COLLOCATION EXTRACTION

- *A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things* (Manning and Schütze 1999: p151)
- The difference between comprehensible and natural-sounding language usage
- CAUTION: "collocation" is an ill-defined term!!!

APPLICATIONS

- Information retrieval (query expansion, query segmentation)
- Language modelling in Speech processing (N-grams)
- Parsing (symbolic, statistical)
- Generation (symbolic, statistical)
- Word sense disambiguation ("one sense per collocation" principle)
- Lexicography (e.g. COBUILD)
- Terminology
- Text simplification
- Machine translation (multi-word translation pairs)

EXTRACTION PARADIGMS

- **Segment-based knowledge-driven/statistical extraction:** extract multi-segments as part of segmentation process
- **Word-based, knowledge-driven extraction:** extract word sequences of pre-defined type (e.g. nominal compounds)

  POS-based regular expressions, structural analysis

- **Word-based, statistical extraction:** extract statistically idiosyncratic word sequences

STATISTICAL TESTS USED IN COLLOCATION EXTRACTION

- Simple frequency: $f(XY)$
- Pointwise/specific mutual information: $\log \frac{P(x,y)}{P(x)P(y)}$
- Dice's coefficient: $\frac{2\, f(x,y)}{f(x)f(y)}$
- (Student's) t score
- (Pearson's) chi-square (χ^2)
- Z score
- Log likelihood
- Selectional association

$\vdots$

BIGRAM RESULTS FROM THE WSJ

Rank	Frequency	Mutual information	χ^2	t te
1	*of the*	*Quadi Doum*	*Posse Comitatus*	*of t*
2	*in the*	*Wrongful Discharge*	*LORIMAR TELEPICTURES*	*in t*
3	*to NUMB*	*Seh Jik*	*Petits Riens*	*to NU*
4	*for the*	*Noo Yawk*	*Wrongful Discharge*	*on t*
5	*to the*	*WESTDEUTSCHE LANDESBANK*	*Tupac Amaru*	*the cor*
6	*of NUMB*	*Naamloze Vennootschap*	*Sary Shagan*	*about N*
7	*on the*	*Caisses Regionales*	*Outlaw Biker*	*said*
8	*NUMB to*	*Centenaire Blanzy*	*GEMINI SOGETI*	*for t*
9	*that the*	*Guillen Landrau*	*Centenaire Blanzy*	*to*
10	*the company*	*Ea Matsekha*	*Smith-Corona Typewriters*	*a sh*
⋮				

WHY STATISTICS?

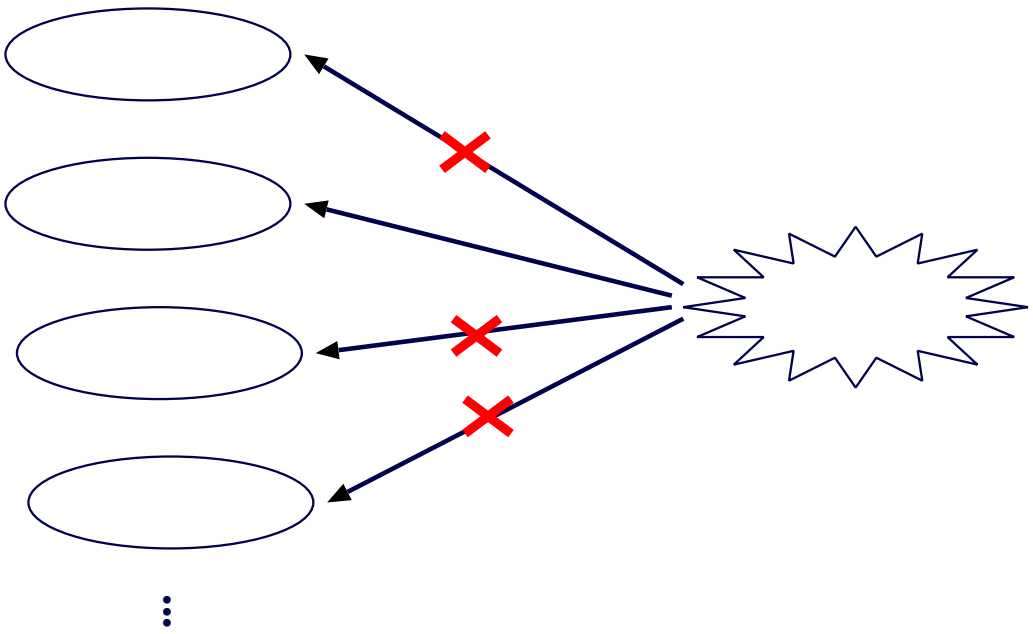
- Pick up on word combinations which occur with "significantly" high relative frequency when compared to the frequencies of the individual words (i.e. $f(x, y)$ as compared to $f(x)$ and $f(y)$)
- BUT WHY SO MANY #$%! STATISTICAL TESTS?
  - complications in evaluation (hard to say which is the "best" test, conflicting results from different researchers)
  - different corpora have different distributional idiosyncracies
  - different tests have different statistical idiosyncracies
- AND WHERE'S THE #$%! LINGUISTICS!!!
  - bear with me!

LINGUISTICS IN COLLOCATION EXTRACTION

- Apply statistical measures to (head) bigrams in a given dependency relation (e.g. subject-verb)
  - filters out stop words, produces "collocations" of pre-defined type for direct use in parsing, etc
- Look beyond contiguous bigrams, to bigrams occurring within a "collocational window" of fixed size (e.g. within 3-4 words of each other)
- Utilise linguistic qualities of collocations:
  - limited internal modifiability (applicable as a post-filter)
  - limited substitutability (contrast with anti-collocations, e.g. *(strong *powerful) coffee*)
  - non-compositional semantics

SUBSTITUTABILITY

Lexicalisation **Concept**



SUBSTITUTABILITY

- Most immediate means of testing substitutability via synonyms
- Synonyms accessible from thesauri, but word sense disambiguation is generally needed to isolate which synset(s) over which to apply substitution test
- Possibilities of getting at synonyms via distributional analysis (based on dependency pairs)???

CONCEPTUAL IDENTIFIABILITY

Lexicalisation **Concept**