

Language Technologies

"New Media and eScience" MSc Programme
Jožef Stefan International Postgraduate School

Winter Semester 2008/09

Lecture I.
Introduction to Human Language
Technologies

Tomaž Erjavec

Technicalities

- Lecturer:
[http://nl.ijs.si/et/
tomaz.erjavec@ijs.si](http://nl.ijs.si/et/tomaz.erjavec@ijs.si)
- Course homepage:
<http://nl.ijs.si/et/teach/mps08-ht/>
- Assessment:
Seminar work
- Next Wednesday: introduction to datasets
- Exam dates

Introduction to Human Language Technologies

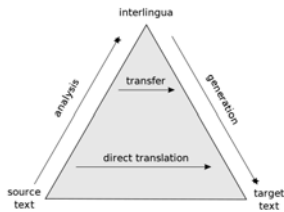
1. Application areas of language technologies
2. The science of language: linguistics
3. Computational linguistics: some history
4. HLT: Processes, methods, and resources

	<h2>I. Applications of HLT</h2>
	<ul style="list-style-type: none"> ■ Speech technologies ■ Machine translation ■ Information retrieval and extraction <ul style="list-style-type: none"> ■ Text summarisation, text mining ■ Question answering, dialogue systems ■ Multimodal and multimedia systems ■ Computer assisted authoring; language learning; translating; lexicology; language research

	<h2>Speech technologies</h2>
	<ul style="list-style-type: none"> ■ speech synthesis ■ speech recognition ■ speaker verification (biometrics, security) ■ spoken dialogue systems ■ speech-to-speech translation ■ speech prosody: emotional speech ■ audio-visual speech (talking heads)

	<h2>Machine translation</h2>
	<p>Perfect MT would require the problem of NL understanding to be solved first!</p> <p>Types of MT:</p> <ul style="list-style-type: none"> ■ Fully automatic MT (babelfish, Google translate) ■ Human-aided MT (pre and post-processing) ■ Machine aided HT (translation memories) <p>Problem of evaluation!</p> <ul style="list-style-type: none"> ■ automatic (BLEU, METEOR) ■ manual (expensive!)

MT approaches



- rule based:
rules +
lexicons
- statistical:
parallel corpora

Statistical MT

- parallel corpora: text in original language + translation
- on the basis of parallel corpora only:
induce statistical model of translation
- very influential approach: now used in Google translate

Information retrieval and extraction

- **Information retrieval (IR)** is the science of searching for documents, for information within documents and for metadata about documents.
 - "bag of words" approach
- **Information extraction (IE)** is a type of information retrieval 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 Extraction**
 - identify names, dates, numeric expression in text

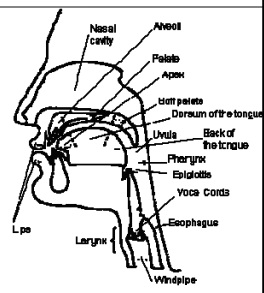
	<h2>II. Background: Linguistics</h2>
	<ul style="list-style-type: none"> ■ What <i>is</i> language? ■ The science of language ■ Levels of linguistics analysis

	<h2>Language</h2>
	<ul style="list-style-type: none"> ■ <i>Act of speaking</i> in a given situation (parole or performance) ■ The <i>abstract system</i> underlying the collective totality of the speech/writing behaviour of a community (langue) ■ The <i>knowledge of this system</i> by an individual (competence) <p><u>De Saussure</u> (structuralism ~ 1910) parole / langue</p> <p><u>Chomsky</u> (generative ling. > 1960) performance / competence</p>

	<h2>What is Linguistics?</h2>
	<p>The scientific study of language</p> <ul style="list-style-type: none"> ■ Prescriptive vs. descriptive ■ Diachronic vs. synchronic ■ Performance vs. competence ■ Anthropological, clinical, psycho, socio,... linguistics ■ General, theoretical, formal, mathematical, computational linguistics

	<h2>Levels of linguistic analysis</h2>
	<ul style="list-style-type: none"> ■ Phonetics ■ Phonology ■ Morphology ■ Syntax ■ Semantics ■ Discourse analysis ■ Pragmatics ■ + Lexicology

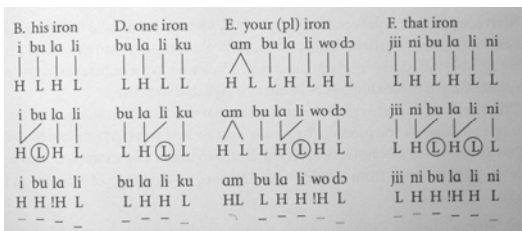
	<h2>Phonetics</h2>
	<ul style="list-style-type: none"> ■ Studies how sounds are produced; methods for description, classification, transcription ■ Articulatory phonetics (how sounds are made) ■ Acoustic phonetics (physical properties of speech sounds) ■ Auditory phonetics (perceptual response to speech sounds)



	<h2>Phonology</h2>
	<ul style="list-style-type: none"> ■ 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 <i>phonemes</i> or <i>distinctive features</i> ■ Segmental vs. suprasegmental phonology ■ Generative phonology, metrical phonology, autosegmental phonology, ... (two-level phonology)

Autosegmental phonology

- A multi-layer approach:



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

Table 3.2: Finite Forms of the Macedonian Verb

The declension of Slovene adjectives

Pridevnik

PRIDEVNIK

I. NAGLAS NA ISTEM ZLOGU OSNOVNIH OBLIK

ed. im. m.	z.	s.	rod. m.	rod. z.	rod. s.	im. im. m.	dat.	pret.	prizemnik	prizemnik pret.
-e	-a	-o	-ega	-o	-i	-i	-i	-e	-ji	-je in -je

Opomba: Pridevnik imajo lahko tudi samo določilo obliko.

II. NAGLAS NA KONČNICI (OZVROMA ZADNJEM ZLOGU)

ed. im. m.	z.	s.	rod. m.	rod. z.	rod. s.	im. im. m.	dat.	pret.	prizemnik	prizemnik pret.
-e	-a	-o	-ega	-o	-i	-i	-i	-e	-ji	-je in -je

Characteristics of Slovene inflectional morphology

- Paradigmatic morphology: fused morphs, many-to-many mappings between form and function:
*hodil-a*_[masculine dual], *stol-a*_[singular, genitive], *sošed-*_[singular, genitive].
- Complex relations within and between paradigms: syncretism, alternations, multiple stems, defective paradigms, the boundary between inflection and derivation,...
- Large set of morphosyntactic descriptions (>1000) Ncmsn, Ncmsg, Ncmpn,...
- MULTEXT-East [tables for Slovene](#)

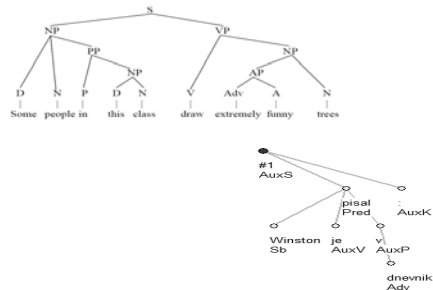
Syntax

- 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)

	<h2>Lexicology</h2>
	<ul style="list-style-type: none"> ■ 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

	<h2>III. The history of Computational Linguistics</h2>
	<ul style="list-style-type: none"> ■ MT, empiricism (1950-70) ■ The Generative paradigm (70-90) ■ Data fights back (80-00) ■ A happy marriage? ■ The promise of the Web

	<h2>The early years</h2>
	<ul style="list-style-type: none"> ■ The promise (and need!) for machine translation ■ The decade of optimism: 1954-1966 ■ <i>The spirit is willing but the flesh is weak ≠ The vodka is good but the meat is rotten</i> ■ 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.

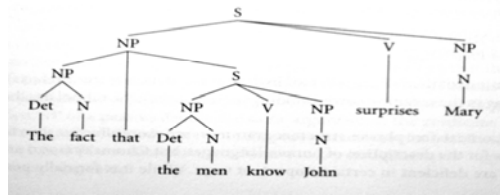
Phrase structure rules and derivation trees

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 overpermissiveness
- Chomsky's Development: Transformational Grammar (1957, 1964), ..., Government and Binding/Principles and Parameters (1981), Minimalism (1995)

	<h2 style="margin: 0;">Problems</h2>
	<p>Disadvantage of rule-based (deep-knowledge) systems:</p> <ul style="list-style-type: none"> ■ Coverage (lexicon) ■ Robustness (ill-formed input) ■ Speed (polynomial complexity) ■ Preferences (the problem of ambiguity: "<i>Time flies like an arrow</i>") ■ 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?

	<h2 style="margin: 0;">Back to data</h2>
	<ul style="list-style-type: none"> ■ Late 1980's: applied methods based on data (the decade of "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)

	<h2 style="margin: 0;">The new millennium</h2>
	<p>The emergence of the Web:</p> <ul style="list-style-type: none"> ■ Simple to access, but hard to digest ■ Large and getting larger ■ Multilinguality <p>The promise of mobile, 'invisible' interfaces; HLT in the role of middle-ware</p>

Processes, methods, and resources

The Oxford Handbook of Computational Linguistics,

Ruslan Mitkov (ed.)

- Text-to-Speech Synthesis
- Speech Recognition
- Text Segmentation
- Part-of-Speech Tagging and lemmatisation
- Parsing
- Word-Sense Disambiguation
- Anaphora Resolution
- Natural Language Generation
- Finite-State Technology
- Statistical Methods
- Machine Learning
- Lexical Knowledge Acquisition
- Evaluation
- Sublanguages and Controlled Languages
- Corpora
- Ontologies
