Disfluency Detection and Parsing of Transcribed Speech of Estonian

Kaili Müürisep, Helen Nigol

University of Tartu
Tartu, Estonia
{kaili.muurisep, helen.nigol}@ut.ee

Abstract

The paper introduces our strategy for adapting a rule based parser of written language to transcribed speech. Special attention has been paid to disfluencies (repairs, repetitions and false starts). A Constraint Grammar based parser was used for shallow syntactic analysis of spoken Estonian. The modification of grammar and additional methods improved the recall from 97.5% to 97.7% and precision from 90.2% to 90.4%. Also, the paper gives a detailed analysis of the types of errors made by the parser while analyzing the corpus of disfluencies.

1. Introduction

The paper introduces our strategy for adapting a rule based parser of written language to transcribed speech.

The corpus of spoken Estonian (1ă065ă000 words, 1703 transcripts) contains 100ă000 part-of-speech-tagged and manually morphologically disambiguated words (Hennoste et al., 2000). Our goal is to provide syntactic annotation to that part of the corpus.

Parsing of spontaneous speech is a serious challenge: spoken language has often different vocabulary, it is hard to determine where the sentence starts from and where is the end due to the lack of capitalized letters and punctuation marks. Spontaneous speech is also rich of disfluencies such as partial words, filled pauses (e.g., *uh*, *um*), repetitions, false starts and self-repairs. One type of disfluency that has proven particularly problematic for parsing is speech repairs: when a speaker amends what he is saying mid-sentence or "stretches of wording in which a speaker begins to realize one grammatical plan, but breaks off and either starts a fresh or continues in conformity to a different plan" (Sampson, 1998).

In this paper, we will focus on the parsing of non-fluent speech using a rule based parser.

The parser for written Estonian (Müürisep, 2001) 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, in which 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.

The figure 1 demonstrates 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 unerroneous). The words which are hard to analyze remain with two or more tags.

The parser is rule based. The grammar consists of 1200

```
Se # this
  see+0 //_P_ dem sg nom //  **CLB @NN>
veranda # veranda
  veranda+0 //_S_ com sg nom //  @SUBJ
on # is
  ole+0 //_V_ main indic pres ps3 sg //  @+FMV
minu # my
  mina+0 //_P_ pers ps1 sg gen //  @P>
meelest # opinion
  meelest+0 //_K_ post #gen //  @ADVL
maailma # world?s
  maa_ilm+0 //_S_ com sg gen //  @NN>
kihvtim # coolest
  kihvti=m+0 //_A_ comp sg nom //  @AN>
asi # thing
  asi+0 //_S_ com sg nom //  @PRD
$.
  . //_Z_ Fst //
```

Figure 1: Syntactically analyzed utterance In my opinion, this veranda is the coolest thing in the world. (@SUBJ - subject, @PRD - predicative or complement of the subject, @+FMV - finite main verb, @ADVL - adverbial, @AN>, @NN> - attributes, @P> - complement of the postposition)

handcrafted rules, described thoroughly in (Müürisep, 2000). The grammar rules try to avoid risks. They rather leave the word form ambiguous than remove the correct tag.

We have achieved promising results adapting this parser for spoken language in our previous experiment described in section 2. We have fixed the weakest point in our experiment, namely the limited size of corpus. For our new experiments, we use different corpora for testing and training the parser and special corpora of disfluencies. The description of corpora is given in section 3.

Finding the smallest appropriate syntactic unit for parsing is the key issue in automatic analysis of spoken language. The problems handling clause boundaries, false starts and overlaps are discussed in section 4.

The detection of self-repairs and repetitions is essen-

tial prior parsing since ungrammatical constructions disturb the analysis of correct parts of utterances. We give an overview of our methodology for discovering repairs and repetitions in section 5. Also we describe the major types of errors still occur in parsed text and give a preliminary evaluation of the performance of disfluency detector.

2. The first experiment

The first experiment to adapt parser of written language to spoken Estonian was made in 2005 (Müürisep and Uibo, 2006). This approach did not pay special attention to disfluencies. The end of dialogue turn was used as the delimiter of utterance. Although the input includes the punctuation marks, they are not reliable. They describe the intonation, not a certain end of the utterance. Also, two additional tags were added to the tag set of the parser: for particles and for words with unknown syntactic function. To adapt the parser for the spoken language, new rules for the sentence internal clause boundary detection were compiled and some of the syntactic constraints were reformulated, taking into account the specific features of the spoken language.

The outcome of the first experiment demonstrated that the adaptation of the written language parser for the spoken language turned out to be easier task than expected. The efficient detection of clause boundaries became the key issue for successful automatic analysis, while syntactic constraints required only minimal modification. Quite surprisingly, the performance of the parser for the spoken language exceeded its original performance for the written language (which can be due to simpler and shorter clauses of spoken language). The output of the parser was compared with manually annotated corpus (2543 words) and the following results have been achieved (the results for parsing the written language are enclosed in parentheses):

1. recall (the ratio of the number of correctly assigned syntactic tags to the number of all correct tags): 97.3% (98.5%).

2. precision (the ratio of the number of correctly assigned syntactic tags to the number of all assigned syntactic tags): 89.2% (87.5%).

The recall describes the correctness of analysis and precision illustrates the level of noisiness.

The similar experiments using CG have been made by Eckhard Bick for parsing spoken Portuguese (Bick, 1998). His grammar achieved almost unambiguous output with correctness rate 95 - 96% (automatic morphological disambiguation included).

3. Corpora

We used morphologically disambiguated texts for the experiments described in this article. The texts were normalized (vaguely articulated or colloquial words have the description of the corresponding word form in the written language) and provided with some transcription annotation (longer pauses, falling or rising intonation). For the assessment of the work of the parser the benchmark corpus of 6700 words was compiled and analyzed manually. This corpus contains both longer narrative dialogues and shorter dialogues where turns alternate swiftly. The automatic syntactic analysis was corrected by a single expert.

Disfluencies		Total
Repairs	Word fragments	53
	Substitutions	50
Repetitions		113
False starts		33
Total		249

Table 1: Occurrence of types of disfluencies

We used separate corpus of 8400-words for training the parser (i.e. generating or modifying rules) and in addition the special corpus of disfluencies (Nigol, 2007) (13,000 words), which was annotated according to principles of the Disfluency annotation stylebook for the Switchboard corpus (Meteer et al., 1995). During the annotation the annotator detects the extent of the disfluency and annotates the reparandum and repair, as well as the editing phase. Different tags are used to specify the disfluency, i.e. whether the subject is a repair - RP, repetition - RE, particle - D, filled pause - F, or non-analyzable unit - X. A false start is marked with '+/'. As a result of annotation, after the removal of the reparandum and the editing phase, the result should be a syntactically well-formed utterance, e.g. consider the following example (1).

(1) a. Original utterance:

meil lihtsalt sellist nii-öelda
we simply this kind of so to say
süvenemiseks pole eriti aega
to indagate do not have not much time

'simply we do not have such time to such so said to indagate'

b. Annotated utterance:
meil lihtsalt [RP sellist + {D nii-öelda} süvenemiseks] pole eriti aega

c. Normalized utterance:
meil lihtsalt süvenemiseks pole eriti aega
'we simply do not have the time to indagate'

The annotation scheme was applied to an information dialogue subcorpus of Estonian, part of the Estonian Dialogue Corpus. 35 randomly selected information dialogues (13,168 words) were analyzed. The shortest dialogue consisted of 31 words and the longest of 1962 words. In Table 1, the occurrence of the types of disfluencies is presented.

Based on this corpus of disfluencies two syntactically annotated corpora were created. The first corpus was parsed in its original form; the second was parsed after its normalization. In the first corpus, disfluencies have been annotated as they were normal parts of the utterance and they had corresponding syntactic tags. If it was impossible to determine the function of the word in the sentence, special tag was used - @T (unknown syntactic function).

In Table 2 the gained recall and precision with the preliminary version of the parser for spoken Estonian

Type of disfluency	Utterances	Recall	Precision
Repairs	original	94.4	84.6
	normalized	96.2	87.3
Repetitions	original	98.2	90.7
	normalized	98.6	91.8
False starts	original	97.4	90.0
	normalized	98.9	93.8

Table 2: Results of the experiment (%)

((Müürisep and Uibo, 2006) is demonstrated. As the morphological disambiguation was made manually, the statistics shows only the problems of syntax.

The results showed significant improvement. For repairs, the recall rate rose 1.8% and precision 2.7%. For repetitions the recall rose slightly, 0.4% and 1.1% accordingly. For false starts, the recall rate rose 1.5% and precision 3.8%. So the detection of disfluencies should improve the overall statistics of the rule-based parser.

Several experiments,for example, (Charniak and Johnson, 2001; Lease and Johnson, 2006), have showed that parsing performance is increased when dis?uencies are removed prior to data-driven parsing. These results prove that this statement is valid also for rule based parsing.

4. Clause boundaries, false starts and overlaps

The inner clause boundaries in the sentence of written language are detected using the conjunction words, punctuation marks and verbs. Parser assigns the tag CLB to the first word of the sentence internal clause. Since the usage of punctuation marks is different in the spoken language transcription, we had to remove original rules and write new ones which are less dependent on punctuation marks. Also the meaning of the clause is different in spoken language. A clause in our interpretation is something like a text chunk which may include a verb but this is not obligatory. For example, a dialog turn (2) is parsed as one unit but the presence of the clause boundary tag allows us write rules which check the context only inside the clause.

(2) A: mitte iga liin see (CLB) sõltub liinist
A: not every line it (CLB) depends on line
'Not every line. It depends on the line.'

The false starts with or without verb are considered as separate chunks. A special attention has been paid to particles characterizing spoken language (noh, pause fillers aa, ee, öö). These particles are used as delimiters if there are finite verbs in both left and right context.

In spite of the fact that all the clause boundary detection rules were reformulated the erroneous clause boundaries remain the main source of errors: more than the third of errors in training corpus were caused by wrong or missing clause boundaries. The thorough inspection of errors indicated that direct speech may occur quite often in spoken language (see Example 3). The new rules try to fix this phenomenon.

	Recall	Precision
Repairs	95.6	85.5
Repetitions	98.4	90.7
False starts	97.9	90.0
Test corpus	97.6	90.2

Table 3: Results with improved clause boundary rules.

(3) ma mõtsin huvitav, endal vaja ei lähe või
I thought interesting yourself need not go or
'I thought interesting don?t you need (these), do you'

The second flaw in rules was the missed beginnings of small clauses with second-person verbs, e.g. (4).

(4) aga selle taga on saad aru selline lähenemine
but this behind is understand this approach
'this approach is used behind this as you understand'

The speaker may drop the verb *be* in utterances:

(5) pesu (on) pestud (on) vaja triikimislaud osta
clothes (is) washed (is) need ironing board buy
'the clothes have been washed, (someone) needs to buy ironing board'

A closer look to the transcription of source texts showed that the longer pauses suit to be clause boundaries.

The rules try to discover false starts and mark these by clause boundary tags but this is possible only if there is a verb in the false start phrase, e.g.,

(6) mul (CLB) on kassetil (CLB)
I-SG-AD be-SG3 tape-SG-AD
oleks ruumipuudus tekkinud
be-COND lack of space arise-PCP
'I have | there would be no space on the tape'

Overlaps break the utterance. If the overlap is short and lasts all the turn then it is possible to move the overlapping turn from the middle of the utterance to the next turn (see Example 7). This repaired 2 errors.

(7) T: /.../ [ma]=aint ‘mõtsin need on ‘õutselt ‘mugavad. [(0.8) et=nad]
/.../ I thought that these are extremely comfortable that they
L: [mh] [((naerab))]
mh (laugh)
T: ‘niigi sobivad ‘kätte. /.../ anyway fit to hand /.../

The results gained using improved grammar are given in Table 3.

5. Repetitions and self-repairs

One of the signals about self-repairs are the words with interruptions (e.g. nor- instead of normal). All these words are tagged with a new tag - @REP. Also the patterns *word*

break- word and *word break- või* were detected and annotated.

Unfortunately, it is not possible to detect these patterns using Constraint Grammar rules and a special external script was used for this purpose.

The example in Figure 2 illustrates detected self-repairs.

```
väga [ADVL]  # very
    väga+0 //_D_ // **CLB @ADVL
nor- [REP]  # nor-
    nor //_T_  #- //   @REP
väga [REP]  # very
    väga+0 //_D_ //   @REP
normaalne [PRD]  # normal
    normaalne+0 //_A_ pos sg nom // @PRD
noh [B]  # noh
    noh+0 //_B_ //    @B
väga [ADVL]  # very
    väga+0 //_D_ //    @ADVL
naiss [T]  # nice
    naiss //_T_ //    @T
$.
    . //_Z_ Fst //
```

Figure 2: Example of analyzed utterance with self-repair

Actually this minor replacement gave a small effect in training corpus, influencing only analysis of couple of sentences.

In the special corpus of 920-words with interrupted words the amount of errors decreased from 36 to 29 (the correctness grew 0.3%) and unambiguity rate grew 0.1%.

Also a grammar external script was used for tagging simple repetitions of a single word (*miks miks miks peab* ...- /why why why one should .../). The repetitions of verb be and numerals may occur in the normal sentence, so we had to consider the part-of-speech tags.

Also, it is possible to repeat the same word in different cases. Some of these repetitions are normal in written texts also (e.g, *samm sammult* /step by step/), but the others signal to the occurrence of a self-repair (see Example 8).

(8) noh erinevatel päevadel on võimalik siis
noh different days is possible then
mägi mäge valida
hill-NOM hill-PART to choose
'so it is possible to choose a hill in different days'

Table 4 demonstrates the results gained in test corpora. The results of parsing of repetitions are as good as in the manually normalized corpus.

The closer look to the error types in the corpus of self-repairs (see Table 5) shows that the repairs and complicated phrases of repetitions are the reason of the most of errors. The corpus consists of 2100 words and the parser made 82 errors in the analysis of the corpus.

Self-repairs are hard to detect and this type of errors is difficult to avoid. In some cases the edited word and editing word have the same form (see Example 9) and maybe

	Recall	Precision
Repairs	96.1	86.3
Repetitions	98.6	92.1
False starts	98.1	91.1
Test corpus	97.7	90.4

Table 4: Results with tagged repetitions and self repairs.

Type of errors	Count	%
self-repairs and repetitions	18	22.0
error in the clause boundary	15	18.3
regular syntactic errors	13	15.8
unfinished words, phrases, clauses	9	11.0
pause fillers	7	8.5
expressions of time	5	6.1
other	15	18.3

Table 5: Analysis of errors in the corpus of self-repairs.

it is possible to refine our detector of disfluencies in further experiments.

(9) see on siin selle kaubahalli uue kaubahalli
this is here this shop new shop
kõrval
next to
'this is here next to the new shop'

The syntactic errors caused by false or undetected clause boundaries are also dominant. The situation where a turn of the dialog consists of separate noun phrase and a clause with a verb is typical (see Example 2).

The utterances in spoken language are often unfinished or consist of unfinished phrases. Their automatic analysis is impossible in most of the cases. If the pause fillers are special particles like *ee*, *noh*, *mmm*, they do not disturb the analysis of the sentence. Unfortunately, the demonstrative pronoun *see* (this) in nominative case is quite often used as a pause filler. It is very hard to detect automatically, and it causes errors in analysis of subjects or objects typically (see Example 10).

(10) meil on olnud see ee
we-ADES have been this-NOM ee
kapslid
capsules-NOM
'we have had these capsules'

And the fifth biggest class of errors characterizing spoken language are the errors related to expressions of time. It is typical that the adverbial of the time is in nominative case in spoken language.

(11) kolmapäev tuleb tööle
Wednesday-NOM come-SG3 work-ALLAT
'he arrives to the office on Wednesday'

Also the hours and minutes are expressed as a sequence of numbers; their analysis needs an extra work.

The "regular syntactic errors" occur in written text also. We should consider removing some heuristic rules which tend to cause errors in spoken language texts. Also, maybe it would be reasonable to remove the rules which declare that there is possible only one uncoordinatedsubject in every clause since we can not trust the clause boundary detection.

Unfortunately, we can not present the exact statistics about the efficiency rate of finding disfluencies. The disfluency detector added 96 @REP tags to the corpus of self-repairs and they all were correct. Also the corpus contains 8 words with unknown morphological information and these words got the @T tag. 24 words which have been annotated manually as unknown syntactic function got some other syntactic tag erroneously. So the efficiency rate of disfluency detector is less than 81.25% (we did not take into account the disfluencies which have been annotated with some other syntactic tag than @T or @REP). Given that the state-of -the-art data-driven edit word detector performs at about 80% f-measure (Johnson and Charniak, 2004), our rule based detector have room for improvement.

6. Conclusions

The experiment for improving the efficiency of the parser demonstrates that the rule-based grammar composed originally for written unrestricted text is suitable for parsing spoken language but one should pay a special attention to the detection of clause boundaries. Also, the automatic identification of disfluencies helps the parser a lot.

Repetitions can be analyzed as well as in normalized corpus, almost half of false starts are well detectable and do not affect the analysis of the rest of the utterance. The hardest type of disfluencies are self-repairs. Their analysis gained only marginal improvements.

The next challenge would be the adaptation of morphological disambiguator to spoken language.

Also we need bigger annotated corpus in order to analyze the syntactic phenomena of disfluencies. We plan to enhance the annotation scheme of our corpora with more exact disfluency information, similar to the works of (Bies et al., 2006).

7. References

Bick, Eckhard, 1998. Tagging Speech Data - Constraint Grammar Analysis of Spoken Portuguese. In *Proceedings of the 17th scandinavian conference of linguistics*. Odense.

Bies, Ann, Stephanie Strassel, Haejoong Lee, Kazuaki Maeda, Seth Kulick, Yang Liu, Mary Harper, and Matthew Lease, 2006. Linguistic resources for speech parsing. In *Fifth international conference on language resources and evaluation (LREC'06)*. Genoa, Italy.

Charniak, Eugene and Mark Johnson, 2001. Edit detection and parsing for transcribed speech. In *Proceedings of the second conference of the north american chapter of the association for computational linguistics (NAACL'01)*.

Hennoste, Tiit, Liina Lindström, Andriela Rääbis, Piret Toomet, and Riina Vellerind, 2000. Tartu University Corpus of Spoken Estonian. In T. Seilenthal, A. Nurk, and T. Palo (eds.), *Congressus nonus internationalis fenno-ugristarum 7.-13. 8. 2000. Pars iv. Dissertationes sectionum: Linguistica I*. Tartu.

Johnson, Mark and Eugene Charniak, 2004. A tag-based noisy-channel model of speech repairs. In *Proceedings of the 42nd meeting of the association for computational linguistics (ACL'04)*. Barcelona, Spain.

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, and Arto Anttila, 1995. *Constraint grammar: a language independent system for parsing unrestricted text*. Berlin and New York: Mouton de Gruyter.

Lease, Matthew and Mark Johnson, 2006. Early deletion of fillers in processing conversational speech. In *Proceedings of the human language technology conference of the naacl (HLT–NAACL'06), companion volume: short papers*. New York City, USA: Association for Computational Linguistics.

Meteer, Marie, A Taylor, R MacIntyre, and R Iver, 1995. *Dysfluency Annotation Stylebook for the Switchboard Corpus.*. LDC.

Müürisep, Kaili, 2000. *Eesti keele arvutigrammatika: süntaks*. Dissertationes Mathematicae Universitatis Tartuensis 22. Tartu: Tartu Ülikooli kirjastus.

Müürisep, Kaili, 2001. Parsing Estonian with Constraint Grammar. In *Online Proceedings of NODALIDA'01*. Uppsala.

Müürisep, Kaili and Heli Uibo, 2006. Shallow Parsing of Spoken Estonian Using Constraint Grammar. In P. J. Henrichsen and P.R. Skadhauge (eds.), *Treebanking for Discourse and Speech. Proceed. of NODALIDA 2005 Special Session on Treebanks for Spoken Language and Discourse*, Copenhagen Studies in Language 32. Samfundslitteratur.

Nigol, Helen, 2007. Parsing Manually Detected and Normalized Disfluencies in Spoken Estonian. In *Proceedings of NODALIDA 2007*. Tartu.

Sampson, Geoffrey, 1998. Consistent annotation of speech-repair structures. In A. Rubio (ed.), *Proceedings of the First International Conference on Language Resources and Evaluation*, volume 2. Granada, Spain.