Control and Cybernetics

vol. **40** (2011) No. 2

Construction of a medical corpus based on information extraction results[*]

by

Małgorzata Marciniak and Agnieszka Mykowiecka

Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

Abstract: The paper presents a method of automatic construction of a semantically annotated corpus using the results of a rule-based information extraction (IE) application. Construction of the corpus is based on using existing programs for text tokenization and morphological analysis and combining their results with domain related correction rules. We reuse the specialized IE system to obtain a corpus annotated on the semantic level. The texts included within the corpus are Polish free text clinical data. We present the documents — diabetic patients' discharge records, the structure of the corpus annotation and the methods for obtaining the annotations. Initial evaluations based on the results of manual verification of selected data subset are also presented. The corpus, once manually corrected, is designed to be used for developing supervised machine learning models for IE applications.

Keywords: corpus, semantic annotation, clinical data, information extraction.

1. Introduction

The vast majority of computational linguistics research and natural language applications are now based on the analysis of real language data. This approach requires access to appropriately annotated text corpora. For several languages (e.g. English, German, Czech, Hungarian, Polish) there exist large general corpora annotated with morphological data, part of speech names and simple syntactic information. There are also some corpora annotated with Named Entities (persons, institutions, geographical names). Domain specific corpora with semantic annotation are much smaller and still not very common. However, their existence is crucial to the construction of natural language processing applications which depend on domain specific knowledge, as such applications have to take into account not only the domain terminology, but also the very many ways of expressing concepts in natural texts. For some areas of interest, like press

[*] Submitted: November 2010; Accepted: June 2011.

news, there are a lot of data available. For others, like medical clinical data, the availability of texts is very restricted. But the most important problem, which limits the development of corpora, apart from data availability and the possibility of their disclosure, is the high cost of the annotation process.

Although research in the domain of biomedicine, including biomedical NLP, is nowadays very intense, there are still few corpora containing clinical data. Only for English the situation is better. One of the most sophisticated annotated English biomedical corpora is CLEF, Roberts et al. (2009). The corpus is a result of *The Clinical E-Science Framework* project and it consists of structured records and free text documents from more than twenty thousand patients of Royal Marsed Hospital. The CLEF gold standard is a small subset of the corpus consisting of clinical narratives, histopathology reports and imaging reports which are annotated with information about clinical entities and the relations between them (e.g. *condition, intervention, locus, has_ target, has_finding*) and temporal information (relations and dates). Other examples of biomedical corpora are: GENIA corpus of Medline[1] abstracts annotated with information about biological entities (Kim et al., 2003) and biological events (Kim et al., 2008), PennBioIE corpus of more than two thousand Medline abstracts annotated for biomedical entity types and part of speech (Mandel, 2006), Yapex corpus of two hundred Medline abstracts annotated for protein names (Franzén et al., 2002), and BioScope corpus (Vincze et al., 2008), which consists of more than twenty thousand sentences taken from three different sources (Pestian et al., 2007), several full research papers and part of GENIA, annotated for the scope of negation and uncertainty. There are also some corpora prepared for various competitions in semantic tagging, but they usually contain only a very limited type of simple annotation (e.g. ICD codes, Pestian et al., 2007). An example of a public resource prepared for information extraction is BioInfer (Pyysalo et al., 2007) — a small corpus of 1100 sentences from abstracts of biomedical articles. It was manually annotated for relationships, name entities and syntactic dependences. Medical corpora are also collected for less spoken languages, e.g. MEDLEX — Swedish medical corpus (Kokkinakis, 2006) annotated with terminology and named entities; IATROLEXI project for Greek (Tsalidis et al., 2007); or Norwegian corpus of patients' histories (Røst et al., 2008). For Polish, medical corpora practically do not exist. We are aware of only one corpus with morphologically annotated Polish medical texts – KorMeDIIS (Piasecki et al., 2006).

Annotation is a fundamental issue in corpora construction. Nowadays, semantic annotation is usually done manually and this process is time consuming and costly. The typical annotation approach requires the preparation of a detailed annotation manual. Realization of this process involves at least two people who annotate each corpus fragment, and a super annotator responsible for re-

[1] MEDLINE (`http://www.nlm.nih.gov/pubs/factsheets/medline.html`) is the U.S. National Library of Medicine's (NLM) premier bibliographic database that contains over sixteen million references to journal articles in life sciences with an emphasis on biomedicine.

solving conflicts. Regardless of how carefully such manual annotation is done, there are always some errors and inconsistencies in the results. Another solution is automatic annotation done by rule based programs or machine learning (ML) techniques. This is the typical situation for morphological and part of speech (POS) annotations, although manual annotation is also proposed, e.g. (Pakhomova et al., 2006). For semantic annotation, this approach is less frequent, as for rule based methods it is difficult to foresee all possible text variants, and for ML techniques there is no annotated training data available. But for some domains, there already exist IE applications which recognize domain related objects or events in free text. Such applications are used to convert text data into database tables (Marciniak and Mykowiecka, 2007; Aramaki et al., 2009) but their results can also be used in annotation. These tasks are similar but not the same. In the task of database filling it is important to recognize that a piece of information is represented in textual form. Whereas, the task of annotating corpora consists in indicating a particular phrase that contains that information.

In this paper we describe how the rule based IE system developed for diabetic patients' discharge records (Mykowiecka et al., 2009) can be used in the semi-automatic process of constructing an annotated corpus, which contains both morphological and semantic annotation, and needs only a relatively small number of manual corrections. The system was implemented using SProUT — general-purpose IE platform (Drożdżyński et al., 2004) enriched with a Polish tokenizer and a morphological analyzer. In the rest of the paper we describe the data included in the corpus, the annotation format, and the method of corpus construction. We present an evaluation of the automatically built resource on a selected data subset. Our claim was that such a corpus could be of a quality good enough for some purposes (e.g. testing machine learning methods of IE) and can be further improved by manual verification if needed at a relatively low cost. The results of the first evaluation, which are shown in the paper (F-measure of 0.965 on about 50 attributes) support this claim. However, both the processing of specialized data and the reuse of applications meant for a slightly different purpose, posed problems which had to be solved during the construction of medical corpora.

2. Data in the corpus

The corpus consists of 460 diabetic patients' discharge reports from the Bródnowski Hospital in Warsaw from the years 2001-2006 (more than 450,000 tokens). These are official documents 1.5—2.5 pages long, written in MS Word, typed with spelling correction, so errors are observed mainly in words that are not included in MS Word's dictionary. Corpora containing clinical data are usually not very big. The corpus of clinical notes manually annotated for POS, described in Pakhomova et al. (2006) consists of 273 clinical notes and contains 100,650 tokens; 100,000 tokens were reported in the German clinical reports corpus (Wermter and Hahn, 2004). The size of some corpora are given only in numbers of patient records, for example CLEF collected 20,000 cancer patient records. Interesting data are reported by Røst et al. (2008), whereby authors collected 616,000 consultation reports concerning 12,000 patients from a general practice center.

The documents collected in the corpus are those that are given to patients as the summary of their hospital treatment. Each document concerns one patient's visit in the hospital. A particular visit is identified by two parameters: an identification number of the visit within the year, and the year itself. Each visit concerns one patient. Original documents contain patients' names and addresses but these are substituted by symbolic identification codes, before making the documents accessible.

Most information is given as free-form text but some data is written in table form, e.g. results of biochemical tests. Each document begins with the identification numbers of the visit and the patient, the age of the patient, and the start/end dates of the visit in hospital. Next, the following information is given in a short form: significant past and current illnesses; diagnoses; patient's health at the beginning of the hospitalization. After these initial data, the document contains results of examinations e.g.: basic data like height, weight, BMI (Body Mass Index), blood pressure; an ultrasound check up of the abdominal cavity; 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 most important part of the document starts from the word *Epikryza* 'Discharge abstract'. Its length is about half a page of text. It contains: data about the 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 insulin and oral medication types and doses.

3. Corpus annotation format

The corpus is annotated with morphological and semantic information. The standard of annotation follows the TEI P5 guidelines recommended for biomedical corpora in Erjavec et al. (2003) and was based on the format accepted for the NKJP corpus (Przepiórkowski and Bański, 2009). According to this scheme, every annotation is described in a separate file. In our corpus, each discharge document is represented by a catalog containing the following five files:

- *xxx.txt* – file with pure text of the anonymized original document,
- *xxx.xml* – file with text of the document (in the form as in *xxx.txt* file) divided into numbered sections (see below) which are in turn divided into paragraphs (every line break begins a new paragraph),
- *xxx_ segm.xml* – file with tokens (boundaries and types),

- *xxx_morph.xml* – file with morphological information (lemmas and morphological feature values),
- *xxx_sem.xml* – semantic labels and boundaries.

The *xxx.txt* file contains the text equivalent of the original MS Word document. The only practically important difference between these files is the representation of tables. They are converted into pure text by placing each table element in a separate paragraph. In the current version of the corpus, the straightforward format of linearized tables by enumerating their elements, was accepted.

The *xxx.xml* file contains text divided into sections and paragraphs. As the general structure of discharge records is regular, a document is automatically divided on the basis of introductory phrases (we use regular expressions allowing for some diversities within the phrase) into six parts i.e. *Introduction, Diagnosis, Examinations results, Treatment, Discharge record* and *Sign.* Some parts might be omitted in some documents. Every part has a *txt_x-div* label where *x* is a subsequent part number. The *div* xml tag has two attributes: *type* of the value *section*, and *title* which has one of the values listed above. For 460 documents this simple algorithm failed only in several cases where the beginning of the recommendation section was marked by a slightly different phrase at the end of a sentence. Each section consists of text paragraphs tagged with *txt_x.y-ab* labels where *x* is a section number and *y* is a number of a paragraph (counting from the file beginning). In Fig. 1 we present a fragment of the *xxx.xml* file[2] with the first three lines from the second section entitled *"Diagnosis"*.

```
<div xml:id="txt_2-div" type="section" title="Diagnose" >

  <ab xml:id="txt_2.9-ab">Rozpoznanie i wyniki badania klinicznego: </ab>
  <!--Diagnosis and clinical examination results:-->

  <ab xml:id="txt_2.10-ab">Cukrzyca typu 1 o wieloletnim
    przebiegu niewyrównana powikłana retinopatią prostą. </ab>
  <!--Diabetes of type 1 long-lasting uncontrolled, simple rethinopathy.-->

 <ab xml:id="txt_2.11-ab">Nadciśnienie tętnicze pierwotne. </ab>
 <!--Primary hypertension.-->

 </div>
```

Figure 1. Top level annotation file fragment

The next file – *xxx_segm.xml* – contains the three level segmentation information.

The first two levels mirror the division into sections and paragraphs made within the *xxx.xml* file. Every *segm_xx-div* label corresponds to a *txt_xx-div*

[2] The only difference is that the corpus itself does not contain English translations.

label and every *segm_xx.yy-p* label to a *txt_xx.yy-ab*. The third level of labels describes segmentation into tokens. Each token has a *segm_yy.zz-seg* tag assigned, where *yy* is the paragraph number and *zz* is the number of the token within the paragraph. The token description consists of a type name, a reference to the *xxx.xml* file and the appropriate character range within a numbered paragraph (there are over twenty token types, see Section 4). The token level annotation of the first three words from the second paragraph in Fig. 1 is given in Fig. 2. In this example, a string 'txt_2.10-ab,0,8' means that the segment 2.10 is related to the text fragment which belongs to the paragraph numbered 2.10 and consists of 8 characters starting from character 0. They relate to the word *cukrzyca* ('diabetes') which is given explicitly in the comment line below.

```
<p xlink:href="d2005-V_021.xml_2.10-ab" xml:id="segm_2.10-p">

  <seg xml:id="segm_10.1-seg">
    <xi:include href="d2005-V_021.xml" xpointer=
      "string-range(txt_2.10-ab,0,8)" type="first_capital_word"/>
  </seg>
  <!--Cukrzyca (Diabetes)-->

  <seg xml:id="segm_10.2-seg">
    <xi:include href="d2005-V_021.xml" xpointer=
      "string-range(txt_2.10-ab,9,4)" type="lowercase_word"/>
  </seg>
  <!--typu (type)-->

  <seg xml:id="segm_10.3-seg">
    <xi:include href="d2005-V_021.xml" xpointer=
      "string-range(txt_2.10-ab,14,1)" type="any_natural_number"/>
  </seg>
  <!--1 -->
  ...
</p>
```

Figure 2. Segmentation annotation fragment

The fourth file — *xxx_morph.xml* contains morphological information about every segment described in the appropriate *xxx_segm.xml* file. At the morphological level, segment tags are related to the segmentation level, i.e. every *morph_yy.zz-seg* points to the *segm_yy.zz-seg* label. Morphological annotation is described with the *fs* structure attached to every morphological segment. The type of the structure is *morph* and it has two *f* tags inside. The first one is named *orth*, and describes the orthographic form of the string. The second one is named *interps*, and contains the internal *fs* tag of the type *lex*. It, in turn, consists of two *f* tags representing base form (*base*) and one morphological tag (named *ctag*). In Fig. 3 we present the morphological annotation of the token *Cukrzyca* ('Diabetes') from the previous example. In this fragment there is a pointer to the tenth paragraph of the d2005-V_021 file which is in the second section of the document (label *morph_2.10-p* points to *segm_2.10-p*). The described segment is the first element of the tenth paragraph and at the morphological level gets the label *morph_10.1-seg*. This label is assigned the structure of type *morph* with two attributes. The *orth* attribute has the appropriate sequence of characters as its value, while the *interps* attribute is assigned to a structure of type *lex* with disambiguated morphological information. This structure consists of two attributes: *base* represents the basic form of a word and *ctag* represents complex morphological tags. The first element of a tag is a part of speech, then values of the appropriate morphological features are given (for nouns like *cukrzyca* 'diabetes' this means: number, case and gender). The morphological tags used and rules of their assignment are described in Section 5.

```
<p xlink:href="d2005-V_021.xml#segm_2.10-p"   xml:id="morph_2.10-p">
  <seg xlink:href="d2005-V_021_segm.xml#segm_10.1-seg"
      xml:id="morph_10.1-seg">
   <fs type="morph">

    <f name="orth">
      <string>Cukrzyca</string>
    </f>
    <!-- Cukrzyca [0,8] -->

    <f name="interps">
      <fs type="lex" xml:id="morph_10.1-lex">
        <f name="base"> <string>cukrzyca</string>
        </f>
        <f name="ctag"> <symbol value="subst:sg:nom:f"/>
        <!-- noun:singular:nominative:feminin -->
        </f>
      </fs>
    </f>

   </fs>
  </seg>
  ...
</p>
```

Figure 3. Morphological annotation fragment

The last file contains semantic information. In *xxx_sem.xml* file only small fragments of the original text get semantic labels. The structure of this file is similar to the previous one. Fragments of text for which a semantic label is assigned, are treated as segments and get *sem_yy.ss-seg* labels, where *yy* is the number of the paragraph and *ss* is the subsequent semantic segment number. Semantic labels are simple or complex. Simple labels are of the *f* type while complex labels are of type *fs*, which in turn contains several *f* tags. In Fig. 4

a simple label consisting of one *f* tag with the name equal *ACC_DISEASE* and value *hypertension_t* is assigned to the *morph_10.1-seg*, which in turn is assigned to the word *nadciśnienie (hypertension).* The complex semantic label will be presented in Section 6.

```
<seg xml:id=“sem_10.167-seg">
  <f name=“ACC_DISEASE">
    <symbol value="hypertension_t"/>
  </f>
<ptr target=“d2006_026_morph.xml#morph_10.1-seg"/>
<!–Nadciśnienie –>
</seg>
```

Figure 4. Semantic annotation example

The ways of obtaining semantic labels as well as types of semantic information represented within the corpus are described in the next section of the paper.

4. Segmentation into tokens

In the corpus, all higher annotation levels relate to the token level, so the decision on how to choose their boundaries is very important. The first option would be to take the results of one of the existing programs. Both of the applications we use to assign morphological and semantic tags have their own embedded tokenizers, but none of them seemed to be completely adequate for our task. The morphological analyzer we used – TaKIPI (Piasecki, 2007) does not differentiate nonword tokens and gives all of them the same tag. Moreover, it sometimes divides text in a non-uniform way (sequences like ‘15m’ are treated as one or as two tokens) and does not divide complex strings like ‘K-L’.[3] SProUT tokenizer distinguishes more token types.

There are several types representing strings of letters with capitalization variations: *all_capital_word, first_capital_word, lowercase_word, mixed_word_first_capital, mixed_word_first_lower, word_with_hyphen_first_capital, word_with_hyphen_first_lower.* There are four types representing strings containing letters as well as digits: *number_word_first_capital, number_word_first_lower, word_number_first_capital, word_number_first_lower*, and the *any_natural_number* type which is assigned to digit sequences. There are also several types representing single characters: *closing_bracket, colon, comma, dot, hyphen, opening_bracket, percentage_tok, question_mark, semicolon* and *slash.* Other character sequences get the *other_symbol* type assigned. This type

[3] There is another tokenization option in this program, which leads to the labeling of more complex structures like date and time descriptions, but the quality of this process on our data turned out to be rather low.

is used to tag both singular symbols not listed above, as well as longer sequences of mixed small letters, capital letters, digits and symbols.

To make our token division relatively flexible, we decided to divide the text into fragments without internal structure (with the exception of lowercase words with hyphens). As a result, token limits are more often equal to TaKIPI tokenization results than to SProUT results, but to maintain the extra information given in the names of SProUT token classes (e.g. capitalization), in our processing pipeline, we start from SProUT results and divide certain tokens into smaller parts. Ultimately, in our corpus we use token classes defined in SProUT with the exception of *number_word_first_capital* and *number_word_first_lower* classes. Tokens of these types frequently represent numbers and units (e.g. *10mm, 100Hz*) and were divided into two tokens of the appropriate classes. Another important change was the division of tokens of *other_symbol* type. Tokens of this type could be quite long and include different kinds of information and strings that are typically written after spaces. To make them more easily accessible, the sequences which include characters "=+/'.,():" or 'x' were divided on these symbols, e.g. the string "x3/4" was divided into four tokens. Introducing a special rule for 'x' was caused by its use as an abbreviation of 'times' in strings like '4x10x12'. Complex tokens (e.g. dates) are not annotated – such constructions, if needed, are recognized further on by IE rules and annotated on a semantic level.

The frequencies of occurrences of token of all types are given in Table 1. The conclusions, which can be drawn from it, are the diversity of texts at the token class level, and the uniformity of texts at the type of language used level. About 18% of tokens are numbers, 25% are different types of symbols (mostly punctuation marks) and 5% are sequences of letters which are probably not part of the general dictionary (classes *all_capital_word, mixed_word_first_capital, mixed_word_first_lower, word_number_first_capital, word_number_first_lower*). On the other hand, the list of different words used within these texts is rather short. All of the texts concern patients whose hospital visit was related to diabetes. All descriptions of the illness and treatment are similar and use limited vocabulary, although they were written by different physicians. For about 240,000 tokens representing words, there are only about 13,000 different forms even if all inflected forms are counted separately.

5. Morphological annotation

5.1. Rules of annotation

Morphological annotation was based on the results obtained by TaKIPI — the publicly available Polish POS tagger that cooperates with the general-purpose morphological analyzer *Morfeusz* SIAT (Woliński, 2006). To each word form, it assigns all the possible interpretations consisting of the base form, and full morphologic characteristic (described in Woliński, 2003). For example a noun form

Table 1. Token types and numbers of occurrences

	numbers of		
token class name	occurences	types	most frequent forms & occurences
all_ capital_ word	18416	303	W (2281), HM, R, N, P, KK, K (467)
any_ natural_ number	87246	-	–
apostrophe	14	1	’
back_ slash	7	1	
closing_ bracket	2663	3	) (2601), > (59),]
colon	12427	1	:
comma	28831	1	,
dot	47269	1	.
exclamation_ sign	49	1	-
first_ capital_ word	43269	2410	Insulina (1465), Actrapid (1257)
hyphen	4720	1	-
lowercase_ word	192368	9420	w(8513), mg (4386), z, j, i, l, do(2667)
mixed_ word_ first_ capital	514	19	NovoRapid (264), HBs (38),
mixed_ word_ first_ lower	1003	26	pH (518), mmHg (160), mU (153)
opening_ bracket	3355	3	((3261), < (79), [(4)
other_ symbol	2868	30	+(1535), HbA1C (457), =, HbA1c
percentage_ tok	4478	1	%
question_ mark	209	1	?
quotation	1	1	"
semicolon	455	1	;
slash	10353	1	/
word_ number_ first_ capital	1195	93	W1 (232), W2, HbA1, T4, V1, HCO3
word_ number_ first_ lower	1854	49	mm3 (1383), m2 (346), pCO2 (38)
word_ with_ hyphen_ first_ capital	163	61	Amox-Clavulan (30) Trimeth-Sulfa
word_ with_ hyphen_ first_ lower	402	144	p-ż (41), korowo-rdzeniowej (40)
all tokens	465004	13423	

dermatologa ‘dermatologist’ is assigned two possible tags: subst:sg:acc:m1, and subst:sg:gen:m1, both describe a singular (sg) noun (subst) of personal masculine gender (m1) in accusative (acc) or in genitive (gen). In case of participles additional morphologic categories are given, for example *niewypełniony* ‘not filled’ also has information on aspect (perfect for this form), and information that the base form — *wypełniony* ‘filled’ — is negated. In our data, this participle is assigned the following tag: ppas:sg:nom:m3:perf:neg. The set of potential morphological tags consists of more than 4,000 elements, while in our data only 450 different tags are represented, and in the general Polish IPIPAN corpus there are over 1,000 (Przepiórkowski, 2005). TaKIPI assigns the *ign* tag to all unrecognized tokens. As we want to have a more precise classification on *ign* tokens, we added several tags which allow us to differentiate numbers, foreign words and misspelled words. The list of all additional tags is given in Table 2.

The process of morphological analysis is described in detail in Marciniak and Mykowiecka (2011). The annotation is done in three steps:

- Documents are analyzed and disambiguated by TaKIPI combined with the *Guesser* module (Piasecki and Radziszewski, 2007) that suggests tags for word forms not recognized by the dictionary. Otherwise 27% of tokens representing words, acronyms and abbreviations would be assigned the unknown description. Its interpretations of domain related terms were relatively good, but typical medical abbreviations, medication names and

Table 2. Additional tags

tag	description	example
number	numerical values	*12, 1000*
acron	acronyms and abbreviations	*r* (year)
unit	units	*mm, mmol*
prefix	prefixes	*hipo* as a part of a word *hipoglikemia* 'hypoglycemia'
sufix	suffix	*ty* from *5-ty* '5th'
foreign	foreign words	*minoris, obesitas*
err_spell	misspeled tokens	*Oniadanie* instead of *śniadanie* 'breakfast'
err_conj	concatenations	*ciała103* 'body103'
err_disj_f	first part of disjoint word	*tyl*
err_disj_r	rest of disjoint word	*no* from *tylno* 'back'
tsym	patient codes	*d2004_023*

test names were usually incorrect. 20% of forms suggested by *Guesser* were analyzed correctly in a test set consisting of 8 manually corrected documents.

- The results of TaKIPI are postprocessed with a set of correction rules created on the basis of a list of all different token descriptions. The rules correct the annotations of domain related tokens like acronyms and abbreviations *BMI, Hg, BUN, kcal, MCV*, medication names *Polcard, Novonorm, Humulin*, and other domain terms like *chemioterapeutyk* 'chemotherapeutic' or *diuretyk* 'diuretic'.[4] Every rule concerns one form description, as it was hard to introduce any generalization (especially concerning lemmas).
- The last step consists in manual correction of the morphology analysis (in progress).

Below we indicate typical problems of TaKIPI tagging. Quite a lot of them can be eliminated by correction rules, but in some cases a context is indispensable.

- Capitalization of medication names lemmas. As they are proper names, we decided to preserve capitalization of their lemmas within the corpus. All lemmas guessed by TaKIPI start with a small letter, e.g., *Actrapid* was assigned the lemma *actrapid* that shall be corrected.
- Abbreviations and domain related acronyms are the most numerous group of incorrectly analyzed tokens. *Guesser* frequently gave a verb POS label for these short tokens, for example *LDL* (acronym of Low Density Lipoproteins) was assigned 'ldlić' base form and verb description (fin).

[4] Both were assigned personal masculine gender.

Sometimes common noun labels are assigned by TaKIPI, e.g. Apo (a part of a medication name) – 'apa' and the tag: subst:sg:voc:f.

- Abbreviations and acronyms are often interpreted as prepositions, e.g. token *w* is a domain abbreviation of *wieczór* 'evening' in insulin doses description while is interpreted as the preposition 'in'; *Na* can be a symbol of sodium or the preposition 'on'. These problems can be resolved only if the context of token usage is known.
- Problems with gender e.g. choosing feminine instead of masculine gender, results in assigning a 'subst:pl:gen:f' morphological tag instead of 'subst:sg:nom:masc3'. This error nearly always results in choosing the wrong lemma for a given word form (e.g. 'accolaty' in place of 'Accolate' or 'acarda' instead of 'Acard'). A very frequent error is to assign 'masc1' (personal) gender to a noun instead of 'masc3' (non-animate), but in these cases lemmas are correct e.g. *pernazin* was assigned 'masc1' gender instead of *Pernazin* with 'masc3' gender.
- TaKIPI does not resolve ambiguities typical for Polish in the recognition of substantives vs. gerunds e.g. *osiągnięcie* can be interpreted as the noun 'achievement' or a gerund 'reaching', similarly participles and adjectives are difficult to differentiate without a context.

It must be noted that the tagging errors described above do not influence the results of semantic annotation, as an IE system does not use tagger results and has access to all possible token descriptions given by the *Morfeusz* dictionary. Moreover, only token lemmas are addressed in rules, while domain terminology, that shall be taken into account in IE rules, is defined in the gazetteer (569 entries) – an IE system dictionary.

5.2. Morphological annotation results

In all analyzed reports there are about 13,000 different word forms (including symbols and abbreviations, excluding numbers). 2,660 of them occur ten and more times while about 1/3 of the forms occur only once. One discharge record contains about 1,100 tokens out of which about 680 are word forms. In discharge reports there are a few spelling errors, but a lot of terminology variants are used (Polish, Latin, abbreviations). Evaluation of the morphological annotation was done on 8 randomly chosen documents from the year 2001 as we already pointed out that the vocabulary of documents does not vary much so the choice of one year cannot influence the results of verification significantly, more extensive evaluation is, however, planned. The results are presented in Table 3. This set contains 8,919 tokens. 3,938 tokens are numbers and other nonword tokens and 4,981 are word forms of 1381 types. The most common part of speech types are nouns (1930 – 39% occurences of 591 types) and adjectives (680, 13.7%). Verbs are much less common – only 45 (0.9%) finite verb forms and 6 (0.12 %) infinitives were encountered. A great number of tokens represent acronyms (558

– 11%) and units – 576 (11.6%). Manual verification of base forms and part of speech assignment showed that 475 forms were assigned the wrong tags. Two annotators gave different suggestions on 250, i.e. 0.06% of the wordform tags. There were 49 orthographic errors within the text including 6 disjoints and 2 incorrect conjuncts. Only 17 of misspelled words were words which are out of both general and biomedical vocabularies, others are forms which are potentially proper but not adequate in a given context. The morphological results analysis showed that efficiency of the tagger on this data is much lower than reported on general ones (91.3% reported by Karwańska and Przepiórkowski, 2009), but after applying our domain based changes we reached 89.4% of tag correctness.

Table 3. Morphological anotation evaluation of 4,972 tokens

type of errors	nb.	most frequent cases wrong-> correct (nb)
all errors	475	
good POS		
only case change	185	acc->nom (139), nom->acc (25)
only gender	32	f->m3 (7), m3->n (7), m1->m3 (6)
other features	24	zasady:subst:sg:gen:f -> subst:pl:acc:f (5)
wrong POS		
only POS	100	prep:acc:nwok -> acron (22), conj -> acron (12) acron -> prep:loc:nwok (18), prep:gen -> unit (10)
morph tag	79	
spell errors	34	xxx-> err_spell(28), xxx-> err_disj(6)

Morphological annotation of medical data with a complicated tagset turned out to be difficult. A great number of specialized terms and acronyms were very frequently assigned wrong, sometimes surprising labels, e.g 'K' (Potassium) was tagged as a very rarely occurring Polish preposition, or as unknown, e.g. an abbreviation 'kk' which turned out to mean either 'limbs' or 'bones'. The analysis of medical texts also requires defining methods of dealing with Latin or English words (which are sometimes inflected just like Polish) and with spelling errors. In our data most errors resulted in correct forms which do not agree with context (such forms are frequently the result of omitting Polish diacritics, which is a type of error very easy to make and not so easy to notice). Such phrases will not be recognized by any Polish grammar assuming, for example, noun-adjective agreement in case, number and gender. We proposed a set of special morphological tags to represent these phenomena.

Even after correcting a larger part of the corpus, then trying to train a specialized tagger, which could also take into account data about frequent errors, because many forms appear in a limited context (enumerations, or pairs of test names and values, short phrases), the efficiency of a tagger would probably still be lower then it is for a general tagger used on general texts.

6. Semantic annotation

6.1. Rules of labels assignment

To realize semantic annotation of the corpus we decided to reuse a rule-based information extraction approach designed to select important data from hospital records of diabetic patients (Marciniak and Mykowiecka, 2007; Mykowiecka et al., 2009). Originally, the system was developed to convert information given in free text into a database, which will serve for statistical or epidemiological purposes. The information extraction grammar consists of 201 rules.

The grammar contains several context rules. They extract information only if it occurs close to an appropriate keyword, see example (1).[5]

(1) *Cukrzyca typu 1 o wieloletnim przebiegu niekontrolowana*
Diabetes type 1 long-lasting uncontrolled

All words in (1): 'type 1', 'long-lasting' and 'uncontrolled' are diabetes features and are represented by the following semantic labels: D(iabetes)_TYPE: *first*, FROM_IN_W(ords): *long_ standing_ rec* and D(iabetes)_CONTROLL: *uncontrolled_ t*. These features should be extracted as important only in the context of the word 'diabetes'. In other cases, they can mean something else e.g., *schizofrenia o wieloletnim przebiegu* — 'long-lasting schizophrenia' or *niekontrolowana reakcja* — 'uncontrolled reaction'. In the analyzed texts concerning diabetic patients, all phrases *typu 1* 'type 1' refer to diabetes, but there are other diseases with similar classification that can appear in medical documents. Thus, the whole phrase from example (1) is recognized by one rule of the grammar. It combines different phrases co-occurring with the keyword *cukrzyca* 'diabetes'.

The context rules in the grammar allow us to recognize diabetic features together with other important data like complications or information about patient's weight or age, which are also correctly interpreted without looking into the context. In example (2), the phrase *with obesity* is inserted inside a phrase describing diabetes itself. It is interpreted within this context in the same way as in any other.

(2) *Cukrzyca typu 2 skojarzona z otyłością, o chwiejnym przebiegu*
Diabetes type 2 with obesity, unstable course

Some information, such as different types of complications, like neuropathy (3), is given by coordinated phrases. In this case, the IE system recognizes the whole information by one rule. The word *obwodowa* 'peripheral' is recognized as a type of neuropathy only in the context of the word *neuropatia* 'neuropathy'. This was forced by the fact that the word *obwodowa* 'peripheral' appears also in other contexts, e.g. *krew obwodowa* 'peripheral blood'.

[5] Lack of punctuation is typical for the texts.

(3) *z neuropatią autonomiczną i obwodową*
with neuropathy autonomic and peripheral

Any complex reason for hospitalization that consists of several elements with key-phrases like: *przyjęty do szpitala z powodu* 'hospitalized because of', is also recognized in the full grammar (by one rule) as one list structure.

The IE system results include boundaries of a whole phrase to which a semantic label (or a list of semantic labels) is attached. However, in the case of complex information recognized by one rule and represented by a list, like in (1), this is insufficient for the task of corpus annotation. To indicate the boundaries of semantic labels attached to phrases more precisely, we prepared a simplified version of the grammar on the basis of the full grammar. It consists of 129 rules and recognizes important phrases without the context.

Sometimes a rule is divided into several rules, recognizing smaller pieces of information. Example (4) shows the case where the original full grammar recognized the whole phrase by one rule but in the simplified grammar there are two rules that recognize two pieces of information indicated by brackets, and represented by two complex attributes H_FROM and H_TO in structure (5). These two rules also recognize the phrases *od dnia ...* 'from the date ...' and *do dnia ...* 'to the date ...' which do not refer to hospitalization.

(4) *Przebywała w Klinice [od dnia 29.07.2006] [do dnia 09.08.2006]*
Hospitalized from 29.07.2006 to 09.08.2006

(5)

$$\begin{bmatrix} \text{hospit\_str} & \\ \text{H\_FROM} & \begin{bmatrix} \text{Y\_DAT} & 2006 \\ \text{M\_DAT} & 07 \\ \text{D\_DAT} & 29 \end{bmatrix} \\ \multicolumn{2}{l}{< \textit{od dnia 29.07.2006} >} \\ \multicolumn{2}{l}{(\textit{from 29.07.2006})} \\ & \\ \text{H\_TO} & \begin{bmatrix} \text{Y\_DAT} & 2006 \\ \text{M\_DAT} & 08 \\ \text{D\_DAT} & 09 \end{bmatrix} \\ \multicolumn{2}{l}{< \textit{do dnia 09.08.2006} >} \\ \multicolumn{2}{l}{(\textit{to 09.08.2006})} \end{bmatrix}$$

The simplified grammar recognizes all three features of diabetes in (1) separately, as indicated in (6). The structure assigned to the phrase with internal information is given in (7).

(6) *Cukrzyca [typu 1] o [wieloletnim] przebiegu [niekontrolowana]*
Diabetes type 1 long-lasting uncontrolled

(7)
$$\begin{bmatrix} \text{feature\_l\_str} \\ \quad [\text{D\_TYPE}\ \ first] \\ \quad < \textit{typu}\ \ \textit{1} > (\textit{type}\ \ \textit{1}) \\ \quad [\text{FROM\_IN\_W}\ \ \textit{long\_lasting\_rec}] \\ \quad < \textit{wieloletnim} > (\textit{long} - \textit{lasting}) \\ \quad [\text{D\_CONTROLL}\ \ \textit{uncontrolled\_t}] \\ \quad < \textit{niekontrolowana} > (\textit{uncontrolled}) \end{bmatrix}$$

All other appearances of the words: *typ 1* ‘type 1’, *wieloletni* ‘long-lasting’, *niekontrolowany* ‘uncontrolled’ are detected by the simplified IE grammar as diabetic features also. For example, in the phrase *pacjentka z wieloletnią chorobą niedokrwienną serca* ’patient with long-lasting ischemic heart disease’ the word *wieloletnią* is recognized as information about diabetes.

Comparing full and simplified IE grammars in terms of the IE task, it can be said that the recall of both systems is very close (or even the same) but the precision of simplified grammar is much lower than the precision of the full one. The simplified grammar recognized too much information. For example, in test data consisting of 46 patients’ records, the word *wieloletni* ‘long-lasting’ occurred 10 times while four of them did not refer to diabetes, but the simplified grammar recognized all of them as diabetes features.

In the next step, the results of both extraction grammars are cleaned up. The non-informative pieces of structures are removed from the results. For example if there is information that the diabetes was recognized in December 1999, which is represented by a date structure having attributes: Y_DATE, M_DATE and D_DATE, the last attribute (representing an unknown day) has *string* value that can be removed. Moreover, if an attribute has assigned a label indicating that its value is unified with the value of another structure, the value is assigned to it. So after cleaning, the output structures contain only values (not labels indicating unified values).

Finally, the results of both IE grammars are compared. The automatic annotation contains combined information from both annotations. Boundaries of the whole phrase recognized by the original grammar as well as the narrower limits of phrases representing a particular feature are preserved in the annotation. However, the annotation from the simplified grammar is valid only in the case when it is enclosed within a phrase recognized by the full grammar. This approach is very similar to an approach based on a cascade of grammars. We chose this approach because we reused some tools prepared for database construction. The simplified grammar was created on the basis of the full grammar by removing context rules and dividing some rules into rules recognizing only fragments of information.

The main goal of the IE system was to find out whether a particular piece of information is present in an analyzed text. The first experiment of corpus creation showed the problem of inconsistently recognized boundaries of phrases. For example, rules recognizing patients’ height and weight sometimes did it with and sometimes without units. So we decided to inspect both grammars to remove all such inconsistencies. We decided that rules should recognize the smallest phrase that includes the desired information. As units used to describe height and weight are constant for all documents, so we decided not to recognize them by rules. This decision was especially important in process of manual verification of the annotation. The instruction for annotators contains a very precise definition of a phrase to which the structure might be assigned. Structures are assigned to continuous phrases, i.e. to all tokens between the first and last tokens of the phrase. The precise definition of phrase boundaries consists in determining the sets of words that may start and end the phrase, and the type of information included in the phrase:

- parsing the text with the full extraction grammar,
- parsing the entire text using the simplified extraction grammar rules which recognize small pieces of information, usually one attribute, only numerical values (e.g. ranges, dates) are still recognized together,
- removing uninformative structures from the output of both extraction grammars,
- comparing and combining the results – only structures that are represented in both results are represented in the final corpus data,
- final postprocessing of data consisting in removing unnecessarily deep hierarchies of structures, recognition of sections,
- combining the semantic information with morphological information to create a set of corpus XML files.

6.2. Semantic labels

In the corpus about 50 simple attributes and 14 complex structures are labeled. They represent the following information:

- Identification of a patient's visit in hospital represented by *id_ str* structure that contains information on ID number and information if it is a main document or a continuation of a document CONT; the date of the document DOC_DAT; dates when the hospitalization took place (*hospit_ str*, see example (5)), and hospitalization reasons resulting from diabetes, represented by the *reason_ l_ str* list of attributes.
- Patient information: *id_ pat_ str* structure containing patient's identifier and sex; simple attributes representing age, weight, height, BMI as numbers and information on weight given in words (W_IN_WORDS).
- Data about diabetes (in some cases grouped in a *feature_ l_ str* structure, see example (7)), e.g.: type (D_TYPE); if the illness is balanced (D_CONTROL); when diabetes was diagnosed in three different formats (RELATIVE_DATA: number and units, 'three month ago'; W_IN_WORDS 'newly recognized' or 'long-lasting'; YEAR_OF_LIFE); results of tests e.g.,:

HbA1c, acetone detection, up to three levels of LDL, levels of microalbuminury and creatinine.

- Complications: COMP indicating a type (e.g., retinopathy); lack of all complications or a particular one (N_COMP); other illnesses including autoimmunology (AUTOIMM_DISEASE) and accompanying illnesses (ACC_DISEASE), which may be correlated with diabetes.
- Diabetes treatment described by *insulin_treat_str* that contains insulin type and its doses (*dose_str* — DOSE_MIN and DOSE_MAX); description of continuous insulin infusion therapy (structures *ins_inf_treat*); description of oral medications (ORAL_TREAT); information that insulin therapy was started (I_THERAPY_BEG). The applied therapy is sometimes given as a list of information that is represented by a *cure_l_str* list of attributes.
- Diet description represented by *diet_str* that contains information on type of diet (DIET_TYPE), and structures describing how many calories are recommended (CAL_MIN, CAL_MAX) and a similar structure representing numbers of recommended meals.
- Information on therapy given in text form, e.g.: patient's education (EDUCATION), observing of diet (DIET_OBSERVE), modification of treatment or diet (THERAPY_MODIFF), self monitoring (SELF_MONITORING).

The semantic annotation of phrase (1) is given in Fig. 5. It contains information that the whole phrase was recognized as a phrase describing diabetes features — the *sem_7.5-seg* is labeled with a structure of the *feature_l_str* type. Within the phrase three pieces of information were recognized. They are represented by *f* attributes. The first is of type D_TYPE and has the *first* value, the second is of FROM_IN_W type and has the *long_standing_rec* value and the third is D_CONTROLL with the *uncontrolled_t* value. All *f* attributes have pointers to the particular words which they describe (pointers to the morphological annotation level labels are used).

6.3. Semantic annotation results

The semantic annotation method described in the paper is roughly as good as the IE system used for this purpose. A detailed evaluation of the IE system is provided in Mykowiecka et al. (2009). The IE test set consisted of 100 discharge records. 55 attributes (from the total number of 68) occurred in these documents. 16 attributes have an F-score above 99% and three attributes have an F-score less than 95%. The most frequent attribute was COMPlication which occurred 369 times and had an F-score of 0.98.

The results of manual correction of semantic annotation of a randomly selected 10% of documents (46 documents) are given in Table 5. The correction was done by two annotators, and one coherent version was elaborated in the inter-annotators' negotiations. The table includes the number of elements

```
<seg xml:id="sem_7.5-seg">
  <fs type="feature_l_str" >
    <f name="D_TYPE">
      <symbol value="first">
    </f>
    <ptr target="d2005-V_21_morph.xml#morph_7.2-seg"/>
    <ptr target="d2005-V_21_morph.xml#morph_7.3-seg"/>
    <!-- typu 1 (type 1) -->
    <f name="FROM_IN_W">
      <symbol value="long_standing_rec">
    </f>
    <ptr target="d2005_V_21_morph.xml#morph_7.5-seg" />
    <!-- wieloletnim (long-lasting) -->
    <f name="D_CONTROLL">
      <symbol value ="uncontrolled_t">
    </f>
    <ptr target="d2005-V_21_morph.xml#morph_7.7-seg" />
    <!-- niewyrównana (uncontrolled) -->
  </fs>
</seg>
<ptr target="d2005-V_021_morph.xml#morph_7.1-seg"/>
...
<ptr target="d2005-V_021_morph.xml#morph_7.7-seg"/>
<!-- Cukrzyca typu 1 o wieloletnim przebiegu niewyrównana -->
<!-- Diabetes type 1 long lasting uncontrolled -->
```

Figure 5. Example of semantic annotation

which were automatically recognized, the number of elements after correction, F-measure, precision and recall for particular attributes and the number of different phrases representing the attribute in the test data. The overall F-measure counted for structure and attribute recognition is equal to 0.965. The worst results were achieved for attributes, whose value can be expressed by many different phrases, and their number of occurrences was low in comparison to the diversity of possible phrases: I_THERAPY_BEG(ining), THERAPY_MODIFICATION. The problem of a low F-measure of the N_COMP attribute stems from the fact that a phrase like *Bez późnych zmian cukrzycowych* 'There were no long-lasting diabetes complications' is differently interpreted if it is included within an eye test (usually as no retinopathy). If, however, it is included in the final part of the document, it means that no diabetic complications were diagnosed.

Table 4. Semantic label occurrences in the test set

structure/attribute	numb of occurrences		F-	prec.	recall	numb of diff.
	org.	correct	measure			phrases
administrative information						
DOC_BEG	46	46	1	1	1	1
DOC_DAT	38	37	0.99	0.97	1	1
id_str	45	46	0.99	1	0.98	–
ID	45	46	0.99	1	0.98	1
CONT	45	45	1	1	1	2
hospit_str	43	46	0.97	1	0.93	–
H_FROM	43	46	0.97	1	0.93	2
H_TO	43	46	0.97	1	0.93	2
EPIKRYZA_BEG	46	46	1	1	1	1
recommendation_str	44	44	1	1	1	–
RECOMMENDATION_BEG	44	44	1	1	1	1
basic patient data						
id_pat_str	45	46	0.99	1	1	–
id_pat_sex	45	46	0.99	1	1	3
ID_PAT	45	46	0.99	1	1	1
ID_P_SEX	45	46	0.99	1	1	2
ID_AGE	45	46	0.99	1	1	6
W_IN_WORDS	5	6	0.91	1	0.83	4
WEIGHT	39	40	0.99	1	0.97	6
BMI	33	33	1	1	1	5
HEIGHT	39	40	0.99	1	0.97	3
basic diabetes data						
D_CONTROLL	27	30	0.95	1	0.90	41
FROM_IN_W	1	0	–	–	–	6
HBA1C	54	59	0.96	1	0.92	8
ACET_D	42	42	1	1	1	4
creatinin_str	41	43	0.98	1	0.95	7
microalbuminury_str	12	13	0.96	1	0.92	5
lipid_str	27	31	0.93	1	0.87	–
LDL1	27	31	0.93	1	0.87	3
feature_l_str	91	91	1	1	1	–
COMP	5	5	1	1	1	53
D_CONTROLL	34	34	1	1	1	41
D_TREAT	24	24	1	1	1	8
D_TYPE	70	70	1	1	1	3
FROM_IN_W	10	10	1	1	1	6
RELATIVE_DATA	18	19	0.97	1	0.95	9
W_IN_WORDS	10	10	1	1	1	4
reason_l_str	27	30	0.95	1	0.90	–
D_CONTROLL	37	40	0.95	1	0.95	41
KETO_D	2	2	1	1	1	2
KWAS_D	1	1	1	1	1	2
RELATIVE_DATA	1	1	1	1	1	9
SELF_MONITORING	1	1	1	1	1	3
complication and acc diseases						
ACC_DISEASE	48	48	1	1	1	3
COMP	132	134	0.97	0.98	0.96	53
N_COMP	15	27	0.71	1	0.56	11

Table 5. cont. Semantic label occurrences in the test set

structure/attribute	numb of occurrences		F-	prec.	recall	numb of diff.
	org.	correct	measure			phrases
therapy						
insulin_treat_str	444	446	0.99	0.99	0.99	–
I_TYPE	436	439	0.99	1	0.99	34
dose_str	440	441	0.99	0.99	0.99	8
corr_str	1	2	0.67	1	0.5	–
DOSE_MODIFF	1	2	0.67	1	0.5	7
THERAPY_MODIFF	1	2	0.67	1	0.5	23
diet_str	44	47	0.97	1	0.94	–
DIET_TYPE	44	47	0.97	1	0.94	4
cal_str	44	47	0.97	1	0.94	–
CAL_MIN	44	47	0.97	1	0.94	4
meals_str	41	45	0.95	1	0.91	–
MEALS_MIN	41	45	0.95	1	0.91	5
ORAL_TREAT	63	63	1	1	1	18
I_THERAPY_BEG	1	4	0.40	1	0.25	4
THERAPY_MODIFF	19	24	0.88	1	0.79	23
DOSE_MODIFF	8	9	0.94	1	0.89	7
DIET_CORRECTION	2	2	1	1	1	2
SELF_MONITORING	0	2	–	–	–	3
EDUCATION	25	27	0.96	1	0.93	20

7. Conclusions

The obtained results show that the IE rule based systems cope well with the task of medical data extraction and can be used in the first, automatic step of corpora semantic annotation. It is a well-known fact that these systems can relatively easily extract rare data, like, e.g., autoimmunology disease (35 occurrences in all our data and none in the evaluated 46 documents). But, rule based IE systems also have limitations — they only extract data specified by the rules. In medicine, there are new medications and tests introduced every year. For example, in the data considered in the paper, there were three new names of insulin medication introduced in the year 2006. As it is well - known, statistical IE systems cope better with new data that occur within patterns. In the next stage of our research, we plan to develop a Conditional Random Field (Sutton and McCallum, 2007) model for the task of semantic labeling of diabetic patients' discharge records.

Morphological annotation of medical data with a complicated tagset turned out to be difficult, but such a detailed description is not indispensable for basic information extraction, so data which are annotated with a more limited tagset, or data, which are not completely correctly annotated, can be also useful for the purpose. As our work showed, for efficient IE and semantic labeling an inflected list of biomedical terminology is much more desirable, but unfortunately does not yet exist for Polish. However, for a limited domain, the list of actually occurring terms is not so long, so it can be prepared on the basis of real life documents manually or using statistical methods.

Although an exact evaluation of the time needed for manual annotation of the data was not already performed, the subjective opinions of all annotators for both morphological and semantic annotation, supported our claim that manual annotation made from scratch would be more time consuming and error prone (preforming token segmentation manually would be rather ridiculous). Of course, one should also account for the time used for IE system creation, but in the case of manual annotation, a lot of time has to be devoted to prepare the rules of annotations. When they are ready, they can be be converted into a rule based system (to some extent at least). Although our work concerns very specific types of texts, the method itself is general – it can be used for any IE system for any language as long as domain and language dependent data exists.

Acknowledgements: The work was partially financed by the European Union Innovative Economy Programme project POIG.01.01.02-14-013/09.

References

Aramaki, E., Miura, Y., Tonoike, M., Ohkuma, T., Mahuichi, H. and Ohe, K. (2009) TEXT2TABLE: Medical Text Summarization System based on Named Entity Recognition and Modality Identification. In: *Proceedings of the Workshop on BioNLP, Boulder, Colorado.* Association for Computational Linguistics, 185–192.

Drożdżyński, W., Krieger, H.U., Piskorski, J., Schäfer, U. and Xu, F. (2004) Shallow Processing with Unification and Typed Feature Structures – Foundations and Applications. *German AI Journal KI-Zeitschrift*, 01/04.

Erjavec, T., Tateisi, Y., Dong Kim, J., Ohta, T. and Tsujii, J. (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.* Association for Computational Linguistics, 97–104.

Franzén, K., Eriksson, G., Olsson, F., Asker, L., Lidén, P.and Cöster, J. (2002) Protein names and how to find them. *International Journal of Medical Informatics*, **67**, 49–61.

Karwańska, D. and Przepiórkowski, A. (2009) On the Evaluation of Two Polish Taggers. In: *The Proceedings of Practical Applications in Language and Computers PALC 2009.* Peter Lang.

Kim, J.D., Ohta, T., Tateisi, Y. and Tsujii, J. (2003) GENIA corpus – a semantically annotated corpus for bio-textmining. *Bioinformatics*, **19**(1), 180–182.

Kim, J.D., Ohtai, T. and Tsujii, J. (2008) Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, **9** (10), doi 10.1186/1471-2105-9-10.

Kokkinakis, D. (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).* ELRA, European Language Resources Association, 1200–1205.

Mandel, M.A. (2006) Integrated Annotation of Biomedical Text: Creating the PennBioIE corpus. In: *Proceedings of the Workshop on Text Mining, Ontologies and Natural Language Processing in Biomedicine: 20-21 March 2006.* http://www-tsujii.is.s.u-tokyo.ac.jp/jw-tmnlpo/MarkMandel.pdf

Marciniak, M. and Mykowiecka, A. (2007) Automatic processing of diabetic patients' hospital documentation. In: *Proceedings of Balto-Slavonic Natural Language Processing ACL 2007 Workshop.* Association for Computational Linguistics, 35–42.

Marciniak, M. and Mykowiecka, A. (2011) Towards Morphologically Annotated Corpus of Hospital Discharge Reports in Polish. *Proc. of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011.* Association for Computational Linguistics, 92–100.

Mykowiecka, A., Marciniak, M. and Kupść, A. (2009) Rule-based information extraction from patients' clinical data. *Journal of Biomedical Informatics*, **42**, 923–936.

Pakhomova, S.V., Codenb, A. and Chutea, Ch.G. (2006) Developing a corpus of clinical notes manually annotated for part-of-speech. *International Journal of Medical Informatics*, **75**, 418–429.

Pestian, J.P., Brew, Ch., Matykiewicz, P., Hovermale, D.J., Johnson, N., Cohen, K.B. and Duch, W. (2007) A shared task involving multilabel classification of clinical free text. In: *BioNLP '07: Proceedings of the Workshop on BioNLP 2007*, Association for Computational Linguistics, 97–104.

Piasecki, M. (2007) Polish Tagger TaKIPI: Rule Based Construction and Optimisation. *Task Quarterly*, **11**(1–2), 151–167.

Piasecki, M., Godlewski, G. and Pejcz, J. (2006) Corpus of medical texts and tools. In: *Proceedings of Medical Informatics and Technologies.* Silesian University of Technology, 281–286.

Piasecki, M. and Radziszewski, A. (2007) Polish Morphological Guesser Based on a Statistical A Tergo Index. In: *Proceedings of the International Multiconference on Computer Science and Information Technology — 2nd International Symposium Advances in Artificial Intelligence and Applications (AAIA'07).* Polskie Towarzystwo Informatyczne, 247–256, http://www.proceedings2007.imcsit.org/pliks/150.pdf.

Przepiórkowski, A. (2005) The IPI PAN Corpus in Numbers. In: Zygmunt Vetulani, ed., *Proceedings of the 2nd Language & Technology Conference*, Poznań, Poland. Wydawnictwo Poznańskie, 27–31.

Przepiórkowski, A. and Bański, P. (2009) XML Text Interchange Format in the National Corpus of Polish. In: *The Proceedings of Practical Applications in Language and Computers PALC 2009.* Peter Lang, 245–250, http://nlp.ipipan.waw.pl/ adamp/Papers/2009-palc-xml/paper.pdf.

Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Jörvinen, J. and Salakoski, T. (2007) BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8.

Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I. and Setzer, A. (2009) Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*, **42**(5), 950–966.

Røst, T.B., Huseth, O., Nytrø, Ø. and Grimsmo, A. (2008) Lessons From Developing an Annotated Corpus of Patient Histories. *Journal of Computing Science and Engineering*, **2**(2), 162–179.

Sutton, Ch. and McCallum, A. (2007) An Introduction to Conditional Random Fields for Relational Learning. In: L. Getoor, and B. Taskar, eds., *Introduction to Statistical Relational Learning*, chapter 4. MIT Press.

Tsalidis, Ch., Orphanos, G., Mantzari, E., Pantazara, M., Diolis, Ch. and Vagelatos, A. (2007) Developing a Greek biomedical corpus towards text mining. In: *Proceedings of the Corpus Linguistics Conference (CL2007)*. University of Birmingham.

Vincze, V., Szarvas, G., Farkas, R., Móra, G. and Csirik, J. (2008) The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. *BMC Bioinformatics*, **9** (Supl 11), 38–45.

Wermter, J. and Hahn, U. (2004) An Annotated German-Language Medical Text Corpus as Language Resorces. In: *Proceedings of the Fourth International Language Resources and Evaluation (LREC'04)*. ELRA, European Language Resources Association, 473–476.

Woliński, M. (2003) System znaczników morfosyntaktycznych w korpusie IPI PAN. *Polonica*, **XXII-XXIII**, 39–55.

Woliński, M. (2006) Morfeusz — a Practical Tool for the Morphological Analysis of Polish. In: M. Kłopotek, S. Wierzchoń, and K. Trojanowski, eds., *Intelligent Information Processing and Web Mining, IIS:IIPWM'06 Proceedings*. Springer, 503–512.