Parsing Manually Detected and Normalized Disfluencies in Spoken Estonian

Helen Nigol
University of Tartu
helen.nigol@ut.ee

Abstract

An experiment with an Estonian Constraint Grammar based syntactic analyzer is conducted, analyzing transcribed speech. In this paper the problems encountered during parsing disfluencies are analyzed. In addition, the amount by which the manual normalization of disfluencies improved the results of recall and precision was compared to non-normalized utterances.

1 Introduction

Müürisep and Uibo (2006) have made the first attempt at analyzing spoken Estonian via adapting the existing Constraint Grammar based syntactic analyzer for written Estonian. Based on a 2543-word corpus of spoken Estonian, the achieved recall rate was 97.3% and precision 89.2%. Results were surprisingly good, but there is room for some improvement. They were focusing on the problems of parsing incomplete and elliptical sentences as well as problems in finding clause boundaries. In this experiment we investigate how the parser copes with analyzing disfluencies, which adds further difficulty to the task of parsing spoken language. In this paper we will concentrate on certain types of disfluencies: repairs and false starts. Under repairs, we distinguish word fragments and substitutions, wherein the speaker corrects or alters the utterance. A false start is where speaker abandons the utterance entirely and starts over. All the instances were manually annotated and some words normalized, i.e. ungrammatical utterances were made grammatical.

Disfluencies have been annotated in several corpora of spoken English; the most well-known is the Switchboard Corpus (Meteer et al, 1995). There are also several studies where the detection and correction of disfluencies have been made automatically. E.g. meta-rules (Hindle, 1983; McKelvie, 1998), a statistical approach (Stolcke and Shriberg, 1996), triggers as indicators of repair (Spilker et al, 2000) and prosodic-acustic cues (Nakatani and Hirschberg, 1993) are used for the detection of disfluencies; for normalization, pattern-matching (Bear et al, 1992; Heeman and Allen, 1994; Kurdi, 2002) is applied; it may also be handled as a machine translation task (Spilker et al, 2000). As the disfluencies are such a heterogenous class of linguistic events, maximum results may only be achieved through combining different methods. So far the best results have been attained when some amount of manual annotation prior to automatic analysis is conducted.

This paper is structured as follows: firstly, the compiled subcorpus is introduced, and the annotation of disfluencies is defined. In section 3, an overview of the parser and the prework for the analysis is given. In section 4, the results of a test run with the Estonian constraint-based parser are presented.

2 Corpus and annotation

Detecting and normalizing disfluencies is a complex task. Classically, the disfluent unit is divided into four: reparandum, interruption point, editing phase, and repair. There are very obvious cases where it is very easy to decide what has been replaced, deleted or added, but there are also very vague cases. During the annotation the annotator detects the extent of the disfluency and annotates the reparandum and repair, as well as the editing phase. The *Disfluency annotation stylebook for the Switchboard corpus* has been used as the model for the annotation of disfluency in spoken Estonian. The tags used in the annotation of spoken Estonian are presented in Table 1. The abbreviations *RP*, *D*, *F*, and *X* specify the content of the brackets, i.e. whether the subject is a repair, particle, filled pause, or non-analyzable unit. A false start is marked with '+/'. As a result of annotation, after the removal of the reparandum and the editing phase (consisting of particle or filled pause), the result should be a syntactically well-formed utterance.

DF class	**Tag**
Repair	[RP…+…]
Particle	{D…}
Filled pause	{F…}
False start	+/
Non-analyzable unit	{X…}

Table 1. The tags used in annotation of spoken Estonian.

The annotation scheme was applied on an information dialogue subcorpus of Estonian, part of the Estonian Dialogue Corpus[1]. 35 randomly selected information dialogues (13 168 words, 1991 utterances) were analyzed. The shortest dialogue consisted of 31 words and the longest of 1962 words. In Table 2, the occurrence of the types of disfluencies is presented.

Disfluencies	**Total**
Word fragments	53
Substitutions	50
False starts	33
Total	**136**

Table 2. Occurrence of types of disfluencies in corpus.

3 Experiment

The experiment was conducted with Estonian constraint-based parser, which was originally designed to analyze written language. The parser gives a shallow surface oriented analysis to a sentence, in which every word is annotated with the tag corresponding to its syntactic function. For analyzing spoken language, two additional tags were adapted: @B – particle; @T – unknown syntactic function. In the adapting process, clause boundary detection rules as well as some syntactic constraints were changed. The process of syntactic analysis consists of three stages: morphological disambiguation, identification of clause boundaries, and identification of the syntactic functions of words. The syntax used in CG is word based, i.e. no hierarchical phrase structure is constructed. (Müürisep and Uibo, 2006)

All utterances containing word fragments, substitutions and/or false starts were analyzed twice. The first run parsed the corpus in its original form; the second run parsed the same corpus after its normalization. The original utterance was retained in the corpus, but the input to the parser did not include the disfluencies. As the analysis of disfluencies is in the preliminary stage and the syntactic analyzer does not count the tags used in annotation, the reparandum and editing phase were manually removed to get the normalized utterances. The corpus of original utterances contained 4701 words and the corpus of normalized utterances 3864 words; thus, 837 words were removed by normalization.

When analyzing the original utterances, it was assumed that mistakes occurring during analysis would be related to the disfluencies, as all utterances contained one of the disfluencies. For normalized utterances the assumption was that as the obvious reasons of mistakes had been removed, the occurred mistakes would be caused by other things, e.g. an incorrect clause boundary detection causing the wrong analysis of a whole utterance. Thus, the encountered mistakes in the original and normalized utterances are not analyzed in further detail. However, the achieved recall (the ratio of number of correct assigned syntactic tags to the number of all correct tags) and precision (the ratio of number of correct assigned syntactic tags to the number of all assigned syntactic tags) of repairs and false starts is compared and the main problems

[1] http://math.ut.ee/~koit/Dialoog/EDiC.html

encountered in parsing word fragments, substitutions and false starts are discussed.

4 Results and analysis

In this section, the problems encountered during parsing the three types of disfluencies are analyzed, as well as the amount by which the manual normalization of disfluencies improved the results. The results of the experiment are given in Table 3. As the morphological disambiguation was made manually, the statistics show only the problems of syntax.

	Repairs		**False starts**	
	Original	**Normalized**	**Original**	**Normalized**
Recall	94.4%	96.2%	97.4%	98.9%
Precision	84.6%	87.3%	90.0%	93.8%

Table 3. Results of the experiment.

The results showed significant improvement. For repairs, the recall rate rose 1.8% and precision 2.7%. For false starts, recall rate rose 1.5% and precision 3.8%.

4.1 Word fragments

In the majority of the cases the analysis of word fragments did not present a problem, as long as the word in question was changed into another.

Original utterance	
sööke	# food
söök+e //_S_// @SUBJ	
nende	# these / pl gen
see+de //_P_// @NN>	
hin-	# pri-
hin+0 //_T_// @T	
selle	# this / sg gen
see+0 //_P_// @NN>	
hinna	# price
hind+0 //_S_// @P>	
sees	# in
sees+0 //_K_// @ADVL	
ei	# is not
ei+0 //_V_// @NEG	
ole	# included
ole+0 //_V_// @+FMV	

Example 1. sööke **[RP** nende hin- + selle hinna**]** sees ei ole ('food is not included in price').

The problems were in analyzing phrases that contained a disfluent element. In this case the word fragment is automatically assigned with the tag @T, but the other element of the phrase is analyzed as if it still were part of the utterance, although it is not, e.g. Example 1.

The reparandum in Example 1 is 'nende hin-' and the speaker has corrected it to 'selle hinna'. But from this analysis it is possible to determine that the word 'hinna' has two attributes, 'nende' and 'selle'. The analysis will be adequate assuming the whole phrase is assigned with the tag @T or is somehow otherwise marked. But as was mentioned earlier, the Constraint Grammar syntactic analyzer is word-based; no phrase structure is constructed.

4.2 Substitutions

Substitutions are more complex to detect and normalize than the word fragments. The syntactic analyzer examines the utterance from left to right, i.e. from the beginning of utterance, but in a disfluent utterance the repair is always situated at the end of the utterance. This proves the need for normalization as a preprocessing task as long as the constraint-based syntactic analyzer is being used. Example 2 shows what happens when the reparandum is part of the utterance.

Original utterance	
erinevatel	
erinev+tel //_A_// @AN>	# different / pl ade
päevadel	
päev+del //_S_// @ADVL	# day / pl ade
on	
ole+0 //_V_// @+FMV	# is
võimalik	
võimalik+0 //_A_// @AN>	# possible
sis	
sis+0 //_D_// @ADVL	# then
mägi	
mägi+0 //_S_// @SUBJ	# mountain / sg nom
mäge	
mägi+0 //_S_// @OBJ	# mountain / sg part
valida	
vali+da //_V_// @-FMV @<INF_N @ADVL	# to choose

Example 2. erinevatel päevadel on võimalik sis [RP mägi + mäge] valida ('every day you may choose a new mountain').

The word 'mägi' evoked that the adjective 'võimalik' is analyzed as an attribute and the verb 'valida' had three different tags: non-finite main verb, infinitive as a modifier and adverbial. The correct parse would have been that 'võimalik' is predicative and 'valida' a subject.

4.3 False starts

Müürisep and Uibo (2006) found that false starts are detectable with rules by marking them with clause boundary tags, but this is possible only if there is a verb in the false start phrase. In Example 3 is presented an utterance where the false start does not have a verb. The utterance with the false start has dictated the analysis of the entire utterance, though the real subject is the word 'seda'.

Original utterance

```
kui
   kui+0 //_J_//   @J                      # if
ta
   tema+0 //_P_//  @SUBJ                   # he/she
seda
   see+da //_P_//  @ADVL @NN>              # this / sg  part
seda
   see+da //_P_//  @PRD @ADVL              # this / sg  part
tükina
   tükk+na //_S_//  @ADVL                  # as a single piece
siin
   siin+0 //_D_//  @ADVL                   # here
ei
   ei+0 //_V_//  @NEG                      # not
ole
   ole+0 //_V_//  @+FMV                    # is
```

Example 3. kui ta seda +/ seda tükina siin ei ole ('we don't have it here as a single piece').

5 Conclusion

In this paper we presented how the Constraint Grammar based syntactic analyzer handles the problems of parsing disfluent spoken Estonian and how much the manual normalization of disfluencies improved the results of recall and precision compared to non-normalized utterances. An experiment was conducted. The utterances containing word fragments, substitutions and/or false starts were analyzed twice. The first run parsed the corpus in its original form; the second run parsed the same corpus after its normalization. For repairs the recall rate rose from 94.4% to 96.2% and precision from 84.6% to 87.3%. For false starts recall rose from 97.4% to 98.9% and precision from 90.0% to 93.8%.

The main objective is to identify regularities in different kinds of disfluencies, focusing on repairs and false starts. The future plan is to increase the annotated corpus and to use it as a training corpus to train the parser to automatically detect disfluencies in order to save time in manual annotation.

References

Bear, John, John Dowding, and Elizabeth Shriberg. 1992. Automatic Detection and Correction of Repairs in Human-Computer Dialog. Proceedings of the DARPA Speech and Natural Language Workshop.

Heeman, Peter, and James Allen. 1994. Tagging Speech Repairs. ARPA Workshop on Human Language Technology, pp. 187–192.

Hindle, Donald. 1983. Deterministic Parsing of Syntactic Nonfluencies. Proceedings of the 21st Meeting of the Association of Computational Linguistics.

Kurdi, Mohamed-Zakaria. 2002. Combining pattern matching and shallow parsing techniques for detecting and correcting spoken language extragrammaticalities. 2nd Workshop on Robust Methods in Analysis of Natural Language Data, Italy, pp. 1–9.

McKelvie, David. 1998. The syntax of disfluency in spontaneous spoken language. Technical Report HCRC/RP-95, Edinburgh University, Edinburgh, Scotland.

Meteer, M., A. Taylor, R. MacIntyre, R. Iver. 1995. Dysfluency annotation stylebook for the Switchboard corpus. Distributed by LDC.

Müürisep, Kaili, and Heli Uibo. 2006. Shallow Parsing of Spoken Estonian Using Constraint Grammar. In Proceedings of NODALIDA special session on treebanking (ed. Peter Juel Henrichsen and Peter Rossen Skadhauge). Copenhagen Studies in Language #33/2006.

Nakatani, Christine, Julia Hirschberg. 1993. A Speech-First Model for Repair Detection and Correction. Proceedings of the 31st Annual Meeting on Association for Computational Linguistics, pp. 46–53.

Spilker, Jörg, Martin Klarner, and Günther Görz. 2000. Processing Self-Corrections in a Speech-to-Speech System. W. Wahlster (ed.) *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, pp. 131–140.

Stolcke, Andreas, and Elizabeth Shriberg. 1996. Statistical Language Modeling for Speech Disfluencies. Proceedings of the International Conference on Audio, Speech and Signal Processing.