Linguateca@Porto

Belinda Maia & Luís Sarmento
Presentation at SINTEF, 10 Sep. 2003

What is Linguateca

- Improve Portuguese processing
- Dissemination
- Resource creation
- Evaluation
- A virtual organization with four nodes
- Oslo, Braga, Lisboa, Porto, … 5 full-time, 3 part-time workers
- Collaboration partners in more locations: Odense, Lisbon, São Carlos, Porto Alegre.
- A follow-up of the Computational processing of Portuguese project, created in 1998, by the then Ministry of Science and Technology

Linguateca activities: the IRE model

- Dissemination of information and resources on Portuguese processing
- Web catalogue with a dedicated search engine
- Forum and a contact service
- Creation of publicly available language resources
- Making the available resources more available: Web services
- Creating new ones; both Web and physical access
- Promotion of joint evaluation using the evaluation context or evaluation campaign model
  - Web site and discussion list [Eval]
  - Organization of a workshop (June 2002) and a conference (AVALON’03)
  - Organization of the first evaluation contest for Portuguese: [ME-191,201,330,741]
  - Evaluation of machine translation systems

The Porto node of Linguateca

- Initiated by Belinda Maia
- Started October 2002
- Embedded in the Arts Faculty of Porto, Linguistics Centre
- Closely related to the Masters and PhD studies in Translation and Technology Studies
- Contacts with several other subject areas (Engineering, Geography, Medicine, Arts History…)
- Already around 25 different local users of the Linguateca reality

The Porto and Oslo nodes compared

<table>
<thead>
<tr>
<th>Porto</th>
<th>Odo</th>
</tr>
</thead>
<tbody>
<tr>
<td>RE</td>
<td>IRE</td>
</tr>
<tr>
<td>Primarily concerned with language of specific domains</td>
<td>Primarily concerned with general language</td>
</tr>
<tr>
<td>Therefore, specialized corpora</td>
<td>Therefore, general corpora</td>
</tr>
<tr>
<td>Therefore, parallel corpora: comparable corpora</td>
<td>Therefore, parallel corpora: translation corpora</td>
</tr>
<tr>
<td>Evaluation of machine translation systems</td>
<td>Evaluation of morphological analysers and other simpler systems</td>
</tr>
</tbody>
</table>

The Porto node of Linguateca: activities

- R Creation of a “corpus manager”
  - to facilitate specific corpus creation and terminology extraction
  - to provide a comparable corpus environment
- R Activities in MT evaluation (ATA)
- R Support to other resource building projects
  - CHAT corpus
  - BNC in CQP format
  - Grammar teaching using corpora
Porto’s activities - background

- Contrastive linguistics (CA) - using literary corpora: e.g. Doctoral thesis (PT/EN) Maia (1994)
- Translation teaching: Same teachers teach CA, TT and Translation – and computer technology applied to translation
- Therefore teaching and research tend to blend

Moving the goalposts in translation teaching methodology

- From ‘general’ and literary translation > special domain translation because 95% of translation internationally is NON-literary
- From ‘Word Processing’ > translation software and the Internet
- From (often non-existent or outdated) ‘specialized dictionaries’ > online databases and specialized corpora building
- Realization of need for change often better understood by teachers whose research involved corpora, contrastive linguistics, use of IT and research into IT e.g. LETRAC report, CULT conferences etc
- Development of new teaching and research methodologies

Changes in teaching methodology

- Student projects begin to include construction of mini ‘do-it-yourself’ / ‘disposable’ corpora
- Corpora work associated with term extraction and glossary building
- Resulting awareness of wide varieties and different levels of text types as well as different functions of texts
- Increased comprehension of importance of observing terminology in context
- The Internet as essential source of information
- Connection between corpora and terminology database building with commercial translation software – e.g. TRADOS, STAR, SDLX, Deja Vu etc

Teaching > Research

- Master’s in Translation Studies included:
  1. Translation and Linguistics – 45 hours
  2. Translation and Information Technology
- Master’s in Terminology and Translation
  > Emphasis on NON-literary translation
  > Terminology project work
  > Corpora building
  > Etc

Master’s in Terminology and Translation - Projects

- 2001-3
  - Geography Department – multilingual dictionary ‘Population Geography’
  - Several small projects with cooperation of Faculty of Engineering – Mechanical Engineering
  - Others: Vinho Verde, History of Wool textiles, Metaphors in Football corpora
- BUT – Lack of IT support – appeal for help to Linguateca

Master’s in Terminology and Translation + Linguateca

- 2002-4
  - Geography department – Natural Hazards > corpora + database
  - Other projects in Civil Engineering, Genetics, Medicine
- 2003-5
  - Continuation and development of existing projects
  - Addition of possible History of Art project
Parallel Corpora in terminology and translation research
- Parallel corpora in specialized domains
- Advantages: IF translations are GOOD and LINEAR
  - E.g. EC documentation, Canadian Parliament, multinational companies’ webpages
- Disadvantages are:
  - Bad translations
  - Specialized domain texts - and especially webpages - are often ADAPTED or LOCALIZED
- Difficult to find
- Uses of GOOD Parallel corpora for research:
  - Automatic term extraction
  - Observation on linguistic phenomena in translation

Comparable Corpora in terminology and translation research
- Degrees of comparability:
  - In ‘general’ and literary texts – date, genre, individual style, male/female authors
  - In special domain texts – everything from general information texts – encyclopedia articles – pedagogical textbooks - specialist-to-specialist communication.
- BUT most work so far on comparable corpora is in specialized domain corpora

Choosing texts for Comparable Corpora
- Select your project and define your ends
  - E.g. ISO standards or Instruction manuals
- Take what you can find!
  - Not always easy to find balanced corpora in an English dominated world – especially at academic levels

Research with Comparable Corpora
- IMMEDIATE
  - Text analysis > contrastive work on text and sentence structure, observation of syntax and use of general lexicon in special domain
  - Terminology extraction based on actual usage
    - Corpora used for a specific project > Disposable as limited ‘shelf life’ and applications
- LONG-TERM
  - IF Good Comparable Corpora are constructed they can become official corpora
    - Approved + copyright clearance etc. > can be used for testing
    - For testing information retrieval, ontologies and document creation tools ETC
  - For projects coordinating corpora with Machine Translation

Our “Corpus Manager” - GC

Live demonstration of the “corpus manager”...
A future comparable corpus manager

- Modules for
  - automatic discovery of terminology candidates
  - automatic discovery of definitions and semantic relations
  - automatic “putting in correspondence” of terms in two languages
  - automatic harvesting of similar texts
- A user-friendly environment for
  - a terminologist to study and create terminologies
  - a translator to learn about cultural conventions in different registers
  - a linguist to study specialized languages
  - an MT researcher to train MT systems in specialized domains
  - an NLP researcher to get lexicon, grammatical and stylistic resources

Evaluation of Machine Translations

- Two motivations:
  - Evaluation of NLP Systems
  - Pedagogical Implications
  - Demonstration of the first initiatives

Evaluation contests for Portuguese

- Model: agree on what should be compared, on the criteria, on the measures and on the workflow. One of the most difficult things is specify an agreement platform where there is consensus.
- Translation is probably the most difficult (human) activity that has to do with language…
- But conversely a layman can (and does) produce quality judgements on translation (subtitles, books, news, TV interviews…)
- No way to start comparing systems in their entirety...
- Let us start to gather a set of specific problems where (machine) translation is considered to fail, and see whether there is some consensus.

Pedagogical uses

- To demonstrate the state-of-the-art in order to:
  - Remove fear of MT
  - Show possibilities of using MT as tool for professional translator
  - To study the lexicon, syntax and semantics of two languages in a special situation and raise the students’ awareness of the difficulties involved
  - To train future translators to work in MT environments – i.e. To:
    - Create and improve specialised lexicons for MT
    - Edit text input and output
    - Understand and study technical writing and controlled writing
    - To create interest in research in this area

Live demonstration of TrAva and CORTA...