Linguistics with CLARIN OpenSONAR

Jan Odijk

LOT Winterschool

Amsterdam, 2015-01-13

Overview

- SONAR
- OpenSONAR
- Methodological Considerations
- Google?

Overview

- **SONAR**
- OpenSONAR
- Methodological Considerations
- Google?

SONAR

- SONAR Dutch corpus
- 500 million tokens
- Written language
  - (for spoken language: CGN)**
- Many different text types
  - Includes `new media‘ (sms, tweets, blogs, …)
  - but not balanced (mainly because of legal restrictions)

SONAR

- FoLIA Format
- Pos, lemma, word properties for each token
- `metadata' for each document
- Created in the STEVIN-project (2004-2011)
- Can be obtained via TST-Centrale
- Reference:
- Oostdijk Nelleke, Martin Reynaert, Véronique Hoste, Ineke Schuurman (2013). `The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch', In [Spyns & Odijk 2013]. [pdf]

SONAR

- Some interesting Annotated Text Corpora
  - English
    - British National Corpus
    - Corpus of Contemporary American English (and many more at BYU)
    - American National corpus
  - Multiple languages
    - CHILDES Corpora
  - German
    - Das Deutsche Referenzkorpus

SONAR

- Some interesting Annotated Text Corpora
  - Spanish
    - Syntactic Spanish Database (SDB) University of Santiago de Compostela. 160,000 clauses / 1.5 million words.
    - Ancora-ES (and Ancora-CA) and others
    - Panacea Annotated Corpus (downloadable)
    - Corpus Molinero (but no annotations)
    - Corpus Tecnic de l'IULA
  - Dutch
    - Corpus Gesproken Nederlands (CGN)
    - SONAR en SONAR Nieuwe Media
    - VU-DNC
    - Discan
  - …

Overview

- SONAR
- **OpenSONAR**
- Methodological Considerations
- Google?

OpenSONAR

- Search interface to the SONAR Corpus
- Some Interfaces to Corpora for other lgs:

Interface	Language(s)
BNCweb interface at Lancaster	British English
IMS Open Corpus Work Bench	German
Corpus of Contemporary American English	American English
Corpus of Contemporary Dutch	Dutch
TrovA	Multiple
Språkbanken	Swedish
Corpuscle	Norwegian
Bwananet	Spanish, Catalan, ..

OpenSONAR

- Search interface to the SONAR Corpus
- Runs on INL (Instituut voor Nederlandse Lexicologie), one of the Dutch CLARIN Centres
- http://opensonar.clarin.inl.nl/
- Login with the account of your institute
  - Federated login, single sign on (CLARIN)
- created in the CLARIN-NL project (2009-2014)
- Available since November 2014 (!)

OpenSONAR

- Back-end based on BlackLab, developed at INL
  - Open Source Software, based on Apache Lucene
  - https://github.com/INL/BlackLab#readme
  - https://github.com/INL/BlackLab/wiki/BlackLab-blog
- Front-end developed by UvT, `Whitelab'
  - Open Source Software
  - https://github.com/INL/WhiteLab

OpenSONAR

- 4 interfaces
  - Simple, extended, advanced, expert
  - Expert = CQP language (CQL)
- Grouping, Restricting by metadata
- Pos-codes:
  - Van Eynde, Frank (2004), `Part Of Speech Tagging en Lemmatisering Van Het Corpus Gesproken Nederlands', Centrum voor Computerlinguïstiek, K.U.Leuven [pdf]

OpenSONAR

- See Scenario demo OpenSONAR

Overview

- SONAR
- OpenSONAR
- ➢ **Methodological Considerations**
- Google?

Methodological Considerations

- Performance (actually used) data
  - Including errors, hesitations, fillers, etc
  - Good for certain research questions
  - Less good for other research questions
- No `negative' data
  - Linguists sometimes want to know what is NOT possible in language
  - More difficult to find non-standard examples (e.g. examples not covered by the grammar used for a treebank)

Methodological Considerations

- Danger of circularity
  - 'Which verbs occur with a predicative adjective?'
  - ➔the verbs that have been specified as such in the grammar underlying a treebank
  - Can be avoided by globally knowing how the relevant grammar works
- No controlled experiments
  - Minimal pairs seldom occur naturally
  - BUT: Corpora/Treebanks can be used to construct minimal pairs on the basis of really occurring examples

Methodological Considerations

- Annotations have mainly been made by automatic programs
  - They make errors
    - `absurd errors'
    - Insufficient information errors
  - People also make errors but different ones
    - `sloppiness errors'

Methodological Considerations

- Large corpora:
  - high frequency results are more reliable results
  - low frequencies are suspect
- Small corpora:
  - human verification and correction is required

Methodological Considerations

- Desired:
  - get all relevant examples (high recall)
  - no or few irrelevant examples (high precision)
- Very difficult to achieve
  - Critical analysis of the results is always required

Methodological Considerations

- User friendly interface implies limitations:
  - Cf. OpenSONAR interface (advanced: no extended pos (inflectional information)
  - Several examples can be given for GrETEL

Methodological Considerations

- Simple cases can be solved by small adaptations in the query, e.g.
  - Start with the graphical interface
  - Adapt in the expert interface
    - Adapting easier than creation from scratch

Overview

- SONAR
- OpenSONAR
- Methodological Considerations
- **Google?**

Google?

Property	Google	What you want
String search	yes	yes
Relation between strings	nearness	Grammatical relations
Search for function words	No / unreliable	Yes
Search for morpho-syntactic and syntactic properties	no	Yes

Google?

Property	Google	What you want
Search within a sentence, paragraph?	No (documents only)	Sentence, paragraph, section etc
results	List of documents	List of sentences, paragraphs, sections, documents
Grouped /sorted (analyzed) results	no	yes
Construction search	no	Yes
Single language only	unreliable	Yes
Size	huge	Huge (but so far there is only small (1m tokens) to large (500m tokens)

Thanks for your Attention!