Towards Morphologically Annotated Corpus of Hospital Discharge Reports in Polish

Małgorzata Marciniak
Institute of Computer Science PAS
ul. J.K. Ordona 21,
01-237 Warszawa, Poland
mm@ipipan.waw.pl

Agnieszka Mykowiecka
Institute of Computer Science PAS
ul. J.K. Ordona 21,
01-237 Warszawa, Poland
agn@ipipan.waw.pl

Abstract

The paper discuses problems in annotating a corpus containing Polish clinical data with low level linguistic information. We propose an approach to tokenization and automatic morphologic annotation of data that uses existing programs combined with a set of domain specific rules and vocabulary. Finally we present the results of manual verification of the annotation for a subset of data.

1 Introduction

Annotated corpora are knowledge resources indispensable to the design, testing and evaluation of language tools. Medical language differs significantly from the everyday language used in newspapers, magazines or fiction. Therefore, general language corpora are insufficient when creating tools for (bio)medical text processing.

There are several biomedical corpora available for English such as GENIA (Kim et al., 2010) — the best known and most used one, containing MEDLINE abstracts annotated on several levels; BioInfer (Pyysalo et al., 2007) targeted at protein, gene, and RNA relationships annotation; or CLEF (Roberts et al., 2009) containing 20,000 cancer patient records annotated with clinical relations. Medical corpora are also collected for lesser spoken languages, e.g. MEDLEX — Swedish medical corpus (Kokkinakis, 2006); IATROLEXI project for Greek (Tsalidis et al., 2007); or Norwegian corpus of patients' histories (Røst et al., 2008). The paper (Cohen et al., 2005) contains a survey of 6 biomedical corpora. The authors emphasize the importance of a standard format and give guidelines for careful annotation and evaluation of corpora.

The immediate goal of the paper is to establish and test a method of annotating Polish clinical data with low level linguistic information, i.e. token and morpheme descriptions. The research is done on a relatively small set of data (more than 450,000 tokens) but to gain the experience necessary to create a much larger annotated corpus of Polish medical texts. We would like to use our corpus to refine and test domain tools for: tagging, Named Entity Recognition or annotation of nominal phrases. We have already annotated the corpus with semantic information (Marciniak and Mykowiecka, 2011) using an existing rule based extraction system (Mykowiecka et al., 2009) and performed experiments with machine learning approaches to semantic labeling (Mykowiecka and Marciniak, 2011). Thus, to enable the realization of various scientific goals, a detailed and universal morphologic annotation of the corpus was introduced.

The division into tokens is the first level of text analysis. It is frequently performed without paying special attention to potential problems, just by dividing text on spaces, line breaks and punctuation marks. In many applications this is quite a satisfactory solution, but in case of texts that contain a lot of non-letter characters, using universal tokenization rules frequently causes problems. Some examples, in the case of using the Penn Treebank tokenization scheme in annotating the GENIA corpus were pointed out in (Teteisi and Tsujii, 2006). Jiang and Zhai (2007) show the importance of tokenization strategies in the biomedical domain, and the in-

Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011, pages 92–100,
Portland, Oregon, USA, June 23-24, 2011.

fluence of this process on the results of information retrieval. Our approach consists of dividing text into simple tokens which can be grouped at subsequent levels of analysis using domain specific knowledge.

For languages with rich inflection, like Polish, morphological annotation is indispensable for further text analysis. As there are no Polish taggers which can analyze medical texts, nor medical lexicons containing inflected forms, we combine a general purpose tagger with a set of domain specific rules referring to a small data induced vocabulary. A portion of the automatically annotated data was checked by two linguists to assess data quality. The results obtained are given in 8. Currently, the entire dataset is undergoing manual verification.

2 Linguistic Characteristics of Texts

The corpus consists of 460 hospital discharge reports of diabetic patients, collected between the years 2001 and 2006 in one of Warsaw's hospitals. These documents are summaries of hospital treatment and are originally written in MS Word with spelling correction turned on, so the errors observed are mainly in words that are not included in the dictionary. The documents are converted into plain text files to facilitate their linguistic analysis and corpus construction. Clinical data include information serving identification purposes (names and addresses) which are substituted by symbolic codes before making the documents accessible for further analysis. The annonymization task was performed in order to make the data available for scientific purposes. We plan to inspect the data manually, to remove all indirect information enabling a patient's identification, and negotiate the terms for making the corpus publicly available.

Each document is 1.5 – 2.5 pages long, and begins with the identification information of the patient and his/her visit in hospital. Next, the following information is given in short form: significant past and current illnesses, diagnoses and patient's health at the beginning of the hospitalization. After these data, the document describes results of examinations such as height, weight, BMI and blood pressure, ophthalmology examinations, blood tests, lipid profile tests, radiology or ultrasound. This part of the document may also contain descriptions of attempts to select the best treatment for the patient. The summary of the document starts from the word *Epikryza* 'Discharge abstract'. Its length is about half a page of text. It contains: data about a patient's diabetes, a description of diabetic complications, and other illnesses, selected examination results and surgical interventions, information about education, diet observed, self monitoring, patient's reactions, and other remarks. Finally, all recommendations are mentioned, including information about prescribed diet, insulin treatment (type and doses) and oral medication.

Most information is given as free-form text, but the vocabulary of these documents is very specific, and significantly differs from texts included in corpora of general Polish like IPIPAN Corpus (Przepiórkowski, 2004) or NKJP (National Corpus of Polish, `http://nkjp.pl`). The texts contain many dates in different formats, and a lot of test results with numerical values, whose descriptions are omitted in NKJP. The texts contains also a lot of medication names, like *Cefepime* or *Acard* not present in any general Polish dictionary. Some of them are multi-word names like *Diaprel MR, Mono Mack Depot, Mixtard 10*. The same medication can be referred to in different ways depending on international or Polish spelling rules (e.g. *Amitriptylinum* and its Polish equivalent *Amitryptylina*). Polish names could be inflected by cases (e.g. *Amitryptyliny*$_{gen}$).

In documents, many diagnoses are written in Latin. In the following examples the whole phrases are in Latin: *Retinopathia diabetica simplex cum maculopathia oc. sin.* 'simple diabetic retinopathy with maculopathy of the left eye'; or *Laryngitis chronica. Otitis media purulenta chronica dex.* 'Chronic laryngitis. Chronic purulent inflammation of the middle right ear'. Sometimes foreign expressions are thrown into a Polish sentences: *Ascites duża ilość płynu w jamie brzusznej między pętlami jelit ...* 'Ascites a lot of fluid in abdominal cavity between intestinal loops ...' — only the first word is not in Polish.

3 Corpus description

The corpus is annotated with morphological and semantic information. The standard of annotation fol-

lows the TEI P5 guidelines advised for annotation of biomedical corpora, see (Erjavec et al., 2003). Our corpus format is based on the one accepted for the NKJP corpus (Przepiórkowski and Bański, 2009). According to this scheme, every annotation is described in a separate file. Each discharge document is represented by a catalog containing the following five files:

- *xxx.txt* – plain text of the original annonymized document;
- *xxx.xml* – text of the document (in the form as in *xxx.txt* file) divided into numbered sections which are in turn divided into paragraphs;
- *xxx_segm.xml* – token limits and types (29 classes);
- *xxx_morph.xml* – morphological information (lemmas and morphological feature values);
- *xxx_sem.xml* – semantic labels and limits.

4 Tokenization

The first level of text analysis is its segmentation into tokens. In general, most tokens in texts are lowercase words, words beginning with a capital letter and punctuation marks. The most common (thus the most important) tokenization problem is then to decide whether a particular dot ends a sentence or belongs to the preceding abbreviation (or both). In some texts there are also many numbers representing dates, time points, time intervals or various numerical values. For texts in which uniform standards of expressing these notions are obeyed, recognizing such complex tokens is much easier and simplifies further text analysis.

In medical texts the problem of non-word tokens is harder than in the case of newspapers or novel content as they constitute a much larger portion of the text itself. Apart from descriptions of time (dates, hours, periods of time) there are numbers that refer to values of different medical tests or medicine doses and sizes. There are also many specific names which sometimes contain non-letter characters (e.g. *Na+*) as well as locally used abbreviations and acronyms. An additional difficulty is caused by the lack of will to obey writing standards. Physicians use different ways of describing dates (e.g. *02.09.2004, 30.09/1.10.2003, 06/01/2004, 14.05.05, 28 .04. 05, 12.05.2005r.*) or time (*8:00 vs 8.00*). They also do not pay enough attention to punctuation rules and mix Polish and English standards of writing decimal numbers. In Polish we use a comma not a dot, but the influence of English results in common usage of the decimal point. Sometimes both notations can be found in the same line of text. Further, the sequence *'2,3'* may mean either *'2.3'* or two separate values: *'2'* and *'3'*.

Two tools used in the process of constructing the corpus have embedded tokenizers. The first one is a part of the information extraction system SProUT (Drożdżyński et al., 2004) which was used to write grammars identifying semantically important pieces of text. The general assumption adopted while building its tokenizer was "not to interpret too much", which means that tokens are relatively simple and do not rely on any semantic interpretation. Their self explanatory names, together with token examples and their frequencies in the entire input data set, are listed in table 1.

Two other tokenization modules are embedded in the TaKIPI tagger used to disambiguate the morphological descriptions of word forms (Piasecki, 2007). The first one divides all character sequences into words and non-words which are assigned the *ign* label. The second tokenizer interprets these non-word sequences and assigns them *ttime, tdate, turi* (for sequences with dots inside) and *tsym* labels. It also applies a different identification strategy for token limits – for all non-word tokens only a space or a line break ends a token. Although treating a date (*15.10.2004r*) or a range (*1500-2000*) as one token is appropriate, in the case of sequences where spaces are omitted by mistake, the resulting tokens are often too long (e.g. *'dnia13/14.07.04', 'iVS-1,5'*).

After analyzing the results given by three different tokenizers we decided to use the token classes identified by the SProUT tokenizer and align its results with the results of the 'simple' TaKIPI tokenizer. SProUT tokens which were longer than TaKIPI tokens, e.g. *'1x2mg', '100mg', '50x16x18'*, were divided into smaller onces. The changes introduced to token limits concern those tokens of the *other_symbol* type which contain punctuation marks. The *other_symbol* class comprises sequences which do not fit into any other class, i.e.

symbols for which separate classes are not defined (e.g. '=') and mixed sequences of letters and digits. In this latter case a token ends only when a space or a line break is encountered. The most typical case when this strategy fails in our data is the sequence *'HbA1c:'* as the name of the test according to the tokenizer rules is classified as an 'other_symbol' the following colon is not separated. There are also other similar sequences: *'HbA1c=9,1%:'* or *'(HbA1C'*. To make the results more uniform we divided these tokens on punctuation characters. This process resulted in replacing 1226 complex tokens by 4627 simple ones. Among these newly created tokens the most numerous class was *lowercase_word* and numbers which were formed after separating numbers and unit names, e.g. *10g, 100cm* and sequences describing repetitions or sizes, like *2x3, 2mmx5mm*. The longest sequence of this kind was *'ml/min.,GFR/C-G/-37,5ml/min/1,73m2'*. This string was divided into 18 tokens by TAKIPI but finally represented as 23 tokens in the corpus. Finally, in the entire data set 465004 tokens (1802864 characters) were identified. The most numerous class represents numbers – 18.8% (9% of characters), all punctuation characters constitute 25% of the total number of tokens (6.5% characters).

Table 1: Token types and number of occurrences

	numbers	
token class name & examples	initial	final
all_capital_word: ALT, B, HDL, HM	18369	18416
any_natural_number	85766	87246
apostrophe	14	14
back_slash	7	7
closing_bracket	2661	2663
colon	12426	12427
comma	28799	28831
dot	47261	47269
exclamation_sign	49	49
first_capital_word: Al, Amikacin, Wysokie	43136	43269
hyphen	4720	4725
lowercase_word: antygen, aorta	192305	193368
mixed_word_first_capital: AgHBs, IIo, NovoRapid	513	514
mixed_word_first_lower: antyHBS, dlAST	989	1003
number_word_first_capital: 200Hz, 14HN	48	0
number_word_first_lower: 100ml, 200r 1kaps	650	0
opening_bracket	3344	3355
other_symbol: (132x60mm), 1,34x3,25, HbA1c=10,3%,	3161	2868
percentage_tok	4461	4478
question_mark	207	209
quotation	1	1
semicolon	455	455
slash	10340	10353
word_number_first_capital: AST34, B6	1195	1195
word_number_first_lower: mm3, pH6	1865	1854
word_with_hyphen_first_capital: B-hCG, Anty-HBs	163	163
word_with_hyphen_first_lower: m-ce, p-ciał	402	402
all tokens	463307	465004

5 Morphological analysies

Morphological annotation was based on the results obtained by the publicly available Polish POS tagger TaKIPI that cooperates with *Morfeusz* SIAT (Woliński, 2006) — a general-purpose morphological analyzer of Polish. For each word, it assigns all possible interpretations containing: its base form, part of speech, and complete morphological characterization (e.g. case, gender, number, aspect if relevant). The description is exhaustive and aimed at further syntactic analyses of texts.

The annotation is done in three steps. In the first one the documents are analyzed and disambiguated by TaKIPI. TaKIPI can be combined with the *Guesser* module (Piasecki and Radziszewski, 2007) which suggests tags for words which are not in the dictionary. We decided to use this module because otherwise 70600 tokens representing words and acronyms that occur in the documents would be assigned an unknown description. The gain from its usage is however not so evident, as tags and base forms suggested by *Guesser* are quite often incorrect – in one test set, only 272 forms out of 1345 were analyzed correctly.

The analyses of TaKIPI results shows that there are many systematic errors. They can be corrected globally. An example of such an error is the description of medication names produced by *Guesser*. Their morphologic tags are often correct, but the problem is with gender assignment in case of masculine forms. In Polish there are three subtypes of masculine gender: personal, animate and inanimate, and *Guesser* quite often uses personal masculine gender instead of the inanimate one while analyzing medication names. The second most common problem concerns base forms, because all base forms created by the module are written with a small letter. So in the case of proper names, all base forms have to be corrected. Moreover, TaKIPI do not disambiguate all tags – certain forms still have more than one possible description.

Thus, to limit the number of manual changes needed in the final version of the corpus, we post-process the results with a set of rules (see section 7) created on the basis of a list of all different token descriptions. The rules mainly correct the annotations of domain related tokens like acronyms and units: *BMI, HbA1c, RR, USG, Hz* or *kcal*; medication names e.g. *Diaprel*, its *diaprel* base form is changed into *Diaprel*; and other domain terms like *dekarboksylazie* ('decarboxylase$_{loc}$') for which the masculine base form was suggested *dekarboksylaz* instead of feminine *dekarboksylaza*. Moreover, tags of misspelled tokens and foreign words are assigned to tokens during this stage and if there is more than one description attached to a token, then the more probable in the domain is chosen.

Finally, the morphology analyses are manually corrected. This is done by two linguists. The results are compared and corrected by a third annotator. The first results are described in section 8.

6 Tags

For each token, TaKIPI assigns its base form, POS, and full morphological description. For example, the token *badania* that has the base form *badanie* 'examination' is classified in all 579 occurrences as a neutral noun. In 566 cases it is classified as a singular form in genitive and is assigned the tag **subst:sg:gen:n** (substantive:singular:genitive:neutral); in 13 cases as a plural noun including 8 nominative forms, 4 accusative and even one vocative (unreliable in medical texts). TaKIPI assigns the unknown tag (ign) to numbers, so we introduced the **number** tag to represent numerical values in the corpus. It is assigned to 18.8% of tokens.

The set of potential morphological tags consists of more than 4000 elements. In our corpus only 450 different tags are represented, in comparison to over 1000 tags used in the general Polish IPIPAN corpus (Przepiórkowski, 2005).

In the rest of this section we describe tags used for the classification of strings that are not properly classified by TaKIPI. If no tag described in the section suits a token, the tag **tsym** is assigned to it. In particular, all patient codes (like *d2005_006*) have the **tsym** tag.

6.1 Errors

Spelling errors in the corpus are left as they are. Misspelled tokens are assigned the base form equal to the token, and one of the following tags depending on the type of error:

- **err_spell** describes misspelled tokens like *bia3ko* instead of *białko* ('protein'). In the corpus we provide additional information with the corrected input token, its base form and morphological tag.
- **err_conj** describes concatenations like *cukrzycowej2000* ('diabetic2000'). In this case we add the correct form *cukrzycowej 2000* to the corpus but do not add its description.
- **err_disj_f** describes the first part of an incorrectly disjointed word. For example the word *ciśnienie* ('pressure') was divided into two parts *ci* and *śnienie*, (by chance, both are valid Polish words).
- **err_disj_r** describes the second part of the incorrectly disjointed word.

The last three categories can be supplemented with **spell** description if necessary. For example the token *Bylaw* is a concatenation of the misspelled word *Była* ('was') with the preposition *w* ('in'). This token has the tag **err_conj_spell**, and the *Była w* correction is added.

6.2 Abbreviations

There are many abbreviations in the documents. Some of them are used in general Polish like *prof* ('professor') or *dr* ('doctor), but there are many abbreviations that are specific to the medical domain. For example in the descriptions of USG examinations the letter *t* denotes *tętnica* ('artery'), while *tt* refers to the same word in plural, although usually there is no number related difference e.g. *wit* ('vitamin') can be used in plural and singular context. Sometimes it is not a single word but the whole phrase which is abbreviated, e.g. *NLPZ* is the acronym of the noun phrase *Niesterydowe Leki PrzeciwZapalne* 'Non-Steroidal Anti-Inflammatory Drugs', and *wpw* is an abbreviation of the prepositional phrase *w polu widzenia* 'in field of view'.

Abbreviations and acronyms obtain the tag **acron**. Moreover, it is possible to insert the full form corresponding to them.

Acronyms denoting units obtain the tag **unit**. Units in common usage are not explained: *mm, kg, h*, but if a unit is typical to the medical domain, its full form is given (e.g. *HBD* means *tydzień ciaży* 'week of pregnancy').

We also distinguish two tags describing prefixes and suffixes. The token *makro* ('macro') in the phrase *makro i mikroangiopatia* ('macro and microangiopathy') has the tag **prefix**, while the **suffix** tag describes, for example, the part *ma* of the string *10-ma* which indicates instrumental case of number 10, like in: *cukrzyca rozpoznana przed 10-ma laty* ('diabetes diagnosed 10 years ago').

6.3 Foreign Words

Foreign words receive the **foreign** tag. This tag can be elaborated with information on the part of speech, so for example, *Acne* has the tag **foreign_subst**. It is possible to attach a Polish translation to foreign words.

7 Correction Rules

Correction rules are created on the basis of a list of different tokens, their base form, and tags that occurred in the corpus. Each rule is applied to all matching form descriptions of tokens in the already tagged data.

We use the method of global changes because we want to decrease the number of manual corrections in the corpus on the final, manual stage. It should be noted that without context it is impossible to correct all morphological tags. We can only eliminate evident errors but we cannot decide, for example, if a particular description of a token *badanie* 'examination' (see section 6) is correct or not. All these tags can be verified only if we know the context where they occurred. However, quite a lot of changes can be made correctly in any context, e.g. changes of gender of a medication name (*Lorinden*$_f$ into *Lorinden*$_{m3}$), or in the prevailing number of cases, e.g. assigning to *zwolnienie* the *gerund* tag 'slowing' (11 occurrences) instead of less frequent in the texts *noun* 'sick leave' only one occurrence (TaKIPI leaves both descriptions).

There are two main types of correction rules of which syntax is given in (1–2). '#' is a separator; the character '>' indicates the new token description that is applied to the corpus; after || additional information can be noted. In case of rule (1) it could be a text that explains the meaning of acronyms, abbreviations or foreign words, while for rule (2), a corrected token, base form and tag can be given. This additional information might be used for creating a corpus without spelling errors, dictionaries of abbreviations or foreign words used in the medical domain.

(1) token#base form#tag#>
token#new base form#new tag#
|| 'string' (optionally)

(2) token#base form#tag#>
token#token#error_spell# ||
corr. token#corr. base form#new tag#

The first scheme is useful for changing the base form or the tag of a token. See example (3) where the first letter of the base form is capitalized and personal masculine gender *m1* is changed into inanimate masculine gender *m3*.

(3) Insulatard#insulatard#subst:sg:nom:m1#>
Insulatard#Insulatard#subst:sg:nom:m3#

The second scheme is applied to a token *graniach* 'ridges' (in mountain) that represents the existing but unreliable word in the medical domain. For all of its occurrences in our data (3 cases) it is substituted by *granicach* 'limits' by the following correction rule:

(4) graniach#grań#subst:pl:loc:f#>
granicach#granicach#err_spell# ||
granicach#granica#subst:pl:loc:f#

If there is more than one interpretation left by TaKIPI, all are mentioned before the character '>'. See example (5) where two different base forms are possible for the token *barku* and both have the same tag assigned. The first base form *bark* ('shoulder') is definitely more probable in the medical domain than the second one *barek* ('small bar' or 'cocktail cabinet'), so the rule chooses the first description.

(5) barku#bark#subst:sg:gen:m3##barek#
subst:sg:gen:m3#>barku#bark#subst:sg:gen:m3#

Table 2 presents the frequencies of top level morphological classes: directly after running the tagger, after changing the token limits and after applying automatic changes. In the last column the number of different forms in every POS class is presented.

Most part of speech names are self explanatory, the full list and description of all morphological tags can be found in (Przepiórkowski, 2004), the newly introduced tags are marked with *. Of all words (all tags apart form *interp*unction, *number* and *tsym*) the most numerous groups are nouns (*subs*tantive) – 54% and *adj*ectives – 15% of wordform occurrences.

Table 2: Morpheme types and numbers of occurrences

	tagger results	after tok. change	final corpus	
POS tag	number of tag occurences			different forms
adj	35305	35041	36848	3576
adv	2323	2323	2437	245
conj	5852	5852	5680	36
prep	29400	29400	26120	71
pron	302	302	142	21
subst	82215	82215	105311	5093
verb forms:	24743	24741	19912	2001
fin	2173	2173	1900	190
ger	9778	9778	4677	423
ppas	5593	5593	6170	551
other	7199	7197	7165	837
qub	4244	4242	2452	67
num	703	703	703	34
ign	160951	163629	0	0
acron*	0	0	30003	678
unit*	0	0	28290	82
prefix*	0	0	13	5
suffix*	0	0	36	6
tsym*	0	0	534	462
interp	115323	116556	116556	21
number*	0	0	87898	1386
err_disj*	0	0	179	129
err_spell*	0	0	560	440
foreign*	0	0	1330	184
total	461361	465004	465004	14537

If we don't take into account *number*, *tsym* and the punctuation tokens, we have a corpus of 348461 tokens (TW) out of which 78854 (29.81%) were changed. The most frequent changes concerned introducing domain related *unit* and *acronym* classes (nearly 72% of changes). Quite a number of changes were responsible for the capitalization of proper name lemmata. In table 3 the numbers of some other types of changes are presented.

Table 3: Morphological tag changes

type of change	number	% of changes	% of TW
base form			
capitalization only	6164	13.8	4.12
other	25503	32.34	9.64
POS			
to acron & unit	56697	71.90	21.43
to other	10547	13.37	3.99
grammatical features (without acron and unit)			
only case	109	0.13	0.04
only gender	1663	2.11	0.62
other	13215	16.75	4.99

Table 4: Manual correction

	basic tags	all tags
all tokens	8919	8919
without numbers and interp	4972	4972
unchanged	4497	4451
changed	475	521
same changes accepted	226	228
same changes not accepted	1	1
different changes none accepted	4	5
different changes. accepted 1	3	4
different changes. accepted 2	40	42
only 1st annot. changes - accepted	15	48
only 2nd annot. changes - accepted	128	124
only 1nd annot. changes - not accepted	47	47
only 2nd annot. changes - not accepted	0	0

8 Manual Correction

The process of manual correction of the corpus is now in progress. It is performed using an editor specially prepared for visualization and facilitation of the task of correcting the corpus annotation at all levels. In this section we present conclusions on the bases of 8 documents corrected by two annotators (highly experienced linguists). In the case of inconsistent corrections the opinion of a third annotator was taken into account. The process of annotation checking took about 2x20 hours.

From a total number of 8919 tokens in the dataset, the verification of 4972 (words, acronyms, units) was essential, the remaining 3947 tokens represent numbers, punctuation and **tsym** tokens. The correction rules changed the descriptions of 1717 (34%) tokens, only 87 cases were limited to the change of a lowercase letter into a capital letter of the base form. Manual verification left 4497 token descriptions unchanged, while 10.6% of descriptions were modified (evaluation of TaKIPI by Karwińska and Przepiórkowski (2009) reports 91.3% accuracy). Kappa coefficient was equal to 0.983 for part of

speech and 0.982 for case assignment (when it is applicable). The results of manual correction are given in table 4. The 'basic tags' column gives the number of changes of the base form and tag, while the 'all tags' column takes into account all changes, including descriptions of the correct word form in case of spelling errors, explanations of acronyms or units.

More detailed analysis of annotation inconsistencies shows two main sources of errors:

- lack of precision in guidelines resulted in choosing different base forms in case of spelling errors and different labeling of cases with the lack of diacritics which resulted in correct but not the desired forms;
- some errors were unnoticed by one of the annotators (just cost of manual work), e.g. in the data there are many strings 'W' and 'w' which may be either acronyms or prepositions.

There are only a few cases that represent real morphological difficulties, e.g. differentiating adjectives and participles (5 cases among the annotators). Some examples of different case and gender assignments were also observed. They are mostly errors consisting in correcting only one feature instead of two, or a wrong choice of a case for long phrases.

9 Conclusions and Further Work

The problems described in the paper are twofold, some of them are language independent like tokenization, description of: abbreviations, acronyms, foreign expressions and spelling errors; while the others are specific for rich-morphology languages. Our experiment showed that analyzing specialized texts written in highly inflected language with a general purpose morphologic analyzer can give satisfactory results if it is combined with manually created global domain dependent rules. Our rules were created on the basis of a sorted list of all token descriptions. That allowed us to analyze a group of tokens with the same base form e.g. an inflected noun. Additional information concerning the frequency of each description, indicated which token corrections would be important.

Unfortunately, the process of rule creation is time-consuming (it took about 90 hours to create them). To speed up the process we postulate to prepare three sets of tokens for which rules will be created separately. The first one shall contain tokens which are not recognized by a morphological analyzer, and hence requiring transformation rules to be created for them. The second set shall contain tokens with more than one interpretation, for which a decision is necessary. Finally we propose to take into account the set of frequent descriptions. Infrequent tokens can be left to the manual correction stage as it is easier to correct them knowing the context.

At the moment our corpus contains three annotation levels – segmentation into tokens, morphological tags and semantic annotation. After the first phase of corpus creation we decided to introduce an additional level of annotation — extended tokenization, see (Marcus Hassler, 2006). Current tokenization divides text into simple unstructured fragments. This solution makes it easy to address any important fragment of a text, but leaves the interpretation of all complex strings to the next levels of analysis. A new extended tokenization is planned to create higher level tokens, semantically motivated. It will allow the annotation of complex strings like: dates (*02.12.2004, 02/12/2004*); decimal numbers; ranges (*10 - 15*, *10-15*); sizes and frequencies (*10 x 15, 10x15*); complex units (mm/h);abbreviations with full stops (*r. – rok* 'year'); acronyms containing non-letter characters (*K+*); complex medication names (*Mono Mack Depot*).

Extended tokens can be recognized by rules taking into account two aspects: specificity of the domain and problems resulting from careless typing. In the case of abbreviations and acronyms, the best method is to use dictionaries, but some heuristics can be useful too. Electronic dictionaries of acronyms and abbreviations are not available for Polish, but on the basis of annotated data, a domain specific lexicon can be created. Moreover, we want to test ideas from (Kokkinakis, 2008), the author presents a method for the application of the MeSH lexicon (that contains English and Latin data) to Swedish medical corpus annotation. We will use a similar approach for acronyms and complex medication name recognition.

References

K. Bretonnel Cohen, Lynne Fox, Philip V. Ogren, and Lawrence Hunter. 2005. Corpus design for biomedical natural language processing. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, pages 38–45, Detroit, June. Association for Computational Linguistics.

Witold Drożdżyński, Hans-Ulrich Krieger, Jakub Piskorski, Ulrich Schäfer, and Feiyu Xu. 2004. Shallow Processing with Unification and Typed Feature Structures — Foundations and Applications. *German AI Journal KI-Zeitschrift*, 01/04.

Toma Erjavec, Yuka Tateisi, Jin dong Kim, Tomoko Ohta, and Jun ichi Tsujii. 2003. Encoding Biomedical Resources in TEI: the Case of the GENIA Corpus. In *Proceedings of the ACL 2003, Workshop on Natural Language Processing in Biomedicine*, pages 97–104.

Jing Jiang and Chengxiang Zhai. 2007. An Empirical Study of Tokenization Strategies for Biomedical Information Retrieval. *Information Retrieval*, 10(4–5):341–363.

Danuta Karwańska and Adam Przepiórkowski. 2009. On the evaluation of two Polish taggers. In *The proceedings of Practical Applications in Language and Computers PALC 2009*.

Jin-Dong Kim, Tomoko Ohtai, and Jun'ichi Tsujii. 2010. Multilevel Annotation for Information Extraction Introduction to the GENIA Annotation. In *Linguistic Modeling of Information and Markup Languages*, pages 125–142. Springer.

Dimitrios Kokkinakis. 2006. Collection, Encoding and Linguistic Processing of a Swedish Medical Corpus – The MEDLEX Experience. In *Proceedings of the Fifth International Language Resources and Evaluation (LREC'06)*, pages 1200–1205.

Dimitrios Kokkinakis. 2008. A Semantically Annotated Swedish Medical Corpus. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 32–38.

Małgorzata Marciniak and Agnieszka Mykowiecka. 2011. Construction of a medical corpus based on information extraction results. *Control & Cybernetics*. in preparation.

Günther Fliedl Marcus Hassler. 2006. Text preparation throughextended tokenization. *Data Mining VII: Data, Text and Web Mining and their Business Applications*, 37.

Agnieszka Mykowiecka and Małgorzata Marciniak. 2011. Automatic semantic labeling of medical texts with feature structures. In *The Text Speech and Dialogue Conference 2011 (submitted)*.

Agnieszka Mykowiecka, Małgorzata Marciniak, and Anna Kupść. 2009. Rule-based information extraction from patients' clinical data. *Journal of Biomedical Informatics*, 42:923–936.

Maciej Piasecki and Adam Radziszewski. 2007. Polish Morphological Guesser Based on a Statistical A Tergo Index. In *2nd International Symposium Advances in Artificial Intelligence and Applications (AAIA'07), wisła, Poland*, pages 247–256.

Maciej Piasecki. 2007. Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly*, 11(1–2):151–167.

Adam Przepiórkowski and Piotr Bański. 2009. XML text intechange format in the National Corpus of Polish. In *The proceedings of Practical Applications in Language and Computers PALC 2009*, pages 245–250.

Adam Przepiórkowski. 2004. *Korpus IPI PAN. Wersja wstępna / The IPI PAN Corpus: Preliminary version*. IPI PAN.

Adam Przepiórkowski. 2005. The IPI PAN Corpus in numbers. In Zygmunt Vetulani, editor, *Proc. of the* 2nd Language & Technology Conference, Poznań, Poland.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Jörvinen, and Tapio Salakoski. 2007. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8.

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, and Andrea Setzer. 2009. Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*, 42(5):950–966.

Thomas Brox Røst, Ola Huseth, Øystein Nytrø, and Anders Grimsmo. 2008. Lessons from developing an annotated corpus of patient histories. *Journal of Computing Science and Engineering*, 2(2):162–179.

Yuka Teteisi and Jun'ichi Tsujii. 2006. GENIA Annotation Guidelines for Tokenization and POS Tagging. Technical report, Tsujii Laboratory, University of Tokyo.

Christos Tsalidis, Giorgos Orphanos, Elena Mantzari, Mavina Pantazara, Christos Diolis, and Aristides Vagelatos. 2007. Developing a Greek biomedical corpus towards text mining. In *Proceedings of the Corpus Linguistics Conference (CL2007)*.

Marcin Woliński. 2006. Morfeusz — a Practical Tool for the Morphological Analysis of Polish. In Mieczysław Kłopotek, Sławomir Wierzchoń, and Krzysztof Trojanowski, editors, *IIS:IIPWM'06 Proceedings, Ustron, Poland*, pages 503–512. Springer.