DISPARA, a system for distributing parallel corpora on the Web

Diana Santos, SINTEF Telecom and Informatics
Computational processing of Portuguese

What is DISPARA?
- A system for distributing
- parallel aligned corpora
- on the Web
- built on top of the IMS CWB
DISPARA was first developed in connection with the COMPARA corpus. COMPARA/DISPARA is a collaboration project with Ana Frankenberg-Garcia, under the framework of the Computational Processing of Portuguese project

Why present DISPARA?
- Focus on the generality of the system
  - so far it has only been presented in connection with COMPARA
- Insist on the importance of making the work available on the Web
  - and therefore OS independent
- Augment the user community (for both corpus users and corpus providers)
  - informing about the work involved

Presentation plan
- Parallel corpora
- What is involved in Web distribution
- DISPARA proper: encoding, search and display options
- DISPARA context: COMPARA
- DISPARA at work: users, use, problems

Uses of parallel corpora
- To study/teach translation
- To study/teach the differences and similarities between languages
- To create systems
  - that do NLP tasks (machine translation, word sense disambiguation, terminology extraction,...)
  - that help translators
- To do cultural studies

Parallel corpora...
- are hard to compile
  - not all kinds of text are translated
  - translation quality differs widely
  - for each text one needs (at least) two more permissions
- come in two flavours
  - aligned (the same content)
  - comparable (the same kind)
Examples of parallel corpora
- MLCC-DEB EU parliamentary debates
- ECI-MCI EU calls
- ENPC / OMC
- COMPARA
- NILC abstracts
- the Web
- multilingual IR test collections

Ways of using parallel corpora
- Exploration
- Hypothesis test
- Evaluation of systems
- Creation of exercises or didactic material
- Dictionary or terminology construction
- Creation of NLP systems

Distributing corpora on the Web
- A set of texts
- A corpus encoding system
- A set of encoding options
- A Web interface

Encoding options - macro level
- More than one translation per original
- Alignment always 1-x (alignment unit)
- Encoding of translation notes
- Encoding of addition, deletions and reorderings at sentence level, as well as alignment type
- Encoding of variant and date (both for original and translation(s))
- Encoding of titles, foreign expressions and simple emphasis

Encoding options - micro level
- What is a token?
- What is a sentence?
- What to do with spelling errors?
- What to do with ordinary notes, or headings?
- What is reordering?
- What is addition?
- What is deletion?

The corpus encoding system
- DISPARA is based on IMS-CWB abilities
- Features heavy AC/DC reuse
- In the IMS Corpus Workbench, alignment is one kind of annotation, previous to actual search
- INTEX is a dynamic annotation system, an environment to develop grammars and lexicons (linguistic resources)
**Search options vs display options**
*two independent things!*
- Search: restrict by all encoded information
  - basic and advanced search modes
- Display: distribution of the results by all encoded information
  - quantitative wrap up; distribution of forms
  - show translation notes
  - special display of sentence reordering
  - context size; maximum number of hits
  - basic error handling

**Users of DISPARA**
1. Us
2. Real users
3. Prospective users
   - translators
   - teachers
   - NLP developers
4. Ideal users
   - people who would also comment / give feedback / help with annotation / gather texts

**Use of DISPARA (serving COMPARA) since May 2000**
- 10,000 searches
- 1,062 - 1,342 different users/IP addresses
- 205,000 results returned
- 75% using the English interface
- 73% using the simple interface
- 9% using complex query restrictions
- 3% requiring special display capabilities

**Short presentation of COMPARA**
- 16 text pairs (10 Portuguese, 4 English originals) (+45 already copyright cleared)
- 284,582 Portuguese words (136,274 original)
- 296,910 English words (146,113 original)
- 17,639 alignment units (95.7% 1-1) (9,400 PE)
- (18 + 79) translation notes
- (250 + 267) titles; (443 + 206) foreign; (152 +191) emphasis

**Feedback/wishes from users**
- Actual reporting of problems and bugs
- Detection of recurrent problems by logs observation
- Asking for deleting examples from a concordance
  - *use a simple text editor*
- Augment the concordance context
  - *(w)kkel in the source but not in the target*
- Get the translation pattern of a particular word
  - *(r)quires word alignment*
- POS tagging, lemmatization, semantic tagging

**Further information**
- Try DISPARA at
  - http://www.portugues.mct.pt/COMPARA/
- Read about its instantiation in COMPARA in several papers and on the Web
- Other instantiations (proof of concept):
  - the SQUIRREL project: Norwegian - Portuguese (raw) - Portuguese (corrected) ...