Parsing Corpus of Estonian Dialects

Liina Lindström
University of Tartu
Estonia
liina.lindstrom@ut.ee

Kaili Müürisep
University of Tartu
Estonia
kaili.muurisep@ut.ee

Abstract

This paper introduces our work for adapting a rule based parser of spoken Estonian to the morphologically unambiguous part of the corpus of dialects. A Constraint Grammar based parser was used for shallow syntactic analysis of Estonian dialects. The recall of the grammar was 96-97% and the precision 87-89%.

1 Introduction

The goal of this research was to find a method for automatic syntactic annotation of the Corpus of Estonian Dialects (CED)[1].

The dialect corpus was compiled by two institutions – the University of Tartu and the Institute of the Estonian Language. The Corpus of Estonian Dialects consists of:

1) dialect recordings;
2) phonetically transcribed dialect texts;
3) dialect texts in simplified transcription;
4) morphologically tagged texts;
5) a database containing information about informants and recordings.

The texts in the corpus are spoken dialect interviews.

By the end of 2008, the corpus contained about 1,000,000 transcribed text words and 500,000 morphologically tagged text words.

We have used morphologically tagged texts as input for the syntactic parser.

The texts of the dialect corpus represent spoken language and have been transcribed using quite similar principles as used for the Corpus of Spoken Estonian (Hennoste et al., 2000). For this reason, we decided to test the parser of spoken language (Müürisep and Nigol, 2007, also Müürisep and Nigol, 2008) on the texts of dialects. It should be noted that the parser of spoken language is an adaption of parser for written language (Müürisep et al., 2003).

The parser for written Estonian is based on Constraint Grammar framework (Karlsson et al., 1995). The CG parser consists of two modules: morphological disambiguator and syntactic parser. In this paper, we presume that the input (transcribed speech) is already morphologically unambiguous and the word forms have been normalized according to their orthographic forms.

The parser gives a shallow surface oriented description to the sentence where every word is annotated with the tag corresponding to its syntactic function (in addition to morphological description). The head and modifiers are not linked directly, only the tag of modifiers indicates the direction where the head may be found.

```
aga                                        ;; but
   aga+0 //_J_ coord  //    **CLB @J
timä                                       ;; he
   tema+0 //_P_ pers ps3 sg nom //   @SUBJ
!!!=
ol'l'                                      ;; was
   ole+0 //_V_ main ps indic impf sg ps3 // @+FMV
latsõst                                    ;; childhood
   laps+0 //_S_ com sg el //   @P>
saan'iq                                    ;; since
   saadik+0 //_K_ post #el //   @ADVL
!!!=
tark                                       ;; clever
   tark+0 //_A_ pos sg nom //   @AN>
poiss                                      ;; boy
   poiss+0 //_S_ com sg nom //   @PRD
```

Fig. 1: An extract from syntactically annotated corpus of dialect Võru**:** *aga timä oll latsõst saaniq tark poiss* 'but he was a clever boy already since childhood'. @J - conjuction, @SUBJ - subject, @+FMV - finite main verb, predicate, @P> - complement of postposition, @ADVL - adverbial, @AN> - premodifying attribute, @PRD - predicative or complement of subject. Morphological tags are between "//"-characters.

[1] see *http://www.murre.ut.ee/korpus.html* (in Estonian)

Figure 1 depicts the format and tag set of syntactically annotated sentence. The parser of written text analyzes 88 - 90% of words unambiguously and its error rate is 2% (if the input is morphologically disambiguated and error-free). The error rate for the corpora of dialects is higher: 3-5%, but approximately 89-92% of words are assigned exactly one syntactic tag. The words which are hard to analyze remain with two or more tags.

As mentioned before, the parser is rule based. The grammar consists of 1200 handcrafted rules. The grammar rules implement a conservative parsing strategy - they rather leave the word form ambiguous than remove the correct tag.

The remainder of this paper is organized as follows. We will give an overview of the Corpus of Estonian Dialects in section 2. Section 3 describes the conversion of texts from XML format to the textual format (see Fig. 1 and 2) and section 4 deals with the modification of the grammar. We will give an overview of the parser evaluation process in section 5. In section 5, we also discuss the main shortcomings of the parser: the error types and ambiguity classes and compare the results of the parser with the results of the spoken language parser.

2 Overview of the Corpus

The Corpus of Estonian Dialects (CED) is an electronic data collection which includes authentic dialect texts from all Estonian dialects. In order to create a solid base for further research, the dialect data in CED were well-chosen and meticulously transcribed. There is roughly the same amount of material from every Estonian dialect in the corpus. The first part of CED was composed from the oldest available tape-recorded dialect texts and contains about 1 million text words.

The corpus is based on dialect recordings which have mainly been made in the 1960s and 1970s. However, the first recordings are much older – they date from 1938. The recordings are usually interviews conducted at the home of the dialect informant.

The dialect texts in Fenno-Ugric phonetic transcription constitute one of the main parts of the corpus. The aim has been to transcribe the texts as accurately as possible; the phenomena accompanying spontaneous speech (e.g. the discourse particles, corrections, repetitions, etc.) have been added to the text which usually have not been considered important in dialect research.

```
<u who="KJ">
<mark><sne>no</sne><msn>no</msn><mrf
slk="Par"/> </mark>
<mark><sne>tsuvvaq</sne><msn>tsuug</
msn><tah>pastel</tah><mrf slk="S">pl
n</mrf></mark>
</u>
```

Fig. 2: Example of morphologically annotated utterance

All of the phonetically transcribed texts have been transformed in one-to-one fashion without information loss into the simplified transcription. In addition, the comments, the text of the informant(s) and the interviewer have been annotated. This annotation is preserved also in morphologically tagged texts.

Texts in the simplified transcription are morphologically tagged. The tagged texts are in XML format. Words have been divided into 26 word classes according to their morphological inflections, syntactic characteristics and semantics. This classification is based on the system of word classes presented in Estonian grammars (Erelt et al., 1995: 14–41); however, more subclasses can be distinguished (e.g. proadverbs, affixal adverbs; see Lindstrom et al., 2006). In addition, the annotation includes 2 numbers, 15 cases and possessive suffixes for nomens, and 25 features and endings for verbs. The XML annotation consists also of meta information (dialect, informant, transcriber, annotator etc.), remarks about background activities, and sometimes also the meaning of the word form.

Figure 2 demonstrates an extract from a short dialogue turn from CED where the informant (<u who="KJ">) says *no tsuvvaq*, *no* is a particle and *tsuvvaq* is a plural noun in nominative case meaning *pastel* 'soft leather shoe'.

According to the traditional approach (cf. Pajusalu, 2003), Estonian dialects are divided into three dialect groups. These dialect groups are further divided into different dialects, the dialects are divided into parish dialects (sub-dialects). The following dialect groups and dialects are represented in the dialect corpus:

1) North Estonian dialect group: Mid, Eastern, Western, Insular dialects;

2) South Estonian dialect group: Võru, Mulgi, Tartu, Seto dialects;

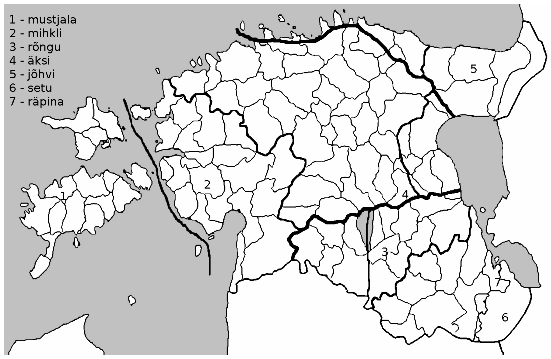

Fig. 3: The map of of the parish dialects used in the experiment

3) North-Eastern Coastal dialect group: North-Eastern (Alutaguse), Coastal dialects.

In our research for automatic syntactic annotation of dialects, we use subcorpus of 19,000 words from 7 different parish dialects (see Fig. 3).

The Äksi parish dialect (4 in the map) represents the central Mid dialect which is also the basis for standard Estonian. Mustjala (1) represents the Insular dialect and Mihkli (2) represents the Western dialect, both belonging to the North-Estonian dialect group. Jõhvi (5) belongs to North-Eastern Coastal dialect group which is rather different from the North Estonian dialect group; also, it has many similarities to Finnish dialects.

Three parish dialects – Rõngu (3), Räpina (7) and Seto (6) – represent the South-Estonian dialect group which is even more different from North Estonian (and standard Estonian) than North-Eastern Coastal dialect. Rõngu belongs to Tartu dialect which has historically had more connections to North Estonian than Räpina and Seto.

Parish dialect	**Word count**
Äksi	3569
Mustjala	1013
Rõngu	1457
Jõhvi	2975
Seto	3122
Räpina	2559
Mihkli	4303
Total	18998

Table 1: The list of used subdialects and their size

Table 1 presents word counts for these corpora.

3 Conversion of the Corpus

In order to apply constraint grammar parser to the corpus of dialects, we had to convert it to the appropriate format (see Fig. 1). As the original format of the corpus was well documented and automatically generated, the transformation process was fairly smooth. The hardest task was the mapping of differencies in word class tagging.

The original annotation did not distinguish modal verbs from main verbs but this information is crucial for syntactic rules. For this reason, every potential modal verb (4 verbs) got an additional morphological reading.

Also, the original mark-up lacks the detailed classification of pronouns. This was added using a special database. Since the dialects may have different pronouns (for example *sjoo* means 'this' in Seto subdialect) there might be a need to update the database before analysing new dialect.

Grammar rules use the valency database of adpositions. Dialect specific adpositions should be added to this before automatic transformation. Before applying the conversion program to a new dialect one should check the list of adpositions.

The tags which exist in the dialect corpus but do not exist in the corpus of spoken language remain in the annotation in the same form (for example, the case of instructive).

All words without morphological annotation, irrelevant transcription tags, records of meanings and remarks are commented out with a special tag *!!!*, so they do not influence the work of the parser (see Fig. 1).

The most substantial difference in the annotation of dialects and spoken language is in the mark up of participles. Namely, the participles which act similarly to adjectives (attributes and predicatives) are annotated as adjectives with extra tag *partic* in the corpora of spoken and written language. The mark up of dialect corpus does not distinguish different types of participles, all participles carry the POS tag of verb. As the participles act in dialects mainly as parts of verb chain (they form perfect and past perfect tense) and quite seldom as attribute or predicative, the introduction of a new morphological ambiguity was not reasonable.

```
aga                                            ;; but
   aga+0 //_J_ coord //   @J
siss                                           ;; then
   siis+0 //_D_ //   @ADVL
!!!=
e                                              ;; ee
   e+0 //_B_ //   @B
!!!$.
!!!   $. //_Z_ Fst //
*pulmad                                        ;; weddings
   pulm+0 //_S_ com pl nom // @REP
*pulmad                                        ;; weddings
   pulm+0 //_S_ com pl nom //   @SUBJ
õlid                                           ;;were
   ole+0 //_V_ main ps indic impf pl ps3 //   @
+FMV
*ikke                                          ;; still
   ikka+0 //_D_ //   @ADVL
*suure+perälised                               ;; marvellous
   suure+pärane+0 //_A_ pos pl nom //   @PRD
minul                                          ;; I
   mina+0 //_P_ pers ps1 sg ad //   @ADVL
küll                                           ;; indeed
   küll+0 //_B_ //   @B
```

Fig. 4: An extract from syntactically annotated corpus of dialect Võru. 'I had indeed marvellous weddings'

4 Conversion of the Grammar

Comparison of dialect texts with texts of spoken language revealed that the largest modifications in grammar should be related to a) inner clause boundary detection rules due to lack of intonation mark up; b) differences in annotation scheme; c) differences in vocabulary.

We inspected all rules for clause boundary detection thoroughly. In addition to the fact that dialect corpus lacks the intonation mark up, we had to consider that dialect texts resemble monologues, the utterances are longer than in everyday conversations or information dialogues.

Two types of pauses were transcribed in the dialect corpus, the shorter and the longer. The experiments showed that the use of shorter pauses as delimiters is dangerous since they occur quite often inside a phrase when a speaker is looking for an appropriate word, and their use was rather an obstacle during parsing.

In most cases the morphological description contains the normalized form of the stem which was mostly the same as in written language. There were some exceptions: we had to amend negational words (*ei* 'not', new words *ep*, *es*), add *nakkama* to the set of *hakkama* 'begin, start', etc. Also, we had to add new items to the sets related to temporal adverbial with folk calendar days like *jüripäev* 'St. George's day', *jaanipäev* 'midsummer day', *mihklipäev* 'St. Michael's day'. Fortunately, these modifications of rules were marginal.

We did not find a good solution for the analysis of participles which have different annotation scheme than used in other text corpora. It turned out that the ratio of precision and recall was best if we left the grammar willingly erroneous since the participles act seldom as attributes or predicatives in dialects.

We had to remove some seemingly correct rules from the grammar since they caused many errors due to erroneous clause boundary detection. First of all this holds for the principle of uniqueness: every main verb may have one uncoordinated subject. The same principle is also valid for objects and predicatives. These rules generate a lot of errors during the analysis of utterances with disfluencies or ellipses (see example (1)).

(1) ja ilus **ein** onn väga ilus
and beautiful hay is very beautiful
ein sin all ...
hay here below ...
'and it is a very beautiful hay here below'

We use the same method for the detection of simpler disfluencies as used for contemporary spoken language: an application of external script which removes repeats and simpler self-repairs before the parsing process and restores them in the output with a special tag after the analysis.

Modification and addition of rules took place with the help of a training corpus of 5700 words which was manually syntactically annotated. The training corpus allowed to research how the rules function and interact on dialect texts, which rules should be modified, which ones should be removed and which ones to be added. The texts of the training corpus were basically from Central, Western and Insular parishes.

During the rule design process, we attempted to minimize their error rate. If the reasonable error rate for written language is below 2% then error rate for dialects turned into 3-3.5%. The further debugging of rules gave only small effect since most of remaining errors had been caused by the phenomena specific to spoken language: disfluencies, elliptical utterances, unfinished utterances, agreement conflicts etc.

5 Evaluation

Table 2 demonstrates the gained results for different corpora. The test corpora have not been used during the process of grammar development. The results have been calculated on the automatic comparison of manually annotated corpora with automatically parsed corpora. Corpora have been annotated mainly by one human expert but the complicated utterances have been discussed by several researchers.

Dialect and type	**Word count**	**Recall**	**Precision**
Mustjala (training)	1013	97.14	86.54
Mihkli (training)	2140	96.87	90.01
Mihkli (test)	2163	96.44	85.88
Rõngu (training)	1457	96.98	89.96
Äksi (training)	977	96.52	88.56
Äksi (test)	2592	96.45	87.81
Jõhvi (test)	2975	96.12	87.35
Seto (test)	3122	95.26	88.59
Räpina (test)	2559	95.82	86.49
Training total	5587	96.89	89.09
Test total	13441	95.93	87.24

Table 2: The precision and the recall of the parser.

The table illustrates that the correctness in test corpora is almost 1% lower than in training corpora, and the precision is lower by 2%. The results are significantly worse on the corpora of Southern Estonian dialects. This may have two reasons: first, Southern Estonian texts were not used during the training and development process of the grammar. On the other hand, the Souther Estonian dialects differ significantly from standard Estonian which is based on North-Estonian central dialect. Also, one should take into account that every dialect text in this experiment represents only one speaker and the results of the dialect parsing depend on the fluency of speech of this speaker. For example, the informant for Jõhvi dialect was an elderly woman who had difficulties with speaking fluently.

The comparison of results of parsing dialects and spoken language indicates that the parser performs 1-2% worse on dialects (see Table 3). But also, we have to consider the influence of the genre to the outcome. For example, everyday conversations are easier to parse than information dialogues (this means that the precision and recall are higher). For this reason, we included a short radio interview to the comparison corpora which has a genre most similar to dialect corpora. The results of parsing this corpus are comparable to the results of parsing dialect corpora.

Corpus	**Type**	**Recall**	**Precision**
Everyday conversation	training	97.46	89.66
	test	97.58	91.84
Information dialogues	training	97.06	87.63
	test	96.77	87.42
Radio interview	test	96.80	88.47
Dialects	training	96.89	89.09
	test	95.93	87.24

Table 3: Comparison of parsing results for spoken language and dialects

5.1 Error types

The analysis of error types has been generated on the basis of subcorpus of Mihkli parish dialect of 2500 words.

We tried to group the errors in a generic fashion, individual cases which were hard to generalize have been categorized as Other. Table 4 gives overview of error types and their occurrence in the subcorpus.

In some cases it is very difficult to detect the clause boundary (see example (2)) and these errors are hard to avoid.

(2) rukis andis ikka väiksema saagi ia ei olnud
rye gave still smaller harvest good not was
'Rye gave a smaller harvest. It wasn't good.'

The errors of syntactic rules may occur also during the analysis of other types of corpora, they may be caused by unusual word order, small unfixed error in context conditions of a rule or some other shortcomings of rules.

Error	**Count**
clause boundary detection	12
syntactic rules	11
a np-phrase before or after a clause	11
ellipse	9
mapping rules	6
kõik/all	6
predicative	4
disfluency detector	2
unknown syntactic error	2
dialect specific	3
other	11
Total	77

Table 4: Count of different error types

An solitary noun phrase causes always confusion since the clause boundary detection rules could not find the border between the phrase and a new clause. Mostly the problematic noun phrases locate before the clause as in example (3).

(3) üks sort need on väga kibedad
one sort these are very bitter
'One sort. These are very bitter.'

But they can also be found after the clause as in example (4).

(4) kui aeg seokke oli seemne
when time such was seed
tegemise aeg
making time
'When time was such. It was time for sowing seeds.'

Ellipse is also a frequent phenomenon in spoken language. Often the missing element is *be*-verb as in example (5).

(5) üks ees teene taga
one before other behind
'One is before, the other is behind"

In some cases, the correct syntactic tag is never added to the word form. Typically this is a case where adjective acts as a noun but in dialect texts, there are also cases where pronouns were used as discourse particles or as a part of exclamation (*oh sa taevas* 'oh you heaven').

Unexpectedly, the word *kõik* 'all' caused a number of errors which are all hard to avoid. *kõik* 'all' can act as a normal pronoun but quite often it is premodifying or postmodifying attribute locating outside the phrase (see example (6)).

(6) pääbad oli jaettud kõik
days were divided all
'All days were divided'

kõik 'all' may also be found as a discourse marker as in example (7).

(7) pangad olid raha täis ja kõik jahh
banks were money full and all yes
'The banks were full of money and ...'

There was a regular pattern of incorrect analysis of predicatives in the test corpus as in example (8).

(8) Põllud ond neokst kitsad
Fields were such narrow
'Fields were such narrow.'

One could consider this as a shortcoming of syntactic rules.

There were only 3 errors which may be classified as dialect specific, 2 of them occur with indefinite pronoun *keegi* 'nobody' which was used instead of *miski* 'nothing'.

Disfluency detector made 2 errors, and 2 errors were related with words which syntactic functions were not possible to decide.

5.2 Ambiguities

As the error rate of the grammar was 3-4% then the second important indicator of parsing efficiency was ambiguity rate. The percentage of remaining syntactic readings is given in Table 5 (on the basis of test corpus of 13,411 words).

92% of words become unambiguous, 5.8% of words have two syntactic tags, and 1.9% of words have 3-5 syntactic tags.

The ambiguity class of subject and object dominates among ambiguity classes (see Table

6), followed by the ambiguity of subject and predicative, adverbial and subject, and finally followed by the ambiguity classes containing attributes.

Count of syntactic tags	Percentage
1	92.36
2	5.80
3	1.56
4	0.23
5	0.05

Table 5: The percentage of the count of syntactic tags in the test corpus

The domination of the ambiguity class of object and subject may be explained by the inexact clause boundary detection - it is not clear which word belongs to which verb and the decisions are made rather by the form of the noun.

Ambiguity class	Count
@OBJ @SUBJ	212
@PRD @SUBJ	134
@ADVL @SUBJ	68
@ADVL @NN>	64
@NN> @OBJ	60
@NN> @SUBJ	57
@ADVL @OBJ	56
@-FMV @ADVL	55
@ADVL @OBJ @SUBJ	53
@OBJ @PRD @SUBJ	36
@ADVL @PRD @SUBJ	35
@<NN @ADVL	30

Table 6: The main ambiguity classes

6 Conclusions

Our experiment of using a parser of spoken language for syntactic analysis of the corpus of dialects can be regarded fairly successful. Although the error rate of the analysis is 1-2% higher than for the spoken language parser, most of the errors are hard to avoid. The parser and its grammar that are based on Constraint Grammar framework are robust enough to deal with non-fluent speech and syntactic constructions specific to dialects. Approximately 10% of words remain ambiguous in the output of the parser but fortunately these ambiguities will not obstruct linguistic research.

We plan to analyze the whole corpus in an automated fashion and make it available on the web. Also, we are planning to create a publicly available search engine for the corpus, in order to facilitate further studies of Estonian syntax and dialects.

References

Erelt, Mati, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael, Silvi Vare. 1995. *Eesti keele grammatika*, vol. 1. Eesti TA Keele ja Kirjanduse Instituut, Tallinn.

Hennoste, T., Lindström, L., Rääbis, A., Toomet, P., Vellerind, R. 2000. Tartu University Corpus of Spoken Estonian. In Seilenthal, T., Nurk, A., Palo, T., eds.: *Congressus Nonus Internationalis Fenno-Ugristarum* 7.-13. 8. 2000. Pars iv. Dissertationes sectionum: Linguistica I, Tartu (2000) 345–351

Karlsson, F., Voutilainen, A., Heikkilä, J., Anttila, A. 1995. *Constraint Grammar: a Language Independent System for Parsing Unrestricted Tex*t. Mouton de Gruyter, Berlin and New York.

Lindström, Liina, Liisi Bakhoff, Mari-Liis Kalvik, Anneliis Klaus, Rutt Läänemets, Mari Mets, Ellen Niit, Karl Pajusalu, Pire Teras, Kristel Uiboaed, Ann Veismann, Eva Velsker. 2006. Sõnaliigituse küsimusi eesti murrete korpuse põhjal. – E. Niit (ed.) *Keele ehe*. Tartu Ülikooli eesti keele õppetooli toimetised 30, Tartu: 154-167

Müürisep, Kaili, Helen Nigol. 2007. Disfluency Detection and Parsing of Transcribed Speech of Estonian. *Proc. of Human Language Technologies as a Challenge for Computer Science and Linguistics*. 3rd Language & Technology Conference (ed. Zygmunt Vetulani). Oct 5-7, 2007, Poznan, Poland. Fundacja Uniwersitetu im. A. Mickiewicza. pp. 483-487.

Müürisep, Kaili, Helen Nigol. 2008. Where Do Parsing Errors Come From: The Case of Spoken Estonian. In Sojka, P.; Horak, A.; Kopecek, I.; Karel, P. (eds.). LNCS 5246. *Text, Speech and Dialogue*. Springer-Verlag. pp. 161 - 168.

Müürisep, Kaili, Tiina Puolakainen, Kadri Muischnek, Mare Koit, Tiit Roosmaa, Heli Uibo. 2003. A New Language for Constraint Grammar: Estonian. *International Conference Recent Advances in Natural Language Processing*. Proceedings. Borovets, Bulgaria, 10-12 September 2003, pp. 304-310.

Pajusalu, Karl. 2003. Estonian Dialects. – Mati Erelt (ed.) *Estonian Language*. Linguistica Uralica supplementary series, vol. 1. Estonian Academy Publishers, Tallinn: 231-272.