DOI:10.15774/PPKE.ITK.2015.012

METHODS FOR PROCESSING NOISY TEXTS AND THEIR APPLICATION TO HUNGARIAN CLINICAL NOTES

Thesis for the Degree of Doctor of Philosophy

Borbála Siklósi

Roska Tamás Doctoral School of Sciences and Technology

Pázmány Péter Catholic University, Faculty of Information Technology and Bionics

Academic advisor:
Dr. Gábor Prószéky

2015

DOI:10.15774/PPKE.ITK.2015.012

DOI:10.15774/PPKE.ITK.2015.012

Methods for processing noisy texts and their application to Hungarian clinical notes

In most hospitals medical records are only used for archiving and documenting a patient's medical history. Though it has been quite a long time since hospitals started using digital ways for written text document creation instead of handwriting and they have produced a huge amount of domain specific data, they later use them only to lookup the medical history of individual patients. Digitized records of patients' medical history could be used for a much wider range of purposes. It would be a reasonable expectation to be able to search and find trustworthy information, reveal extended knowledge and deeper relations. Language technology, ontologies and statistical algorithms make a deeper analysis of text possible, which may open the prospect of exploration of hidden information inherent in the texts, such as relations between drugs and other treatments and their effects. However, the way clinical records are currently stored in Hungarian hospitals does not even make free text search possible, the look-up of records is only available referring to certain fields, such as the name of the patient. Aiming at such a goal, i.e. implementing an intelligent medical system requires a robust representation of data. This includes well determined relations between and within the records and filling these structures with valid textual data. In this research I was trying to transform raw clinical records written in the Hungarian medical language into a normalized set of documents to provide proper input to such higher-level processing methods.

DOI:10.15774/PPKE.ITK.2015.012

DOI:10.15774/PPKE.ITK.2015.012

Acknowledgements

“Mid van amit nem kaptál volna?”

I. Korinthus 4,7.

“What do you have that you did not receive?”

1 Corinthians 4:7

DOI:10.15774/PPKE.ITK.2015.012

DOI:10.15774/PPKE.ITK.2015.012

Contents

DOI:10.15774/PPKE.ITK.2015.012

DOI:10.15774/PPKE.ITK.2015.012

1

Introduction

In which the topic of this Thesis is introduced highlighting the main problems it is about to solve in the forthcoming Chapters. But I would like to prepare the Reader not to recall their painful memories at the sight of some textual realization of uneasy medical incidents...

DOI:10.15774/PPKE.ITK.2015.012

Processing medical texts is an emerging topic in natural language processing. There are existing solutions, mainly for English, to extract knowledge from medical documents, which thus becomes available for researchers and medical experts. However, locally relevant characteristics of applied medical protocols or information relevant to locally prevailing phenomena can be extracted only from documents written in the language of the local community.

As Meystre et al. (2008) point out, it is crucial to distinguish between clinical and biomedical texts. Clinical records are documents created at clinical settings with the purpose of documenting every-day clinical cases or treatments. The quality of this type of text stays far behind that of biomedical texts, which are also the object of several studies. Biomedical texts, mainly written in English, are the ones that are published in scientific journals, books, proceedings, etc. These are written in the standard language, in accordance with orthographic rules (Sager et al., 1994; Meystre et al., 2008). On the contrary, clinical records are created as unstructured texts without using any proofing tools, resulting in texts full of spelling errors and nonstandard use of word forms in a language that is usually a mixture of the local language (Hungarian in our case) and Latin (Siklósi et al., 2012, 2013). These texts are also characterized by a high ratio of abbreviated forms, most of them used in an arbitrary manner. Moreover, in many cases, full statements are written in a special notational language (Barrows et al., 2000) that is often used in clinical settings, consisting only, or mostly of abbreviated forms.

Another characteristic of clinical records is that the target readers are usually the doctors themselves, thus using their own unique language and notational habits is not perceived to cause any loss in the information to be stored and retrieved. However, beyond the primary aim of recording patient history, these documents contain much more information which, if extracted, could be useful for other fields of medicine as well. In order to access this implicit knowledge, an efficient representation of the facts and statements recorded in the texts should be created.

This noisy and domain-specific character of clinical texts makes it much more challenging to process than general and even biomedical texts. **The goal of my Thesis research was to create preprocessing methods designed explicitly to Hungarian clinical records, preparing them to reach a normalized representation suitable for deeper processing and information extraction.** The performance of any text processing algorithm depends on the quality of the input text created by humans (doctors). Relevant and correct information is only extractable if it is present in the input. My goal was to reconstruct the intended information in these clinical documents from their noisy and malformed state to provide them to higher level processing units.

There are two main approaches in processing clinical documents (Meystre et al., 2008). Methods falling into the first category apply rule-based algorithms. These are usually small, domain-specific applications that are expensive and time-consuming to build. The second group includes statistical algorithms. Though such methods are more and more popular, the main drawback is the need of large datasets, which are usually hard to obtain due to ethical issues. Moreover, supervised methods also require high quality annotations needed to be created manually by domain experts (in our case having both linguistic and medical expertise). Moreover, applications used for processing domain-specific texts are usually

DOI:10.15774/PPKE.ITK.2015.012

supported by some hand-made lexical resources, such as ontologies or vocabularies. In the case of less-resourced languages, there are very few such datasets and their construction needs quite an amount of human work.

In order to be able to support the adaptation of existing tools, and the building of structured resources, I examined a corpus of Hungarian ophthalmology notes. In this research, statistical methods are applied to the corpus in order to capture as much information as possible based on the raw data. Even though the results of each module are not robust representations of the underlying information, these groups of semi-structured data can be used in the real construction process.

DOI:10.15774/PPKE.ITK.2015.012

DOI:10.15774/PPKE.ITK.2015.012

2

The Hungarian Clinical Corpus

In which the mysterious language of medical practitioners is introduced. Having it read, the highly literate Reader will be more comfortable by having a bunch of scientific reasons why they do not understand documents received at the scene of medical treatments. The clinical language will be compared to general Hungarian along three characteristics: their syntactic behaviour, the frightening ratio of spelling errors and the not so frightening, but still very high ratio of abbreviations.

Contents

DOI:10.15774/PPKE.ITK.2015.012

Research in the field of clinical record processing has advanced considerably in the past decades and applications exist for records written in English. However, these tools are not readily applicable to other languages. In the case of Hungarian, agglutination and compounding, which yield a huge number of different word forms, and free word order in sentences render solutions applicable to English unfeasible.

Creutz et al. (2007) have compared the number of different word forms encountered in a corpus as a function of corpus size for English and agglutinating languages like Finnish, Estonian or Turkish. They found that while the number of different word tokens in a 10 million word English corpus is generally below 100 000, in Finnish it is well above 800 000. However, the 1:8 ratio does not correspond to the ratio of the number of possible word forms between the two languages: while there are about 4-5 different inflected forms for an English word, there are several hundred or thousand in any of these languages.

Similarly to these agglutinating languages, a corpus of a certain size is much less representative for Hungarian than it is for English. Moreover, existing tools for processing general Hungarian texts perform very poorly when applied to documents from the medical domain. Compared to a general Hungarian corpus, there are significant differences between the two domains, which explains the inapplicability of such tools. These differences are not only present in the semantics of the content, but in the syntax and even in the surface form of the texts and fall into three main categories discussed in the following subsections. The corpus used in the comparison as general text was the Szeged Corpus (Csendes et al., 2004), containing 1 194 348 tokens (70 990 sentences) and the statistics related to this corpus was taken from Vincze (2013).

2.1 Syntactic behaviour

The **length of the sentences** used in a language can reflect the complexity of the syntactic behaviour of utterances. In the general corpus, the average length of sentences is 16.82 tokens, while in the clinical corpus it is 9.7. However, in the case of clinical records, this difference does not mean that the sentences are simpler. Rather, the length of the sentences is reduced at the cost of introducing incomplete grammatical structures, which make the text more difficult to understand. Doctors tend to use shorter and rather incomplete and compact statements. This habit makes the creation of the notes faster, but being in lack of crucial grammatical constituents, most parsers fail when trying to process them.

Regarding the **distribution of part-of-speech** (pos) in the two domains, there are also significant differences. While in the general corpus the three most frequent types are nouns, verbs and adjectives, in the clinical domain nouns are followed by adjectives and numbers in the frequency ranking, while the number of verbs in this corpus is just one third of the number of the latter two. Another significant difference is that in the clinical domain, determiners, conjunctions, and pronouns are also ranked lower in the frequency list. These occurrence ratios are not surprising, since a significant portion of clinical documents record a statement (*something has a property*, which is expressed in Hungarian with a phrase containing only a

DOI:10.15774/PPKE.ITK.2015.012

noun phrase without a determiner and an adjective), or the result of an examination (*the value of something is some amount*, i.e. a noun phrase and a number). Furthermore, most of the numbers in the clinical corpus are numerical data. Table 2.1 shows the detailed statistics and ranking of pos tags in the two corpora.

	NOUN	ADJ	NUM	VERB	ADV	PRN	DET	POSTP	CONJ
CLIN	43,02%	13,87%	12,33%	3,88%	2,47%	2,21%	2,12%	1,03%	0,87%
SZEG	21,96%	9,48%	2,46%	9,55%	7,60%	3,85%	9,39%	1,24%	5,58%

	NOUN	ADJ	NUM	VERB	ADV	PRN	DET	POSTP	CONJ
CLIN	1	2	3	4	5	6	7	8	9
SZEG	1	3	8	2	5	7	4	9	6

Table 2.1: The distribution and ranking of part-of-speech in the clinical corpus (CLIN) and the general Szeged Corpus (SZEG)

2.2 Spelling errors

A characteristic of clinical documents is that they are usually created in a rush without any proofreading. The medical record creation and archival tools used at most Hungarian hospitals provide no proofing or structuring possibilities. Thus, the number of misspellings is very high and a wide variety of error types occur. These mistakes are due not only to the complexity of the Hungarian language and orthography, but also to characteristics typical of the medical domain and the situation in which the documents are created. The most frequent types of errors are the following:

- mistyping, accidentally swapping letters, inserting extra letters or just missing some,
- lack or improper use of punctuation marks (e.g. no sign of sentence boundaries, missing commas, no space between the punctuation mark and the neighbouring words),
- grammatical errors,
- sentence fragments,
- domain-specific and often ad hoc abbreviations, which usually do not correspond to any standard
- Latin medical terminology not conforming to orthographic standards.

A common feature of these phenomena is that the prevailing errors vary with the doctor or assistant typing the text. Thus, it is possible that a certain word is mistyped and should be corrected in one document, while the same word is a specific abbreviation in another one, which does not correspond to the same concept as the corrected one.

Compared to the Szeged Corpus, there are two main differencies: the ratio of abbreviations and that of spelling errors. While 0.08% of the tokens of the Szeged Corpus are abbreviations, this ratio is 7.15% in the clinical corpus Siklósi and Novák (2013). This ratio was determined

DOI:10.15774/PPKE.ITK.2015.012

from a 15 278-token-long portion of the clinical corpus. In this portion, misspellings were also marked manually. Thus, more detailed statistics describing the characteristics of misspellings in these documents could be measured from this, which is included in Table 2.2. The ratio of misspellings in the clinical corpus is 8.44%, which is only 0.27% in the Szeged Corpus. The subcorpus of the latter one containing texts written by primary school students has also a much lower ratio (0.87%) of misspelled words compared to the clinical corpus. In the case of the clinical documents, more than half of these errors are punctuation errors (the most frequent is the missing period at the end of abbreviations). Errors of joining or separating words occur with similar frequencies, adding up 10% of all the errors. Beside leaving the period from abbreviations, punctuation errors still occur very frequently in clinical texts. While only 1.04% of the sentences in the Szeged Corpus lack sentence final period (titles and headings), this ratio is 48.28% in the clinical corpus. Sentence initial capitalization shows similar problems: while only 0.42% of sentences in the Szeged Corpus do not start with capitalized letters, this ratio is 12.81% in the clinical corpus, which makes the task of sentence segmentation a challenging problem (Orosz et al., 2013).

Regarding punctuations, however, it should be noted that such marks might have different roles in the clinical documents. For example, the question mark can stand for some uncertainty of a finding or result even in the middle of sentences. For example: *'H-tokon nagy capsulot. (?) nyilás (rupture?)'*, meaning that these are hypothetical diagnoses, but the doctor is not certain about them.

	misspelled	punctuation	joining	separating	other
Szeged Corpus	0,27%	-	-	-	-
Szeged Corpus – primary school	0,87%	-	-	-	-
Clinical corpus	8,44%	46,55%	5,66%	5,59%	42,2%

Table 2.2: The ratio of different types of misspellings found in a subcorpus of clinical documents and in the Szeged Corpus

2.3 Abbreviations

The use of a kind of notational text is very common in clinical documents. This dense form of documentation contains a high ratio of standard or arbitrary abbreviations and symbols, some of which may be specific to a special domain or even to a doctor or administrator. These short forms might refer to clinically relevant concepts or to some common phrases that are very frequent in the specific domain. For the clinicians, the meaning of these common phrases is as trivial as the standard shortened forms of clinical concepts due to their expertise and familiarity with the context. They do not rely on orthographic features that would isolate abbreviations from unabbreviated words. Thus, word final periods are usually missing, abbreviations are written with varying case (capitalization) and in varying length. For example the following forms represent the same expression, *vörös visszfény* 'red reflection': *vvf, vvfény, vörösvfény.*

DOI:10.15774/PPKE.ITK.2015.012

Another characteristic feature of the abbreviations in these medical texts is the partially shortened use of a phrase, with a diverse variation of choosing certain words to be used in their full or shortened form. The individual constituents of such sequences of abbreviations are by themselves highly ambiguous, especially if all tokens are abbreviated. Even if there were an inventory of Hungarian medical abbreviations, which does not exist, their detection and resolution could not be solved. Moreover, the mixed use of Hungarian and Latin phrases results in abbreviated forms of words in both languages, thus the detection of the language of the abbreviation is another problem.

From the perspective of automatic spelling correction and normalization, the high number of variations for a single abbreviated form is the most important drawback. Table 2.3 shows some statistics about the different forms of an abbreviated phrase occurring in our corpus. Although there is a most common abbreviated form for each phrase, some other forms also appear frequently enough not to be considered as spelling errors. For a more detailed description about the behaviour of medical abbreviations see Chapter 5.

The difference in the ratio of abbreviations in the general and clinical corpora is also significant, being 0.08% in the Szeged Corpus, while 7.15% in the clinical corpus, which means that the frequency of abbreviations is two orders of magnitude larger in clinical documents than in general language.

oculus sinister	freq.	oculus dexter	freq.	oculi utriusque	freq.
o. s.	1056	o. d.	1543	o. u.	897
o.s.	15	o.d.	3	o.u.	37
o. s	51	o. d	188	o. u	180
os	160	od	235	ou	257
O. s.	118	O. d.	353	O. u.	39
o. sin.	348	o. dex.	156	o. utr.	398
o. sin	246	o. dex	19	o. utr	129
O. sin	336	O. dex	106	O. utr	50
O. sin.	48	O. dex.	16	O. utr.	77

Table 2.3: Corpus frequencies of some variations for abbreviating the three phrases *oculus sinister*, *oculus dexter* and *oculi utriusque*, which are the three most frequent abbreviated phrases.

2.4 The domain of ophthalmology

In a broad sense, there are two sources of clinical documents regarding the nature of these textual data. First, they might be produced through an EHR (Electronic Health Records) system. In this case, practitioners or assistants type the information into a predefined template, resulting in structured documents. The granularity of this structure might depend on the actual system and the habit of its users. The second possibility is that the production of these clinical records follows the nature of traditional hand-written documents, i.e. even though they are stored in a computer, it is only used as a typewriter, resulting in raw text,

DOI:10.15774/PPKE.ITK.2015.012

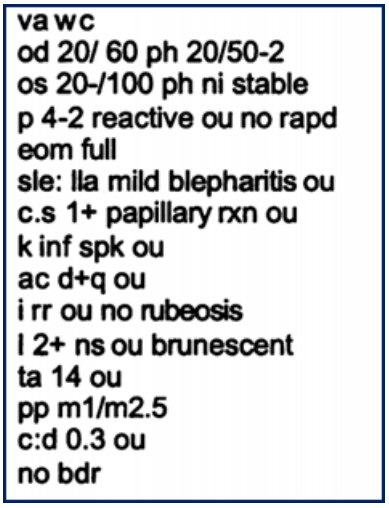

va wc
od 20/ 60 ph 20/50-2
os 20-/100 ph ni stable
p 4-2 reactive ou no rapd
eom full
sle: lla mild blepharitis ou
c.s 1+ papillary rxn ou
k inf spk ou
ac d+q ou
i rr ou no rubeosis
l 2+ ns ou brunescent
ta 14 ou
pp m1/m2.5
c:d 0.3 ou
no bdr

Figure 2.1: A portion of an ophthalmology record in English

having some clues of the structure only in the manual formatting. Of course, these are the two extremes, and the production of such records is usually somewhere in between, depending on institutional regulations, personal habits and the actual clinical domain as well. Whatever the format of the source of these documents are, the value of the content is the same, thus it is the processing methodology that should be adjusted to the constraints of the source.

In Hungarian hospitals, the usage of EHR systems is far behind expectations. Assistants or doctors are provided with some documentation templates, but most of them complain about the complexity and inflexibility of these systems. This results in keeping their own habit of documentation, filling most of the information into a single field and manually copying patient history.

Moreover, ophthalmology has been reported to be a suboptimal target of application of EHR systems in several surveys carried out in the US (Chiang et al., 2013; Redd et al., 2014; Elliott et al., 2012). The special requirements of documenting a mixture of various measurements (some of them resulting in tabular data, while others in single values or textual descriptions) make the design of a usable system for storing ophthalmology reports in a structured and validated form very hard.

Another unique characteristic of documentation in the field of ophthalmology is that the documents are created in a rush, during the examination. Thus, the adhoc use of abbreviations, frequent misspellings and the use of a language that is a mixture of English, Latin and the local language are very common phenomena. Moreover, even in the textual descriptions, essential grammatical structures are missing, making most general text parsers fail when trying to process these texts. This is true even at the lowest levels of processing, such as tokenization, sentence boundary detection or part-of-speech tagging. Figure 2.1 shows an example of an English ophthalmology note demonstrating the complexity of the domain, which is even worse for Hungarian due to the complexity of the language itself. My methods, described in the following chapters, are designed to satisfy all these constraints.

DOI:10.15774/PPKE.ITK.2015.012

3

Accessing the content

"A typewriter is a mechanical or electromechanical machine for writing in characters similar to those produced by printer's movable type by means of keyboard-operated types striking a ribbon to transfer ink or carbon impressions onto the paper. Typically one character is printed per keypress. The machine prints characters by making ink impressions of type elements similar to the sorts used in movable type letterpress printing."[i]
And some treat computers in a similar manner, without the paper part. This Chapter will describe how to turn documents created with such an attitude into machine readable records.

Contents

[i] Definition is from the typewriter section of Wikipedia, 2015

DOI:10.15774/PPKE.ITK.2015.012

We were provided anonymized clinical records from various medical fields, and ophthalmology was chosen out of them to build the pivot system that can be extended later to other fields as well. The first phase of processing raw documents was to compensate the lack of structural information. Due to the lack of a sophisticated clinical documentation system, the structure of raw medical documents can only be inspected in the formatting or by understanding the actual content. Besides basic separations - that are not even unified through documents - there were no other aspects of determining structural units. Moreover a significant portion of the records were redundant: medical history of a patient is sometimes copied to later documents at least partially, making subsequent documents longer without additional information regarding the content itself. However, these repetitions will provide the base of linking each segment of a long lasting medical process. See Figure 3.1 as an example for an original document in raw text format.

In order to be able to process these documents, their content had to be extracted while preserving the clues of the original structure. Thus, first the overall structure was defined by an XML scheme and was populated by the documents. Then, those parts that contained textual information were further divided into sentences and words in order to be able to consider such units as the base of any higher level processing. However, the extraction of these basic elements was not a trivial task either.

```
                          A M B U L Á N S   K E Z E L O L A P

Státusz
  2010.10.19 12:28
  Olvasó szemüveget szeretne. Néha könnyeznek a szemei.
          //S/he would like reading glasses, eyes are sometimes watering.

  V:0,7+0,75Dsph=1,0
    1,0 +0,5 Dsph élesebb

  +2.0 Dsph mko Cs IV

  St.o.u: halvány kh, ép cornea, csarnok kp mély tiszta, iris ép békés, pupilla
          // St.o.u: blanch conj, intact cornea, chamber deep clean, iris intact, calm, pupil
  rekciók rendben, lencse tiszta, jó vvf.
          //reactions allright, clean lens, good rbl.
  Átfecskendezés mko sikerült.
          //Successful squishing at both side

  Olvasó szemüveg javasolt: +2.0 Dsph mko.
          //Reading glasses are suggested: +2.0 Dsph both side
  Éjszakánként mukönnygél ha szükséges.
          //Artificial tears can be used at night if necessary
  Kontroll: panasz esetén
          //Contorl: in case of further complaints

Diagnózis //Diagnosis
  DIAGNÓZISOK megnevezése                    Kód    Dátum      Év   K V T
  Látászavar, k.m.n.                         H5390  2010.10.19          3

Beavatkozások //Treatments
   Kód      Megnevezés                                  Menny.   Pont
   11041    Vizsgálat                                        1    750
```

Figure 3.1: A clinical record in its original form. Lines starting with '//' are the corresponding English translations. In order to exemplify the nature of these texts, spelling errors and abbreviations are kept in the translation.

DOI:10.15774/PPKE.ITK.2015.012

3.1 XML structure

Wide-spread practice for representing structure of texts is to use XML to describe each part of the document. In our case it is not only for storing data in a standard format, but also representing the identified internal structure of the texts which are recognized by basic text mining procedures, such as transforming formatting elements to structural identifiers or applying recognition algorithms for certain surface patterns. After tagging the available metadata and performing these transformations the structural units of the medical records are the followings:

- content: parts of the records that are in free text form. These should have been documented under various sub-headings, such as *header*, *diagnoses*, *applied treatments*, *status*, *operation*, *symptoms*, etc. However, at this stage, all textual content parts are collected under this content tag.
- metadata: I automatically tagged such units as the type of the record, name of the institution and department, diagnoses represented in tabular forms and standard encodings of health related concepts.
- simple named entities: *dates*, *doctors*, *operations*, etc. The medical language is very sensitive to named entities, that is why handling them requires much more sophisticated algorithms, which are a matter of further research.
- medical history: with the help of repeated sections of medical records related to one certain patient, a simple network of medical processes can be built. Thus, the identifiers of the preceding and following records can be stored.

3.2 Separating Textual and Non-Textual Data

The resulting structure defines the separable parts of each record, however this separation is not yet satisfactory for accessing those parts of the documents that can be handled as texts. The documents of ophthalmology investigated in this research were especially characterized by nontextual information inserted into sections containing texts as well. These (originally tabular) data behave as noise in such context. Non-textual information inserted into free word descriptions are laboratory test results, numerical values, delimiting character series and longer chains of abbreviations and special characters. See Figure 3.2 for some specific examples of textual and nontextual data. Though these statements contain relevant information from the aspect of the actual case, filtering them out was necessary to create a textual corpus as the base of further preprocessing steps. However, as these statements do not follow any standard patterns even by themselves and they further vary by documents of different style, doctor or assistant, I could not define any rules or pattern matching algorithms to perform the filtering.

To solve this issue, unsupervised clustering methods were applied. Prior to categorization, units of statements had to be declared. The documents were exported from the original system in a way that kept the fixed width of the original input fields. Thus, linebreaks were

DOI:10.15774/PPKE.ITK.2015.012


```
V -1,0 Dsph -1,5Dcyl 180° =0,5 -1,25 Dsph -1,0 Dcyl 70° =0,25
CFF: 37/37 Hz
Felvételkor o.d.  2mou -4,5Dsph -0,75Dcyl 120° =0,6 Tappl:15/14 Hgmm
o.s.  0,1-2,5Dsph -0,5Dcyl 75° =1,0
Távozáskor o.d.  0.1-3.0 Dsph -0.75Dcyl 120f =0.8 Tdig:n/n
o.s.idem
```

(a) Tabular data mixed into textual parts

```
Betegünk szemészeti anamnézisében bal szem szürkehályog ellenes műtéte
szerepel.
Jelen felvételére jobb szemen lévő szürkehályog műtéti megoldása céljából
került sor.
o.utr:  békés elülső segment.
Lencsében finom maghomály.
Üvegtestben sűrűn kristályok.
Dg:  Myopia c.ast.  o.utr., Cat.incip.  o.utr., Asteroid hyalosis o.utr.
```

(b) Lines considered to be texts

Figure 3.2: Examples for nontextual (a) and textual (b)data found in the documents in a mixed manner. The separation is the result of the clustering algorithm.

inserted to the text at certain positions corresponding to this width. In order to restore the original units intended to be single lines, these linebreaks were deleted from the end of a line which could be continued by the next one. That is, if the second line does not start with capital letter, does not start with whitespace and if the length of the actual line plus the length of the first word of the second line is larger than the fixed width (hyphenation was not implemented in the system, thus if a word would pass the right margin, then the whole word is transmitted to a new line). Moreover, lines containing tabular data were also recognized during this processing step. The units of categorization were these concatenated lines

Thus, such short textual fragments were kept together with more representative neighbours avoiding them to be filtered out by themselves, since their feature characteristics are very similar to those of non-textual lines.

Then, I applied k-means clustering algorithm to these lines to group them as either text or nontext. Each line was represented by a feature vector containing the characteristics of the given line, such as the number of words, number of characters, number of alphanumeric characters, number of punctuation marks, ratio of vowels and consonants, ratio of numbers, number of capital letters, etc. The goal was to create two disjunct sets, however setting $k = 2$ did not yield in an efficient separation. Since the results could not be improved by modifying the feature set either, I increased the number of resulting clusters. Finally, the best grouping was achieved in the case of $k = 7$. Out of these seven sets, two contained real texts, the other five contained nontextual data of different types. Using the same feature set and having the clustered lines as training data, a classifier could also be set up in order to be able to classify new data without reclustering the whole corpus. A simple Naïve Bayes classifier performed with 98% accuracy on a test set of 100 lines.

DOI:10.15774/PPKE.ITK.2015.012


```
<sent>
     <surf>Azarga th. kezdünk </surf>
     <w NE="b-MED" id="102.0.0" type="">
          <orig>Azarga</orig>
          <corr>Azarga</corr>
          <lemma>Azarga</lemma>
          <pos>[N][NOM]</pos>
     </w>
     <w NE="" id="102.0.1" type="abbr">
          <orig>th.</orig>
          <corr>th.</corr>
          <lemma>th</lemma>
          <pos>[N][NOM]</pos>
     </w>
     <w NE="" id="102.0.2" type="">
          <orig>kezdünk</orig>
          <corr>kezdünk</corr>
          <lemma>kezd</lemma>
          <pos>[V][PL1]</pos>
     </w>
</sent>
```

Figure 3.3: The xml representation of the sentence "Azarga th. kezdünk"

Digging deeper into the textual contents of the documents, a more detailed representation of these text fragments was necessary. That is why I store each word in each sentence in an individual data tag, augmented with several information. Such information are the *original form* of the word, the *corrected form*, its *lemma* and *part-of-speech* tag, and some phrase level information such as different types of *named entities*. The lemma and PoS information are produced by PurePos (Orosz and Novák, 2012), the named entities are produced by the system created for the Master Thesis of Pirk (2013). The sentence *Azarga th. kezdünk* ('We start Azarga th.') is represented in Figure 3.3. As shown in the example, the statement is considered a sentence, even though the sentence-final period is missing. The `<surf>` tag contains the surface (the original) form of the statement. Then, each word is represented by a `<w>` tag that has several attributes, such as whether the word is part of a named entity (the attribute values are defined by IB-tags), whether it is an abbreviation and each word has a unique identifier as well. Moreover, four forms of each word is stored, i.e. its original form, its corrected form (produced automatically by the system described in Chapter 4), its lemma and its part-of-speech tag. In the example, the word *Azarga* is a named entity of type medication, while *th.* is an abbreviated form (of the word *therapia* 'therapy'). It should be noted that *therapy* in this sentence should be in the accusative case in Hungarian (i.e. *therapiát* instead of *therapia*), but this is not explicit in the written form, thus the part-of-speech tagger is not able to assign the correct label to this word and it is stored as a noun in the nominative case. As described in Chapter 5, disregarding orthographic standards, inflectional suffixes are hardly ever attached to abbreviated words.

DOI:10.15774/PPKE.ITK.2015.012

3.3 Structuring and categorizing lines

Having the documents preprocessed by the methods described above, an enriched representation of the corpus was achieved. However, the textual content segments, each intended to appear under various headings, still remained as a mixture under the *content* tag. The original sections under these headings (*header*, *diagnoses*, *applied treatments*, *status*, *operation*, *symptoms*, etc.) contain different types of statements requiring different methods of higher-level processing. Moreover, the underlying information should also be handled in different ways, unique to each subheading. Thus, the categorization of the content to these structural units was unavoidable. This was performed in two steps. First, formatting clues were recognized and labelled. Second, each line was classified into a content unit defined on statistical observations from the corpus.

3.3.1 Structuring

Even though the documentation system used when creating these documents did provide a basic template for labelling some sections of the document to be created, these were very rarely followed by the administrative personnel. However, some of these system generated labels were printed into the final documents, which I could consider as 'clues' of the intended structure. These system generated labels followed a consistent pattern, and as such, could easily be recognized based on features such as the amount of white space at the beginning of the line, capitalization, and the recurring text of the headline. Thus, such structural units were identified and their beginning was labelled with a `PART` tag (referring to different *parts* of the document).

Similarly, tables of codes were also printed by the system in a predefined format. These tables contain the BNO-codes (the Hungarian translation of ICD coding) of diagnoses and the applied treatments. Such tables, though printed as raw text, could also be recognized by the spacing used in them and were labelled with an `SPART` (structured part) tag distinguishing them from textual parts of the document.

3.3.2 Detecting patient history

We found it very often that findings about a patient recorded in documents of earlier visits were copied to the actual record, and in some cases minor adjustments were also introduced during the replication. Thus, although these partial recurrences contain only redundant information, they could not be recognized by simply looking for exact matches. Moreover, the short and dense statements of findings are often formatted the same way in the case of different patients or even doctors. Thus, in order to filter these copied sections, first we detect all date stamps in each document. Date stamps may occur in the headers, in the notation of some examinations, in the tables of codings or might be inserted manually at any point in the documents. The dates were labelled with a `DATE` tag. Then, the contents

DOI:10.15774/PPKE.ITK.2015.012

between these tags were ordered in increasing order and partial matches were found by comparing the `md5` coded form of each part. Those sections that had a matching under an earlier date stamp, were labelled with a `COPY` tag. Furthermore, these `DATE` tags were used to partition each document corresponding to separate visits. Thus, patient history could be retrieved by referring to the same ID and each date. All the information that was originally in a single document can thus be retrieved in order.

3.3.3 Categorizing statements

Even though the `PART` tags have labelled each part according to the documentation template of the system, the title of these fields is rarely in accordance with the content. For example, the status field is frequently used to include all the information, be it originally anamnesis, treatment, therapy, or any other comments. Thus, it was necessary to categorize each statement in each part of the documents. The units of categorization were the concatenated lines (see Chapter 3.2). Moreover, lines containing tabular data were also recognized during this processing step based on the indentation at the beginning of a line and the amount and appearance of whitespace within a line.

The set of these categories, or intentional subheadings, was defined with the help of an ophthalmologist. The categories, their definitions and an example sentence is shown in Table 3.1. These parts are, however, not always present in all the records and there is no compulsory order of these types of statements. However, by tradition, anamnesis is usually in the beginning, while the diagnoses and opinions are at the end. Almost every document includes visual acuity measurements (sometimes nothing else). The original documentation system also provides the possibility to type these data into different fields, but the granularity of these templates is much less sophisticated, and doctors do not tend to use them.

First, using the preprocessed version of the texts, some patterns were identified based on part-of-speech tags and the semantic concept categories assigned to the most frequent entities. For example, due to the rare use of verbs, if a past tense verb was recognized in a sentence, it was a good indicator of being part of the anamnesis or the complaints of the patient. See Chapter 6.4

Second, some indicator words were extracted from the documents. At the first place, these were those line initial words and short phrases that started with capital letter and were followed by a colon and some more content. These phrases were then ordered by their occurrence frequencies. Then, they were manually assigned a category label referring to the type of the statement that the phrase could be an indicator of. For example the phrase *korábbi betegségek* ‘previous illnesses’ was given the label `Ana` referring to anamnesis. Table 3.2 shows some more examples of tags and phrases labelled by them. After having all the phrases occurring at least 10 times in the whole corpus labelled, they were matched against the lines of each document that were found in `PART` sections and were not recognized as tabular data. If the line started with a phrase or any of its variations (case variations, misspellings, punctuation marks and white spaces were allowed differences), then the line was labelled

DOI:10.15774/PPKE.ITK.2015.012

tag	meaning	definition and example
Tens	Tension	Eye pressure expressed in millimeters of mercury, representing the pressure inside the eye. e.g.: T:18 Hgmm
V/Refr	Refraction	Refraction/Visus/visual acuity refers to the clarity of vision. The examined correction is measured in spherecal and cylindric dioptre. e.g.: Refr: +1.25Dsph -2.5Dcyl 5'
Ana	Anamnesis	Anamnesis includes patient history, family history and the description of the problem and the cause by the patient. May also include allergy information. e.g.: Hamar elfárad a szeme,vibráló fényeket kb.fél éve lát, nem tudja, melyik szemében. // His eyes quickly get tired, has seen flashing lights for about half a year, doesn't know in which eye.
Dg	Diagnosis	The diagnoses found during the examination (or during previous examinations). e.g.: Dg: Cat. incip. ou., Dystrophia con. lu., Keratitis ou., Háttér retinophatia és retinális elváltozások. // Dg: Cat. incip. ou., Dystrophia con. lu., Keratitis ou., Background retinophatia and retinal anomalies.
Beav	Treatment	The applied treatment during, before or after the actual examination. e.g.: 2010.05.27 08:25 - (2SOCT) OCT + FLAG
Vél	Opinion	The opinion or the suggestions of the doctor. It is not an official diagnosis, but might contain the description of the diagnosis. e.g.: vél: jelenleg szemészeti teendö nincs // op: no further action is needed
St	Status	The actual state of the patient. Usually includes the results of the performed examination but without the diagnoses. e.g.: Jelen státusz: // Present state:
Ther	Therapy	Applied or prescribed therapy e.g.: 2x Humapent
BNO	BNO (ICD)	The BNO code of a disease or treatment. e.g.: BNO: H00100 CHALAZION
T	Test	Performed tests, except refraction measurements. Most commonly slit lamp or ultrasound tests are applied. e.g.: Látótér od beszükült, de korábbinál jobb // Field of vision is narrower, but better than before.
V	Visus	Refraction/Visus/visual acuity refers to the clarity of vision. The examined correction is measured in spherecal and cylindric dioptre. e.g.: V: 0.8 +1.0 Dsph -2.5Dcyl 30' =1.0
Rl	Slit lamp	Slit lamp is used to examine the inner parts of the eye. The state of the different parts are described as seen by the doctor using the slit lamp. e.g.: Fundus: éles szélü, jó színü papilla nívóban, maculatáj fénytelen, sclerotikus erek, perif. ép. // Fundus: sharp edges, good colored standard papil, the area of the macula is dim, sclerotic veins, perif. intact
Kontr	Control	The decision about the next visit or control examination. e.g.: Kontroll 2-3 hónap múlva vagy panasz esetén // Control in 2-3 months or in the case of complaints.
Műtét	Operation	Applied or prescribed operations. e.g.: Phaco + PCL impl. o.sin. (Dr R. Zs.)
XXX	-	e.g.: Tisztelt Háziorvos! // Dear Family Doctor,

Table 3.1: The tags with their meaning definitions, and an example sentence

DOI:10.15774/PPKE.ITK.2015.012

with the tag the phrase belonged to. These first two steps were able to categorize 34% of the concatenated lines in the documents.

tag	phrase	English translation
	egyéb betegség	other illness
	panasz	complaint
Ana	család	family
	korábbi	earlier
	hypertonia	hypertonia
	anamnézis	anamnesis
	eredmény	result
T	ultrahang	ultrasound
	Topo	Topo
	Schirmer	Schirmer
	réslámpa	slit lamp
	macula	macula
RL	fundus	fundus
	rl	sl (for slit lamp)
	lencse	lens
	th	th (for therapy)
	szemcsepp	eyedrop
Ther	terápia	therapy
	rendelés	prescription
	javasolt	recommended

Table 3.2: Examples of tags and some of the phrases labelled by the tag.

In the third step, the rest of the lines were given a label. In order to do this, all lines labelled in the first two steps were collected for each tag (they will be referred to as tag collections). Then, for each line, the most similar tag collection was determined and the tag of this collection was assigned to the actual line. The similarity measure applied was the tf-idf weighted cosine similarity between a line (l) and a tag collection (c) defined by Formula 3.1.

$$sim(\vec{l},\vec{c}) = \frac{\sum\limits_{w\in l,c} tf_{w,l}tf_{w,c}(idf_w)^2}{\sqrt{\sum\limits_{l_i\in l}(tf_{l_i,l}idf_{l_i})^2}\times\sqrt{\sum\limits_{c_i\in c}(tf_{c_i,c}idf_{c_i})^2}} \tag{3.1}$$

where $\vec{l}$ contained the normalized set of words in line l, and $\vec{c}$ the normalized set of words contained in the tag collection c. During normalization, stopwords and punctuation marks were removed and numbers were replaced by the character x, so that the actual numerical values do not mislead the representation. As a result, all lines within `PART` sections were labelled by a tag. Finally, tabular lines were assigned the tag `Vis`, since these contained the detailed information about the visual acuity of the patient.

DOI:10.15774/PPKE.ITK.2015.012

3.3.4 Results

The labels of 1000 lines were checked manually. This testset was selected randomly only from `PART` sections, since the categorization was applied only to these portions of the documents. However, the label `XXX` was also allowed in the system when it was not able to assign any meaningful labels. The rest of the lines were assigned one of the 14 labels shown in Table 3.1. In the evaluation setup, these labels were considered either as correct, non-correct or undecidable. Lines of this latter category either did not include enough information referring to the content, or it was too difficult even for the human evaluator to decide what category the line belonged to. The label `XXX` was accepted as correct, if the line did not belong to any category (e.g. a single date). Out of the 1000 lines in the test set, its 7.8% could not be categorized by the human expert. For the rest of the lines, 81.99% of these lines were assigned the correct label and only 18.01% the incorrect one.

Regarding the errors, most of them were due to the lack of contextual information for the algorithm. For example, if the anamnesis of a patient included some surgery, then the label for surgery was assigned to it, which is correct at the level of standalone statements, but incorrect in the context of the whole document. The other main source of the errors was that some longer lines included more than one types of statements and the system was unable to choose a correct one. In these cases, the human annotation assigned the "more relevant" tag as correct. Thus, a significant part of these errors could be eliminated by a more accurate segmentation for separating each statement and by the incorporation of contextual features to the categorization process.

Figure 3.4 shows a document at this stage of processing. The beginning of structural parts is labelled with tags starting with # symbols, line category labels are written at the beginning of textual lines. The 'C symbols indicate possible line concatenations, which is only applied if the 'C symbol is preceded by a @ symbol at the end of the previous line.

DOI:10.15774/PPKE.ITK.2015.012


```
###DOCTYPE:AMBULÁNS KEZELOLAP
‘T                                A M B U L Á N S    K E Z E L O L A P

###PART:Státusz //Status
St       Státusz

##DOCDATE##
##DATE-TIME##
XXX ‘T        2010.10.19 12:28

Beav ‘C        Olvasó szemüveget szeretne. Néha könnyeznek a szemei.
                    //S/he would like reading glasses, eyes are sometimes watering.
V         V:0,7+0,75Dsph=1,0
V           1,0 +0,5 Dsph élesebb

V\Refr         +2.0 Dsph mko Cs IV

St         St.o.u: halvány kh, ép cornea, csarnok kp mély tiszta, iris ép békés, pupilla@
                // St.o.u: blanch conj, intact cornea, chamber deep clean, iris intact, calm, pupil
    ‘C   rekciók rendben, lencse tiszta, jó vvf.
                //reactions allright, clean lens, good rbl.
Ther         Átfecskendezés mko sikerült.
                //Successful squishing at both side
V\Refr ‘C       Olvasó szemüveg javasolt: +2.0 Dsph mko.
                //Reading glasses are suggested: +2.0 Dsph both side
Vél ‘C        Éjszakánként mükönnygél ha szükséges.
                //Artificial tears can be used at night if necessary
Kontr         Kontroll: panasz esetén
                //Contorl: in case of further complaints


###SPART:Diagnózis  //Diagnoses
      Diagnózis
        DIAGNÓZISOK megnevezése                          Kód    Dátum      Év   K V T
##DOCDATE##
        Látászavar, k.m.n.                               H5390  2010.10.19         3

###SPART:Beavatkozások  //Treatments
      Beavatkozások
         Kód      Megnevezés                                    Menny.   Pont
         11041    Vizsgálat                                          1    750
```

Figure 3.4: A document tagged with structural labels and line category labels.

DOI:10.15774/PPKE.ITK.2015.012

DOI:10.15774/PPKE.ITK.2015.012

4

Context-aware automatic spelling correction

In which a machine translation system is described.
Th sstem is albe to traanslat suc htext To normal.
... or at least to a quasi-standard form by applying methods of statistical machine translation.
The previous sentence might as well be a real example, but the Chapter will return to the medical domain.

Contents

DOI:10.15774/PPKE.ITK.2015.012

In Hungarian hospitals, clinical records are created as unstructured texts, without any proofing control (e.g. spell checking). Moreover, the language of these documents contains a high ratio of word forms not commonly used: such as Latin medical terminology, abbreviations and drug names. Many of the authors of these texts are not aware of the standard orthography of this terminology. Thus the automatic analysis of such documents is rather challenging and automatic correction of the documents is a prerequisite of any further linguistic processing.

The errors detected in the texts fall into the following categories: errors due to the frequent (and apparently intentional) use of non-standard orthography, unintentional mistyping, inconsistent word usage and ambiguous misspellings (e.g. misspelled abbreviations), some of which are very hard to interpret and correct even for a medical expert. Besides, there is a high number of real-word errors, i.e. otherwise correct word forms, which are incorrect in the actual context. Many misspelled words never or hardly ever occur in their orthographically standard form in our corpus of clinical records.

Moreover, it is a separate task to detect whether an unknown token is a variation of an abbreviated form or a misspelled form. In the latter case, it should be corrected to one of its standard forms. Text normalization might include the resolution of abbreviations, but in order to have them resolved, all misspelled forms must be corrected.

In this Chapter, I present a method for considering textual context when recognizing and correcting spelling errors. My system applies methods of Statistical Machine Translation (SMT), based on a word-based system for generating correction candidates. First the context-unaware word-based approach is described for generating correction suggestions, then its integration into an SMT framework is presented. I show that my system is able to correct certain errors with high accuracy, and, due to its parametrization, it can be tuned to the actual task. Thus the presented method is able to correct single errors in words automatically, making a firm base for creating a normalized version of the clinical records corpus in order to apply higher-level processing.

4.1 Automatic spelling correction

4.1.1 The word-based setup

First, a word-based system (Siklósi et al., 2012) was implemented that generates correction candidates for single words based on several simple word lists, some frequency lists and a linear scoring system. The correction process, as illustrated in Figure 4.1, has two phases, and it can be summarized as follows.

At the beginning of the correction process, word forms that are contained in a list of stopwords and abbreviations are identified. For these words, no suggestions are generated. For the rest of the words, the correction suggestion algorithm is applied. For each word, a list of suggestion candidates was generated that contains word forms within one edit distance (Levenshtein, 1965) from the original form. Table 4.1 summarizes the possible cases of one

DOI:10.15774/PPKE.ITK.2015.012

Edit type	Input word	Output word
insertion of a single character	presure	pressure
deletion of a single character	inculude	include
substitution of a character with another one	syght	sight

Table 4.1: Possible single-character edits

Figure 4.1: The word-based system (w's stand for words, a's for abbreviations, c's are correction candidates and (c, s)'s are correction candidate, score pairs. Misspelled words are signed with an asterisk.)

edit distance variations. The possible suggestions generated by a wide-coverage Hungarian morphological analyzer (Prószéky and Kis, 1999; Novák, 2003) are also added to this list.

In the second phase, these candidates are ranked using a scoring method. Thus a ranked list of correction candidates is generated to all words in the text (except for the abbreviations and stopwords). However, only those are considered to be relevant, where the score of the first ranked suggestion is higher than that of the original word (w_2 and w_4 in the example shown in Figure 4.1).

First, the word lists (and the resources these are built from), then the scoring method is described in the following subsections.

4.1.1.1 Word lists

Several models were built on the original data set and on external resources. Some of these models are simple word lists, while others also contain frequency information. These models are listed below. The first two of them (the stopword list and the abbreviation list) are used as prefilters before suggesting corrections, the rest were used to generate the suggestions.

- **stopword list** (sw list): a general stopword list for Hungarian (containing articles, prepositions, function words, etc.) was extended with the most frequent content words present in our medical corpus. After creating a frequency list, these items were manually selected from the words occurring more times than a predefined threshold.

DOI:10.15774/PPKE.ITK.2015.012

- **abbreviation list** (ABBR LIST): after automatically selecting possible abbreviations in the corpus Siklósi et al. (2014), the generated list was manually filtered to include the most frequent abbreviations.
- **list of word forms licensed by morphology** (LICENSED WORDLIST): word forms that are accepted by the Hungarian morphological analyzer were selected from the original corpus, creating a list of potentially correct word forms. To be able to handle different forms of medical expressions, the morphology was extended with names of medicines and active ingredients [i], the content of the Orthographic Dictionary of Hungarian Medical Language Fábián and Magasi (1992) and the most frequent words from the corpus. A unigram model was built from these accepted word forms including the number of occurrences of each word in the corpus.
- **list of word forms not licensed by morphology** (NON-LICENSED WORDLIST): the frequency distribution of these words were taken into consideration in two ways when generating suggestions. Those appearing just a few times in the corpus were classified as forms not to be accepted (transforming their frequency value to 1 - original frequency). The ones, however, whose frequency was higher than the predefined threshold, were considered to be valid forms, even though they were not accepted by the morphology. Actually, it is possible that a word is misspelled the same way several times resulting in an erroneous form. However, this is less probable than that word form being correct in spite of not being licensed by our morphology.
- **general and domain-specific corpora** (SZEGED KORPUSZ and ICD LIST): unigram models were built, similar to that of the above-described licensed word forms, from the Hungarian Szeged Korpusz (Csendes et al., 2004) and from the descriptions of the entities in the ICD code system documentation. I assumed that both corpora contained only correct word forms.

4.1.1.2 Scoring method

Having a list of correction candidates, a score based on (1) the weighted linear combination of scores assigned by several different frequency lists, (2) the weight coming from a confusion matrix of single-edit-distance corrections, (3) the features of the original word form, and (4) the judgement of the morphological analyzer was derived for each suggestion. Some examples with the first five top-ranked correction candidates and their scores is shown in Table 4.2. The system is parametrized to assign much weight to frequency data coming from the domain-specific corpus, which ensures not coercing medical terminology into word forms frequent in general out-of-domain text. The weights for each component were tuned to achieve the best results on the development set, based on metrics described in the evaluation section of this paper, in accordance with the following theoretical considerations:

- **domain-specific models**: two lists of words were generated from the clinical corpus, separating morphologically justified words from unknown forms. Since these models are the most representative for the given corpus, these were taken with the highest weight.

[i] `http://www.ogyi.hu/listak/`, retrieved in October, 2011.

DOI:10.15774/PPKE.ITK.2015.012

implatatumot	telefonnegbeszélés	Meibm	mirgy	kupakszeráűen	túvel
implantatumot	telefonmegbeszélés	meibom	mirigy	kupakszerűen	tűvel
5.60144e-05	5.87158e-05	0.0001056	9.03702e-05	5.87158e-05	5.88118e-05
implatatumot	telefonnegbeszélés	meibm	miragy	kupakszervűen	tevel
5.33130e-05	5.33130e-05	5.06116e-05	5.87158e-05	5.87158e-05	5.87158e-05
ímplatatumot	telefónnegbeszélés	meíbm	mirgy	kupakszeráűen	tővel
1.875e-05	1.875e-05	1.875e-05	5.06116e-05	5.06116e-05	5.87158e-05
implatatumót	telefonnegbeszéléz	meybm	mírgy	kúpakszeráűen	túvel
1.875e-05	1.40625e-05	1.40625e-05	1.875e-05	1.875e-05	5.06116e-05
implatatúmot	telefonnegbessélés	meilbm	myrgy	kupakszeráüen	tuvel
1.875e-05	1.40625e-05	4.6875e-06	1.40625e-05	1.875e-05	1.875e-05

Table 4.2: Ranked suggestion lists for some misspelled words. The numbers are the scores given by the system to each correction candidate.

- **models built from external resources**: these models are larger, but they are more general, thus word forms are not that relevant for medical texts. The results reflect that though these models contribute to the quality of the corrections, they must have relatively low weights in order to keep the scores of medical words higher.
- **original form**: the original form of the words received two kinds of weighting. First, if the original word was licensed by the morphology, then it also received a certain extra weight. Second, a weight was given to the original word form in the suggestion list, regardless of its correctness. This second weight type was introduced so that the system would not "correct" an incorrect word form to another incorrect form, but rather keep the original one if no real suggestions can be provided.
- **morphological judgment on suggestions**: each generated suggestion licensed by the morphology received a higher weight to ensure that the final suggestions are valid words.
- **weighted Levenshtein generation**: when generating word forms that are one Levenshtein edit distance far from the original one, special weighting was given for more probable phenomena, such as swapping neighbouring letters on the keyboard (e.g.: *n-m*, *s-d*), improper use of long and short forms of Hungarian vowels (e.g.: *o-ó*, *u-ú*, *ö-ő*), or mixing characteristic letters of Latin (e.g.: *t-c*, *y-i* as for example in the word *dysfunctio*, which is frequently written as *disfunctio*). While the cost of an edit operation was 1, in general, an *y-z* swap and neighbouring keys swaps were counted with a cost coefficient of 0.2, the Latin-Hungarian swaps with a cost of 0.6 and a long versus short vowel swap with a cost of 0.8.

4.1.2 Application of statistical machine translation

When generating correction suggestions, the word-based system ignores the lexical context of the words to be corrected. Since my goal is to perform correction fully automatically, rather than offering the user a set of corrections that they can choose from, the system should be able to select the most appropriate candidate. In order to achieve this goal, the ranking of the word-based system based on morphology and word frequency data is not enough.

DOI:10.15774/PPKE.ITK.2015.012

To improve the accuracy of the system, lexical context also needs to be considered. To satisfy these two requirements, I applied Moses (Koehn et al., 2007), a widely used statistical machine translation (SMT) toolkit. During "translation", I consider the original erroneous text as the source language, while the target is its corrected, normalized version. In this case, the input of the system is the erroneous sentence: $E = e_1, e_2 \dots e_k$, and the corresponding correct sentence $C = c_1, c_2 \dots c_l$ is the expected output. Applying the noisy-channel model terminology to my spelling correction system: the original message is the correct sentence and the noisy signal received at the end of the channel data is the corresponding sentence containing spelling errors. The output of the system trying to decode the noisy signal is the sentence $\hat{C}$, where the $P(C|E)$ conditional probability takes its maximal value according to Formula (1).

$$\hat{C} = argmax P(C|E) = argmax \frac{P(E|C)P(C)}{P(E)} \tag{4.1}$$

Since P(E) is constant, the denominator can be ignored, thus the product in the numerator can be derived from the statistical translation model ($P(E|C)$) and the target-language model ($P(C)$).

These models in a traditional SMT task are built from a parallel corpus of the source and target languages based on the probabilities of phrases corresponding to each other. In my case, however, such a parallel corpus of erroneous and corrected medical texts does not exist, thus the training step was replaced by the word-based system, where correction candidates were included into the translation model. The language model responsible for checking how well each candidate generated by the translation model fits the actual context is built using the SRI Language Modeling Toolkit (SRILM) (Stolcke et al., 2011). Figure 4.2 shows the process of correcting documents by the context-aware system.

4.1.2.1 Translation models

Three translation (correction) models were applied according to three categories of words and errors. The first one handles general words, the second one is applied to possible abbreviations and the third one can split erroneously joined words. In the following subsections, I describe each of these models.

Translation model for errors in general words The translation model is based on the output of the word-based system. For each word, except for abbreviations and stopwords, the first 20 suggestions were considered. Taking more than 20 candidates would have caused noise rather than increasing the quality of the system. The scores used for ranking these suggestions in the word-based system are normalized as a quasi-probability distribution, so that the probabilities of all possible corrections for a word would sum up to 1. This method was applied instead of learning these probabilities from a parallel corpus. It should be noted that though suggestions are generated for each word, these suggestions usually include the original form (if its score in the word-based ranking was high enough). The scoring ensures

DOI:10.15774/PPKE.ITK.2015.012

Figure 4.2: The context-aware SMT-based system

that if the original form was correct, then it will receive a higher score, thus the decoder will not modify the word.

Table 4.3 contains a common word that is misspelled in the input text. The word *hosszúságu* should be written as *hosszúságú* 'of length ...'. Another word form, *hosszúsági* 'longitudinal' is ranked higher by the original context-insensitive scoring algorithm, because it is also a correct and more frequent Hungarian word. Furthermore, the *u:i* correspondence is also a frequent error beside *u:ú*, since *u* and *i* are neighboring letters on the keyboard. Though the rest of the words in the example are also correct candidates, they received a lower score, since either the resulting word form is not that typical to the domain, or the type of the mistake that would have caused the actual misspelling is less probable. Thus, without considering the context, all the others would also be correct at the word level. The language model will be responsible for making the contextually optimal choice.

Translation model for abbreviations Clinical documents contain much more abbreviations than general texts (see Section 2.3). Applying the models above to abbreviations is difficult due to two main reasons. On the one hand, the same word or phrase usually appears in several different abbreviated forms in the text according to the individual custom of the author or just due to accidental variation. On the other hand, most abbreviations are very

DOI:10.15774/PPKE.ITK.2015.012

original form (e)	correction candidate (c)	$P(c\|e)$
hosszúságu	hosszúsági	0.01649
hosszúságu	hosszúságú	0.01560
hosszúságu	hosszúsága	0.01353
hosszúságu	hosszúságuk	0.01317
hosszúságu	hosszúságul	0.01292
hosszúságu	hosszúságé	0.01284
hosszúságu	hosszúság	0.01034

Table 4.3: A fragment of the translation model for a misspelled common word, its possible candidate corrections and their probabilities.

original form (e)	correction candidate (c)	$P(c\|e)$
soronkívül	soron kívül	0.02074
soronkívül	soronkívül	0.01459

Table 4.4: Extract from the translation model for multiword errors

short, and, in most cases, the suggestion generator would prefer to transform the original abbreviation to a very frequent similar common word. Due to their high frequency and the fact that the morphology would also affirm their correctness, such "corrections" would practically ruin the semantics of the original text.

Handling joining errors Since the Moses SMT toolkit is usually used as a phrase-based translation tool in traditional translation tasks, a general feature of the translation models is that the translation of one (or more) words can also be more than one word. Thus the system can be used to generate multi-word suggestions for a single word in a straightforward manner. This way my system can split erroneously joined words. Probability estimates for these phrases are also derived from the scores assigned by the suggestion generation system. When inserting a space into a word, the models used for creating the ranking scores are calculated for both words separately and the geometric mean of these values is assigned to the phrase as a score. This final score then corresponds to the scale of the rest of the single word suggestions. An example for correction candidates for erroneously joined words is shown in Table 4.4. Since the correction process is carried out word-by-word, the method for joining two erroneously split words is not implemented (though theoretically available in the system), but the occurrence of such errors is (about six times) less frequent than the other way round.

4.1.2.2 Language model

The language model is responsible for taking the lexical context of the words into account. In order to have a proper language model, it should be built on a correct, domain-specific corpus by acquiring the required word n-grams and the corresponding probabilities. Since the only manually corrected portions of the corpus were the development and test sets, such a model could not be built. Though there are orthographically correct texts of other, mostly

DOI:10.15774/PPKE.ITK.2015.012

general domains, the n-gram statistics of these would not correspond to the characteristics of the clinical domain due to the differences described in Chapter 2. That is why such texts were not used to build the language model. Nevertheless, the results of some experiments performed by using general texts to build the language model are also described in the evaluation section of this paper.

I assumed that the frequency of correct occurrences of a certain word sequence can be expected to be higher than that of the same sequence containing a misspelled word. Of course, the development and test sets used for evaluation were separated from the corpus prior to building the language model. Otherwise, the word sequences would have corresponded to these, and no correction would have been made.

The documents in the corpus were split into sentences at hypothesized sentence boundaries along with applying tokenization as a preprocessing step using the system of Orosz et al. (2013). However, finding sentence boundaries was often quite challenging in our corpus. The average length of these quasi-sentences is 9.7 tokens. Thus a 3-gram language model was used, because in these relatively short sentences longer mappings cannot be expected. The measurements also confirmed this: choosing a higher-order language model resulted in worse accuracy. This is not only caused by the shortness of the sentences, but also by the nature of the corpus used for building the language model. Since no correct corpus from this domain exists in Hungarian and manually correcting a large enough portion of the documents would have been a very time consuming task, I used the original, noisy texts for creating the language model, which was still an unusually small corpus for this purpose. Thus, the variety of longer n-grams was lower and if these contained some errors, these could have matched the sequence of words in the input text, leaving it uncorrected.

4.1.2.3 Decoding

The result of Formula 4.1 is determined by the decoding algorithm of the SMT system based on the above models. To carry out decoding, I used the widely-used Moses toolkit (Koehn et al., 2007). The parameters of decoding can be changed easily in order to adapt the system to new circumstances and weighting schemes, because they can be set in a simple configuration file. The list of these parameters and their adaptation procedure are detailed below. During decoding, each input sentence is corrected by creating the translation models sentence by sentence. These models are based on the suggestions generated for the words occurring in the actual sentence, and on the pre-built abbreviation translation model. The parameters for decoding were set as follows:

- **Weights of the translation models**: since the contents of the phrase tables do not overlap, their weights could be set independently. As mentioned earlier, the correction of the texts was meant as a normalization process rather than adjusting them to a strict orthographic standard. In the case of correcting abbreviations, the goal was to choose the same abbreviated form for each concept appearing in different forms in the original text. To guarantee a high probability for these normalized forms, the abbreviation translation model was given a higher weight.

DOI:10.15774/PPKE.ITK.2015.012

- **Language model**: a 3-gram language model was applied, which was given a lower weight than the translation models in order to prevent the harmful effect of the possibly erroneous n-grams due to the incorrect word forms in the corpus that were used for building this model.
- **Reordering constraint**: when translating between different languages in a traditional translation task, the reordering of some words within a sentence might be necessary. However, my system is designed to correct spelling errors within words only (including the possibility of splitting words), not grammatical errors at the sentence level. Thus, word order changes are not allowed, the structure of the sentence cannot be changed. That is why monotone decoding was applied.
- **Penalty for difference in the length of sentences before and after correction**: since the length of a sentence measured in number of tokens cannot change significantly during correction, there is no need to apply a penalty factor of the decoder for this parameter. (The theoretical maximum in the change of the length for a sentence is doubling it by inserting a space to each and every word, but the necessary number of space insertions was at most two per sentence in the test set.)

4.1.3 Data sets

We were provided with a set of anonymized clinical documents from various departments of a Hungarian hospital. Though the departments belong to the same institution, both the structure and the use of the language is unique to each department. Therefore my system is trained to be applicable only to one of these, namely ophthalmology. The ophthalmology portion of the corpus consists of 50,394 tokens, out of which a set of documents were separated for testing purposes. The size of the test set was 3722 tokens. The test set contained 89 different misspelled words, which were manually corrected providing the gold standard for the evaluation.

Since the word-based system, which is also used for generating correction candidates in the SMT system, had several weighting parameters to be tuned, a development set was also necessary in order to avoid overtraining. For this, 2000 sentences (17243 tokens / 6234 types) were randomly selected from the whole clinical corpus (from various departments). Having such a mixed development set was necessary for two reasons. First, the ophthalmology portion was too small to be used by itself. Second, the suggestion generation system uses some frequency lists built from general corpora. In order to set the weights of these frequency lists properly, a big and more general development set was more appropriate.

The remaining part of the corpus was used for building the domain-specific frequency lists for the word-based system and the language models for the SMT system. In the latter case, two separate language models were built: one created from texts of various departments and one containing only the ophthalmology portion of the corpus.

All sets of sentences contained only free-text parts of clinical reports. Tabular laboratory data, measurement results, headers, ICD codes and other structured content had previously been filtered out. In spite of this prefiltering, there was still a high number of sentences both

DOI:10.15774/PPKE.ITK.2015.012

	Error detection			Error correction
	Precision	Recall	F-measure	Accuracy
WB	38.46%	60.24%	41.45%	78.00%
SMT-MEDLM	69.11%	56.62%	66.19%	87.23%
SMT-GENLM	24.88%	63.85%	28.34%	77.35%

Table 4.5: Performance of the two systems (the context-unaware word-based (WB), the context-aware SMT with language model from the medical domain (SMT-MEDLM) and with a general language model (SMT-GENLM)) on the test set

in the training and test sets that hardly contained real words. These sentences consist of sequences of abbreviations and numbers while having a clearly Hungarian syntax.

4.2 Results

Two aspects of the performance of the systems were evaluated: (i) error detection and (ii) error correction. Error detection was measured by the common metric of F-measure, according to Formula 4.2.

$$F_\beta = \frac{(1+\beta^2)\cdot true\,positive}{(1+\beta^2)\cdot true\,positive + \beta^2 \cdot false\,negative + false\,positive} \tag{4.2}$$

The parameter β was chosen to be 0.5 in order to prefer precision over recall. This way, the system that recognized erroneous and only erroneous words as misspelled was preferred to a hypothetical setup that would overwrite all words, even the correct ones. Since error detection and error correction were not done separately, a word was considered to be classified as erroneous if the system made a change in it during the correction process. In this case the correctness of the change did not matter. The quality of error correction was measured by accuracy, i.e. the ratio of the number of properly corrected words, over the number of all words changed. Table 4.5 shows the numerical results for both systems.

In the case of the word-based system, the performance was evaluated in a setup that simply replaced each word with the top-ranked suggestion that the suggestion generation system generated. Thus, the correction was done at word level, correcting ambiguous mistakes always to the same form, no matter what their context was. This method performed quite well in the case of long words containing one misspelling. However, in the case of short words, it was seldom able to rank the actually correct form first. Regarding the ranking of the suggestions, in 99.12% of the development set, the 5 best suggestions contained the real correction. Moreover, the precision of this system in the case of error detection was quite low (38.46%), which was due to the high number of false positives, i.e. correct words classified as erroneous. Even though the system could achvieve a higher value for recall (60.24%), the F-measure emphasizing precision was still 41.45% for error detection. Considering the task

DOI:10.15774/PPKE.ITK.2015.012

of error correction, 78.00% of the erroneous words were changed to the correct form in this case.

On the other hand, the SMT based system though detected slightly less number of errors (i.e. 56.62% recall), the precision of this system was significantly higher, 69.11% of its alerts were real errors. Thus, F-measure was also near to this value, i.e. 66.19%. The accuracy of this system when correcting erroneous words was 87.23%. These results prove the beneficial effect of contextual information in the process of correcting errors.

Since the language model built from the clinical domain was small and noisy, an experiment was also performed using another language model built from a general corpus of Hungarian (for details about this corpus see Chapter 2). Even though the resulting language model was larger and contained less spelling errors than the medical one, the performance of the error correction system was significantly worse with this setting. The number of false positives was 8 times higher than in the previous case, causing the precision of the system for error detection to fall down to 24.88%, which is even worse than that of the word-based system. Using a general language model forces the system to overwrite correctly used domain-specific word forms to similar words frequent in the general corpus. This result is in accordance with the observations in Chapter 2 concerning the huge differences between the language of our corpus of medical records and general Hungarian.

The comparison of the word-based and the SMT systems based on the automatic performance measures affirms the improvement due to taking lexical context into account in the correction process. However, investigating the actual corrections manually reveals even more sophisticated differences not reflected by the numerical results. Table 5.5 shows some originally erroneous sentences with their automatic corrections generated by each system and the reference correction as well. The examples are chosen so that they contain different types of sentences occurring in the corpus (i.e. the first one is a real sentence, the second contains hardly any real words, the third contains punctuation and encoding errors, and the last one is a mixture of Latin and Hungarian). As presented in these examples, there are some words properly corrected, some others are altered to an improper form, while others are left in their original misspelled form. Moreover, the behaviour of the word-based and the SMT system differs regarding these phenomena.

It should be noted that, in some cases, I had to accept some non-standard forms that were consistently used throughout the whole corpus, without the standard form appearing at all. I believe that the retrieval of concepts in the texts and their normalization do not require that the normalized version of each word be the orthographically standard form, but mapping variants to a single representation is sufficient.

4.2.1 Shortcomings of both systems

When performing manual evaluation, it was found that even though there are several cases where none of the correction systems is able to find the correct form of a word, the SMT-based context-aware system created words that are much “closer” to the real correction than the ones selected by the word-based system. Even in cases when an originally correct word is

DOI:10.15774/PPKE.ITK.2015.012

ORIG	csppent előírés szerint ,
WB	cseppent előír és szerint ,
SMT	cseppent előírás szerint ,
REF	cseppent előírás szerint ,
ORIG	th : mko tovább 1 x duotrav 3 ü-1 rec , íb : 2 x azoipt 3 ü-1 rec
WB	th : mko tovább 1 x duotrav 3 ü-1 sec , kb : 2 x azoipt 3 ü-1 sec
SMT	th. : mko tovább 1 x duotrav 3 ü-1 rec , kb : 2 x azopt 3 ü-1 rec
REF	th. : mko. tovább 1 x duotrav 3 ü-1 rec , kb. : 2 x azopt 3 ü-1 rec
ORIG	/alsó m?fogsor .
WB	/alsó műfogsor .
SMT	alsó műfogsor .
REF	alsó műfogsor .
ORIG	vértelt nyállkahártyák , kp erezett conjuctiva , fehér sclera .
WB	vértelt nyálkahártyák , kp erezett conjunctiva , fehér sclera .
SMT	vértelt nyálkahártyák , kp. erezett conjunctiva , fehér sclera .
REF	vértelt nyálkahártyák , kp. erezett conjunctiva , fehér sclera .

Table 4.6: Originally erroneous sentences (ORIG) with the automatic correction of the context-unaware word-based (WB) and the SMT systems and the manually corrected reference (REF)

ORIG	homályos látást panaszol .	*(s/he complains about blurred vision)*
SMT	homályos látás panaszok .	*(complaints of blurred vision)*
ORIG	panasz nem volt .	*(there were no complaints)*
SMT	panasza nem volt .	*(s/he didn't have any complaints)*

Table 4.7: Examples for the transformation of a correct sentence (ORIG) to another correct sentence with very similar meaning, but different words (SMT)

modified, the SMT system results in a word that is appropriate in the given context. On the contrary, the word-based system usually replaces these words with some meaningless strings. Such instances are usually real-word errors, when an originally correct word form is transformed to another correct word. Another case is, when the original form is not correct, and it might be corrected to a word that is correct and grammatically appropriate in the sentence. Nevertheless, it is still not the actually expected correction (Table 4.7). These incorrect solutions mainly originate from the language model, built from the clinical corpus itself, that also contains some improper n-grams.

On the other hand, errors not handled at other levels of processing are also present and could not be corrected as spelling errors. Such problems arise from incorrect tokenization or the inconsistent usage of measurement results. For example the phrases *07.23.án* or *2010.08.-hó* were given to the spelling correction system as single tokens, but the gold standard correction of these are *07. 23-án* and *2010. 08. hó* respectively, thus the spelling error could have been corrected only if the tokenization had been correct. So is the case with measurement results, for example *0,15?-1,0d*. For such units there is no standard tokenization scheme.

DOI:10.15774/PPKE.ITK.2015.012

Original form	Correction	English translation
dúrva	durva	rough
feltünnek	feltűnnek	they appear
tizta	tiszta	clean
felszinéhez	felszínéhez	to its surface
tágitás	tágítás	expansion
konzilium	konzílium	consultation
presens	praesens	present (in Latin)
felirva	felírva	prescribed

Table 4.8: Some examples for words corrected properly by both systems

4.2.2 Errors corrected by both systems properly

As mentioned earlier, both the word-based and the context-aware SMT systems were able to correct those words properly that were either long or frequent words within the corpus or in the general word lists. Such words are shown in Table 4.8 showing the original form, the proper correction (achieved by both systems) and their English translation. For such words, the two implementations did not show too much difference.

4.2.3 Errors corrected by one of the systems

As opposed to common, full word forms, in the case of shorter terms or abbreviations and domain-specific words, the behaviour of the two systems were different, especially in the task of error detection. The word-based system tended to change these words to some other forms incorrectly, while the SMT system either left them in their original form if they had been correct already, or corrected them to the proper form. Some examples are listed in Table 4.9.

Since abbreviations and shortened forms can be disambiguated only in their context (Siklósi and Novák, 2013), their correction also requires contextual information ensured by the domain-specific language model. Similarly, the proper correction of special medical terms and Latin expressions is impossible without contextual information. The word-based system was either not able to suggest a correction ranked higher than the original form, or changed such words to some common terms that gained a higher score due to their high frequency in texts from other domains. In some cases even originally correct words were overwritten. For example, the last line in Table 4.9 shows the results for the word *vannas*. This is a special type of scissors used in surgery, called “vannas scissors”. The word-based system altered this word to *vannak* ’are’, which is a very common Hungarian word.

DOI:10.15774/PPKE.ITK.2015.012

Original form	Word-based	SMT	Gold st.	English translation
szemhéjszélt	szemhéjszéli	szemhéjszélt	szemhéjszélt	'side of eyelid'
tu	tó	tu.	tu.	short form of 'tumor'
inf	in	inf.	inf.	short form of 'inferioris'
elasticum	elasticus	elasticum	elasticum	'elasticum'
ell	el	ell.	ell.	short form of 'check'
cover	over	cover	cover	'cover' (a medical test)
skia	skin	skia	skia	'skia' (a medical test)
deg	meg	deg.	deg.	short form of 'degenerate'
jav	ja	jav.	jav.	short form of 'correct'
dec	de	dec.	dec.	short form of 'December'
tonopen	tonogen	Tonopen	Tonopen	'Tonopen' (a medication)
ill	áll	ill.	ill.	short form of 'or'
amb	ab	amb.	amb.	short form of 'ambulatory'
vannas	vannak	vannas	vannas	'vannas' (a medical tool)

Table 4.9: Some examples for words corrected properly or untouched by the SMT system, but altered incorrectly by the word-based system

DOI:10.15774/PPKE.ITK.2015.012

DOI:10.15774/PPKE.ITK.2015.012

5

Identification and resolution of abbreviations

In which it is shown how much extra letters we use in our every-day language. Doctors need less. Still, in order to be able to understand their message, some character refilling methods are described with special emphasis on the opththalmology domain. To make it more exciting, try to figure out what 'Cat. incip. o. utr.' stands for. By the end of this Chapter, this will be revealed.

Contents

DOI:10.15774/PPKE.ITK.2015.012

The task of abbreviation resolution is often treated as word sense disambiguation (WSD) (Navigli, 2012). The best-performing approaches of WSD use supervised machine learning techniques. In the case of less-resourced languages, however, neither manually annotated data, nor an inventory of possible senses of abbreviations are available, which are prerequisites of supervised algorithms (Nasiruddin, 2013). On the other hand, unsupervised WSD methods are composed of two phases: word sense induction (WSI) must precede the disambiguation process. Possible senses for words or abbreviations can be induced from a corpus based on contextual features. However, such methods require large corpora to work properly, especially if the ratio of ambiguous terms and abbreviations is as high as in the case of clinical texts. Due to confidentiality issues and quality problems, this approach is not promising either.

In this Chapter, I introduce the behaviour of abbreviations in Hungarian clinical documents. Then, a corpus-based approach is described for the resolution of abbreviations with using the very few lexical resources available in Hungarian. As this method did not provide acceptable results, the construction of a domain-specific lexicon was unavoidable. Instead of trying to create huge resources covering the whole field of medical expressions, it is shown that small domain-specific lexicons are satisfactory and the abbreviations to be included can be derived from the corpus itself. Finally, an analysis of the combination of these methods is presented.

5.1 Clinical abbreviations

The use of a kind of notational text is very common in clinical documents. This dense form of documentation contains a high ratio of standard or arbitrary abbreviations and symbols, some of which may be specific to a special domain or even to a doctor or administrator. These shortened forms might refer to clinically relevant concepts or to some common phrases that are very frequent in the specific domain. For the clinicians, the meaning of most of these common phrases is as trivial as the standard shortened forms of clinical concepts due to their expertise and familiarity with the context. Some examples for abbreviations falling into these categories are shown in Table 5.1.

5.1.1 Series of abbreviations

Even though standalone abbreviated tokens are highly ambiguous, they more frequently occur as members of multiword abbreviated phrases, in which they are usually easier to interpret unambiguously. For example *o.* could stand for any word either in Hungarian or in Latin, starting with the letter *o*, even if limited to the medical domain. However, in our corpus of anonymized ophthalmology reports, *o.* is barely used by itself, but together with a laterality indicator, i.e. in forms such as *o. s., o. d.*, or *o. u.* meaning *oculus sinister* ‘left eye’, *oculus dexter* ‘right eye’, or *oculi utriusque* ‘both eyes’, respectively. In such contexts, the meaning of the abbreviated *o.* is unambiguous. It should be noted, that these are

DOI:10.15774/PPKE.ITK.2015.012

Domain	Abbr.	Resolution	in Hungarian	in English
standard	o. d.	oculus dexter	jobb szem	right eye
	med. gr.	mediocris gradus	közepes fokú	medium grade
domain-specific	o.	oculus	szem	eye
	o.	os	csont	bone
domain-specific common	sü	saját szemüveg	saját szemüveg	own glasses
	fén	fényérzés nélkül	fényérzés nélkül	no sense of light
	n	normál	normál	normal
common words	köv	következő	következő	next
	lsd	lásd	lásd	see

Table 5.1: Some examples for the use of simple abbreviations. Some of them are commonly known standard forms, usually of Latin origin, some others, though related to the clinical domain, might have several meanings depending on the specific sub-domain. The rest are abbreviated common words, usually of Hungarian origin, and might also refer to both clinical phrases or common words.

not the only representations for these abbreviated phrases. For example, *oculus sinister* is also abbreviated as *o. sin., os, OS*, etc. Table 5.2 shows the ratio of unique abbreviation sequences of different lengths detected automatically in the corpus with the method described in Section 5.3.1 The number of different single-token abbreviations is roughly equal to the number of all multi-token abbreviations.

Length:	1	2	3	4	5	>5
Number:	49.53%	26.34%	15.00%	5.95%	2.16%	0.98%

Table 5.2: The ratio of unique abbreviation series of different lengths detected automatically in the corpus.

Thus, when performing the resolution of abbreviations, I considered series of such shortened forms instead of single tokens. A series is defined as a continuous sequence of shortened forms without any unabbreviated word breaking the sequence. These series are not necessarily coherent phrases. The individual elements of such sequences of abbreviations are by themselves highly ambiguous, and even if there were an inventory of Hungarian medical abbreviations, which does not exist, their resolution could not be solved. Moreover, the mixed use of Hungarian and Latin phrases results in abbreviated forms of words in both languages, thus the detection of the language of the abbreviation is another problem. For example, in the "sentence"

Dg : Tu. pp. inf et orbitae l. dex. , Cataracta incip. o. utr. , Hypertonia

the abbreviation spans are the following:

Dg | Tu. pp. inf | l. dex. | incip. o. utr..

DOI:10.15774/PPKE.ITK.2015.012

5.1.2 The lexical context of abbreviation sequences

In the above example, the last section is misleading, since the token *incip.* is part of the phrase *Cataracta incip.*, i.e. it is related to its preceding neighbour, which is not included in this list as part of an abbreviation. This mixed use of a phrase is very common in the documents, with a diverse variation in using certain words in their full form or in some shortened form instead. In order to save such phrases and to keep the information relevant for the resolution of multiword abbreviations, the context of a certain length is attached to the detected series. In my experiments, the length of the context taken from both the left and right sides of the abbreviations ranged from 0 to 3 tokens. Since the average length of sentences in the corpus is 9.7 (Orosz et al., 2013), considering a larger context could span across sentences, but that would make no sense.

Beside completing such mixed phrases, the context also plays a role in the process of disambiguation. The meaning (i.e. the resolution) of abbreviations of the same surface form might vary in different contexts. My experimental results showed that this does not require a larger window of sampling either.

5.2 Resources

5.2.1 External lexicon

Even though there are no structured lexical resources for Hungarian, the official coding system for diseases, anatomical structures and medical procedures is available (similar to the ICD systems in English). Thus, a simple dictionary was built from the ophthalmology sections of these descriptions. The final list of phrases contained 3329 entries. However, these phrases are written in the language of the official terminology, which is different in several respects from that used in the clinical texts.

5.2.2 Handmade lexicon

Since the official descriptions turned out not to be of much use, a domain-specific lexicon seemed to be necessary. The first step of designing such a resource is to decide what phrases to include. I assumed that the most frequent abbreviations occurring in the corpus without their expanded form ever being written out have one unambiguous resolution within a narrow domain. For example, in the domain of ophthalmology, the abbreviation *o. d.* always stands for the phrase *oculus dexter*, meaning 'right eye'. Even though it appears in various shortened forms, it is never spelled out in its full form. Thus, a frequency list of abbreviations and abbreviation series was created from the whole corpus. Then, a threshold value was defined experimentally and the abbreviations with a relative frequency above this threshold

DOI:10.15774/PPKE.ITK.2015.012

were included into the set of domain-specific abbreviations. Finally, the resolution of these abbreviations were defined with the help of a medical expert.

This approach can be applied to other domains as well. Thus, my method of abbreviation resolution is applicable to new domains with a relatively small amount of manual effort. In our case of ophthalmology records, rather good coverage can be achieved even with a small lexicon of 44 entries. Adding more items to the list further improves the quality of the resolution, however, the improvement achieved by adding new items diminishes quickly.

5.3 Methods

The primary objective of the research described here is finding spans in a sequence of abbreviations that can be unambiguously resolved together. Since sometimes whole statements or even sentences are written using this kind of heavily abbreviated notation, it is important to find an optimal partitioning of the tokens into meaningful spans. In the previous example, the fragment *incip. o. utr.* should be divided into the spans of *incip.* and *o. utr.*, even if the abbreviation *incip.* is not relevant by itself, but still its meaning is not related to the rest of the abbreviation sequence. However *o. utr.* can be resolved with high confidence.

5.3.1 Detection of abbreviations

The first problem to solve when trying to handle abbreviations within running text is detecting them. Since these texts usually do not follow standard orthographic and punctuation rules, especially in the case of highly abbreviated notational text, the detection of abbreviations cannot be based on patterns formulated according to standard rules of forming abbreviations. The ending periods are usually missing, abbreviations are written with varying case (capitalization) and in varying length. For example the following forms represent the same expression, *vörös visszfény* 'red reflection': *vvf, vvfény, vörösvfény.* I applied some heuristic rules to derive relevant features as indicators of a token being an abbreviation. These features were based on the following characteristics: the presence or absence of a word-final period, the length of the token, the ratio of vowels and consonants within the token, the ratio of upper- and lowercase letters, and the judgment of a Hungarian morphological analyzer, the lexicon of which was expanded with medical terminology (Novák, 2003; Prószéky and Kis, 1999).

Table 5.3 shows the final results for the detection of abbreviations. Most of the errors in the detection arose from misspelled forms, tokenization errors or Latin abbreviations.

DOI:10.15774/PPKE.ITK.2015.012

Precision	Recall	F-measure
95.99%	97.12%	96.55%

Table 5.3: Evaluation results for abbreviation detection

5.3.2 Resolving abbreviations based on external resources

In the first approach, once having a sequence of abbreviations, a maximum coverage resolution suggestion process is carried out. For each possible partitioning of the tokens into non-overlapping spans, a regular expression pattern is generated, which is then matched against lexicons. The patterns are created by general abbreviation rules, such as each letter in the abbreviated form represents the starting letter of each word in the expanded phrase. Or, in the case of multiword abbreviations, each member represents the beginning of each word (not just the first letter) in the interpretation. Some pattern generation rules are presented in Table 5.4.

abbr	regexp	matching expansion	regexp	matching expansion
o. s.	`o[^ ]* s[^ ]*`	oculus sinister		
os	`os[^ ]*`	osteoporosis	`o[^ ]* s[^ ]*`	oculus sinister

Table 5.4: Some of the simplest patterns generated from two short abbreviated phrases. The complexity and variability of these patterns is proportional to the length of the original abbreviation sequence.

The lexicons, in which these patterns are looked up were created from the ophthalmology sections of the descriptions of the official coding system for diseases, anatomical structures and medical procedures, along with a medical dictionary. These resources were used to create a final list of phrases containing 3329 entries. The regular expressions are matched against possible resolution candidates from these lists. In addition to these official descriptions, a small domain-specific lexicon was also created manually with the help of a medical expert. This list contains frequent non-official abbreviations specific to the corpus. The resolution candidates generated for each span are ranked according to the source lexicon they originate from.

During this process, the optimal division of the abbreviation series and its resolution are carried out in one step. This is ensured by the scoring method. The different partitionings are ranked according to three features, optimizing for the longest coverage and best resolution of the abbreviation sequence. The features used for ranking are 1) the number of all tokens in the sequence covered by a resolved form, 2) the size of the longest span covered, 3) the size of the shortest span covered. If the sequence *Exstirp. tu. et reconstr. pp. inf. l. d.*, is partitioned as | *Exstirp. tu.* | *et* | *reconstr. pp. inf.* | *l. d.* |, with having a resolution candidate for each partition except for the word *et*, then the value of the features are 7, 3 and 2, respectively. Finally, these values given for each feature are transformed into a percentage score by a weighted combination of them.

DOI:10.15774/PPKE.ITK.2015.012

5.3.3 Unsupervised, corpus-induced resolution

The main drawback of the previous method is that only official descriptions can be resolved. However, in clinical documents there are freestyle statements written in abbreviated forms that cannot be matched to any phrase in the lexicons. However, due to the limited domain of the corpus, one may assume that such statements, or at least some fragments, are contained within the corpus itself in their expanded form. Thus the second approach also utilizes the aforementioned fragment matching strategy along with all partitioning possibilities. In the first round, however, the generated patterns are not matched against the lexicons, but are sought for in the corpus itself. In this case, single token fragments are not taken into account, which would rather increase the noise in the candidate list. During the generation of this list, the results are pruned using some threshold values based on the length of the covered span and the corpus frequency of the matched result.

Once this corpus-specific resolution is done, the lexicon lookup procedure is carried out, now on the partially resolved sequences, in order to finish the still improper resolutions or to find the missing ones. For example, the simple phrase *o. s.* is never present in the corpus as the fully resolved form of *oculus sinister*, but might be overwritten by the form *o. sin.* due to the higher frequency of that variant of the abbreviation as part of the actual phrase. In either case, the lexicon lookup will find the full resolution. Replacing *o. s.* by *o. sin.* might also have an effect on covering the rest of the sequence that would in some cases not be matched if frequent subexpressions were not normalized first.

The steps of the algorithm are the following:

1. For each possible partitioning of the tokens into non-overlapping spans, a regular expression pattern is generated. The patterns are created by general abbreviation rules, such that each letter in the abbreviated form represents the starting letter of each word in the expanded phrase (assuming that it is an acronym). Or, in the case of multiword abbreviations, each member represents the beginning of each word (not just the first letter) in the interpretation. Some pattern generation rules are presented in Table 5.4.
2. The regular expressions for each span are matched against the corpus resulting in full or partial resolutions.
3. The regular expressions are refreshed based on the results retrieved from the corpus.
4. These expressions are then matched against the lexicons.
5. The results of the spans are concatenated to cover the whole series and each such merged resolution candidates are given a score appropriate for ranking.
6. The highest ranked resolution is considered as the final one.

DOI:10.15774/PPKE.ITK.2015.012

5.4 Results and experiments

A test set of 22 documents was used for evaluation purposes both for the task of abbreviation detection and resolution. The abbreviations in this set were labeled manually and resolved by a medical expert. Finally, the number of tokens labeled as abbreviations was 765. The actual meaning for 56 of them could not be specified. These included initials of doctors; author-specific shortened forms; tokenization errors, etc. Thus, the remaining 709 abbreviations were considered when evaluating my methods for abbreviation resolution.For evaluating the resolution methods on abbreviation series, these were also separated from the test set. 44 unique abbreviation series were considered, which consist of at least 2 tokens and occur at least twice in the test corpus.

Performance was measured in terms of precision, recall and F-measure. For abbreviation detection, precision was calculated as the number of true positives divided by the sum of true positives and false positives and recall was defined as the number of true positives divided by the sum of true positives and false negatives. On the other hand, for the resolution task, precision was defined as the number of correctly resolved tokens divided by the number of all resolved tokens, while recall is the number of correctly resolved tokens divided by the number of all abbreviations (Cohen, 2003). Thus, if having one abbreviation resolved correctly without touching anything else, a precision of 100% could be achieved, which does not reflect the real performance. That is why F_2-measure was defined as the harmonic mean between precision and recall biased towards recall.

5.4.1 Fine-tuning the parameters

The algorithm described above uses three resources where the resolution candidates are searched for: 1) a lexicon containing the ophthalmology section of the official ICD coding system, 2) the handmade lexicon, and 3) the corpus itself. In my experiments, I investigated how the performance of the algorithm is influenced by the availability of these resources to the program. The goal of these experiments were threefold. First, since there is a lack of structured external resources, I wanted to investigate to what extent I could rely on a raw, domain-specific corpus to resolve abbreviations. In order to do this, the size of the corpus in which the regular expressions were matched was changed incrementally. Second, I wanted to check the hypothesis that a small, manually built lexicon (containing the most frequent abbreviations) can be built and utilized and in an effective manner. Moreover, I wanted to identify a threshold for the collected entry candidates to such a lexicon for an arbitrary domain. Third, the best performing combination was evaluated.

DOI:10.15774/PPKE.ITK.2015.012

5.5 Performance on resolving abbreviations

Table 5.5 shows some example sequences with their fragmentation and resolution according to the two methods described above. In the case of abbreviation resolution, the effect of varying four parameters were investigated:

- The first parameter was the size of the context taken into consideration when resolving abbreviation sequences.
- Second, I investigated the effect of changing the size of the corpus used for pattern matching.
- Third, the effect of changing the size of the handmade lexicon and finding the optimal threshold value to decide what to include in it.
- And fourth, the performance of the system using the best combination of these parameters was evaluated.

	Cat. incip. o. utr.
1st method	cat. incip. oculi utriusque
2nd method	cataracta incipiens oculi utriusque
gold standard	cataracta incipiens oculi utriusque
	Myopia c. ast. o. utr.
1st method	myopia kritikus fúziós frekvencia ast. oculi utriusque
2nd method	myopia cum astigmia oculi utriusque
gold standard	myopia cum astigmia oculi utriusque
	myop. maj. gr. o. u.
1st method	myop. maj. gr. oculi utriusque
2nd method	myopia maj grad. oculi utriusque
gold standard	myopia major gradus oculi utriusque
	med. gr. cum
1st method	med. gr. cum
2nd method	med. gr. cum
gold standard	medium gradus cum

Table 5.5: Examples for expanding some abbreviation sequences with each method compared to the manually created gold standard.

Figure 5.1 shows the results of three experimental setups ((a),(b) and (c)). In each of them, the size of the manually created lexicon was kept at a fixed size (0, 44, and 136 entries). The size of the corpus was increased in units of around 3000 sentences in each step and the performance for context sizes 0 to 3 tokens were measured. Replacing each abbreviation with its definition from the lexicon (if it was included in the lexicon) was considered as the baseline. In the first case, without the lexicon, this baseline was an F-measure of 0%, in the second case 60.52%, and in the third case, it was 75.71%. As it can be seen from the graphs, these values are quite below the performance of my final combined system. In each case, considering the tokens without any context always performs worst, however, taking a context larger than one token before and after the abbreviation has a positive effect only if the manually created lexicon is not used. In this case, the system with a context size of three tokens performed best.

DOI:10.15774/PPKE.ITK.2015.012

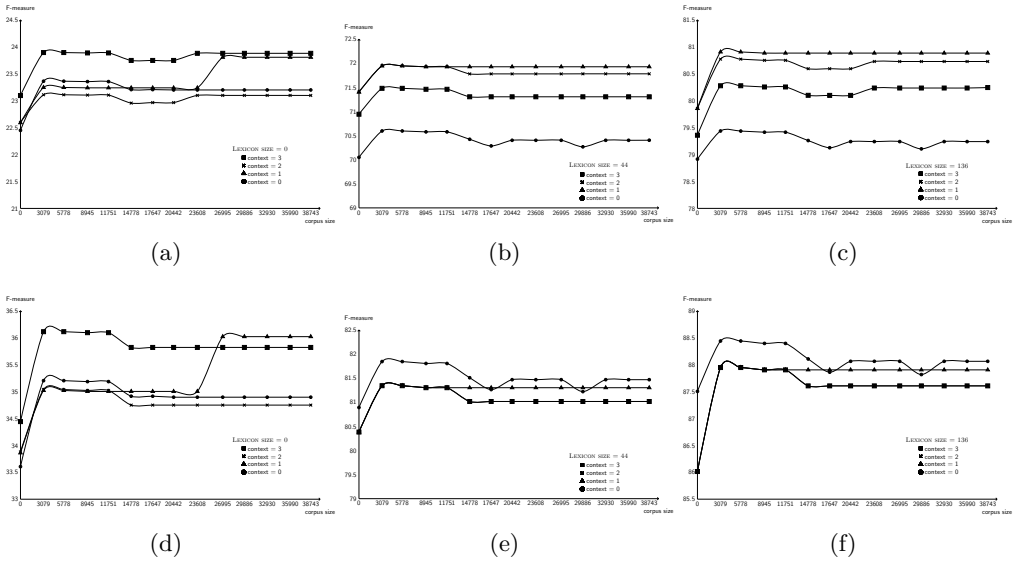


Figure 5.1: The performance results as a function of the corpus size for different context sizes and using a fixed portion of the handmade lexicon (0, 44 and 136 entries respectively). Graphs (a), (b) and (c) represent the results for all abbreviation series, while graphs (d), (e) and (f) represent the results for multi-token abbreviation sequences only.

Increasing the size of the corpus had a similar effect. When the domain-specific lexicon is available, then the only significant change in performance occurred when adding the first portion of the corpus. Further increasing its size did not influence the performance of these setups. This is due to the relatively small size of the corpus and the noisy nature of the texts.

It is clear from the results that relying only on the corpus itself, without any lexical resources results in very poor performance because the most frequent abbreviations are never resolved in the corpus. External lexicons are required to achieve an acceptable performance, but the method of using only external lexical resources performs significantly worse than applying the corpus search as a preprocessing step.

Although performance was adversely affected for both methods when reducing the size of the handmade, domain-specific lexicon, the drop in performance is significantly lower if the corpus is also used for matching, where much of the coverage lost by deleting part of the manually created lexicon is regained dynamically by the system itself. Figure 5.2 shows the learning curve of the second method when gradually increasing the size of the training corpus in the case of resolving abbreviations series of at least two tokens. Point 0 on axis x (i.e. corpus size 0) corresponds to the first method. At this point, the effect of reducing my lexicon is quite significant, however, this difference is made up by learning from the corpus.

Thus I also performed the evaluation tests on abbreviation sequences of length greater than 1 (see graphs (d), (e) and (f) of Figure 5.1). Comparing the behaviour of the algorithm for all and for multiple-token abbreviations, there are two main differences. First, the performance values are higher for longer series of abbreviations. Second, taking a context of any size performs worse than having only the abbreviation by itself in the case of such longer series.

DOI:10.15774/PPKE.ITK.2015.012

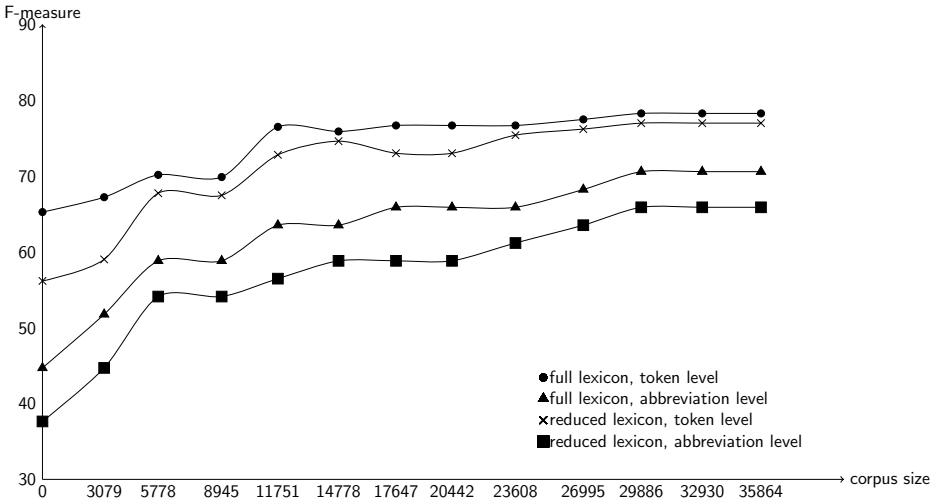


Figure 5.2: The learning curve of each combination as a function of the size (in sentences) of the training corpus.

Moreover, I found that, when using the lexicon, adding too much of the corpus will generate noise for the resolution process instead of enhancing the quality.

In order to find an optimal relative occurrence frequency threshold value for abbreviations that should be included in the handmade lexicon, the largest corpus size was used and the abbreviations were added to the lexicon incrementally. Figure 5.3 shows the change in the threshold and the performance as a function of the number of entries in the lexicon. The results are in accordance with my assumption that including the most frequent abbreviations in the lexicon has a more significant role than creating a large, more detailed lexicon. Adding only the first 10 most frequent abbreviations to the lexicon results in a 30% increase in the performance. Even though the performance grows further if adding more entries, an optimal threshold can be set by cutting the long tail of the graphs. Thus, in our case, this cut was done where abbreviations with relative frequency higher than 0.0025 were added to the lexicon. It resulted in a lexicon size of 44 entries.

Even though the above investigations are important in order to be able to generalize the system to other domains or languages, and filling the gap caused by the lack of structured resources, my goal was also to achieve the best results in resolving abbreviations in Hungarian clinical records in the domain of ophthalmology. The best results compared to those achieved by using the above described threshold are shown in Table 5.6. In the case of the final setup, pattern matching was applied to the whole corpus with taking a one-token context around each abbreviation, and using an enlarged version of the lexicon to 136 entries. (I have no data about further increasing this size.) Thus, an F-measure of 80.88% was achieved for all abbreviations and 88.05% for abbreviation series consisting of multiple tokens.

DOI:10.15774/PPKE.ITK.2015.012

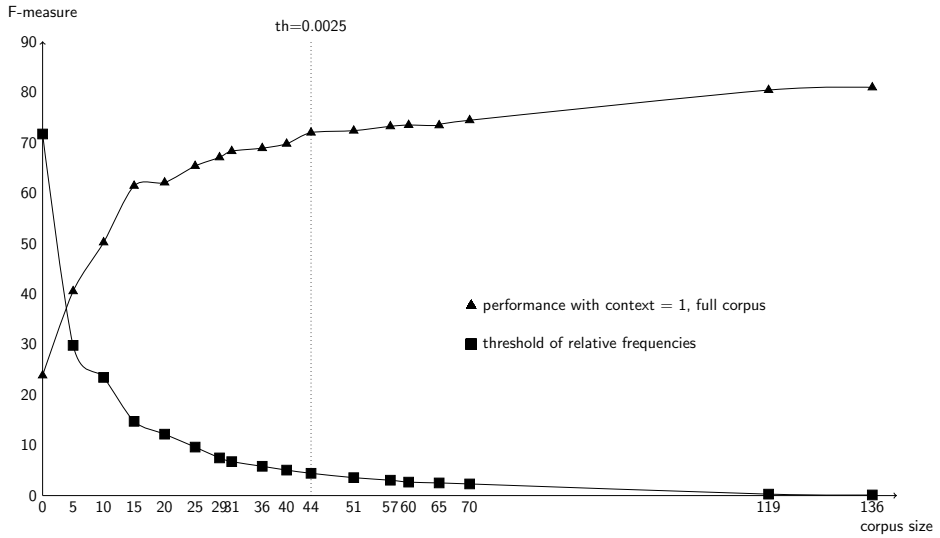


Figure 5.3: The change in the threshold and the performance as a function of the number of entries in the lexicon. Decreasing the threshold (measured in relative corpus frequency) below the value of 0.0025 does not produce a significant increase in the performance relative to the manual effort needed to define the meaning of these abbreviations. The F-measure values here correspond to a context size of one token and the whole corpus is used for pattern matching.

length	lexicon size	precision	recall	f2
all	136	93.23%	78.29%	80.88%
	44	88.37%	68.73%	71.92%
>1	136	96.22%	86.23%	88.05%
	44	90.63%	79.46%	81.46%

Table 5.6: The best performance achieved for the ophthalmology corpus for all abbreviations and for abbreviation series of length greater than 1.

DOI:10.15774/PPKE.ITK.2015.012

6

Identifying and clustering relevant terms in clinical records using unsupervised methods

Believe it or not, these clinical documents still contain meaningful concepts in the form of single or multiple words. Moreover, some of them are more similar to each other than the rest, gathering into groups of semantic units. Just like people, who follow similar customs, these terms behave similarly. Finally, we are going to take a walk in a forest of trees.

Contents

DOI:10.15774/PPKE.ITK.2015.012

In order to be able to create a formal representation of knowledge in the clinical records, a normalized representation of concepts should be defined. This can be done by mapping each record to an external ontology or other semantic resources. However, as Zhang (2002) showed, there is a significant difference between the representation of concepts in such an artifical system and the cognitive behaviour of knowledge. Moreover, in the case of clinical documents, Patel et al. (2002) have shown that the representation of medical concepts by doctors and patients were also different. Thus, it is more reasonable to create a representational schema from the texts themselves rather than enforcing these documents to adjust to a predefined ontology of concepts, which, by the way, does not even exist for Hungarian. Having such an initial system of knowledge extracted from the documents, a domain expert can then justify the results and build a system of concepts that is in accordance with the documents the system is applied to.

In this chapter, statistical methods applied to the corpus are described in order to capture as much information as possible based on the raw data. The aim of this part of my research was to create a transformation of clinical documents into a semi-structured form to aid the construction of hand-made resources and the annotation of clinical texts. In order to achieve a semi-structured representation of raw clinical documents, first normalization procedures were carried out (see the previous Chapters), then two more modules were created that extract multiword terms from the documents and create a distributional model of relevant concepts of the domain. Even though the results of each module are not robust representations of the underlying information, these groups of semi-structured data can be used in the real construction process. The core of each module is based on statistical observations from the corpus itself, augmented by some linguistic rules or resources at just a very few points.

6.1 Extracting multiword terms

In the clinical language (or in any other domain-specific or technical language), there are certain multiword terms that express a single concept. These are important to be recognized, because a disease, a treatment, a part of the body, or other relevant information can be in such a form. Moreover, these terms in the clinical reports could not be covered by a standard lexicon. For example, the word *eye* is a part of the body, but by itself it does not say too much about the actual case in an opthalmology corpus, where most phenomena are related to the eye. Thus, in this domain the terms *left eye, right eye* or *both eyes* are single terms, referring to the exact target of the event the note is about. Moreover, the word *eye* seldom occurs in the corpus without a modifier. This would indicate the need to use some common method for collocation identification.

DOI:10.15774/PPKE.ITK.2015.012

6.1.1 The C-value approach

In my work I used a modified version of the C-value method described by Frantzi et al. (2000). The method discovers multiword terms in raw texts (annotated with part-of-speech tags) and returns a list of terms ranked by their c-value indicating their termhood.

The algorithm applies as follows. First, a list of 1 to n grams are extracted from the corpus, where n can be chosen arbitrary. In my implementation, it was set to 20 (the reason for choosing such a large number will be explained later). Then, a linguistic filter and a stopword filter were applied and the corpus frequency values for each term candidate in the remaining list were determined. Finally, C-value is counted from the longest to the shortest candidates, which will return a number. The higher this number is, the more probable it is that the candidate can be considered as a domain-specific term.

This method exploits statistics derived from the corpus itself, thus does not require external lexical resources. However, the linguistic filter contains some hand-made language-specific rules in order to guarantee that proper terms are extracted.

6.1.1.1 The linguistic filter and the stopword list

After the list of n-grams are produced from the corpus, these are filtered by the linguistic and the stopword filters. The linguistic filter is applied in order to ensure that the resulting list of terms contains only well-formed phrases. Even though, there might be relevant technical terms expressed by verb phrases or even adjectives, the complexity of handling these together with nouns would be intractable. Thus, I dealt with noun phrases only. It should be noted that the goal of this step was not to extract linguistically proper phrases, but only to filter the list of n-grams. This was necessary because the task would have been computationally too complex if all n-grams (up to n=20) would have been used in the calculations and also to separate verbal or nominal phrases. However, the final termhood was not defined by their correspondence to this filtering pattern, but by their c-value, described in the following section. As Frantzi et al. (2000) have shown, the linguistic rules can be set either to be more strict (allowing only nouns) or less strict at the cost of precision over recall. My experiments justified their findings, extending this perspective to the length of the n-grams extracted in the previous step. That is why, n-grams of 1 to 20 were allowed, which lead to more robust statistics when finding the C-value and resulting in more complex terms. However, most of these longer term candidates received a very low C-value, thus they did not remain in the final list of multiword terms.

The base of the linguistic filter was of the following:

$$\{Noun|Adjective|PresentParticiple|Past(passive)Participle\}^{+}Noun$$

This pattern ensures that only noun phrases are extracted and excludes fragments of frequent cooccurrences. This pattern was applied on the part-of-speech tags of words and the

DOI:10.15774/PPKE.ITK.2015.012

lemmatized forms were returned with a few exceptions. These were possessive phrases (e.g. *szürkehályog műtéti megoldása*, 'surgical solution of cataracta') when the possessor was also present (here *szürkehályog*), but if the possessor was not present, then these phrases were lemmatized as well (e.g. *műtéti megoldás*, 'surgical solution'). Participles were also kept in their original inflected form.

The only drawback of this linguistic filter is its language-specific behaviour. As the regular expression describing noun phrases is constructed to apply on Hungarian grammatical structures, it excludes phrases corresponding to Latin constructions. As the morphological analyzer and the pos-tagger had been adapted to handle word forms of Latin origin, these are tagged properly, but do not match the above pattern. For example the pos structure of the phrase 'oculus dexter' is in the form of 'Noun, Adjective', thus it is not extracted as opposed to its Hungarian translation *jobb|Adj szem|Noun*, 'right eye'. The problem is also present when the mixture of the two languages are used.

A stopword list was also applied on the extracted n-grams in order to filter out general phrases. This list was manually and iteratively created by examining the resulting list of terms and selecting words making irrelevant terms in the results. However, the size of the stopword list is also a matter of balance, as explained in Frantzi et al. (2000). The list of stopwords in my implementation contained words such as *megnevezés, kód, menny, diagnózis, beavatkozás, dátum, státusz, k, v, t, h, év, jelen, felvétel, friss, mai, nap, utáni, utóbbi.*

6.1.1.2 Counting C-value

After collecting all n-grams matching the above pattern and passing the stopword filter, the corresponding C-value is calculated for each candidate, which is an indicator of the termhood of a phrase. The C-value is based on four components:

- the frequency of the candidate phrase;
- the frequency of the candidate phrase as a subphrase of a longer one;
- the number of these longer phrases;
- and the length of the candidate phrase.

These components are then combined according to Formula 6.1.

$$C - value(a) = \begin{cases} log_2 |a| \cdot f(a), & \text{if } a \textit{ is not nested} \\ log_2 |a| \cdot (f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b)), & \text{otherwise} \end{cases} \tag{6.1}$$

where

a is the candidate phrase, $f(a)$ is its frequency in the corpus, T_a is the set of longer candidate terms containing a and $P(T_a)$ is the number of these candidate terms.

Thus, the algorithm prefers nested terms occurring independently from longer term candidates, but frequently enough. Let us consider the following examples:

DOI:10.15774/PPKE.ITK.2015.012

Term	English translation	C-value
bal szem	'left eye'	2431.708
ép papilla	'intact papilla'	1172.0
tiszta törőközeg	'clean refractive media'	373.0
békés elülső szegmentum	'calm anterior segment'	160.08
hátsó polus	'posterior pole'	47.5
tompa sérülés	'faint damage'	12.0

Table 6.1: Multiword terms extracted from a document with their corresponding C-value

bal szem látása //vision of the **left eye**	békés **tiszta törőközeg** //calm **clean refractive media**
bal szem sérülése //injury of the **left eye**	egyebekben **tiszta törőközeg** //otherwise **clean refractive media**
bal szem műtéte //operation of the **left eye**	egyebekben **békés elülső szegmentum** //otherwise **calm anterior segment**
bal szem állapota //state of the **left eye**	ép **békés elülső szegmentum** //intact **calm anterior segment**

From the phrases listed in the first coloumn, one can suspect that *bal szem* ('left eye') is a term, because it appears with different words in longer substrings. In the second coloumn *tiszta törőközeg* and *békés elülső szegmentum* are expected terms. The indication is that *bal szem* appears in every term of the first set, and *tiszta törőközeg* or *békés elülső szegmentum* in two-two terms of the second set. We have no such indication for the other substrings, such as *szem látása, szem sérülése, szem műtéte, szem állapota, békés tiszta, egyebekben tiszta, ép békés, egyebekben békés.* Since *bal szem* appears in four longer terms, it can be considered as independent from the longer substrings, and so can be *tiszta törőközeg* and *békés elülső szegmentum.* The substring *szem látása*, however, appears only in one term. The higher the number of longer terms a candidate term appears in, the higher the probability that it is a multiword term. (This is reflected by $P(T_a)$ used in the denominator in the second factor in Equation 6.1.)

The statistics for frequency values are derived from the whole corpus of clinical notes. The details of the algorithm are found in Frantzi et al. (2000).

Table 6.1 shows some multiword terms extracted from a document with their corresponding C-value. Since the linguistic filter allows longer phrases and the n-grams collected at the beginning are also long enough to contain such examples, the resulting list of terms might also contain terms not strictly satisfying criteria of medical terminology. For example the phrase 'tompa sérülés' or the aforementioned 'bal szem' are not real medical terms. However, from the aspect of preprocessing clinical documents these are very useful units and handling them as single terms in further processing steps contribute to a valid representation of information found in the documents.

DOI:10.15774/PPKE.ITK.2015.012

Author(s)	Definition
Harris (1954)	*"Difference of meaning correlates with difference of distribution."*
Firth (1957)	*"You shall know a word by the company it keeps."*
Wittgenstein (1953)	*"Meaning is use."*
Rubenstein and Goodenough (1965)	*"Words which are similar in meaning occur in similar contexts."*
Schütze and Pedersen (1995)	*"Words with similar meanings will occur with similar neighbors if enough text material is available."*
Landauer and Dumais (1997)	*"A representation that captures much of how words are used in natural context will capture much of what we mean by meaning."*
Pantel (2005)	*"Words that occur in the same contexts tend to have similar meanings."*

Table 6.2: Paraphrases of the distributional hypothesis (the list is not complete)

6.2 Distributional semantic models

In order to understand the meaning of the documents, or statements in these documents, a semantic model is needed. It is out of the scope of this work to create a complete model, however the foundations of such a semantic description have been laid.

The construction of hand-made semantic resources for a language is very expensive, requires language and domain-specific expertise and is not always in accordance with the cognitive representation of knowledge (Zhang, 2002). While the domain-specific validation is unavoidable, the other two problems can partially be handled by applying unsupervised methods for ontology learning and recognition of semantic patterns of a sublanguage, such as medical language.

The unsupervised approach to semantics is called distributional semantic models, which captures the meaning of terms based on their distribution in different contexts. As Cohen and Widdows (2009) state, such models are applicable to the medical domain, since the constraints regarding the meaning of words and phrases are more tight than in general language. Pedersen et al. (2007) have shown that in the medical domain distributional methods outperform the similarity measures of ontology-based semantic relatedness.

The theory behind distributional semantics is that semantically similar words tend to occur in similar contexts (Firth, 1957) i. e. the similarity of two concepts is determined by their shared contexts. Table 6.2 shows some paraphrases of the distributional hypothesis.

6.2.1 Distributional relatedness

The context of a word is represented by a set of features, each feature consisting of a relation (r) and the related word (w'). For each word (w) the frequencies of all (w, r, w') triples are

DOI:10.15774/PPKE.ITK.2015.012

determined. In other studies, these relations are usually grammatical relations, however in the case of Hungarian ophthalmology texts, grammatical analysis performs poorly, resulting in a rather noisy model. Carroll et al. (2012), suggest using only the occurrences of surface word forms within a small window around the target word as features. In this research, a mixture of these ideas was used by applying the following relations to determine the features for a certain word:

- prev_1: the previous word
- prev_w: words preceding the target word within a distance of 2 to 4
- next_1: the following word
- next_w: words following the target word within a distance of 2 to 4
- pos: the part-of-speech tag of the actual word
- prev_pos: the part-of-speech tag of the preceding word
- next_pos: the part-of-speech tag of the following word

Words in this context are the lemmatized forms of the original words on both sides of the relations. To create the distributional model of words, a similarity measure needs to be defined over these features. Based on the results of Lin (1998), the similarity measure we used was pointwise mutual information, which prefers less common values of features to more common ones, emphasising that the former characterize a word better than the latter (Carroll et al., 2012).

First, each feature is associated with a frequency determined from the corpus. Then, the information contained in a triple of (w, r, w'), i.e. the mutual information between w and w' w.r.t. the relation r. (Hindle, 1990) can be computed according to Formula 6.2:

$$I(w, r, w') = log \frac{||w, r, w'|| \times ||*, r, *||}{||w, r, *|| \times ||*, r, w'||} \tag{6.2}$$

While $||w, r, w'||$ corresponds to the frequency of the triple (w, r, w') determined from the corpus, when any member of the triple is a $*$, then the frequencies of all the triples corresponding the rest of the triple are summed over. For example, $||*, next\_1, szem||$ corresponds to the sum of the frequencies of words followed by the word *szem* 'eye'.

Then, the similarity between two words (w_1 and w_2) can be counted according to Formula 6.3

$$SIM(w_1, w_2) = \frac{\sum\limits_{(r,w) \in T(w_1) \bigcap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum\limits_{(r,w) \in T(w_1)} I(w_1, r, w) + \sum\limits_{(r,w) \in T(w_2)} I(w_2, r, w)} \tag{6.3}$$

where $T(w)$ is the set of pairs (r, w') such that $I(w, r, w')$ is positive.

It should be noted that even though these models can be applied to all words in the raw text, it is reasonable to build separate models for words of different part-of-speech. Due to the relatively small size of our corpus and the distribution of part-of-speech as described in

DOI:10.15774/PPKE.ITK.2015.012

Figure 6.1: The heatmap of pairwise similarities of terms extracted from a single document. The lighter a square is, the more similar the two corresponding phrases are.

Chapter 2, I only dealt with nouns and nominal multiword terms that appear at least twice in the corpus.

Moreover, in order to avoid the complexity arising from applying this metric between multiword terms, these phrases were considered as single units, having the [N] tag when comparing them to each other or to single nouns. Figure 6.1 shows a heatmap where the pairwise similarities of terms found in a single ophthalmology document are shown. The lighter a square is, the more similar the two corresponding phrases are. As it can be seen on the map, the terms *"tiszta törőközeg"* ('clean refractive media') and *"békés elülső segmentum"* ('calm anterior segment') are similar with regard to their distributional behaviour, while for example the term *"neos mksz"* ('Neo-Synephrine to both eyes') is just slightly related to a few other terms in the particular document.

The results show that the model does indeed identify related terms, however due to the nature of distributional models, the semantic type of the relation may vary. These similarities are paradigmatic in nature, i.e. similar terms can be replaced by each other in their shared contexts. As this is true not only for synonyms, but also for hypernyms, hyponyms and even antonyms, such distinctions can not be made with this method. This shortcoming, however, does not prohibit the application of this measure of semantic relatedness when creating conceptual clusters as the basis of an ontology for the clinical domain. This can be done, because in this sublanguage the classification of terms should be based on their medical relevance, rather than on the common meaning of these words in every day language.

DOI:10.15774/PPKE.ITK.2015.012

Thus, no matter what their meaning might be, terms in the semantic group of e.g. 'signs and symptoms' are all related from the point of view of the semantics characterizing our specific clinical domain.

6.3 Conceptual clusters

Based on the pairwise similarities of words and multiword terms, a conceptual hierarchy can be built. In order to create the hierarchy, we applied agglomerative clustering on the most frequent terms. Each term was represented by a feature vector containing its similarity to all the other terms. Formally, the c_i element of the vector $c(w)$ corresponding to term w is $SIM(w, w_i)$ as defined in Formula 6.3. The clustering algorithm was then applied on these vectors. The linkage method was chosen based on the cophenet correlation between the original data points and the resulting linkage matrix (Sokal and Rohlf, 1962). The best correlation was achieved when using Ward's distance criteria (Ward, 1963) as the linkage method. This resulted in small and dense groups of terms at the lower level of the resulting dendrogram.

However, we needed not only the whole hierarchy, represented as a binary tree, but separate, compact groups of terms, i.e. well-separated subtrees of the dendrogram. The most intuitive way of defining these cutting points of the tree is to find large jumps in the clustering levels. To put it more formally, the height of each link in the cluster tree is to be compared with the heights of neighbouring links below it in a certain depth. If this difference is larger than a predefined threshold value (i.e. the link is inconsistent), then the link is a cutting point.

We applied this cutting method twice. First, we used a lower threshold value to create small and dense groups of terms. At this stage, the full hierarchy was kept and the nodes below the threshold were collapsed, having these groups of terms as leaf nodes (see for example node `1403` in Figure 6.2). In the second iteration, the hierarchy was divided into subtrees by using a higher threshold value. After this step, the hierarchy was only kept within these subtrees, but they were treated as single clusters. Each node in the tree was given a unique concept identifier.

Table 6.3 shows some examples of collapsed groups of terms. The resulting groups contain terms of either similar meaning, or ones having a semantically similar role (e.g. names of months or medicines, etc.), even containing some abbreviated variants as well (e.g. *"bes"*, *"békés elülső szegmens"*, *"békés elülső szegmentum"*, *"békés es"* – all standing for the term 'calm anterior segment'). Beside these semantically related groups of terms, there are some more abstract ones as well, which contain terms related to certain medical processes or phases of medical procedures. For example the term *"éhgyomor"* ('empty stomach') was grouped together with terms related to time and appointments, or *strab* ('strabism') and *párhuzamos szem* ('parallel eyes') were grouped together based on their medical relatedness. Figure 6.2 shows an example of a subtree with root identifier *2016* and the hierarchical organization of groups of terms in the leaf nodes.

DOI:10.15774/PPKE.ITK.2015.012

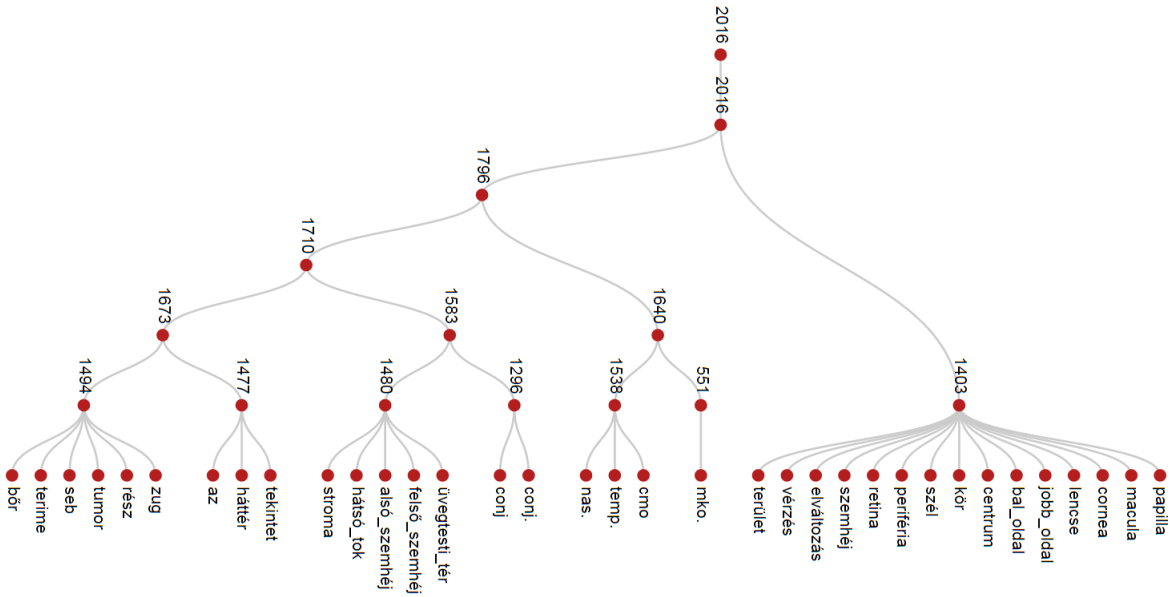


Figure 6.2: A subtree cut out from the whole hierarchy containing groups of terms on the leaf nodes.

I1403	papilla, macula, cornea, lencse, jobb oldal, bal oldal, centrum, kör, szél, periféria, retina, szemhéj, elváltozás, vérzés, terület *papil, macula, cornea, lens, right side, left side, centre, circle, verge, periphery, retina, eyelid, change, blooding, area*
I1636	hely, kh, kötőhártya, szaru, conjunctiva, szemrés, szempilla, pilla, könnypont *place, cj, conjunctiva, white of the eye, conjunctiva, eyelid opening, eyelash, lash, punctum*
I1549	tbl, medrol, üveg, szemcsepp, gyógyszer *tbl, medrol, glass, eyedrop, medication*
I1551	folyamat, kivizsgálás, érték, idegentest, gyulladás, retinaleválás, látásromlás *process, examination, value, foreign body, imflamation, retinal detachment, worse vision*
I1551	ép papilla, halvány papilla, jó színű papilla, szűk ér, ép macula, fénytelen macula, kör fekvő retina, fekvő retina, rb, tiszta törőközeg, bes, békés elülső szegmentum, békés es *intact papil, vague papil, good colored papil, narrow vein, intact macula, dim macula, retina laying around, laying retina, ok, clean refractive media, cas, calm anterior segment, calm as*

Table 6.3: Some example groups of terms as the result of the clustering algorithm

DOI:10.15774/PPKE.ITK.2015.012

6.4 Discovering semantic patterns

The clustering and ordering of terms extracted from clinical documents might be used directly as an initial point of a Hungarian medical (ophthalmological) ontology, containing phrases used by practitioners in their daily cases. However, since each group (and each node in the hierarchy) was given a unique identifier, words and phrases in the original text can be annotated or replaced by these concept ID's. Thus, a higher-level abstract representation can be created for the documents. Then, frequent patterns can easily be determined at this abstract level, no matter what the actual content and the frequency at actual realizations of the pattern are. Table 6.4 contains some example sentence pairs of the abstract and the original sentences. Using cluster identifiers, the sentences not only became simpler, but these patterns were also easy to extract from them.

For example the pattern `1889|2139 1327 1627` characterizes expressions that contain some data about the state of one or both of the patient's eyes. The most frequent realizations of this pattern are: *"st. o. u.", "st. o. s.", "st. o. d.", "moct o. d.", rl. o. u.", "rl. o. sin.", "status o. s.", "távozáskor o. d." ('at leaving'), "b-scan o. d."*, etc. Another characteristic of this pattern is that it appears at the beginning of a sentence, thus this information can be used if this template is used as a separator between different statements in the documents.

These frequent patterns can also be used to classify sentences in each record. In Hungarian clinics, there are EHR (Electronic Health Record) systems, but there are no serious regulations about the use of such software. Thus, doctors or assistants, who are still used to the tradition of writing unstructured documents, tend to type all information into a single text box. Thus, various types of data that should appear under their own heading in the final document are mixed under a single (or no) heading. Thus, it is a separate task to find statements that should belong to different sections, such as anamnesis, lab tests, treatments, opinion, diagnosis, etc.

Thus, we have performed an experiment on defining patterns to find anamnesis sentences. In order to be able to exploit all the information we have already annotated our corpus with, we set up a simple search interface. Then, we were able to define patterns which include constraints applied to different annotation levels, such as the original text, the lemmas of words, part-of-speech, morphological analysis and cluster identifiers.

Anamnesis statements are the ones referring to the reason for a patient's visit to the doctor, their medical history or the aim of admission to hospital. We defined three patterns to cover such statements:

1. Including current complains and the reason of admitting the patient to the clinic. Such sentences include past tense singular verbs between any members of the cluster `1436` on both sides (not necessarily immediate neighbours). The pattern can be defined as {`I1436 .. VERB|Past|singular .. I1436`}. Sentences belonging to this group were like the one in the first example in Table 6.5.

2. Including some past events, or medical history. This pattern is similar to the previous one, differing only in the first part, i.e. there is no need for a preceding constraint before

DOI:10.15774/PPKE.ITK.2015.012

1518 1706 1706 : ***2016*** tiszta üti *2007* , *2045* , szemfenék-szerte *2007* , *1956* , a macula_kemény_exsudatum , *2007* . *fu. o. u* : ***mko.*** tiszta üti *tér* , *ép_papilla* , szemfenék-szerte *ma-k* , *pontszerű_vérzés* , a macula_kemény_exsudatum , *oedema* .
2071 1706 1706 : ***2016*** felett és nasalisan szivárgó , ischaemiás_terület , kis neovasc._burjánzás , *2049* *flag o. d.* : ***macula*** felett és nasalisan szivárgó , ischaemiás_terület , kis neovasc._burjánzás , *macula_oedema*
2016 sima , csillogó , *2007* és a *1789* tiszta . ***cornea*** sima , csillogó , *állománya* és a *hátlapja* tiszta .
2016 tizsta . *1706* friss *1884* nem látható . ***lencse*** tizsta . *funduson* friss *kóros* nem látható .
2016 tiszta , *1789* tiszta , *1789* tiszta , *1789* békés , *1789* jól reagál . ***stroma*** tiszta , *hátlap* tiszta , *csarnok* tiszta , *iris* békés , *pupilla* jól reagál .
2016 nem vizsgálható erős_fénykerülés miatt . ***perifféria*** nem vizsgálható erős_fénykerülés miatt .
1889 1706 1706 : *1998* , halvány_conjunctiva , ***2016*** epithelialis pontszerű_kiemelkedő_szürkésfehér_laesi_(j»b, balon csak NUM-NUM laesio) , a cornea_mély_rész épek ,transparens , *1812* mély , tiszta , *1789* békés , *1812* tág , kerek , centrális , *2007* jól reagál . *st. o. u* : *ép_védőszerv* , halvány_conjunctiva , ***corneán*** epithelialis pontszerű_kiemelkedő_szürkésfehér_laesi_(j»b, balon csak NUM-NUM laesio) , a cornea_mély_rész épek , transparens , *csarnok_kp* mély , tiszta , *iris* békés , *pupilla_kp* tág , kerek , centrális , *fényre* jól reagál .

Table 6.4: Examples of sentences where terms are replaced by their cluster identifiers

DOI:10.15774/PPKE.ITK.2015.012

1. I1436..VERB|Past|singular..I1436
Betegünk/N glaucomás/N kivizsgálás/N céljából/N került/V|Past felvételre/N.
1436 1930 1551 1434 1436
'Our patient was admitted for the investigation of glaucoma.'
2. VERB|Past|singular..I1436
Tegnap/N óta/P begyulladt/V|Past a/Det szeme/N.
2249 1436
'His eyes are inflamed since yesterday.'
3. I1436..VERB|Cond
Úgy/Adv érzi/V, mintha/C valami/N lenne/V|Cond a/Det szemében/N.
2187 1436
'He feels as if there was something in his eyes.'

Table 6.5: Example sentences for each pattern describing anamnesis statements.

the verb resulting in the form of {VERB|Past|singular .. I1436}. An example is the second one in Table 6.5.

3. Including some conditional or wish statements. These are sentences describing some uncertain descriptions of the feelings of the patient or a certain wish. These are covered by the pattern {I1436 .. VERB|Cond}. But, since we found only few examples of such sentences, a simpler form of this pattern was also tried (i.e. including only the conditional verb), which produced more results, but the precision was much lower in this case at the expense of recall. See the third example in Table 6.5 for a sentence covered by this pattern.

Table 6.6 shows the performance results of recognizing anamnesis sentences based on these patterns. ALL is the number of results, out of which TP (true positive) is the ratio of sentences classified correctly, and the rest are the ERRORS. Erroneously classified sentences fall into three categories. (1)FP (false positive) is the ratio of erroneous sentences classified incorrectly. (2)There were some sentences that were too ambiguous or malformed, so that we could not decide whether its content is anamnesis or not (UD). (3)There were also some sentences labelled as P/S, which contained some errors at a lower preprocessing level, such as a misspelling or bad pos-tag. However, only those errors were counted here, which caused the sentence to be erroneously classified. For example past participles are frequently mistagged as past tense verbs, which lead to the high ratio of errors in the second case.

DOI:10.15774/PPKE.ITK.2015.012

PATTERN	ALL(#)	TP(%)	ERRORS(%)		
			FP	UD	P/S
1. `I1436..VERB|Past|singular..I1436`	147	0.972	0.4	0.4	0.2
2. `VERB|Past|singular..I1436`	192	0.961	0.16	0	0.84
3a. `I1436..VERB|Cond`	11	1.0	0	0	0
3b. `VERB|Cond`	145	0.889	0.75	0.187	0.062

Table 6.6: Results of recognizing anamnesis sentences based on multilevel patterns.

6.5 Concept trees of stuctural units

Building a concept tree based on the textual parts of the corpus was useful for detecting patterns and getting a general overview of the concepts in the documents. However, when these trees were investigated to be used as a starting point for building an ontology or system of concepts, several problems have risen:

- The heterogeneous content resulted in a rather noisy structure, containing terms from the various sections of the documents.
- The distribution of concepts are unbalanced due to the individual writing style of certain doctors, suppressing the language use in documents written by less frequent doctors in the corpus.
- Tokenization errors, misspellings and abbreviations might be clustered to improper subtrees.
- The coverage of the concept tree depends on the actual documents in the corpus, thus it might miss some infrequent phenomena in ophthalmology.

In order to have a more robust system of concepts, I applied the same clustering algorithm on different inputs.

6.5.1 Concept trees of structural units

First, I dismantled the original corpus to portions each containing texts of the same structural unit from all documents. As shown in Chapter 6.5, the content of the documents were classified into 15 different classes. Thus, I created the 15 subcorpora, each containing statements characteristic of the certain structural part. For example, while the anamnesis subcorpus contains some stories the patient told, or general diseases and family history, the diagnosis subcorpus contains mostly Latin names of diseases of the eye. See Table 3.1 for examples of each part. When creating a concept tree from the whole corpus, both the terms found in the different parts and the distributional model were mixed. Thus, I created these models for each subcorpus separately and built the hierarchical organization with different parameters for each part. The resulting trees were more dense and proper representations as it can be seen in Figure 6.3. As the examples reveal, the nature of the concepts differ in the two subtrees taken from the hiearchy built from the structural units of Therapy (`Ther`) and Slit lamp (`RL`). While the first one contains different types of medications and related terms

DOI:10.15774/PPKE.ITK.2015.012

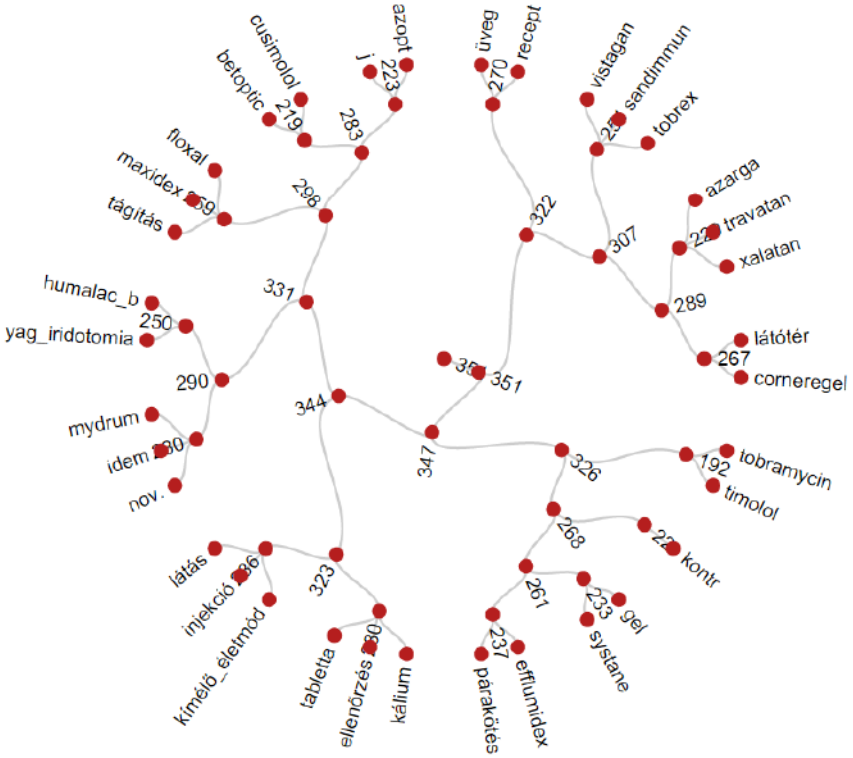
(e.g. injekció 'injection', üveg 'glass', recept 'prescription', etc), the other contains terms related to the actual investigation of the patient.

6.5.2 Concept trees of the ophthalmology science

Though the trees built from the different parts of the documents represent these units of the documentation of ophthalmological visits well, they might include some out-of-domain terminology, while miss some relevant terms due to the unbalanced distribution of words. Thus, I created another hierarchy of concepts by applying the same algorithm on the content of an official ophthalmology book in Hungarian (Süveges, 2010), which is also available online in a digital edition[i]. This book contains not only the functional anatomy of the eye and the description of its disorders, but is used in the practical education of doctors. Thus, its attitude is close to the original documents of the ophthalmology corpus. However, it has several advantages: it is written by a single author, thus the language use is more homogeneous. Moreover, its content is not biased by the repetitive mention of frequent diseases, as it happens in the corpus of visitation notes. Since it is an edited book (with a size of 125 824 tokens), it contains mostly proper sentences making tokenization and part-of-speech tagging perform better than in the original corpus. However, regarding misspellings, I found that despite the proofreading and printed edition, the text contained a surprisingly high ratio of spelling errors. Nevertheless, the final hierarchy built from its concepts provided a comprehensive structure of the terminology of this domain. Figure 6.4 shows an example subtree cut from the whole hierarchy of concepts built from this book.

[i] http://www.tankonyvtar.hu/hu/tartalom/tamop425/2011_0001_524_szemeszet/adatok.html

DOI:10.15774/PPKE.ITK.2015.012



(a)

(b)

Figure 6.3: Two examples of subtrees of concepts built from the structural units Therapy (a) and Slit lamp (b)

DOI:10.15774/PPKE.ITK.2015.012

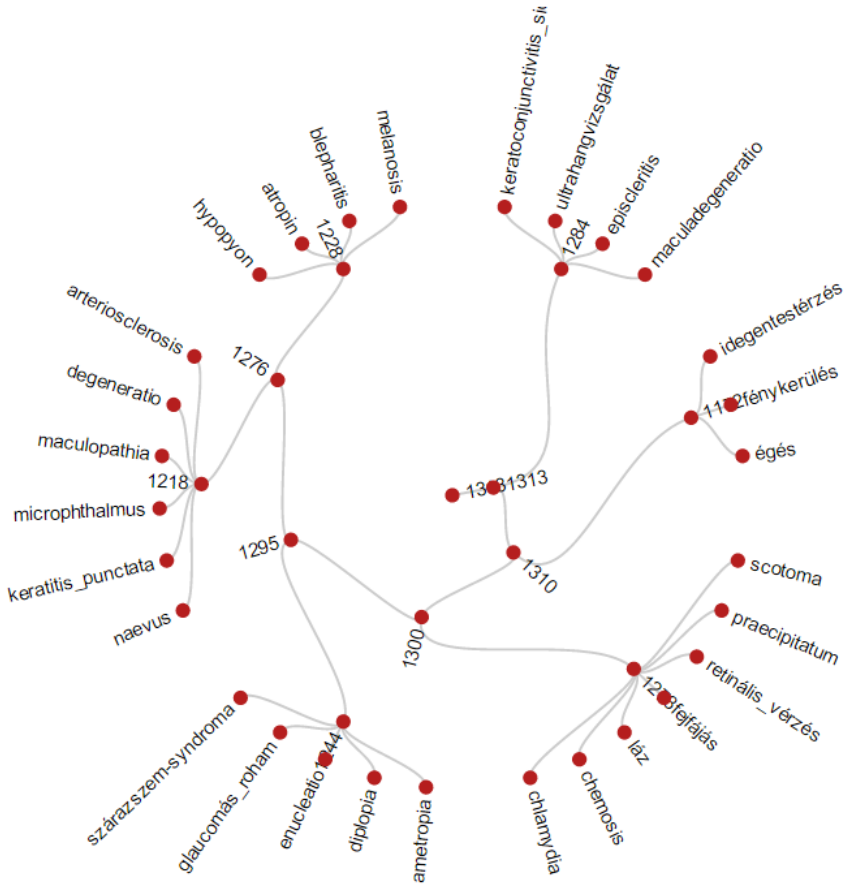


Figure 6.4: A subtree from the hierarchy of the Szemészet (Süveges, 2010) book

DOI:10.15774/PPKE.ITK.2015.012

DOI:10.15774/PPKE.ITK.2015.012

7

Related work

I am not the only one inventing the wheel. There are several excellent researchers struggling with similar problems. I relied on their results, but of course, none of them can be compared to each other, not even to mine... Or mine to theirs...

Contents

DOI:10.15774/PPKE.ITK.2015.012

Processing clinical records (also mentioned as Electronic Health Records (EHR) in the literature, but since Hungarian clinics use far less sophisticated systems, in our case the name would be inadequate) has been an area of growing interest in the field of natural language processing (NLP). As a huge amount of information is stored in digital documents describing everyday cases of patient care, the practical knowledge in these documents is a valuable resource, the only barrier hiding it from even medical practitioners is the amount of noise and the unstructured and unsearchable way they are stored. Thus, the goal of processing such documents is to gain access to the information stored in them.

Much of the research done related to medical language processing is applied to biomedical texts, which include scientific articles, books, i.e. proper, proofread literature. However the language of biomedical literature is very different from that of clinical documents, which are written in a special notational language used in clinical settings, containing a lot of abbreviations, misspellings and incomplete grammatical structures as shown in Chapter 2. Thus, these texts require different methods, and it has been shown that when trying to apply general linguistic applications to clinical records, there is a significant performance drop (Meystre et al., 2008; Hassel et al., 2011; Orosz et al., 2014; Dalianis et al., 2009). Moreover, there is a difference in the quality of clinical texts in different countries depending on institutional or state regulations on the expected content and quality of clinical records. However, in Hungary there is no such a constraint, thus producing documents hardly understandable does not have any consequences for the practitioners. This makes the simple adaptation of general tools insufficient and methods of preprocessing and normalization not always necessary in such depths for other languages unavoidable.

One of the earliest studies in processing clinical narratives, also mentioned in a comprehensive report on clinical text processing by Meystre et al. (2008), is that of Sager et al. (1994), relying on the sublanguage theory by Harris (2002). Based on this research, Friedman et al. (1995) developed MedLEE (Medical Language Extraction and Encoding System) that is used to extract information from clinical narratives to enhance automated decision-support systems. These systems are capable of creating complex representations of events found in clinical notes. Furthermore, they fulfil the expectations of extracting trustworthy information and revealing extended knowledge as well as deeper relations found in these texts. However, all of these methods rely on proper, well-formed, and correct input documents and are applicable to English texts. Since the goal of my research was to achieve a preprocessed state of documents suitable for deeper analysis and information extraction, I will not cover a detailed review of systems functioning at such higher levels, rather methods regarding preprocessing steps and the theoretical background of my research.

7.1 Corpora and resources

Accessing large amounts of clinical documents must face serious limitations. First, confidentiality provisions inhibit clinics to provide their documents even for research. Beside the problem of anonymization, which is made more complex in Hungarian clinical documentation systems due to their adhoc usage, there is a general distrust in both doctors and patients.

DOI:10.15774/PPKE.ITK.2015.012

In other languages there are some already publicly available clinical corpora (Bourke et al., 2004; Dalianis et al., 2009; Savova et al., 2010), but for Hungarian this is not the case.

Beside corpora, clinical text processing methods also rely on external resources. Even for normalization purposes, such ontologies and thesauri are used for mapping various word forms and abbreviations to their corresponding entries in these databases. The largest such resources are UMLS, MeSH and SNOMED, to name a few. For encoding purposes the system of ICD-10 is also used. Only the latter has a Hungarian translation (the BNO system), however, the entries and its artificial hierarchy is very hard to apply on the documents. Nevertheless, I also used this resource in my research. However, the lack of structured resources for less-resourced languages, such as Hungarian, makes it very hard to produce results comparable to those achieved by solutions for major languages. One way to overcome this problem could be the translation of these resources, however, doing it manually would require a huge amount of work, and automated methods that could support the translation effort are also of low quality for these languages.

7.2 Spelling correction

Kukich (1992) partitions the problem of spelling correction to three subcases as (a) non-word error detection; (b) isolated-word error correction; and (c) context-dependent word correction. However, most of the techniques described in the study by Kukich (1992) rely on a lexicon-based approach that is not applicable to agglutinating languages such as Hungarian. The problems of spelling correction for agglutinative languages is described by Oflazer and Güzey (1994). One way of handling an infinite vocabulary is applying finite-state automata or transducers, which are used in implementations by Park and Levy (2011), Noeman and Madkour (2010) and Pirinen and Lindén (2010). In my work, I aim at performing all three tasks in one step, that is, recognizing and correcting misspellings in context. Since the adequate use of this clinical language is never present in the documents, the goal to achieve is a quasi-standard representation (i.e. each concept represented by the same string for all occurrences) even if that spelling does not correspond to the orthographic standard.

In order to achieve this goal, a hybrid approach was chosen. Statistical or hybrid approaches are reported to outperform the previously prevailing rule-based methods. A widespread solution is to apply the noisy-channel model. The systems of Church and Gale (1991) and Brill and Moore (2000) apply variants of this model using different error models and probability scoring. Furthermore, Boswell (2004) emphasizes the beneficial use of a contextual language model in the case of spelling correction while adopting the noisy-channel model.

Although the idea of the noisy-channel model is the basis of statistical machine translation (SMT) algorithms, only very few studies use SMT implementations directly (Brockett et al., 2006; Ehsan and Faili, 2013). Still, the task of spelling correction can be modelled as a translation task, where the source language is the erroneous text and the target language is the corrected one.

DOI:10.15774/PPKE.ITK.2015.012

Similar models are used by Turchin et al. (2007), where misspelled words are identified by comparing them to some predefined list of words. This baseline method is extended by doing prevalence analysis, i.e. determining the frequency ratio of a word and its one-edit-distance alternatives in the corpus. Mass noun errors in English as a Second Language texts are corrected by a similar technique by Brockett et al. (2006). However, that is a grammatical rather than an orthographic problem.

The system most similar to my approach is that of Ehsan and Faili (2013), in which a traditional SMT algorithm performs spelling error correction. In that implementation, the translation model is based on a parallel corpus of proofread and erroneous texts into which errors were introduced artificially. However, these random errors might not model human-made mistakes satisfactorily. Furthermore, the system needs a correctly written corpus in the first place, which I do not have in the clinical domain. Training the system on general texts would not be applicable on clinical documents due to the differences detailed in Chapter 2.

There are some approaches aiming at the specific problem of spelling correction in the clinical domain as well. A research published by Patrick and Nguyen (2011) uses several knowledge bases of English clinical terms beside applying statistical methods. Crowell et al. (2004) described their implementation for rescoring the ranked candidates of different correction suggestion methods.

7.3 Detecting and resolving abbreviations

There are several studies, most of them applied to English texts, that address the specific task of the detection and resolution of abbreviations found in the free-text parts of clinical documents. As opposed to biomedical literature, where the first mention of an abbreviated form is usually preceded by its expanded form or definition, in clinical records this is not the case. That is why simple abbreviation-definition patterns are not applicable to clinical notes as described by Xu et al. (2007). The same study compares some machine learning approaches, all achieving considerable results, but even using already existing external resources, the authors admit the need of a manually created inventory. A recent study of Wu et al. (2012) compared the performance of some biomedical text processing systems trained on biomedical literature on the task of resolving abbreviations in clinical texts. All the systems (MetaMap, MedLEE and cTAKES) achieved suboptimal results calling for more advanced abbreviation recognition modules.

Most approaches to resolving clinical abbreviations carried out in English rely on some very common medical lexical resources. Even though Xu in Xu et al. (2007) showed that the sense inventories generated from the UMLS covered only about 35% of the abbreviations they had extracted from their corpus, these already contain definitions and possible interpretation candidates. Thus the problem can be reduced to abbreviation disambiguation, as it is carried out in Wu et al. (2012); Xu et al. (2009); Pakhomov et al. (2005). These methods focus primarily on supervised machine learning approaches, where a part of the training corpus

DOI:10.15774/PPKE.ITK.2015.012

is labelled manually. Pakhomov in Pakhomov (2002) described a semi-supervised method to build training data for Maximum Entropy modeling of abbreviations automatically. In most of these studies, both training and evaluation of the systems are performed on a few manually chosen abbreviations and their disambiguation.

7.4 Identification of multiword terms

There are a lot of research on automatic term recognition (ATR) aiming at extracting multiword terms from general or domain-specific collections of texts. The main branches of ATR methods are that of rule-based linguistic approaches, purely statistical ones and hybrid combinations of the two. Although there are approaches to adapt existing multiword extraction methods to special domains (e.g. Nagy T. et al. (2011)), the unique characteristics of different sublanguages require specific implementations optimized to recognize the types of multiword terms characteristic to the actual domain. Thus, there have been applications designed for the specific domain of biomedicine, ecology, mathematics, social networks, banking, natural sciences, information technology, etc. (Lossio-Ventura et al., 2014b).

There are examples of different methods applied in the clinical or biomedical domain as well. Gaizauskas et al. (2000) define rules based on morphosyntactic and lexical characteristics. Statistical methods ranging from the baseline application of *tfidf* (term frequency, inverse document frequency) calculations borrowed from information extraction to the application of statistical tests (Lossio-Ventura et al., 2014a).

Beside the method applied, term extraction approaches can be categorized according to their view on *unithood* and *termhood*. The former emphasizes the recognition of units of phrases in the texts being possible candidates as terms. The latter ranks these candidates. Hybrid approaches tend to apply some linguistic patterns and measures such as mutual information in order to extract term candidates and then use statistics to rank them (Zhang et al., 2008). Weirdness was introduced as a measure that relies on the difference between the distribution of words in a specific and in general domains (Ahmad et al., 1999).

However, as Zhang et al. (2008) has shown, methods that do not apply the identification of unithood and termhood in separate steps, perform best. The most well-known such measure is the C/NC-value method (Frantzi et al., 2000). They use frequency data to identify nested terms in a list of candidate terms extracted by n-gram and linguistic filters. Thus, unithood incorporates frequency data as well. The NC-value is an extension of the C-value measure, which retrieves the context of multiword terms. It has been shown that in a comparative evaluation, the C-value method outperforms other hybrid approaches (Zhang et al., 2008). Moreover, it has been applied in the clinical domain (Frantzi et al., 2000; Lossio-Ventura et al., 2014a; Ittoo and Bouma, 2013).

DOI:10.15774/PPKE.ITK.2015.012

7.5 Application of distributional methods for inspecting semantic behaviour

Semantics is needed to understand language. Even though, most applications apply semantic models as one of the latest modules of a language processing pipeline, in my case some basic semantic approaches were reasonable to apply as preprocessing steps. Still, my goal was not to create a fully functional semantic representation, thus related literature was also investigated from this perspective.

There are two main branches of computationally handling semantic behaviour of words in free texts: mapping them to formal representations, such as ontologies as described by Jurafsky and Martin (2000) and applying various models of distributional semantics (Deerwester et al., 1990; Schütze, 1993; Lund and Burgess, 1996) ranging from spatial models to probabilistic ones (for a complete review of empirical distributional models consult Cohen and Widdows (2009)). Handmade resources are very robust and precise, but are very expensive to create and often their representation of medical concepts do not correlate to the real usage in clinical texts (Zhang, 2002; Cohen and Widdows, 2009). On the other hand, the early pioneers of distributional semantics have shown that there is a correlation between distributional similarity and semantic similarity, which makes it possible to derive a representation of meaning from the corpus itself, without the expectation of precise formal definitions of concepts and relations (Cohen and Widdows, 2009).

Beside the various applications of distributional methods in the biomedical domain, there are approaches, when these are applied to texts from the clinical domain. Carroll et al. (2012) create distributional thesauri from clinical texts by applying distributinal models in order to improve recall of their manually constructed word lists of symptoms and to quantify similarity of terms extracted. In their approach, the context of terms is considered as the surrounding words within a small window, but they do not include any grammatical information as opposed to my definition of features representing context. Still, they report satisfactory results for extracting candidates of thesaurus entries of nouns and adjectives, producing somewhat worse results in the latter case. However, the corpus used in their research was magnitudes larger than in my approaches. As Sridharan and Murphy (2012) have shown, either a large corpus or a smaller one with high quality is needed for distributional models to perform well, emphasising the quality over size. This explains the slightly lower, but still satisfactory results in my case. The similarity measure used in the work of Carroll et al. (2012) was based on the one used by Lin (1998). In that study it is also applied to create thesauri from raw texts, however there it is done for general texts and is exploiting grammatical dependencies produced by high-quality syntactic parsers. A detailed overview of distributional semantic applications can be found in Cohen and Widdows (2009) and Turney and Pantel (2010) and on its application in the clinical domain in Henriksson (2013).

DOI:10.15774/PPKE.ITK.2015.012

8

Conclusion
New scientific results

The most important part of this thesis work, summarizing new scientific results described in the previous chapters. Includes the main points of the whole Thesis putting them most densely in the Thesis sentences 1 to 5.b.

Contents

DOI:10.15774/PPKE.ITK.2015.012

Processing medical texts is an emerging topic in natural language processing. There are existing solutions mainly for English to extract knowledge from medical documents, which will be available for researchers and medical experts. However, locally relevant characteristics of applied medical protocols or information relevant to locally prevailing epidemic data can be extracted only from documents written in the language of the local community. In the case of less-resourced languages, such as Hungarian, the lack of structured resources, like UMLS, SNOMED, etc. makes it very hard to produce results comparable to those achieved by solutions for major languages. One way to overcome this problem could be the translation of these resources, however, doing it manually would require a huge amount of work, and automated methods that could support the translation effort are also of low quality for these languages.

Moreover, the quality of the documents created in the clinical settings is much worse, than that of general texts. Thus, the goal of this research was to transform raw clinical documents to a normalized representation that is appropriate for further processing. The methods applied are based on statistical algorithms, exploiting the information found within the corpus itself even at such preprocessing steps.

8.1 Representational schema

Wide-spread practice for representing structure of texts is to use XML to describe each part of the document. In my case it is not only for storing data in a standard format, but also representing the identified internal structure of the texts which are recognized by basic text mining procedures, such as transforming formatting elements to structural identifiers or applying recognition algorithms for certain surface patterns.

The resulting structure defines the separable parts of each record; however there are still several types of data within these structural units. Non-textual information inserted into free word descriptions are laboratory test results, numerical values, delimiting character series and longer chains of abbreviations and special characters. I filtered out these expressions to get a set of records containing only natural text. To solve this issue, unsupervised clustering algorithms were applied.

Digging deeper into the textual contents of the documents, a more detailed representation of these text fragments was necessary. That is why I stored each word in each sentence in an individual data tag, augmented with several information. Such information are the *original form* of the word, the *corrected form*, its *lemma* and *part-of-speech* tag, and some phrase level information such as different types of *named entities*.

At this point, the textual content segments, each intended to appear under various subheadings, still remained as a mixture under a *content* tag. The original sections under these subheadings (*header*, *diagnoses*, *applied treatments*, *status*, *operation*, *symptoms*, etc.) contain different types of statements requiring different methods of higher-level processing. Moreover, the underlying information had to be handled in different ways, unique to each subheading. Thus, I implemented a method for categorizing lines of statements into their

DOI:10.15774/PPKE.ITK.2015.012

intended subheading. This was performed in two steps. First, formatting clues were recognized and labelled. These labelled lines were used as the training set for the second step, in which unlabelled lines were categorized by finding the most similar tag collection based on the tf-idf weighted cosine similarity measure.

THESIS 1:

I defined a flexible representational schema for Hungarian clinical records and developed an algorithm that is able to transform raw documents to the defined structure.

Related publications: 1, 4, 10, 16, 17

8.2 Automatic spelling correction

In Hungarian hospitals, clinical records are created as unstructured texts, without any proofing control (e.g. spell checking). Moreover, the language of these documents contains a high ratio of word forms not commonly used: such as Latin medical terminology, abbreviations and drug names. Many of the authors of these texts are not aware of the standard orthography of this terminology. Thus the automatic analysis of such documents is rather challenging and automatic correction of the documents was a prerequisite of any further linguistic processing.

The errors detected in the texts fall into the following categories: errors due to the frequent (and apparently intentional) use of non-standard orthography, unintentional mistyping, inconsistent word usage and ambiguous misspellings (e.g. misspelled abbreviations), some of which are very hard to interpret and correct even for a medical expert. Besides, there is a high number of real-word errors, i.e. otherwise correct word forms, which are incorrect in the actual context. Many misspelled words never or hardly ever occur in their orthographically standard form in our corpus of clinical records.

I prepared a method for considering textual context when recognizing and correcting spelling errors. My system applies methods of Statistical Machine Translation (SMT), based on a word-based system for generating correction candidates. First a context-unaware word-based approach was created for generating correction suggestions, then I integrated this into an SMT framework. My system is able to correct certain errors with high accuracy, and, due to its parametrization, it can be tuned to the actual task. Thus, the presented method is able to correct single errors in words automatically, making a firm base for creating a normalized version of the clinical records corpus in order to apply higher-level processing.

8.2.1 The word-based correction suggestion system

First, a word-based system was implemented that generates correction candidates for single words based on several simple word lists, some frequency lists and a linear scoring system.

DOI:10.15774/PPKE.ITK.2015.012

At the beginning of the correction process, word forms that are contained in a list of stopwords and abbreviations are identified. For these words, no suggestions are generated. For the rest of the words, the correction suggestion algorithm is applied. For each word, a list of suggestion candidates are generated that contains word forms within one edit distance from the original form. The possible suggestions generated by a wide-coverage Hungarian morphological analyzer (Prószéky and Kis, 1999; Novák, 2003) are also added to this list.

In the second phase, these candidates are ranked using a scoring method based on (1) the weighted linear combination of scores assigned by several different frequency lists, (2) the weight coming from a confusion matrix of single-edit-distance corrections, (3) the features of the original word form, and (4) the judgement of the morphological analyzer. The system is parametrized to assign much weight to frequency data coming from the domain-specific corpus, which ensures not coercing medical terminology into word forms frequent in general out-of-domain text. Thus a ranked list of correction candidates is generated to all words in the text (except for the abbreviations and stopwords). However, only those are considered to be relevant, where the score of the first ranked suggestion is higher than that of the original word. This system was able to recognize most spelling errors and the list of the 5 highest ranked automatically generated corrections contained the actually correct one in 99.12% of the corrections in the test set.

8.2.2 Application of statistical machine translation to error corrections

Since my goal was to create fully automatic correction, rather than offering the user a set of corrections, the system should be able to automatically find the most appropriate correction. In order to achieve this goal, the ranking of the word-based system based on morphology and word frequency data proved to be insufficient. To improve the accuracy of the system, lexical context also needed to be considered.

To satisfy these two requirements, I applied Moses (Koehn et al., 2007), a widely used statistical machine translation (SMT) toolkit. During "translation", the original erroneous text is considered as the source language, while the target is its corrected, normalized version. In this case, the input of the system is the erroneous sentence: $E = e_1, e_2 \dots e_k$, and the corresponding correct sentence $C = c_1, c_2 \dots c_k$ is the expected output. Applying the noisy-channel model terminology to my spelling correction system: the original message is the correct sentence and the noisy signal received at the end of the channel data is the corresponding sentence containing spelling errors. The output of the system trying to decode the noisy signal is the sentence $\hat{C}$, where

$$\hat{C} = argmax P(C|E) = argmax \frac{P(E|C)P(C)}{P(E)} \tag{8.1}$$

conditional probability takes its maximal value. Since $P(E)$ is constant, the denominator can be ignored, thus the product in the numerator can be derived from the statistical translation and language models.

DOI:10.15774/PPKE.ITK.2015.012

These models in a traditional SMT task are built from a parallel corpus of the source and target languages based on the probabilities of phrases corresponding to each other. However, in my case there was no such a parallel set of documents. Thus, the creation of the translation models was substituted by three methods: (1) the word-based correction candidate generation system, (2) transformation of the distribution of various forms of abbreviations, and (3) inserting a table containing joining errors. These phrase tables are generated online, for each sentence that is to be corrected. The language model responsible for checking how well each candidate generated by the translation models fits the actual context is built using the SRILM toolkit (Stolcke et al., 2011). I have shown that the context-aware system outperformed the word-based one regarding both error detection and error correction accuracy.

THESIS 2:

I created an advanced method to automatically correct single spelling errors with high accuracy in Hungarian clinical records written in a special variant of domain-specific language containing expressions of foreign origin and a lot of abbreviations. I showed that applying a statistical machine translation framework as a spelling correction system with a language model responsible for context information is appropriate for the task and can achieve high accuracy.

Related publications: 1, 2, 6, 16, 17

8.3 Detecting and resolving abbreviations

Abbreviations occurring in clinical documents are usually ambiguous regarding not only their meaning, but the variety of the different forms they can take in the texts (for example *o.sin./o sin/o.s./os/OS*, etc.). Moreover, the ambiguity is further increased by the several resolution candidates a single abbreviated token might have (e.g. *o./f./p.*). Thus, after detecting abbreviations with the help of rules described by regular expressions, I investigated these shortened forms in the lexical context they appear in. When defining detection rules, I had to consider the non-standard usage of abbreviations, which is a very frequent phenomenon in clinical texts. The word-final period is usually missing, capitalization is used in an ad-hoc manner, compound expressions are abbreviated in several ways.

When performing the resolution of the detected abbreviations, I considered series of shortened forms (i.e. series of neighbouring tokens without any full word breaking the sequence) as single abbreviations. In such units, the number of possible resolutions of single, ambiguous tokens is reduced significantly. My goal was to find an optimal partitioning and resolution of these series in one step, i.e. having a resolved form corresponding to as much tokens as possible, while having as few partitions as possible.

Thus, in this research, a corpus-based approach was applied for the resolution of abbreviations with using the very few lexical resources available in Hungarian. Even though the first

DOI:10.15774/PPKE.ITK.2015.012

approach was based on the corpus itself, it did not provide acceptable results, thus the construction of a domain-specific lexicon was unavoidable. But, instead of trying to create huge resources covering the whole field of medical expressions, I have shown that a small, domain-specific lexicon is satisfactory, and the abbreviations to be included can be derived from the corpus itself.

Having this lexicon and the abbreviated tokens detected, the resolution was based on series of abbrevitaions. Moreover, in order to save mixed phrases (when only some parts of a multiword phrase is abbreviated) and to keep the information relevant for the resolution of multiword abbreviations, the context of a certain length was attached to the detected series. Beside completing such mixed phrases, the context also plays a role in the process of disambiguation. The meaning (i.e. the resolution) of abbreviations of the same surface form might vary in different contexts.

These abbreviation series were then matched against the corpus, looking for resolution candidates, and only unresolved fragments were completed based on searching in the lexicon. I have shown that having the corpus as the primary source is though insufficient, but provides more adequate resolutions in the actual domain, resulting in a performance of 96.5% f-measure in the case of abbreviation detection and 80.88% f-measure when resolving abbreviations of any length, while 88.05% for abbreviation series of more than one token.

THESIS 3:

I prepared an algorithm that is able to detect and resolve abbreviations in Hungarian clinical documents without relying on robust lexical resources and hand-made rules, rather applying statistical observations based on the clinical corpus.

THESIS 3.a:

I have shown that ambiguous abbreviations are much easier to be interpreted as members of abbreviation series, moreover, adding a one token long context to these series has also beneficial effect on the performance of the disambiguation process.

THESIS 3.b:

I have shown that the presence of a domain-specific lexicon is crucial, however it does not need to be a large, detailed knowledgebase. A small lexicon can be created by defining the resolution for the most frequent abbreviations found in a corpus of a narrow domain.

Related publications: 1, 7, 12, 13, 14

DOI:10.15774/PPKE.ITK.2015.012

8.4 Semi-structured representation of clinical documents

Clinical documents represent a sublanguage regarding both the content and the language used to record them. However, one of the main characteristics of these texts is the high ratio of noise due to misspellings, abbreviations and incomplete syntactic structures. It has been shown that for a less-resourced language, such as Hungarian, there is a lack of lexical resources, which are used in similar studies to identify relevant concepts and relations for other languages. Thus, such lexicons should be built manually by human experts. However, an initial preprocessed transformation of the raw documents makes the task more efficient. Due to the availability of efficient implementations, statistical methods can be applied to a wide variety of text processing tasks. That is why, I have shown that corpus-based approaches (augmented with some linguistic restrictions) perform well on multiword term extraction and distributional similarity measures. Applying such methods can result in a semi-structured representation of clinical documents, appropriate for further human analyses.

8.4.1 Extracting multiword terms

In the clinical language (or in any other domain-specific or technical language), there are certain multiword terms that express a single concept. These are important to be recognized, because a disease, a treatment, a part of the body, or other relevant information can be in such a form. Moreover, these terms in the clinical reports could not be covered by a standard lexicon. This indicates the need to use some method for collocation identification. I used a modified version of the c-value algorithm (Frantzi et al., 2000). First, I defined a linguistic filter that is applied in order to ensure that the resulting list of terms contains only well-formed phrases. Phrases of the following forms were allowed:

$$(Noun|Adjective|PresentParticiple|Past(passive)Participle)^{+}Noun$$

This pattern ensures that only noun phrases are extracted and excludes fragments of frequent cooccurrences. After collecting all n-grams matching this pattern, the corresponding C-value is calculated for each of them, which is an indicator of the termhood of a phrase. The C-value is based on four components:

- the frequency of the candidate phrase;
- the frequency of the candidate phrase as a subphrase of a longer one;
- the number of these longer phrases;
- and the length of the candidate phrase.

These statistics are derived from the whole corpus of clinical notes, thus the resulting list of multiword terms are well suited in the domain and reflect their usage in the Hungarian medical language.

DOI:10.15774/PPKE.ITK.2015.012

THESIS 4:

I have shown that corpus-based approaches augmented with some linguistic restrictions perform well on multiword term extraction from Hungarian clinical documents, resulting in a list of domain-specific terminology phrases ranked according to their termhood.

Related publications: 1, 4, 5, 11

8.4.2 Distributional behaviour of the clinical corpus

Creating groups of relevant terms in the corpus requires a similarity metric measuring the closeness of two terms. Instead of using an ontology for retrieving similarity relations between words, I applied the unsupervised method of distributional semantics. Thus, the similarity of terms is based on the way they are used in the specific corpus.

The context of a word is represented as a set of features, each feature consisting of a relation (r) and the related word (w'). In other studies these relations are usually grammatical relations, however in the case of clinical texts, the grammatical analysis performs poorly, resulting in a rather noisy model. In Carroll et al. (2012), Carroll et al. suggest using only the occurrences of surface word forms within a small window around the target word as features. In my research, a mixture of these ideas was used by applying relations based on the word forms and part-of-speech tags.

Each feature was associated with a frequency determined from the corpus. From these frequencies the amount of information contained in a tuple of (w,r,w') was computed by using maximum likelihood estimation. This is equal to the mutual information between w and w'. Then, to determine the similarity between two words (w_1 and w_2) the similarity measure described in Lin (1998) was used, i.e.:

$$SIM(w_1, w_2) = \frac{\sum_{(r,w) \in T(w_1) \bigcap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w) \in T(w_1)} I(w_1, r, w) + \sum_{(r,w) \in T(w_2)} I(w_2, r, w)}$$

where T(w) is the set of pairs (r,w') such that I(w,r,w') is positive.

Having this metric, the pairwise distributional similarity of any two terms can be counted. In my research, however, I only dealt with nouns that appear at least twice in the corpus and multiword expressions.

The results showed that the resulting similarity relations are valid between terms, which made the application of this measure of semantic relatedness appropriate for creating conceptual clusters as the base of an ontology for the clinical domain. In order to create such a hierarchy, I applied an agglomerative clustering algorithm on the most frequent terms and nouns, where each term was represented by a feature vector containing its similarity to all the other terms. The clustering and ordering of terms extracted from clinical documents can be used directly as an initial point of a Hungarian medical ontology containing phrases used by practitioners

DOI:10.15774/PPKE.ITK.2015.012

in their daily cases. Moreover, since each group (and each node in the hierarchy) was given a unique identifier, these can also be replaced into the original texts. Thus, a higher-level abstract representation of the documents were created, transforming each document to a set of normalized patterns.

THESIS 5:

I have applied the methods of distributional semantics to create a similarity measure between multiword terms and nouns and used it to create a hierarchical representation of the most relevant concepts in Hungarian clinical documents.

THESIS 5.a:

I created a method for automatically constructing a system of concepts that can be used as an aid in the creation of hand-made domain-specific resources and is flexible to be parametrized regarding its granularity.

THESIS 5.b:

I have shown that the resulting system of concepts is appropriate for creating an abstract representation of raw documents by mapping various occurrences of a cluster member to an identifier, thus resulting in documents containing a normalized set of patterns.

Related publications: 1, 4, 5, 11, 14

DOI:10.15774/PPKE.ITK.2015.012

DOI:10.15774/PPKE.ITK.2015.012

9

List of Papers

Journal publications

1 **Borbála Siklósi**, Attila Novák, Gábor Prószéky (2016): Context-aware correction of spelling errors in Hungarian medical documents, *Computer Speech & Language*, Vol.35, pp. 219-233, ISSN 0885-2308.

2 **Borbála Siklósi**, Attila Novák, György Orosz, Gábor Prószéky (2014): Processing noisy texts in Hungarian: a showcase from the clinical domain, *Jedlik Laboratories Reports*, Vol. II, no.3, pp. 5-62, ISSN 2064-3942

3 László János Laki, Attila Novák, **Borbála Siklósi**, György Orosz (2013): Syntax-based reordering in phrase-based English-Hungarian statistical machine translation. *International Journal of Computational Linguistics and Applications*, Vol. 4 no. 2. pp. 63–78, ISSN 0976-0962.

Book chapters

4 **Borbála Siklósi** (2015): Clustering Relevant Terms and Identifying Types of Statements in Clinical Records, In: A. Gelbukh (Ed.), *Lecture Notes in Computer Science Volume 9042: Computational Linguistics and Intelligent Text Processing* Springer International Publishing, Berlin Heidelberg. Part II pp. 619–630. ISBN 978-3-319-18116-5.

5 **Borbála Siklósi**, Attila Novák (2014): Identifying and Clustering Relevant Terms in Clinical Records Using Unsupervised Methods. In: Besacier, L.; Dediu, A.-H. and Martín-Vide, C. (Eds.), *Lecture Notes in Computer Science Volume 8791: Statistical Language and Speech Processing* Springer International Publishing, Berlin Heidelberg. pp. 233-243 ISBN 978-3-319-11396-8.

DOI:10.15774/PPKE.ITK.2015.012

6 **Borbála Siklósi**, Attila Novák, Gábor Prószéky (2013): Context-Aware Correction of Spelling Errors in Hungarian Medical Documents. In: Adrian-Horia Dediu, Carlos Martín-Vide, Ruslan Mitkov, Bianca Truthe (Eds.), *Lecture Notes in Computer Science Volume 7978: Statistical Language and Speech Processing* Springer, Berlin Heidelberg. pp. 248–259 ISBN 978-3-642-39592-5.

7 **Borbála Siklósi**, Attila Novák (2013): Detection and Expansion of Abbreviations in Hungarian Clinical Notes. In: F. Castro, A. Gelbukh, M.G. Mendoza (Eds.): *Lecture Notes in Computer Science, Vol. 8265: Advances in Artificial Intelligence and Its Applications.* Springer, Berlin Heidelberg. pp. 318–328. ISBN 978-3-642-45114-0

8 György Orosz, László János Laki, Attila Novák, **Borbála Siklósi** (2013): Improved Hungarian Morphological Disambiguation with Tagger Combination. In: Habernal, Ivan; Matousek, Vaclav (Eds.) *Lecture Notes in Computer Science, Vol. 8082: Text, Speech, and Dialogue* Pilsen, Czech Republic. Springer, Berlin Heidelberg. pp. 280–287. ISBN: 978-3-642-40584-6.

Conference proceedings

9 Novák Attila, **Siklósi Borbála** (2015): Automatic Diacritics Restoration for Hungarian. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics. pp. 2286–91.

10 **Siklósi Borbála**, Novák Attila (2015): Restoring the intended structure of Hungarian ophthalmology documents. BioNLP Workshop at the 53^{rd} Annual Meeting of the Association for Computational Linguistics, ACL 2015. Beijing, China, July 26-31, 2015

11 **Siklósi Borbála**, Novák Attila (2015): Nem felügyelt módszerek alkalmazása releváns kifejezések azonosítására és csoportosítására klinikai dokumentumokban. In: Tanács Attila, Varga Viktor, Vincze Veronika (szerk.) *XI. Magyar Számítógépes Nyelvészeti Konferencia.* Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 237-248

12 **Borbála Siklósi**, Attila Novák, Gábor Prószéky (2014): Resolving Abbreviations in Clinical Texts Without Pre-existing Structured Resources. In: *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM 2014).* Reykjavík

13 **Siklósi Borbála**, Novák Attila (2014): Rec. et exp. aut. Abbr. mnyelv. KLIN. szöv-ben – rövidítések automatikus felismerése és feloldása magyar nyelvű klinikai szövegekben. In: Tanács Attila, Varga Viktor, Vincze Veronika (szerk.) *X. Magyar Számítógépes Nyelvészeti Konferencia.* Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 167–176. ISBN 978-963-306-246-3

DOI:10.15774/PPKE.ITK.2015.012

14 **Siklósi Borbála**, Novák Attila (2014): A magyar beteg. In: Tanács Attila, Varga Viktor, Vincze Veronika (szerk.) *X. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 188–198. ISBN 978-963-306-246-3

15 **Siklósi Borbála**, Novák Attila, Prószéky Gábor (2013): Helyesírási hibák automatikus javítása orvosi szövegekben a szövegkörnyezet figyelembevételével. In: Tanács Attila; Vincze Veronika (szerk.) *A IX. Magyar Számítógépes Nyelvészeti Konferencia előadásai*. SZTE, Szeged. pp. 148–158 ISBN 978-963-306-189-3

16 **Borbála Siklósi**, György Orosz, Attila Novák, Gábor Prószéky (2012): Automatic structuring and correction suggestion system for Hungarian clinical records. In: *LREC-2012: SALTMIL-AfLaT Workshop on "Language technology for normalisation of less-resourced languages"*. Istanbul, Turkey, 2012. pp. 29–34

17 **Siklósi Borbála**, Orosz György, Novák Attila (2011): Magyar nyelvű klinikai dokumentumok előfeldolgozása. In: Tanács Attila, Varga Viktor, Vincze Veronika (szerk.) *VIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2011)*. Szegedi Tudományegyetem, pp. 143–340

18 Laki László, Novák Attila, **Siklósi Borbála** (2013): Hunglish mondattan – átrendezésalapú angol-magyar statisztikai gépifordító-rendszer. In: Tanács Attila; Vincze Veronika (szerk.) *A IX. Magyar Számítógépes Nyelvészeti Konferencia előadásai*. SZTE, Szeged. pp. 71–82 ISBN 978-963-306-189-3

19 György Orosz, László János Laki, Attila Novák, **Borbála Siklósi** (2013): Combining Language-Independent Part-of-Speech Tagging Tools. In: J. P. Leal, R. Rocha, and A. Simoes (Eds.), *2nd Symposium on Languages, Applications and Technologies*. Porto: Schloss Dagstuhl–Leibniz-Zentrum für Informatik. pp. 249–257 ISBN 978-3-939897-52-1

20 László János Laki, Attila Novák, **Borbála Siklósi** (2013): English-to-Hungarian Morpheme-based Statistical Machine Translation System with Reordering Rules. In: Marta R. Costa-jussa, Reinhard Rapp, Patrik Lambert, Kurt Eberle, Rafael E. Banchs, Bogdan Babych (Eds.) *Proceedings of the Second Workshop on Hybrid Approaches to Machine Translation (HyTra)*. Association for Computational Linguistics. pp. 42–50

DOI:10.15774/PPKE.ITK.2015.012

DOI:10.15774/PPKE.ITK.2015.012

List of Figures

DOI:10.15774/PPKE.ITK.2015.012

DOI:10.15774/PPKE.ITK.2015.012

List of Tables

DOI:10.15774/PPKE.ITK.2015.012

DOI:10.15774/PPKE.ITK.2015.012

Bibliography

Ahmad, K., Gillam, L., and Tostevin, L. (1999). University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER). In *The Eighth Text REtrieval Conference (TREC-8)*.

Barrows, J. R., Busuioc, M., and Friedman, C. (2000). Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. *Proceedings of the AMIA Annual Symposium*, pages 51–55.

Boswell, D. (2004). *CSE 256 (Spring 2004) Language Models For Spelling Correction*. PhD thesis.

Bourke, A., Dattani, H., and Robinson, M. (2004). Feasibility study and methodology to create a quality-evaluated database of primary care data. *Informatics in primary care*, 12(3):171–177.

Brill, E. and Moore, R. C. (2000). An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 286–293, Stroudsburg, PA, USA. Association for Computational Linguistics.

Brockett, C., Dolan, W. B., and Gamon, M. (2006). Correcting ESL Errors Using Phrasal SMT Techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 249–256, Sydney, Australia. Association for Computational Linguistics.

Carroll, J., Koeling, R., and Puri, S. (2012). Lexical acquisition for clinical text mining using distributional similarity. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 7182 of *Lecture Notes in Computer Science*, pages 232–246. Springer Berlin Heidelberg, Berlin, Heidelberg.

Chiang, M. F., Read-Brown, S., Tu, D. C., Choi, D., Sanders, D. S., Hwang, T. S., Bailey, S., Karr, D. J., Cottle, E., Morrison, J. C., Wilson, D. J., and Yackel, T. R. (2013). Evaluation of electronic health record implementation in ophthalmology at an academic medical center (an american ophthalmological society thesis). *Trans Am Ophthalmol Soc*, 111:70–92.

Church, K. W. and Gale, W. A. (1991). Probability scoring for spelling correction. *Statistics and Computing*, 1(2):93–103.

Cohen, T. (2003). Performance Metrics for Word Sense Disambiguation. In *Proceedings of the Australasian Language Technology Workshop (ALTW)*, pages 49–56, Melbourne, Australia.

Cohen, T. and Widdows, D. (2009). Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42(2):390 – 405.

Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraçlar, M., and Stolcke, A. (2007). Morph-based Speech Recognition and Modeling of Out-of-vocabulary Words Across Languages. *ACM Trans. Speech Lang. Process.*, 5(1):3:1–3:29.

Crowell, J., Zeng, Q., Ngo, L., and Lacroix, E. (2004). A frequency-based technique to improve the spelling suggestion rank in medical queries. *Journal of the American Medical Informatics Association*, 11(3):179–85.

Csendes, D., Csirik, J., and Gyimóthy, T. (2004). The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In Sojka, P., Kopecek, I., and Pala, K., editors, *Text, Speech, and Dialog*, volume 3206 of *Lecture Notes in Computer Science*, pages 19–23. Springer.

Dalianis, H., Hassel, M., and Velupillai, S. (2009). The Stockholm EPR Corpus - Characteristics and Some Initial Findings. In *Proceedings of ISHIMR 2009, Evaluation and implementation of e-health and health information initiatives: international perspectives. 14th International Symposium for Health Information Management Research*, pages 14–16, Kalmar, Sweden. Awarded best paper.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

DOI:10.15774/PPKE.ITK.2015.012

Ehsan, N. and Faili, H. (2013). Grammatical and context-sensitive error correction using a statistical machine translation framework. *Software – Practice and Experience*, 43(2):187–206.

Elliott, A., Davidson, A., Lum, F., Chiang, M., Saaddine, J. B., Zhang, X., Crews, J. E., and Chou, C.-F. (2012). Use of electronic health records and administrative data for public health surveillance of eye health and vision-related conditions. *Am J Ophthalmol*, 154(6 0):S63–S70.

Fábián, P. and Magasi, P. (1992). *Orvosi helyesírási szótár [Orthographic Dictionary of Hungarian Medical Language]*. Akadémiai Kiadó, Budapest.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1952-59:1–32.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Friedman, C., Johnson, S., Forman, B., and Starren, J. (1995). Architectural requirements for a multipurpose natural language processor in the clinical environment. *Proc Annu Symp Comput Appl Med Care*, pages 347–51.

Gaizauskas, R., Demetriou, G., and Humphreys, K. (2000). Term Recognition and Classification in Biological Science Journal Articles. In *Proc. of the Computional Terminology for Medical and Biological Applications Workshop of the 2 nd International Conference on NLP*, pages 37–44.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Harris, Z. (2002). The structure of science information. *Journal of Biomedical Informatics*, 35(4):215–221.

Hassel, M., Henriksson, A., and Velupillai, S. (2011). Something Old, Something New : Applying a Pre-trained Parsing Model to Clinical Swedish. In *18th Nordic Conference of Computational Linguistics NODALIDA 2011*. Northern European Association for Language Technology (NEALT).

Henriksson, A. (2013). *Semantic Spaces of Clinical Text : Leveraging Distributional Semantics for Natural Language Processing of Electronic Health Records*. Number 13-009 in DSV Report Series. Stockholm University, Department of Computer and Systems Sciences.

Hindle, D. (1990). Noun Classification from Predicate-argument Structures. In *Proceedings of the 28th Annual Meeting on Association for Computational Linguistics*, ACL '90, pages 268–275, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ittoo, A. and Bouma, G. (2013). Term extraction from sparse, ungrammatical domain-specific documents. *Expert Syst. Appl.*, 40(7):2530–2540.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180, Prague. Association for Computational Linguistics.

Kukich, K. (1992). Techniques for Automatically Correcting Words in Text. *ACM Comput. Surv.*, 24(4):377–439.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240.

Levenshtein, V. (1965). Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1:8–17.

DOI:10.15774/PPKE.ITK.2015.012

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics - Volume 2*, COLING '98, pages 768–774, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2014a). Biomedical Terminology Extraction: A new combination of Statistical and Web Mining Approaches. In *JADT'2014 : Journées internationales d'Analyse statistique des Données Textuelles*, pages 421–432, Paris, France.

Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2014b). Yet Another Ranking Function for Automatic Multiword Term Extraction. In *9th International Conference on Natural Language Processing, PolTAL'14*, volume 8686, pages 52–64. Springer.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, 28(2):203–208.

Meystre, S., Savova, G., Kipper-Schuler, K., and Hurdle, J. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, 35:128–44.

Nagy T., I., Vincze, V., and Berend, G. (2011). Domain-Dependent Identification of Multiword Expressions. In Angelova, G., Bontcheva, K., Mitkov, R., and Nicolov, N., editors, *Recent Advances in Natural Language Processing, {RANLP} 2011*, pages 622–627, Hissar, Bulgaria. RANLP 2011 Organising Committee.

Nasiruddin, M. (2013). A State of the Art of Word Sense Induction: A Way Towards Word Sense Disambiguation for Under-Resourced Languages. *CoRR*, abs/1310.1425.

Navigli, R. (2012). A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches. In *Proceedings of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM)*, pages 115–129, Spindleruv Mlyn, Czech Republic.

Noeman, S. and Madkour, A. (2010). Language independent transliteration mining system using finite state automata framework. In *Proceedings of the 2010 Named Entities Workshop*, NEWS '10, pages 57–61, Stroudsburg, PA, USA. Association for Computational Linguistics.

Novák, A. (2003). Milyen a jó Humor? [What is good Humor like?]. In *I. Magyar Számítógépes Nyelvészeti Konferencia*, pages 138–144, Szeged. SZTE.

Oflazer, K. and Güzey, C. (1994). Spelling correction in agglutinative languages. In *Proceedings of the fourth conference on Applied Natural Language Processing*, ANLC '94, pages 194–195, Stroudsburg, PA, USA. Association for Computational Linguistics.

Orosz, G. and Novák, A. (2012). PurePos – an open source morphological disambiguator. In Sharp, B. and Zock, M., editors, *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science*, pages 53–63, Wroclaw.

Orosz, G., Novák, A., and Prószéky, G. (2013). Hybrid text segmentation for Hungarian clinical records. In Castro, F., Gelbukh, A., and González, M., editors, *Advances in Artificial Intelligence and Its Applications*, volume 8265 of *Lecture Notes in Computer Science*, pages 306–317. Springer Berlin Heidelberg, Heidelberg.

Orosz, G., Novák, A., and Prószéky, G. (2014). Lessons learned from tagging clinical Hungarian. *International Journal of Computational Linguistics and Applications*, 5(1):159–176.

Pakhomov, S. (2002). Semi-Supervised Maximum Entropy based Approach to Acronym and Abbreviation Normalization in Medical Texts. In Isabelle, P., editor, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 160–167, Philadelphia, USA. Rochester, NY: ACL Press.

Pakhomov, S., Pedersen, T., and Chute, C. (2005). Abbreviation and Acronym Disambiguation in Clinical Discourse. In Friedman, C., Ash, J., and Tarczy-Hornoch, P., editors, *Proceedings of the AMIA Annual Symposium*, pages 589–593, Washington DC, USA. Bethesda, MD: AMIA Press.

DOI:10.15774/PPKE.ITK.2015.012

Pantel, P. (2005). Inducing Ontological Co-occurrence Vectors. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 125–132, Stroudsburg, PA, USA. Association for Computational Linguistics.

Park, Y. A. and Levy, R. (2011). Automated whole sentence grammar correction using a noisy channel model. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1 of *HLT '11*, pages 934–944, Stroudsburg, PA, USA. Association for Computational Linguistics.

Patel, V. L., Arocha, J. F., and Kushniruk, A. W. (2002). Patients' and Physicians' Understanding of Health and Biomedical Concepts: Relationship to the Design of EMR Systems. *J. of Biomedical Informatics*, 35(1):8–16.

Patrick, J. and Nguyen, D. (2011). Automated Proof Reading of Clinical Notes. In Gao, H. H. and Dong, M., editors, *PACLIC*, pages 303–312. Digital Enhancement of Cognitive Development, Waseda University.

Pedersen, T., Pakhomov, S. V. S., Patwardhan, S., and Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288 – 299.

Pirinen, T. A. and Lindén, K. (2010). Finite-State Spell-Checking with Weighted Language and Error Models. In *Proceedings of the Seventh SaLTMiL workshop on creation and use of basic lexical resources for less-resourced languagages*, pages 13–18, Valletta, Malta.

Pirk, E. (2013). *Névkifejezések automatikus felismerése orvosi szovegekben.* MSc Thesis, Pázmány Péter Catholic University, Budapest.

Prószéky, G. and Kis, B. (1999). A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 261–268, College Park, Maryland. Association for Computational Linguistics.

Redd, T. K., Read-Brown, S., Choi, D., Yackel, T. R., Tu, D. C., and Chiang, M. F. (2014). Electronic health record impact on productivity and efficiency in an academic pediatric ophthalmology practice. *Journal of AAPOS*, 18(6):584–589.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual Correlates of Synonymy. *Commun. ACM*, 8(10):627–633.

Sager, N., Lyman, M., Bucknall, C., Nhan, N., and Tick, L. J. (1994). Natural Language Processing and the Representation of Clinical Data. *Journal of the American Medical Informatics Association*, 1(2).

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., and Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA*, 17(5):507–513.

Schütze, H. (1993). Word Space. In Giles, L. C., Hanson, S. J., and Cowan, J. D., editors, *Advances in Neural Information Processing Systems 5*, pages 895–902. San Francisco, CA: Morgan Kaufmann.

Schütze, H. and Pedersen, J. (1995). Information Retrieval based on Word Senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, USA.

Siklósi, B. and Novák, A. (2013). Detection and Expansion of Abbreviations in Hungarian Clinical Notes. In Castro, F., Gelbukh, A., and González, M., editors, *Advances in Artificial Intelligence and Its Applications*, volume 8265 of *Lecture Notes in Computer Science*, pages 318–328. Springer Berlin Heidelberg, Heidelberg.

Siklósi, B., Novák, A., and Prószéky, G. (2013). Context-Aware Correction of Spelling Errors in Hungarian Medical Documents. In Dediu, A.-H., Martín-Vide, C., Mitkov, R., and Truthe, B., editors, *Statistical Language and Speech Processing*, volume 7978 of *Lecture Notes in Computer Science*, pages 248–259. Springer Berlin Heidelberg.

DOI:10.15774/PPKE.ITK.2015.012

Siklósi, B., Novák, A., and Prószéky, G. (2014). Resolving abbreviations in clinical texts without pre-existing structured resources. In *Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing, LREC 2014*.

Siklósi, B., Orosz, G., Novák, A., and Prószéky, G. (2012). Automatic structuring and correction suggestion system for Hungarian clinical records. In De Pauw, G., De Schryver, G.-M., Forcada, M., M. Tyers, F., and Waiganjo Wagacha, P., editors, *8th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages*, pages 29–34, Istanbul, Turkey.

Sokal, R. R. and Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, 11(2):33–40.

Sridharan, S. and Murphy, B. (2012). Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon. pages 53–68, Mumbai, India. The COLING 2012 Organizing Committee.

Stolcke, A., Zheng, J., Wang, W., and Abrash, V. (2011). SRILM at Sixteen: Update and Outlook. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Waikoloa, Hawaii.

Süveges, I. (2010). *Szemészet*. Medicina Könyvkiadó Zrt.

Turchin, A., Chu, J. T., Shubina, M., and Einbinder, J. S. (2007). Identification of Misspelled Words without a Comprehensive Dictionary Using Prevalence Analysis. *AMIA Annual Symposium proceedings*, 2007:751–755.

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Vincze, V. (2013). Domének közti hasonlóságok és különbségek a szófajok és szintaktikai viszonyok eloszlásában. In *IX. Magyar Számítógépes Nyelvészeti Konferencia*, pages 182–192, Szeged. Szegedi Tudományegyetem.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244.

Wittgenstein, L. (1953). *Philosophical Investigations*. Basil Blackwell, Oxford.

Wu, Y., Denny, J. C., Rosenbloom, S. T., Miller, R. A., Giuse, D. A., and Xu, H. (2012). A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. *Proceedings of the AMIA Annual Symposium*, 2012:997–1003.

Xu, H., Stetson, P., and Friedman, C. (2007). A Study of Abbreviations in Clinical Notes. In Teich, J., Suermondt, J., and Hripcsak, G., editors, *Proceedings of the AMIA Annual Symposium*, pages 821–825, Washington DC, USA. Bethesda, MD: AMIA Press.

Xu, H., Stetson, P., and Friedman, C. (2009). Methods for Building Sense Inventories of Abbreviations in Clinical Notes. *Journal of American Medical Informatics Association*, 16(1):103–108.

Zhang, J. (2002). Representations of health concepts: a cognitive perspective. *Journal of Biomedical Informatics*, 35(1):17 – 24.

Zhang, Z., Iria, J., Brewster, C., and Ciravegna, F. (2008). A Comparative Evaluation of Term Recognition Algorithms. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 2108–2113, Marrakech, Morocco. European Language Resources Association (ELRA).