Overview of the lecture

1. Background
2. Corpus compilation and markup
3. Morphosyntactic tagging

Background

- What is a corpus?
- Using corpora
- Characteristics of a corpus
- Typology of corpora
- History
- Slovene language corpora
A corpus is:

- a large collection of texts
- in digital format
- language “as it is”
- a sample of the language it is meant to represent
- used for describing language (descriptive/empirical linguistics)

A more precise definition

- **Corpus** (plural **corpora**) is Latin for body
- 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.
- For computer scientists: a dataset

Using corpora

- **Applied linguistics**:
  - Lexicography: making dictionaries (first users of corpora)
  - Translation studies: translation equivalents with contexts, translation memories, machine aided translations
  - Language learning: real-life examples, curriculum development
- **Corpus linguistics**:
  - Linguistics based not on introspection, but on observation of real data
- **Language technology**:
  - Testing set for developed methods;
  - Training set for inductive learning (statistical natural language processing)
Characteristics of a (good) 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 metadata

Typology of corpora I.

- **Medium:**
  - written language
  - spoken language (spoken, but in writing / transcription)
  - speech corpora (actual speech signal)
- **Content:**
  - reference corpora (representative), e.g. BNC
  - sub-language corpora (specialised), e.g. COLT
- **Structure:**
  - corpora with integral texts
  - corpora or of text samples (historical and legal reasons) e.g. Brown

Typology of corpora II

- **Time:**
  - static corpora
  - monitor corpora (language change)
- **Languages:**
  - monolingual corpora
  - multilingual parallel corpora (e.g. Hansard, Europarl, JRC Acquis)
  - multilingual comparable corpora
- **Annotation:**
  - plain text corpora
  - annotated corpora
Reference corpora

- Characteristics:
  - a sample of the "complete" language
  - large, expensive, detailed and explicit design criteria
  - typically of contemporary language
  - documented and annotated
  - legally clean, available (but usu. only via a concordancer)

- Criteria for including texts:
  - representativeness:
    - corpus includes "all" text types
  - balance:
    - the sizes of text type samples are in proportion to their "importance" for the speakers of the language

- methodology v.s. practical constraints

History of 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; Croatian HNB (100M) 1999...
  - Slovene reference corpora: FIDA (100M), Nova Beseda (100M) 1998; FIDAT (600M) 2006; gigaFIDA (2011?)

- EU corpus oriented projects in the '90: NERC, MULTEXT-East...

- Language resources brokers: LDC 1992, ELRA 1995

- Web as Corpus (2000..): ukWaC, itWaC, ... slWaC

- more, larger, for more languages, with diverse annotations: EUROPARL, PDT, ...

Slovene language corpora

- Monolingual reference corpora:
  - IRC SAZU: Beseda, 1998; Nova Beseda, 2000
  - US, FF: FIDA corpora

- Parallel corpora:
  - NL: Eurokorpis
  - FF: TRANST, 2002

- Speech corpora:
  - Laboratory for Digital Signal Processing, University of Maribor: speechDat, ONOMASTICA...
  - Laboratory of Artificial Perception, Systems and Cybernetics, University of Ljubljana: SQEL, COPOLIS...
II. Compilation and markup of corpora

- Steps in the preparation of a corpus
- What annotation can be added to the text
- Computer coding of corpora
- Markup Methods

Before making your own corpus

check if an appropriate corpus is already available
- google
- corpora@lists.uib.no
- LDC, ELRA

Steps in the preparation of a corpus

1. Choosing the component texts and acquiring digital originals
2. Up-translation to standard format
3. Linguistic annotation
4. Documentation
5. Use and Dissemination
Getting the text

1. Choosing the component texts:
   linguistic and non-linguistic criteria; availability; simplicity; size
2. Copyright
   sensitivity of source (financial and privacy considerations); agreement with providers; usage, publication
3. Acquiring digital originals
   OCR; digital originals; Web
   - BootCat

Processing

1. Conversion to common format
   consistency; character set encodings; structure
   - Web as Corpus: Wacky tools
2. Documentation
   e.g. TEI header; Open Archives etc.
3. Linguistic annotation
   language dependent methods; errors

Use and dissemination

- Using the corpus:
  - concordancer (linguists)
    e.g. FidaPLUS, SKE, iKorpus, JOS, IMP
  - statistics extraction
  - development of new methods for analysis
- Dissemination:
  - legalities (source copyright, corpus use agreement)
  - mode: concordancer or dataset
Computer coding of corpora

- Encoding must ensure
  - durability
  - interchange between computer platforms
  - interchange between applications
- Basic standard: XML
  - companion standards: W3C Schema, ISO Relax NG, XSLT, XPath, XQuery,
- XML vocabulary of annotations of arbitrary texts:
  Text Encoding Initiative, TEI
- ISO TC 37 „Terminology and other language resources”: many standards for text encoding

Corpus annotation

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

Example: TEI header

```xml
<teiHeader id="ecmr.H" type="text" lang="sl-en" creator="ET" status="update"
date.created="1999-04-13" date.updated="1999-06-22">
<titleStmt>
<title lang="sl">Ekonomsko ogledalo; 13 & scarono;tevilk 98/99</title>
<title lang="en">Slovenian Economic Mirror; 13 issues, 98/99</title>
<respStmt>
<name>Andrej Skubic, FF</name>
<resp lang="sl">Zagotovitev digitalnega originala, poravnava</resp>
<resp lang="en">Provision of digital original, alignment</resp>
<name>Tomaž Erjavec, IJS</name>
<resp lang="sl">Tokenizacija, pretvorba v TEI</resp>
<resp lang="en">Tokenisation, conversion to TEI</resp>
</respStmt>
</teiHeader> ...
Example: text structure

<quote id="Osl.1.8.18" rend="center;it">
  <lj id="Osl.1.8.18.1">Tam pod kostanjevim drevesom</lj>
  <lj id="Osl.1.8.18.2">izdala si me</lj>
  <lj id="Osl.1.8.18.3">izdal sem te</lj>
  <lj id="Osl.1.8.18.4">ne da bi trenila z očesom</lj>
</quote>

<p id="Osl.1.8.19">
  <s id="Osl.1.8.19.1">Trije možje se niti ganili</s>...
  <s id="Osl.1.8.19.2">Toda ko je <name>Winston</name> znova pogledal v Rutherfordov propadli obraz, je opazil, da so njegove oči polne solz.</s>...
</p>

Example: morphosyntactic tagging

<w lemma="biti" ana="Vcps-sma">Bil</w>
<w lemma="biti" ana="Vcip3s--n">je</w>
<w lemma="jasen" ana="Afpmsnn">jasen</w><c>,</c>
<w lemma="mrzel" ana="Afpmsnn">mrzel</w>
<w lemma="aprilski" ana="Aopmsn">aprilski</w>
<w lemma="dan" ana="Fcmn">dan</w><c>
<w lemma="ura" ana="Ncfpn">ura</w><c>
<w lemma="biti" ana="Vcip3p--n">so</w>
<w lemma="biti" ana="Vmps--pfa">bile</w><c>
<w lemma="trinajst" ana="Mcnpnl">trinajst</w>. ...
Methods for linguistic markup

- Hand annotation: documentation, first steps
  - generic (XML, spreadsheet) 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, decision trees, inductive logic programming,...
  - machine, with inductively built models from un-annotated data:
    "unsupervised learning": clustering techniques

- Overview of the field

III. Morphosyntactic tagging

- Better known as part-of-speech (PoS) tagging
- Tagging is the task of labeling each word in a sequence of words with its appropriate part-of-speech
- Words are often ambiguous with respect to their POS:
  - saw → singular noun "I brought a saw"
  - saw → past tense of verb "I saw a tree"

- Purposes and applications (examples):
  - pre-processing step for further analyses:
    - lemmatisation
    - syntactic structure, etc.
  - text indexing, e.g. nouns are more useful than verbs
  - pronunciation in speech processing

Steps in tagging

- For each word token in text the tagger needs to know all its possible tags
  (ambiguity class)
    → a morphological lexicon
- Given the context in which the word appears in, the tagger must decide in the correct tag:
  - he saw/V a man carrying a saw/N
- So, tagging performs limited syntactic disambiguation
Example: Penn Treebank

Under/IN the/DT proposal/NN , Delmed/NNP would/MD issue/VB about/IN 123.5/CD million/CD additional/IJ Delmed/NNP common/IJ shares/NNS to/TO Fresenius/NNP at/IN an/DT average/IJ price/NN of/IN about/IN 65/CD cents/NNS a/DT share/NN , though/IN under/IN no/DT circumstances/NNS more/IIR than/IN 75/CD cents/NNS a/DT share/NN .

PoS taggers

- Most taggers induce the language model from a hand-annotated corpus
- Typically, two resources are induced:
  - lexicon, giving the ambiguity class of a word and their frequencies in the training corpus
  - tag n-grams

Tagging with Markov Models

- Sequence of tags in a text is regarded a Markov chain
- Limited horizon: A word’s tag only depends on the previous tag: $p(x_{i+1} = t_j | x_1, \ldots, x_i) = p (x_{i+1} = t_j | x_i)$
- Time invariant: This dependency does not change over time: $p(x_{i+1} = t_j | x_i) = p(x_2 = t_j | x_1)$
- Task: Find the most probable tag sequence for a sequence of words
- Maximum likelihood estimate of tag $t'$ following $t$: $p(t' | t) = f(t, t') / f(t)$
- Optimal tags for a sentence: $t_{i}^{*} = \arg \max p(t_{i}^{*} | w_{i-1}) = \Pi p(w_i | t) p(t_i | t_{i-1})$
Most popular Markov model tagger

- TnT (Trigrams ’n Tags)
- induces lexicon and tag trigrams from the training corpus
- has heuristics to tag unknown words
- has no problem with large tagsets
- fast in training and tagging
- freely available for non-commercial use
- but only as a Linux executable
- OS alternative: hunpos

Yet another Tagger

For a while, trying out new approaches to tagging was in fashion
- Maximum Entropy taggers
- Support Vector Machine taggers
- Memory based taggers
- ...

Tagsets

- A tagset is a set of part-of-speech tags
- Classical 8 classes (Thrax, 100 BC): noun, verb, article, participle, pronoun, preposition, adverb, conjunction
- But all tagset use more tags than that!
- Criteria:
  - specifiability: degree to which humans use the tagset uniformly on the same text
  - accuracy: evaluation of output on tagged text
  - suitability for intended application
Tagsets for English

- For English, there exist several tagsets: Brown, CLAWS, Penn, ...
- English tagsets include PoS + some other morphological (inflectional) properties: 30-80 tags
- Penn Treebank Tagset for English: 37 tags, e.g.
  - JJ adjective, positive
  - JJR adjective, comparative
  - JJS adjective, superlative
  - NN non-plural common noun
  - NNS plural common noun
  - NNP non-plural proper name
  - NNPS plural proper name
  - IN preposition
  - ...

Morphosyntactic tagsets

- For inflectionally rich languages (such as Slavic languages), tagsets contain much more information than just PoS
- Slovene, Czech, etc. > 1,000 different morphosyntactic tags
  - gender, number, case, animacy, definiteness, ...
- Efforts to standardise tagsets across languages:
  - Eagles
  - MULTEXT
  - MULTEXT-East

MULTEXT-East

- EU project in '90s: development of language resources for Central and East-European languages
- Several later releases, V4 in 2010 (17 languages)
- Development of morphosyntactic specifications, lexica and annotated corpus
- Parallel annotated corpus: Orwell's 1984
- Web site: http://nl.ijs.si/ME/
MULTEXT-East morphosyntactic specifications

- Specify
  - what morphosyntactic features particular languages distinguish,
  - what their names and values are,
  - how they can be mapped to tags (morphosyntactic descriptions, MSDs)
- e.g. that Ncmts is:
  - a valid for Slovene
  - is equivalent to PoS:Noun, Type:common, Gender:masculating, Number:singular

JOS project

- JOS language resources are meant to facilitate developments of human language technologies and corpus linguistics for the Slovene language
- Morphosyntactic specifications
- Two annotated corpora (morphosyntactic descriptions and lemmas)
  - jos100k (hand validated)
  - jos1M (partially hand validated)
- Sampled from FidaPLUS corpus
- jos100k: syntactic and semantic levels of linguistic description
- Two web services
  - concordancer
  - text annotation tool
- Encoded in TEI P5
- Freely available (CC): http://nl.ijs.si/jos/

2.2. GLAGOL

Preposlovna zabilježka navadne izmed JOS

2.2. GLAGOL

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Jos100k encoding

Processing Historical Language

- Interesting for diachronic linguistics and better access to digital libraries
- Problems:
  - Difficult to obtain good transcriptions
  - Great variation in spelling
  - No resources for tool training

Historical slv:
- Late standardisation (XIX ≠ XX)
- Before 1850: š → š, šh → šh, šh → šh, z → ž, ž
- No corpora/lexica of historical Slovene

Background

- AHLib (2004–08) / Deutsch-slowenische/kroatische Übersetzung 1848–1918
  - Scans + correction + (lemmatisation) of ger−slv books
  - AAS & Karl-Franzens University, Graz (prof. Erich Prunč)
  - JSI: correction & lemmatisation environment

- EU IP IMPACT (2010–2011)
  - Better OCR for historical texts
  - NUK, GTD transcriptions
  - JSI: (semi)manual lexicon construction

- Google award (2011)
  - Developing language models for historical Slovene
    - ZGC SAGU: transcriptions of old texts
    - JSI: annotating a corpus of XXth century Slovene

Jos100k encoding
Producing the IMP corpus

- Representative & balanced, sampled
- Corpus element: unbroken & contiguous text from 1 page
- Sampled by decade & text
- Target size: 1,000 pages (~200,000 words)
- Encoded in TEI P5
- Automatically annotated
- Tool for manual annotation: IMPACT INL Cobalt
- Annotator training & management: May
- Manual correction: June–November

Annotation tool

Approach:
- Modernise, then process as contemporary language
- Language independent (trainable) modules

Steps:
1. Tokenisation (mlToken)
2. Transcription (Vaam)
3. Tagging (Trt)
4. Lemmatisation (CLOG)

= ToTrTaLe
- Pipeline in Perl
- TEI P5 I/O

Conclusions

- What is a corpus
- How to make it
- How to annotate it
- Case studies: MULTEXT-East, JOS, IMP