

# Statistical Natural Language Processing



**Instructor:** Gideon Dror

**e-mail:** [gideon@mta.ac.il](mailto:gideon@mta.ac.il)

Objectives of the Course  
and  
Preliminaries



# Why study Natural Language Processing (NLP)?

---

- NLP is a very important current area of investigation as it is necessary to many useful applications.
- These applications include: information retrieval, extraction, and filtering; intelligent Web searching; spelling and grammar checking; automatic text summarization; pseudo-understanding and generation of natural language; and multi-lingual systems including machine translation....



# Why study NLP Statistically?

---

- Up to about 5-10 years ago, NLP was mainly investigated using a **rule-based** approach.
- However, rules appear too strict to characterize people's use of language.
- This is because people tend to stretch and bend rules in order to meet their communicative needs.
- Methods for making the modeling of language more accurate are needed and **statistical methods** appear to provide the necessary flexibility.



# Subdivisions of NLP

---

- **Parts of Speech and Morphology** (words, their syntactic function in sentences, and the various forms they can take).
- **Phrase Structure and Syntax** (regularities and constraints of word order and phrase structure).
- **Semantics** (the study of the meaning of words (*lexical semantics*) and of how word meanings are combined into the meaning of sentences, etc.)
- **Pragmatics** (the study of how knowledge about the world and language conventions interact with literal meaning).



# Topics Covered in this course

---

- **Studying Words:**

- Collocations
- Statistical inference
- Word sense disambiguation
- Lexical Acquisition

- **Studying Grammars:**

- Markov Models
- Part-of-Speech Tagging
- Probabilistic Context Free Grammars

- **Applications:**

- Text categorization
- Information retrieval
- Machine translation



# Tools and Resources Used

---

- **Probability/Statistical Theory**: Statistical Distributions, Bayesian Decision Theory.
- **Linguistics Knowledge**: Morphology, Syntax, Semantics and Pragmatics.
- **Corpora**: Bodies of marked or unmarked text to which statistical methods and current linguistic knowledge can be applied in order to discover novel linguistic theories or interesting and useful knowledge organization.



# Course Requirements

---

- About 3 written and programming assignments (30%)
- A final exam (70%)

# Textbook and other useful information

- **Foundations of Statistical Natural Language Processing**, by Chris Manning and Hinrich Schütze, MIT Press.
- **Speech and Language Processing**, by Dan Jurafsky and James Martin
- **Class' Website:**  
<http://www2.mta.ac.il/~gideon/nlp.html>
- Check the class' website for a companion website for the textbook and other statistical NLP resources.

