Morphological analysis of a non-standard language variety

Heiki-Jaan Kaalep
University of Tartu
Heiki-Jaan.Kaalep@ut.ee

Kadri Muischnek
University of Tartu
Kadri.Muischnek@ut.ee

Abstract

This article introduces a corpus-based method for improving the process of automatic morphological analysis of a non-standard text variety. More precisely, our paper is concerned with the morphological analysis of Estonian chatroom texts. First, the morphological analyzer designed for the standard written Estonian is used for the analysis of chatroom texts. On the basis of output error analysis a method for improving the process is proposed. We take advantage of the fact that there are deviations with high token frequency, but low type frequency, on the one hand, and deviations with low token frequency, but high type frequency, on the other hand. The first group has to be manually compiled into a user lexicon, whereas the second group of errors can be taken care of by automatic means: automatic preprocessing of texts and automatic complementation of the user lexicon. As a result, the percentage of unknown tokens in the output of the morphological analyzer decreases from 27 to 10.5.

1 Introduction

Recently new text types have emerged where the language and orthography used differ considerably from the conventions of the standard written language. E-mails, chatroom texts, internet forums and blogs are some of the examples of these new text types. The expressions ‘language of the computer-mediated communication’, ‘Internet language’ or ‘Netspeak’ (e.g. Crystal 2001) are used to designate all of them.

Morphological analysis is an inevitable prerequisite for any kind of further automatic analysis of a morphologically complex language like Estonian. For standard written Estonian the automatic morphological analysis is nowadays more or less a solved problem (Kaalep, Vaino 2001), but the language and orthography of the aforementioned new text types differs considerably from that of the standard written Estonian and thus the morphological analyzer can be expected to perform poorly while analyzing the Internet language texts.

In this article we report on the experiments on the morphological analysis of the Estonian chatroom (or Internet Relay Chat, IRC) texts. We present the results of the automatic morphological analysis using a tagger meant for the standard written language, group and analyze the errors and introduce a corpus-based method for compiling a customized lexicon for the analyzer for the text currently at hand.

Estonian language belongs to the Finnic group of the Finno-Ugric language family. Typologically Estonian is an agglutinating language but more fusional and analytic than the languages belonging to the northern branch of the Finnic languages. Written Estonian uses phonemic orthography. One can find a detailed description of the grammatical system of Estonian in (Erelt 2003).

The rest of this article is structured as follows. Section 2 gives an overview of the material we have used: the corpus of Estonian chatroom texts and the morphological analyzer *etmrf*. Section 3 presents some related research on the linguistic properties and automatic morphological analysis of Internet language varieties. In Section 4 we analyze the main deviations of the language used in Estonian chatrooms from the standard written Estonian and in Section 5 we put forward different strategies for coping with these deviations, namely preprocessing of texts prior to the morphological analysis and compiling a user lexicon. In Section 6 we present the results of running the morphological analyzer together with preprocessor and user lexicon; in Section 7 we mention some perspectives for the future research and Section 8 concludes.

2 Material: corpus of chatroom texts and the morphological analyzer *etmrf*

For our experiments we used a corpus of Estonian chatroom texts from 2003 and 2006[1] consisting of ca 7 million tokens annotated according to the TEI P5 guidelines.[2] The annotation of the corpus dwells from the opinion that the chatroom conversation is like a kind of staged drama text: there are actors who enter the stage, produce their lines, and leave. All chatters i.e. their nicknames are annotated with the tag <speaker> and the text produced by the chatter is annotated with the paragraph tag <p>. The nicknames have not been changed in any way for anonymisation, but e-mail addresses, URLs and phone numbers have been masked (by substituting parts with ‘xxxx’).

These chatrooms are an environment for general leisurely chatting. The chatters try to be witty. Language play, including play with orthography and nicknames, is an integral part of this type of communication.

The sentences are not annotated. A sentence splitter developed for the standard written Estonian fails to find sentence boundaries in the chatroom texts, as a typical sentence there does not begin with a capital letter nor end with a punctuation mark. But the text entered by one chatter at one time and annotated as a paragraph is typically short and can be treated as one sentence for the morphological analysis.

For the morphological analysis we used *etmrf*, a tool developed by Filosoft Ltd; the demo version of the program can be found at www.filosoft.ee. *Etmrf* can be used both as a morphological analyzer and as a morphological disambiguator; we performed only morphological analysis.

Etmrf is a convenient tool for our purposes as its behaviour as a morphological analyzer can be modified with a customized user lexicon – a text file that contains word-forms and their preferred analyses. While analyzing a word-form in the text with a user lexicon, *etmrf* first turns to the user lexicon and only in the case the word-form is not present there, the ordinary process of morphological analysis starts. Thus the user lexicon enables us to give analysis to the word-forms not present in the standard written language, or give alternative readings to word-forms which usage in chatroom texts differs from that of the standard written language, e.g. representing a different part of speech.

3 Internet language

From the linguistic and sociolinguistic point of view, David Crystal (2001) gives a comprehensive overview of the language used in the computer-mediated communication and some of its subtypes. He characterizes the language used on the Internet as identical neither to speech nor writing, but selectively and adaptively displaying properties of both (Crystal 2001: 47).

Crystal has a separate chapter on the language of the chatgroups, that is a generic term he uses for chatrooms, newsgroups, mailing lists and other multi-participant electronic discourse, whether real-time or not (Crystal 2001: 129).

Among the distinctive features of the language used in computer-mediated communication Crystal mentions distinctive graphology, especially the strong tendency to use lowercase everywhere, minimalist punctuation and multiplying of vowels and consonants to express the “ferocity” of the expression. He also notes that chatgroups make a great deal of use of phonetic spelling; and presents a long list of genre-specific abbreviations (2001: 85-86) compiled of the initial letters of the words in a phrase (e.g. *btw* ‘by the way’, *cu* ‘see you’).

In an overview of the Internet language, Naomi S. Baron (2003) lists the usage of emoticons, abbreviations and acronyms as distinctive features of the computer-mediated communication.

Lari Kotilainen (2002) analyzes the usage of English words and phrases in Finnish chatrooms. Finnish chatters like to use phonetic spelling while writing English text (e.g. *how aar juu*). Kotilainen shows that English in Finnish chatrooms is used mostly in the form of fixed expressions.

Mark Myslín and Stephan Th. Gries (2010) conduct a study of Spanish Internet orthography. They conclude that the spelling used by the “speakers” of Internet Spanish reflects two interrelated rules: ‘modify words

[1] The corpus is available at http://www.cl.ut.ee/korpused/segakorpus/uusmeedia/jututoad.php?lang=en

[2] http://www.tei-c.org/Guidelines/P5/

that have special pragmatic functions and if you are really determined to modify a common word then make big/several changes'.

Vincent Ooi (2002) has done work on the morphological analysis of English Internet Relay Chat texts. The author experiments with two taggers, namely CLAWS and AUTASYS, both probabilistic taggers.

Ooi notes the frequent usage of discourse particles (which he calls discourse markers) in the genre of IRC. Analyzing the output and listing frequent errors he mentions that the taggers he used could not handle emoticons or extra-linguistic acts like *lol* – an abbreviation for laughing out loud. Other frequent errors were caused by proper nouns beginning with a lower-case letter and therefore not recognized as proper nouns, non-standard spelling and using digits for syllables, e.g. *2* for *too*. Ooi concludes that clearly the lexicon of the part-of-speech tagger needs to be modified in order to handle the language of the computer-mediated communication.

We are not aware of any previous work, resulting in a large morphologically tagged corpus of internet language. Experiments have either led to an analysis of errors, as in (Ooi 2002), or to a manually corrected small corpus, as in (Forsyth, Martell 2007).

4 Main deviations from standard literary language

As an initial experiment, we performed morphological analysis of the chatroom texts using simply the morphological analyzer *etmrf* as it is. In the output text, 27% of the tokens were tagged as unknown words. This is much worse than the 2% reported for *estmorf* (a previous implementation of *etmrf*) for standard literary Estonian by (Kaalep, Vaino 2001).

By and large, our chatrooms show the same types of deviation from the standard written form as do other languages (cf Section 3).

4.1 Parts of Speech

4.1.1 Discourse particles

Discourse particles are a part of speech widely recognized in spoken language, but not present in the word-class system of *etmrf*. Analyzing the frequent word-forms that had received the label of an unknown token from *etmrf*, it was clear that we should follow the word-class system used for analyzing the spoken language (e.g. Hennoste et. al. 2002) and create a special part-of-speech tag for discourse particles, i.e. these short or shortened word-forms that often constitute a clause alone and if used in a clause are not syntactically part of it. They have mostly no clear semantic content but a pragmatic (interactional or emotive) function, e.g. *tre* (a shortened form of *tere*) 'hello' or *kle* (a shortened form of *kuule*) 'hey, listen'. Discourse particles are a frequent part of speech in chatroom text making up 5.8% of the tokens.

```
(1)kle   krizzy   mis  teed
particle propname what do-
2.ps.sg
‘hey krizzy, what are you
doing?’
```

4.1.2 Emoticons

Emoticons are iconic signs combined of punctuation marks that are used for expressing emotions, e.g. :P , :-) , :D. As emoticons contribute a certain meaning or a meaning nuance to the text, just like words do, it would be reasonable to analyze them as words and give them a word-class tag. If analyzed as word-forms, emoticons make up 3.2% of the tokens in chatroom texts.

4.2 Orthography

In languages with non-phonetic spelling like English and French, chatgroups make a great deal of use of phonetic spelling.

Estonian spelling, in contrast, is very close to phonetic. Still, dropping h from the beginning of a word (e.g. *ommik* for *hommik* 'morning') might be called an instance of pronunciation affecting spelling as the word-initial *h* is not articulated in spoken Estonian.

4.2.1 Character substitution

This is a wide-spread phenomenon in chatroom texts. Frequent substitutions include using *ff* for *hv* (e.g. *raffas* pro *rahvas* 'people'), *x* for *ks* (e.g. *näitex* pro *näiteks* 'for example'), *y* for *ü* (e.g. *kyll* pro *küll* 'enough'), *c* for *ts* (e.g. *täica* pro *täitsa* 'entirely'), *2* for *ä* (e.g. *h2sti* pro *hästi* 'well'), *6* for *õ* (e.g. *h6be* pro *hõbe* 'silver') and *8* for *ö* (e.g. *t88* for *töö* 'work').

The tradition of substituting non-ASCII characters like *äöüõ* with some other symbol has originally arisen from the wide usage of non-customized keyboards, but this need has largely disappeared as we can see for example from spellings like *ykskõik* pro *ükskõik* 'no matter'.

A question that this material brings forth is: what is the meaning of all these alternations? Why do the chatters bother to alternate the orthography of Estonian, although (differently from, say, English) it is phonetic already? Using *ff* for *hv* does not make the typing quicker or the orthography more phonetic. One explanation would be they need to signal the informality of chatroom communication as opposed to other registers of the written language.

Multiplication of characters, mostly in order to express emotion (e.g. *ahhhhh*) could also be included in this group.

The question important for compiling the user lexicon is, whether these substitutions occur in a closed set of frequent word-forms, or they are productive, i.e. used also in an open set of non-frequent word-forms. In the first case we could simply list them in the user lexicon of the morphological analyzer; in the second case we should think of some kind of an algorithm for normalizing the words prior to the standard morphological analysis.

A look at the frequency lists of words with these substitutions revealed that their frequency profiles differ from each other. For example *ff* is used for *hv* in only certain high-frequency word-forms, but *x* is used for *ks* in non-frequent as well as frequent word-forms, e.g. in the grammatical endings of the translative case of a noun (e.g. *kirurgix* ‘surgeon-sg.transl’) or subjunctive mood of a verb (e.g. *ärkax* ‘awake-ps.subj).

4.2.2 Non-capitalized proper nouns

Proper nouns are typically not written with a capital letter in the chatroom texts (capitalization is used for other purposes, namely for emphasizing). Proper nouns make up 6.5% of the tokens in chatroom texts. They are so frequent because they are used as a direct address in order to explicitly show the adressee of the message and to catch the addressee’s attention, e.g. *krizzy kle...*’Krizzy (proper noun) listen...’

4.2.3 Typos

Typos are frequent in chatroom text as the messages are often typed in a hurry and there is no time and actually also no need for editing the text. Word-forms with typos are frequent as a class but this class consists mostly *of hapax legomena*.

4.3 Vocabulary

4.3.1 Foreign-language words

Foreign languages, mostly English, but also Russian and other languages are used in chatroom texts both in the form of single words or phrases in an Estonian sentence or whole foreign-language sentences. Similarly to Finnish (see Section 3), foreign-language text can be written with phonetic spelling, e.g. *luk huus tooking* ’look who’s talking’ as a certain form of language play.

There are no chatroom-specific abbreviations, compiled of the initial letters of the words in a phrase (e.g. English *btw* ‘by the way’, *cu* ‘see you’) for frequent Estonian phrases; English loans are used instead.

4.3.2 Neologisms and genre-specific vocabulary

The chatroom texts contain a lot of genre-specific (i.e. chatroom-specific) vocabulary, mostly fresh loanwords, but also innovative derivatives. Such vocabulary includes e.g. verbs *privama* ‘hold a private conversation in chatroom’ or *ruulima* ‘rule’; nouns like *friik* ‘freak’, adjectives like *feik* ‘fake’ and adverbs like *loogish* ‘logical’.

Of course, it is a bit complicated to make a clear distinction between a new loanword and a foreign word in an Estonian-language sentence, cf. e.g. (2). The governing principle has been to regard a word-form following the rules of Estonian inflection a loanword, e.g. the inflected form of the verb *chillima* ‘chill’ in (3). But the problem remains with uninflecting words like adverbs and inflecting words in a non-marked form, e.g. a nominal in nominative case like the adjective *cool* in (2).

```
(2)tahan  ka   cool olla
want-1.sg also cool be-inf
‘I want to be cool too’
(3) no mis chillite siin
    so what chill-2.pl here
‘so what are you chilling/doing
here?’
```

4.3.3 Dialect and colloquial word-forms

Dialect and colloquial, non-standard word variants are frequent in chatroom texts, perhaps signaling the informality of the interaction. Besides colloquial word-forms also colloquial inflectional endings are used. For example, the standard ending of the active past participle form of a verb is *–nud* (e.g. *teinud* ‘done’) but in chatroom texts often the colloquial form ending with *–nd* (e.g. *teind* ‘done’) is used. This ending is productive, i.e. used also for forming past participle forms of verbs occurring only 1-2 times in the corpus.

5 Strategies for achieving better morphological analysis

From the viewpoint of automatic processing, the non-standard word-forms described in Section 4 should be divided into two groups: those that have to be included in the user lexicon manually, and those that can be normalized using some kind of rewriting rules prior to the morphological analysis, or added to the user lexicon automatically. This division corresponds roughly to that of frequent, irregular, non-productive on one side, and infrequent, regular, productive morphological or orthographic changes, on the other side.

Further in this section we will present these two solutions in more detail. Words that are frequent in chatroom texts but not present in the standard written Estonian are included in the user lexicon. The automatic methods are twofold: preprocessing of texts prior to the morphological analysis and automatic complementing of the user lexicon.

5.1 Preprocessing

The preprocessing of the corpus started from the reducing of repeated characters or syllables. Such repetitions are frequently used in chatroom texts and their function is mostly intensification, e.g *eieieieieiei* (for intensive *ei* ’no’) or *jaaaaaaaaaa* for intensive *ja* ’yes’). The fact of repetition (being an intensification) may be of importance for the further linguistic analysis but the exact number of repetitions is probably not. So we reduced all multiplied characters or syllables to three repetitions, so *eieieieieiei* became *eieiei*.

Analogous multiplying occurs in emoticons, whereas a punctuation mark in an emoticon could be repeated more than hundred times. These repetitions were also reduced to three during the preprocessing step, thus keeping the original intention of the chatter.

Next, we tried to eliminate flood, i.e. repeated chunks of text, often with nonsense meaning, entered by some of the chatters in order to disturb the conversation or simply as a bad joke. Distinctive features of the flood messages are their length and repetitions – they are usually longer than ordinary messages and/or have a distinctly repetitious nature. E.g. if a message contained three identical 20-character spans in a row, or a message was repeated at least 5 times, while being at least 110 characters long, then it was classified as flood.

Flood deletion erased 16,000 tokens, and repetition deletion diminished the word type count by 18,000, to 390,000.

5.2 User lexicon

As mentioned previously, *etmrf* is a convenient tool for our purposes as it has an in-built option of a user lexicon.

5.2.1 Manual complementation of the user lexicon

Among the groups described in Section 4, the discourse particles (Section 4.1.1) and emoticons (Section 4.1.2) are the most frequent ones in texts. The traditional system of the parts of speech of Estonian does not recognize particles (and of course not the emoticons), nor does the morphological analyzer *etmrf*. For the morphological analysis of chatroom texts two new part-of speech tags were introduced for them.

As for handling the variability of the particles (e.g. the particle with literary meaning ‘listen’ could be written as *kuule*, *kule* or *kle*), we assumed that it is better not to overgeneralize and kept the possible variants of particles apart, so *kuule*, *kule* and *kle* are tagged as three different particles, not variants of the same particle.

The treatment of particles in the output of etmrf with user lexicon is not very systematic, though. The user lexicon contains mostly those word-forms that had received the analysis of an unknown token during the morphological analysis of the chatroom corpus using *etmrf* without the user lexicon. Some word-forms, that on the basis of the usage in spoken Estonian could be suspected to be used as particles also in the chatroom texts, were checked in the corpus and if used as particles, added to the user lexicon.

So, for example the present plural 1st person form of the verb ’say’ *ütleme* and the present conditional form of the same verb *ütleks* are used as particles in spoken Estonian and also in Estonian chatrooms. But a systematic study of the particles in chatroom texts has not been conducted, so there certainly are word-forms in the output text that are used as particles but have been tagged with some other part-of-speech tag.

The user lexicon also gives a special part-of-speech tag to the emoticons. There are 100 different emoticons in the user lexicon; this relatively great amount is due to the fact that during the preprocessing of the texts multiplied punctuation marks up to three repetitions were left as they were and more repetitions were diminished to three, so the user lexicon contains separate entries for :) , :)) and :)))

Frequent new loanwords (Section 4.3.2) were also entered into the user lexicon. As for foreign words and phrases (Section 4.3.1), we do not think that the user lexicon is a proper way to solve their problem. Instead, some kind of a language identification program should be used.

Dialectal and colloquial variants of standard words (Section 4.3.3) were entered into the user lexicon so that their lemmas are those of the standard written language. Neologisms and genre-specific words (Section 4.3.2), being out of standard written language vocabulary, naturally have their own lemmas in the lexicon.

One can easily see that drawing a strict line between these two groups is somewhat problematic. For example, is the word-form *plix* a genre-specific variant of the standard Estonian word *plika* ‘girlie’ or a different, genre-specific word which lemma should be *pliks*?

The manually complemented lexicon has less than 300 entries.

5.2.2 Automatic complementation of the user lexicon

The remaining groups of deviations from the Standard Written Estonian, namely lower-case proper nouns (Section 4.2.2) and word-forms written with character substitutions (Section 4.2.1), including the phonology-related word-forms with omitted initial *h*, show high type frequency, but low token frequency.

Proper nouns are as a rule not capitalized in the chatroom texts and are frequently used as direct address in order to catch the adressee’s attention. Fortunately a program can scan a chatroom transcript prior to analyzing it morphologically and compile a list of proper nouns used as nicknames in every single chatroom, as the nicknames have been annotated in the corpus, as described in Section 2. This list can then be automatically turned into a subpart of the user lexicon.

If a nickname is homonymous with some Estonian word-form, the user lexicon leaves the word-form ambiguous between the readings of a proper noun and the other reading. If the chatter has entered his/her nickname with a capital letter, it still should be included in the user lexicon also with a small initial letter – other chatters tend not to use the capital letter while addressing her/him or chatting about her/him.

So, for example the examples (4-5) add three entries to the user lexicon:

Dammu as a proper noun

dammu as a proper noun

kakuke (‘bun’) as a general noun and a proper noun – the general noun reading is present in the actual lexicon of *etmrf*, but as the user lexicon “overrides” the original lexicon, we have to repeat it here.

```
(4) <speaker> Dammu </speaker>
<p> heihei </p>
(5) <speaker> kakuke </speaker>
<p> dammu mis teed </p>
```

The nicknames used in one chatroom are temporarily included in the user lexicon just for the morphological analysis of the same text only. The reason for this is that nicknames are a very heterogeneous class, containing also word-forms that are used as common nouns and/or even pronouns, e.g. *keegi* ’someone’. Thus, including them in the user lexicon for analysis of all the chatroom texts would result in a purposeless increase of ambiguity.

For non-standard word-forms with character substitutions, the user lexicon entries were generated in the following cyclical way.

Etmrf analyzed the text, using the user dictionary it had at the moment. The unknown words were collected, and modified with some character substitution rule reversed, thus undoing the substitutions, described in section 4.2.1, e.g. *c* was changed to *ts* in *kick* and *viici*. The examples resulted in word-forms *kitsk* (nonce word) and *viitsi* (present personal negative form of verb *viitsima* ’bother’) and were given to *etmrf* for analysis again. If *etmrf* gave the word-form some analysis other than that of an unknown word, the original word-form and the analysis of its rewritten variant were made into an entry of the user lexicon. The process continued for several cycles, trying different character substitution rules in every cycle, from the more likely ones to less likely ones, and eventually making several substitutions on the same word, e.g. *viicix* pro *viitsik*s (present personal conditional form of the verb *viitsima* ‘to bother’).

Note that this cyclical process results in including in the user dictionary also the variants of a chatroom-specific word that has been included in the user dictionary manually, e.g. *sau, sauu, sauh, tsau, zauu, tzau* etc. for *tsau* (‘ciao’).

The result of this automatic complementation was a user lexicon of over 30 000 entries (excluding the nicknames) for the whole corpus.

The automatically generated part of the user lexicon is presumably not very useful for the analysis of new chatroom texts or the texts of the other genres of the new media. More likely, it is the methodology of creating the user lexicon that is of some value to the future work: rewriting the word-forms unknown for the morphological analyzer, possibly several times, until the morphological analysis succeeds, and using the annotation of nicknames for the analysis of only the text where they occur.

6 Experiment: morphological analysis with the user lexicon

Finally, we performed morphological analysis of the whole corpus of chatroom texts with *etmrf*, using preprocessing and the user lexicon, described in Section 5. That is, we used the same set of texts we had been using for developing the strategies in the first place.

In the output text, 10.5% of the tokens still remained tagged as unknown words. This is a clear improvement from the initial 27%, when we used *etmrf* “as is”.

As for the types of the unknown words, the foreign-language word-forms, especially frequent English words like *the*, *is*, *to*, *in*, *my*, *it* etc constituted the most numerous group.

In order to evaluate the quality of the morphological analysis (in addition to coverage), we manually checked certain excerpts of the output. These excerpts, originating from different chatroom texts contained 3281 tokens altogether. 3.4% of these tokens had received a wrong analysis from our customized *etmrf*. We counted as errors also the occasions if a word-form had been attached the label of the part-of-speech it typically has in the standard written language, but its usage in the chatroom text would have suggested another part-of-speech reading. For example, the most frequent error, making up ca 50% of all errors, concerned the word-form *tere* ‘hello’. *etmrf* analyzed it as an interjection, in line with the grammar of standard literary Estonian, although it should be tagged as a particle, just like its shortened counterpart *tre* that was given the part-of-speech tag of a particle by means of the user lexicon.

As we worked only with lists of unknown tokens while compiling the user lexicon and developing the preprocessor, it could be anticipated that the other frequent type of errors was analysing a foreign-language word-form homonymous with some Estonian word-form like an Estonian one; e.g. *me* is a short form of 1st person plural pronoun ‘we’ in nominative or genitive case and *mind* is a partitive case form of 1st person singular pronoun ‘I’; both of them are also frequent tokens in English text. So in the output text all instances of *me* and *mind* while part of an English phrase were erroneously tagged as Estonian pronouns.

Also, if a genre-specific version of some Estonian word-form coincides with a Standard Written Estonian word-form, it receives an erroneous reading during the morphological analysis. For example, *ikke,* a colloquial form of *ikka* ‘still’ is homonymous with singular genitive case form of the word *ike* ‘yoke’ and has been given that analysis by *etmrf.*

The tagged version of the chatroom corpus can be queried at http://www.keeleveeb.ee

7 Unsolved issues

In spite of the efforts to recognize and delete foreign-language passages during the compilation process of the chatroom corpus, the texts still contain a considerable amount of foreign sentences, also foreign phrases and words as parts of Estonian sentences. Foreign language written with phonetic spelling is also common enough to need some special attention. A solution could be applying a language identification program that could identify as short excerpts of non-Estonian text as possible. The foreign language written with phonetic spelling needs some special attention here; perhaps compiling a small corpus of such sentences for (re-)training a language identification program is needed.

The other problem we have not found a good solution is that of the typos. As described in Section 4.2.3, typos are frequent as a type of errors but infrequent as word-forms. Perhaps we should think of a solution similar to that we used for character alternations: for unknown word-forms make some changes for fixing the common types of typos and try to perform the morphological analysis again.

By common types of typos we mean changing the order of two adjacent characters, e.g. *tow* for *two*, typing a neighboring character from the keyboard, e.g. *teo* for *two* and misplacing the space, e.g. *twot imes* instead of *two times*.

8 Conclusion: lessons learnt

This article focused on the process of customizing the morphological analyzer originally designed for the purposes of standard written language to meet the needs of a non-standard language variety; namely that of the chatroom texts.

Our main contribution lies in proposing a practical solution for coping with massive deviations from standard language, using a tool designed for analysis of this standard language.

The language of chatrooms is a variant of written Estonian. At first sight, it looks very different from the standard literary language – over a quarter of the tokens could not be analysed by a program, meant for the standard literary Estonian. A closer look, however, reveals that, roughly speaking, the differences are either systematic or concern a small set of words. This is to be expected – a (sub)language has to be learnable and usable, meaning here that unsystematic deviations from the standard language have to be limited to a small set of high-frequency words, just like irregularly inflected words have to have a low type frequency and a high token frequency.

The idea that most of the deviations result from some regular, productive modifications of the standard orthography might serve as guidance for future work. The place for looking for these regular modifications is hapax legomenon, the set of tokens that occur in the corpus only once.

In a way the chatroom corpus is self-contained: one can extract data that can be used for analyzing the data itself. It is the frequency profiles of different words that let us decide which words should be added to a user lexicon manually, and what productive rules might be at work here. It is the nicknames of the chatters that give us clues for analyzing much of the vocabulary.

References

Baron, Naomi S. 2003. Language of the Internet. In: Ali Farghali (ed.) *The Stanford Handbook for Language Engineers*. Stanford:CSLI Publications, pp. 59-127

Crystal, David 2001. *Language and the Internet*. Cambridge:University Press

Forsyth, Eric N., Martell, Craig H. 2007. Lexical and Discourse Analysis of Online Chat Dialog. *International Conference on Semantic Computing*, Irvine, California, pp. 19-26

Erelt, Mati (editor) 2003. *Estonian Language.* Linguistica Uralica Supplementary Series vol 1. Estonian Academy Publishers, Tallinn.

Hennoste, Tiit, Liina Lindström, Olga Gerassimenko, Airi Jansons, Andriela Rääbis, Krista Strandson, Piret Toomet, Riina Vellerind 2002. Suuline kõne ja morfoloogiaanalüsaator. In: Pajusalu, R.; Hennoste, T. (eds). *Tähendusepüüdja.* Tartu: Tartu Ülikooli Kirjastus, pp. 161-171

Kaalep, Heiki-Jaan, Vaino, Tarmo 2001. Complete Morphological Analysis in the Linguist's Toolbox. *Congressus Nonus Internationalis Fenno-Ugristarum* Pars V, Tartu, pp. 9-16,

Kotilainen, Lari 2002. Moi taas, ai äm päk. Lauseet, tilanteet ja englanti suomenkielisessä chat-keskustelussa. In: Ilona Herlin, Jyrki Kalliokoski, Lari Kotilainen and Tiina Onikki-Rantajääskö (eds). *Äidinkielen merkitykset*. Helsinki: Suomalaisen Kirjallisuuden seura, pp. 191-209.

Myslín, Mark and Stefan T. Gries. 2010. k dixez? A corpus study of Spanish Internet orthography. *Literary and Linguistic Computing*, Vol. 25, No. 1, pp. 85-104.

Ooi, Vincent B. Y. 2002. Aspects of computer-mediated communication for research in corpus linguistics. In: Peters, P., Collins P., Smith, A. (eds.) *New Frontiers of Corpus Research: papers form the twenty-first International Conference on English Language Research on Computerized Corpora*. Amsterdam: Rodopi, pp. 91-104.