Corpus Linguistics and Methodologies for Human Annotation: Introduction

Karën Fort
karen.fort@inist.fr

November 18, 2011

People involved

People involved

You?

Tell us

- which corpus/corpora do you know?
- what are their characteristics?
- what are they used for?

What will you study here?

Corpus Linguistics

What will you study here?

Corpus Linguistics

Linguistics (in a nutshell)?

Linguistics (in a nutshell)?

Scientific study of human language
(Wikipedia, as of Nov. 18, 2010)

What will you study here?

Corpus Linguistics

What will you study here?

Corpus Linguistics

Corpus (more interesting here)?

Corpus (more interesting here)?

A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic [and/or extra-linguistic] criteria in order to be used as a sample of the language
[Sinclair, 1996]

More generally

- "The study of language based on examples of 'real life' language use" [McEnery and Wilson, 1996]
- Not only a methodology, but also a theory
- Not by itself a linguistic branch (unlike phonology, syntax or semantics), but transversal (in particular from the methodological point of view)

What will you study here?

Computer Corpus Linguistics [Leech, 1997], with particular emphasis on **corpus annotation**

and why should you study this?

- More and more corpora available
- More and more languages covered
- Bigger and bigger corpora available

$\Rightarrow$ More and more used in NLP

(Annotated) Corpora Usage in NLP

- Supervised learning - training and evaluation
- Unsupervised learning - evaluation
- Hand-crafted systems - learning and evaluation
- Analysis of text

Rules, thank you!

- Everybody arrives **on time**
- Everybody attends all the classes
- If you talk, talk to all of us
- No Facebook, no mail, no telephone
- Do not hesitate to ask questions: there are silly questions, but I'll answer them (once) too!
- Do not hesitate to tell me if I'm going too fast or too slow

And please, bring your laptop!

Content

This course will not only be about what's in the title. I hope it will also be about a more important thing:

What is being a researcher?

- Doubt and distance
- Ethics
- References, references, references

Course

- 20 hours of courses, in English
- Some practical activities, including:
  - a bit of oral corpus transcription
  - some corpus annotation
  - crowdsourcing, using AMT and PhraseDetective
  - inter-annotator agreement computation
  - presentations

More details

1. Corpus linguistics presentation
   - Course 1: Introduction and History
   - Course 2: Corpora Characteristics and Most Well-Known Corpora
2. Human annotation
   - Course 3: Practical Course, Transcribing with Transcriber (prereq.: install Transcriber and bring earphones)
   - Course 4: Practical Course, Annotating with GATE and Glozz (prereq.: install GATE and Glozz)
   - Course 5: Annotation: Introduction, Methodology, Formats
   - Course 6: Solutions for Annotation
   - Course 6/7: Practical Course, Crowdsourcing, using AMT and PhraseDetective
3. Evaluation
   - Course 7: Principles and Inter-annotator Agreement
   - Course 8: Practical course: Computing the Inter-annotator Agreement in an Annotation Campaign
4. Presentations by the students
   - Course 9/10: Presentations by the Students (20 min. each).

Practical courses: transcribing

transcribing, using *Transcriber*

Practical courses: annotating

- dealing with an annotation project
- annotating, using *GATE* and/or *Glozz*
- annotating using crowdsourcing
- evaluating

Practical courses: presenting

- Search an annotated (if available) corpus in YOUR language
- Gather information about its creation, availability, etc.
- Present it to us (20 minutes), discussion

$\Rightarrow$ Will be part of your final grade (1/3)

$\Rightarrow$ Questions will be asked about those during the exam

Practical courses: please!

- install Transcriber (bring earphones!): http://trans.sourceforge.net/
- install GATE: http://gate.ac.uk/
- install Glozz (I have logins for all of you): http://www.glozz.org/

Glozz logins

- kfort_etudiant_1 = 37099
- kfort_etudiant_2 = 93711
- kfort_etudiant_3 = 55394
- kfort_etudiant_4 = 77852
- kfort_etudiant_5 = 21240
- kfort_etudiant_6 = 59557
- kfort_etudiant_7 = 46526
- kfort_etudiant_8 = 8209
- kfort_etudiant_9 = 64821
- kfort_etudiant_10 = 35913

Presentations: please!

Handouts

Available after each course (I'll send you the URL, someone please send me your Emails).


Leech, G. (1997).
Corpus annotation: Linguistic information from computer text corpora, chapter Introducing corpus annotation, pages 1–18.
Longman, London.


McEnery, T. and Wilson, A. (1996).
Corpus linguistics.
Edinburgh University Press.


Sinclair, J. (1996).
Preliminary recommendations on corpus typology.
Technical report, Eagles.