Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 35 (2016) 219–233

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Context-aware correction of spelling errors in Hungarian medical documents☆

Borbála Siklósi[a,*], Attila Novák[a,b,**], Gábor Prószéky[a,b]

[a] *Pázmány Péter Catholic University, Faculty of Information Technology and Bionics, 50/a Práter Street, 1083 Budapest, Hungary*
[b] *MTA-PPKE Hungarian Language Technology Research Group, 50/a Práter Street, 1083 Budapest, Hungary*

Received 8 November 2013; received in revised form 3 July 2014; accepted 6 September 2014
Available online 16 September 2014

Abstract

Owing to the growing need of acquiring medical data from clinical records, processing such documents is an important topic in natural language processing (NLP). However, for general NLP methods to work, a proper, normalized input is required. Otherwise the system is overwhelmed by the unusually high amount of noise generally characteristic of this kind of text. The different types of this noise originate from non-standard language use: short fragments instead of proper sentences, usage of Latin words, many acronyms and very frequent misspellings.

In this paper, a method is described for the automated correction of spelling errors in Hungarian clinical records. First, a word-based algorithm was implemented to generate a ranked list of correction candidates for word forms regarded as incorrect. Second, the problem of spelling correction was modelled as a translation task, where the source language is the erroneous text and the target language is the corrected one. A Statistical Machine Translation (SMT) decoder performed the task of error correction. Since no orthographically correct proofread text from this domain is available, we could not use such a corpus for training the system. Instead, the word-based system was used to create translation models. In addition, a 3-gram token-based language model was used to model lexical context. Due to the high number of abbreviations and acronyms in the texts, the behaviour of these abbreviated forms was further examined both in the case of the context-unaware word-based and the SMT-decoder-based implementations.

The results show that the SMT-based method outperforms the first candidate accuracy of the word-based ranking system. However, the normalization of abbreviations should be handled as a separate task.

Keywords: Spelling correction; Medical text processing; Agglutinating languages

1. Introduction

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

☆ This paper has been recommended for acceptance by R.K. Moore.

* Corresponding author. Tel.: +36 303346802.
** Corresponding author at: Pázmány Péter Catholic University, Faculty of Information Technology and Bionics, 50/a Práter Street, 1083 Budapest, Hungary.
E-mail addresses: siklosi.borbala@itk.ppke.hu (B. Siklósi), novak.attila@itk.ppke.hu (A. Novák), proszeky.gabor@itk.ppke.hu (G. Prószéky).

http://dx.doi.org/10.1016/j.csl.2014.09.001
0885-2308/

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.

As Meystre et al. (2008) point out, it is crucial to distinguish between clinical and biomedical texts. Biomedical text processing methods, which are developed for handling published documents consider well-formed, proofread texts as their object, while this paper concentrates on clinical texts written by clinicians in clinical settings. Owing to the radical differences between these two types of textual input, methods of biomedical text processing cannot be applied to our corpus.

One of the earliest studies in processing clinical narratives, also mentioned in the report 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 fulfill the expectations of extracting trustworthy information and revealing extended knowledge as well as deeper relations found in these texts. All these methods rely on proper, well-formed, and correct input documents.

In Hungarian hospitals, however, 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 purpose of this paper is to present methods that can create a normalized representation of raw Hungarian clinical documents.

We investigated anonymized clinical records of a Hungarian clinic. 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.

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 our work, we 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. Moreover, a language model can be used in order to model lexical context, which is of crucial importance when choosing the appropriate item from the list of correction candidates. In a traditional SMT setup, a translation model is built from parallel training corpora and a language model from a target-language-side monolingual corpus. Since no such training data is available in our case, the translation model is replaced by a ranked list of correction candidates. These are produced by a hybrid system based on a rule-based morphological analyzer and several general and domain-specific statistical models.

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 our 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 we 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 Section 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.

Beside spelling errors, clinical records are also characterized by a high ratio of abbreviated forms. The use of some of these abbreviations follows some standard rules, but most of them are used in an arbitrary manner. This great variation in their written forms is caused by non-standard usage and by misspellings. Thus 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. Even then, simply matching these abbreviations to a lexicon is not satisfactory. Moreover, due to the special notational language (Barrows et al., 2000) often used in clinical settings where full statements are written using only, or mostly abbreviated forms, the lexical context of an abbreviation is not of much help.

In this paper, we present a method for considering textual context when recognizing and correcting spelling errors. Our system applies methods of SMT, based on a word-based system for generating correction candidates. After describing the characteristics of the clinical records and the language they were written in, first the context-unaware word-based approach is described for generating correction suggestions, then its integration into an SMT framework is presented. We show that our 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.

2. Language- and domain-specific difficulties

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).

Table 1
The distribution and ranking of part-of-speech in the clinical corpus (CLIN) and the general Szeged Corpus (SZEG).

	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%
CLIN	1	2	3	4	5	6	7	8	9
SZEG	1	3	8	2	5	7	4	9	6

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 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 1 shows the detailed statistics and ranking of pos tags in the two corpora.

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. Latin medical terms usually have a standard form based on both Latin and Hungarian orthography. However what we find in the documents is often an inconsistent mixture of the two (e.g. *tensio/tenzio/tensió/tenzió* ‘tension’). Even though the spelling of these forms is standardized, doctors tend to develop their own habits which they use inconsistently. Another difficulty is the complete lack of correctly written clinical documents that could be used for creating appropriate language and error models.

Table 2
Corpus frequencies of some variations for abbreviating the three phrases *oculus sinister*, *oculus dexter* and *oculi utrisque*, which are the three most frequent abbreviated phrases.

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

Compared to the Szeged Corpus, the ratio of misspelled words was 0.27% in the general Hungarian texts, while 8.44% in the clinical notes. Moreover, the general corpus has several subcorpora, including one of primary school essays, which still had only an 0.87% error rate.

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*.

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 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 Siklósi and Novák (2013).

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.

3. Automatic spelling correction

3.1. 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 Fig. 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

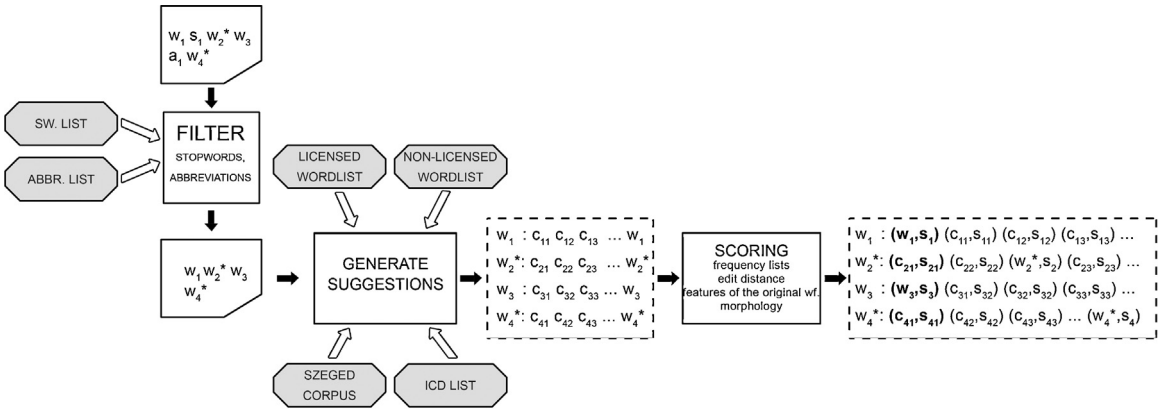

Fig. 1. The word-based system (w's stand for words, a's for abbreviations, c's are correction candidates and (c, s)'s are correction candidate, score pairs. Misspelled words are signed with an asterisk.)

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. 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 Fig. 1).

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

3.1.1. 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.
- *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 our 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, [1], 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

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

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. We assumed that both corpora contained only correct word forms.

3.1.2. 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. 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.
- *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*).

3.2. Application of statistical machine translation

When generating correction suggestions, the word-based system ignores the lexical context of the words to be corrected. Since our 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. To improve the accuracy of the system, lexical context also needs to be considered. To satisfy these two requirements, we applied Moses (Koehn et al., 2007), a widely used statistical machine translation (SMT) toolkit. During "translation", we 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 our 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).

$$\hat{C} = \operatorname{argmax} P(C|E) = \operatorname{argmax} \frac{P(E|C)P(C)}{P(E)} \tag{1}$$

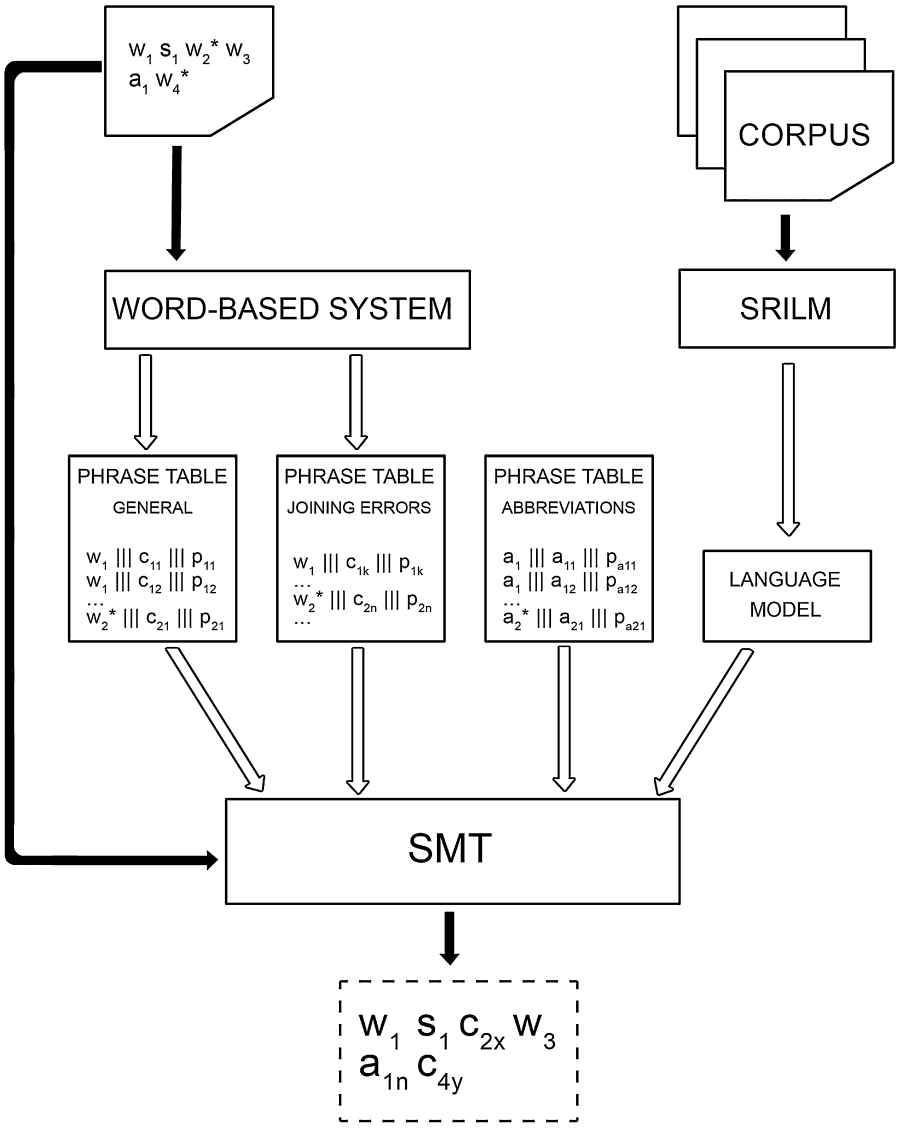

Fig. 2. The context-aware SMT-based system.

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

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 our 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 SRILM toolkit (Stolcke et al., 2011). Fig. 2 shows the process of correcting documents by the context-aware system.

3.2.1. 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, we describe each of these models.

Table 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)	$P(c\|e)$
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

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

Table 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. Our 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 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 our 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. Since the correction process is carried out word-by-word, the method for joining two erroneously split words is not

Table 4
Extract from the translation model for multiword errors.

Original form (e)	Correction candidate (c)	$P(c\|e)$
soronkívül	soron kívül	0.02074
soronkívül	soronkívül	0.01459

implemented (though theoretically available in the system), but the occurrence of such errors is (about six times) less frequent than the other way round.

3.2.2. 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 our corpus were the development and test sets, such a model could not be built. Though there are orthographically correct texts of other, mostly general domains, the *n*-gram statistics of these would not correspond to the characteristics of the clinical domain due to the differences described in Section 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.

We 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.

3.2.3. Decoding

The result of formula (1) is determined by the decoding algorithm of the SMT system based on the above models. To carry out decoding, we used the widely-used Moses toolkit (Koehn et al., 2007). The parameters of decoding can be set in the Moses configuration file, thus they can be changed easily in order to adapt the system to new circumstances and weighting schemes. 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.
- *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, word order changes are not allowed in our application, since modifications can only occur within words or by splitting some words. The structure of the sentence cannot be changed. Thus a 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.)

3.3. 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

Table 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.

	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

each department. Therefore our 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 (17 243 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 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. Results

Two aspects of the performance of our systems were evaluated: (i) error detection and (ii) error correction. Error detection was measured by the common metric of F-measure, according to formula (2).

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

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 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 achieve a higher value for recall (60.24%), the F-measure emphasizing precision was still 41.45% for error detection. Considering the task of error correction, 78.00% of the erroneous words were changed to the correct form in this case.

Table 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	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, fb: 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.

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 Section 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 Section 2 concerning the huge differences between the language of our corpus of medical records and general Hungarian.

The explanation of the relatively low performance measures for error detection and a deeper error analysis is presented in Section 5 below.

5. Discussion

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 6 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, we had to accept some non-standard forms that were consistently used throughout the whole corpus, without the standard form appearing at all. We 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.

Table 7
Examples for the transformation of a correct sentence (ORIG) to another correct sentence with very similar meaning, but different words (SMT).

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)*

5.1. 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 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 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.

5.2. 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 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.

5.3. 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 9.

Table 8
Some examples for words corrected properly by both systems.

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 9
Some examples for words corrected properly or untouched by the SMT system, but altered incorrectly by the word-based system.

Original form	Word-based	SMT	Gold standard	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)

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 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.

6. Conclusion

In our paper, we presented 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 a lot of abbreviations. The aim of this research was to create a system optimized for this specific domain, thus we did not intend to create a general error correction system, rather a normalizer tool for Hungarian clinical records by considering the specific characteristiscs of the language used in these documents. Besides applying morphological rules and statistics on the word level, lexical context is also considered during the correction process. Due to the lack of a corpus normalized to proper standard orthography, a practical goal in our work was to consider frequently used word forms as a quasi-standard. Applying our method to raw clinical free-text data, a normalized representation can be achieved that is of crucial importance for further processing steps. We showed that applying an SMT framework as a spelling correction system is appropriate and can achieve high accuracy compared to a context-unaware word-based method of generating a ranked list of correction candidates for each word separately.

Acknowledgements

The authors would like to thank Nóra Wenszky for her useful comments and suggestions, which helped to improve on the clarity and the English language quality of the article.

This research was partially supported by the project grants TÁMOP–4.2.1./B–11/2-KMR-2011-0002 and TÁMOP–4.2.2./B–10/1-2010-0014.

References

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

Boswell, D., 2004. CSE 256 (Spring 2004) Language Models for Spelling Correction.

Brill, E., 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, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 286–293.

Brockett, C., Dolan, W.B., 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, Association for Computational Linguistics, Sydney, Australia, pp. 249–256.

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

Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraçlar, M., Stolcke, A., 2007. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. ACM Trans. Speech Lang. Process. (TSLP) 5, 3.

Crowell, J., Zeng, Q., Ngo, L., Lacroix, E., 2004. A frequency-based technique to improve the spelling suggestion rank in medical queries. J. Am. Med. Inform. Assoc. 11, 179–185.

Csendes, D., Csirik, J., Gyimóthy, T., 2004. The Szeged Corpus: a POS tagged and syntactically annotated Hungarian natural language corpus. In: Sojka, P., Kopecek, I., Pala, K. (Eds.), Text, Speech and Dialogue: 7th International Conference (TSD). Springer, pp. 41–48.

Ehsan, N., Faili, H., 2013. Grammatical and context-sensitive error correction using a statistical machine translation framework. Softw. Pract. Exp. 43, 187–206.

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

Friedman, C., Johnson, S., Forman, B., Starren, J., 1995. Architectural requirements for a multipurpose natural language processor in the clinical environment. Proc. Annu. Symp. Comput. Appl. Med. Care, 347–351.

Harris, Z.S., 2002. The structure of science information. J. Biomed. Inform. 35, 215–221.

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., Herbst, E., 2007. Moses: open source toolkit for statistical machine translation. In: Proceedings of the ACL 2007 Demo and Poster Sessions, Association for Computational Linguistics, Prague, pp. 177–180.

Kukich, K., 1992. Techniques for automatically correcting words in text. ACM Comput. Surv. 24, 377–439.

Levenshtein, V., 1965. Binary codes capable of correcting spurious insertions and deletions of ones. Probl. Inf. Transm. 1, 8–17.

Meystre, S., Savova, G., Kipper-Schuler, K., Hurdle, J., 2008. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb. Med. Inform. 35, 128–144.

Noeman, S., Madkour, A., 2010. Language independent transliteration mining system using finite state automata framework. In: Proceedings of the 2010 Named Entities Workshop, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 57–61.

Novák, A., 2003. What is good Humor like? In: I. Magyar Számítógépes Nyelvészeti Konferencia. SZTE, Szeged, pp. 138–144.

Oflazer, K., Güzey, C., 1994. Spelling correction in agglutinative languages. In: Proceedings of the Fourth Conference on Applied Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 194–195.

Orosz, G., Novák, A., Prószéky, G., 2013. Hybrid text segmentation for Hungarian clinical records. In: Vol. 8265 of Lecture Notes in Artificial Intelligence. Springer-Verlag, Heidelberg, pp. 306–317.

Park, Y.A., 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 – Vol. 1, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 934–944.

Patrick, J., Nguyen, D., 2011. In: Gao, H.H., Dong, M. (Eds.), PACLIC, Digital Enhancement of Cognitive Development. Waseda University, pp. 303–312.

Pirinen, T.A., 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 Languages, Valletta, Malta, pp. 13–18.

Prószéky, G., 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, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 261–268.

Sager, N., Lyman, M., Bucknall, C., Nhan, N., Tick, L.J., 1994. Natural language processing and the representation of clinical data. J. Am. Med. Inform. Assoc., 1.

Siklósi, B., Novák, A., 2013. Detection and expansion of abbreviations in Hungarian clinical notes. In: Vol. 8265 of Lecture Notes in Artificial Intelligence. Springer-Verlag, Heidelberg, pp. 318–328.

Siklósi, B., Orosz, G., Novák, A., Prószéky, G., 2012. Automatic structuring and correction suggestion system for Hungarian clinical records. In: in: 8th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, pp. 29–34.

Siklósi, B., Novák, A., 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.

Stolcke, A., Zheng, J., Wang, W., Abrash, V., 2011. SRILM at sixteen: update and outlook. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, Waikoloa, Hawaii.

Turchin, A., Chu, J.T., Shubina, M., Einbinder, J.S., 2007. Identification of misspelled words without a comprehensive dictionary using prevalence analysis. AMIA Annual Symposium Proceedings 2007, 751–755.

Vincze, V., 2013. Domének közti hasonlóságok és különbségek a szófajok és szintaktikai viszonyok eloszlásában. In: IX. Magyar Számítógépes Nyelvészeti Konferencia, pp. 182–192.