Computational processing of Portuguese

Diana Santos
Signe Oksefjell

Overview of the talk

- Aims and rationale of this project
- What we have been doing
- Signe: Problems in compiling the catalogue
- Signe: Example of language resources and evaluation of one such language-aware resource
- Diana: Discussion on making resources available
- Diana: Several possible futures

Background

- The Portuguese Ministry of Science and Technology is aware of the importance of computational language processing
- There is very little going on in Portugal at the moment in this area
- In order to avoid accusations of partiality, they wanted to have someone outside the milieu -- but quite well informed about it -- to establish a plan and to study the problems

Diana had been leader of an NLP group in Portugal for ten years and had moved to Norway for personal reasons. She was willing to do this task

Project at SINTEF

- SINTEF provided the best working environment in Norway
- the Portuguese Ministry of Science and Technology (MCT) fully finances the project (one year, with planned extension of one further year, no longer commitments so far)
- future developments may prove that it is a good idea to be established at SINTEF
  - both projectwise, if more general NLP will be pursued
  - and from the point of view of human resources: hovedfag students in computer science
- project started end of May 1998

Main goals

- Survey the area, identifying
  - main needs
  - the participants
  - the resources available
- Prepare a plan
  - cooperation with companies
  - dedicated PhD program
  - other measures
- Write a document for discussion in the R&D community
- Try to bring about some changes
  - establish bridges among different groups
  - suggest projects
  - make available some resources

What we have been doing

- producing a Web catalogue of Portuguese language resources (including Brazilian) on the Web
  http://www.portugues.mct.pt/recursos.html
- contacting several industrial partners, distribution agencies, and researchers
- writing internal reports / recommendations / reviews for the MCT
- preparing the ground for
  - a resource network
  - some evaluation of the resources
  - further distribution of corpora and texts

Papers: Santos, Diana. "Disponibilização de corpora através da WWW", Proceedings of Computational Linguistics Workshop (Lisbon, May 1988)

Oksefjell, Signe and Santos, Diana. "Breve panorâmica dos recursos de português mencionados na Web", Proceedings of the 3rd conference on the processing of Portuguese (PA, November 1998)

PROBLEMS in cataloguing the field on the Web

- How to find the relevant Web pages?
  - ⇨Search engines
  - ⇨Cooperation
  - ⇨Coverage
  - ⇨Time perspective (information rot)
- How to control the quality of information in each site?
- How to maintain the pages; add new sites, avoid dead links, etc.?
- How to structure the pages after having found the links you wish to include?
  - ⇨Common features
  - ⇨Relative importance
  - ⇨Presentation of information (groups, projects, results)
  - ⇨Terminology

Translation Corpus A database of translated texts

"Vi har det visst ikke helt bra," hvisker Herman, og ser opp på moren som kaver med hendene, akkurat som om hun også er innestengt i et akvarium.

"Og hvor gammel er du da?"

Doktoren snakker merkelig, som om han forveksler Herman med en puddel.

"Jeg tror jeg nesten har kommet ut av tellingen.

Men jeg har fødselsdag til neste år også." (*Herman*, Lars Saabye Christensen)

"We are not completely fine," whispers Herman, looking up at Mother, who is wringing her hands as if she is also confined in an aquarium.

"And how old are you?"

The doctor is speaking strangely, as if he mistakes Herman for a poodle.

"I think I may have lost count, but I have another birthday next year too."

The Translation Corpus Aligner

(Program that automatically aligns an original text with the translated version of that same text, on the sentence level.)

The alignment program calculates which sentences to link. The matching depends on:

- the number of matching words in each sentence pair, based on a so-called anchor word list;
- the number of matching proper names;
- some special characters, notably punctuation marks some tags and attributes: <head>, , rend=italic/bold;

The values may be adjusted as a result of correspondence, or non-correspondence, in length as measured in number of characters.

Aligned text

<s id=LSC1.4.s95 corresp=LSC1T.1.4.s103>Doktoren snakker merkelig, som om han forveksler Herman med en puddel.</s>

<s id=LSC1T.1.4.s103 corresp=LSC1.4.s95>The doctor is speaking strangely, as if he mistakes Herman for a poodle.</s>

Matches in word list:

doctor* / doktor* lege*

poodle* puddel*

speak* snakke*

Evaluation of the English-Portuguese Aligner

Evaluate how important the English-Portuguese anchor word list is in the process of linking the sentences

1) Need aligned and proofread version of the original and translated texts;
2) Align the original and the translation WITH and WITHOUT the word list;
3) Match both versions in (2) with the already final version as described in (1);
4) View discrepancies; what did the alignment program do differently in the two versions;
5) Compare both versions with the final version; what went wrong WITH the word list, and what went wrong WITHOUT it?

Data and program resources

- Making resources available boosts work in the field
- It is no point making it all over again
- Sharing resources leads to improved resources (feedback, criticism, problems in merging lead the way)
- you may never get to USE your resources if you don't share them
- Having made resources available is a plus for further funding
- Making resources available is giving benefits to competition
- You have to do it any way all over again, other people's data are never good enough
- You spend all the time helping other people using your things instead of using it yourself
- Not showing the resources protects us
- Having owned resources is a plus for further funding ("nobody else has this...")

Large resources

- Large collections need continuous development / maintenance
  - users (other people trying to make use of what we did) can lead us in the right directions
  - users can provide further information (further ways to classify, more data, etc.)
- Large collections require cooperation between different groups
- Better SOME version than none at all
- Large collections need continuous development / maintenance
  - there is never a time when you can say they're finished
  - other people using previous versions prevent us from making improvements
- Large collections need strong management and guidelines: you can't leave people to do what they please with it
- Better no data than bad, unreliable data

Specific issues

- Should one use existing (printed dictionaries) in order to get computational lexicons?
- Should one try to merge the large number of existing texts in electronic form – in different formats and encoding schemes – to create computer corpora?
- Should one use available recordings to build speech corpora?
- Should one create dedicate programs or customize and use the ones already available?

Even though common sense would answer YES, the actual answer in most of the cases has been NO

And we all know too well that there is a great amount of unacknowledged work involved when adapting other people's code, cleaning dictionary and text files, etc.

Possible futures

- Go on with a specific project for Portuguese, involving commercial products
  - writing a computational grammar for Portuguese
  - producing application-specific Portuguese customization
- Creating an expertise center in a specific NLP technology, later on to be applied to further languages
- Broaden the scope of our interests in order to cooperate with SINTEF in particular projects
  - digital libraries
  - speech processing
  - multilingual tools
- Eventually making part of SINTEF and being fully financed in SINTEF ways