The IPI PAN Corpus in Numbers

Adam Przepiórkowski

Institute of Computer Science
Polish Academy of Sciences
ul. Ordona 21, Warsaw, Poland
adamp@ipipan.waw.pl

Abstract

The aim of this article is to present the IPI PAN Corpus (cf. http://korpus.pl/), a large morphosyntactically annotated XML encoded corpus of Polish developed at the Institute of Computer Science, Polish Academy of Sciences. Various quantitative information about the corpus and its publicly available subcorpora is given including: sizes in terms of orthographic words and interpretable segments, tagset size measured in types and tokens, etc., but also information reflecting interesting facts about Polish, i.e., frequencies of words of different lengths and frequencies of grammatical classes and some grammatical categories.

1. Introduction

The IPI PAN Corpus, a large morphosyntactically annotated XML[1] encoded corpus of Polish, is one of the results of a corpus project financed by the State Committee for Scientific Research (Polish: Komitet Badań Naukowych; project number 7 T11C 043 20) from mid-2001 to mid-2004, carried out at the Institute of Computer Science, Polish Academy of Sciences (Polish: Instytut Podstaw Informatyki PAN; hence, the IPI PAN Corpus). Other results of the project, documented elsewhere, include: the design of a flexemic tagset for Polish (Woliński, 2003; Przepiórkowski, 2003; Przepiórkowski and Woliński, 2003), a statistical tagger (Dębowski, 2003; Dębowski, 2004) and the Poliqarp search engine featuring an expressive query language (Przepiórkowski et al., 2004; Przepiórkowski, 2004).

The aim of this paper[2] is to present some hitherto unpublished quantitative information about the IPI PAN Corpus and, especially, about two publicly available subcorpora: the smaller (around 15 million segments; see below) and more balanced sample corpus, searchable via the Internet at http://korpus.pl/, and a larger (around 70 million segments) corpus downloadable for searching from the same site. Similar data about frek, the older tiny (0.5 million running words) corpus of the "Frequency dictionary of contemporary Polish" (Kurcz et al., 1990), manually re-tagged within the current project, are also cited.

All quantitative information given below pertains to the so-called 'preliminary' version of the IPI PAN Corpus (Przepiórkowski, 2004) of June 2004.

[1] More precisely: XCES, i.e., XML Corpus Encoding Standard; cf. (Ide et al., 2000).

[2] This article is a preliminary and abbreviated version of a longer paper, *The Potential of the IPI PAN Corpus*, which will appear in a special issue of *Poznań Studies in Contemporary Linguistics* devoted to the potential of linguistic corpora. An electronic version of that paper will probably be available from http://www.ipipan.waw.pl/~adamp/.

2. Segmentation and Tagset

In order to appreciate the quantitative results presented below, it is necessary to understand basic assumptions about the segmentation (tokenisation) procedure and the tagset employed in the IPI PAN Corpus. Both are described in detail in (Woliński, 2003; Przepiórkowski, 2003; Przepiórkowski and Woliński, 2003), with the final version of the tagset presented in (Przepiórkowski, 2004). The present section offers a concise summary of the most important aspects of segmentation and tagging.

2.1. Segmentation

Segments are those sequences of orthographic characters to which tags are assigned. Segments in this sense are often understood as orthographic words (words "from space to space") and, in fact, segments of the IPI PAN Corpus are never longer than such orthographic words. However, in some special cases, segments may be shorter than orthographic words and, moreover, some non-words sequences of characters, especially punctuation marks, are assigned tags, i.e., they are treated as segments.

Cases where orthographic words are split into smaller segments include first and second person past forms such as *łgał*|*eś* 'lied-you', *długo*|*śmy* 'long time-we', *tak*|*em* 'so-I', etc., words containing particles such as *by* (subjunctive particle), *-ż(e)* (emphatic particle) and *-li* (question particle), e.g., *przyszedł*|*by* 'come-would', *napisała*|*by*|*m* 'write-would-I', *chodź*|*że* 'come-Emph', *potrzebował*|*że*|*by*|*ś* 'need-Emph-would-you', *znasz*|*li* 'know-Q', prepositions incorporating post-prepositional weak pronominal form *-ń*, as in *do*|*ń* 'to-him' or *ze*|*ń* 'with-him', and also some words containing the hyphen, i.e., words such as *polsko*|-|*niemiecki* 'Polish-German' and double names, e.g., *Kowalska*|-|*Nowakowska*.

2.2. Tagset

In the IPI PAN Tagset, each morphosyntactic tag is a sequence of colon-separated values, e.g.: subst:sg:nom:m1 for the segment *chłopiec* 'boy'. The first value, e.g., subst, determines the *grammatical class*, i.e., roughly, part of speech (POS), while the values that follow

it, e.g., sg, nom and m1, are the values of grammatical categories appropriate for that grammatical class.

Grammatical categories adopted here are: *number* (sg, pl), *case* (nom, gen, dat, acc, inst, loc, voc), *gender* (m1, m2, m3, f, n), *person* (pri, sec, ter), *degree* (pos, comp, sup), *aspect* (imperf, perf), *negation* (aff, neg), *accentability* (akc, nakc), *post-prepositionality* (praep, npraep), *accommodability* (congr, rec), *agglutination* (nagl, agl) and *vocalicity* (wok, nwok).

Grammatical classes are based on the notion of *flexeme* introduced in (Bień, 1991; Bień, 2004) — they are inflectionally uniform subsets of more traditional lexemic classes (POSs). The following grammatical classes are used in the IPI PAN Corpus: nominal classes: noun (subst), depreciative form (depr); pronominal classes: non-3rd person pronoun (ppron12), 3rd-person pronoun (ppron3), strong reflexive pronoun SIEBIE (siebie); numeral (num); adjectival classes: adjective (adj), adadjectival adjective (adja), post-prepositional adjective (adjp); adverb (adv); (de-)verbal classes: non-past form (fin), future BYĆ (bedzie), agglutinate BYĆ (also called *mobile inflection*) (aglt), l-participle (praet), imperative (impt), impersonal (imps), infinitive (inf), contemporary adverbial participle (pcon), anterior adverbial participle (pant), gerund (ger), active adjectival participle (pact), passive adjectival participle (ppas), winien (winien), predicative (pred); functional classes: preposition (prep), conjunction (conj), particle-adverb (qub); other classes: nominal alien (xxs), other alien (xxx), unknown form (ign) and punctuation (interp).

3. Segments

The IPI PAN Corpus as a whole is heavily unbalanced: most of the text in the corpus comes from newspapers, transcripts of parliamentary sessions and legal texts. Also `wstepny` consists mainly of parliamentary proceedings (over 68%) and newspapers (almost 21%), with only 6.5% of artistic prose, 3% of scientific texts and 1% of legal texts.

Some effort towards so-called representativeness was put into the make-up of the `sample` corpus, which consists of scientific texts (10%), contemporary artistic prose (10.6%), older (late XIX and early XX century) artistic prose of the kind read at schools (9.7%), legal texts (4.9%), transcripts of parliamentary sessions (15.5%) and various newspaper texts (49.3%). As is well known, also the `frek` corpus of the "Frequency dictionary of contemporary Polish" is supposed to be balanced, with 20% of popular science, 20% of news dispatches, 20% of editorials and longer articles, 20% of artistic prose, and 20% of artistic drama.

	segments	words	ratio
IPI PAN Corpus	360,446,336	291,187,457	1.24
`wstepny`	70,492,786	58,317,809	1.21
`sample`	15,252,022	12,198,241	1.25
`frek`	659,511	545,970	1.21

Table 1: Corpus sizes measured in segments and orthographic words.

The sizes of the IPI PAN Corpus as a whole, the two subcorpora `sample` and `wstepny`, as well as the `frek` corpus, are given in Table 1. The **segments** column contains the exact number of segments (including punctuation), while the **words** column shows the exact number of orthographic words (excluding punctuation) in each of these corpora. The final column gives the segment-to-word ratio calculated on the basis of the previous two columns.[3]

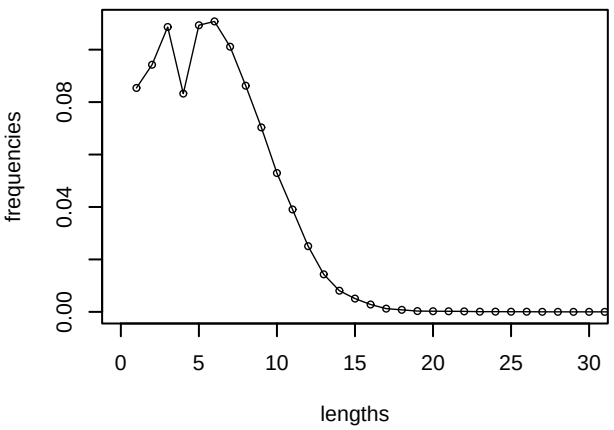

Figure 1: Frequencies of lengths of words in `frek`.

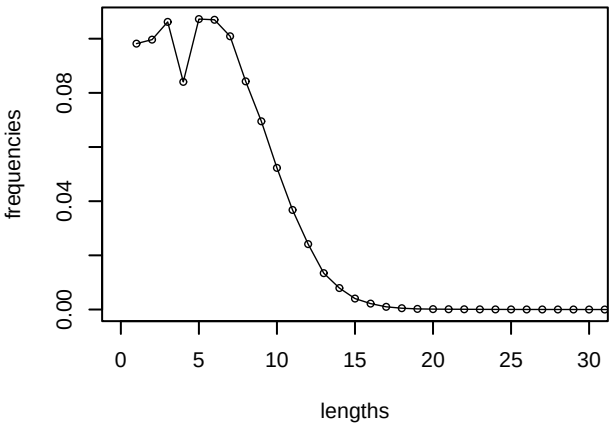

Figure 2: Frequencies of lengths of words in `sample`.

Additionally, Fig. 1–2 show the frequencies of words of various lengths in `frek` and `sample`. It is clear from these figures that the most frequent number of letters in words is 6 (`frek`, as well as `wstepny` and the whole IPI PAN Corpus; the last two not illustrated here) or perhaps 5 (`sample`), and that, interestingly, words of length 4 are

[3] Note that the number of words for `frek` in Table 1 differs from the declared 500,000. The difference probably stems from the fact that, as noted in (Kurcz et al., 1990) (see also (Czerepowicka and Saloni, 2004)), in the original version of `frek`, multiple orthographic words were sometimes treated as single wordforms containing a space, e.g., the reflexive marker *się* following an adjectival participle or a gerundial form, foreign surnames containing *de*, *von*, etc., fossilised sequences of prepositions and bound words, etc.

conspicuously less frequent than words of neighbouring lengths.

Finally, let us mention that the mean length of words calculated on the basis of these corpora are: 5.92 (frek), 5.78 (sample), 5.94 (wstepny) and 5.90 (the whole IPI PAN Corpus).

4. Tags

As mentioned above, each complete tag is a list whose first element is a grammatical class and the other elements are values of grammatical categories appropriate for this class. For example, each nominal tag has the form subst:*number*:*case*:*gender*, where *number* is sg or pl, *case* is one of 7 cases and *gender* is one of 5 genders, i.e., there are 70 potential nominal tags. Overall, the current tagset allows for 4179 potential tags, although many of the potentially possible combinations of grammatical classes and grammatical categories are never realised. The number of different tags actually found in corpora is given in Table 2.

	known		unknown		all	
	dis.	all	dis.	all	dis.	all
wstepny	945	1149	259	368	946	1150
sample	912	1131	237	357	913	1132
frek	–	–	–	–	1642	1642

Table 2: Actual tags — only those disambiguated or all tags proposed by the morphological analyser Morfeusz — for segments known by Morfeusz, for unknown segments, whose morphological interpretation was guessed on the basis of (Tokarski, 1993), and for all segments.

The most conspicuous difference between the rows for frek on the one hand and for sample and wstepny on the other hand stems from the fact that frek is annotated manually, with tags which are correct in a given context (we will call them 'disambiguated tags'), hence, there is no difference in frek between disambiguated tags and all tags. This also implies that, in frek, there are no segments marked as unknown to the morphological analyser, hence the '–'s.

On the other hand, as explained in more detail in (Przepiórkowski, 2004), not only does the IPI PAN Corpus contain appropriately marked disambiguated tags, but it also retains all other tags proposed by the morphological analyser used in the project, i.e., by Morfeusz (by Zygmunt Saloni and Marcin Woliński). Moreover, since Morfeusz is a dictionary-based analyser, there are words unknown to it.[4] In such cases, a guesser derived from (Tokarski, 1993) is used, which proposes interpretations on the basis of endings of words. As Table 2 shows, the repertoire of tags proposed by the guesser is quite limited with respect to the tags returned by Morfeusz (e.g., 368 vs. 1149 for wstepny); for example, the guesser never proposes pronominal, conjunctive or prepositional tags. Also, the only tag produced by the guesser which is not produced by Morfeusz is ign (hence, the difference of one between the **known** and **all** column pairs).

Another observation which begs explanation is that quite many tags returned by Morfeusz are never selected by the statistical disambiguator of (Dębowski, 2004): there are $1149 - 945 = 204$ such tags in wstepny and $1131 - 912 = 219$ in case of sample. In both cases the source of discrepancy is roughly the same; for example, in the latter case, 152 of those tags are participial tags (almost all adjectival participles), 36 — pronominal tags mainly with genders m2 and m3, 20 — comparative and superlative adjectival tags, 7 — vocative numeral and nominal tags, and a few odd prepositional and gerundial tags. It may be hypothesised that this difference reflects some systematic errors made by the disambiguator.[5]

Finally, the difference between the tags present in frek (1642) and all the tags present in, e.g., sample (1132) is caused by 930 tags present in frek but not in sample, and 420 tags present in sample, but not in frek. The majority of the 930 tags found in frek only are manual annotation errors resulting in such inconsistent tags as ppron3:pl:acc:f:pri, i.e., a 3rd person pronoun in plural number, accusative case, feminine gender and... first person. Three other main classes of frek-only tags are: numerals with explicit congr/rec information (65; automatic annotation often left those underspecified), vocative forms (50) probably not known to Morfeusz and various alien tags (automatic tagging was not able to distinguish alien forms from other unknown forms). On the other hand, surprisingly, the 420 sample-only tags are mainly the 264 adjectival participial (ppas and pact) tags and 130 first and second pronoun (ppron12) tags.

The results reported above show that much work still needs to be carried out to alleviate the annotation problems introduced both by manual and by automatic annotation.

Let us finish by comparing the frequencies of the main groups of grammatical classes, of case values and of gender values in frek and sample. In Fig. 3–4, the group *noun* comprises classes subst and depr, *verb* — all (de-)verbal classes enumerated in §2.2., *adjective* — the three adjectival classes mentioned there, *pronoun* — the pronominal classes ppron12, ppron3 and siebie, while *adverb*, *numeral*, *preposition*, *conjunction*, *particle* and *punctuation* are adv, num, prep, conj, qub and interp, respectively. Since the breakdown presented in these figures is made on the basis of disambiguated tags of known forms, classes xxs, xxx and ign are not included. Two clear differences between these two corpora concern pronouns and numerals, both more frequent in frek. The difference in the frequencies of numerals might be caused by the fact that numbers are converted to words in frek and treated as numerals,[6] while there is no such conversion in the IPI PAN Corpus, which results in the assignment of ign. On the other hand, the differences in the frequencies of pronouns might be the result of the high percentage of pronoun-rich 'artistic drama' in frek (although, on the

[4] As reported in (Piskorski et al., 2004), Morfeusz does not recognise about 5% of wordforms.

[5] (Dębowski, 2004) reports the 9.4% error rate of the disambiguator.

[6] (Czerepowicka and Saloni, 2004) note that this 'conversion' was actually an interpretative and non-deterministic process.

other hand, the IPI PAN Corpus contains a high percentage of parliamentary proceedings, which also to some extent approximates spoken language).

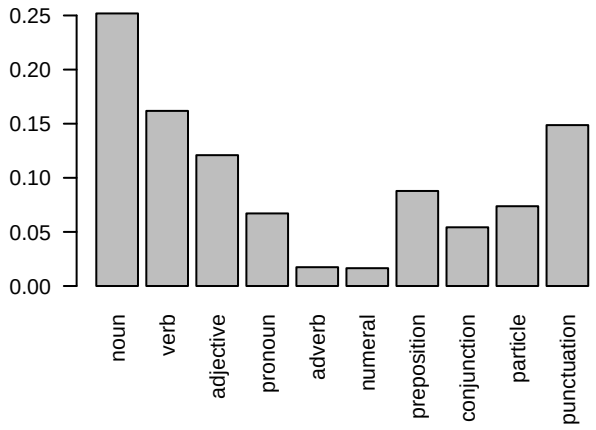

Figure 3: Frequencies of grammatical classes in `frek`.

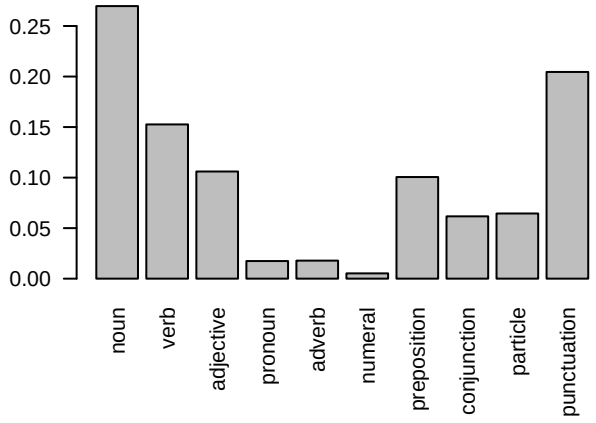

Figure 4: Frequencies of grammatical classes in `sample`.

Finally, Fig. 5–8 compare the frequencies of cases and genders in `frek` and `sample`. As can be seen from these figures, `sample` contains a conspicuously lower percentage of dative forms and animate masculine forms. The question whether this difference reflects any real change in Polish between 1960ies (`frek`) and now (over 90% of texts in `sample` date from the last 5 years), or whether it simply results from the difference in genre breakdown and annotation procedure, is left for future research.

5. Conclusion

The aim of this article, apart from presenting the IPI PAN Corpus to the Computational Linguistic community, was to provide its basic quantitative characteristics. In particular, we gave the number of potentially possible tags and the number of tags actually used in particular corpora, and we compared the frequencies of grammatical classes and two grammatical categories in `frek` and `sample`. The brief investigation of the differences between the tags found in various corpora led us to the identification of some of the annotation errors both in the manually annotated `frek` corpus and in the automatically annotated IPI PAN Corpus.

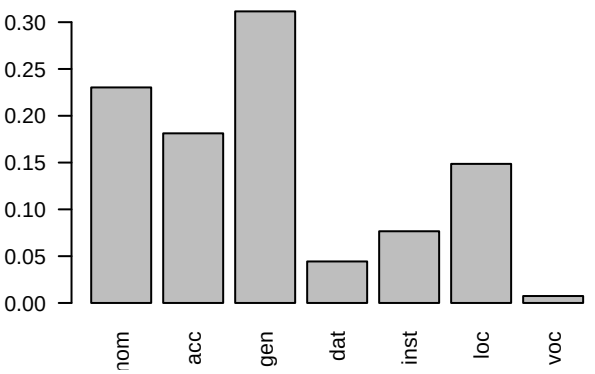

Figure 5: Frequencies of grammatical cases in `frek`.

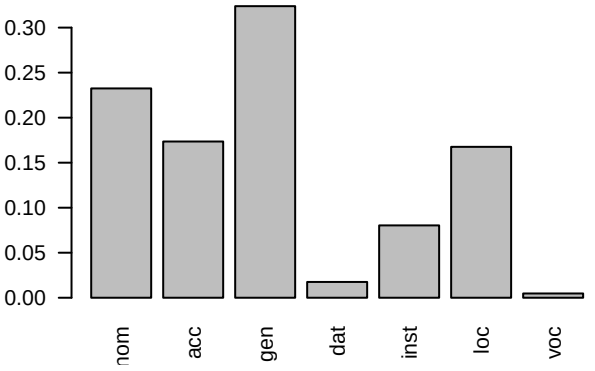

Figure 6: Frequencies of grammatical cases in `sample`.

6. References

Bień, Janusz S., 1991. *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji*, volume 383 of *Rozprawy Uniwersytetu Warszawskiego*. Warsaw: Wydawnictwa Uniwersytetu Warszawskiego.

Bień, Janusz S., 2004. An approach to computational morphology. In (Kłopotek et al., 2004), pages 191–199.

Czerepowicka, Monika and Zygmunt Saloni, 2004. Co skreślano i co dopisywano w korpusie *Słownika frekwencyjnego polszczyzny współczesnej*. In Irena Kamińska-Szmaj (ed.), *Od starożytności do współczesności. Język–literatura–kultura. Księga poświęcona pamięci profesora Jerzego Woronczaka*. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego, pages 381–391.

Dębowski, Łukasz, 2003. A reconfigurable stochastic tagger for languages with complex tag structure. In *Proceedings of* Morphological Processing of Slavic Languages, *EACL 2003*.

Dębowski, Łukasz, 2004. Trigram morphosyntactic tagger for Polish. In (Kłopotek et al., 2004), pages 409–413.

Ide, Nancy, Patrice Bonhomme, and Laurent Romary, 2000. XCES: An XML-based standard for linguistic corpora. In *Proceedings of the Linguistic Resources and Evaluation Conference*. Athens, Greece.

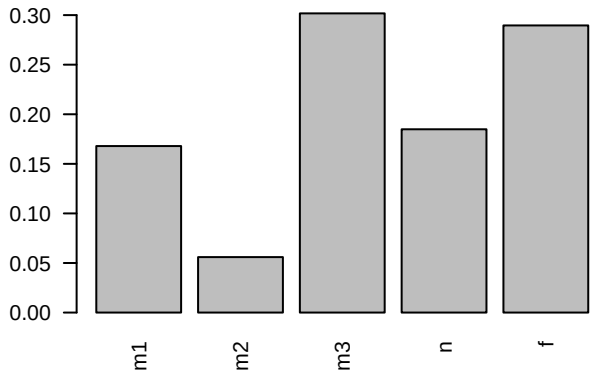

Figure 7: Frequencies of grammatical genders in `frek`.

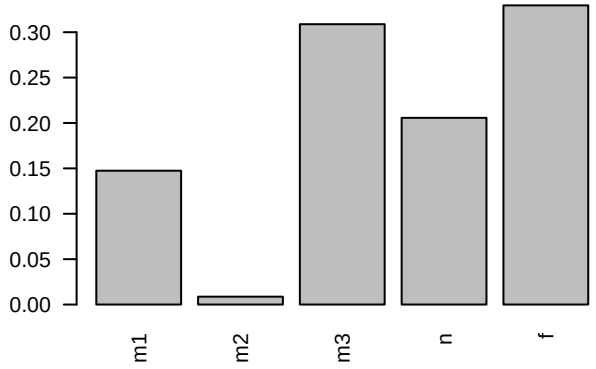

Figure 8: Frequencies of grammatical genders in `sample`.

Kurcz, Ida, Andrzej Lewicki, Jadwiga Sambor, Krzysztof Szafran, and Jerzy Woronczak, 1990. *Słownik frekwencyjny polszczyzny współczesnej*. Kraków: Wydawnictwo Instytutu Języka Polskiego PAN.

Kłopotek, Mieczysław A., Sławomir T. Wierzchoń, and Krzysztof Trojanowski (eds.), 2004. *Intelligent Information Processing and Web Mining*. Advances in Soft Computing. Berlin: Springer-Verlag.

Piskorski, Jakub, Peter Homola, Małorzata Marciniak, Agnieszka Mykowiecka, Adam Przepiórkowski, and Marcin Woliński, 2004. Information extraction for Polish using the SProUT platform. In (Kłopotek et al., 2004), pages 227–236.

Przepiórkowski, Adam, 2003. Składniowe uwarunkowania znakowania morfosyntaktycznego w korpusie IPI PAN. *Polonica*, XXII–XXIII:57–76.

Przepiórkowski, Adam, 2004. *The IPI PAN Corpus: Preliminary version*. Warsaw: Institute of Computer Science, Polish Academy of Sciences.

Przepiórkowski, Adam, Zygmunt Krynicki, Łukasz Dębowski, Marcin Woliński, Daniel Janus, and Piotr Bański, 2004. A search tool for corpora with positional tagsets and ambiguities. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*. Lisbon: ELRA.

Przepiórkowski, Adam and Marcin Woliński, 2003. 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*.

R Development Core Team, 2004. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Woliński, Marcin, 2003. System znaczników morfosyntaktycznych w korpusie IPI PAN. *Polonica*, XXII–XXIII:39–55.