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:
  Stefan Gries (2009) What is corpus linguistics?, *Language and Linguistics Compass*, 3:1-1.
- 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<NN> is<BEZ> an<AT> excellent<JJ> virtue<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 distiguish 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/` including BNC, COCA, CHCA
- English Internet Corpus (110 million words, POS) `http://corpus.leeds.ac.uk/internet.html`
- Copora available with NLTK `http://www.nltk.org/`
- CHILDES: Child Language Data Exchange System `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?

	Non-complements	Complements	Totals
Verb: *remember*	295 (row perc.: 74%)	104	399
Verb: *forget*	131 (row perc.: 79%)	35	166
Totals	426	139	565

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

- StatSoft, Inc. (2010) Electronic Statistics Textbook. Tulsa, OK: StatSoft. `http://statsoft.com/textbook`
- 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 varibles 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

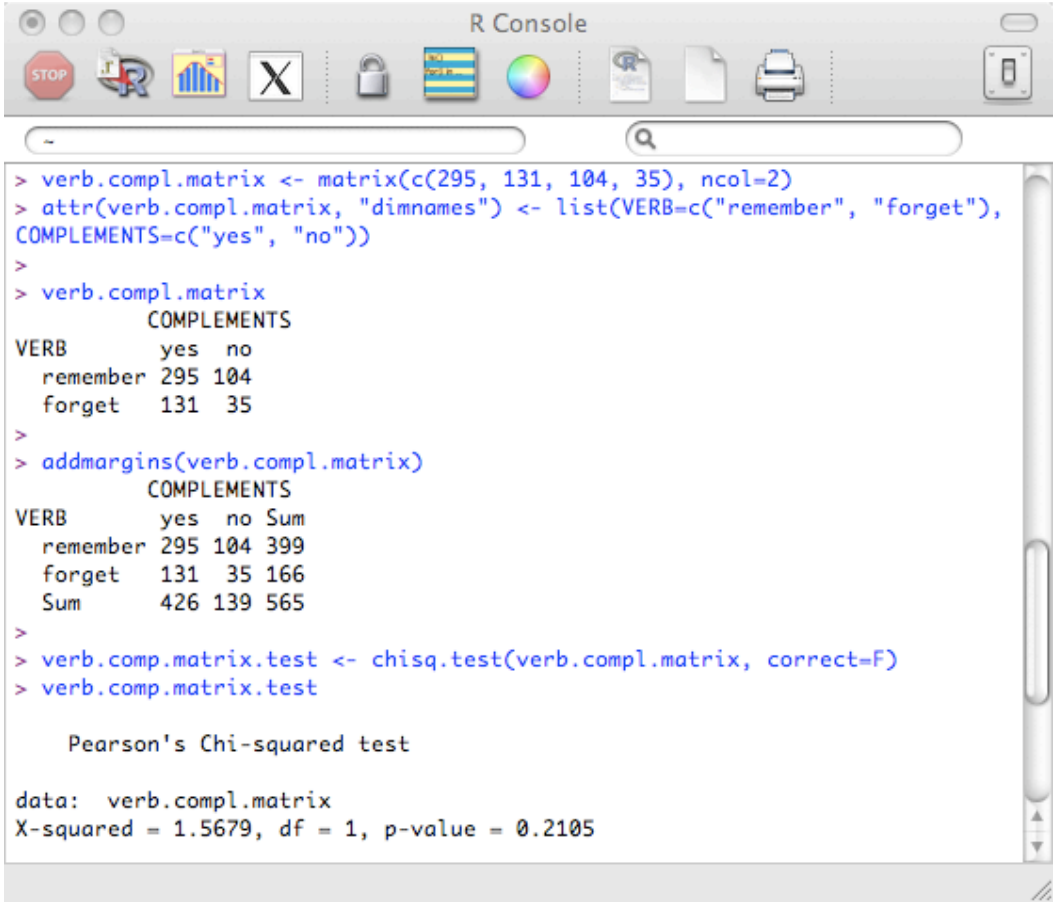
- If the two variables of interest are categorical (conjunctive/disjunctive, negative/positive) we can use the Pearson χ^2 (chi-squared) statistics for testing the significance of the relationship between the two variables.
- The χ^2 test computes the expected frequencies in a two-way table (i.e., frequencies 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''

χ^2 in R

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:



$p = 0.21$, 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 varibles 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.

independent variable	**dependent variable**
conjunctive/disjunctive	# ‘‘with respect to’’ (relative freq.)
negative/positive	# nominalisations (relative freq.)
relative/absolute	# ‘‘totally’’ (relative freq.)
old/modern	# comparative forms (rel.freq.corpus size)
english/esperanto	sentence length

- 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`