**COMPARA, the largest ever post-edited English-Portuguese parallel corpus**

www.linguateca.pt/COMPARA/

**Contents**
- Translated published fiction 1810 - 2002
- Open ended, currently (version 7.0) 72 pairs, 33 authors and 45 translators
- Several varieties of Portuguese (Brazil, Portugal, Angola, Mozambique) and English (UK, USA, South Africa)
- Around 30% of each book
- More than one translation per work
- Meta-information: source language variety, translation language variety, source date, translation date
- Alignment units: one sentence in the source text, with whatever its translation is
- Kind of alignment
- Ordering changes, additions
- Translation notes
- Paragraph shifts
- Typographically salient material classified as titles, foreign words or expressions, named entities, or other emphasis

**Measuring PoS ambiguity in Portuguese**

<table>
<thead>
<tr>
<th>PoS</th>
<th>Types</th>
<th>Tokens</th>
<th>ambiguous types</th>
<th>ambiguous tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>nouns</td>
<td>19,845</td>
<td>268,636</td>
<td>1378</td>
<td>6.943</td>
</tr>
<tr>
<td>adjectives</td>
<td>9,641</td>
<td>64,493</td>
<td>1378</td>
<td>14.29</td>
</tr>
<tr>
<td>PoS</td>
<td>Types</td>
<td>Tokens</td>
<td>ambiguous types</td>
<td>ambiguous tokens</td>
</tr>
<tr>
<td>adverbs</td>
<td>1,343</td>
<td>67,380</td>
<td>94</td>
<td>3.648</td>
</tr>
<tr>
<td>adjectives</td>
<td>9,641</td>
<td>64,493</td>
<td>49</td>
<td>0.508</td>
</tr>
<tr>
<td>total</td>
<td>10,953</td>
<td>131,873</td>
<td>49</td>
<td>2.42</td>
</tr>
<tr>
<td>verbs</td>
<td>35,253</td>
<td>281,925</td>
<td>471</td>
<td>1.903</td>
</tr>
<tr>
<td>adjectives</td>
<td>9,641</td>
<td>64,493</td>
<td>671</td>
<td>6.559</td>
</tr>
<tr>
<td>total</td>
<td>44,223</td>
<td>346,418</td>
<td>671</td>
<td>1.517</td>
</tr>
<tr>
<td>PoS</td>
<td>Types</td>
<td>Tokens</td>
<td>ambiguous types</td>
<td>ambiguous tokens</td>
</tr>
<tr>
<td>verbs</td>
<td>35,253</td>
<td>281,925</td>
<td>1149</td>
<td>3.259</td>
</tr>
<tr>
<td>nouns</td>
<td>19,845</td>
<td>268,636</td>
<td>1149</td>
<td>5.789</td>
</tr>
<tr>
<td>total</td>
<td>53,949</td>
<td>548,561</td>
<td>1149</td>
<td>2.129</td>
</tr>
</tbody>
</table>

**Design**
- Parallel interface (in the 2 languages)
- Two modes for interaction
- Strict separation between selection and presentation options
- Constant concern with usability
- Large documentation effort
- No need for registration

**Grammatical annotation**
- Automatic parsing with PALAVRAS (Bick, 2000)
- Human revision and documentation

**Users**
- Large number of users
- Constant increase in use

**Post-editon**
- Sentence separation
- Sentence alignment
- PoS classification
- Tokenization (including MWE)
- Morphosyntactical features
- Future: Syntactic function

**May 2006: 3 million words**

Diana Santos  Diana.Santos@sintef.no  Susana Inácio susinacio@sapo.pt