Detecting semantic overlap: 
Announcing a Parallel Monolingual 
Treebank for Dutch

Erwin Marsi & Emiel Krahmer

CLIN 2007, Nijmegen
What is semantic overlap?

- Zangeres Christina Aguilera heeft eindelijk verteld waarom haar buik zo dik is. [NOS]

- Christina Aguilera heeft in het Amerikaanse tijdschrift Glamour bevestigd dat zij zwanger is. [AD]

- Christina Aguilera heeft eindelijk bevestigd wat de hele wereld al wist: ze is zwanger. [NOVUM]

- Iedereen wist het al, maar nu zou Christina Aguilera het zelf voor het eerst hebben toegegeven: ze is zwanger. [The Agenda]
Why bother?

- Similar information can be expressed ("paraphrased") in many different ways.

- Major **stumbling block** for robust NLP applications such as IE, IR or QA.

- **Resources** exist on the word level (e.g., Wordnet), but are mostly lacking for more complex phrases.

- The Stevin **Daeso** (Detecting and Exploiting Semantic Overlap) project intends to fill this gap.
The Daeso corpus

- Building a 1M word parallel monolingual treebank.

- Basic idea: look for pairs of sentences where there is an independent criterion that there will be some amount of overlap.

- The corpus should contain different text genres and different amounts of overlap.

- 500K manually aligned and corrected [now]; 500K automatic [2009]
Corpus collection

- **Autocue - Subtitling** (NOS, TwNC)

- **Parallel translations** into Dutch

- **Google Headlines** (mined by Wauter Bosma)

- **Different press releases** (ANP, Novum) about the same (Dutch) event.

- **Potential sets of answers** to different questions (from the IMIX project).
## Corpus data

<table>
<thead>
<tr>
<th>Corpus Type</th>
<th>Manual</th>
<th>Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autocue-subtitles</td>
<td>125k</td>
<td>192k</td>
</tr>
<tr>
<td>Book translations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Darwin 1</td>
<td>25k</td>
<td>154k</td>
</tr>
<tr>
<td>Darwin 2</td>
<td>25k</td>
<td>191k</td>
</tr>
<tr>
<td>Montaigne 1</td>
<td>25k</td>
<td>462k</td>
</tr>
<tr>
<td>Montaigne 2</td>
<td>25k</td>
<td>~500k</td>
</tr>
<tr>
<td>Saint-Exupéry 1</td>
<td>15k</td>
<td>15k</td>
</tr>
<tr>
<td>Saint-Exupéry 2</td>
<td>15k</td>
<td>15k</td>
</tr>
<tr>
<td>News headlines</td>
<td>24k</td>
<td>&gt; 900k</td>
</tr>
<tr>
<td>Press releases</td>
<td>ANP</td>
<td>125k</td>
</tr>
<tr>
<td></td>
<td>Novum</td>
<td>125k</td>
</tr>
<tr>
<td>QA system output</td>
<td>1k</td>
<td>1k</td>
</tr>
</tbody>
</table>
Pre-processing and annotation steps

1. XML TEI format (Text Encoding Initiative).

2. Sentence splitting and tokenization with the DCOI tokenizer for Dutch (Reynaert 2007).

3. Dependency parsing with the Alpino parser (van Noord et al.).

4. Alignment at text and sentence level.

5. Alignment of dependency trees.
Sentence alignment

- Standard alignment methods (e.g., Gale and Church 1993) assume alignment is mostly 1-to-1 and that crossing alignments and unaligned sentences are rare.

- These assumptions are often violated.
  - Obviously in comparable texts
  - But also in e.g., translations of Darwin’s Origin of Species

- Developed:
  - A new alignment method to boost manual annotation.
  - A new annotation tool to check sentence alignments
Automatic sentence alignment

- Tricky for comparable texts

- As a first approximation: low level, multiple pass, shallow features.

- Experiments with:
  - types vs token;
  - different overlap metrics (MaxSim, Cosine, Jaccard, Dice, Tanimoto, ...);
  - tf-idf weighting (Nelken & Schieber 2006)

- Ongoing...
Hitaext: Tool for text and sentence alignment

First public release (October 2007): http://daeso.uvt.nl/hitaext/
Alignments of words and phrases

- Given two dependency trees for two aligned sentences: align nodes and label the alignment relation.
  - “Christina Aguilera” equals “Christina Aguilera”
  - “zwanger” restates “in verwachting”
  - “de zangeres Aguilera” specifies “Aguilera”
  - “Aguilera” generalizes “de zangeres Aguilera”
  - “Christina Aguilera en Beyoncé” intersects “Beyoncé en Pink”

- Marsi & Krahmer (2005): for first five chapters of “Le Petit Prince”, two annotators reached an F-score of .98 on relations and .95 on labels.
Algraeph: Tool for aligning nodes and labeling alignments
State of affairs

- Work on manual alignment of words and phrases currently *ongoing*.

- Other work on the corpus is now *finished*.

- **Further activities:**
  - *Sentence fusion*: combine two related sentences into a single grammatical sentence. New results on question-driven fusion just in.
  - *Multi-document summarization*: currently building a baseline multi-document summarization system for Dutch, to be extended with Daeso tools later on.
About the Daeso Stevin project

- **People involved:** Paul van Pelt, Jurry de Vos, Iris Hendricks, Walter Daelemans, Jakub Zavrel, Maarten de Rijke, Erwin Marsi, Emiel Krahmer

- **More info:**

  http://daeso.uvt.nl/