SPRING SCHOOL

Empirical methods in Usage-Based Linguistics

University of Lille 3, 13 & 14 May, 2013

WORKSHOP 1

Corpus linguistics workshop

Dagmar Divjak

d.divjak@sheffield.ac.uk
http://www.shef.ac.uk/russian/staff/profiles/divjakd

Objectives

The hands-on workshop is aimed at absolute beginners: it does not presuppose any knowledge or experience of corpora, but does expect participants to be well-versed in linguistic description.

Programme

Day 1:

Morning:

The theory/method/discipline: What is corpus linguistics? What kind of questions can it answer (and which questions can it not answer)?

Afternoon:

The research question: How to ask a question that can be answered by means of corpus data

Day 2:

Morning:

The data: How to extract relevant data from an existing corpus and annotate it so that it sheds light on your question

Afternoon:

The results: How to summarize the trends found in your annotated dataset and prepare your data for statistical analysis

Bring along

1. Laptop with text editor (such as TextPad http://textpad.com/) and spreadsheet software (such as Excel)
2. Corpus of interest or information on how to access corpus of interest if available on-line
3. Research question to pursue.

WORKSHOP 2

Corpus linguistics : Multivariate Statistics for Semantics and Pragmatics

Dylan Glynn
http://person2.sol.lu.se/DylanGlynn/

Objectives

1. **Introduce the method of usage-feature analysis.**
   This method allows traditional research questions
2. **Introduce the computer program R**
   This program is entirely free and it probably the most widely used application for doing statistics
3. **Introduce Multivariate Statistics for Categorical data**
   Multivariate statistics is needed in all empirical science. It is a vital tool for identifying complex patterns in data and then determining how representative those patterns are.

Outline

1. **Multivariate Categorical Statistics in Linguistics**
   Presentation of the statistical techniques covered in workshop, how they can be used in Linguistics research
2. **Usage-Feature analysis**
   After consultation with group, a sample of language will be analysed. The results of this analysis will form the basis of the data used in the statistical analysis
3. **Introduction to R**
   R will be installed and the data will be loaded so that the participants can familiarise themselves with R, we will prepare the data fro analysis
4. **Exploratory Statistics: Correspondence Analysis**
   Look for associations between things / identify what is similar and different in complex situation
5. **Exploratory Statistics: Cluster Analysis**
   Sort something relative to something(s) else / identify what is similar and different given *a constant of comparison*.
6. **Confirmatory Statistics: Binary Logistic Regression**
   Modelling the data, identifying significant predictors of a binary outcome, determining the descriptive strength and accuracy of the model
7. **Confirmatory Statistics: Loglinear Regression and Multinomial Logistic Regression**
   Modelling the data and predicting associations and complex outcomes

Pre-requisites

1. Participants are required to bring their own laptop; please download R before the workshop (http://www.r-project.org/); installing it will be done in the workshop
2. No previous experience with R or with statistics or with corpora is required
3. The workshop will be run in English, but the instructor speaks French and will use French when needed or desired.

WORKSHOP 3

Annotation and analysis of multimodal data using ELAN

Mark Tutton

mark.tutton@univ-nantes.fr
http://lling.univ-nantes.fr/index.php?option=com_profils&user=80&lang=fr

This workshop focuses on the analysis of gesture in the multimodal annotation program, ELAN. It is divided into three major parts, each of which is outlined below.

1. Gestures: forms and functions
What is gesture, and can gestures be broken down into different types? What are the different gesture phases? What particular elements of gestures should be analysed (for example, the hand used, the form of the gesture), and how should these be coded (i.e. by creating a set list of sub-categories?) This part of the workshop looks at *what* gestures you should analyse and *how* you should code them. Hence, it addresses methodological questions that need to be answered *before* analysing data in ELAN.

2. Using Elan to code gestural data
The second part of the workshop looks at how to use ELAN to code your gestural data. More specifically, it targets the following areas:

- Identifying the types of video files that can be imported into ELAN
- Creating a new Elan file
- Developing an annotation template that can be applied to multiple files
- Creating tiers and dependencies to analyse relevant gesture variables (for example, the hand(s) used, hand form)
- Devising a 'controlled vocabulary' that provides you with a handy list of set coding possibilities for the gesture variables being analysed
- Using frame-by-frame analysis to analyse gestures
- How to export data from ELAN
- Analysing data across multiple files

3. Using ELAN to analyse your own data
The third part of the workshop requires participants to apply the skills learned earlier on to the analysis of their own data. They will create their own ELAN template and analyse a portion of their video data in the program, allowing them to troubleshoot for any problems that may arise in their analysis.

Requirements
No previous experience in gesture analysis is required – just motivation to learn! Participants should bring with them a laptop and audiovisual data to be analysed (a five-minute clip that features (a) gesturing speaker(s) will be sufficient to practice on). ELAN accepts a range of data file types, including .avi, .mov and .mpg formats. Participants should download and install 'ELAN' **before** the workshop. They can do so by visiting the following website: http://tla.mpi.nl/tools/tla-tools/elan/

WORKSHOP 4

Transcription and analysis of oral data using CLAN

Outline of the workshop

1. CLAN, CHILDES and sharing data

Presentation of the CHILDES initiative: corpus, tools for child language and other linguistic data (TalkBank).

- Ground Rules for data sharing for all TalkBank databases
  - Using data
  - Providing data
- Corpus databases available
  - 37 different languages
- Downloading and visualising databases
  - Transcriptions
  - Media
- Tools: Clan, Phon

2. Creating your own oral language corpus

Transcribing with CLAN

- What does it mean to transcribe oral language data?
- The CHAT Transcription Format
  - Organisation of a transcript
  - Coding oral language
  - Supplementary annotations
- Linking sound or video with the transcription

3. Using a corpus for research with the CLAN commands

- What does it mean to have a corpus and use it for research?
- Using the clan commands on the transcriptions
  - Searching data
    - Navigation within the database
    - Searching multiples tiers
  - Creating a lexicon
  - Syntactic analysis and morphosyntactic coding
  - Modifying and coding the data
  - Interface with worksheet processors
  - Interface with textometric software and other corpus software

Requirements

Participants should bring a laptop. They can bring their own data, either sound (wave or mp3 format) or video (Quicktime format; this includes MOV and most MP4 formats – please check that your video can be displayed using Quicktime). CLAN and Quicktime should be installed before the workshop. Both are free, work on PC and MAC, and can be download at:
CLAN: http://childes.psy.cmu.edu/clan/
Quicktime: http://www.apple.com/quicktime/download/

WORKSHOP 5

Experimental methods of linguistic research

Efstathia Soroli

efstathia.soroli@univ-lille3.fr

http://stl.recherche.univ-lille3.fr/sitespersonnel

1. Overview

Current scientific research in all disciplines requires strong empirical support, data for validation (or refutation) and deep knowledge of new experimental techniques. Furthermore, data requires the solid foundation of a theory for interpretation. Although experimental support is a central element in order to answer fundamental questions of linguistic theory little is still used in linguistics. Gradually, linguists have come to realize the importance of corpus-based reasearch and experimental methods. In making the transition to objective/controlled empirical methods, linguists are confronted with problems linked to the framing of their scientific questions, the operationalization of their hypotheses, issues related to the experimental design, the implementation and the (qualitative and quantitative) analysis.

2. Aims

This workshop aims (a) to inform participants about the possibilities that psycholinguistic research offers and (b) to help researchers :

- to acquire the fundamental logic of experimental design
- to understand and use an experiment-builder tool (E-Prime) for design and setup.
- to implement and interpret the results of various kinds of experiments (categorization tasks, production, acceptability judgements / various measures: verbal, eye movements, reaction times, etc.).

3. Prerequisites/requirements

Participants must be able to understand English (English sessions), and have a good background in language theories and statistical methods. They are also required to bring a laptop computer (for the practice sessions on day 2).

4. Sessions and topics to be covered

PART I : Theory (day 1)

- Linguistics and empirical data : a brief historical overview (session 1)
- Psycholinguistics and basic empirical methods (session 2)
- Building a controlled *corpus-based* research protocol (session 3)
- Framing a scientific question, and setting up a question that can be answered through an experiment (session 4)

PART II : Practice (day 2)

- Fundamental concepts of experimental design (session 5)
- Experimental design using E-prime (session 6-7)
- Analysing and interpreting the results (session 8).