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

Lecture 17

A COURSE SUMMARY

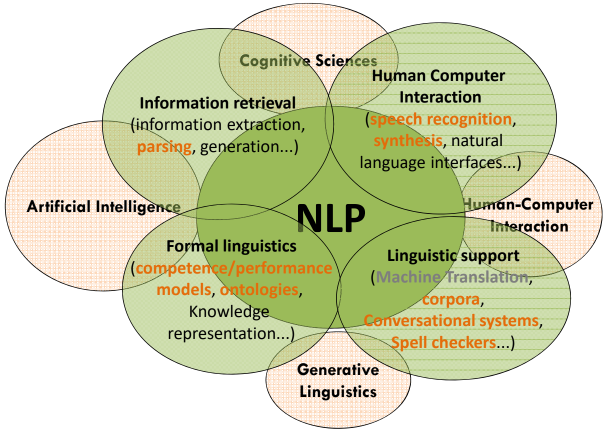
General Information

- **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
- **Lecture materials**
  - Lectures and Labs materials are available here: http://elearning.unisi.it/moodle/course/view.php?id=2085
- **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 (required for oral exam!)**
  - 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), 15, (16, 17, 18), 19, (20) (optional chapters)
- **Extended References (optional!)**
  - Advanced readings are included in each lecture header.
    Those readings won't be included in the oral exam, but they should help you in shaping your project and better understand various aspects of NLP (Natural Language Processing) and CL (Computational Linguistics)

NLP: Natural Language Processing



Cognitive Sciences

Information retrieval (information extraction, parsing, generation...)

Human Computer Interaction (speech recognition, synthesis, natural language interfaces...)

Artificial Intelligence

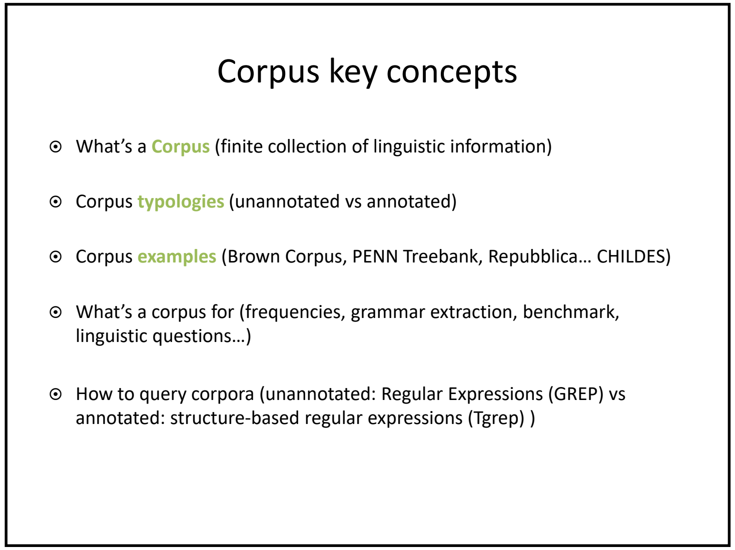
NLP

Human-Computer Interaction

Formal linguistics (competence/performance models, ontologies, Knowledge representation...)

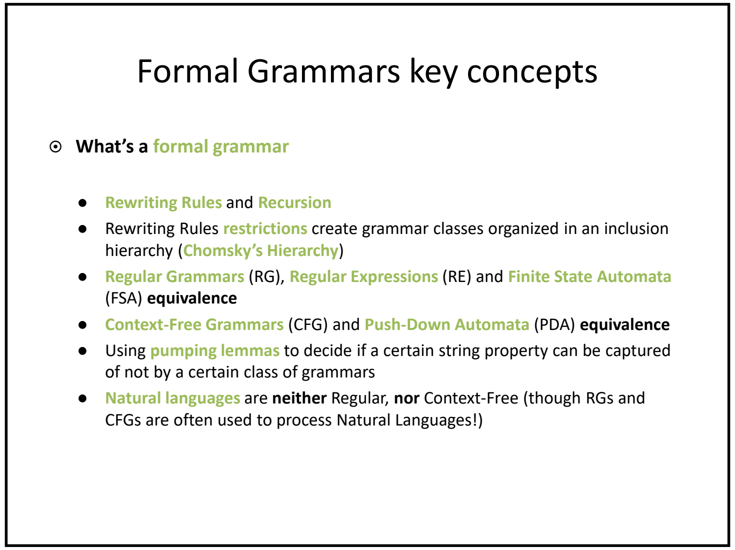
Linguistic support (Machine Translation, corpora, Conversational systems, Spell checkers...)

Generative Linguistics



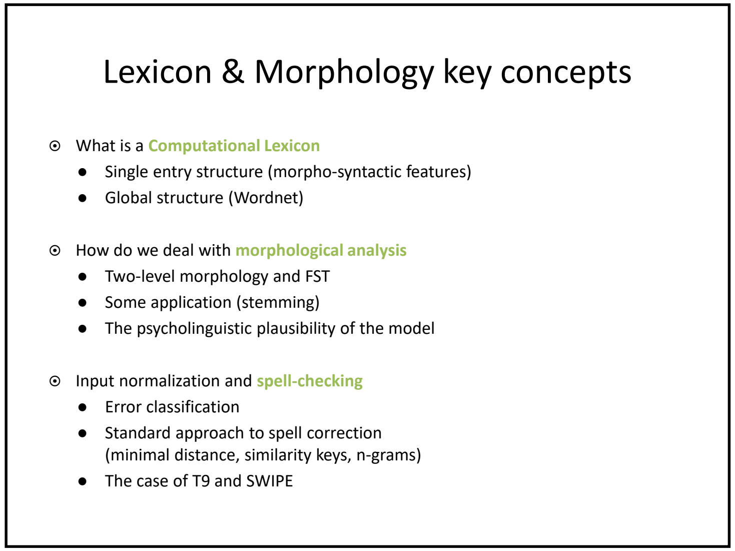
Corpus key concepts

- What's a Corpus (finite collection of linguistic information)
- Corpus typologies (unannotated vs annotated)
- Corpus examples (Brown Corpus, PENN Treebank, Repubblica… CHILDES)
- What's a corpus for (frequencies, grammar extraction, benchmark, linguistic questions…)
- How to query corpora (unannotated: Regular Expressions (GREP) vs annotated: structure-based regular expressions (Tgrep))



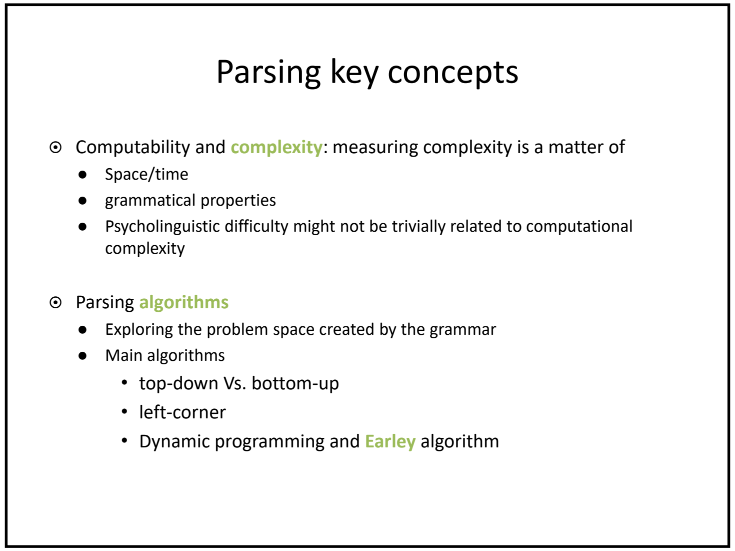
Formal Grammars key concepts

- **What's a formal grammar**
  - **Rewriting Rules** and **Recursion**
  - Rewriting Rules restrictions create grammar classes organized in an inclusion hierarchy (Chomsky's Hierarchy)
  - **Regular Grammars** (RG), **Regular Expressions** (RE) and **Finite State Automata** (FSA) **equivalence**
  - **Context-Free Grammars** (CFG) and **Push-Down Automata** (PDA) **equivalence**
  - Using pumping lemmas to decide if a certain string property can be captured of not by a certain class of grammars
  - Natural languages are **neither** Regular, **nor** Context-Free (though RGs and CFGs are often used to process Natural Languages!)



Lexicon & Morphology key concepts

- What is a Computational Lexicon
  - Single entry structure (morpho-syntactic features)
  - Global structure (Wordnet)
- How do we deal with morphological analysis
  - Two-level morphology and FST
  - Some application (stemming)
  - The psycholinguistic plausibility of the model
- Input normalization and spell-checking
  - Error classification
  - Standard approach to spell correction (minimal distance, similarity keys, n-grams)
  - The case of T9 and SWIPE



Parsing key concepts

- Computability and complexity: measuring complexity is a matter of
  - Space/time
  - grammatical properties
  - Psycholinguistic difficulty might not be trivially related to computational complexity
- Parsing algorithms
  - Exploring the problem space created by the grammar
  - Main algorithms
    - top-down Vs. bottom-up
    - left-corner
    - Dynamic programming and Earley algorithm

Advanced parsing key concepts

- **On CFG-rules inefficiency:**
  - They are language specific
  - Too many rules are difficult to control
  - They can't be possibly learned (explanatory adequacy flaw)

- **Alternatives to CFG-rules:**
  - Principle and Parameters (Fong 1991, but principles are inefficient from the computational point of view)
  - Minimalist Grammars (Stabler 2007, but merge and move operate in opposition to the parsing direction; deductive parsing is not psycholinguistically plausible)
  - Phase-based Minimalist Grammars (Chesi 2004-15, top-down derivations are cognitively plausible and can Merge and Move can be implemented this way; locality can be captured too)

Neural Network NLP key concepts

- Which problems are better approached using **sub-symbolic methods**
  - **Why** certain problems better suit sub-symbolic approaches

- **Neural networks**
  - Neurophysiology
  - Artificial Neural Networks (ANN)
  - **Past tense** learning

- NLP with neural networks: the case of Simple Recurrent Networks (SRN)
  - **Short term memory**
  - **Learning** a language = predicting next word
  - **Recursive properties** acquisition with SRN

Last Lecture

- Your Project Presentation
  - Simple problem, precisely stated
  - State of the art of the approach to the problem
  - (Computational) resources available / needed (corpora, algorithms, lexicon…)
  - Sketch of a «solution» for the problem

Some idea for your **master thesis**

- **Processing-friendly** grammatical descriptions
  do you want to use grammatical formalisms, such as **CFG**, both in **parsing** and in **generation**? Did you find that **Earley parsing** is a very interesting strategy? Then you might be interested in this project!

References

- *Chesi, C. (2015). On directionality of phrase structure building. Journal of psycholinguistic research, 44(1), 65-89);*
- *Earley, J. (1970). An efficient context-free parsing algorithm. Communications of the ACM, 13(2), 94-102.*
- *Jurafsky, D., & Martin, J. H. (2014). Speech and language processing (Vol. 3). London: Pearson. (chapters 10,11, 12 & 13: https://web.stanford.edu/~jurafsky/slp3/)*

Some idea for your **master thesis**

- **Refinement of complexity metrics**
  Do you want to interpret performance data on-line, both gathered with psycholinguistic experiments or as a result of neurophysiological investigation? Then you might be interested in this approach.

References

- Boschi V., E. Catricalà, M. Consonni, C. Chesi, A. Moro, S. Cappa (2017) *Connected speech in neurodegenerative language disorders: a review*.Front. Psychol. 8:269.doi:10.3389/fpsyg.2017.00269
- Chesi C., P. Canal (2017) Feature Retrieval Cost and on-line/off-line complexity in clefts. Proceedings of AMLAP, Architectures and Mechanisms of Language Processing. Lancaster, September 7-9 2017 http://www.nets.iusspavia.it/dox/papers/chesi_canal-2017-amlap-poster-frc.pdf
- Chesi C., A. Moro (2014) Computational complexity in the brain. In: Newmeyer FJ, Preston LB, eds. Measuring grammatical complexity, Oxford: Oxford University Press. pp. 264–280.

Some idea for your **master thesis**

- **Evidence for the acquisition of specific particles**
  Are you curious about how language growths in children? when do they start mastering quantifiers? in which context do they omit the auxiliaries or other particles? If you want to use CHILDES, find quantitative evidence of usage of specific particles such as negative markers, quantifiers, specific tense inflections… this project is for you!

References:

- MacWhinney, B. (2014). *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.
- Hurewitz, F., Papafragou, A., Gleitman, L., & Gelman, R. (2006). Asymmetries in the acquisition of numbers and quantifiers. *Language learning and development*, *2*(2), 77-96.
- Pleh, C., Vinkler, Z., & Kálmán, L. (1997). Early morphology of spatial expressions in Hungarian children: A CHILDES study. *Acta Linguistica Hungarica*, 249-260.
- Sokolov, J. L., & Snow, C. E. (Eds.). (1994). *Handbook of research in language development using CHILDES*. Hillsdale, NJ: Lawrence Erlbaum.

Some idea for your **master thesis**

- **Sentiment analysis of social comments**
  Would you like to guess automatically if a comment on twitter is positive or negative? Are you willing of investigating billions of social interactions in a blink of an eye? You might be interested in this project then.

References

- Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., & Patti, V. (2016, December). Overview of the evalita 2016 sentiment polarity classification task. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*.

Some idea for your **master thesis**

- **Conversational systems**
  Not happy with **Alexa**, **Cortana**, **Siri** and **Google**? Try a better approach!
  If you succeed you can make big money:
  https://developer.amazon.com/alexaprize

References

- Cassell, J., Sullivan, J., Churchill, E., & Prevost, S. (Eds.). (2000). *Embodied conversational agents*. MIT press.
- Khan, R., & Das, A. (2018). Build Better Chatbots. *A complete guide to getting started with chatbots*.
- Abdul-Kader, S. A., & Woods, J. C. (2015). Survey on chatbot design techniques in speech conversation systems. *International Journal of Advanced Computer Science and Applications*, *6*(7).
- Shawar, B. A., & Atwell, E. S. (2005). Using corpora in machine-learning chatbot systems. *International journal of corpus linguistics*, *10*(4), 489-516.