Disfluency Detection and Parsing of Transcribed Speech of Estonian

Kaili Müürisep

and

Helen Nigol

University of Tartu

The corpus

- 1,065,000 words
- 1,703 transcripts
- 100,000 words manually POS-tagged
- Different types of spoken language: everyday and institutional conversations, spontaneous and planned speech, monologues and dialogues
- Max. authentic situations

Example

H: [ee:] oskate ´öelda kas ´homme oleks võimalik pool´teljed Renool ´ära vahetada.

V2: m ´homme ´kindlasti ei ole võimalik seda ´as[ja]

H: [ei]=´ole=jah   (1.0)

V2: jah   (.) ja mis ´auto teil ´ültse [on {---} ei=no]

H: [RE´NOO RE´NOO.   Renoo üks´teist.]

V2: ja=ja=ja mis prob´leem teil on.   üks[´teist]=ä

H: [ta]

H: jah tal on: mõlemad pool ´teljed nagu vaja ´ära vahetada.

Disfluencies

- **False starts**

  A sentences is aborted before completion and a new sentence is started.

  – I should it is time to finish
- **Repetition**

  A word or a phrase is repeated twice or more.

  – This this this will take a long time
- **Self-repair**

  A word or a phrase is corrected by deleting, substituting or inserting words

  – Please two tickets to Wroclaw no I mean to Poznan

Estonian CG Parser

- In the style of the first version of Constraint Grammar
- Designed for written language and then adapted for spoken language
- Tagset: SUBJ OBJ PRD ADVL +FMV -FMV +FCV -FCV P> <P Q> <Q NN> <NN AN> <AN PN> <PN etc.
- Ca 1200 syntactic constraints and 50 clause boundary detection rules.
- 88-90% unambiguos, 2% errors

Parsing (1)

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

Parsing (2)

- (@w=s! (@P>) (0 Subst) (0 Gen)(1 Postp)(1 #Gen))
  - Select @P> if the word form is substantive genitive and the next word form is a postposition and accepts only genitive complements
- (@w=s0 (@OBJ) (NOT *-1 TrVerb)(NOT *1 TrVerb) **CLB)
  - Remove @OBJ if there is no transitive verb in left and right context inside inner clause boundaries

Parsing (3)

```
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 //
```

Adaption of grammar

- New POS – special particles – *ahah*, *mhmh*, *hurraa*, *jess*, *ee*, *õõ*, *noh* etc.
- New syntactic labels:
  @B – syntactically independent uninflected words;
  @T – unknown syntactic function.
- compile new rules for the sentence internal clause boundary detection
- fix the syntactic constraints (slight modifications of less than 100 rules from 1200)

Preliminary results

- The word count in the corpus: 2194
- Errors: 68
- Recall: 96.9% (98.5%)
  - correctly found / ideal
- Precision: 89.5% (87.5%)
  - correctly found / all found
- Unambiguity rate: 92.9% (89.5%)

New corpora

- 8400 words training corpus
- 6700 words benchmark corpus
- 13000 words disfluency corpus
  - Repairs
  - Repetitions
  - False starts
    - Annotated: I should +/ it is time to finish
    - Normalized: it is time to finish
    - Input of the parser: I should it is time to finish

Detection of clause boundaries

- Parser considers a speech turn in dialogues as a unit of analysis (sentence).
- Pauses are marked by punctuation marks – parser uses them for detecting clause boundaries.
- Some particles and adverbs are used in the beginnings or ends of clauses.
- Addition of rules for special cases gave growth of 0.2% in correctness and 0.2% in unambiguity rate.

False starts

- False starts are detectable if they contain a verb:

H: ei ma: tahakski: umbes +/ ma uurin praegu ´hinda

H: no I would_like approximately +/ I am doing background search on prices

False starts

```
K           ####
$<s>
muna        muna+0 //_S_ com sg nom //   **CLB @SUBJ                                   ;; egg
noh         noh+0 //_B_ //   @B                                                        ;; well
see         see+0 //_P_ dem sg nom //   @<NN                                           ;; this
siia        siia+0 //_D_ //   @ADVL                                                    ;; here
asemele     asemele+0 //_D_ //   @ADVL                                                 ;; instead of
tuleks      tule+ks //_V_ main cond pres ps3 sg ps af #FinV #Intr //   @+FMV           ;; should
leida       leid+a //_V_ main inf #NGP-P //   @OBJ                                     ;; find
midagi      miski+dagi //_P_ indef sg part //   @OBJ                                   ;; something
muud        muu+d //_P_ indef sg part //   @<NN                                        ;; other
ma          mina+0 //_P_ pers ps1 sg nom //   **CLB-C @SUBJ                            ;; I
soovitaks   soovita+ks //_V_ main cond pres ps1 sg ps af  //   @+FMV                   ;; suggest
hapukoort   hapu_koor+t //_S_ com sg part //   @OBJ                                    ;; sour cream
$.           . //_Z_ Fst //
$</s>
```

Repetitions

- miks miks miks peab
  - *why why why one should*
- aga sa aga sa peaksid
  - *but you but you should*
- noh see see on tähtis
  - *noh this this is important*

Self-repairs

- Fragments of words as markers of self-repairs

  väga nor- väga normaalne noh väga naiss
  very nor- very normal noh very nice

- Triggers of repair

  V: ee ja programmis on=ää, (1.2) tändab hinna=sees on=nüd=lennu´piletid

  V: ee and in the program there are ää (1.2) this means the flight tickets are in the price

- Undetectable

  X: aga mingit firmat kes seal seda prügiveoga tegeleb

  X: but any company who there this deals with garbage collecting

Self-repairs (2)

```
väga [ADVL]
    väga+0 //_D_ //     **CLB @ADVL
!!!nor- [T]
!!!    nor //_T_  #- //
!!!väga [ADVL]
!!!    väga+0 //_D_ //
normaalne [PRD]
    normaalne+0 //_A_ pos sg nom //    @PRD
noh [B]
    noh+0 //_B_ //    @B
väga [ADVL]
    väga+0 //_D_ //    @ADVL
naiss [T]
    naiss //_T_ //    @T
```

Self-repairs (3)

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

Results

	Written	Before	Now
Words		2543	6717
Recall	98.5	97.3	97.7
Precision	87.5	89.2	90.4
Unambiguity	89.5	91.5	93.0

Results (2)

	Normalized Prec/Rec	**Original** Prec/Rec	**Now** Prec/Rec
Repairs	87.6/96.4	84.9/94.6	85.5/95.0
Repetitions	91.8/98.6	90.7/98.2	92.1/98.6
False starts	93.8/98.9	90.0/97.4	91.1/98.1

Conclusions

- Clause boundary detection is the key issue.
- Automatic identification of disfluencies would help a lot.
- Next challenge is the adaption of morphological disambiguator to spoken language and to focus more on self-repairs.
- Look also the other structural incompletenesses of utterances.