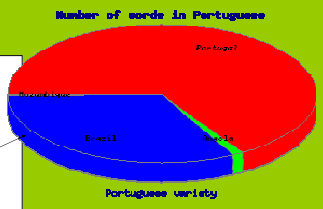
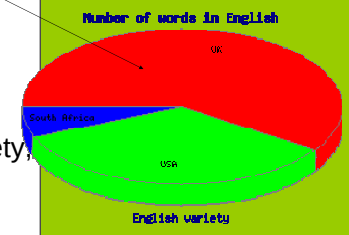
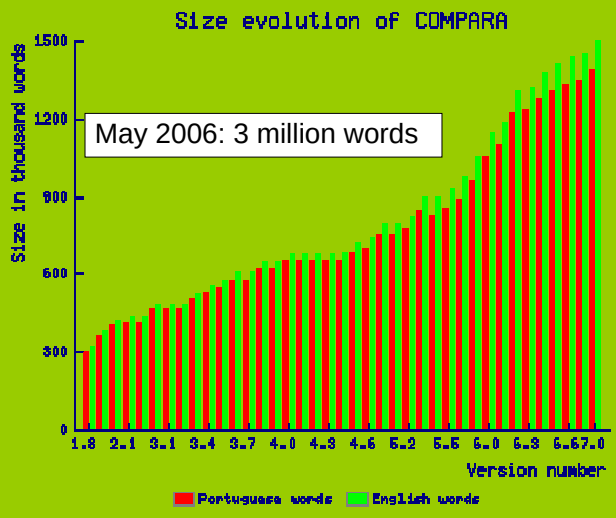
COMPARA, the largest ever post-edited English-Portuguese parallel corpus

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

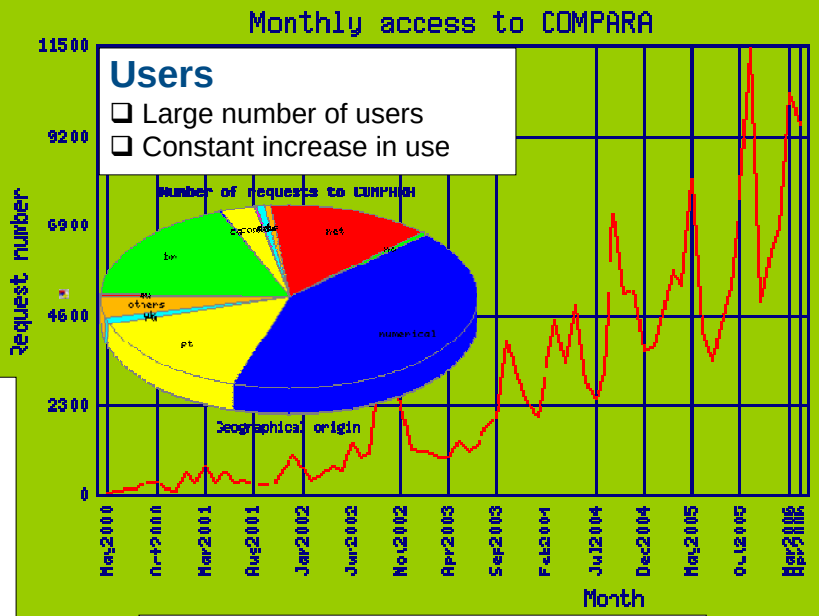
- ❑ Meta-information: source language variety, translation language variety, source date, translation date

- ❑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

Grammatical annotation

- ❑ Automatic parsing with PALAVRAS (Bick, 2000)
- ❑ Human revision and documentation



Users

- ❑ Large number of users
- ❑ Constant increase in use

Post-edition

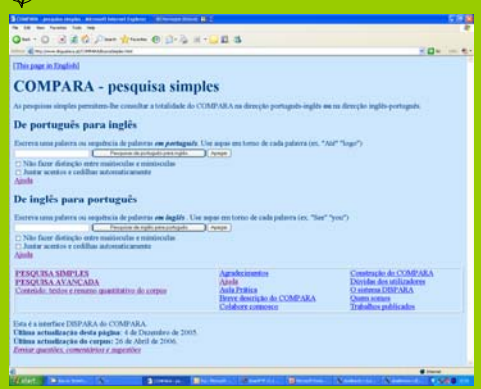
- ❑ Sentence separation
- ❑ Sentence alignment
- ❑ PoS classification
- ❑ Tokenization (including MWE)
- ❑ Morphosyntactical features
- ❑ Future: Syntactic function

Measuring PoS ambiguity in Portuguese

PoS	Types	Tokens	ambiguous types		ambiguous tokens	
			#	%	#	%
nouns	19,845	266,636	1378	6.943	48,193	18.07
adjectives	9,641	64,493	1378	14.29	48,193	74.72
total	28,108	331,129	1378	4.902	48,193	14.55
PoS	Types	Tokens	ambiguous types		ambiguous tokens	
			#	%	#	%
adverbs	1,343	67,380	49	3.648	10,784	16.00
adjectives	9,641	64,493	49	0.508	10,784	16.72
total	10,935	131,873	49	2.42	10,784	8.177
PoS	Types	Tokens	ambiguous types		ambiguous tokens	
			#	%	#	%
verbs	35,253	281,925	671	1.903	17,969	6.37
adjectives	9,641	64,493	671	6.959	17,969	27.86
total	44,223	346,418	671	1.517	17,969	5.19
PoS	Types	Tokens	ambiguous types		ambiguous tokens	
			#	%	#	%
verbs	35,253	281,925	1149	3.259	82,787	34.57
nouns	19,845	266,636	1149	5.789	82,787	31.04
total	53,949	548,561	1149	2.129	82,787	15.09

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



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