Introduction to Human Language Technologies

Tomaž Erjavec

Karl-Franzens-Universität Graz

Lecture 2: Corpora
16.11.2007

Overview
1. what are corpora
2. historical perspective
3. how they are annotated

What is a corpus?
The Collins English Dictionary (1986):
1. a collection or body of writings, esp. by a single author or topic.

Guidelines of the Expert Advisory Group on Language Engineering Standards, EAGLES:

Corpus: A collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.

Computer corpus: a corpus which is encoded in a standardised and homogeneous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance.
Using corpora

- Applied linguistics:
  - Lexicography: mono-lingual dictionaries, terminological, bi-lingual
  - Language studies: hypothesis verification, knowledge discovery (lexis, morphology, syntax, ...)
  - Translation studies: a source translation equivalents and their contexts, translation memories, machine aided translations
  - Language learning: real-life examples, systematic teaching*, curriculum development
- Language technology:
  - Testing set for developed methods;
  - Training set for inductive learning
  - (Statistical Natural Language Processing)

Characteristics of a corpus

- Quantity:
  the bigger, the better
- Quality:
  the texts are authentic; the mark-up is validated
- Simplicity:
  the computer representation is understandable, with the markup easily separated from the text
- Documentation:
  the corpus contains bibliographic and other meta-data

Typology of corpora

- Corpora of written language, spoken and speech corpora (authenticity/price)
  e.g. the agency ELRA catalog
- Reference corpora (representative) and sub-language corpora (specialised)
  e.g. BNC, ICE, COLT
- Corpora with integral texts or of text samples (historical and legal reasons)
  e.g. Brown
- Static and monitor corpora (language change)
- Monolingual and multilingual parallel and comparable corpora
  e.g. Hansard, Europarl
- Plain text and annotated corpora
The history of computer corpora:

• First milestone: Brown (1 million words) 1964; LOB (also 1M) 1974
• Cobuild Bank of English (monitor, 100..200..M) 1980
• The spread of reference corpora: BNC (100M) 1995; Czech CNC (100M) 1998; Slovene: FIDA (100M), Nova Beseda (100M...) 1998; Croatian HNK (100M) 1999;
• EU corpus oriented projects in the ’90: NERC, MULTEXT-East...
• Language resources brokers: LDC 1992, ELRA 1995
• Web as Corpus (2002...): Sharoff’s corpora, Sketch Engine

Literature on corpora

• Uvod v korpusno jezikoslovje, Vojko Gorjanc. Domžale: Izolit, 2005
• LREC conferences: Fifth international conference on Language Resources and Evaluation. LREC2010

Steps in the preparation of a corpus

• Choosing the component texts:
  - linguistic and non-linguistic criteria; availability; simplicity; size
• Copyright
  - sensitivity of source (financial and privacy considerations); agreement with provider; usage, publication
• Acquiring digital originals
  - Web transfer; visit, OCR
• Up-translation
  - conversion to standard format; consistency; character set encodings
• Linguistic annotation
  - language dependent methods; errors
• Documentation
  - TEI header; Open Archives etc.
• Use / Download
  - (Web-based) concordancers for linguists
  - download needed for HLT use
  - licences for use
What annotation can be added to the text of the corpus?

- Annotation = interpretation
- Documentation about the corpus
- Document structure
- Basic linguistic markup: sentences, words, punctuation, abbreviations
- Lemmas and morphosyntactic descriptions
- Syntax
- Alignment
- Terms, semantics, anaphora, pragmatics, intonation, ...

Markup Methods

- **hand annotation**: documentation, first steps
  - generic editors or specialised editors
- **semi-automatic**: morphosyntactic and other linguistic
  annotation
  - cyclic approach: machine, hand, validate, correct, machine, ...
- **machine, with hand-written rules**: tokenisation
  - regular expression
- **machine, with inductively built models from annotated data**: "supervised learning"; HMMs, machine learning
- **machine, with inductively built models from unannotated data**: "unsupervised learning"; clustering techniques
- **overview of the field**

Computer coding of corpora

- Many corpora encoded in simple tabular format
- A good encoding must ensure durability, enable interchange between computer platforms and applications
- The basic standard used is *Extended Markup Language, XML*
- There are a number of companion standards and technologies: XML transformations (XSLT), data definition (DTD, XML Schema, ISO Relax NG), addressing and queries (XPath, XQuery), ...
- The vocabulary of annotations for corpora and other language resources are defined by the *Text Encoding Initiative, TEI*
Examples of use

• Concordances
• Collocations
  "You shall know a word by the company it keeps." (Firth, 1957)
• Induction of multilingual lexica
• Automatic translation

The future of corpus and data-driven linguistics

• Size:
  − Larger quantities of readily accessible data (Web as corpus)
  − Larger storage and processing power (Moore law)
• Complexity:
  − Deeper analysis:
    − syntax, deixis, semantic roles, dialogue acts, ...
    − Multimodal corpora:
      − speech, film, transcriptions, ...
  − Annotation levels and linking:
    − co-existence and linking of varied types of annotations; ambiguity
  − Development of tools and platforms:
    − precision, robustness, unsupervised learning, meta-learning