corpus, medical text, annotation
morpho-syntactic tagging,
natural language processing

Grzegorz GODLEWSKI[*], Maciej PIASECKI*, Jerzy PEJCZ[**]

CORPUS OF MEDICAL TEXTS AND TOOLS[1]

There is only one large corpus of Polish annotated with morpho-syntactic information, namely The IPI PAN Corpus (IPIC). This situation is a big obstacle in creation of tools for natural language processing dedicated to the domain of medical texts. However, the real life medical texts exhibit features making them very distinct from the most of the texts stored in IPIC. In the paper, the attempts to create a larger corpus of medical texts called KorMedIIS are presented. The subsequent phases of texts gathering, their lexical analysis and morpho-syntactic annotation are discussed. The works on preparation of extended version of tools for processing natural language, namely a morphological analyser and a tagger, are presented. Different linguistic problems encountered are discussed. A special tool for the edition of the corpus, searching and correcting it is also presented.

1. INTRODUCTION

Processing of natural language on practical scale needs several tools like morfo-syntactic analysers or parsers. The tools are commonly constructed by application of Machine Learning (ML) methods on the basis of *language resources* where the most important is the corpus of texts (henceforth, *corpus*). For the needs of ML, the corpus must be *annotated*, i.e. the linguistic information must be attached to words or sequences of words. The simplest form of annotation is the annotation expressing the morpho-syntactic description of words: *grammatical class* [2] (a more detailed division than Parts of Speech, there are 32 classes defined for Polish) and *values* of the *grammatical categories*, e.g. case, number, gender etc. The annotation is often expressed in XML based format. The example of such format is given in Fig. 2. All possible morpho-syntactic analyses are assigned to a word. However, the number of different analyses can be as a large as 20 for Polish. In many processing tasks we need to know the one appropriate analysis for a word. Thus we need to *disambiguate* the corpus by marking the proper choice in the annotation of each word. The corpus can be annotated also by other types of information, e.g. syntactic information concerning the syntactic structures of phrases and sentences. However, even the simpler morpho-syntactically

[*] Institute of Applied Informatics, Wrocław University of Technology,

ul. Wybrzeże Wyspiańskiego 27, Wrocław, Poland.

[**] Szpital im. Świętej Jadwigi Śląskiej, Trzebnica

[1] This work was financed by the Ministry of Education and Science projects No 3 T11E 005 28.

annotated and disambiguated corpus can be a basis for many solutions including construction of a *language model* used as a support for OCR.

There is only one morpho-syntactically disambiguated corpus of Polish, namely The IPI PAN Corpus [2] (henceforth, IPIC). IPIC includes over 260 millions of *tokens*. Simplifying, a token is a word or a symbol occurring in the corpus. A part of IPIC including 885 669 has been disambiguated manually, the rest was disambiguated automatically by a application of a *morpho-syntactic tagger* (a computer program), called TaKIPI (the Polish acronym for *the tagger of IPIC*) [1]. IPIC includes a wide variety of texts, but is intended to be a representation of a `standard Polish'. The medical texts we deal with in the project are documents created by doctors during the medical treatment in hospital. They are created during work, in order to note down information, and are very different than carefully prepared publications dominating in IPIC. The analysis of the medical texts is presented in Sec. 2. Thus, language tools trained exclusively on IPIC can exhibit a decreased accuracy when applied to typical medical texts.

The goal of this work are: to develop a procedure of creation of a representative corpus in a selected domain of medical texts and next to construct tools dedicated to this class of texts.

2. LINGUISTIC CHARACTERISTICS OF TEXTS

The texts collected from the database of the hospital belong to several categories: *initial diagnosis*, *medical history*, *objective examination*, *accessory examinations*, *final diagnosis*, *treatment*, and *epicrisis*. The epicrisis is a short description of a patient stay in a hospital. It is written when patient is discharged from the hospital. A typical epicrisis includes larger passages of text, consists of several sentences and/or shorter phrases, reports some details of the patient stay and treatment, and often copies after the other documents. The epicrisises contain typical elements from other types of the medical documents, including terminology and structure of sentences. We have chosen the epicrisises as the basis of the medical corpus, called KorMedIIS (the Polish acronym for *the medical corpus of the Institute of Applied Informatics*) because of their universality. KorMedIIS includes presently 15 961 epicrisises (2 millions of tokens). KorMedIIS was automatically morpho-syntactically annotated and disambiguated, see Sec. 4. 2 127 of the epicrisises were manually disambiguated by a linguist. We plan to extend the corpus with other types of medical documents.

As we expected the collected texts contains a lot of proper words which were not recognised (and annotated) by the standard version of the morphological analyser used in IPIC called *Morfeusz* [4]. The process of extension of Morfeusz is described in Sec. 3.

The collected texts contain quite numerous spelling errors. Most of them were caused by misspelled single characters (especially Polish) but some rather strange ones have a permanent character and were probably performed during the introduction of the data to the database, e.g. the Polish word “mężczyzna” (*man*) was constantly misspelled to “męższczyzna” for about 161 times. We created a statistic dictionary of misspelled forms including 419 words down to 16 occurrences.

Other problem was that authors of the documents use different ways of creating abbreviations. They seem to be often created *ad hoc*, e.g. expression “ciąża druga” (second pregnancy), is written: CII, C II, cII, c II, c2, c2, c I I, etc. A dictionary of abbreviations was created (666 words / 21)

Surprisingly, the syntactic constructions in KorMedIIS are mostly standard, i.e. there are mostly proper sentences or at least phrases. Only some forms of enumerations in points or `bullets' can be met. There is also a tendency to use long coordinated constructions (joined by conjunctions), e.g. expressing different symptoms, with the same grammar case imposed for all parts by a verb.

3. LEXICAL RESOURCES

Morfeusz recognises surprisingly many words specific to medical terminology, but not all.

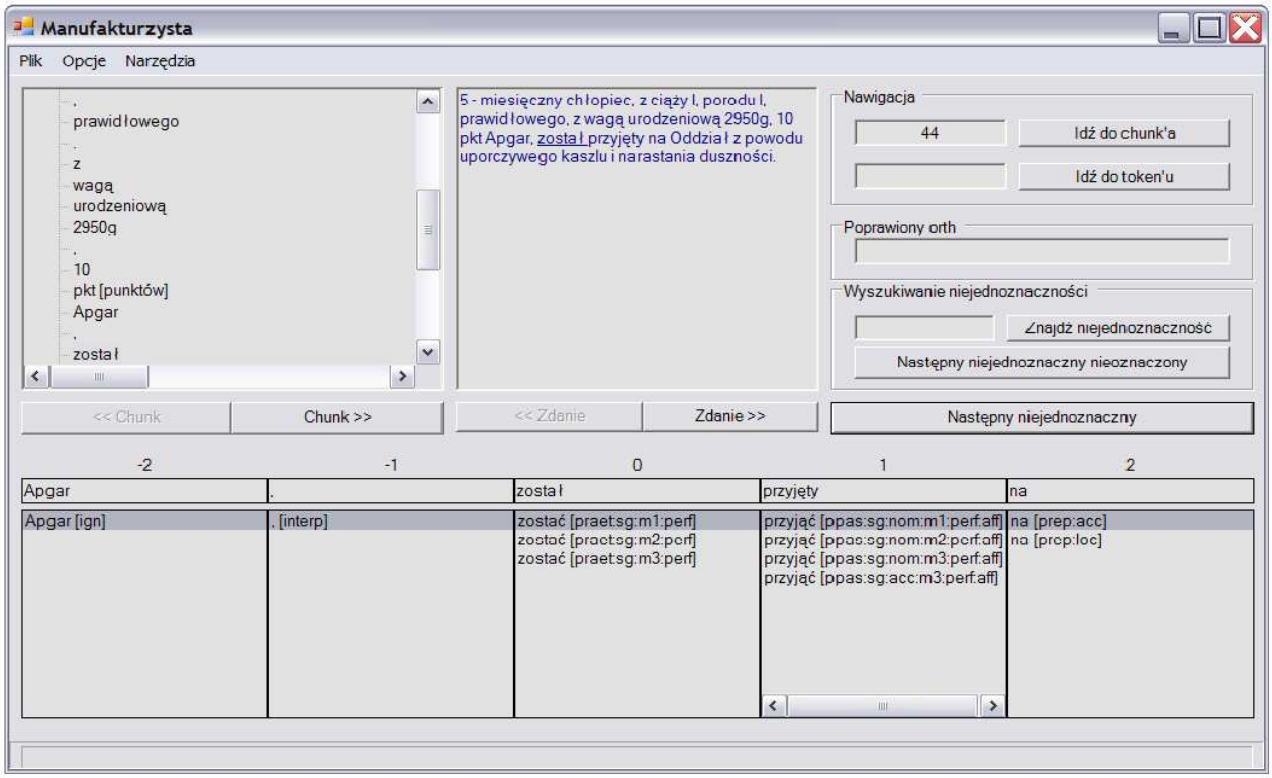

Figure 1. Annotations editor - Manufacturer

Firstly, we applied Morfeusz to the corpus in order to get a list of many thousands of unrecognised word forms. Next, a frequency list of unrecognised forms was examined manually starting from the most frequent forms. The list was divided by a linguist into four lists: proper forms, Latin or uncertain forms, abbreviations, and misspelled forms. The list of Latin or uncertain forms was later automatically filtered for the presence of proper Latin forms, and finally the forms from the rest of the list were described by a medical specialist (by automatic analysis of Latin, we spared his time).

About 1050 forms identified as forms of the proper words were added manually to the extended version of Morfeusz. For each form an inflectional paradigm was identified and the complete set of inflectional forms were defined with the help of a tool called *Klikadełko* constructed by Marcin Woliński (Polish grant 7T11C04320 — IPIC). The dedicated extended version of Morfeusz was prepared on the basis of the data described thanks to the kind help of him. The extended Morfeusz was used for further processing of KorMedIIS.

4. MORPHOSYNTACTIC ANNOTATION

KorMedIIS was automatically morphos-syntactically annotated with the help of TaKIPI tagger. As the accuracy of the general version of TaKIPI was about 92%, the number of words wrongly disambiguated was significant (the problem of the tagger is described in Sec 5). In order to

create a reliable pattern of the medical language, we decided to correct manually as large part of KorMedIIS as possible. Thus, the annotation generated by TaKIPI for 2127 epicrisises was manually checked and corrected by a linguist.

In MedKorIIS we preserve the standard of annotation defined in IPIC [2]. IPIC is at the moment the largest annotated corpus of the Polish language, and it is pointing towards to be representational, so it is reasonable to use the same system of annotation. An example of an IPIC tag (and the same of KorMedIIS tag) is presented in Fig. 2.

```
<tok>
  <orth>mężczyźnie</orth>
   <lex><base>mężczyzna</base>
              <ctag>subst:sg:dat:m1</ctag></lex>
   <lex><base>mężczyzna</base>
              <ctag>subst:sg:loc:m1</ctag></lex>
</tok>
```

Figure 2. An example of ICS PAS Corpus notation's structure.

In Fig. 2, XML the `tok` tag includes the whole description of a token from the corpus, `orth` represents the exact orthographic form present in text. Each `lex` tag represents a possible morpho-syntactic analysis, where `base` defines the basic morphological form of a word, and `ctag` contains a grammatical class (the first position) and values of all appropriate grammatical categories encoded in the IPIC standard.

As the XML based format is not especially readable for a human and the size of KorMedIIS is 375 MB, a special tool called *Manufakturzysta* (*manufacturer*) for the corpus correction and searching was developed. It is presented in Fig. 1. However, the main tasks to be supported by *Manufakturzysta* is the manual disambiguation of the corpus. When Manufakturzysta is used it to annotate the text, one chooses the correct tag from the list in the middle-bottom of the screen. It is also possible to choose several tags, when we cannot prefer one tag against the others, as all corresponding mopho-syntactic forms seem to be correct on the basis of the syntax of a sentence.

The applied procedure of manual disambiguation was simply manual correction of the automatic disambiguation generated by the standard version of TaKIPI. It speeded up the work greatly (e.g. in the comparison to the experience of IPIC). The linguist could go across only ambiguous words, going from one ambiguous word directly to the next one. The ambiguous word is put in the centre of the context (the bottom part of the screen in Fig. 1). The linguist can check the tags chosen for a word automatically and next can correct some possible mistake. All actions done are recorded. The linguist can also go to any word in a given *chunk* (a kind of generalised paragraph) of a text any time. Manufakturzysta was successfully used in the correction of disambiguation of 330 181 tokens (61,2 MB, but 31 507 unrecognised tokens). It means that the `manual' part of KorMedIIS is as large as one third of the manually disambiguated part of IPIC.

Manufakturzysta makes possible also the edition of the corpus or even its manual creation directly from text files. It can be also used to review the corpora sequentially, sentence by sentence, or *chunk* of text by chunk of text.

We have implemented several methods to search the corpus: by a number of a chunk or a token, by a specific word or by a type of ambiguity, e.g. searching all words ambiguous between

verb and noun and in the same time between singular and plural number. The search of ambiguity is very helpful in linguistic analysis of texts, e.g. for the needs of a tagger or parser construction.

The morpho-syntactic annotation of KorMedIIS gives an opportunity for creation of a well supported stochastic language model, which can be next used as correction tool in handwriting recognition.

5. TAGGER

At the beginning of the work, we were using our standard version of TaKIPI tagger [1] for the automatic annotation of the corpus. However TaKIPI was created using the manually disambiguated part of IPIC [2], which contains 885 669 tokens coming from different types of texts, (mostly written language), mostly quite carefully prepared texts.

The problem was that sentences observed in KorMedIIS are often non-standard. Sentences are mostly short, often discontinuous. One can often observe sequences of clauses or even phrases instead of proper sentences. Especial problem is the high percentage of use of acronyms and abbreviations for both medical terms but also as shorter versions of some common words. The percentage of words unknown to the morphosyntactic analyser or misspelled was much higher than in the texts of IPIC. When tagger TaKIPI is annotating the sentences, it takes into the consideration the whole sentence (as identified by Morfeusz). The presence of any form corresponding to the cases listed above is perceived as a perturbation in the inner sentence's structure. TaKIPI was originally trained with mostly complete description of a sentence and every "gap" is a possible source of its mistake. However, surprisingly, the accuracy of the standard TaKIPI, as measured on the `manual' part of KorMedIIS, appeared to be only a slightly lower than the normal one.

The problem of abbreviations was resolved twofold. Firstly, we added a new module to our tagger for recognition of the abbreviations. For the sake of efficiency, it has been implemented as a *transducer* (i.e. an automat encoding the abbreviations). If a string is recognised, then its full description, potentially ambiguous, i.e. a set of tags, is taken from the dictionary of abbreviations on the basis of the identifier returned by the transducer. Both, the transducer and the dictionary are successively extended by the most frequent abbreviations (as identified during the analysis of the corpus, see Sec. 2).

Secondly, as the IPIC annotation does not take abbreviations into account, we had to propose an extension to the IPIC standard. The extension describes an abbreviation in an expanded form. Each recognised abbreviation is described in KorMedIIS as both: the exact form of abbreviation and the full expanded word or sequence of words. The proposed notation allows also for the representation of ambiguous cases, when a string of characters can be the abbreviation, as well, as a proper word, e.g. "im." which can be: a pronoun *im* (*them*) and the full stop ".", or can be the abbreviation of *imienia* (*~under the name*).

The short form (i.e. the abbreviation itself) is used for the needs of representation (it is just a part of the original text, and as such is preserved), but for the needs of text processing the expanded form, annotated and disambiguated, can be read from the corpus. For example the if we encounter word "np" (*for example*), the following notation will be generated:

```
<chunk type="abbreviation" id="np">
    <tok>
```

```
        <orth>na</orth>
          <lex dissamb="1"><base>na</base>
                                <ctag>prep:acc</ctag></lex>
          <lex><base>na</base><ctag>prep:loc</ctag></lex>
    </tok>
    <tok>
        <orth>przykład</orth>
           <lex><base>przykład</base>
              <ctag>subst:sg:nom:m3</ctag></lex>
           <lex dissamb="1"><base>przykład</base>
              <ctag>subst:sg:acc:m3</ctag></lex>
    </tok>
</chunk>
```

We hope that this will allow the tagger (and later some parser) to better analyse the sentence without loosing its inner structure.

6. FURTHER RESEARCH

The transducer created for the processing of abbreviations can also be applied to the recognition of collocations that are numerous in this domain. Some phrases which are constantly used in text, can be assigned the same analysis in all occurrences, and the accuracy of the tagger should be improved after introducing the mechanism of their recognition.

KorMedIIS includes presently only epicrisises, but now, after weeks spent on the analysis of texts, we are aware that they are not so ideal representation of all types of texts as we had thought at the beginning. Thus we plan to extend the corpus with other types of texts. It is very simple in the case of automatically annotated part, and it will be done by our tools, but we want also to extend the manually disambiguated part of KorMedIIS with the new types of texts. The latter will be time consuming.

KorMedIIS will be used for the creation of a stochastic language model defined on the level of sequences of word classes. The model used e.g. in OCR. Next, the annotation of KorMedIIS will be further extended with the information concerning the main syntactic parts of a sentence or clause. This version of the corpus will be used in adapting a shallow parser (i.e. recognising the general structure) to the domain of medical texts.

REFERENCES

[1] PIASECKI M., GODLEWSKI G., Reductionistic, tree and rule based tagger for Polish. In Kłopotek, M.A., Wierzchoń, S.T., Trojanowski, K., eds. Proceedings of Intelligent Information Processing and Web Mining 2006, Springer Verlag, 2006.

[2] PRZEPIÓRKOWSKI A., The IPI PAN Corpus Preliminary Version. Institute of Computer Science PAS, 2004.

[3] Marcin Woliński, *System znaczników morfosyntaktycznych w korpusie IPI PAN*, Polonica, XXII–XXIII, s. 39–55, 2003

[4] WOLIŃSKI M., Morfeusz --- a practical tool for the morphological analysis of Polish. In Kłopotek, M.A., Wierzchoń, S.T., Trojanowski, K., eds. Proceedings of Intelligent Information Processing and Web Mining 2006, Springer Verlag, 2006.