Language Technologies

Module "Knowledge Technologies"
Jožef Stefan International Postgraduate School
Winter 2011 / Spring 2012

Lecture I.
Introduction to Language Technologies

Tomaž Erjavec

Basic info

- 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/mps11-hlt/

Assesment

- Seminar work on topic connected with HLT
  - ½ quality of work
  - ½ quality of report
- Today: presentation of some possible topics + choosing the topic by students
- Next lecture: March 28th
  - Presentation by students on work / problems so far
- May: submission of seminar
- Each student can have 1 hr of consultations

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?

- AI-complete:
  - To solve NLP, you'd need to solve all of the problems in AI
- Turing test
  - Engaging effectively in linguistic behavior is a sufficient condition for having achieved intelligence.
- …But little kids can "do" NLP…

Problems

Languages have properties that humans find easy to process, but are very problematic for computers

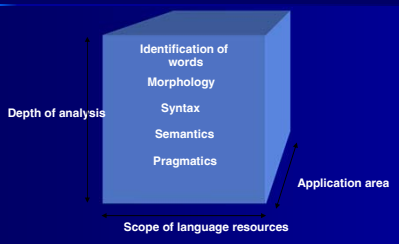
- Ambiguity: many words, syntactic constructions, etc. have more than one interpretation
- Vagueness: many linguistic features are left implicit in the text
- Paraphrases: many concepts can be expressed in different ways

Humans use context and background knowledge; both are difficult for computers

Ambiguity

- "I scream" vs. "ice cream"
- It's very hard to recognize speech.
  It's very hard to wreck a nice beach.
- Squad helps dog bite victim.
  Helicopter powered by human flies.
- Jack invited Mary to the Halloween ball.

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*)
  - Machine Learning: "*data-driven automatic inference of rules*"

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
- Empiricism 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:

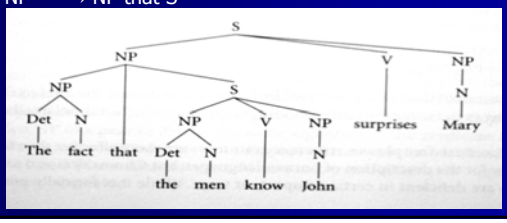
- an underlying, more abstract form, termed 'deep structure',
- the actual form of the sentence produced, called 'surface structure'.

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
- Strict formal specifications (at first), but problems of overpremissivness
- Chomsky's Development: Transformational Grammar (1957, 1964), …, Government and Binding/Principles and Parameters (1981), Minimalism (1995)

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)

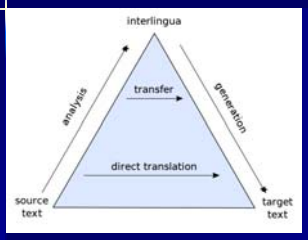
Machine translation

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

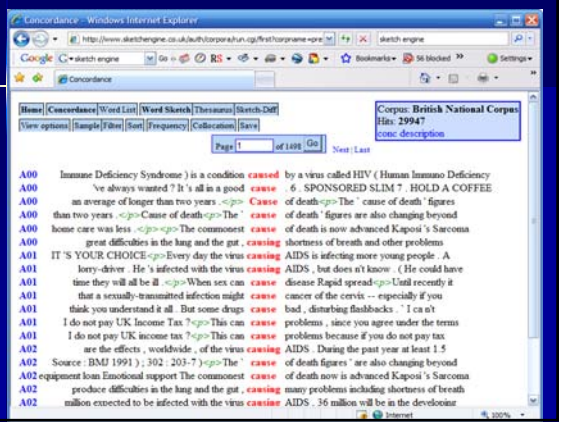
Information retrieval and extraction

- **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!)
  - traning or testing of programs

Concordances

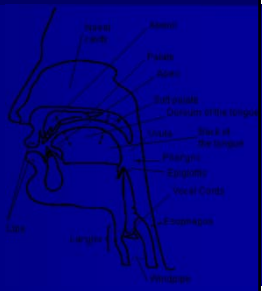


IV. Levels of linguistic analysis

- 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)



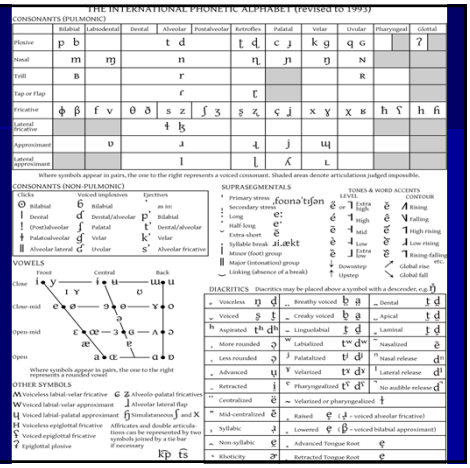
Phonology

- 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

	t	z	m	l	i
anterior	+	+	+	+	–
coronal	+	+	–	+	–
labial	–	–	+	–	–
distributed	–	–	–	–	–
consonantal	+	+	+	+	–
sonorant	–	–	+	+	+
voiced	–	+	+	+	+
approximant	–	–	–	+	+
continuant	–	+	–	+	+
lateral	–	–	–	+	–
nasal	–	–	+	–	–
strident	–	+	–	–	–

IPA



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* → *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

	Present			Imperfect				Aorist			
		I	III		I	II	III		I	II	III
A. padn- "fall"											
1SG	padn		-am	padn	-e	-v		padn	-a	-v	
2SG	padn	-e	-š	padn	-e		-še	padn	-a		
3SG	padn	-e		padn	-e		-še	padn	-a		
1PL	padn	-e	-me	padn	-e	-v	-me	padn	-a	-v	-me
2PL	padn	-e	-te	padn	-e	-v	-te	padn	-a	-v	-te
3PL	padn		-at	padn	-e		-a	padn	-a		-a
B. nos- "carry"											
1SG	nos		-am	nos	-e	-v		iznos	-i	-v	
2SG	nos	-i	-š	nos	-e		-še	iznos	-i		
3SG	nos	-i		nos	-e		-še	iznos	-i		
1PL	nos	-i	-me	nos	-e	-v	-me	iznos	-i	-v	-me
2PL	nos	-i	-te	nos	-e	-v	-te	iznos	-i	-v	-te
3PL	nos		-at	nos	-e		-a	iznos	-i		-a
C. id- "go"											
1SG	id		-am	id	-e	-v		id	-o	-v	
2SG	id	-e	-š	id	-e		-še	id	-e		
3SG	id	-e		id	-e		-še	id	-e		
1PL	id	-e	-me	id	-e	-v	-me	id	-o	-v	-me
2PL	id	-e	-te	id	-e	-v	-te	id	-o	-v	-te
3PL	id		-at	id	-e		-a	id	-o		-a

Table 3.2: Finite Forms of the Macedonian Verb

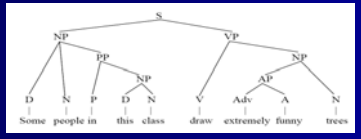
Syntax

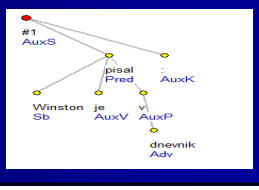
- How are words arranged to form sentences?
  **I milk like*
  I saw the man on the hill with a telescope.
- The study of rules which reveal the structure of sentences (typically tree-based)
- A "pre-processing step" for semantic analysis
- Common terms:
  Subject, Predicate, Object,
  Verb phrase, Noun 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





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
  spinster = unmarried female → **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)
- **Multiulingual 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/