Introduction to Human Language Technologies

Tomaž Erjavec

Karl-Franzens-Universität Graz

Lecture 1: Overview
9.11.2007

Overview

1. a few words about me
2. a few words about you
3. introduction to HLT
4. lab work: first steps with Python

Lecturer

- Tomaž Erjavec
  Department of Knowledge Technologies
  Jožef Stefan Institute
  Ljubljana
- http://nl.ijs.si/et/
- tomaz.erjavec@ijs.si
- Work: corpora and other language resources, standards, annotation, text-critical editions
- Web page for this course: http://nl.ijs.si/et/teach/graz07/hlt/
- assessment

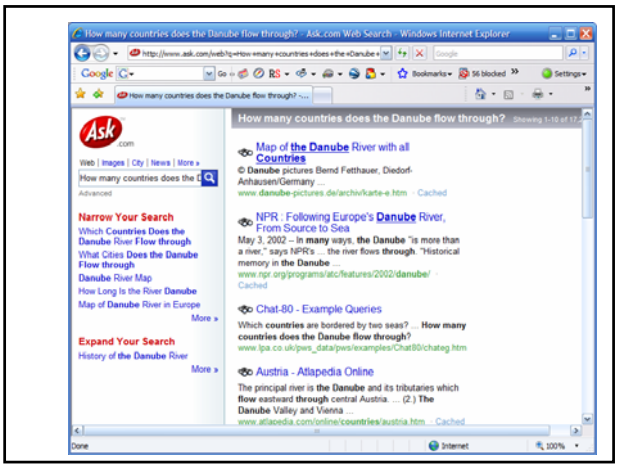
Students

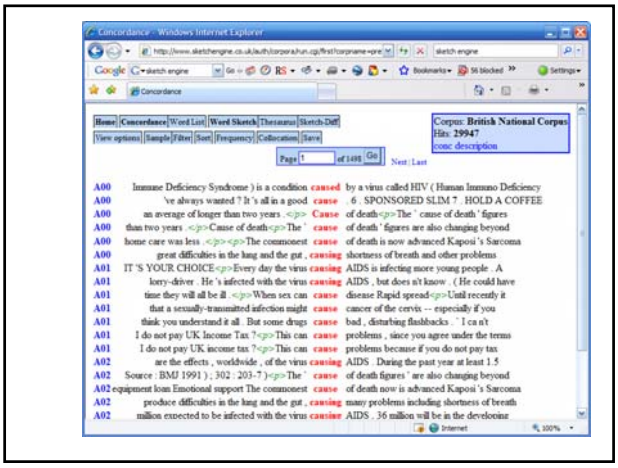
- background: field of study
- exposure to
  - linguistics?
  - corpus linguistics?
  - programming?
- emails

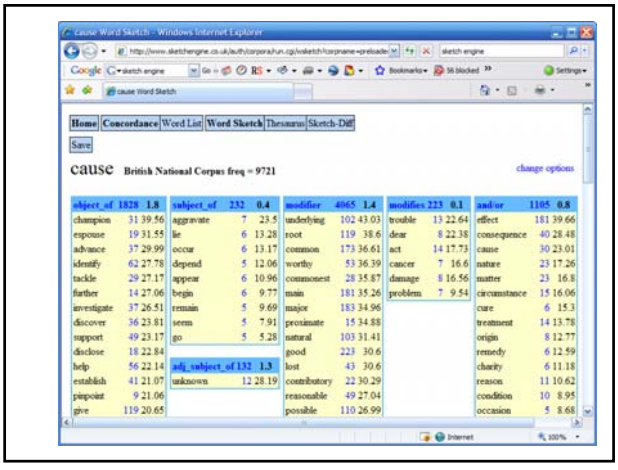
Overview of the course

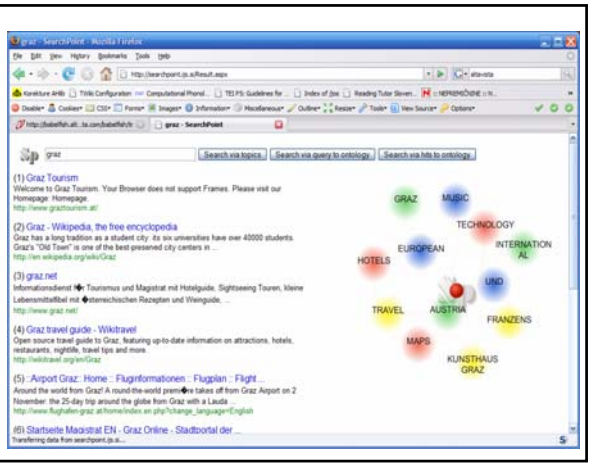
1. Introduction
2. Basic processing of text
3. Working with corpora
4. Multilingual applications
5. Lexical semantics
6. …

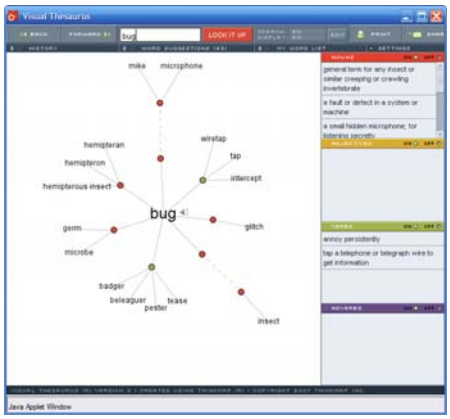
Lectures + work with NLTK


How many countries does the Danube flow through? - Ask.com Web Search - Windows Internet Explorer
How many countries does the Danube flow through?
Narrow Your Search
Which Countries Does the Danube River Flow through
What Cities Does the Danube Flow through
Danube River Map
How Long Is the River Danube
Map of Danube River in Europe
More »
Expand Your Search
History of the Danube River
More »
Map of the Danube River with all Countries
© Danube pictures Bernd Fetthauer, Diedorf-Anhausen/Germany ...
www.danube-pictures.de/archiv/karte-e.htm - Cached
NPR : Following Europe's Danube River, From Source to Sea
May 3, 2002 -- In many ways, the Danube "is more than a river," says NPR's ... the river flows through. "Historical memory in the Danube ...
www.npr.org/programs/atc/features/2002/danube/ - Cached
Chat-80 - Example Queries
Which countries are bordered by two seas? ... How many countries does the Danube flow through?
www.lpa.co.uk/pws_data/pws/examples/Chat80/chateg.htm
Austria - Atlapedia Online
The principal river is the Danube and its tributaries which flow eastward through central Austria. ... (2.) The Danube Valley and Vienna ...
www.atlapedia.com/online/countries/austria.htm - Cached


Concordance - Windows Internet Explorer
Home | Concordance | Word List | Word Sketch | Thesaurus | Sketch-Diff
View options | Sample | Filter | Sort | Frequency | Collocation | Save
Corpus: British National Corpus
Hits: 29947
conc description
Page 1 of 1495 Go
Next | Last
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 . Cause of death The ' cause of death ' figures
A00 than two years . Cause of death The ' cause of death ' figures are also changing beyond
A00 home care was less . 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 .
A01 IT 'S YOUR CHOICE 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 . When sex can cause disease Rapid spread 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 ca n't
A01 I do not pay UK Income Tax ? This can cause problems , since you agree under the terms
A01 I do not pay UK income tax ? 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) 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


cause Word Sketch - Windows Internet Explorer
Home | Concordance | Word List | Word Sketch | Thesaurus | Sketch-Diff
Save
cause British National Corpus freq = 9721
change options
object_of 1828 1.8
champion 31 39.56
espouse 19 31.55
advance 37 29.99
identify 62 27.78
tackle 29 27.17
further 14 27.06
investigate 37 26.51
discover 36 23.81
support 49 23.17
disclose 18 22.84
help 56 22.14
establish 41 21.07
pinpoint 9 21.06
give 119 20.65
subject_of 232 0.4
aggravate 7 23.5
lie 6 13.28
occur 6 13.17
depend 5 12.06
appear 6 10.96
begin 6 9.77
remain 5 9.69
seem 5 7.91
go 5 5.28
adj_subject_of 132 1.3
unknown 12 28.19
modifier 4065 1.4
underlying 102 43.03
root 119 38.6
common 173 36.61
worthy 53 36.39
commonest 28 35.87
main 181 35.26
major 183 34.96
proximate 15 34.88
natural 103 31.41
good 223 30.6
lost 43 30.6
contributory 22 30.29
reasonable 49 27.04
possible 110 26.99
modifies 223 0.1
trouble 13 22.64
dear 8 22.38
act 14 17.73
cancer 7 16.6
damage 8 16.56
problem 7 9.54
and/or 1105 0.8
effect 181 39.66
consequence 40 28.48
cause 30 23.01
nature 23 17.26
matter 23 16.8
circumstance 15 16.06
cure 6 15.3
treatment 14 13.78
origin 8 12.77
remedy 6 12.59
charity 6 11.18
reason 11 10.62
condition 10 8.95
occasion 5 8.68


graz - SearchPoint - Mozilla Firefox
graz
Search via topics
Search via query to ontology
Search via hits to ontology
(1) Graz Tourism
(2) Graz - Wikipedia, the free encyclopedia
(3) graz.net
(4) Graz travel guide - Wikitravel
(5) : Airport Graz: Home :: Fluginformationen :: Flugplan :: Flight ...
(6) Startseite Magistrat EN - Graz Online - Stadtportal der ...
GRAZ
MUSIC
TECHNOLOGY
EUROPEAN
INTERNATIONAL
HOTELS
UND
TRAVEL
AUSTRIA
FRANZENS
MAPS
KUNSTHAUS GRAZ


Visual Thesaurus
bug
LOOK IT UP
mike
microphone
wiretap
tap
intercept
hemipteran
hemipteron
hemipterous insect
bug
germ
glitch
microbe
badger
beleaguer
tease
pester
insect
Java Applet Window


Searching the Japanese-Slovene Dictionary jaSlo - Mozilla Firefox
Japanese-Slovene Dictionary jaSlo
Find
ponos
Part of Speech: All Noun Verb Adjective Phrase
Search in: All Headwords Headword translations
5 entries:
(jisonshin) じそんしん 【自尊心】 (N)
samospoštovanje
ponos
NIVO 1
(jiman) じまん 【自慢】 (Vs)
ponos
ponašanje
bahavost
ponašati se
みんなに自分（じぶん）のバイクを自慢する
vsem se ponašati s svojim motorjem
この子供（こども）はみんなの自慢でした。
To otrok je bil ponos vseh.
NIVO 2
Done

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 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*)
  - 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 post-natal, etc.*

Problems

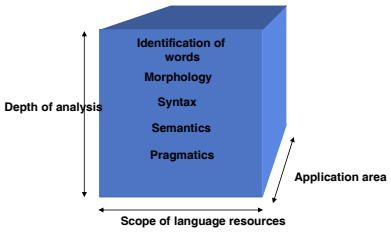
Languages have properties that humans find easy to process, but are very problematic for computers

- 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

- 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 (written and spoken)
  - Generalisations are possible only 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
- Empirical turn supported by rise in processing speed of computers and their amount of storage, and the revolution in the availability of machine-readable texts (the word-wide web)

The history of Computational Linguistics

- MT, empiricism (1950-70)
- Structuralism: 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:

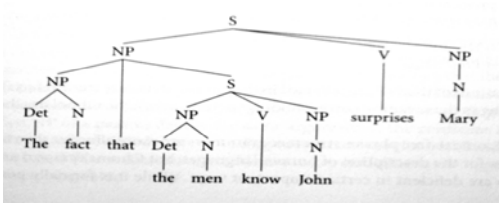
- an underlying, more abstract form, termed 'deep structure',
- the actual form of the sentence produced, called 'surface structure'.

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 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 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

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

HLT applications II.

- 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

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

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

Further reading

- Language Technology World
  http://www.lt-world.org/
- The Association for Computational Linguistics
  http://www.aclweb.org/ (c.f. Resources)
- Interactive Online CL Demos
  http://www.ifi.unizh.ch/CL/InteractiveTools.html
- Natural Language Processing – course materials
  http://www.cs.cornell.edu/Courses/cs674/2003sp/