

Language Technologies

“New Media and eScience” MSc Programme
Jožef Stefan International Postgraduate School

Winter/Spring Semester, 2007/08

Lecture I. Introduction to Human Language Technologies

[Tomaž Erjavec](#)

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

Applications of HLT

- Speech technologies
- Machine translation
- Information retrieval and extraction, text summarisation, text mining
- Question answering, dialogue systems
- Multimodal and multimedia systems
- Computer assisted:
authoring; language learning;
translating; lexicology; language
research

Speech technologies

- speech synthesis
- speech recognition
- speaker verification (biometrics, security)

- spoken dialogue systems
- speech-to-speech translation
- speech prosody: emotional speech
- audio-visual speech (talking heads)

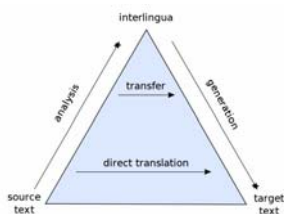
Machine translation

Perfect MT would require the problem of NL understanding to be solved first!

Types of MT:

- Fully automatic MT ([babelfish](#))
- Human-aided MT (pre and post-processing)
- Machine aided HT (translation memories)

MT approaches



- rule based: rules + lexicons
- statistical: parallel corpora
- problem of evaluation

Background: Linguistics

- What is language?
- The science of language
- Levels of linguistics analysis

Language

- *Act of speaking* in a given situation (**parole** or **performance**)
- The *abstract system* underlying the collective totality of the speech/writing behaviour of a community (**langue**)
- The *knowledge of this system* by an individual (**competence**)

[De Saussure](#)

(structuralism ~ 1910) parole / langue

[Chomsky](#)

(generative ling. > 1960) performance / competence

What is Linguistics?

The scientific study of language

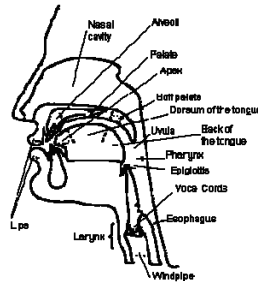
- Prescriptive vs. descriptive
- Diachronic vs. synchronic
- Performance vs. competence
- Anthropological, clinical, psycho, socio,... linguistics
- General, theoretical, formal, mathematical, computational linguistics

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*
- Segmental vs. suprasegmental phonology
- Generative phonology, metrical phonology, autosegmental phonology, ... (two-level phonology)

Distinctive features

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

I
P
A

THE INTERNATIONAL PHONETIC ALPHABET (REVISED TO 1993)

CONSONANTS (PULMONIC)											
Place	Labial	Dental/alveolar	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b	t d					c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	n		ɳ			ɲ	ŋ	ɴ		
Fricative		f v	s z	ʃ ʒ		x ɣ		χ ʁ	ħ ʕ		
Approximant				ɹ			j				
Glottal											h
Labial-velar											w
Lateral approximant				l							ʎ

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

CONSONANTS (NON-PULMONIC)

Click	Ceased ejectives	Ejectives	Primary stress	Secondary stress	Long	Half-long	Extra-short	Halfable break	Voiceless glottal fricative	Voiceless velar fricative	Voiceless uvular fricative	Voiceless pharyngeal fricative	Voiceless epiglottal fricative	Voiceless alveolar fricative	Voiceless dental fricative	Voiceless labiodental fricative	Voiceless labial fricative	Voiceless bilabial fricative
ʘ	ɕ	ɛ	ˈ	ˌ	ː	ˑ	ː̥	ɬ	ɸ	ɕ	ʁ̥	ħ̥	ʕ̥	h̥	ɬ̥	ɸ̥	ʘ̥	ʘ̥

VOWELS

OTHER SYMBOLS

ɰ	voiceless labial-velar fricative	ɱ	labiodental fricative	ɯ	voiceless labial-uvular fricative	ɷ	labiodental fricative	ɸ	voiceless labial-velar fricative	ɸ̥	voiceless labial-velar fricative	ɸ̥̥	voiceless labial-velar fricative	ɸ̥̥̥	voiceless labial-velar fricative	ɸ̥̥̥̥	voiceless labial-velar fricative	ɸ̥̥̥̥̥	voiceless labial-velar fricative
---	----------------------------------	---	-----------------------	---	-----------------------------------	---	-----------------------	---	----------------------------------	----	----------------------------------	-----	----------------------------------	------	----------------------------------	-------	----------------------------------	--------	----------------------------------

Generative phonology

A consonant becomes devoiced if it starts a word:

[C, +voiced] → [-voiced] / # ____
 e.g. #vlak# → #flak#

- Rules change the structure
- Rules apply one after another (feeding and bleeding)
- (in contrast to two-level phonology)

Autosegmental phonology

- A multi-layer approach:

B. his iron i bu la li H L H L	D. one iron bu la li ku L H L L	E. your (pl) iron am bu la li wo dɔ H L L H L H L	F. that iron jii ni bu la li ni L H L H L L
i bu la li H L H L	bu la li ku L H L L	am bu la li wo dɔ H L L H L H L	jii ni bu la li ni L H L H L L
i bu la li H H H L	bu la li ku L H H L	am bu la li wo dɔ HL L H H H L	jii ni bu la li ni L H H H H L

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,p]
- 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. <i>padn</i> - "fall"										
1SG	<i>padn</i>	-am			<i>padn</i>	-o	-v	<i>padn</i>	-a	-v
2SG	<i>padn</i>	-e	-š		<i>padn</i>	-e	-še	<i>padn</i>	-a	
3SG	<i>padn</i>	-e			<i>padn</i>	-e	-še	<i>padn</i>	-a	
1PL	<i>padn</i>	-e	-me		<i>padn</i>	-e	-me	<i>padn</i>	-a	-me
2PL	<i>padn</i>	-e	-te		<i>padn</i>	-e	-te	<i>padn</i>	-a	-te
3PL	<i>padn</i>	-at			<i>padn</i>	-e	-a	<i>padn</i>	-a	-a
B. <i>nos</i> - "carry"										
1SG	<i>nos</i>	-am			<i>nos</i>	-o	-v	<i>nos</i>	-i	-v
2SG	<i>nos</i>	-i	-š		<i>nos</i>	-e	-še	<i>nos</i>	-i	
3SG	<i>nos</i>	-i			<i>nos</i>	-e	-še	<i>nos</i>	-i	
1PL	<i>nos</i>	-i	-me		<i>nos</i>	-e	-me	<i>nos</i>	-i	-me
2PL	<i>nos</i>	-i	-te		<i>nos</i>	-e	-te	<i>nos</i>	-i	-te
3PL	<i>nos</i>	-at			<i>nos</i>	-e	-a	<i>nos</i>	-i	-a
C. <i>id</i> - "go"										
1SG	<i>id</i>	-am			<i>id</i>	-o	-v	<i>id</i>	-o	-v
2SG	<i>id</i>	-e	-š		<i>id</i>	-e	-še	<i>id</i>	-o	
3SG	<i>id</i>	-e			<i>id</i>	-e	-še	<i>id</i>	-o	
1PL	<i>id</i>	-e	-me		<i>id</i>	-o	-me	<i>id</i>	-o	-me
2PL	<i>id</i>	-e	-te		<i>id</i>	-o	-te	<i>id</i>	-o	-te
3PL	<i>id</i>	-at			<i>id</i>	-o	-a	<i>id</i>	-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

Pridevnik na -e, -a, -o imajo samo nedoločno obliko.

Pridevnik na -i, -i, -i in -i imajo štiri dvojne priloge na -i (tudi če je v predmetu samo oblika na -i).

Pridevnik z oblikami, ki ne pomenijo opredeljenosti, so naglasni v naglavju. Prizemniki pridevnik, ki jih imajo še posebej dopolnilni pridevnik v obliki -e, -a, -o, -i, se delajo po obliki. Opišni prizemniki v bolj slovar posebej se razlikujejo, ker so moški, če pomen dopolnila, od vseh pridevnikov in od določnikov na -i, -a, -o.

Pridevnik, ki so po obliki prizemniki (bilo), imajo samo prazenik.

Prizemnik priloga nastopa le ob prizemniku pridevnik.

Užbenik	Užbenik	Užbenik	Užbenik	Užbenik	Užbenik	Užbenik	Užbenik	Užbenik	Užbenik	Užbenik

Oprema: Pridevnik imajo lahko tudi samo določno obliko.

II. NAGLAS NA KONČNICI (OZVROMA ZADNEM 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-u*_[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) Ncmsgn, Ncmsg, Ncmprn,...
- MULTTEXT-East [tables for Slovene](#)

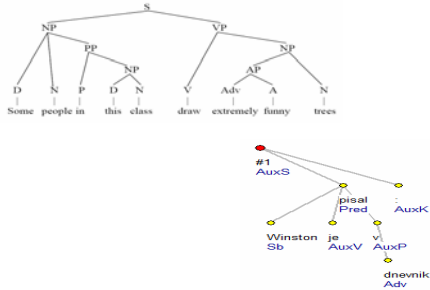
Syntax

- 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)

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

The history of Computational Linguistics

- MT, empiricism (1950-70)
- The Generative paradigm (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.

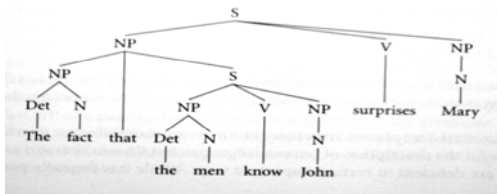
Phrase structure rules and derivation trees

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

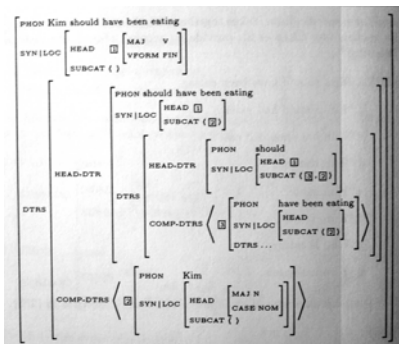
Computational linguistics

- 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,..)

Unification-based grammars

- Based on research in artificial intelligence, logic programming, constraint satisfaction, inheritance reasoning, object oriented programming,..
- The basic data structure is a feature-structure: attribute-value, recursive, co-indexing, typed; modelled by a graph
- The basic operation is unification: information preserving, declarative
- The formal framework for various linguistic theories: GPSG, HPSG, LFG,...
- Implementable!

An example HPSG feature structure



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 (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, Candide)
- Importance of evaluation (resources, methods)

The new millennium

The emergence of the Web:

- Simple to access, but hard to digest
- Large and getting larger
- Multilinguality

The promise of mobile, 'invisible' interfaces;
HLT in the role of middle-ware

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
