T-(538|725)-MALV, Natural Language Processing Corpora

Hrafn Loftsson¹  Hannes Högni Vilhjálmsson¹

¹School of Computer Science, Reykjavik University

September 2009
Outline

1. Corpora

2. Examples of corpora

3. Utility of corpora
- A corpus (í. málheild) is a collection of texts or speech stored in an electronic (machine-readable) format.
- A corpus often contains material compiled using certain rules decided upon in advance.
- Called a text collection (í. textasafn), rather than a corpus, if it contains randomly selected texts.
- Huge corpora, tens (or hundreds) of millions of words, are available in many languages today.
Types of corpora

Genres

- Specific genres, e.g., law, science, novels, news text, etc.
- Wider variety of texts:
  - To survey comprehensively and accurately a language usage.
  - “Balancing a corpus”.
  - Costly task.
- Linguistic Data Consortium [http://www.ldc.upenn.edu/](http://www.ldc.upenn.edu/)

Annotations

- Either, raw text without annotations, or
- Text with annotations (i.e., merkingar/skýringar).
Types of corpora

Genres

- Specific genres, e.g. law, science, novels, news text, etc.
- Wider variety of texts:
  - To survey comprehensively and accurately a language usage.
  - “Balancing a corpus”.
  - Costly task.
- Linguistic Data Consortium http://www.ldc.upenn.edu/

Annotations

- Either, raw text without annotations, or
- Text with annotations (í. merkingar/skýringar).
## Corpora with annotations

### What kind of annotations?

- Each word labeled with a linguistic tag (í. míðfræðilegt mark)
- For example, **part-of-speech** (í. orðflokkur), **constituent** (í. setningararliður), **semantic category** (í. merkingarflokkur)
- Carried out manually and/or semi-automatically.

### Treebank (í. trjábanki)

- A corpus, in which the syntactic structure of sentences is shown.
  - For example, a collection of parse trees.
- Penn Treebank (University of Pennsylvania) is probably the best known treebank.
Corpora with annotations

What kind of annotations?

- Each word labeled with a linguistic tag (í. málfræðilegt mark)
- For example, part-of-speech (í. orðflokkur), constituent (í. setningarliður), semantic category (í. merkingarflokkur)
- Carried out manually and/or semi-automatically.

Treebank (í. trjábanki)

- A corpus, in which the syntactic structure of sentences is shown.
  - For example, a collection of parse trees.
- Penn Treebank (University of Pennsylvania) is probably the best known treebank.
1 Corpora
2 Examples of corpora
3 Utility of corpora
An example of a corpus

Penn Treebank

- http://www.cis.upenn.edu/~treebank/
- About 5 million words.
- POS tagged with a tagger.
  - http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
- Syntactically annotated with a parser.
An example from the Penn Treebank

POS tagged text

- The/Dt grand/JJ jury/NN commented/VBD on/IN a/Dt number/NN of/IN other/JJ topics/NNS ./
- DT=Determiner(Article) (í. ákvæðisörd (greinir))
- JJ=Adjective (í. lýsingarörd)
- NN=Noun (í. nafnorð)
- VBD=Verb, past tense (í. sögn í þátíð)
- IN=Preposition or subordinating conjunction (í. forsetning eða aukatenging)
- NNS=Noun, plural (í. nafnorð í fleirtölu)
Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.
An example of a corpus

The British National Corpus (BNC)

- http://www.natcorp.ox.ac.uk/
- 100 million words.
- A balanced corpus.
- Tagged with a tagger.
  - http://www.natcorp.ox.ac.uk/docs/bnc2postag_manual.htm
An example of a corpus

The Negra Corpus

- http://www.natcorp.ox.ac.uk/
- 355,000 tokens of German newspaper text.
- Tagged with part-of-speech.
- Annotated with syntactic structures.
<table>
<thead>
<tr>
<th>Word</th>
<th>POS</th>
<th>Case</th>
<th>Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Es</td>
<td>PPER</td>
<td>3.Sg.Neut.Nom</td>
<td>PH</td>
</tr>
<tr>
<td>spielt</td>
<td>VVFIN</td>
<td>3.Sg.Pres.Ind</td>
<td>HD</td>
</tr>
<tr>
<td>eben</td>
<td>ADV</td>
<td>MO</td>
<td></td>
</tr>
<tr>
<td>keine</td>
<td>PIAT</td>
<td>Fem.Akk.Sg</td>
<td>NK</td>
</tr>
<tr>
<td>Rolle</td>
<td>NN</td>
<td>Fem.Akk.Sg.*</td>
<td>NK</td>
</tr>
<tr>
<td>,</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>ob</td>
<td>KOUS</td>
<td>CP</td>
<td></td>
</tr>
<tr>
<td>die</td>
<td>ART</td>
<td>Def.Fem.Nom.Sg</td>
<td>NK</td>
</tr>
<tr>
<td>Musik</td>
<td>NN</td>
<td>Fem.Nom.Sg.*</td>
<td>NK</td>
</tr>
<tr>
<td>gefällig</td>
<td>ADJD</td>
<td>Pos</td>
<td>PD</td>
</tr>
<tr>
<td>ist</td>
<td>VAFIN</td>
<td>3.Sg.Pres.Ind</td>
<td>HD</td>
</tr>
</tbody>
</table>
An example of a corpus

The Icelandic Frequency Dictionary (IFD) (í. Íслensk orðtíðnibók)

- ≈ 590,000 tokens.
- A balanced corpus:
  - Icelandic fictions, translated fictions, biographies, educational material, children and teenager books.
- Tagged with a tagger (by Stefán Briem) and hand-corrected.
An example from the IFD

ég fp1en // word tag
stökk sfg1eþ // See explanation of tags
á aa // in a document under ‘‘Other material’’
eftir aþ // on the course web page
strætó nkeþ
og c
veifaði sfg1eþ
, ,
vagnstjórinn nkeng
sá sfg3eþ
mig fp1eo
og c
stoppaði sfg3eþ
..
An example of a corpus

A large Icelandic corpus

- Being compiled at The Árni Magnússon Institute of Icelandic studies (í. Stofnun Árna Magnússonar í íslenskum fræðum).
  - http://www.arnastofnun.is/page/arnastofnun_frontpage_en
- 900 text snippets, 25 million words.
- http://www.lexis.hi.is/malheild.htm
1 Corpora

2 Examples of corpora

3 Utility of corpora
Utility of corpora

- The construction of word lists and dictionaries.
- Research in linguistics; corpus linguistics.
- A precondition for the development of various LT tools, e.g.:
  - Taggers
  - Syntactic parsers
  - Machine translation systems (which often utilise parallel corpora).