The structure of this workshop

PART 1: Intro

- Introducing ICE-GB and ICECUP
  - The ICE-GB corpus, its structure and analysis
  - ICECUP, queries and FTFs
- Introduction to statistics and experimental design
  - Why should we do it? • How do we do it? • What does it mean?

PART 2: Group work

- Lexical and grammatical examples on a small data set
  - Exercise 1: sociolinguistics ⇒ grammar
  - Exercise 2: grammar ⇒ grammar
- Discussion of statistics over this data
  - Testing for significance • Size of effect • Problems
- Presentations of group work – convince us!
  - Anticipating the devil’s advocate
  - Developing a programme of research

Introducing ICE-GB and ICECUP

The British Component of the International Corpus of English

- ♦ Sampling
  - • Spoken and written: 60% spoken • 500 × 2,000-word texts = 1Mw
- ♦ Analysis scheme
  - • Structural markup, tagging and parsing (based on Quirk *et al.* 1985)

The ICE Corpus Utility Program

- ♦ Software dedicated to exploring a parsed corpus
  - • Three levels of browsing: - overview - text - sentence
  - • Search by sociolinguistic variable, text string or FTF
- ♦ Fuzzy Tree Fragments (FTFs)
  - • An intuitive model-based grammatical query system
- ♦ Performing experiments with a parsed corpus
  - • Sufficiently expressive for a huge range of experiments
  - • Ask questions we could not consider before
  - • No programming required...
  - • ...but we still have to think...

Statistics and experimental design

- ♦ Why should we be interested in statistical argument?
  - A: To generalise evidence from a corpus to “Real Language”
- ♦ What is a scientific experiment?
  - **A test of a hypothesis**.
  - A hypothesis consists of an
    - independent variable (**IV**)
    - dependent variable (**DV**)
  - ie. Does the value of the IV have an effect on the value of the DV?
  - **Null hypothesis** = the prediction that there is no effect.
- ♦ An example
  - Q: Is “whom” used more often than “who” in written English?
  - **IV** = *genre* {spoken, written}, **DV** = *choice* {“whom”, “who”}
- ♦ Note the use of *relative frequencies*:
  - “whom” vs. “who” **given the choice**
  - = A move away from frequency per thousand words...
    - What is the likelihood that the speaker says “whom”?

Statistics and experimental design (II)

- Absolute vs. relative frequencies
  - **An absolute frequency** can tell you how common a word is in the corpus. But the reason that it is there might depend on many irrelevant factors.
  - **A relative frequency** focuses on variation where there is a choice. It tells you how often the speaker or writer chooses to use one word over another. It lets us focus on a specific type of **linguistic event**.
- Specificity vs. generality
  - By defining the linguistic event broadly or narrowly, experiments can be specific or general.
    - General experiments invite devil advocacy
    - Specific experiments risk the “so what?” factor
  - Linguistic argument should define
    - what to look for - and can you classify it?
    - how to relate it back to examples in the corpus
    - how the community debates the results
- Experiments must be defensible and reproducible

The one-slide experiment guide

- Choose IV and DV: does the IV predict the DV ?
- Construct a contingency table (IV × DV) – below
- Get data from the corpus using a series of queries
- Complete the table, including totals

		dependent variable			
		DV = x	**DV** = y	…	**TOTAL**
independent variable	**IV** = a	$a \land x$	$a \land y$		$a \land (x \lor y \lor ...)$
	IV = b	$b \land x$	$b \land y$		$b \land (x \lor y \lor ...)$
	…				
	TOTAL	$(a \lor b \lor ...) \land x$	$(a \lor b \lor ...) \land y$		$(a \lor b \lor ...) \land (x \lor y \lor ...)$
		observed			***expected***

- Compare *observed* with *expected* results using a statistical test
  - for example (above) do speakers positively choose x ?

Performing a statistical test

		DV		
		who	*whom*	**TOTAL**
IV	*spoken*	**150**	**50**	**200**
	written	**60**	**40**	**100**
	TOTAL	**210**	**90**	**300**
			***observed* O**	***expected* E**

- χ^2 (chi-square)
  - cf. observed vs expected distributions:
  - Simple, specific value of DV: one obs. column (e.g. *who*)
    - Observed **O** = specific value of DV
    - Expected **E** = total value of DV, scaled down
  - OR all values of DV: sum all columns
  - Formula:

$$\text{chi-square } \chi^2 = \sum \frac{(o-e)^2}{e} \text{ where } o \in \mathbf{O} \text{ and } e \in \mathbf{E}.$$

  - Test: is this greater than a threshold value χ^2_{crit} ?
- Critical values of χ^2 depend on
  - degrees of freedom $df = r\text{-}1$
    - or $(r\text{-}1) \times (c\text{-}1)$ where c = columns
  - probability of error
    - typically p = 0.05, 0.01

df	$p = 0.05$	$p = 0.01$
1	3.841	6.635
2	5.991	9.210
3	7.815	11.345
4	9.488	13.277
5	11.070	15.086

A worked example

♦ Is a preference for *whom* affected by text category?

Observed **O** = {50, 40}, scale factor SF = 90/300 = 0.3,
expected **E** = {200 × 0.3, 100 × 0.3} = {60, 30}.
Chi-square $\chi^2 = \Sigma(o\text{-}e)^2/e = 10^2/60 + 10^2/30 = \mathbf{5.000}$.
Chi-square critical value ($df = 1$, error level $p = 0.05$) = $\chi^2_{crit}(1, 0.05) = 3.841$.

- Since χ^2 > critical value, the result is significant
  - and the null hypothesis, *i.e.*, that *whom* does not correlate with variation of **text category**, is rejected = **YES**

♦ How big is the result?

- A quick measure is *percentage swing*:
  - swing(*dv*, *iv*) = *pr*(*dv* | *iv*) − *pr*(*dv*)

    swing(*whom*, **written**) = *pr*(*whom* | **written**) - *pr*(*whom*)
    = 40/100 - 90/300 = +0.1

♦ Significance and size are not the same thing:

- If you have enough data, small effects will be significant
- Significance means it is probably reproduced in “Real Language”

Exercise 1: sociolinguistics ⇒ grammar

- ♦ Examples
  - Does speaker gender, age, role... affect the choice of a construction?
- ♦ Issues
  - Have we specified the null hypothesis correctly?
  - Have we listed all possible outcomes?
  - Are we really dealing with the same linguistic choice?
  - Do we have enough different speakers?
- ♦ Method, using ICE-GB and ICECUP
  - Enumerate outcomes and construct table
  - Complete the table by:
    - Creating an FTF for each grammatical outcome
    - Performing FTF queries
    - Dragging and dropping sociolinguistic contexts to combine values
    - Calculating the TOTAL column
  - Perform χ^2 and measure size of effect
- ♦ Justify your results through examples in the corpus

Exercise 2: grammar ⇒ grammar

- Examples
  - Does the 'mood' of a clause predict its transitivity?
  - How does one element within a clause or phrase affect another?
- Issues
  - We must specify the *case* (eg. the clause or phrase)
  - We have to consider unmarked cases, eg. with absent features
  - Do cases *interact* with one another (eg. an NP in an NP)?
  - ⇒ Use FTFs to establish the proportion of cases that are strictly independent
  - ⇒ Multiply total χ^2 by this proportion
  - Are the IV and DV measuring different aspects of the same thing?
- Method, using ICE-GB and ICECUP
  - Enumerate outcomes and construct table
  - Complete the table by:
    - Performing an FTF for each different cell
    - Calculating TOTAL or 'missing value' columns and rows
  - Perform χ^2 and measure size of effect, and test for case interaction.
- Justify your results through examples in the corpus

Now for the hard part: convincing others

- The seduction of numbers
  - But what do they mean?
  - An experimental result may give you evidence for an argument, but...
    - Is the argument the right one?
    - Lay out your method so that your reader can repeat your experiment.
    - Show examples from the corpus to make your point.
- Advocating for the devil
  - Correlations don’t prove causes
    - There may be other explanations for the result, so anticipate your critics.
    - Is the result dependent on the particular grammar?
  - Are sentences correctly and completely analysed?
    - No, but how serious is the problem?
- And moving on:
  - What future work is suggested by your results?
  - Is it worth broadening or narrowing your set of cases?
  - Testing your hypotheses against other corpora