Computational Linguistics
A.Y. 2018/19 – C. Chesi

Lecture 1

AN INTRODUCTION TO **NATURAL LANGUAGE PROCESSING** (**NLP**) AND **LINGUISTIC COMPUTATION**

General Information

- **Goals**
  - Deep understanding of what's needed for fully describing a natural language
  - What's a corpus and how it can be used
  - How linguistic data can be (semi)automatically processed
  - Be independent in reading advanced papers in this field
- **Teaching**
  - Lectures (lecture notes and course information will be available at: http://www.ciscl.unisi.it/master/materials.htm)
  - Labs
- **Evaluation**
  - Class participation (20% of final grade)
  - Project presentation (40% of final grade)
  - Oral exam (40% of final grade) on course topics (see References)

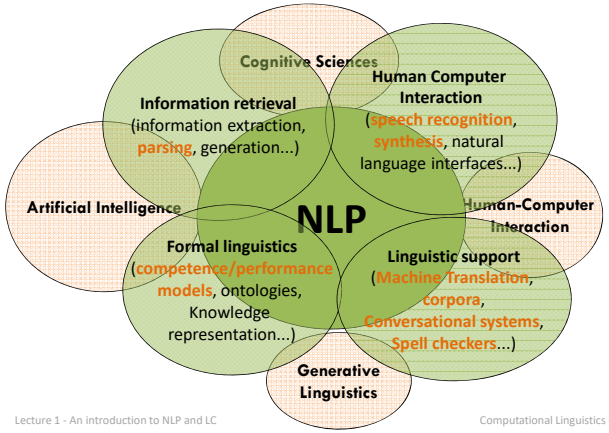
References

- **Essential references**
  - Jurafsky, D. & Martin, J. H. (2009)
    Speech and Language Processing. Prentice-Hall. (2nd edition)
    http://www.cs.colorado.edu/~martin/slp.html
    chapters: 1, 2, 3, 4, 5, 12, 13, 14, 16, 17, 18, 19, 20
- **Extended References**
  - Advanced readings will be presented at each lecture.
    Those readings won't be included at the oral exam, but they should help you in shaping your project and better understanding various aspects of NLP and LC

Today

- Introduction to Natural Language Processing (NLP)
  - Basic ideas and tools
  - Approaching natural language processing from a cognitive perspective

NLP: Natural Language Processing



NLP e Human-Computer Interaction (HCI)

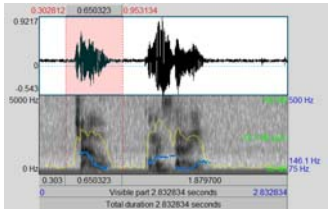
- **Eliza** (Weizenbaum, 1966):
  Human: Well, my boyfriend made me come here.
  ELIZA: *Your boyfriend made you come here?*
  Human: He says I'm depressed much of the time.
  ELIZA: *I am sorry to hear you are depressed.*
  Human: It's true. I'm unhappy.
  ELIZA: *Do you think coming here will help you not to be unhappy?*

- **HAL 9000** (Kubrick & Clarke, 2001 A Space Odyssey; 1968):
  Dave: Open the Pod bay doors, HAL.
  Hal: *I'm sorry, Dave. I'm afraid I can't do that.*

- **Wolfram Alpha** (Wolfram Co., 2016)
  question: "what's the size of an apple?"
  WA answer: *apple | maximum recorded trunk diameter > (data not available)*

What HAL 9000 should have known

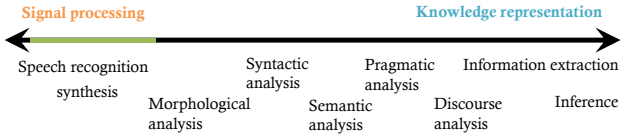
- **speech recognition / synthesis**
  - analysis/production of speech signal, formants identification, syllabification, word segmentation, prosodic contours



What HAL 9000 should have known

- **natural language understanding / generation**
  - **morphology**
    dogs = dog + s
  - **Syntax**
    [[the boy] [eats [an apple]]])
  - **Semantics**
    what does a word mean? And a sentence?
  - **Pragmatics**
    is there a communicative intention beyond the literal meaning?
  - **Discourse**
    interpreting phrases across sentences
  - **Information Extraction / Retrieval**
  - **Inference …**

A naïf plot of NLP applications



What can we do now?

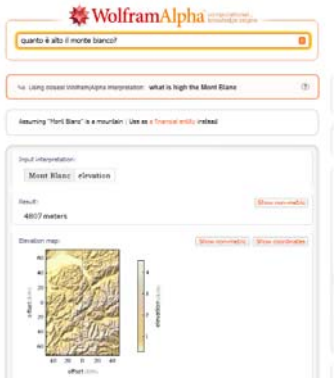
- **Word Processing**
  - **Syllabification** (extremely good)
    ex. Linguistics > Lin-guis-tics
  - **Spell-checking** (good)
    ex. houze > house
    (T9, Swipe…)
  - **Grammar-checking** (bad)
    ex. John sing > John sings
  - **Stylistic correction** (bad…)
    ex. the nail gets removed from the board > the nail is removed from the board
  - **Synonyms**, **Opposites** (thesaurus)
  - **Single word translation**
  - **…**

What can we do now?

- **Human Computer Interaction**
  - **Speech recognition** (Apple Siri, o Microsoft Cortana, Google Now)
    ex. /kasa/ > casa
  - **pseudo-comprehension**
    Where could I find a Chinese restaurant nearby? >
    [opening map with precise]
  - **Natural language generation**
    [previous context] > I'm opening your calendar…
  - **Information filtering/retrieval/extraction**
    es. "these days China Stock Exchange collapsed" >
    entity: Beijing Stock Exchange
    status: lowering;
    time: end of October2015

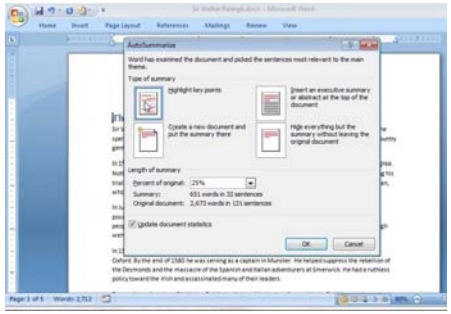
What can we do now?

- **Question answering**
  ex. How high is Monte Bianco??
  (4.810,40 m s.l.m.)



What can we do now?

- **Text summary** (introduced in Word 2003, hidden in Word 2007 e removed in Word 2010)



What can we do now?

- Human Computer Interaction
  - **Machine translation** (sufficiently good, but…)
    e.g. Google Translator :
    - tanto va la gatta al lardo che ci lascia lo zampino
      so the cat goes to the fat that it leaves its handle (April 2013)
      The pitcher goes so much that it leaves its handle (February 2014, October 2015)
      The pitcher goes so that it leaves its handle (October 2016)
      "the pitcher goes so often to the well that it leaves its handle"
    - it's raining cats and dogs

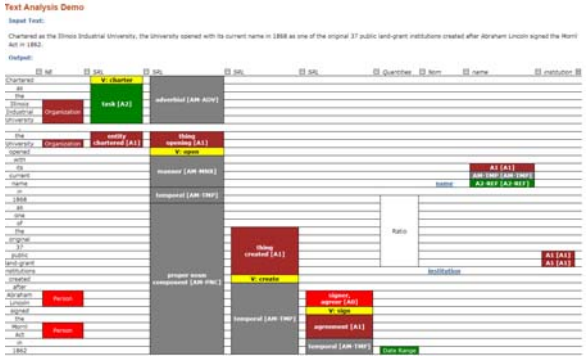
piove cani e gatti	(April 2013, February 2014)
piove a catinelle	(October 2015)
piove a secchiate	(since October 2016)

    - la vecchia legge la regola

the old law rule	(April 2013, February 2014)
the old one reads the rule	(October 2018)

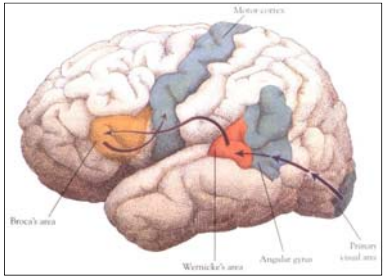
What can we do now?

- Text Analysis (e.g. Linguistic Annotation tool at CNR Pisa)



Cognitive-computational approach

- Describing cognitive modules (vision, equilibrium, movement...)
  Language is an expression of a **competence** (data-structure)
- **Processing** (specific linguistic task that uses competence models)
- **Performance** (competence usage under limited resources)

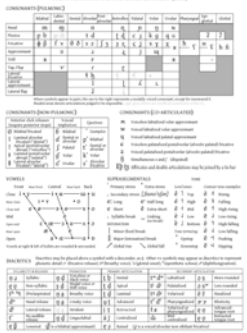


Representation of the linguistic problem

- **Competence** (data-structure)
  - What kind of information structure do we need?
    - A word can start by *wo...* (*word*) but not by *wb...*
    - The s in "sings" is different from the one in "roses"
    - "the roses are beautiful" Vs. *"the are beautiful roses"
    - "The cat chases the dog" > *subj*: cat(agent); *verb*: chase(action); *obj*: dog(patient)
    - ?the television chases the cat
    - "the houses" Vs. "some house"

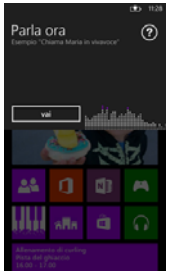
Representation of the linguistic problem

- **Competence** (data-structure)
  - Specification of primitives and features at every level:
    - **Phonemes** – distinctive features…



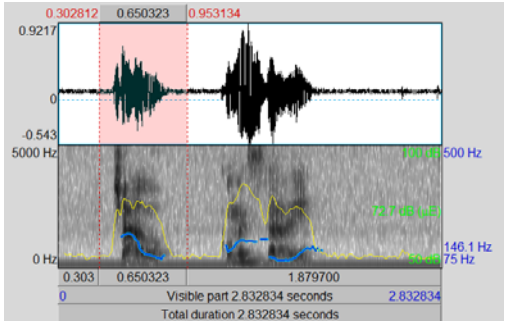
Representation of the linguistic problem

- **Speech recognition**



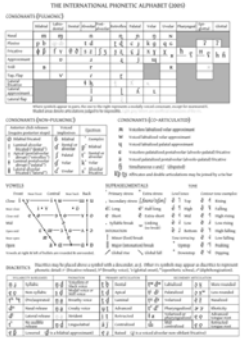
Representation of the linguistic problem

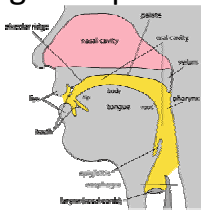
- **Speech recognition**

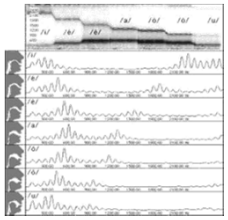


Representation of the linguistic problem

- **Speech recognition**







Representation of the linguistic problem

- **Speech Recognition**

PC	IPA	Es. 1	Es. 2
e	e	eroico	venti (number)
eh	ɛ	elle	venti (“wind” plural)
sc	ʃ	scemo	Gramsci
o	o	obesità	botte (“barrel”)
oh	ɔ	otto	botte (“strokes”)

Representation of the linguistic problem

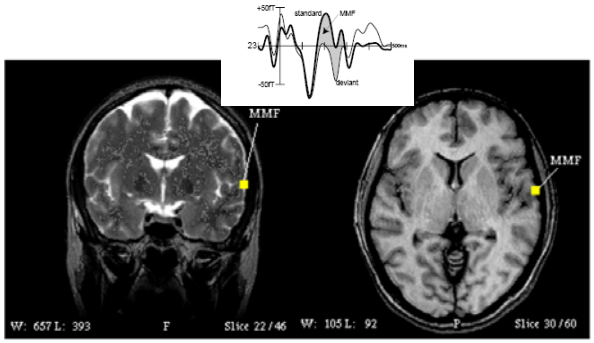


Figure 5: MRI Overlay of single-dipole localization of /dæ/ mismatch field, 190ms latency, 30-year old female subject

Phillips C. et al. (2000) Auditory Cortex Accesses Phonological Categories: An MEG Mismatch Study. Journal of Cognitive Neuroscience

Speech Recognition

- **Input**
  - «I’ll get there at 4:00»
- **Data base**
  - Features extraction
  - Annotated corpora
- **Matching**
  - Probabilistic phonetic alignment models
  - Probabilistic phrasal models

Representation of the linguistic problem

- **Competence** (data-structure)
  - Primitives:
    - **Phonemes**– distinctive features
    - **Morphemes** – combinatorial rules
    - **Words** – significant morphemes bundles
    - **Phrases** – natural groups of words
    - **Thematic roles** – agent, patient …

Representation of the linguistic problem

- **Processing** (competence in use)
  - Combinatorial principles; how do we use primitives components of competence:
    - **Phonological level –** phonotactic restrictions
    - **Morphological level** – inflectional (say > says / said) and derivational rules (easy > easily)
  - **processing** is not just **performance** (that is the use of competence under resources limitation)

Representation of the linguistic problem

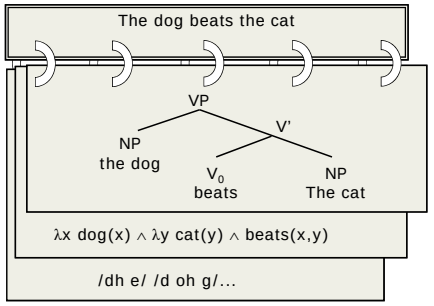
- An historic example: **Sutra for Sanskrit**

Panini (400-600AC): 1700 base elements divided in classes (e.g. nouns, verbs ecc.) + combination rules (about 4000) > Sanskrit description.

Representation of the linguistic problem

- **Lexicon**

spiral notebook model
(how do we map levels?)



Representation of the linguistic problem

- The **complexity** of the problem derives from non-univocal mapping:
  - **Lexical ambiguity** (e.g. Buffalo buffalo buffalo Buffalo buffalo)
  - **Syntactic ambiguity** (e.g. I saw the man with the binocular)
  - **Semantic ambiguity** (e.g. sheets get dried)
- A problem is harder if we have multiple choices all equally plausible (**non-determinism)**.

Representation of the linguistic problem

- Doing **Parsing** means accepting or rejecting an input;
  in case of acceptance a structural description should be provided
  - **lexical** (tagger): house > Part-of-Speech = Common noun
  - **morphological**: house > {N, countable, singular}
  - **Syntactic** (parser): $[_S\ [_{VP}\ [_{DP}$ John$]\ [_{V'}$ loves $[_{DP}$ Mary$]\ _{V'}]\ _{VP}]\ _S]$
  - **Semantic**: f(agent, patient) > loves(John, Mary)
  - ...

«Empiricist» and «rationalist» approach

- Galilean approach to problems: we do not need infinite observations of hundred of thousands of falls of leaves, or stones or apples for deducting a physical law (e.g. gravitation)
- Few observations (sometimes **just one**!) could be sufficient to validate or invalidate a theory
- Generative grammars is a theory that **uses infinite times finite means** and it is always **based on a finite number of observations**

Next lecture

- **Corpus Linguistics**
  - Historical background
  - Main theoretical models
  - Some corpora samples
  - Possible queries on corpora