



Language technology
at University of Tartu –
joint research of
computer scientists and linguists

Kaili Müürisep & Co



Human Language Technology

- Language technology is often called human language technology (HLT) or natural language processing (NLP) and consists of computational linguistics (or CL) and speech technology as its core but includes also many application oriented aspects of them.
Language technology is closely connected to computer science and general linguistics [Wikipedia].



The players

Institute of Computer Science

~10 people

+

Institute of Estonian and General Linguistics

~10 people

=

Informal research group of computational linguistics



Language resources

- Corpora = (structured) amounts of text
- Corpus of Written Estonian 1890-1990
- The Mixed Corpus of Estonian
 - Balanced corpus (newspaper texts+fiction+science texts)
 - fiction, newspapers, journals, legal documents, parliament transcriptions, chat rooms, Ph.D. theses ...
- Annotated corpora
- Corpus of Old Written Estonian
- Corpus of spoken language



Morphological analyzer

- Find all sentences containing word “mees” (man) from corpus



Morphological analyzer (motivation)

- Find all sentences containing word “mees” (man) from corpus
- But how to find “mehed” (men)?
- Estonian grammar has 70-100 model words, 2x14 cases.
- mehed
 - mees+d // S pl nom //



Morphological analyzer

- bases on the lexicon and morphological rules
- 66,000 entries
- still 5% of words unknown (names, abbreviations, non-Estonian words, but also nomens and verbs)
- *klikkima*, *blogima*, *guugeldama*, but also *ostlema*



Morphological analyzer – compound words

- ca 24,000 compound words in lexicon
 - täisarv = täi_sarv
 - laekaunistus = laeka_unistus
 - pihuarvutitest = pihu_arvu_titest
 - fondihaldus = fond_ihaldus
 - Lainersiga, Perelõgina



Morphological Disambiguation

- We looked for the word “mees”
 - mees+0 // S sg nom //
 - mesi+s // S sg in //
- The correct reading depends on the context
- 50% of word forms are ambiguous
- some of word forms may have up to 10 readings



Morphological Disambiguation

- Rule based approach (1200 rules)
- Morphological system of Estonian is too complicated for data driven approaches (>800 combinations of tags)



Morphological Disambiguation

- 15% remain ambiguous, 2% errors
- **Otsi vead üles!**
- **Keel** on võimas realiteet
- **Jälgi** koolimatemaatika arenguteel
- Täpselt **neid** asju muidugi ei tea, kuid tasuks uurida.



Shallow Syntactic Tagging

- Estonian has almost free word order -
 - the knowledge of syntactic functions is extremely valuable
 - subject, object, predicative, attribute, adverbial ...
 - alus, sihitis, öeldistäide, täiend, määrus ...
 - ~1200 regular expression like rules
 - shallow = no direct linkage between words
 - professori (NN>) nahast (NN>) portfell



Example

<s>

Mitmekesisus

mitme_kesi=sus+0 // _S_ com sg nom #cap // **CLB @SUBJ

on

ole+0 // _V_ main indic pres ps3 sg ps af #FinV #Intr // @+FMV

elu

elu+0 // _S_ com sg gen // @NN>

vaieldamatu

vaieldamatu+0 // _A_ pos sg nom // @AN>

omapära

oma_pära+0 // _S_ com sg nom // @PRD

\$.

. // _Z_ Fst //

</s>



Shallow Syntactic Tagging

- Estonian has almost free word order -
 - the knowledge of syntactic functions is extremely valuable
 - the system must be robust

*ee ja kui kaua h mi mis ajal peab puh noh enne
kaitsmist ütleme mul on viieteistkümnes millal ma
pean selle bakalaureusetöö esitama*



Parsing

- To build tree like structures
 - phrase structure trees
 - dependency trees



Syntax Learning

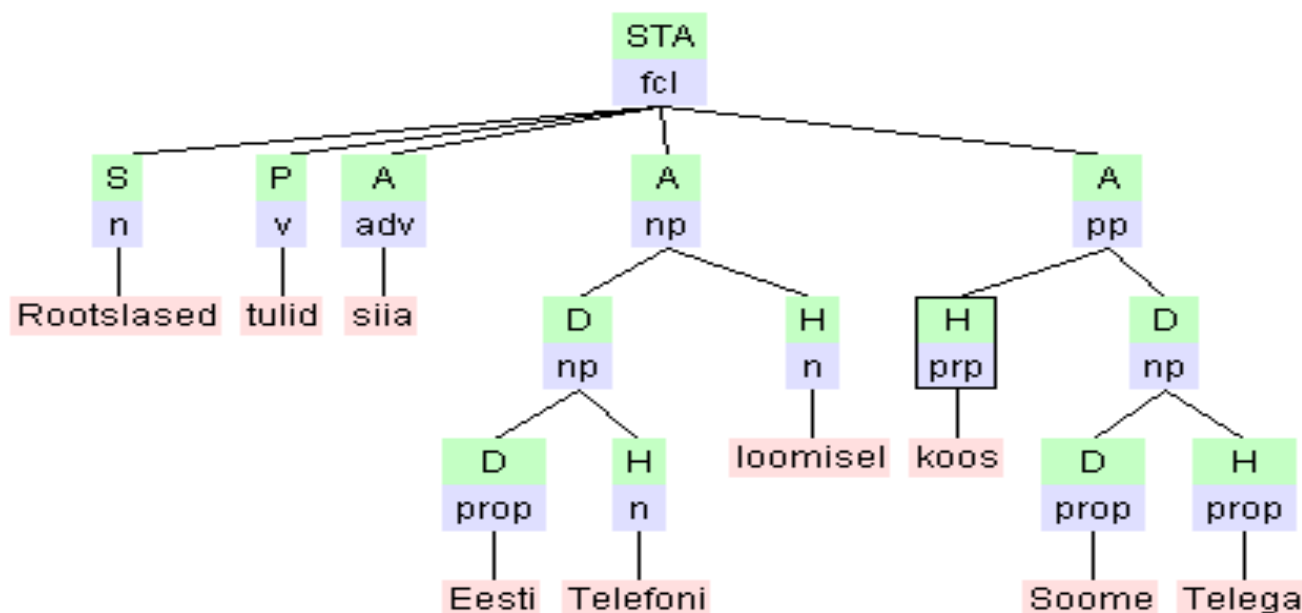
Language Settings Tools Help

Rootslased tulid siia Eesti Telefoni loomisel koos Soome Telega

Op As Ao Cs Co fA fApass fC fCsta fCvoc H DN DNc DNapp DA DAcom DP Dfoc P Vm Vaux Vp

v-fin v-inf v-pcp1 v-pcp2 art pron adj adv prp num conj-s conj-c intj infm np ap pp vp fcl icl acl pa

head preposition ("koos+0" pre %kom)





Semantic analysis

- Wordnet
 - train < vehicle
- Semantic frames
 - Pall lendas aknasse
- Semantic disambiguation
 - Teeme ühe pitsi



Dialogue Systems

- Dialogues have an inner structure
- No every question is followed by answer
- Kas sa tead palju kell on?



Toy tools

- Reisiagent
- Teatriagent
- Zelda & Voldemar



Applications

- EstSum – newspaper text summarizer
- Grammar Checker
- Learners aids – error detectors, grammar games



Machine Translation

- käesolev määrus jõustub järgmisel päeval pärast selle avaldamist euroopa ühenduste teatajas .
- this directive shall enter into force on the day of its publication in the official journal of the european communities .



Machine translation

- the european economic community and the swiss confederation of agreements between the application of the joint committee shall apply in the community decision no 5 / 81



Machine translation

- [the european economic community and the swiss confederation] [of agreements between] [the application] [of the joint committee] [shall apply in the community] [decision no 5 / 81]
- euroopa majandusühenduse ja šveitsi konföderatsiooni vaheliste kokkulepete kohaldamisel rakendatakse ühenduses ühiskomitee otsust nr 5 / 81



Machine translation

- [the european economic community and the swiss confederation]₃ [of agreements between]₂ [the application]₁ [of the joint committee] [shall apply in the community]₅ [decision no 5 / 81]₄
- **for the purposes of** application of the agreements between the european economic community and the swiss confederation decision no 5 / 81 of the joint committee shall apply in the community