

# Building Pattern Dictionaries with Corpus Analysis

*Patrick Hanks*

Faculty of Informatics, Masaryk University, Brno

Riga: course introduction

**26 November, 2007**

# Themes of the course

- A lexical approach to language
  - How do people use words to make meanings?
  - What is meaning, anyway?
  - Regularities in human linguistic behaviour
  - Regularities in language as system
  - A “bottom-up” approach to language analysis

# Topics to be covered (1)

- Attempts to capture the lexicon:
- Dictionaries and thesauruses
  - Their role in the history of European culture
  - The Renaissance
  - The Enlightenment
  - 19th-century philology
  - 20th-century computer technology
    - The Internet, the Semantic Web

# Topics to be covered (2)

- Corpus linguistics and corpora
  - British National Corpus
  - 100,000,000 tokens
  - c.1,000,000 types
  - Of these, c. 500,000 are “noise” (typing errors, etc.)
  - c.300,000 occur only once (names, etc.)
  - Residue: c. 200,000 types
  - Distribution is Zipfian
- Corpus-based analysis
- Hands-on
  - WordNet
  - FrameNet (w.i.p.)
  - CPA in Brno (w.i.p.)

# Topics to be covered (3)

- Syntagmatic analysis
- Words, syntax, collocations
- Mapping meaning onto use

# Topics to be covered (4)

- What kind of the theory accounts for the data?
  - A Theory of Norms and Exploitations
    - Anyone (human or computer) learning a language has to acquire competence in two kinds of rule-governed behaviour:
    - The competence to use words normally and idiomatically
    - The competence to exploit norms creatively (ellipsis, metaphor, etc.)

# Topics to be covered (5)

- Literal meaning
- Metaphor
- Similes

# Topics to be covered (6)

- Writing dictionary entries