Annotating language data

Tomaž Erjavec

Institut für Informationsverarbeitung
Geisteswissenschaftliche Fakultät
Karl-Franzens-Universität Graz

Lecture 1: Corpora
3.11.2006

Overview

1. a few words about me
2. a few words about you
3. introduction to corpora and annotation

Lab work:
exploring publicly accessible corpora

Lecturer

- Tomaž Erjavec
  Department of Knowledge Technologies
  Jožef Stefan Institute
  Ljubljana
- http://nl.ijs.si/et/
- tomaz.erjavec@ijs.si
- Work: corpora and other language resources, standards, annotation, text-critical editions
- Web page for this course:
  http://nl.ijs.si/et/teach/graz06/annotation/
- Big debt to Sabine Schulte im Walde, Heike Zinsmeister:
  Introduction to Corpus Resources, Annotation and Access, Foundational Course at ESSLLI 2006 18th European Summer School in Logic, Language and Information

Students

- background: field of study, exposure to corpus linguistics?
- emails?

Overview of the course

1. Introduction to corpora and encoding
2. Morphosyntactic annotation
3. Syntactic annotation
4. Word-sense annotation
5. Multilingual Alignment

Overview of this lecture

1. empirical linguistics
2. computer corpora
3. corpus annotation
4. corpus sustainability

Two approaches to linguistics

Linguistics: characterisation and explanation of linguistic observations.

Competing approaches:

- rationalism vs. empiricism
- competence vs. performance
- Deductive method: from the general to specific; rules are derived from axioms and principles; verification of rules by observations
- Inductive method: from the specific to the general; rules are derived from specific observations; falsification of rules by observations

Empirical approach

- Describing naturally occurring language data
- Objective (reproducible) statements about language
- Quantitative analysis: common patterns in language use
- Creation of robust tools for Natural Language Processing (NLP) by applying statistical and machine learning approaches to large amounts of language data.
- Empirical turn supported by rise in processing speed of computers and their amount of storage, and the revolution in the availability of machine-readable texts (the word-wide web)

Historical perspective

- (Computational) linguistic paradigms:
- 1950 -- 1960: empiricism
  weak computers: frequency lists
- 1970 -- 1980: cognitive modelling (generative approaches, artificial intelligence)
  deep analysis / "basic science": computational linguistics
- 1990 -- ...: empiricist revival, also combined approaches
  quantity / usefulness: language technologies
- 2000 -- ...: The Web

Empirical resources

- Corpora: large amounts of text
- Dictionaries and thesauri, e.g. Longmans Dictionary of English (LDOCE), Roget's thesaurus
- Morphological databases and analysers
- Semantic hierarchies, e.g. WordNet
- Annotation tools, e.g. TreeTagger, MALT parser
- Processing tools, e.g. TIGERSearch, GIZA++

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.*

So: an arbitrary collection of documents is not a corpus!

Attested language

- Naturally occurring language
- Spoken language: performance errors such as slips of the tongue, hesitations, corrections
- Written language: errors such as misspellings, misediting, missing/additional words
- Creativity of language
- Context-dependency of language, e.g. ellipsis

Sample of a language

- Corpora give only a partial description of a language
- they are incomplete
  - older corpora don't include vocabulary related to the word wide web or e-mail
- they are biased
  - due to availability reasons they can contain disproportionate amount of a specific text type
- they include ungrammatical sentences
  - typos, copy-paste errors, conversion errors

→ Sample a corpus according to design criteria such that it is balanced and representative for a specific purpose

Specific purpose: Example

- Task: developing a machine translation system for dialogues on meeting arrangements
- Creation of a corpus to assist this task (as training and testing data)
- Sampling frame:
  - telephone-based dialogues on meeting arrangements
  - different types of meetings
  - different speakers (varying features such as gender, acquaintance, nationality, etc.)

→ Verbmobil corpus

Purpose: Reference corpus

- Task: create a representative corpus of British English
- Sampling frame:
  - 100 million words
  - 90% written language
    - time of creation: 1960-1993
    - medium: book, newspaper, un-published material
    - theme: informative, imaginative,…
    - language level
    - information on the author and on the "audience"
    - samples of < 40,000 words per text
  - 10% spoken language
    - topic: educational, business, institutional, leisure,…
    - demographic parameter: age, social group, gender, region, type of interaction (monologue/dialogue…)

→ British National Corpus (BNC)

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
    "idiomatic 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
- *Documented*:
  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 catalogue
- *Reference* corpora (general) and *sub-language corpora* (specialised, terminological)
  e.g. BNC, ICE, COLT
- Corpora with *integral* texts or of text *samples* (historical and legal reasons, but also a question of balance)
  e.g. Brown
- *Static* and *monitor* corpora (language change)
- *Monolingual* and multilingual *parallel* and *comparable* corpora
  e.g. Hansard, Europarl, JRC-ACQUIS
- *Plain text* and *annotated* corpora

The history of computer corpora:

- First milestones: Brown (1 million words) 1964; LOB (also 1M) 1974
- The spread of reference corpora: Cobuild Bank of English (monitor, 100..200..M) 1980; 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
- the Web as a corpus (Baroni: BootCat)

Annotation

- The practice of adding interpretative, linguistic information to an electronic corpus of spoken or written language
- The end-product of this process: the linguistic symbols which are attached to, linked with, interspersed with the electronic representation of the language itself.
- Question of granularity: how much detail should be encoded through annotation?

Annotation: motivation

- Extraction linguistic information
  - language is ambiguous → disambiguation by annotation
    e.g. "my left hand" (JJ), "on your left" (NN), "I left early" (VBD).
  - way to study more complex grammatical phenomena
    e.g. a direct object modified by a non-adjacent relative clause
    "I met friends in Rome, who were there for the first time."
- Re-usability
  - annotation is time-consuming and expensive.
  - automatic annotation often requires contextual (annotation) information
  - higher-level annotation often relies on lower-level annotation
- Multi-functionality
  - same corpus can be used of various applications, e.g. lexicography, parser (syntactic analyzers) training, speech synthesis, machine-aided translation, information retrieval.

What annotation can be added to the text of the corpus?

- Documentation about the corpus (example)
- Document structure (example)
- Basic linguistic markup: sentences, words (example), punctuation, abbreviations (example)
- Lemmas and morphosyntactic descriptions (example)
- Syntactic categories or dependencies: treebanks (example)
- Senses and semantic roles
- Information structure: topic/focus
- Anaphora and co-reference
- Named entities
- Terms
- Alignment of sentences, words, phrases (example)
- Time, emotions, prosody,...

Annotation: principles of good practice

- The raw corpus (primary data) should be recoverable.
- Annotations should be extricable from the corpus, to be stored independently if there is a need.
- Easy access to documentation:
  - annotation scheme
  - how, where, by whom the annotation was applied
  - some account of the quality of annotation

Annotation: representation

- Column-based, vertical format: LOB corpus

```
A014010 AT    a
A014020 NN    move
A014030 TO    to
A014040 VB    stop
A014050 NPT   \0Mr
A014060 NP    Gaitskell
```

- Horizontal format: LOB corpus

```
A014 ^ a_AT move_NN to_TO stop_VB
\0Mr_NPT Gaitskell_NP
```

- These formats are problematic if the annotation becomes more complex

Annotation: representation

- Hierarchical structure annotation:
  Penn Treebank bracketing format

```
( (S
    (NP-SBJ (DT This) )
    (VP (VBZ means)
        (SBAR  (-NONE- 0)
            (S
                (NP-SBJ (DT the) (NNS returns) )
                (VP (MC can)
                    (VP (VB vary)
                        (NP (DT a) (JJ great)  (NN deal) ))))))
     ( . .)  ))
```

tags: phrasal category, function, part-of-speech

Annotation: representation

XML inline representation (TüBa corpus):

```
<sentence editor="korder" date="1998080418:46:20"
   origin="cd15e1.export"
   comment="%%&lt;g001acn1_015_AAJ_150000_E&gt;">
   <node cat="S" func="--" parent="0">
     <node cat="np" func="SBJ">
       <word form="I" pos="PP" func="HD/>
     </node>
     <word form="have" pos="VBP" func="HD"/>
     <node cat="VP" func="COMP">
       <word form="to" pos="TO" func="HD"/>
       <node cat="VP" func="COMP">
         <word form="go" pos="VB" func="HD"/>
          …
</sentence>
```

Annotation: representation

XML stand-off representation (TüBa corpus):

```
<terminals>
  <t id="s42_1" word="I" pos="PP"/>
  <t id="s42_2" word="have" pos="VBP"/>
  <t id="s42_3" word="to" pos="TO"/>
  <t id="s42_4" word="go" pos="VB"/>
  <t id="s42_5" word="to" pos="IN"/>
  <t id="s42_6" word="Berlin" pos="NP"/> …
</terminals>
<nonterminals>
 <nt id="s42_500" cat="NP">
    <edge label="HD" idref="s42_1"/>
  </nt>
 <nt id="s42_508" cat="S">
   <edge label="SBJ" idref="s42_500"/>
   <edge label="HD" idref="s42_2"/>
   <edge label="COMP" idref="s42_507"/>
  </nt> …
</nonterminals>
```

Annotation scheme

A detailed specification of the annotation

- A list of symbols used in the annotation, such as terminals (e.g. parts-of-speech), non-terminals (e.g. syntactic category labels), and other symbols
- A basic definition of the symbols, e.g. “JJ = adjective”
- A detailed description of how the symbols are applied to text, e.g. how do the annotators know what is a noun phrase and what isn’t?

Annotation scheme types

- Comprehensive grammar
  - difficult to write and update
  - difficult for annotators to use
- Set of guidelines
  - evolving laws of precedence
  - written as annotators’ manual
- Reference annotated corpus (benchmark corpus)
- Mixed form
  - cross-referenced guidelines and examples

Annotation: principles of good practice

Additional maxims:

- Annotation schemes made available to research community (on caveat emptor principle)
- Annotation should depend on consensual or theory neutral analyses.
- No annotation scheme should claim authority as an absolute standard

Exploitation of annotated corpora in NLP

- corpora give access to quantitative data
- disambiguation is a key problem in many areas of NLP: corpora provide disambiguated data
  - TAGGIT based on hand crafted rules and used for tagging the Brown corpus → accuracy ~77%
  - A subset of the Brown corpus was used to compute the probabilities of tag uni- and bi-grams
  - CLAWS used these probabilities for choosing the correct tag given the local context. Used for tagging the LOB Corpus → accuracy ~97%

Exploitation of annotated corpora in NLP

- tagger, lemmatiser, parser induction
- terminology extraction
- bi-lingual lexicon extraction
- induction of machine-translation models
- word-sense learning
- test-bed for evaluation of NLP systems

Annotation Methods

- *hand annotation*: documentation, first steps
  generic (XML, spreadsheet) editors or specialised editors
- *machine, with hand-written rules*: tokenisation
  regular expression
- *machine, with inductively built models from annotated data*:
  "supervised learning"; Hams, decision trees, inductive logic programming,...
- *semi-automatic*: morphosyntactic and other linguistic annotation
  cyclic approach: machine, hand, validate, correct, machine, ...
- *machine, with inductively built models from un-annotated data*:
  "unsupervised leaning"; clustering techniques
- overview of the field

Metadata

A corpus contains different kinds of data:

- Primary data: the text
- Annotation: linguistic interpretation of the primary data
- Metadata: contextual information about the primary data and annotations
  - documentation for subsequent users
  - key to retrieve particular types of primary data
- Meta-metadata: contextual information about the metadata: who created the metadata and when

The TEI header

- The TEI header gives the meta-data on the TEI document and consists of four main parts (only first is obligatory):
- <fileDesc>
  - *file description*, containing a full bibliographical description of the computer file itself; it includes information about the source or sources (<sourceDesc>) from which the electronic text was derived.
- <encodingDesc>
  - *encoding description*, which describes the relationship between an electronic text and its source or sources: it allows for detailed description of whether (or how) the text was normalized during transcription, how the encoder resolved ambiguities in the source, what levels of encoding or analysis were applied, etc.
- <profileDesc>
  - *text profile*, containing classificatory and contextual information about the text, e.g. its subject matter, the individuals described by or participating in producing it, etc. It is of particular use in structured composite texts such as corpora, where it is often desirable to enforce a controlled descriptive vocabulary or to perform retrievals from a body of text in terms of text type or origin.
- <revisionDesc>
  - *revision history*, which allows the encoder to provide a history of changes made during the development of the electronic text. It is important for version control and for resolving questions about the history of a file.

TEI header example

```
<?xml version="1.0" encoding="utf-8" ?>
- <teiHeader id="svez.teiHeader" type="text" creator="et" status="update" lang="en" date.created="2004-04-26"
  date.updated="2004-10-01">
- <fileDesc>
  - <titleStmt>
      <title lang="en" id="svez.title">SVEZ-IJS English-Slovene ACQUIS Corpus</title>
      <!- title lang="sl" id="svez.title">Angleško-slovenski korpus SVEZ-IJS ACQUIS</title ->
    - <respStmt>
        <name>Jasna Belc, SVEZ</name>
        <resp>Provision of the translation memory.</resp>
      </respStmt>
    - <respStmt>
        <name>Tomaž Erjavec, IJS</name>
        <resp>TEI encoding and linguistic annotation.</resp>
      </respStmt>
    - <principal>
      - <name>
          <xref url="http://nl.ijs.si/et/">Tomaž Erjavec</xref>
        </name>
      - <address>
          <addrLine>Dept. of Knowledge Technologies</addrLine>
          <addrLine>Jožef Stefan Institute</addrLine>
          <addrLine>Jamova 39</addrLine>
          <addrLine>SI-1000 Ljubljana</addrLine>
          <addrLine>Slovenia</addrLine>
        </address>
      </principal>
  </titleStmt>
- <editionStmt>
    <edition>Version 1.0</edition>
  </editionStmt>
  <extent>10 million words</extent>
- <publicationStmt>
  - <address>
    - <addrLine>
        <xref url="http://nl.ijs.si/svez/">http://nl.ijs.si/svez/</xref>
      </addrLine>
    </address>
  - <availability status="restricted">
      <p>This corpus is made available under the condition that it shall be used for non-commercial
      purposes only and that the creators of the corpus - the Office of the Government of the Republic
      of Slovenia for European Affairs and the Jožef Stefan Institute - shall be acknowledged in any
      work making use of the corpus.</p>
    </availability>
  </publicationStmt>
```

Sustainability

- New developments in computer technology allow to capture, store, annotate and disseminate digital data.
- Uncritical adoption of new technologies compromises ability to preserve data.
- Desired: portability of digital language resources across environments, scholarly communities, domains of applications and passage of time

Sustainability: problem areas

- Content: information content of the resource
  - Coverage: unbalanced, low recording quality
  - Terminology: ambiguous / unknown terms
- Format: electronic representation
  - Openness: proprietary format often restricted to specific hardware or operating system
  - Encoding: idiosyncratic representation of characters
  - Markup: idiosyncratic, unstable representation of character string:

    chien **n** dog.

    `chien: [n] dog.`

Sustainability: problem areas

- Discovery: the problem of finding existing resources and knowing whether they are relevant.
  - Documents are “hidden” in linguists personal collection of computer files
  - No publicly available description
- Access: unclear scope and process.
- Citation: Broken URLs, confusion of different versions of the same resources.
- Preservation
  - Formats become obsolete
  - Absence of supporting hardware (e.g. 5.25” floppy disks)
  - Lifespan of physical medium (digital media: 5 years)
- Rights: what a potential user is permitted to do with the resource.

Sustainability: principles of best practice

- Content
  - Documents methods, provide original resources (e.g. recordings)
  - Map terminology and abbreviations to a common ontology of linguistic terms
- Format
  - use open formats, free tools, published proprietary formats
  - use Unicode for encoding
  - use XML for markup
- Discovery
  - list resource in e.g. OLAC repository
  - include metadata and keywords for search engines

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 providers; 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 use in HLT or "deep" linguistics
  - licences for use

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

Literature

- This lecture a subset of :
  Sabine Schulte im Walde, Heike Zinsmeister: Introduction to Corpus Resources, Annotation and Access. Foundational Course at ESSLLI 2006 18th European Summer School in Logic, Language and Information
  (the course material also contains literature and further pointers)
- *Books:*
  - *Corpus Linguistics* by Tony McEnery and Andrew Wilson. Edinburgh: Edinburgh University Press, 1996
  - *An Introduction to Corpus Linguistics* by Graeme D. Kennedy. Studies in Language and Linguistics, London, 1998
  - *Corpus Linguistics: Investigating Language Structure and Use* by Douglas Biber, Susan Conrad, Randi Reppen. Cambridge University Press, 1998
- Conference proceedings:
  - LREC conferences:
    Fifth international conference on Language Resources and Evaluation, LREC'06
  - Slovenian Conferences on LANGUAGE TECHNOLOGIES 2006, 2004,2002, 2000, 1998

Practical work

Find one or two corpora on the web and discuss where they belong according to the presented corpus typology and characteristics, how they are annotated, what the format of the annotation is, what kind of meta-data is available for them what kind of access is provided and what have/could they be useful for.