Lecture I.
Introduction to Language Technologies

Tomaž Erjavec
Technicalities

- Lecturer: http://nl.ijs.si/et/tomaz.erjavec@ijs.si
- Work: language resources for Slovene, linguistic annotation, standards, digital libraries
- Course homepage: http://nl.ijs.si/et/teach/mps10-hlt/
- Assessment: seminar work ½ quality of work, ½ quality of report
- Next lecture: May 12th
  - Presentation on topics we are working on at JSI
  - Possible seminar topics
Overview of the lecture

1. Computer processing of natural language
2. Some history
3. Applications
4. Levels of linguistic analysis
I. Computer processing of natural language

• Computational Linguistics:
  • a branch of computer science, that attempts to model the cognitive faculty of humans that enables us to produce/understand language

• Natural Language Processing:
  • a subfield of CL, dealing with specific computational methods to process language

• Human Language Technologies:
  • (the development of) useful programs to process language
Languages and computers

How do computers “understand” language?

(written) language is, for a computer, merely a sequence of characters (*strings*)

1. Tokenisation – splitting of text into tokens (words):
   - words are separated by spaces
   - words are separated by spaces or punctuation
   - words are separated by spaces or punctuation and space
   - $[2,3H]dexamethasone, $4,000.00, pre- and postnatal, etc.
Languages have properties that humans find easy to process, but are very problematic for computers

- Ambiguity: many words, syntactic constructions, etc. have more than one interpretation
- Vagueness: many linguistic features are left implicit in the text
- Paraphrases: many concepts can be expressed in different ways

Humans use context and background knowledge; both are difficult for computers
• Time flies like an arrow.
• I saw the spy with the binoculars. He left the bank at 3 p.m.
The dimensions of the problem

Many applications require only a shallow level of analysis
Structuralist and empiricist views on language

- **The structuralist approach:**
  - Language is a limited and orderly system based on rules.
  - Automatic processing of language is possible with rules.
  - Rules are written in accordance with language intuition.

- **The empirical approach:**
  - Language is the sum total of all its manifestations.
  - Generalisations are possible only on the basis of large collections of language data, which serve as a sample of the language (*corpora*).
Other names for the two approaches

• Rationalism vs. empiricism
• Competence vs. performance
• Deductive vs. Inductive:
  • 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 by applying statistical and machine learning approaches to large amounts of language data
- Basis for empirical approach: corpora
- Empirical turn supported by rise in processing speed and storage, and the revolution in the availability of machine-readable texts (WWW)
II. The history of Computational Linguistics

- MT, empiricism (1950-70)
- Structuralism: generative linguistics (70-90)
- Data fights back (80-00)
- A happy marriage?
- The promise of the Web
The early years

- The promise (and need!) for machine translation
- The decade of optimism: 1954-1966
- *The spirit is willing but the flesh is weak ≠ The vodka is good but the meat is rotten*
- ALPAC report 1966: no further investment in MT research; instead development of machine aids for translators, such as automatic dictionaries, and the continued support of basic research in computational linguistics
- also quantitative language (text/author) investigations
The Generative Paradigm

Noam Chomsky’s Transformational grammar: *Syntactic Structures* (1957)

Two levels of representation of the structure of sentences:
- an underlying, more abstract form, termed 'deep structure',
- the actual form of the sentence produced, called 'surface structure'.

Deep structure is represented in the form of a hierarchical tree diagram, or "phrase structure tree," depicting the abstract grammatical relationships between the words and phrases within a sentence.

A system of formal rules specifies how deep structures are to be transformed into surface structures.
Phrase structure rules and derivation trees

S → NP V NP
NP → N
NP → Det N
NP → NP that S
Characteristics of generative grammar

- Research mostly in syntax, but also phonology, morphology and semantics (as well as language development, cognitive linguistics)
- Cognitive modelling and generative capacity; search for linguistic universals
- First strict formal specifications (at first), but problems of overpremissivness
Computational linguistics

• Focus in the 70’s is on cognitive simulation (with long term practical prospects..)
• The applied branch of CompLing is called *Natural Language Processing*
• Initially following Chomsky’s theory + developing efficient methods for parsing
• Early 80’s: unification based grammars (artificial intelligence, logic programming, constraint satisfaction, inheritance reasoning, object oriented programming,..)
Problems

Disadvantage of rule-based (deep-knowledge) systems:

- Coverage (lexicon)
- Robustness (ill-formed input)
- Speed (polynomial complexity)
- Preferences (the problem of ambiguity: “Time flies like an arrow”)
- Applicability? (more useful to know what is the name of a company than to know the deep parse of a sentence)
- EUROTRA and VERBMOBIL: success or disaster?
Back to data

- Late 1980’s: applied methods based on data (language resources)
- The increasing role of the lexicon
- (Re)emergence of corpora
- 90’s: Human language technologies
- Data-driven shallow (knowledge-poor) methods
- Inductive approaches, esp. statistical ones (PoS tagging, collocation identification)
- Importance of evaluation (resources, methods)
The new millennium

The emergence of the Web:
- Large and getting larger
- Multilinguality
- Simple to access, but hard to digest → Semantic Web

The promise of mobile, ‘invisible’ interfaces; HLT in the role of middle-ware
III. HLT applications

- Speech technologies
- Machine translation
- Question answering
- Information retrieval and extraction
- Text summarisation
- Text mining
- Dialogue systems
- Multimodal and multimedia systems

- Computer assisted:
  authoring; language learning; translating; lexicology; language research
More HLT applications

- **Corpus tools**
  - concordance software
  - tools for statistical analysis of corpora
  - tools for compiling corpora
  - tools for aligning corpora
  - tools for annotating corpora

- **Translation tools**
  - programs for terminology databases
  - translation memory programs
  - machine translation
Speech technologies

- speech synthesis
- speech recognition
- speaker verification
- spoken dialogue systems
- speech-to-speech translation
- speech prosody: emotional speech
- audio-visual speech (talking heads)
Perfect MT would require the problem of NL understanding to be solved first!

Types of MT:
- Fully automatic MT (Google translate, babel fish)
- Human-aided MT (pre and post-processing)
- Machine aided HT (translation memories)

Problem of evaluation:
- automatic (BLEU, METEOR)
- manual (expensive!)
Rule based MT

- Analysis and generation rules + lexicons
- Altavista: babel fish
- Problems: very expensive to develop, difficult to debug, gaps in knowledge
- Option for closely related languages
Statistical MT

- Parallel corpora: text in original language + translation
- Texts are first aligned by sentences
- On the basis of parallel corpora only: induce statistical model of translation
- Noisy channel model, introduced by researchers working at IBM: very influential approach
- Now used in Google translate
- Difficult getting enough parallel text
Information retrieval and extraction

- **Information retrieval (IR)** searching for documents, for information within documents and for metadata about documents.
  - “bag of words” approach
- **Information extraction (IE)** a type of IR whose goal is to automatically extract structured information, i.e. categorized and contextually and semantically well-defined data from a certain domain, from unstructured machine-readable documents.
- Related area: **Named Entity Recognition**
  - identify names, dates, numeric expression in text
Corpus linguistics

- Large collection of texts, uniformly encoded and chosen according to linguistic criteria = **corpus**
- Corpora can be (manually, automatically) annotated with linguistic information (e.g. PoS, lemma)
- Used as datasets for
  - linguistic investigations (lexicography!)
  - training or testing of programs
Concordances

Corpus: British National Corpus

Page 1 of 1498

A00 Immune Deficiency Syndrome is a condition caused by a virus called HIV. (Human Immuno Deficiency
A00 've always wanted? It's all in a good cause .6. SPONSORED SLIM 7. HOLD A COFFEE
A00 an average of longer than two years. <p> Cause of death <p> The 'cause of death' figures
A00 than two years. <p> Cause of death <p> The 'cause of death' figures are also changing beyond
A00 home care was less. <p> The commonest cause of death is now advanced Kaposi's Sarcoma
A00 great difficulties in the lung and the gut, causing shortness of breath and other problems
A00 IT'S YOUR CHOICE<p> Every day the virus causing AIDS is infecting more young people. A
A01 lorry-driver. He's infected with the virus causing AIDS, but does n't know. (He could have
A01 time they will all be ill. <p> When sex can cause disease Rapid spread <p> Until recently it
A01 that a sexually-transmitted infection might cause cancer of the cervix -- especially if you
A01 think you understand it all. But some drugs cause bad, disturbing flashbacks. 'I can't
A01 I do not pay UK Income Tax? <p> This can cause problems, since you agree under the terms
A01 I do not pay UK income tax? <p> This can cause problems because if you do not pay tax
A02 are the effects, worldwide, of the virus causing AIDS. During the past year at least 1.5
A02 Source: BMJ 1991); 302: 203-7) <p> The 'cause of death figures' are also changing beyond
A02 equipment loan Emotional support The commonest cause of death now is advanced Kaposi's Sarcoma
A02 produce difficulties in the lung and the gut, causing many problems including shortness of breath
A02 million expected to be infected with the virus causing AIDS. 36 million will be in the developing
IV. Levels of linguistic analysis

- Phonetics
- Phonology
- Morphology
- Syntax
- Semantics
- Discourse analysis
- Pragmatics
- + Lexicology
Phonetics

- Studies how sounds are produced; methods for description, classification, transcription
- Articulatory phonetics (how sounds are made)
- Acoustic phonetics (physical properties of speech sounds)
- Auditory phonetics (perceptual response to speech sounds)
Phonology

- Studies the sound systems of a language (of all the sounds humans can produce, only a small number are used distinctively in one language)
- The sounds are organised in a system of contrasts; can be analysed e.g. in terms of *phonemes* or *distinctive features*
## Distinctive features

<table>
<thead>
<tr>
<th>Feature</th>
<th>t</th>
<th>z</th>
<th>m</th>
<th>l</th>
<th>i</th>
</tr>
</thead>
<tbody>
<tr>
<td>anterior</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>coronal</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>labial</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>distributed</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>consonantal</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>sonorant</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>voiced</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>approximant</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>continuant</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>lateral</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>nasal</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>strident</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
# THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993)

## CONSONANTS (PULMONIC)

<table>
<thead>
<tr>
<th></th>
<th>Bilabial</th>
<th>Labiodental</th>
<th>Dental</th>
<th>Alveolar</th>
<th>postalveolar</th>
<th>Retractal</th>
<th>Palatal</th>
<th>Velar</th>
<th>Uvular</th>
<th>Pharyngeal</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plosive</td>
<td>p</td>
<td>b</td>
<td>t</td>
<td>d</td>
<td>t̪</td>
<td>d̪</td>
<td>c</td>
<td>j</td>
<td>k̞</td>
<td>g</td>
<td>ḏ</td>
</tr>
<tr>
<td>Nasal</td>
<td>m</td>
<td>n̞</td>
<td>n̞</td>
<td>n̞</td>
<td>n̞</td>
<td>n̞</td>
<td>n̞</td>
<td>n̞</td>
<td>n̞</td>
<td>n̞</td>
<td>n̞</td>
</tr>
<tr>
<td>Trill</td>
<td>b</td>
<td>r</td>
<td>r̪</td>
<td>r̪</td>
<td>r̪</td>
<td>r̪</td>
<td>r̪</td>
<td>r̪</td>
<td>r̪</td>
<td>r̪</td>
<td>r̪</td>
</tr>
<tr>
<td>Tap or Flap</td>
<td>r̪</td>
<td>r</td>
<td>r̪</td>
<td>r̪</td>
<td>r̪</td>
<td>r̪</td>
<td>r̪</td>
<td>r̪</td>
<td>r̪</td>
<td>r̪</td>
<td>r̪</td>
</tr>
<tr>
<td>Fricative</td>
<td>f̪</td>
<td>v̪</td>
<td>θ̪</td>
<td>ð̪</td>
<td>s̪</td>
<td>z̪</td>
<td>θ̪</td>
<td>ð̪</td>
<td>s̪</td>
<td>z̪</td>
<td>s̪</td>
</tr>
<tr>
<td>Lateral</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>l̪</td>
<td>j̪</td>
<td>l̪</td>
<td>j̪</td>
<td>l̪</td>
<td>j̪</td>
<td>l̪</td>
</tr>
<tr>
<td>Approximant</td>
<td>v̪</td>
<td>i̪</td>
<td>ũ</td>
<td>l̪</td>
<td>l̪</td>
<td>i̪</td>
<td>l̪</td>
<td>i̪</td>
<td>l̪</td>
<td>i̪</td>
<td>l̪</td>
</tr>
<tr>
<td>Lateral</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>l̪</td>
<td>i̪</td>
<td>l̪</td>
<td>i̪</td>
<td>l̪</td>
<td>i̪</td>
<td>l̪</td>
</tr>
</tbody>
</table>

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

## CONSONANTS (NON-PULMONIC)

<table>
<thead>
<tr>
<th>Clicks</th>
<th>Voiced implosives</th>
<th>Ejectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>O Bilabial</td>
<td>ɓ ɓ̪ ɓ̱</td>
<td>ɓ̱</td>
</tr>
<tr>
<td>I Dental</td>
<td>d̪ ḏ ɗ ɗ̱</td>
<td>ḏ</td>
</tr>
<tr>
<td>I (Postalveolar)</td>
<td>t̪ ṯ ɗ̱</td>
<td>ṯ</td>
</tr>
<tr>
<td>Palatoalveolar</td>
<td>ḳ ḵ ɗ̱</td>
<td>ḵ</td>
</tr>
<tr>
<td>Alveolar lateral</td>
<td>g̣ g̱ ɗ̱</td>
<td>g̱</td>
</tr>
<tr>
<td>Uvular</td>
<td>ṣ s̱ ɗ̱</td>
<td>s̱</td>
</tr>
<tr>
<td>Alveolar fricative</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## VOWELS

<table>
<thead>
<tr>
<th>Front</th>
<th>Central</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>i̯</td>
<td>i̯</td>
<td>u̯</td>
</tr>
<tr>
<td>ɪ̯</td>
<td>ɪ̯</td>
<td>ʊ̯</td>
</tr>
<tr>
<td>Open-mid</td>
<td>æ̯</td>
<td>ø̯ ə̯ o̯</td>
</tr>
<tr>
<td>Open</td>
<td>æ̯ ɶ̯ ɐ̯ ə̯ ɔ̯ ɑ̯ ɔ̯</td>
<td>æ̯ ɶ̯ ɐ̯ ə̯ ɔ̯ ɑ̯ ɔ̯</td>
</tr>
</tbody>
</table>

Where symbols appear in pairs, the one to the right represents a rounded vowel.

## OTHER SYMBOLS

- Voiceless labial-velar fricative  Constraints...
- Voiced labial-velar approximant  Constraints...
- Voiced labial-velar approximant  Constraints...
- Voiceless epiglottal fricative  Constraints...
- Voiceless epiglottal fricative  Constraints...
- Epiglottal plosive  Constraints...

## SUPRASEGMENTALS

- Primary stress  Constraints...
- Secondary stress  Constraints...
- Long  Constraints...
- Half-long  Constraints...
- Extra-short  Constraints...
- Syllable break  Constraints...
- Minor (foot) group  Constraints...
- Major (intonation) group  Constraints...
- Linking (absence of a break)  Constraints...
- Downstep  Constraints...
- Upstep  Constraints...

## TONES & WORD ACCENTS

- LEVEL
- CONTOUR

## DIACRITICS

- Voiceless n̞ d̞  Constraints...
- Breathy voiced b̞ a̞  Constraints...
- Dental t̞ d̞  Constraints...
- Voiced s  Constraints...
- Creaky voiced b̞ a̞  Constraints...
- Apical t̞ d̞  Constraints...
- Aspirated t̪ d̪  Constraints...
- Linguolabial t̪ d̪  Constraints...
- Laminal t̪ d̪  Constraints...
- More rounded  Constraints...
- Labialized t̪ d̪  Constraints...
- Nasalized e̞  Constraints...
- Less rounded  Constraints...
- Palatalized t̪ d̪  Constraints...
- Nasal release d̞  Constraints...
- Advanced u̯ y̞  Constraints...
- Veralized t̪ d̪  Constraints...
- Lateral release d̞  Constraints...
- Retracted  Constraints...
- Pharyngealized t̪ d̪  Constraints...
- No audible release d̞  Constraints...
- Centralized e̞  Constraints...
- Veralized or pharyngealized t̪  Constraints...
- Mid-centralized e̞  Constraints...
- Raised  Constraints...
- Lowered b̞  Constraints...
- Syllabic  Constraints...
- Non-syllabic  Constraints...
- Advanced Tongue Root  Constraints...
- Rhoticity  Constraints...
- Retracted Tongue Root  Constraints...
Morphology

- Studies the structure and form of words
- Basic unit of meaning: *morpheme*
- Morphemes pair meaning with form, and combine to make words:
  - *e.g.*  *dogs* → *dog/DOG,Noun* + *-s/plural*
- Process complicated by exceptions and mutations
- Morphology as the interface between phonology and syntax (and the lexicon)
Types of morphological processes

- Inflection (syntax-driven):
  - *run*, *runs*, *running*, *ran*
  - *gledati*, *gledam*, *gleda*, *glej*, *gledal*,...

- Derivation (word-formation):
  - *to run*, *a run*, *runny*, *runner*, *re-run*, ...
  - *gledati*, *zagledati*, *pogledati*, *pogled*, *ogledalo*,...

- Compounding (word-formation):
  - *zvezdogled*,
  - *Herzkreislaufwiederbelebung*
Inflectional Morphology

- Mapping of form to (syntactic) function
- $dogs \rightarrow dog + s / DOG \ [N,pl]$
- In search of regularities: $talk/walk; \ talks/walks; \ talked/walked; \ talking/walking$
- Exceptions: $take/took, \ wolf/wolves, \ sheep/sheep$
- English (relatively) simple; inflection much richer in e.g. Slavic languages
Macedonian verb paradigm

<table>
<thead>
<tr>
<th></th>
<th>Present</th>
<th></th>
<th>Imperfect</th>
<th></th>
<th>Aorist</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>III</td>
<td>I</td>
<td>II</td>
<td>III</td>
</tr>
<tr>
<td>A. padn- “fall”</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1SG</td>
<td>padn</td>
<td>-am</td>
<td>padn</td>
<td>-e</td>
<td>-v</td>
</tr>
<tr>
<td>2SG</td>
<td>padn</td>
<td>-e</td>
<td>padn</td>
<td>-e</td>
<td>-še</td>
</tr>
<tr>
<td>3SG</td>
<td>padn</td>
<td>-e</td>
<td>padn</td>
<td>-e</td>
<td>-še</td>
</tr>
<tr>
<td>1PL</td>
<td>padn</td>
<td>-e</td>
<td>padn</td>
<td>-e</td>
<td>-v</td>
</tr>
<tr>
<td>2PL</td>
<td>padn</td>
<td>-e</td>
<td>padn</td>
<td>-e</td>
<td>-v</td>
</tr>
<tr>
<td>3PL</td>
<td>padn</td>
<td>-at</td>
<td>padn</td>
<td>-e</td>
<td>-a</td>
</tr>
<tr>
<td>B. nos- “carry”</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1SG</td>
<td>nos</td>
<td>-am</td>
<td>nos</td>
<td>-e</td>
<td>-v</td>
</tr>
<tr>
<td>2SG</td>
<td>nos</td>
<td>-i</td>
<td>nos</td>
<td>-e</td>
<td>-še</td>
</tr>
<tr>
<td>3SG</td>
<td>nos</td>
<td>-i</td>
<td>nos</td>
<td>-e</td>
<td>-še</td>
</tr>
<tr>
<td>1PL</td>
<td>nos</td>
<td>-i</td>
<td>nos</td>
<td>-e</td>
<td>-v</td>
</tr>
<tr>
<td>2PL</td>
<td>nos</td>
<td>-i</td>
<td>nos</td>
<td>-e</td>
<td>-v</td>
</tr>
<tr>
<td>3PL</td>
<td>nos</td>
<td>-at</td>
<td>nos</td>
<td>-e</td>
<td>-a</td>
</tr>
<tr>
<td>C. id- “go”</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1SG</td>
<td>id</td>
<td>-am</td>
<td>id</td>
<td>-e</td>
<td>-v</td>
</tr>
<tr>
<td>2SG</td>
<td>id</td>
<td>-e</td>
<td>id</td>
<td>-e</td>
<td>-še</td>
</tr>
<tr>
<td>3SG</td>
<td>id</td>
<td>-c</td>
<td>id</td>
<td>-c</td>
<td>-še</td>
</tr>
<tr>
<td>1PL</td>
<td>id</td>
<td>-e</td>
<td>id</td>
<td>-e</td>
<td>-v</td>
</tr>
<tr>
<td>2PL</td>
<td>id</td>
<td>-e</td>
<td>id</td>
<td>-e</td>
<td>-v</td>
</tr>
<tr>
<td>3PL</td>
<td>id</td>
<td>-at</td>
<td>id</td>
<td>-e</td>
<td>-a</td>
</tr>
</tbody>
</table>

Table 3.2: Finite Forms of the Macedonian Verb
Syntax

- How are words arranged to form sentences?
  *I milk like
  I saw the man on the hill with a telescope.

- The study of rules which reveal the structure of sentences (typically tree-based)

- A “pre-processing step” for semantic analysis

- Common terms:
  Subject, Predicate, Object,
  Verb phrase, Noun phrase, Prepositional phr.,
  Head, Complement, Adjunct,...
Syntactic theories

• Transformational Syntax
  N. Chomsky: TG, GB, Minimalism
• Distinguishes two levels of structure: deep and surface; rules mediate between the two
• Logic and Unification based approaches (’80s) : FUG, TAG, GPSG, HPSG, ...
• Phrase based vs. dependency based approaches
Example of a phrase structure and a dependency tree

```plaintext
S
  NP
    D N P
     Some people in
  PP
    N P D N
      this class
  VP
    V
      draw
    AP
      Adv A
        extremely funny
    NP
      N
        trees
```

```
#1
  AuxS
  pisal
    Pred
      AuxK
  Winston
    Sb
      je
        AuxV
          v
            AuxP
              dnevnik
                Adv
```
Semantics

- The study of *meaning* in language
- Very old discipline, esp. philosophical semantics (Plato, Aristotle)
- Under which conditions are statements true or false; problems of quantification
- The meaning of words – lexical semantics
  
  \textit{spinster} = unmarried female $\rightarrow \ast \text{my brother is a spinster}$
Discourse analysis and Pragmatics

• Discourse analysis: the study of connected sentences – behavioural units (anaphora, cohesion, connectivity)

• Pragmatics: language from the point of view of the users (choices, constraints, effect; pragmatic competence; speech acts; presupposition)

• Dialogue studies (turn taking, task orientation)
Lexicology

- The study of the vocabulary (lexis / lexemes) of a language (a lexical “entry” can describe less or more than one word)
- Lexica can contain a variety of information: sound, pronunciation, spelling, syntactic behaviour, definition, examples, translations, related words
- Dictionaries, mental lexicon, digital lexica
- Plays an increasingly important role in theories and computer applications
- Ontologies: WordNet, Semantic Web
HLT research fields

- **Phonetics and phonology**: speech synthesis and recognition
- **Morphology**: morphological analysis, part-of-speech tagging, lemmatisation, recognition of unknown words
- **Syntax**: determining the constituent parts of a sentence (NP, VP) and their syntactic function (Subject, Predicate, Object)
- **Semantics**: word-sense disambiguation, automatic induction of semantic resources (thesauri, ontologies)
- **Multi-lingual technologies**: extracting translation equivalents from corpora, machine translation
- **Internet**: information extraction, text mining, advanced search engines
Further reading

- Language Technology World
  http://www.lt-world.org/
- The Association for Computational Linguistics
  http://www.aclweb.org/ (c.f. Resources)
- Natural Language Processing – course materials
  http://www.cs.cornell.edu/Courses/cs674/2003sp/