Language and Computation

week 1 (January 14, 2014)

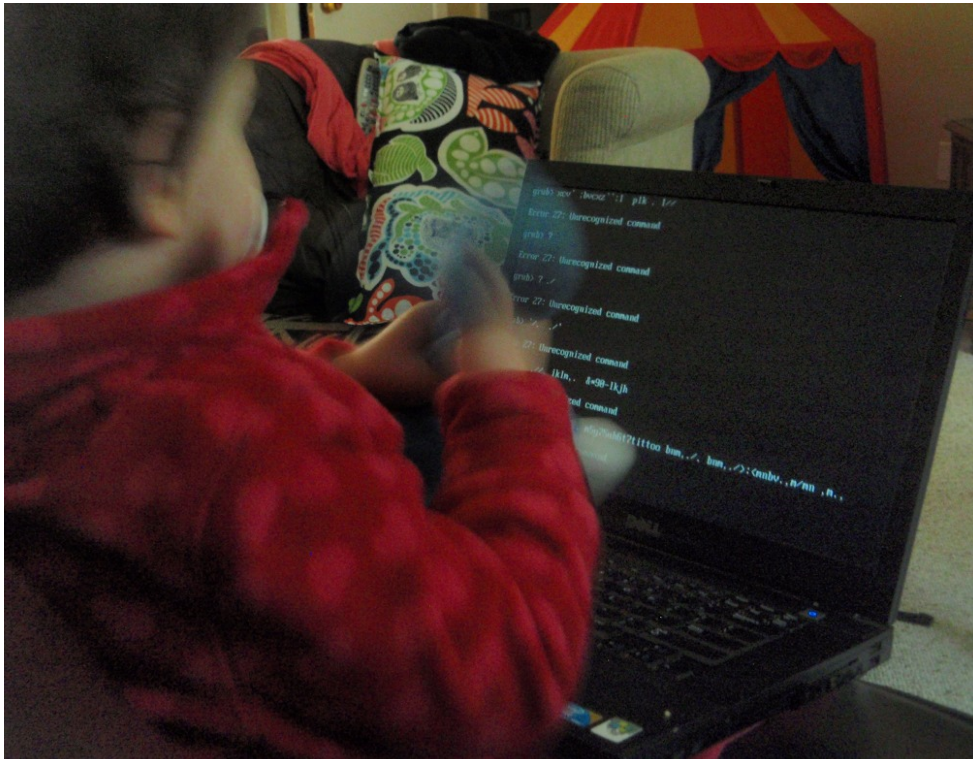
Tamás Biró
Yale University
tamas.biro@yale.edu

http://www.birot.hu/courses/2014-LC/

Tamás Biró

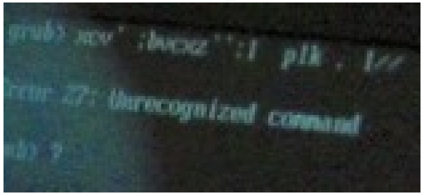
- www.birot.hu
- http://ling.yale.edu/people/tam-s-bir
- http://www.birot.hu/courses/2014-LC/
- Budapest → Amsterdam → New Haven
- Call by first name, most pronunciations accepted.

Human-machine interaction: no common language?



Please, launch
'ABC song'
on Youtube!

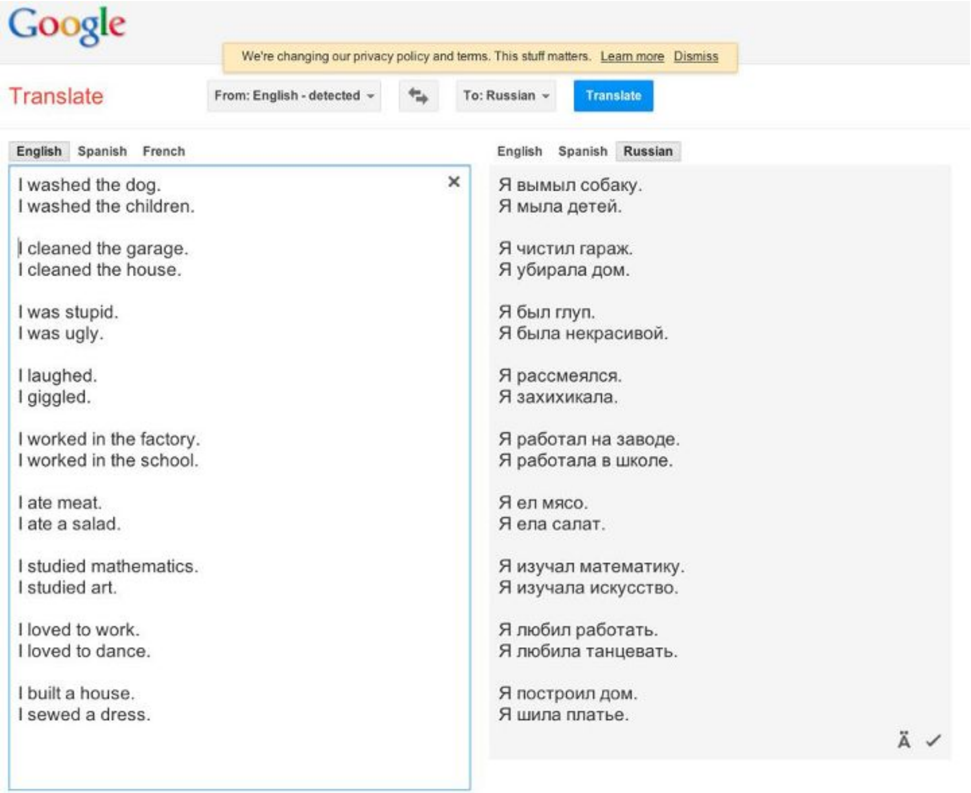
```
bcz'':| pIk , l //
```



```
grub> xcv' :bvcxz'':l plk , l//
rror 27: Unrecognized command
ub> ?
```

```
Unrecognized
command
```

The oldest problem: Machine Translation (MT)



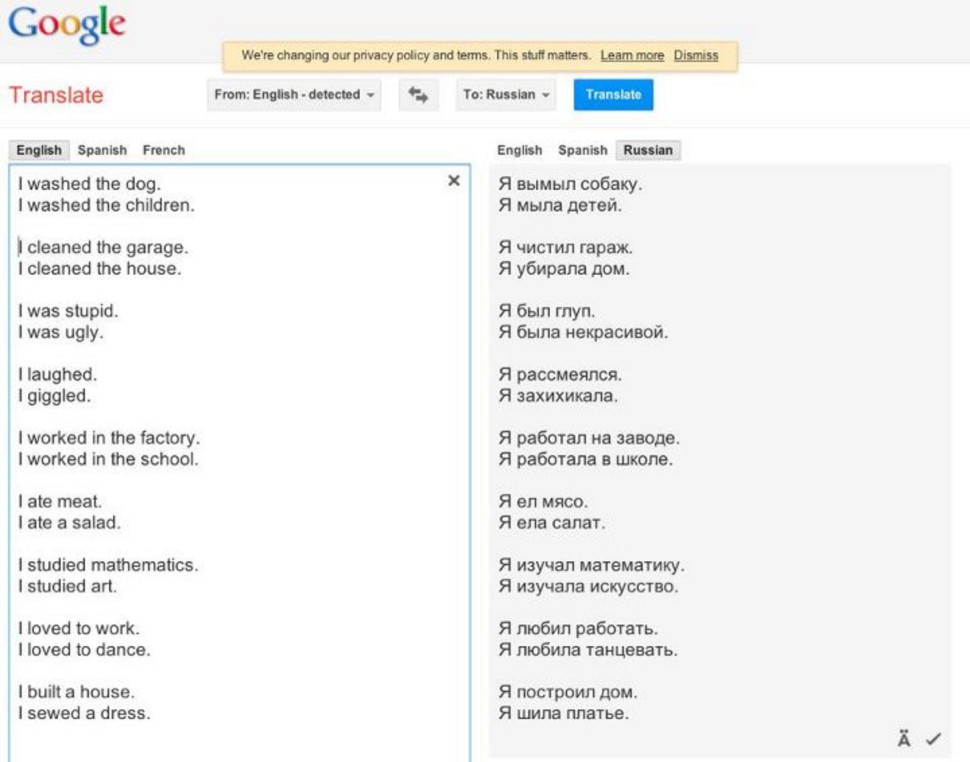
We're changing our privacy policy and terms. This stuff matters.

From: English - detected

To: Russian

I washed the dog.	Я вымыл собаку.
I washed the children.	Я мыла детей.
I cleaned the garage.	Я чистил гараж.
I cleaned the house.	Я убирала дом.
I was stupid.	Я был глуп.
I was ugly.	Я была некрасивой.
I laughed.	Я рассмеялся.
I giggled.	Я захихикала.
I worked in the factory.	Я работал на заводе.
I worked in the school.	Я работала в школе.
I ate meat.	Я ел мясо.
I ate a salad.	Я ела салат.
I studied mathematics.	Я изучал математику.
I studied art.	Я изучала искусство.
I loved to work.	Я любил работать.
I loved to dance.	Я любила танцевать.
I built a house.	Я построил дом.
I sewed a dress.	Я шила платье.

The oldest problem: Machine Translation (MT)



We're changing our privacy policy and terms. This stuff matters. Learn more Dismiss

I washed the dog.
I washed the children.

I cleaned the garage.
I cleaned the house.

I was stupid.
I was ugly.

I laughed.
I giggled.

I worked in the factory.
I worked in the school.

I ate meat.
I ate a salad.

I studied mathematics.
I studied art.

I loved to work.
I loved to dance.

I built a house.
I sewed a dress.

Я вымыл собаку.
Я мыла детей.

Я чистил гараж.
Я убирала дом.

Я был глуп.
Я была некрасивой.

Я рассмеялся.
Я захихикала.

Я работал на заводе.
Я работала в школе.

Я ел мясо.
Я ела салат.

Я изучал математику.
Я изучала искусство.

Я любил работать.
Я любила танцевать.

Я построил дом.
Я шила платье.

Gender bias:

Reflects corpus.
Good translation?

What is our goal?

satisfy linguists
or satisfy users ?

By the end of the course:
understand complexities underlying tasks such as MT.

Ambiguity

(1)
The thieves stole the paintings.
They were subsequently found.

(2)
Les voleurs ont volé les peintures.
Ils ont été trouvés plus tard.

(3)
Les voleurs ont volé les peintures.
Elles ont été trouvées plus tard.

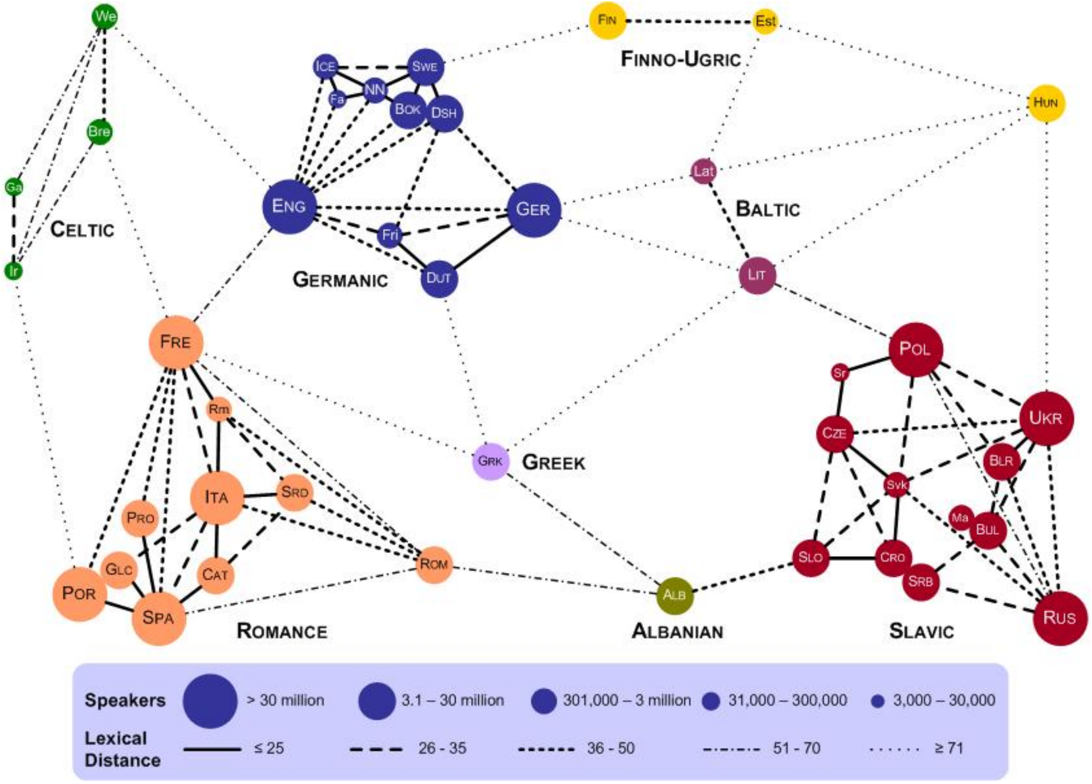
(B. Bird, E. Klein, E. Loper: *Natural Language Processing with Python*. O'Reilly, 2009.)

Ambiguity

I saw the man with the telescope.

I made her duck.

A recent craze on Facebook



`http://elms.wordpress.com/2008/03/04/lexical-distance-among-languages-of-europe/`

Language & Computation: at the intersection of

- Linguistics

  but we often have a perspective different from the linguists'.

- Cognitive science

  but we often do not care about language in human mind.

- Engineering

  but language does not follow standards.

Questions

What is language?

-
-
-
-
-

What is computation?

-
-
-
-
-

Questions

What is language?

- Speech?
- Communication?
- Texts?
- Words?
- Meaning?

What is computation?

- Using numbers?
- Using mathematics?
- Using algorithms?
- Using computers, as a tool?
- Using computers, as an aim?

"Computational Linguistics"

- "Computer Linguistics"? Ger. *Computerlinguistik* Russian *kompyuternaya lingvistika*. Fr. *Linguistique informatique*.
- NLP = Natural Language Processing
- Language technology and speech technology
- Formal / mathematical linguistics
  Or: computational aspects of (theoretical) linguistics?

Computational approaches to **natural** language.

“Computer Linguistics”

- Computers supporting research:
  e.g., counting verbs and nouns in a large corpus.

- Computers enabling research:
  e.g., creating a tree bank from a large corpus.

- Computers as the goal of the research:
  e.g., annotating a large corpus with part-of-speech tags.

- Computing, such as in a computer or as in the brain:
  e.g., algorithms for part-of-speech tagging.

"Computer Linguistics"

- Computers supporting research:
  e.g., counting speech errors in a spoken corpus.

- Computers enabling research:
  e.g., counting speech errors on the web.

- Computers as the goal of the research:
  e.g., natural language generation that is perfect or natural.

- Computing, such as in a computer or as in the brain:
  e.g., an algorithm that generates language with the same errors as humans do.

Computational Linguistics (CL)

The linguist's perspective:

- Can my linguistic theory be formalized mathematically?
- Is my linguistic theory equivalent to another theory? Is it stronger or weaker? What does it predict?
- Can my linguistic theory be implemented in a computational system, such as the human brain or a personal computer?
- Can I use it for both production and interpretation (parsing)?
- Is my linguistic theory learnable?
- Can it predict speech errors, production time, etc.?

Computational Cognitive Science

. . . applied to linguistic tasks (*computational psycholinguistics*)

The cognitive scientist's perspective:

Similar questions, but very different frameworks, which have not been inspired by linguistic scholarly traditions and linguistic observations, rather by research in non-linguistic (psychological, general cognitive) domains.

Natural Language Processing (NLP)

The computer scientist's perspective:

I am processing information, which can be

- numbers → *numerical analysis*
- program code → *programming language theory*
- databases → *database theory*
- etc.
- natural language → *NLP*

NB: natural language (as well as images, etc.) contain information to be processed that is *not* in a "well-defined format".

Language technology

The computer engineer's perspective:

I develop software with users in mind, which will be sold on the market. Therefore, my system must outperform its competitors. My software. . .

- interacts with humans
- reads, processes or returns words, sentences or texts
- performs a task that is somehow related to language

Speech technology

The electrical engineer's perspective:

I develop systems (for users on the market, which is full of competitors) that process or produce sound. In some cases, the sound entering or exiting my system is human speech sound.

Therefore, my system should be able to. . .

- "recognize" speech sound → automatic speech recognition
- generate speech-like sound → text-to-speech (voice synthesis)

Research questions and unbounded possibilities

- Can we use the Internet as a corpus?
- Language technology for endangered languages!
- Can we crack the language software in the brain?
- **Text categorization:** who authored the *Federalist Papers*?
- **Dialectometry:** automatically detect dialect boundaries?
- **Digital humanities:** NLP-supported humanities research.

Language technology applications

- Military and intelligence
- Human-computer interaction
- Search engines
- New media: language detection, emotion detection, etc.
- Biomedical information extraction
- Plagiarism detection
- etc.

What do we cover in this introductory course?

Topics:

- Levels of language: word, sound, sentence, discourse.
- Tasks: spell checking, speech recognition, question answering, word-sense disambiguation, machine translation. . .
- Computational concepts handling these levels of language: regular expressions, formal languages, stochastic models, machine learning. . .
- Some basic approaches (algorithms) to solve these tasks.

What do we cover in this introductory course?

Skills:

- The computational linguist's perspective on language.
- The computational linguist's skills: understanding a problem, understanding a proposed solution, understanding a formula, understanding an algorithm. . .
  Requires notion such as: probability, precision and recall. . .
- . . . and implementing it.
  (How to program: *Python* for language processing.)

Form of this course

- **Lectures:** Tu and Th, 4:00-5:15, in WLH 113.
- **Readings:** "pre-readings" and "post-readings".
- **Sections:** optional, for consultation.
  TA: Jen Runds. Mo 1:30-2:20 and 2:30-3:20, WLH 205.
- **Programming sections:** optional, priority to linguists.
  If interested, email me by Monday 01/20!
  Suggestion: We 4:30-6:00, CLAY-lab (t.b.c.).

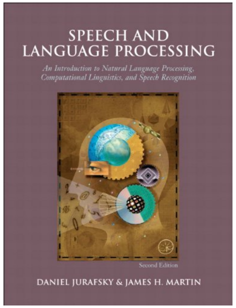
Requirements

- **Class participation** (10%).
- **Homework assignments:** approx. biweekly (40%).
- **Midterm:** take-home, due: March 25. (20%).
- **Final:** in-class, on-paper, open-book. Fr 5/2, 9 am (30%).
- **Term paper** for *grad students* (contact me a.s.a.p.).

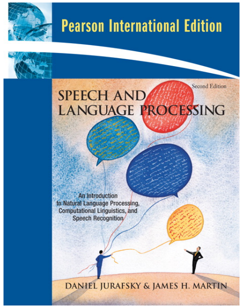
Policies: see syllabus.

A note: exams are also part of the learning process.

Books: Natural Language Processing

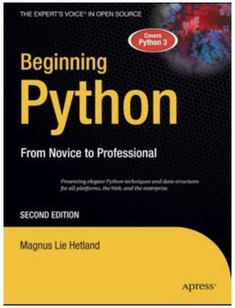

SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN

or

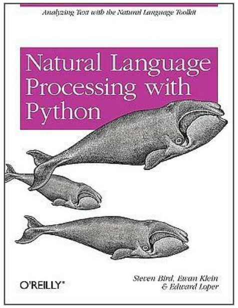

Pearson International Edition
SPEECH AND LANGUAGE PROCESSING
Second Edition
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
DANIEL JURAFSKY & JAMES H. MARTIN

Daniel Jurafsky, James H. Martin: *Speech and Language Processing*, 2nd Edition. Pearson Prentice Hall, 2008.

Books: Python



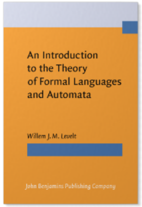
Magnus Lie Hetland: *Beginning Python: From Novice to Professional*, 2nd Edition. Apress, 2008.


Analyzing Text with the Natural Language Toolkit
Natural Language Processing with Python
O'REILLY
Steven Bird, Ewan Klein & Edward Loper

B. Bird, E. Klein, E. Loper: *Natural Language Processing with Python*. O'Reilly, 2009.

Complementing this course

- Robert Frank: *Formal Foundations of Linguistic Theories*, LING 224/624 (fall semester)

- 

  An Introduction to the Theory of Formal Languages and Automata

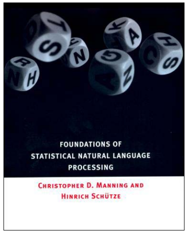
  Willem J.M. Levelt

  John Benjamins Publishing Company

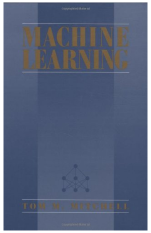
  Formal language theory and automata theory, including advanced topics, especially *mildly context-sensitive languages* (TAG: Tree-Adjoining Grammars, and equivalent formalisms).

- Theoretical linguistic frameworks with formal/computational motivation / background: LFG, HPSG, OT, etc.

Complementing this course

- 

C. Manning and H. Schütze (1999). *Foundations of Statistical Natural Language Processing*.

- 

Machine learning
As an introduction, Tom Mitchell: *Machine Learning*, 1997.
As well as many more recent techniques, e.,g, *Support Vector Machines*.

- Specific topics: e.g., speech technology, dialogue systems, question answering, etc.

Complementing this course

ESSLLI:
European Summer School in Logic, Language and Information
`http://www.esslli2014.info/`

NASSLLI:
North-American Summer School in Logic, Language and Information
`http://www.nasslli2014.com/`

Complementing this course

Association for Computational Linguistics:

- Journal: *Computational Linguistics*
- Conferences: ACL, EACL, NAACL
- ACLWiki (`http://aclweb.org/aclwiki/`) and Anthology
- `http://www.aclweb.org/`

. . . as well as many more: see J&M, end of Chapter 1.

Next:

This Thursday:

language **as** *computation*.

Next week:

regular expressions, finite state automata and transducers.

To prepare for *next week*:

- **Post-reading:** J&M Chapter 1.

  Re sections 1.3 and 1.6: Focus on the “general picture”, while do not care about the details that you do not understand. These are either pointers to later chapters in the book, or even to more advanced topics in the literature.

- **Pre-reading:** J&M Chapter 2.

  Focus on the basics and on the general picture. We’ll discuss more details in the class, and you will be able to learn the rest as post-reading.

- **Python:** Hetland, Chapter 1. Install Python.

See you on Thursday!