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: \( \frac{2f(x,y)}{f(x)f(y)} \)
- (Student’s) \( t \) score
- (Pearson’s) chi-square \( (\chi^2) \)
- \( Z \) score
- Log likelihood
- Selectional association
### BIGRAM RESULTS FROM THE WSJ

<table>
<thead>
<tr>
<th>Rank</th>
<th>Frequency</th>
<th>Mutual information</th>
<th>( \chi^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>of the</td>
<td>Quadi Doum</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>in the</td>
<td>Wrongful Discharge</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>to NUMB</td>
<td>Seh Jik</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>for the</td>
<td>Noo Yawk</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>to the</td>
<td>WESTDEUTSCHE LANDESBANK</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>of NUMB</td>
<td>Naamloze Vennootschap</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>on the</td>
<td>Caisses Regionales</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>NUMB to</td>
<td>Centenaire Blanzy</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>that the</td>
<td>Guillen Landrau</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>the company</td>
<td>Ea Matsekha</td>
<td></td>
</tr>
</tbody>
</table>

**t test**

- LORIMAR TELEPICTURES
- Petits Riens
- Wrongful Discharge
- Tupac Amaru
- Sary Shagan
- Outlaw Biker
- GEMINI SOGETI
- Centenaire Blanzy
- Guillen Landrau
- Smith-Corona Typewriters
- Ea Matsekha
- Sary Shagan
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