Your boldest wishes concerning online corpora: OpenSoNaR and you

Martin Reynaert

TiCC, Tilburg University and CLST, Radboud Universiteit Nijmegen

TiCC Colloquium, Tilburg University. October 16th, 2013

TILBURG UNIVERSITY

Outline

1. Section 1: SoNaR Overview
2. Section 2: SoNaR-500 Content Highlights
3. Section 3: Towards OpenSoNaR
4. Section 4: A peek at other online corpora
5. Section 5: Your own role in OpenSoNaR

SoNaR: Patron & Consortium

- Patron:
  - Dutch Language Union (Nederlandse Taalunie) in the STEVIN programme
- Consortium:
  - Netherlands: Radboud University Nijmegen, Tilburg University, Twente University & Utrecht University
  - Belgium (Flanders): University College Ghent, KU Leuven

STEVIN Nederlandstalig Referentiecorpus (SoNaR) project

- Directed at the actual construction of a 500 MW reference corpus of contemporary standard written Dutch as encountered in texts originating from the Dutch speaking language area in Flanders and the Netherlands as well as translations published in and targeted at this area
- Where possible, (re)use of formats, tools, protocols, and annotation schemes previously developed (D-Coi, but also COREA, DPC, Lassy)

500 MW Reference corpus (SoNaR-500)

- includes native speaker language and the language of (professional) translators
- approx. two-thirds of the data originate from the Netherlands and one-third from Flanders
- only texts included from the year 1954 onwards, in practice: only born-digital texts
- comprises full texts rather than text samples
- includes texts from a wide range of 40 text types incl. texts from the new media
- for (almost) all data IPR has been arranged

SoNaR-500: Automatic linguistic annotations

- automatic sentence splitting and tokenization: ILKTOK
- automatic POS tagging and lemmatisation and morphological analysis using FROG
    - for all data
    - tagset is essentially the set used in the CGN project
- automatic NE labeling by means of NERD
    - Named Entity Recognition for Dutch: NE classifier developed on the basis of SoNaR-1

Contents

- Besides the texts: all available metadata has been recorded
- Over 30 text types
  - old and new media
  - all contemporary, written Dutch. Essentially all born-digital.
  - everything from books to tweets...
- Only discuss a few of the highlights here...

Old Media: Highlights: Books

- Largest donation by Dutch publisher Maarten Muntinga
  - 145 titles: travel, thrillers, chick lit, life style, etc.
  - almost 14 MW, almost no IPR-restrictions
  - great news for KNAW project "The riddle of literary quality"
- 4 versions of the Bible
  - 3 protestant, 1 catholic
  - N-gram frequency lists are a shambles!!

Old Media: Highlights: Magazines/Periodicals

- Largest donation by Flemish publisher Roularta
  - all CD-roms published between 1991 and 2004
  - weekly delivery of new issues of their several titles
  - no commercial use
  - no language restriction: huge comparable corpus Dutch-French a possibility!
- Largest Dutch donation: De Groene Amsterdammer
  - 12 years' editions

Old Media: Highlights: Newspapers

- Largest Flemish donation: Mediargus
  - 1.3 BW ’available’ to the community, no IPR-agreement
  - IPR-restriction 1: ’Take only what is required’
  - have taken Edition 2006 of 4 main Flemish newspapers.
  - IPR-restriction 2: no commercial use
- Largest Dutch donation: PCM / De Persgroep
  - Selection from Twente News Corpus
  - 60MW

Old Media: Highlights: Reports

- NIOD (Institute for War, Holocaust and Genocide Studies)
  - Srebrenica report
  - over 7,000 pages of text
  - probably best known report in Dutch ever
  - potential for building a parallel corpus: English version also online

Old/New Media Highlights: Subtitles

- Largest Flemish donation: VRT
  - wealth of TV-series, a.o. De Kampioenen, ...
  - 18.5MW
  - even more still in ’surplus’
- Open Subtitle Corpus (Tiedemann)
  - balanced set of film subtitles, produced by volunteers
  - classified as both Dutch and Flemish
  - huge surplus of over 384MW available...

New Media: Highlights: Discussion Lists

- Politics.be
  - largest internet forum in Flanders. IPR-settled.
  - site delivered over 480MW, over 2 years ago...
  - Incorporated only 25MW...
- TMF: Flemish youth forum
  - highly dialectical...
  - about 70% of posts classified by language recognizer TextCat as: "‘I don’t know; Perhaps this is a language I haven’t seen before?’"
  - 20MW were incorporated...

New Media: Highlights: Chats

- Chats
  - chats NL with IPR-agreement and metadata: 0.5MW
    - elicited, mainly collected in schools and within a research team
  - chats NL no IPR-agreement and without metadata: 8 M
    - natural
  - chat VL with IPR-agreement, without metadata: 9 M
    - natural

New Media: Highlights: Tweets & SMS

- Twitter:
  - did not yet exist in 2005, no target in SoNaR...
  - 19 MW incorporated
  - token count prior to tokenization...
- SMS
  - collected in two parallel campaigns in NL and VL
  - a prize was offered to donators
  - a well balanced (in terms of 1/3 VL vs. 2/3 NL) set of over 50K SMS were collected
  - these represent about 620K SMS-words

Concluding remarks on the SoNaR corpus

- SoNaR project represents an excellent return on investment:
  - SoNaR-500 definitely allows to ’sound’ most dimensions of contemporary, written Dutch
  - SoNaR-1 has yielded world-class, large gold standards for semantic annotation
- SoNaR fills major gaps in the Dutch language resources infrastructure
  - BUT: Its 2.1M text files accompanied by 2.1M metadata files present a major practical obstacle to most anyone
  - Which is why we are building OpenSoNaR: a system for “Online Personal Exploration and Navigation of SoNaR"

Opening remarks on the OpenSoNaR project

- We want SoNaR online in as open a fashion as possible: usable for schoolkids and researchers alike
    - You, researchers, are hereby invited to tell us what you expect from this system
    - Please provide us with your wish lists, user cases
- The Dutch Language Union and the IPR agreements with text providers impose some restrictions: these we will honour
- The whole corpus is essentially free for non-commercial research purposes

Intentions of OpenSoNaR

- Address this corpus exploration and exploitation problem
- Develop and make available an open corpus exploration and exploitation environment
- Tailor the front-ends to the desiderata of 4 CLARIN-NL priority groups of users

Means employed by OpenSoNaR

- Intended originally to build on the well-known Corpus Workbench back-end
- Turned out INL had been developing a new alternative
- Java system BlackLab, based on Apache Lucene, open source via GitHub
- OpenSoNaR will build Whitelab front-ends according to the user groups specifications

OpenSoNaR User Groups in more detail

CLARIN-NL posited 6 ‘priority user groups’. OpenSoNaR addresses priority groups 2 to 5:

- Literary Sciences: Huygens ING, Karina van Dalen-Oskam, assisted by colleagues at Utrecht University
- Cultural Sciences: Meertens Institute, Nicoline van der Sijs, assisted by Ewoud Sanders
- Linguistics: TiCC: Jan Renkema & Ad Backus & colleagues
- Communication and Media Studies: TiCC: Fons Maes & Emiel Krahmer & colleagues

Nederlab

- URL: https://www.nederlab.nl/onderzoeksportaal/
- NWO Groot project
- Goes far beyond OpenSoNaR
- 4M euro budget vs. 150K, runs 5 years vs. 1
- Diachronic vs. synchronic: ‘all’ Dutch corpora, since about the year 800
- Laboratory for research
    - Exploration of the corpora: through time, space, genres: metadata
    - Exploitation of the corpora: possible to make your own subcorpus, put that in your private or privately shared work space
    - Analysis of the corpora: tools for all kinds of text/topic mining, comparison making, measuring particular features, ...
- Let us take a look at the Clariah Demonstrator...

Språkbanken (the Swedish Language Bank)

- http://spraakbanken.gu.se/eng/start
- 147 corpora to date, 1.4 Billion word tokens
- ‘Korp’ (E.: crow): the corpora
- ‘Karp’ (E: carp): the dictionaries
    - From bird’s eye view of the text to below surface information
    - Mediated by wonderful little icons
    - Below surface information from dictionaries, Wordnet, automatic parsing etc.
- Based on older technology: Corpus Work Bench
- Uses Corpus Query Language
    - More elaborate than OpenSoNaR
    - Less so than Nederlab

Brieven als Buit: “Why does he not write?"

- URL: http://brievenalsbuit.inl.nl/zeebrieven/page/search
- About 1,000 17-18th century Dutch letters
- Subset of the ’Gekaapte Brieven’: mail taken as loot by privateers and confiscated by the High Court of Admiralty during the wars fought between The Netherlands and England
- Digitized and annotated by Dutch Institute of Lexicology (INL)
- INL corpus back-end system BlackLab also the basis for OpenSoNaR
- OpenSoNaR will at first have the functionalities of this site
- Let’s get acquainted!

OpenSoNaR vs. these three online resources

To summarize:

- Språkbanken
  - Definitely a great source for inspiration and emulation
- Brieven als Buit
  - Great start that already will provide good basic functionality to further build on
- Nederlab
  - If the OpenSoNaR project provides, Nederlab will no doubt be enhanced by its accomplishments

OpenSoNaR: online soon

- We are setting up a server for the OpenSoNaR website
    - URL: http://opensonar.uvt.nl
- Will host the ‘rudimentary’ OpenSoNaR soon, corpus is being indexed at INL
- We will keep you informed, of course
- Practical guidelines for submitting wish-lists, case descriptions and for getting feedback via our issue tracker will be made available asap.

openSoNaR

openSoNaR

Thanks!!

Thanks for your attention!

Papers about SoNaR are available at:
`http://ilk.uvt.nl/`

Your boldest wishes concerning online corpora:
OpenSoNaR and you

Martin Reynaert

TiCC, Tilburg University and CLST, Radboud Universiteit Nijmegen