Computing and Informatics, Vol. 28, 2009, 319–338

INCREASING QUALITY OF THE CORPUS OF FREQUENCY DICTIONARY OF CONTEMPORARY POLISH FOR MORPHOSYNTACTIC TAGGING OF THE POLISH LANGUAGE

Marcin Kuta, Paweł Chrzaszcz, Jacek Kitowski

Institute of Computer Science
AGH University of Science and Technology
al. Mickiewicza 30, Cracow, Poland
e-mail: {mkuta, kito}@agh.edu.pl

Manuscript received 18 December 2007; revised 21 February 2008
Communicated by Michal Laclavík

Abstract. The paper is devoted to the issue of correction of the erroneous and ambiguous corpus of Frequency Dictionary of Contemporary Polish (FDCP) and its application to morphosyntactic tagging of the Polish language. Several stages of corpus transformation are presented and baseline part-of-speech tagging algorithms are evaluated, too.

Keywords: Corpora preparation, part-of-speech tagging, natural language processing, machine learning

Mathematics Subject Classification 2000: 68T50, 68T05, 68T35

1 INTRODUCTION

Modern computer linguistics is not possible without machine readable resources like corpora, dictionaries or wordnets. In the paper we analyse requirements posed on corpora when applied to part of speech (POS) tagging. Our aim is to provide the linguistic community with corpus of the Polish language, which would be suitable in a satisfactory way for applications in the NLP area, especially for POS tagging tasks. We base our work on the existing corpus of Frequency Dictionary of Contemporary Polish (FDCP), which suffers many deficiencies but is the only available corpus for the mentioned tasks. Our efforts are focused on decreasing the number of errors and reducing the ambiguity level. The purpose, therefore, is to construct a modified FDCP corpus with better characteristics. Corpus correction is usually hard and expensive as it is a time consuming process, requiring big human teams and deep linguistic knowledge. We present general strategy of corpus correction, useful when few human correctors are available and fast results expected. Our approach is applicable to large tagsets, what is the case of corpora of highly inflecting Slavic languages. Next we evaluate accuracy of selected tagging algorithms on the modified corpus and investigate an influence of corpus quality on accuracy of these algorithms.

To make clear further discussion we introduce some concepts. A token is the smallest entity being subject of tagging. A tag describes linguistic characteristics of a token. A tag consists of a list of attributes. Each attribute describes different morphological category. The first attribute in the list is the most important one and represents grammatical class, called also part of speech class (POS class or shortly POS). According to POS of a token, tags differ in size and attributes they take. All inflectionally related forms of a word are related together by a main, uninflected form, called a base form or a lemma. A set of all distinct tags appearing in a corpus is called a tagset. An ambiguity occurs when a token is assigned different tags in different contexts. An inherent ambiguity occurs when a token is assigned many tags in one context. A token is classified as a word segment if it contains at least one alphanumeric character (including Polish diacritics) or a digit. Remaining tokens represent punctuation marks.

The rest of the paper is organized as follows. Section 2 gives examples of important corpora. Section 3 presents Polish language resources, especially the corpus of Frequency Dictionary of Contemporary Polish. Section 4 describes in more detail a process of preparation of an improved corpus. Obtained results of tagging experiments are given in Section 5. Conclusions close the paper.

2 IMPORTANT CORPORA RESOURCES

Corpora resources are indispensable elements in computationally based natural language processing. Pure text corpora provide accurate statistics about the considered language. Tagged corpora are collections of words (or entities) marked up mainly with parts of speech, but other information like semantics (e.g. with WordNet senses) or dysfluency annotation is also possible. Corpora suitable for development of parsers, called treebanks, additionally carry information about the whole sentence structure. For the purpose of machine translation aligned corpora in two or more languages are necessary.

Beginning of the corpus linguistics is related to compilation of the Brown Corpus [8], the first major corpus of English. The Brown Corpus is a 1 million word collection of samples from 500 witten texts from 15 genres (newspaper, novels, non-fiction, academic, etc.) in American English assembled in 1963–1964. The corpus was first tagged with the TAGGIT program and then hand-corrected. The original Brown tagset contains 87 tags. The Susanne Corpus (Surface and Underlying Structural Analysis of Natural English) is a 130 000 word parsed subset (treebank) of the Brown Corpus. In 1978 the Lancaster-Oslo/Bergen (LOB) corpus [14] was completed as a British English equivalent to the Brown corpus. The Brown Corpus gave also rise to development of corpora of other flavours of English like Australian, New Zealand and Indian English.

The Penn Treebank project has produced treebanks from the Brown Corpus material, the Wall Street Journal (WSJ) texts, ATIS (Air Travel Information Service) translations and the Switchboard corpus of telephone conversations [10] under the common name Penn Treebank [22]. The Penn Treebank is the most significant English treebank and the WSJ section probably the most evaluated and cited resource of the English language. The Penn Treebank tagset, which evolved from the Brown tagset, contains 45 tags and attained great popularity due to its simplicity. The Penn Treebank is available from the Linguistic Data Consortium (LDC), its current version is Treebank-3.

The British National Corpus (BNC) [18] contains 100 million word tagged with the CLAWS (Constituent Likelihood Automatic Word-tagging System) tagger with tags from 61-tag C5 tagset.

In the area of German language the NEGRA corpus [2] and newer TIGER Treebank [4] are the most significant. The NEGRA corpus served for studies on usefulness of POS tagging algorithms of German [26] and development of the TnT tagger [3]. Product of another project, Verbmobil, is a treebank containing over 30 000 sentences in 3 languages supporting efforts of bidirectional speech translation between German and English and between German and Japanese.

The result of the MULTEXT-East project (Multilingual Text Tools and Corpora for Eastern and Central European Languages) [6] is a hand-validated small parallel corpus of Orwell's 1984 novel, speech, fiction and newspapers texts in the following Central European languages: Czech, Romanian, Hungarian, Estonian, Bulgarian, Slovene and English as a hub language. The TELRI project (Trans European Language Resources Infrastructure) added resources for Lithuanian, Croatian, Serbian and Russian.

The Prague Dependency Treebank (PDT) [11], whose annotation scheme is inspired by dependency grammars, contributed to results for the Czech (western Slavic) language [12]. The Hungarian National Corpus [27] and the Manually Annotated Hungarian Corpus (the Szeged Corpus) [1] are interesting examples of resources representing agglutinative language.

3 LANGUAGE RESOURCES FOR POLISH

The IPI PAN corpus [30] elaborated at the Polish Academy of Sciences contains 250 million words in the second issue. The binary version is searchable via Internet or dedicated tool called Poliqarp. Source version of the first issue may be acquired, but under a license excluding any research. The FDCP corpus is a set of 10 000 text samples gathered between 1963–1967 and published in the Internet in 2001. The corpus is broadly discussed in Section 3.1. The PWN corpus [33] of Polish Scientific Publishers is an untagged (plain text), balanced, 40 million word collection of texts from 386 books, 977 issues of 185 newspapers and magazines, 84 talk recordings, 207 web pages and several hundred leaflets. The corpus is available as an online Internet service but only abridged (7.5 million words) version without charge. The corpus, being balanced, represents language well, but is unsuitable for supervised learning algorithms. Provided that the corpus would be tagged with external high accuracy tagger, relatively big size and balance make it adequate resource for techniques like text-based general ontology discovery or creating formal specifications from natural language descriptions [23].

PELCRA (Polish English Language Corpora for Research and Application) [32] is patterned on the BNC 100 million word corpus of written and spoken language mainly from years 1992–2003. The PELCRA corpus is available only via web interface. The aim of authors of the corpus is to create the Polish national corpus.

Sometimes mentioned HPSG (Head-Driven Phrase Structure Grammar) small size set [20] serves as a test set for parsers evaluation and should not be treated as a development resource.

Other important issue is development of tagsets structuring corpora. A few tagsets for Polish have been worked out: IPI PAN, SFPW, LEM, SAM, XeLDA and others, undocumented but exploited by morphological analysers. Their broader characteristics are gathered in [28].

3.1 The Corpus of Frequency Dictionary of Contemporary Polish

The FDCP corpus [29], balanced between five distinct genres: scientific texts, news, essays, fiction and plays, represents different styles of the language. It is the only corpus of the Polish language both available with no restrictions for the community and suitable for supervised learning.

The FDCP corpus comes in two flavours: as annotated with the IPI PAN tagset [30] or with the SFPW tagset [9] (under the name of the enriched corpus of Frequency Dictionary of Contemporary Polish). As the IPI PAN tagset, which evolved from the SFPW tagset, is better suitable for the tagging paradigm, only the FDCP corpus with this tagset is further pondered.

The authors of the IPI PAN tagset provided the FDCP corpus with the IPI PAN source tagging established in automatic manner by a run of a morphological analyser followed by a run of a disambiguating program.

The FDCP corpus consists of 659 511 tokens (where 92 942 are different tokens), 552 739 word segments, 40 862 sentences and contains 1 270 different tags. 49.93 % of tokens is ambiguous with mean token ambiguity equal to 3.4. There are 35 739 different base forms.

The considerable deficiency of the corpus is its inherent token ambiguity – 20 601 tokens (3.12 % of the corpus) are assigned more than one tag, 3.92 tags on average. In particular some tokens are assigned more than one POS class or more than one base form.

3.2 IPI PAN Tagset

The IPI PAN tagset describes grammatical classes which are finer-grained than traditional parts of speech like nouns or verbs. These 32 grammatical classes (for brevity we will further call them POS) represent

- nouns (subst, depr),
- numerals (num, numcol),
- adjectives (adj, adja, adjp),
- adverbs (adv),
- pronouns (ppron12, ppron3, siebie),
- verbs (fin, bedzie, aglt, praet, impt, imps, inf, pcon, pant, ger, pact, ppas, winien),
- other categories (pred, prep, conj, qub, xxs, xxx, ign, interp).

Morphological categories described within the tagset and their possible values are presented in Table 1.

morphological category	values
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
postprepositionality	praep npraep
accommodability	congr rec
agglutination	agl nagl
vocalicity	wok nwok

Table 1. List of possible values of morphological categories

An example of an annotation of a token `Dopadł` (a form of the verb *catch up*) with the IPI PAN tagset encoded in the XML format is given below.

```
<tok>
   <orth>Dopadł</orth>
```

```
  <lex disamb="1">
    <base>dopaść</base><ctag>praet:sg:m1:perf</ctag>
  </lex>
  <lex disamb="1">
    <base>dopaść</base><ctag>praet:sg:m2:perf</ctag>
  </lex>
  <lex disamb="1">
    <base>dopaść</base><ctag>praet:sg:m3:perf</ctag>
  </lex>
</tok>
```

A token, its lemmata and tags are indicated by `<orth>`, `<base>`, `<ctag>` markers. Attributes of a tag are separated by colons.

4 PREPARATION OF MODIFIED FDCP CORPUS

A tagged corpus suitable for NLP tasks should satisfy several criteria.

Availability: To allow the linguistic community exchange of experience and straight comparison of research results, a corpus should be widely attainable in the text (non binary) format, directly searchable (access by web interface or dedicated tools should not be obligatory), preferably cost free.

Unambiguity: is a basic requirement of all tagging algorithms.

Size: The optimal size of a corpus depends on its intended application. Despite efforts for building ontology from short texts [16] very large size corpora (over 100 million words) are the most suitable for text to ontology approach and semantic networks building [15]. For morphosyntactic tagging much smaller corpora are sufficient.

Error free: Undertaken assumptions and rules should be applied consequently during preparation of a corpus. This condition is difficult to fulfil due to observed 96 %–97 % limit of human taggers accuracy [21]. Next, if a corpus is developed by a team, its members differ in linguistic and computer science knowledge. In addition, the language paradigm may change during development of a corpus.

The above criteria are internally contradictory, e.g. a big, error free corpus requires huge amount of work, what is difficult to reconcile with the availability condition. The FDCP corpus satisfies the availability requirement (open, cost free access). Its size, although over 100 times smaller than that of corpora like the BNC, enables application in POS tagging experiments. However, ambiguity and numerous errors necessitate amelioration of the corpus. This laborious task has been divided into several stages: preliminary disambiguation, tags flattening, tagset validation and correction, proper disambiguation, outlined in detail below (Sections 4.1–4.4) together with the modified corpus presented in Section 4.5.

4.1 Preliminary Disambiguation

The aim of this phase is to have each token of the FDCP corpus unambiguously assigned a POS class and a base form. Such approach quickly eliminates many tags without superfluous effort.

The corpus comprises some easy-to-resolve cases as obviously wrong possibilities (e.g. tags with POS class equal to grammatical class xxx mentioned in Section 3.1) or clear-to-disambiguate on the base of a context. A frequent choice between particle-adverb (qub) and conjunction (conj) annotation is judged by analysis of a whole sentence and function of a considered token. Sometimes a lemma has to be selected among an active adjective participle (pact) and an infinitive form (inf). The proper form is selected depending upon whether a sentence describes an action (infinitive form of lemma chosen) or a feature (participle chosen).

The result of preliminary disambiguation is as follows:

- elimination of 60 base forms ambiguities and 129 pos class ambiguities
- removal of 12 damaged sentences containing typos, logical inconsistencies or incomplete.

Since tokens are preliminarily disambiguated, we can also abandon the XML format and encode the corpus in more convenient notation (called the full notation), as shown below.

$$\underbrace{\texttt{Dopadł}}_{\text{token}}\underbrace{\texttt{praet}}_{\substack{\text{POS}\\ \text{class}}}\texttt{:[}\underbrace{\texttt{sg:m1:perf}}_{\text{tag 1}}\texttt{.}\underbrace{\texttt{sg:m2:perf}}_{\text{tag 2}}\texttt{.}\underbrace{\texttt{sg:m3:perf}}_{\text{tag 3}}\texttt{]\{}\underbrace{\texttt{dopaść}}_{\text{base form}}\texttt{\}}$$

4.2 Tags Flattening

The full notation is still too verbose for next corpus transformations and tags flattening has to be performed. Our example, encoded in the more concise flat format, looks as follows:

$$\underbrace{\texttt{Dopadł}}_{\text{token}}\underbrace{\texttt{praet:sg:m1.m2.m3:perf}}_{\text{tag1, tag2, tag3 in flattened form}}\texttt{\{}\underbrace{\texttt{dopaść}}_{\text{base form}}\texttt{\}}$$

In order to convert token annotation to the flat format two conditions must be fulfilled:

- tags describing a token must have equal number of attributes and corresponding attributes must be values of the same morphological category
- for each token, a set of tags resulting from the flat notation must be equal to a set of tags in the full notation.

The above conditions do not hold in few cases. Then tags are flattened manually and some of them (possibly correct) are discarded before conversion. This loss of information is a reasonable price for obtaining the corpus in the flat format.

As an outcome of the tags flattening phase:

- Tags of 20 257 tokens have been flattened automatically and tags of 140 tokens, not suitable for automatic procedure, manually.
- 26 sentences have been removed, as too difficult for any correction.

4.3 Tagset Validation and Correction

The corpus in the flat format is suitable for a tagset validation, i.e., a procedure of checking the correctness of tags with the specification of the IPI PAN tagset given in [24]. The constraints imposed on the tags are recalled in Table 2. The specification admits that in some cases certain attributes are optional. Therefore Table 2 is interpreted in a simplified way. If a tag passes the tagset validation procedure positively, it means that its list of attributes is equal to all attributes marked + or − and appropriate for POS class of the token. It cannot, however, be inferred that values of attributes describe correctly a token being subject to annotation.

The tagset consistency checking procedure revealed inconsistencies, which are summarised in Table 3. According to the gathered data, among total number of 11 344 occurrences of pronouns there are 11 337 third person pronouns (ppron3) and only 7 non-third person pronouns (ppron12). This lack of balance indicates errors in corpus annotation. To fix them, word forms *Ja, Mnie, My, Nam, Nas, ja, mi, mnie, mną, nam, my, nam, nami, nas* (inflected forms of the pronoun *I*), tagged originally as third person pronouns, have been reannotated as non-third person pronouns with the category person *pri* and the base form *ja*. Similarly word forms *Ciebie, Tobie, Ty, Was, Wam, Wy, ci, ciebie, cię, tobie, tobą, ty, wam, wami, was* (inflected forms of the pronoun *you*) have been redefined as non-third person pronouns with the category person *sec* and the base form *ty*.

Next 16 tokens tagged incorrectly as pronouns have been assigned appropriate (different from ppron12 or ppron3) POS classes. Finally two occurrences of tags with a missing definition of the category person have been corrected manually.

Inconsistencies taking into consideration the above transformations are gathered in Table 4. Inconsistent annotations are caused by 5 morphosyntactic categories. Missing definitions concern accentability, postprepositionality, accommodability and agglutination attributes while redundant ones concern vocalicity and postprepositionality attributes. The problem of redundant definitions is obvious to solve – it is sufficient to remove these attributes. Missing attributes have to be examined carefully.

Agglutination, applying only for pseudoparticiples (praet), is correctly fixed in 6 554 occurrences and missing 21 351 times. Large continuous fragments of text without this attribute within the tags suggest that some annotators preparing the FDCP corpus took into account the attribute while others omitted it. Indeed, there is no clear rule governing the presence of the attribute. For the majority of tags with the missing agglutination attribute, the solution is to omit the attribute in all cases.

POS	morphological categories											
	number	case	gender	person	degree	aspect	negation	accentability	postprepositionality	accommodability	agglutination	vocalicity
subst	+	+	–									
depr	–	+	–									
num	–	+	+							+		
numcol	–	+	–							+		
adj	+	+	+		+							
adja												
adjp												
adv					+							
ppron12	–	+	+	–				+				
ppron3	+	+	+	–				+	+			
siebie		+										
fin	+			+		–						
bedzie	+			+		–						
aglt	+			+		–						+
praet	+		+			–					+	
impt	+			+		–						
imps						–						
inf						–						
pcon						–						
pant						–						
ger	+	+	–			–	+					
pact	+	+	+			–	+					
ppas	+	+	+			–	+					
winien	+		+			–						
pred												
prep		–										
conj												
qub												
xxs	+	+	–									
xxx												
ign												
interp												

Table 2. Specification of the correct tags of the IPI PAN tagset; + means that POS is fully inflecting by a given morphological category, – means that POS is described by fixed value of a morphological category (but for different words values may be different), empty field means that a morphological category is not applicable to POS

POS class	type of inconsistency	number of occurrences
adj	correct	77 843
adja	correct	467
adjp	correct	363
adv	correct	12 424
aglt	correct	5 156
bedzie	correct	1 558
conj	correct	38 880
depr	correct	29
fin	correct	32 296
ger	correct	7 117
imps	correct	1 386
impt	correct	1 955
inf	correct	10 343
interp	correct	106 642
num	correct	8 385
num	missing definition of accommodability	591
pact	correct	2 955
pant	correct	122
pcon	correct	1 536
ppas	correct	6 376
ppron12	correct	1
ppron12	missing definition of accentability	6
ppron3	correct	2 672
ppron3	missing definition of accentability	3 176
ppron3	missing definition of accentability and postprepositionality	5 728
ppron3	missing definition of person and accentability	2
ppron3	missing definition of postprepositionality	2 431
praet	correct	65 54
praet	missing definition of agglutination	21 351
pred	correct	3 336
prep	correct	30 374
prep	redundant definition of vocalicity	32 433
qub	correct	52 880
siebie	correct	1 154
subst	correct	178 543
winien	correct	316
xxs	correct	171
xxx	correct	1 197

Table 3. Types of inconsistencies with the specification appearing in the corpus in the flat format

POS class	type of inconsistency	number of occurrences
num	missing definition of accommodability	591
ppron12	missing definition of accentability	3 592
ppron12	redundant definition of postprepositionality	118
ppron12	missing definition of accentability and redundant definition of postprepositionality	148
ppron3	missing definition of accentability	3 030
ppron3	missing definition of accentability and postprepositionality	2 126
ppron3	missing definition of postprepositionality	1
praet	missing definition of agglutination	21 351
prep	redundant definition of vocalicity	32 433

Table 4. Types of inconsistencies with specification appearing in the corpus after correction of POS class in tags defined originally as third person pronouns (ppron3), correct occurrences not shown

The accommodability attribute, defined for cardinal numerals (num) and collective numerals (numcol), is assigned correctly 8 385 times and misses 591 times. Despite a domination of correct annotations we decided to remove the attribute. The large number of tags with the missing attributes precludes manual correction, taking into account our small annotator team and reasonable time horizon. During the analysis of the attribute it has also been discovered that there are no tokens in the corpus defined as collective numerals, all numerals including collective ones are annotated as cardinal numerals. This rule, as being consequently applied, does not break credibility of future taggers outcomes but indicates, however, inconsistency with specification of the IPI PAN tagset.

The next problematic attributes are accentability and postprepositionality. The analysis of attributes is hindered by their optional character for pronouns. Careful examination has revealed that these attributes are really missing in many cases, where they certainly should be present. As previously, removal of the attributes turned out to be the only reasonable solution.

Cutting off 4 attributes may seem a radical move, but anyway we must get rid of the described errors and manual correction would require too much time.

4.4 Proper Disambiguation

Inherent ambiguity of the FDCP corpus appears as:

- real ambiguity, occurring when many tags are equally correct as the annotation of a given token, considering the whole sentence context
- erroneous ambiguity, being effect of annotator mistakes or simply superficial preparation of the corpus.

The number of inherent ambiguities of tokens is presented in Table 5.

category	ambiguity	occurrences
number	pl.sg	80
case	acc.nom	93
	acc.dat.gen.inst.loc.nom.voc	63
	nom.voc	44
	gen.nom	19
	acc.gen	16
	acc.nom.voc	9
	dat.gen	6
	gen.inst	5
	acc.loc	4
	loc.voc	3
	dat.gen.loc	2
	dat.nom	1
	dat.gen.loc.nom.voc	1
	acc.inst	1
	acc.dat.gen	1
	gen.loc	1
gender	m1.m2.m3	8 867
	f.m1.m2.m3.n	6 936
	f.m2.m3.n	1 191
	m1.m2.m3.n	978
	m1.m2	149
	m3.n	105
	m2.m3	98
	f.m3	94
	f.n	84
	f.m1	40
	m2.m3.n	37
	f.m3.n	24
	m1.n	21
	m1.m3	17
	f.m1.n	6
	f.m2.n	6
	m1.m2.n	4
	m1.m3.n	2
	f.m1.m2.n	1
person	pri.sec.ter	521
degree	comp.pos.sup	112
aspect	imperf.perf	35
negation	aff.neg	5

Table 5. Inherent ambiguities of tokens with respect to morphological categories

Two approaches to inherent ambiguity problem were pondered.

- Disambiguation in random manner, i.e., drawing a tag from a list of possible tags. This method introduces bias to an outcome corpus, thus making tagging algorithms less effective.
- Selection of a tag according to a certain deterministic rule. The method leads to information loss and disturbs tags distribution.

Each of the above solutions has drawbacks, so we proceed in the following way. If an ambiguity occurs 50 times or more often, it is treated as a new value of the considered morphological category, e.g. ambiguity sg.pl occurs 80 times and thus we assume morphological category number takes 3 possible values: sg, pl and sg.pl. Note that from now a value like sg.pl is not only notational convention but denotes tokens for which both sg and pl values are equally correct. These new values of attributes are taken into consideration when counting the tagset size of the modified corpus. Ambiguities less frequent than 50 occurrences are resolved manually. 311 ambiguities falling under this case were processed, accounting to reduction of total number of ambiguities from 19 682 to 19 439; 12 sentences have been removed.

The encountered doubtful cases are resolved consequently by rules among which the following turns out to be the most useful. Gender of the pronoun *ja* is fixed to feminine or masculine personal (f.m1). Gender of unknown abbreviations is fixed to masculine inanimate (m3). Gender of numerals not related to nouns is fixed to masculine inanimate (m3).

4.5 The m-FDCP Corpus

Influence of particular phases on the corpus is summarised in Table 6.

stage	all tokens	inherently amb. tokens	inherently amb. tags	sentences
Original corpus	659 511	20 601	80 764	40 862
Preliminary disambiguation	659 250	20 433	80 209	40 850
Tagset flattening	658 749	20 288	78 817	40 824
Tagset validation and correction	658 749	19 240	76 529	40 824
Proper disambiguation (modified corpus)	658 656	19 025	75 899	40 812

Table 6. Impact of consecutive transformations on selected corpus parameters

As an outcome of described transformations we obtained a corpus, referred to the modified FDCP corpus (shortly the m-FDCP) hereafter. The main parameters of the modified corpus are summarised in Table 7 (fourth column). The m-FDCP corpus is available at [31] site.

	m-FDCP			FDCP
	Training 90 %	Test 10 %	Full 100 %	Full 100 %
tokens	592 729	65 927	658 656	659 511
word segments	496 907	55 139	552 046	552 739
sentences	36 601	4 211	40 812	40 862
different tokens	87 097	19 557	92 872	92 942
different base forms	33 860	10 207	35 708	35 739
	Simple tagset			
tagset size	30	30	30	30
ambiguous tokens, %	26.15	26.19	26.16	26.43
mean token ambiguity	1.44	1.43	1.44	1.48
	Complex tagset			
tagset size	1 191	724	1 243	1 270
ambiguous tokens, %	47.76	47.65	47.74	49.93
mean token ambiguity	3.12	3.12	3.12	3.40

Table 7. Parameters of the m-FDCP and the FDCP corpora

5 EVALUATION OF M-FDCP CORPUS

We evaluate the m-FDCP corpus with four baseline tagging algorithms given in Table 8. All the taggers used in experiments are gratuitous for research purposes. For overview of the algorithms the reader is referred to [19] and for taggers descriptions to [3, 5, 7, 25].

Beside evaluation of the m-FDCP corpus annotated with the IPI PAN tagset (referred to a complex tagset) we also examine the corpus annotated with the reduced version of the tagset with only POS class considered (called a simple tagset further).

Algorithm	Tagger name
HMM	TnT
Maximum entropy	MXPost [1]
Transformation based	fnTBL
Memory based	MBT

Table 8. Taggers used in experiment

The experiment was conducted with split of the full m-FDCP corpus into a training set and a test set, representing 90 % and 10 % of the corpus, respectively. The split was carried out in a way preserving balanced representation of five genres both in the test set and the training set. More details of data preparation are reported in [17] for a similar experiment and the main parameters of the training and test

[1] we will refer to this tagger as MXP

sets summarised in Table 7 (columns 2 and 3), compared with overall parameters of the FDCP corpus (column 5) and in Table 9.

unseen tokens, %	9.37
tokens with unseen lemmas, %	3.22
unseen different lemmas, %	18.11
Simple tagset	
tokens with unseen tags	312
unseen different tags, %	0
Complex tagset	
tokens with unseen tags	2 262
unseen different tags, %	7.18

Table 9. Parameters of the test set in relation to the training set of the m-FDCP corpus

The experiments were performed at the ACC Cyfronet AGH-UST site on the SGI Altix 3 700 machine.

Assuming universally a test corpus contains n tokens, i-th token's correct annotation is a tag t_i and guessed annotation is a tag g_i, the accuracy, *acc*, is defined as follows:

$$acc \overset{\text{df}}{=} \frac{\#\text{correctly tagged tokens}}{\#\text{all tokens}} = \frac{\sum_{i=1}^{n} \delta(g_i, t_i)}{n} \quad , \tag{1}$$

where δ is the Kronecker delta function.

Detailed results for taggers trained on the training set of the m-FDCP corpus are gathered in Table 10. An ideal process of splitting a text into sentences and splitting words into tokens is assumed.

For each tagger, given its accuracy computed on the FDCP corpus, acc_{org}, and accuracy computed on the modified corpus, acc_{mod}, the error reduction, Δ_{Err}, is defined in relation [13] to errors of the original corpus as follows:

$$\begin{aligned}\Delta_{Err} &\overset{\text{df}}{=} \frac{\#\text{errors of original corpus} - \#\text{errors of modified corpus}}{\#\text{errors of original corpus}} \\ &= \frac{acc_{\text{mod}} - acc_{\text{org}}}{1 - acc_{\text{org}}} \quad .\end{aligned} \tag{2}$$

Accuracy on the FDCP corpus for the TnT and fnTBL taggers is already provided in [17]. Results for the MXP and MBT taggers (not shown) were obtained similarly. Error rate reduction of the considered taggers is given in Table 11.

6 CONCLUSIONS

Based on the experiments with the FDCP corpus and the modified corpus the following conclusions can be drawn.

	TnT	MXP	fnTBL	MBT
	Simple tagset			
All tokens	96.20	96.30	96.51	95.74
Known tokens	96.98	97.01	97.51	97.10
Unknown tokens	88.65	89.43	86.89	82.60
Ambiguous tokens	89.50	91.09	91.36	89.94
Word segments	95.46	95.57	95.83	94.91
Word segments with known tags	96.94	96.79	97.55	97.08
Word segments with unknown tags	0.00	28.21	3.85	0.00
Unknown word segments	88.65	89.43	86.90	82.60
Sentences	61.48	62.15	63.71	58.54
	Complex tagset			
All tokens	86.33	85.00	86.79	82.31
Known tokens	88.97	87.53	89.76	85.75
Unknown tokens	60.86	60.55	58.09	49.06
Ambiguous tokens	78.66	78.71	80.34	72.48
Word segments	83.66	82.07	84.21	78.85
Word segments with known tags	90.73	87.44	90.27	86.58
Word segments with unknown tags	0.00	29.84	30.50	0.66
Unknown word segments	60.86	60.55	58.09	49.05
Sentences	28.95	26.88	29.87	22.51

Table 10. Tagging accuracy, *acc*, for taggers trained on 90 % of the m-FDCP corpus, [%]

	TnT	MXP	fnTBL	MBT
Simple tagset	0.08	0	0.29	0.05
Complex tagset	13.10	11.30	12.75	9.61

Table 11. Error rate reduction, Δ_{Err}, of taggers trained on the m-FDCP corpus in relation to the same taggers trained on the FDCP corpus, [%]

- The presented strategy of corpus correction, consisting of 4 phases, turned out to be successful. Preliminary disambiguation, exploiting sort of outliers analysis, excluded at the very beginning many incorrect tags with moderate effort. Tagset flattening introduced more compact notation, more convenient both for automatic processing and human correctors. Tagset validation allowed identifying and removing numerous errors in a fast manner. The methodology may be adapted especially to corpora with large tagsets, where the carried information is too large for humans to tackle. This is the case of corpora of the Slavic languages.
- For the complex tagset we note error rate reduction from 9.6 % to 13.1 % if moving from the original corpus to the modified one. The highest error reduction is recorded for the most accurate taggers. These outcomes should be conceded a high improvement, taking into account that only evaluated resources but no algorithms were changed.

- Apart from the higher accuracy the taggers trained on the modified corpus provide additional information as attributes of tags may be multi-valued. Multi-valued attributes in contrast to removing inherent ambiguities in random manner provide full repeatability and comparability of tagging experiments.
- When constraining to the simple tagset, we do not observe practical improvement of accuracy. The results are, however, already comparable with the state-of-the-art accuracy for English.
- Corpus quality has strong influence on the accuracy of applied taggers.

  Furthermore, preparation of training and test sets by removing inherent ambiguities in a random manner, as in our previous experiment [17], turned out to be a bad choice. Such a method introduces unwanted noise and hinders taggers to acquire right patterns.
- Experiments on the m-FDCP corpus confirmed that, similarly to the FDCP corpus, the fntTBL and TnT taggers achieve the highest accuracy. Results for all taggers turned out to be stable, what was verified by 9-fold cross validation on the training set.

To summarize, the results of tagging experiments may differ significantly, depending on profile and quality of resources used for evaluation. That makes performance analysis of examined algorithms difficult and indicates need for the gold standard corpus of Polish to measure accuracy with. Such a corpus would play for Polish similar role to the Penn Treebank for English.

Acknowledgments

The research is supported by the University of Science and Technology (Cracow, Poland) grant No. 11.11.120.777. ACC Cyfronet AGH is acknowledged for the computing time.

REFERENCES

[1] Alexin, Z.—Csirik, J.—Gyimothy, T.—Bibok, K.—Hatvani, C.—Proszeki, G.—Tihanyi, L.: Manually annotated Hungarian corpus. Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL '03), pp. 53–56, Budapest, Hungary, 2003.

[2] Brants, T.—Skut, W.—Uszkoreit, H.: Syntactic annotation of a German newspaper corpus. Proceedings of the ATALA (Association pour le Traitement Automatique des Langues) Treebank Workshop, pp. 69–76, Paris, France, June 18–19, 1999.

[3] Brants T.: TnT – A statistical part-of-speech tagger. Proceedings of the 6th Applied Natural Language Processing Conference (ANLP 2000), pp. 224–231, Seattle, Washington, USA, April 29–May 4, 2000.

[4] Brants, S.—Dipper, S.—Hansen, S.—Lezius, W.—Smith, G.: The TIGER Treebank. Proceedings of the First Workshop on Treebanks and Linguistic Theories, pp. 24–42, Sozopol, Bulgaria, 2002.

[5] Daelemans, W.—Zavrel, J.—van der Sloot, K.—van den Bosch, A.: MBT: Memory Based Tagger. Version 3.0. Reference Guide. ILK Technical Report 07-04, Tilburg University, The Netherlands, July 10, 2007, http://ilk.uvt.nl/downloads/pub/papers/.

[6] Erjavec, T.—Ide, N.: The MULTEXT-East corpus. Proceedings of the First International Conference on Language Resources and Evaluation (LREC 1998), pp. 971–974, Granada, Spain, 1998.

[7] Florian, R.—Ngai, G.: Fast Transformation-Based Learning Toolkit manual. John Hopkins University, USA, 2001, http://nlp.cs.jhu.edu/ rflorian/fntbl.

[8] Francis, W.—Kucera, H: Frequency Analysis of English Usage: Lexicon and Grammar. Houghton Mifflin, Boston, USA, ISBN 0-395-32250-2, 561 pp., 1982.

[9] Głowinska, K.: Morphological taxonomy for frequency dictionary. Unpublished work, Warsaw, Poland, 2001, http://www.mimuw.edu.pl/polszczyzna/pl196x/doc/files/taksonomia.pdf.

[10] Godfrey, J.—Holliman, E.—McDaniel, J.: SWITCHBOARD: Telephone Speech Corpus for Research and Development. Proceedings of the IEEE International Conference on Acoustics, Speechand Signal Processing (ICASSP '92), Vol. 1, pp. 517–520, San Francisco, USA, 1992.

[11] Hajic, J.: Building a Syntactically Annotated Corpus: the Prague Dependency Treebank. Issues of Valency and Meaning (Studies in Honor of Jarmila Panevova), pp. 106–132, Charles University Press, Prague, Czech Republic, 1998.

[12] Hajic, J.—Krbec, P.—Kveton, P.—Oliva, K.—Petkevic, V.: Serial Combination of Rules and Statistics: A Case Study in Czech Tagging. Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001), pp. 260–267, Toulouse, France, July 9–11, 2001.

[13] van Halteren, H.—Zavrel, J.—Daelemans, W.: Improving Accuracy in Word Class Tagging Through the Combination of Machine Learning Systems. Computational Linguistics, Vol. 27, 2001, No. 2, pp. 199–229.

[14] Johansson, S.—Atwell, E.—Garside, R.—Leech, G.: The Tagged LOB Corpus: User's Manual. Norwegian Computing Centre for the Humanities, Bergen, Norway, 1986, http://www.comp.lancs.ac.uk/ucrel/local/lob.

[15] Khorsi, A.: Towards Hybridization of Knowledge Representation and Machine Learning. Computing and Informatics, Vol. 26, 2007, No. 2, pp. 123–147.

[16] Kuta, M.—Polak, S.—Palacz, B.—Miłos, T.—Słota, R.—Kitowski, J.: TeToN – A Jena-Based Tool for Text-To-Ontology Approach. Proceedings of the 6th Cracow Grid Workshop, Vol. 2, pp. 98–105, Cracow, Poland, 2006.

[17] Kuta, M.—Chrzaszcz, P.—Kitowski, J.: A Case Study of Algorithms for Morphosyntactic Tagging of Polish Language. Computing and Informatics, Vol. 26, 2007, No. 6, pp. 627–647.

[18] Leech, G.—Garside, R.—Bryant, M.: CLAWS4: The Tagging of the British National Corpus. Proceedings of the 15th International Conference on Computational Linguistics (COLING 94), pp. 622–628, Kyoto, Japan, 1994.

[19] Manning, C. D.—Schutze, H.: Foundations of Statistical Natural Language Processing. ISBN 0-262-13360-1, 718 pp., MIT Press 1999.

[20] Marciniak, M.—Mykowiecka, A.—Przepiorkowski, A.—Przepiorkowski, A.—Kupsc, A.: An HPSG-Annotated Test Suite for Polish, Treebanks. Building and Using Parsed Corpora. ISBN 1-402-01334-5, 440 pp., Kluwer Academic Publishers, pp. 129–146, 2003.

[21] Marcus, M.—Santorini, B.—Magerman, D.: First Steps Towards an Annotated Database of American English. Readings for Tagging Linguistic Information in a Text Corpus, Tutorial for the 28th Annual Meeting of the Association for Computational Linguistics, University of Pittsburgh, Pittsburgh, USA, June 6–9, 1990.

[22] Marcus, M.—Santorini, B.—Marcinkiewicz, M.: Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics, Vol. 19, 1993, No. 2, pp. 313–330.

[23] Mauco, M.—Leonardi, M.: A Derivation Strategy for Formal Specifications from Natural Language Requirements Models. Computing and Informatics, Vol. 26, 2007, No. 4, pp. 421–445.

[24] Przepiorkowski, A.: The IPI PAN Corpus: Preliminary version. Polish Academy of Sciences, Warsaw, Poland, 2004.

[25] Ratnaparkhi, A.: A Maximum Entropy Model for Part-Of-Speech Tagging. Proceedings of the First Conference on Empirical Methods in Natural Language Processing, pp. 133–142, University of Pennsylvania, USA, 1996.

[26] Schroder, I.: A Case Study in Part-of-Speech Tagging Using the ICOPOST Toolkit. Technical report FBI-HH-M-314/02, Department of Computer Science, University of Hamburg, Hamburg, Germany, 2002.

[27] Varadi, T.: The Hungarian National Corpus. Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), pp. 385–396, Las Palmas de Gran Canaria, Spain, 2002.

[28] Wolinski, M.—Przepiorkowski, A.: Project of a Morphosyntactic Tagset for Polish. IPI PAN report No. 938, Polish Academy of Sciences, Warsaw, Poland, December 2001 (in Polish).

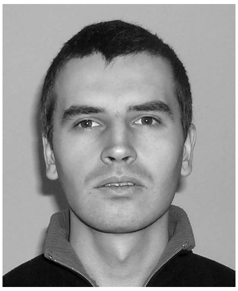
[29] The corpus of Frequency Dictionary of Contemporary Polish (FDCP corpus). `http://www.mimuw.edu.pl/polszczyzna`.

[30] The IPI PAN Corpus resources. `http://korpus.pl`.

[31] The Modified Corpus of Frequency Dictionary of Contemporary Polish (m-FDCP Corpus). `http://nlp.icsr.agh.edu.pl`.

[32] The PELCRA Corpus Resources. `http://korpus.ia.uni.lodz.pl`.

[33] The PWN Corpus Resources. `http://korpus.pwn.pl`.



Marcin Kuta received the M. Sc. degree in computer science in 2001 at the AGH University of Science and Technology in Krakow (Poland). Since 2003 he works at the Institute of Computer Science of the AGH University of Science and Technology, where he teaches compiler techniques and formal languages. His research interests comprise natural language processing, machine learning techniques, ontology usage, question answering systems and knowledge engineering.

Paweł Chrzaszcz is currently a masters student of computer science at the AGH University of Science and Technology in Krakow (Poland). His research interests are in natural language processing, concurrent programming (especially Erlang), parallel and distributed computing environments and computational complexity problems.

Jacek Kitowski is the Head of the Computer Systems Group at the Institute of Computer Science of the AGH University of Science and Technology in Cracow, Poland. He also works for the Academic Computer Centre CYFRONETAGH, being responsible for developing high-performance systems. He is the author or co-author of about 200 scientific papers. His topics of interest include large-scale computations, multiprocessor architectures, high availability systems, network computing, Grid services and Grid storage systems, knowledge engineering. He participates in program committees of many conferences, and has been involved in many national and international (EU) projects: CrossGrid, Pellucid and K-WfGrid. At present he participates in GREDIA and int.eu.grid projects.