# Visualisation of word's collocations

## SemWeb seminar

Pavel Rychlý

Faculty of Informatics
Masaryk University
Brno, Czech Republic

December 1, 2008

# Words are elements of meaning

- a word without context – no meaning
- a word in different contexts – different meanings
- words in *similar* contexts – OK
- what is context?

# Words are elements of meaning

- a word without context – no meaning
- a word in different contexts – different meanings
- words in *similar* contexts – OK
- what is context?
  collocations

# Word Sketch

A corpus-derived one-page summary of a word's grammatical and collocational behaviour

- list of grammatical relations
- lists of collocations
- used mainly by lexicographers

# Programming interface

- Python/Perl/Java/Ruby API
- command line
- web service – XML, RDF

# Programming interface

- Python/Perl/Java/Ruby API
- command line
- web service – XML, RDF

Graphical visualisation of RDF data – visual browser.

# Comparison of frequencies of selected words

Words were selected in random from a random page in a Macmillan English Dictionary.

| | Susanne | BNC | BiWeC | BiWeC/BNC |
|---|---|---|---|---|
| Size [thousands] | 150 | 111,000 | 5,500,000 | 49× |
| heavy (adj) | 11 | 9,089 | 252,305 | 27× |
| hector (v) | | 37 | 956 | 25× |
| hedge (n) | | 1,525 | 19,526 | 12× |
| hedonism (n) | 1 | 63 | 1,757 | 27× |
| heebie-jeebies | | | 151 | |

# More data = better data

For different kinds of applications, different amounts of occurances (of words) are required.

concordances 20

word sketches 300

thesaurus 1500

Number of words with a frequency greater than ...

| | Susanne | BNC | BiWeC |
|---|---|---|---|
| 100 | 120 | 27,959 | 440,224 |
| 1,000 | 13 | 6,703 | 97,096 |
| 10,000 | 1 | 1,111 | 21,675 |