowników pospolitych we współczesnej polszczyźnie pisanej. Volume 122 of Prace językoznawcze. Zakład Narodowy im. Ossolińskich, Wrocław (1989)
7. Wołosz, R.: Efektywna metoda analizy i syntezy morfologicznej w języku polskim. Akademicka Oficyna Wydawnicza EXIT (2005)
8. Koskenniemi, K.: Two-Level Morphology: A General Computational Model for Word Form Recognition and Production. Volume 11 of Publication. Helsinki University, Helsinki (1983)
9. Prószéky, G., Tihanyi, L.: Humor: High-speed unification morphology and its applications for agglutinative languages. La tribune des industries de la langue 10(28-29) (1995) OFIL, Paris
10. Mańczak, W.: Ile rodzajów jest w polskim? Język Polski XXXVI(2), 116–121 (1956)
11. Saloni, Z.: Kategoria rodzaju we współczesnym języku polskim. In: Kategorie gramatyczne grup imiennych we współczesnym języku polskim, pp. 41–75. Ossolineum, Wrocław (1976)
12. Woliński, M.: Morfeusz — a practical tool for the morphological analysis of Polish. In: Kłopotek, M., Wierzchoń, S., Trojanowski, K., eds.: Intelligent Information Processing and Web Mining, IIS:IIPWM'06 Proceedings, Springer, pp. 503–512. (2006)
13. Gruszczyński, W.: Rzeczowniki w słowniku gramatycznym współczesnego języka polskiego. In: Gruszczyński, W., Andrejewicz, U., Bańko, M., Kopcińska, D., eds.:

Nie bez znaczenia... Prace ofiarowane Profesorowi Zygmuntowi Saloniemu z okazji jubileuszu 15000 dni pracy naukowej, Białystok, pp. 99–116 (2001)

14. Gruszczyński, W., Saloni, Z.: Notowanie informacji o odmianie rzeczowników w projektowanym *Słowniku gramatycznym języka polskiego*. In: Bobrowski, I., Kowalik, K., eds.: Od fonemu do zdania. Prace dedykowane Profesorowi Romanowi Laskowskiemu, pp. 203–213, Kraków (2006)
15. Saloni, Z.: O tzw. formach nieosobowych [rzeczowników] męskoosobowych we współczesnej polszczyźnie. Biuletyn Polskiego Towarzystwa Językoznawczego XLI, 155–166 (1988)
16. Saloni, Z.: Wstęp do koniugacji polskiej. Wydawnictwo Uniwersytetu Warmińsko-Mazurskiego, Olsztyn (2000)
17. Saloni, Z.: Kategorie gramatyczne liczebników we współczesnym języku polskim. In: Studia gramatyczne I, pp. 145–173, Wrocław (1977)
18. Saloni, Z.: Klasyfikacja gramatyczna leksemów polskich. Język Polski LIV, fasc. 1, 3–13, fasc. 2, 93–101. (1974)