university of groningen

Topics I would like to Work On

Gosse Bouma

Centre for Language and Cognition
University of Groningen
g.bouma@rug.nl

Oslo SynSem, Feb 2017

Parsing Closely Related Languages

university of groningen

State of the Art in Parsing

We are getting pretty good at standard language:

> *"Grammar parsing is relatively advanced; it is the domain of the well-established field of 'natural-language processing'."*
>
> (The Economist Tech Review, Jan 2017)

Parsing Closely Related Languages

university of groningen

State of the Art in Parsing

We are getting pretty good at standard language:

> *"Grammar parsing is relatively advanced; it is the domain of the well-established field of 'natural-language processing'."*
>
> (The Economist Tech Review, Jan 2017)

Other?

- Historical Text
- Minority Languages
- Social Media

Treebank for Frisian?

university of groningen

Language Situation

- Frisian is a language spoken in Fryslan (province of the Netherlands), approx 500.000 speakers
- Recognized minority language, established spelling, newspaper, TV, twitter, Frisian Academy

Treebank for Frisian?

university of groningen

Language Situation

- Frisian is a language spoken in Fryslan (province of the Netherlands), approx 500.000 speakers
- Recognized minority language, established spelling, newspaper, TV, twitter, Frisian Academy

Language Technology?

- Frisian present in Google Translate!
- Can we build a lemmatizer, POS-tagger, Treebank for Frisian?
- Resources for Dutch (and German and English) exist, can they help?

Grammar Profiling

university of groningen

Grammatical Profiles

Can we characterize different genre's, languages, or users in terms of the (frequency of the) grammatical constructions they use?

Examples

- Differences in Literary Genre's
- Language of essays written by high school children

Grammar Profiling

university of groningen

Detecting cross-linguistic Syntactic Differences Automatically (PhD project)

- Investigate the possibility of automatic detection of syntactic differences between languages by using on-line parallel corpora
- Which differences do we find in the Germanic languages with respect to the structural positions of verbs and with respect to verbal inflection?
- Is the existing theory of verb placement and inflection as it has been developed since the late eighties of the 20th century capable of capturing these facts?

Grammar Profiling

university of groningen

Essays and Writing Style

- One of the secrets of writing coherent and attractive prose is varying the word order and sentence types across sentences (e.g. Myhill 2008; Christie 2009).
- We will build a tool that aims to capture this word order variation...an extension to the well-known dependency parser Alpino.
- It will be useful to provide counts such as the following:
  - 'elliptical' sentences lacking subjects or heads
  - sentences with direct, indirect and prepositional objects
  - sentences with non-initial subjects
  - sentences followed by sentences with the same initial unit (e.g. two subject-initial sentences)
  - ...

Psychological Relevance

university of groningen

More Grammar Profiling....

> *Many recent models of language comprehension have stressed the role of distributional frequencies in determining the relative accessibility or ease of processing associated with a particular lexical item or sentence structure. However, there exist relatively* ***few comprehensive analyses of structural frequencies****, and little consideration has been given to the appropriateness of using any particular set of corpus frequencies in modeling human language.*

D. Roland et al, *Frequency of basic English grammatical structures: A corpus analysis*, J of Mem and Lang, 2007

Psychological Relevance

university of groningen

My boss thinks [(that) I'm absolutely crazy].

That-deletion and Information Density

Uniform Information Density predicts that language production is affected by a preference to distribute information uniformly across the linguistic signal... production is probability-sensitive, in that speakers' preferences are affected by the contextual probability of syntactic structures.

*speakers should be more likely to produce full complement clauses (CCs with **that**) than reduced CCs (without **that**), the higher the information of the CC onset in its context*
$I(CC|context)$ *is estimated as* $-log\ p(CC|matrix\ verb)$

F Jaeger, Redundancy and reduction: Speakers manage syntactic information density, Cog Sci, 2010