

# **Statistical Natural Language Processing**

Detlef Prescher

Language and Inference Technology Group  
Institute for Logic, Language and Computation  
University of Amsterdam

November, 2003

# Applications of NLP

---

(NLP = Natural Language Processing)

Natural language and NLP play a central role in systems that

- **augment textual or spoken data with information** (e.g. automatic transcription of speech signals, part-of-speech tagging, named-entity recognition, parsing/chunking, word-sense disambiguation)
- **transform textual or spoken data** (e.g. text-to-speech, speech-to-text, spelling correction, text summarization, machine translation)
- **extract information from textual or spoken data** (e.g. information retrieval, question answering, information extraction, data mining)
- **communicate with people** (dialog systems)

## The aim of NLP

---

**Scientific:** Build models reflecting the human use of language and speech.

**Technological:** Build models that serve in technological applications.

The main NLP questions are:

1. What are the kind of things that people say and write?
2. What do these things mean?
3. How to incorporate the knowledge about these things into algorithms?

## The elements of NLP

---

**Phonetics/Phonology:** map acoustic signals to phoneme sequences and vice versa (speech recognition/synthesis)

**Morphology:** analyze the structure of words (morphological analysis)

**Syntax:** identify the category of words (POS tagging), analyze the structure of sentences (parsing/generation)

**Semantics:** calculate the meaning of words/sentences (lexical/compositional semantics)

**Discourse:** analyze the structure of dialog or text (discourse representation)

**Pragmatics:** incorporate world knowledge, cultural conventions, a specific use of language.

## How to build models of NLP?

---

Nowadays, there are two views on modeling what people write and say:

**Competence** (Chomsky, ~ 1960): The traditional linguistic view on NLP

- A *language* is a set of word sequences,
- A *sentence* is a sequence of words in the language,
- A *grammar* is a device defining the language,
- So *grammaticality of sentences* is defined via a set membership test.

**Performance** (~ 1990): Given a specific NLP task and a specific domain of language use, the human language-behavior is nowadays modeled by a

**(black-box) function: input  $\longrightarrow$  output**

The output that humans perceive as the most plausible for a given input.

## From competence to performance — What changed NLP?

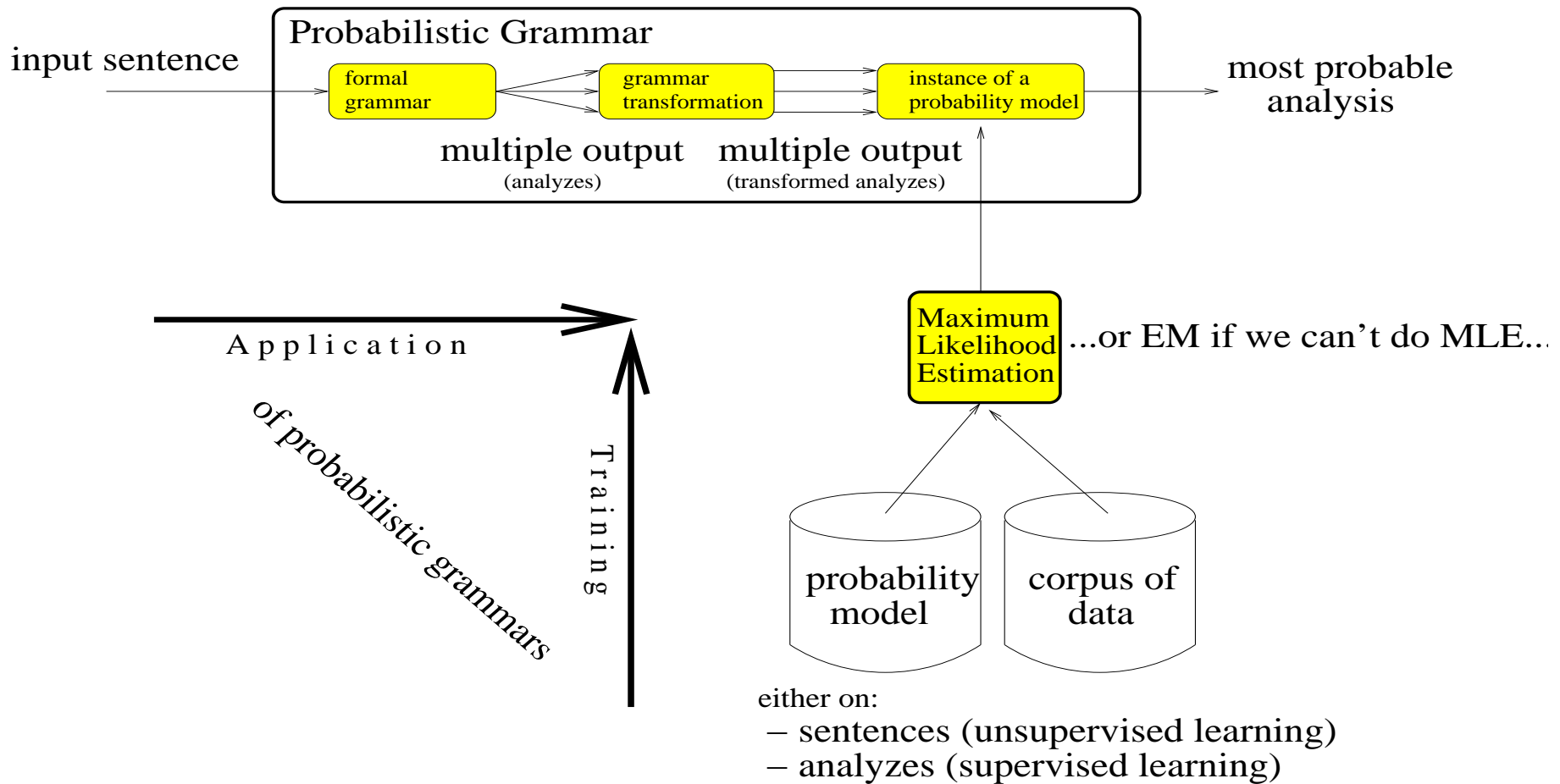
**Competence models:** In contrast to people, the linguistic view of language as a set does not care about problems caused by ambiguity/uncertainty. It

- cannot resolve multiple output
- cannot handle multiple input (resulting e.g. from noisy utterances)
- cannot express multiple levels of grammaticality

**Performance models** mimic people's language behavior and are specifically designed to resolve ambiguity. The majority of these models do the job by

- + **generalizing competence models.** Competence models serve simply as components of performance models.
- + **dealing naturally with uncertainty** concerning input/output and grammaticality, exploiting properly **Probability Theory and Statistics.**
- + incorporating the potential to model even extra-linguistic factors.

# Statistical NLP



## Past LIT-Research on NLP

---

- **Computational Linguistics, Probability Theory, Statistics, and Machine Learning** is our daily bread.
- **Probabilistic models for NLP:** We developed the world-best or at least state-of-the-art models for
  - Phonology (English, German) [CELEX-based probabilistic grammars]
  - Morphology (Hebrew)
  - Syntax (English, Dutch, German) [phrase-structure grammars]
  - Lexical Semantics (English, German) [semantically annotated lexica]

Beside this, we have a bit of experience in question-answering and machine-translation systems...

## Future LIT-Research on NLP

---

- **Deep Parsing:** Development of probabilistic grammars/parsers for several European languages (Dutch, English, German). Our main interest is to integrate in our current probabilistic parsers: morphology, predicate-argument structure, and semantically annotated lexica.
- **Learning Stochastic Tree Grammars from Treebanks:** STGs work on the basis of tree-fragment probabilities. Compared to other grammars (PCFGs)
  - + STGs are more expressive and achieve better empirical results
  - Current STG estimation has several flaws: bias and inconsistency
- **Creating NLP Tools for Question Processing**
  - + QA systems: Desired information is completely specified by the question
  - Unfortunately, current NLP tools are of limited use for question processing

## Download: Introduction to Statistical NLP

---

- The course material of two courses given at the *15th European Summer School in Logic, Language and Information*, Vienna, August 2003:
  - **Probabilistic Models of NLP** (sponsored by EACL)
  - **Probabilistic Parsing**

<http://staff.science.uva.nl/~simaan/ESSLLI03.html>

- A Tutorial on the Expectation-Maximization Algorithm Including Maximum-Likelihood Estimation and EM Training of Probabilistic Context-Free Grammars:

<http://staff.science.uva.nl/~prescher/papers/em/>