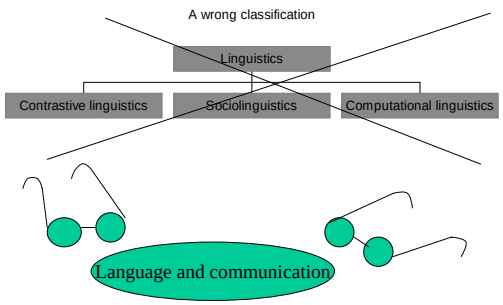
Contrastive linguistics and NLP

Diana Santos
Linguateca
Oslo, SINTEF

- what is NLP? what is CL?
- linguistics and NLP
- contrastive linguistics and NLP
- corpora...
  - methodology
  - evaluation
- communication
- some advice/pointers

What is the relationship



What is NLP / language technology / language engineering?

- ultimate goal
  - try to make computers
    - handle natural language
    - communicate with people
    - "understand" people
    - mediate among people
- meanwhile
  - help people in their linguistic/cognitive tasks
  - help linguistically-oriented people in their tasks

NLP to help people with

- linguistic / cognitive tasks
  - find information
  - write clearly and fast
  - understand something in a foreign language
  - book their plane tickets by phone
- linguistically-oriented people's tasks
  - compile dictionaries
  - write grammars
  - create/update a technical manual
  - translate

What is contrastive linguistics

- The study of the differences and similarities of two languages
  - to know both languages better
  - to know "language" better
  - to teach/learn the languages
  - to develop applications dealing with the two languages
  - to avoid misunderstandings / cultural pitfalls

Linguistics for NLP and vice-versa

- A vicious circle? "computer-aided"
  - linguistics / language related activities
    - language teaching
    - translators training etc.
- NLP for real world applications
  - real text: real problems?
    - translation and interpretation
    - writing in own and foreign language
    - finding information (IR)
    - talking on the phone
    - creating documentation etc.

Contrastive linguistics for NLP

- Obviously useful for many-language related activities
- But useful for monolingual studies as well
  - from a theoretical and methodological angle
  - and from a practical point of view
    - monolingual lexicography
    - semantics
    - information retrieval

NLP for (contrastive) linguistics

- Help select the interesting problems with a view of practical application
- Help debug linguistic hypotheses by providing running systems
- Help to uncover the "everything is linked" syndrome.
  - Assume X is solved, attack Y
  - Assume Y is solved, attack X

  } system requires X and Y working

The main difference is the angle

- The result for NLP: a system that does something / improves due to language understanding
- The result for L: improved understanding of what language does, *ergo* what we can do with language (creativity, power/persuasion, war, recreation, teaching (knowledge transmission), ...)
- Merging in the end, in that more and more of what we do is computer-mediated

Corpus contrastive linguistics or contrastive corpus linguistics

- Nobody drives a bay/chariot these days: cars took over
- Very soon, nobody is going to make a linguistic contrastive claim without corpus support: astrology vs. astronomy
- Good or bad? Depends on how well you use the tool - the corpus, how well you use the computer - the mediator

Parallel corpora

- You compile them because you are interested in the relationships:
  - what is it that remains the "same"?
  - what is it that changes? may want to identify patterns of change (in order to devise applications that help people change, or change themselves)

- narratives /myths
- several versions of a literary work
- original/censored
- translations
- news reuse
- wrong and corrected
- communicative intent
  - touristic brochures
  - scientific papers

What are the problems of using corpora

- Too much information
- Methodology is still in the craddle
- Not every corpus is suited for every claim
- Social inadaptation between arts scholars and the ubiquous computer science skills: programming languages, query languages
- Paradigm change, "buzz words" ... corpus studies are old - what is new is REUSE

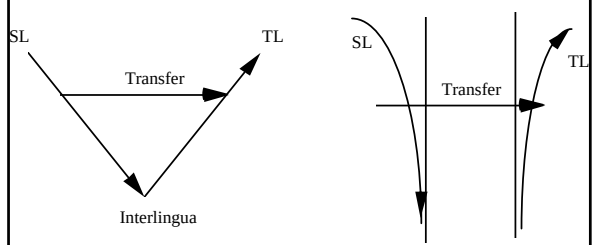
Requirements for corpus-based research

- Can phenomenon A be illuminated by using corpora?
- How can we evaluate?
- Based on what?
- Is the right procedure to compile when you want to investigate something?
  - Are there other corpora around where you can compare / partially check your findings?

Digression: mismatches between corpora and claims

- *web language on the whole is dramatically skewed toward dense, academic-like prose* (Ide et al., 2002)
- there is actually little variability in Norwegian (Rosén, 2001)
- punctuation studies in COMPARA ...
- ...

Are the assumptions correct?



Are the arguments based on hypotheses you can independently justify, subscribe to? (example of MT architecture from Santos 98)

Evaluate a corpus

- What is the information provided?
  - markup, annotation, extratextual information
- What is the fidelity to the original text?
  - better, what was changed/standardized/cut
- What were the selection criteria?
- Has it been validated? Evaluated? Quality-proofed?
- Version, date, problems reported?

Maintenance issues

See Santos & Gasperin (2002)

Kinds of data provided by a corpus

- Concordances
- Frequency
- Distribution
- Cross-correlation

- Bilingual concordances
- Translation frequency
- Bilingual distribution
- Cross-correlation
- Translation strategies
  - reordering
  - addition or deletion
  - translation notes
  - proper name handling

Kinds of data provided by a corpus

- Concordances in difference languages
- Frequency in different languages
- Distribution in different languages
- Cross-correlation between distributions

But interaction is the most important !!!

Across corpora

- A plea for making the same (minimum) kinds of information available
- The need to have data against which to weigh our own data
- The observed measurement is a property of ...?
  - P vs. E
  - S vs. T
  - SP vs. TE

(does it depend on the language pair?)

Across corpora 2

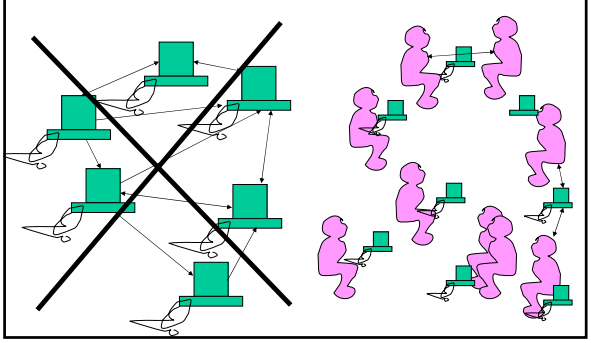
- Are the measures / data / studies you performed on ONE corpus valid across corpora?
- Are your results corpus-dependant?
- Can you replicate the studies?

See Santos & Oksefjell (1999) on this

Sources

- Quine's *Word and object*
  - indeterminacy of correlation. There is less basis of comparison -- less sense in saying what is good translation and what is bad -- the farther we get away from sentences with visibly direct conditioning to non-verbal stimuli and the farther we get off home ground (p.78)
- Keenan in Guenthner's *Meaning and Translation*
  - it would surely be surprising, and a very strong empirical claim, that different languages using different means to express 'meanings' always arrived at exactly the same end" (p.166).

This kind of information society



Communication is the bottleneck

- *jeg stoler på deg, lille venn...*
- different CVs for different occasions
- mailing lists and the cc: problematic
- the language of publication
- metaphors we program by
- knowledge extraction? evaluation/validation
- how to efficiently share work?

Never forget why you are doing what you are doing

- Clarity concerns
- Evaluation concerns
- Application concerns
- Take your stance about every one: there is nothing worse for science/linguistics if you uncritically accept authorities. ”Det er lov å" disagree with the most learned person

Basic standpoints

- All science has application (Kuhn)
- All research has an underlying question and a goal
- Languages are different systems
  - it's no use postulating they are the same
- Multilinguality is not more than bilinguality

Three kinds of corpus researchers

- Compilers
- Users
- Tool developers

- And also people concerned with evaluation of NLP systems

Take the users in consideration

- if you build a corpus
- if you write a paper
  - can the reader replicate the study
  - can the reader disconfirm / put into question the results
- if you perform a study
  - does it bring progress to the community
  - does it amass more data to a common pool
  - does it provide users with a more informed resource

References 1

- Ide, Nancy, Randi Reppen & Keith Suderman. "The American National Corpus: More Than the Web Can Provide". *Proceedings of the Third Language Resources and Evaluation Conference (LREC)*, Las Palmas, 2002, pp. 839-44.
- Keenan, Edward L. "Some Logical Problems in Translation". F. Guenthner & M. Guenthner-Reutter (eds.), *Meaning and Translation: Philosophical and Linguistic Approaches*, Duckworth, 1978, pp.157-89.
- Quine, W.O. *Word and Object*, The MIT Press, 1960
- Rosén, Victoria. "Er norsk et naturlig språk?", in Andersen, Øivin, Kjersti Fløttum & Torodd Kinn (eds.), *Menneske, språk og fellesskap. Festskrift til Kirsti Koch Christensen på 60-årsdagen 1. desember 2000*. Oslo: Novus, 2000, pp. 157-73.

References 2

- Santos, Diana. "Punctuation and multilinguality: Reflections from a language engineering perspective". *Working Papers in Applied Linguistics* 4/98, redigert av Jo Terje Ydstie og Anne C. Wollebæk. Oslo: Department of Linguistics, Faculty of Arts, University of Oslo, pp.138-60
- Santos, Diana & Caroline Gasperin. "Evaluation of parsed corpora: experiments in user-transparent and user-visible evaluation", *Proceedings of the Third Language Resources and Evaluation Conference (LREC)*, Las Palmas, 2002, pp. 597-604.
- Santos, Diana & Signe Oksefjell. "Using a Parallel Corpus to Validate Independent Claims", *Languages in contrast*, Vol. **2(1)**, 1999, pp.117-132. John Benjamins Publishing Co.