



Evaluating word sketches and corpora

Adam Kilgarriff

Lexical Computing Ltd

Lexicography MasterClass Ltd

Universities of Leeds and Sussex



Word sketches

- Over 10 years
 - Since 1999
- Feedback
 - Good but anecdotal
- Formal evaluation



Goal

- Collocations dictionary
 - Model: Oxford Collocations Dictionary
 - Publication-quality
- Ask a lexicographer
 - For 42 headwords
 - For 20 best collocates per headwords
 - ***“should we include this collocation in a published dictionary?”***



Sample of headwords

- Nouns verbs adjectives, random
- **High (Top 3000)**
 - *N* space solution opinion mass corporation leader
 - *V* serve incorporate mix desire
 - *Adj* high detailed open academic
- **Mid (3000- 9999)**
 - *N* cattle repayment fundraising elder biologist sanitation
 - *V* grieve classify ascertain implant
 - *Adj* adjacent eldest prolific ill
- **Low (10,000- 30,000)**
 - *N* predicament adulterer bake bombshell candy shellfish
 - *V* slap outgrow plow traipse
 - *Adj* neoclassical votive adulterous expandable



Precision and recall

- We test precision
- Recall is harder
 - How do we find all the collocations that the system should have found?
 - Current work
 - 200 collocates per headword
 - Selected from
 - All the corpora we have
 - Various parameter settings
 - Plus just-in-time evaluation for 'new' collocates



Four languages, three families

- Dutch
 - ANW, 102m-word lexicographic corpus
- English
 - UKWaC, 1.5b web corpus
- Japanese
 - JpWaC, 400m web corpus
- Slovene
 - FidaPlus, 620m lexicographic corpus



User evaluation

- Evaluate whole system
 - Will it help with my task
 - Eg preparing a collocations dictionary
- Contrast: developer evaluation
 - Can I make the system better?
 - Evaluate each module separately
 - Current work



Components

- Grammar
- NLP tools
 - Segmenter, lemmatiser, POS-tagger
- Sketch grammar
- Statistics



Practicalities

- Interface
 - Good, Good-but
 - Merge to **good**
 - Maybe, Maybe-specialised, Bad
 - Merge to **bad**
- For each language
 - Two/three linguists/lexicographers
 - If they disagree
 - Don't use for computing performance



Results

- Dutch 66%
- English 71%
- Japanese 87%
- Slovene 71%



Corpus evaluation

- Collocation-finding
 - Typical corpus task
- Recall
- Hold all else constant
 - Statistic, NLP tools, grammar
 - ***Best results: best corpus***
 - (for collocation-finding)
- Pomikalek: de-duplication