Against multilinguality

Diana Santos, 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 confortable 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)

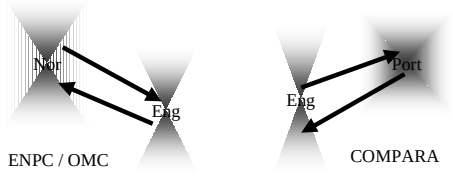
Multilingual corpora

- 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

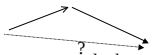
- contrast two prepositions (standard translations of each other)
- Portuguese *com* and Norwegian *med* (both translated by *with*)
- A fictive trilingual corpus



ENPC / OMC COMPARA

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. hfhgsa!gfahga:jhhkjhkj. 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 instrospection
  - Not necessarily everything is possible to document
- Give others access and obvious options become questioned
- Let users, not compilers, define how.