Finding multiwords of more than two words

The Sketch Engine Team

SKEW-3

Brno, March 2012

Multiwords

- Lexical items with spaces in

  (Western languages)

Two-word multiwords

- Church and Hanks 1989
  - Mutual information
  - A statistic that finds multiwords in a corpus
- Since
  - Other statistics
    - T-score, Log-likelihood, Dice, Fishers Exact Test
  - Evaluation
    - Krenn and Evert 2001, many others since
  - ***Better with grammar***
    - Wermter and Hahn 2006
- Problem solved

More than two words

- Problem 1: what to count
- Problem 2: statistics
- Attempts include
  - Dias 2002
  - Petrovic Snajder Basic 2010
- Not convincing
  - No *prima facie* validity to results
  - Stats only; no grammar

Responses

- Principle:
  - ***Word sketches work very well.*** Extend them
    - Commonest match
      - Vit Baisa
    - Multiword sketches
      - Vojta Kovar
    - COLLOCS directive
- Other
  - N-gram lists: soon
  - Multi-level tokenisation: at SKEW-2