

## Against multilinguality

Diana Santos, [www.linguateca.pt](http://www.linguateca.pt)

## Contents

- Introduction
- Multilinguality as an ontological monster
- Multilingual resources
- Multilingual corpora
- Bilingual methodological issues

## Motivation

- Presentation of Linguateca aims and *raison d'être*
- Healthy discussion on what corpora are for
- What weight should be assigned to building instead of using?
- Not comfortable with the multilingual label (or language-independent)

## What is Linguateca

- Improve Portuguese processing
  - Dissemination
  - Resource creation
  - Evaluation
- A virtual organization with four nodes
  - Oslo, Braga, Lisbon, Oporto, ...
  - Collaboration partners in more locations: Odense, Lisbon, São Carlos, Porto Alegre, ...
- Originally created as the *Computational Processing of Portuguese* project, in 1998, by the then Ministry of Science and Technology

## Assumptions of Linguateca

- First things first
  - Find out what are the problems and bottlenecks of Portuguese processing
- International entities or bodies cannot solve our problems
  - In any case not better than us
  - Resource building is time consuming, and "market driven"
- Language (and not region, or nation) should be the unit for natural language processing
  - So Brazil and Portugal should cooperate closely
- Public resources are a must for scientific progress
  - There are enough barriers already

## But: what I am doing here?

- Personal awareness of contrastive studies as one key for the advancing language understanding (and processing capabilities)
  - the subject of my PhD
  - parallel corpus methodology a key issue
- Machine translation as the NLP task par excellence
- To be primarily concerned with one's own language does not mean lack of interest for the processing of other languages
- Lately engaged in a Portuguese-English bilingual corpus, COMPARA
  - to my knowledge the largest edited corpus for this pair
- Starting a comparable corpus endeavour with the Oporto node

## A multilingual corpus is for

- studying multilingual tasks
- understanding multilingual contexts
- investigate issues that cannot be found or investigated in bilingual or monolingual corpora

what are they ???

## Multilingual tasks?

- What knowledge of multilinguality above bilinguality can help?
- In what tasks can I profit from a multilingual expert?
- What multilingual resources would I like to use in my everyday life?

## What is multilinguality? a multilingual system?

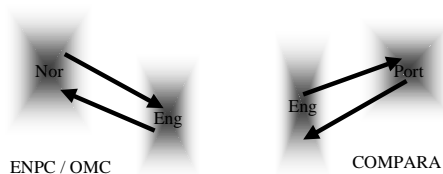
- *The linguistic behaviour of human beings who use several different natural languages in their daily life...*
- Handling several languages well enough? ... *too weak*
- Involves translation between more than two languages
  - translation is an eminently bilingual task
- Involves a "language-independent", above languages, concept representation
  - this is a myth
- Involves language as a variable or as an output
  - "meta-natural language" applications, such as language identification, or cross-language tasks (the ranking part)

## Multilingual corpora

- Use the information explicit in other languages to do monolingual studies / processing --- fair enough, but expensive!
- Study the Web
  - a subset
  - different sampling requirements
  - how little a part will be comparable, translation, commensurable
- What is a multilingual corpus?
  - Texts in more than two languages ... *not enough*
  - Texts (explicitly) related to other texts in more than one language
  - Search in more than one language (pair) at the same time
  - Texts explicitly encoded with a language value

## One example of using multilingual corpora

- contrast two prepositions (standard translations of each other)
- Portuguese *com* and Norwegian *med* (both translated by *with*)
- A fictive trilingual corpus



## Comments

- This was a realistic situation
  - but it was not really a multilingual corpus
- Quantitative data for two language pairs
  - does not allow one to infer the pattern for the third one
- Not even consensual what conclusions to take from a bilingual case, due to
  - lack of enough data (across other language pairs)
  - lack of firm methodological ground



## Multilingual corpus encoding

- Most decisions as how to put things in correspondence are defined by the particular pair
- If one has different pairs for which corresponding units differ, there is simply no single encoding where you may be both faithful to monolingual practice of the different languages and have consistent encoding of the correspondences

■ fhasdha: hagdh; hgad      2-2      ■ gsahas. hfhsa!gfahga;jhhkj. B

■ fhasdha: hagdh; hgad      1-1      ■ fhasdha hagdh, hgad. C

A

## The use of the corpus

- There is no general purpose dictionary
- There is no general purpose problem solver
- There is no general purpose (multilingual) corpus
  
- what are words? sentences? clauses? errors? meanings?
- this is all-important for corpus processing (and encoding) of
  - sentence alignment
  - word correspondence
  - clause correspondence
  - semantic studies
  - translation studies

## Concluding remarks

- *Why* build multilingual corpora is preliminary to *how*?
- How are bilingual corpora going to be used?
- We should be pursuing methodological issues
- We should be gathering data (results) for subsequent theory building
- Standards for what?
  - Different projects / corpora have different goals
  - If a format is documented it is easy to convert it to others, if the need arises to exchange it
  - Most corpus encoding has to rely on subjective introspection
  - Not necessarily everything is possible to document
- Give others access and obvious options become questioned
- Let users, not compilers, define how.

Table 13. Kinds of mismatch when *com* and not *with*

Kind	translation from E->P (random in 5 pairs COMPARA)		translation from E->P (all in 2.5 pairs, ENPC, En-Po)	translation from P to E (random in 10 pairs COMPARA)	Total	
Different preposition	61	30.5%	21	27	109	27%
Direct object vs. PP( <i>com</i> )	29	14.5%	16	13	58	14.5%
Adverb	29	14.5%	15	18	60	15.5%
Adjective	21	11.5%	10	3	34	8.5%
Verb vs verb+PP( <i>com</i> )	13	6.6%	6	3	22	5.3%
Verb vs. PP( <i>com</i> )	4	2%	3	4	11	2.8%
discourse	12	6%	4	7	23	5.8%
<i>and</i> vs. <i>com</i>	2	1%		4	6	1.5%
reordering	29	14.5%	25	21	75	18.8%
» explicit args	2		6			
» head swithcing	2		6	7		
» <i>parecido com</i>			2			
Total	200		100	100	400	

Are these (relevant) results?

In a Portuguese-to-Norwegian bilingual dictionary...

- ...88% of the occurrences of *com* do not have *med* in the translation
- ...94% of the occurrences of *med* do not originate from entries having *com*
- ...Only 22% of the occurrences of *com* and 16% of the occurrences of *med* have respectively *med* or *com* counterpart

In Portuguese corpora ...

- ...the frequency of the word *com* ranges from 43 to 112 occurrences per 10,000 words.
- ... the frequency of occurrence of *com* ranges from 12 to 26 occurrences per 100 sentences
- ...around 30% of PPs with *com* have an adverbial function
- ...around 17% of PPs with *com* modify NPs
- ...around 10% of PPs are prepositional objects (arguments) of verbs
- ...around 6% of PPs are N<PRED (postnominal nexus predicative in small clause; predicative adjunct) / discourse
- ...around 3% of PPs with *com* modify AJs

In Norwegian corpora...

- ... the frequency of the word *med* ranges from 93 to 108 occurrences per 10,000 words

In English-Portuguese parallel corpora...

- ...63% of *com* are translated by *with*
- ...74% of *with* are translated by *com*
- ...45% of *com* in translated text does not correspond to *with*
- ...44% of *with* in translated text does not correspond to *com*

In English-Norwegian parallel corpora...

- ...55% of *med* are translated by *with*
- ...72% of *with* are translated by *med*
- ...51% of *med* in translated text does not correspond to *with*
- ...27% of *with* in translated text does not correspond to *med*

When *com* is not translated by *with*...

- ...27% are simply alternative prepositions
- ...3.3% are translated by *and*

When there is the pattern *with* in English and not *com* in Portuguese...

- ...48% are due to alternative prepositions
- ...1.7% is translated by conjunction (*e*)

## We need

- similar information in order to compare different corpora
    - what is a word
    - what is a sentence
    - what is a translation-unit
  - huge amounts of quantitative information about contrastive data so that we can assess what is normal, what is special
    - per language pair
    - per original - translation
    - per phenomenon
    - per kind of text
- > Criteria to base contrastive studies**
- huge amounts of quantitative information about monolingual data
    - per language
    - per syntactic phenomenon
    - per word
- > Criteria to base monolingual studies**