

# **Finding Semantically Related Words in Large Corpora**

**Pavel Smrž and Pavel Rychlý**

**Faculty of Informatics, Masaryk University**

**Botanická 68a, CZ-602 00 Brno, Czech Republic**

**E-mail: {smrz,pary}@fi.muni.cz**

## Motivation

- Assumption that semantically related words behave similarly
- Human users:
  - Dictionaries arranged according to topics
  - Foreign language learning
- Machine processing:
  - Selectional preferences on particular type of verb arguments
  - Word sense disambiguation
  - Machine translation
  - Information retrieval
  - Document classification

## How To Measure Semantic Relatedness

Words that are likely to co-occur within similar contexts

How to characterize the fact that words co-occur “frequently”:

- Statistical tests defining the probability of events
  - t-test (or t-score)
  - z-score
  - Pearson’s  $\chi^2$  (chi-square) test
- Likelihood ratio
- MI – (pointwise) mutual information – amount of information provided by the occurrence of one entity about the occurrence of the other one.

Exclusion of the low-frequency events

## Finding Collocations

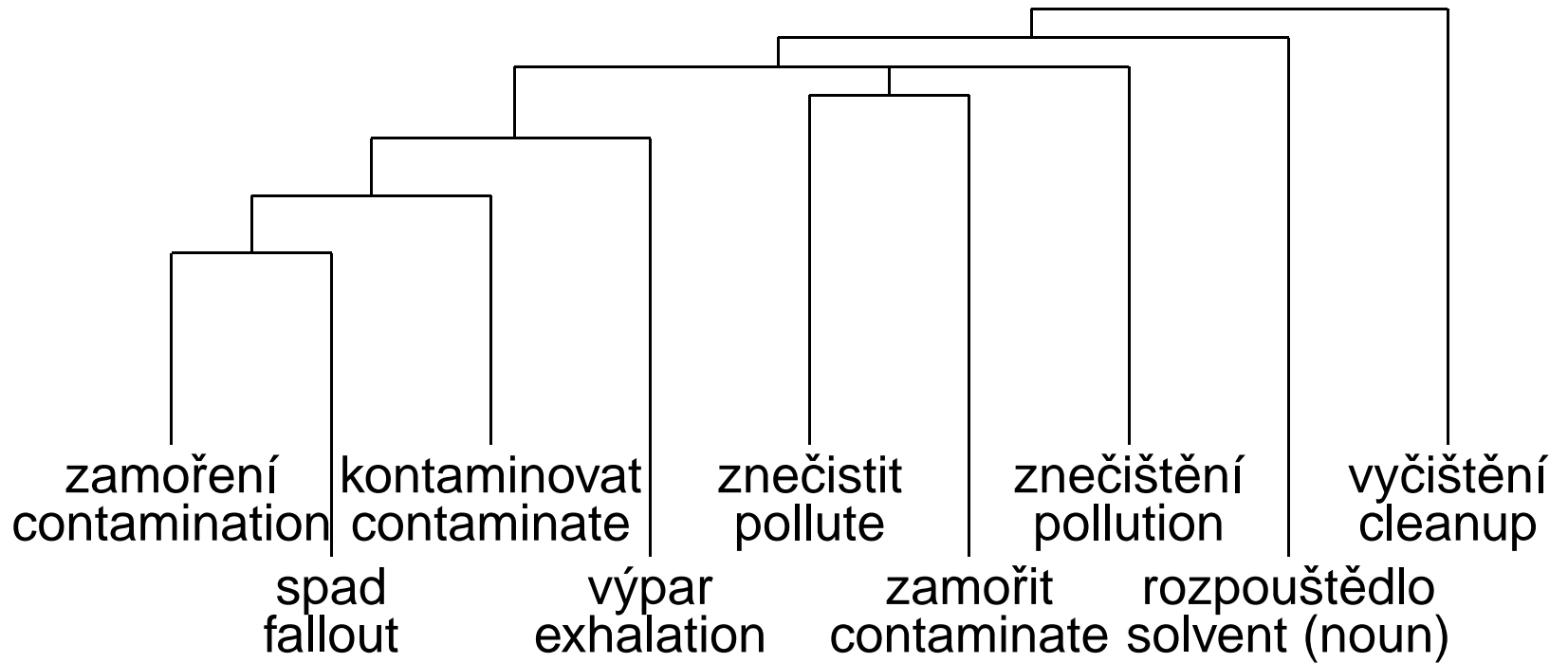
- Collocations of a given word are statements of the habitual or customary places of the word.
- Not contaminate clusters of semantically related words by collocations
- MI-score measure
- Successive formation of  $n + 1$ -word collocations from  $n$ -word collocations.
- Word sense disambiguation effect – soft clustering (a word can belong to one cluster as a part of one collocation and to other cluster as a part of another collocation)

## Characteristics of corpora used in experiments

# of	Czech	English
tokens	121,493,290	119,888,683
types	1,753,285	490,734
documents	329,977	4,124

# of	
different lemmata	1,071,364
lemmata with frequency $\geq 5$	218,956
lemmata with frequency $\geq 20$	95,636
bigrams with frequency $\geq 5$	25,009,524
lemmata in bigrams with frequency $\geq 5$	72,311

## An Example of Resulting Dendrogram



**“Path” from word *výpar/exhalation*  
to *vyčištění/cleanup***

výpar/exhalation

kontaminovat/contaminate (zamoření/contamination spad/fallout)

znečistit/pollute zamořit/contaminate

znečištění/pollution

rozpouštědlo/solvent(noun)

\_\_\_\_\_

vyčištění/cleanup