Empirically Motivated Logical Representations in Lexical Semantics

Raquel Fernández & Galit Sanssoon

MoL Project, January 2010

Session 3
Plan for today

• Discussion of the following paper:

• Corpora as a source of linguistic data

• Statistical significance
Corpora as a source of linguistic data

- Data is central to linguistics.
- Corpus data is a source of empirical evidence that can complement other sources of information, such as acceptability judgements, experimental data, etc. as well as analytic thinking.
- A prototypical corpus is (Gries 2006):
  - a machine-readable collection of (spoken or written) language
  - representative with regard to a particular variety/register/genre
  - balanced with regard to a particular variety/register/genre
  - produced in a natural communicative setting
- Corpora can be raw or annotated with different kinds of information: phonological info, POS, semantic classes, syntactic trees, anaphoric relations, dialogue acts...

hospitality\text{\texttt{NN}} \text{\texttt{BEZ}} an\text{\texttt{AT}} excellent\text{\texttt{JJ}} virtue\text{\texttt{NN}}}
The web as corpus

• Some advantages:
  – large amounts of data, content is constantly added
  – inherently machine-readable
  – universally and freely available
  – diverse data: many topics/registers/genres, and multi-lingual

• Some disadvantages:
  – no control for native vs. non-native speakers
  – counts are often distorted:
    ▶ difficult to distinguish page counts from word counts
    ▶ multiple copies of identical documents
    ▶ cache of search engines distorts results
    ▶ non-permanence of data rules out replicability
  – limited searchability and no linguistic annotations
  – questionable representativity and balance; e.g. prominence of patterns particular to only the internet genre
Some available corpora

- 6 online corpora [http://corpus.byu.edu/](http://corpus.byu.edu/) including BNC, COCA, CHCA
- English Internet Corpus (110 million words, POS) [http://corpus.leeds.ac.uk/internet.html](http://corpus.leeds.ac.uk/internet.html)
- CHILDES: Child Language Data Exchange System [http://childes.psy.cmu.edu/](http://childes.psy.cmu.edu/)
Corpora as a source of linguistic data

• Amongs other things, corpus data can be used to inform our theoretical claims with quantitative evidence from language use, and to refute or validate a theoretical hypothesis.
• But only if quantitative data is evaluated carefully with appropriate tools from statistics.
• There is no point in evaluating quantitative data intuitively!
• When is a result statistically significant?

<table>
<thead>
<tr>
<th></th>
<th>Non-complements</th>
<th>Complements</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb: remember</td>
<td>295 (row perc.: 74%)</td>
<td>104</td>
<td>399</td>
</tr>
<tr>
<td>Verb: forget</td>
<td>131 (row perc.: 79%)</td>
<td>35</td>
<td>166</td>
</tr>
<tr>
<td>Totals</td>
<td>426</td>
<td>139</td>
<td>565</td>
</tr>
</tbody>
</table>

*Table 3: Postverbal elements in remember/forget clauses (after Tao 2003:80)*

The sentence immediately following these data is "[c]omparing the postverbal elements in the two verbs, we can see that the proportion of non-complements for forget is higher than remember: 79% vs. 74%" (Tao 2003:81). Just as with Aijmer's study, I do not wish to challenge
Some statistic resources


• R software: http://www.r-project.org/

• Online statistics calculator: http://faculty.vassar.edu/lowry/VassarStats.html
Statistical Significance (p-value)

- The statistical significance of a result is the probability that an observed relationship (e.g. between variables) or difference (e.g. between means) in a sample occurred simply by chance and hence doesn’t exist in the population.
- It tells us something about the degree to which the result is true (in the sense of being “representative of the population”).
- The p-value represents the probability of error that is involved in accepting our observed result as valid.
- In many areas of research, a p-value of .05 is considered the threshold statistical significance or acceptable error level.
- Typical p-values reported, in increasing level of significance: \( p \leq .05, \ p \leq .01, \ p \leq .001 \)
Statistical Significance tests

Parametric vs. non-parametric statistic tests:

• Parametric tests are more powerful and precise, but require variables that are normally distributed

• We can use a parametric test without knowing the type of distribution of our variables if the sample size is big enough (e.g. 100 or more observations).

• In linguistics, often variables are not normally distributed, or we do not have information about the shape of the distribution
  – If the sample size is small ($n < 100$), use non-parametric methods
  – If the sample size is big ($n > 100$), prefer parametric methods.

• We will see a couple of examples:
  – Relationships between two variables
  – Comparing central tendencies of two categories
Relationships between two variables

- If the two variables of interest are categorical (conjunctive/disjunctive, negative/positive) we can use the Pearson $\chi^2$ (chi-squared) statistics for testing the significance of the relationship between the two variables.
- The $\chi^2$ test computes the expected frequencies in a two-way table (i.e., frequencies that we would expect if there was no relationship between the variables).
- Significance increases as the numbers deviate further from the expected pattern.
- It requires that the expected frequencies are bigger than 5; if they are smaller Yates Correction can be applied.

Possible examples of variables to check:
- Conjunctive/disjunctive vs. positive/negative
- Relative/absolute vs. open/close scale
- Healthy/sick vs. ‘’P except’’/’’¬P except’’
Here is how to do a chi-squared test in R to check if there is a significant relation between the two variables in table on page 6:

```r
> verb.compl.matrix <- matrix(c(295, 131, 104, 35), ncol=2)
> attr(verb.compl.matrix, "dimnames") <- list(VERB=c("remember", "forget"),
   COMPLEMENTS=c("yes", "no"))
>
> verb.compl.matrix
   COMPLEMENTS
VERB   yes  no
   remember 295 104
   forget    131  35
>
> addmargins(verb.compl.matrix)
   COMPLEMENTS
VERB   yes  no Sum
   remember 295 104 399
   forget    131  35 166
   Sum        426 139 565
>
> verb.compl.matrix.test <- chisq.test(verb.compl.matrix, correct=F)
> verb.compl.matrix.test

Pearson's Chi-squared test

data:  verb.compl.matrix
X-squared = 1.5679, df = 1, p-value = 0.2105
```

\[ p = 0.21 \], hence there isn’t a statistically significant relation \((p > 0.05)\)
Comparing central tendencies of two categories

- The t-test is the most commonly used method to evaluate the differences in means between two groups.
- Use it if the variables are normally distributed or if the sample size is large.
- It is recommended to always report the standard, two-tailed t-test probability.
- We need a nominal independent variable that defines the grouping, and at least one numeric dependent variable.

<table>
<thead>
<tr>
<th>independent variable</th>
<th>dependent variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>conjunctive/disjunctive</td>
<td># ‘‘with respect to’’ (relative freq.)</td>
</tr>
<tr>
<td>negative/positive</td>
<td># nominalisations (relative freq.)</td>
</tr>
<tr>
<td>relative/absolute</td>
<td># ‘‘totally’’ (relative freq.)</td>
</tr>
<tr>
<td>old/modern</td>
<td># comparative forms (rel.freq.corpus size)</td>
</tr>
<tr>
<td>english/ esperanto</td>
<td>sentence length</td>
</tr>
</tbody>
</table>

- A non-parametric alternative: two-sample Wilcoxon test.
Some possible examples

How to encode your data:

| un-paired t-test (independent) or un-paired Wilcoxon test (paired=F) |
|---------------|-----------------|-------|
| INSTANCE | LANGUAGE | LENGTH |
| 1 | english | 9 |
| 2 | english | 12 |
| 33 | esperanto | 15 |
| 34 | esperanto | 7 |

| paired t-test (correlated) or paired Wilcoxon test (paired=T) |
|-----------------|---------------------|-----------------|
| ADJ | OLD | MODERN |
| tall | 0.02 | 0.4 | normalised by total # ‘‘tall’’ and corpus size |
| sick | 0.01 | 0.03 | normalised by total # ‘‘sick’’ and corpus size |

The relevant R functions are t.test() and wilcox.test().
You can also use the online statistics calculator:
http://faculty.vassar.edu/lowry/VassarStats.html