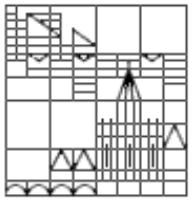
SPONSORED BY THE
Federal Ministry
of Education
and Research



Visual Analytics for Linguists

Miriam Butt & Chris Culy

ESSLII 2014, Introductory Course

Tübingen

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

PHILOSOPHISCHE
FAKULTÄT
Seminar für Sprachwissenschaft

Day 4 – Hands On

- Data Preparation
- Interacting with the software
- Possible Tasks

Hands On

- We can work with the following visual analytic tools:
  - The WALS Explorer (Mayer/Rohrdantz)
  - PhonMatrix (Mayer/Rohrdantz)
  - Cluster Visualization (Lamprecht et al.)
  - Diachronic Corpora via Glyphs (Sacha/Rohrdantz)
  - DoubleTreeJS (Culy)
- In the following we explain
  - how to work with them
  - the type of data needed

WALS Explorer

- The WALS Explorer can be accessed on-line:
  - http://th-mayer.de/wals/
- You cannot use your own data here.
- It is meant for an exploration of the World Atlas of Language Structure (http://wals.info).
- Task/Interaction Suggestion:
  - pick a phenomenon you are interested in
  - see what you can find out about it
  - think critically about the visualization and the interactive possibilities

PhonMatrix

- PhonMatrix can be accessed on-line:
  - http://paralleltext.info/phonmatrix/
- A demo set of data is provided (Finnish).
- You can also upload your own data.
  - It needs to be utf-8.
  - The file needs to contain one word per line.
  - All of the rest of the preprocessing that is necessary is done by PhonMatrix (you can set filters)
  - We have provided several other sets of data (most courtesy of Thomas Mayer).

PhonMatrix

Task/Interaction Suggestion:

- go through the demo first
- upload the Bambara file and see what you can find out about this language
- think critically about the visualization and the interactive possibilities

Cluster Visualization

- You are provided with a Java program in class.
- The software is still under development, so if you want to use it for purposes outside of this class, please contact Miriam.
- It should start by just clicking on it.
- A Readme file will guide you through what needs to be done.

Cluster Visualization

Data:

- The data needs to be in a txt file.
- The data points need to be separated by a symbol (e.g. “,”)
- We have provided sample data from our work on Urdu
  - Motion verbs courtesy of Annette Hautli – this works quite nicely and quickly.
  - Urdu N-V complex predicates – this file takes longer to load
- We have also provided some data based on Levin’s verb classes (levin-classes.txt).
- Feel free to add to this data as you wish.
- Some information on Levin’s verb classes is provided in levin-verbs-lawler.txt.

Cluster Visualization

Task/Interaction Suggestion:

- work with the Urdu motion verbs or the Levin verb classes file to get a feel for the visualization
- experiment with different numbers of clusters
- experiment with different visualizations of the data points (glyphs, star glyphs)
  - the Levin verb classes file contains three errors (three verbs contain wrong information)
  - see if you can spot that via the visualization
- think critically about the visualization and the interactive possibilities

Cluster Visualization

Task/Interaction Suggestion:

- enter your own data into a file by using the existing ones as a model
- you need to think about how to encode your data so that the system can compute with it
- Example: you may be interested in properties like whether a noun takes a certain case marker
  - Noun1: accusative, instrumental
  - Noun2: accusative, no instrumental
- this can be encoded as:

  NounType, accusative, instrumental

  Noun1, 1, 1

  Noun2, 1, 0

ClusterVis without verbs

- **Idea:** Sometimes we can use visualizations with data other than what they were originally designed for.

- **Example:** ClusterVis was designed to analyse properties of verbs, but it can be used to analyze any similarly encoded properties, no matter what those properties are for.

- **To try:** bierce-freq.txt and bierce-freq-2.txt contain information about letters that Ambrose Bierce wrote. The features include things like the number of pronouns (PP), and the number of words longer than 6 characters. Are there any clusters? If so, can you interpret them? (The data are from Chris who doesn‘t know the answer.)

Diachronic Visualization

- You are provided with a Java program in class.
- You are also provided with the entire IcePaHC corpus for Icelandic (under “data”).
- The software is still under development, so if you want to use it for purposes outside of this class, please contact Miriam.
- It should start by just clicking on it.
- A Readme file will guide you through what needs to be done.

Diachronic Visualization

- There are two pieces of software.
  - One is specialized for V1 in Icelandic
  - The other is looking at dative subjects in Icelandic.
- You can, in principle, feed your own data into this visualization, but many preprocessing/analytic steps are assumed.
- Task Suggestion
  - Work with the software as is.
  - Think critically about the visualization/interactive possibilities.
  - See if you can identify patterns from the visualization without necessarily knowing anything about Icelandic or the phenomenon (we could).

Exploring Corpora with DoubleTreeJS and KWICis

http://www.sfs.uni-tuebingen.de/~cculy/software/DoubleTreeJS/index.html
http://www.sfs.uni-tuebingen.de/~cculy/software/KWICis/index.html

- Explore the examples provided with the visualizations
  - What are advantages/disadvantages of each?
  - What would you like them to do that they can't?
- Get 2-3 books from Project Gutenberg, from different authors, or from same authors.
  - Use DoubleTreeJS, KWICis to compare them.
  - Data: Perhaps use js_corpus_tools (class site) for tokenization, tagging
  - There are different ways to do the comparisons. What are the advantages/ disadvantages?
- Try the data from the Bambara wikipedia.
  - What can you find, even without knowing the language?

Exploring Corpora with Structured Parallel Coordinates

http://www.eurac.edu/en/research/institutes/multilingualism/Projects/LInfoVis/StructuredParallelCoordinates.html

- Explore the examples provided
- Try your own data
  - Perhaps use js_corpus_tools for bigram frequencies
- Is this useful? If so, in what ways? If not, what would be better?

Exploring bigrams with MagicTable
Medium-advanced: basic programming

http://magic-table.googlecode.com/svn/trunk/magic-table/google_visualisation/example_1.html

- Use the MagicTable visualization from Google charts to look at bigram co-occurrences
  - cell row,column is for the bigram: row column

- Data:
  - maybe look at POS tag bigrams
  - have to count and normalize

Droplet visualization as Sankey Chart Advanced programming

Re-create the Droplet visualization using a Sankey diagram as a starting point

http://www.sfs.uni-tuebingen.de/~cculy/courses/ESSLLI2014/CuC_slides/reveal-based/vis_techniques.html#/graphs_networks

Word co-occurence network advanced

- Create some word co-occurrence data and visualize it using one or more network visualizations

- To consider:
  - What is the window size? Fixed? User-specified?
  - What about indicating the *strength* of the co-occurrences?
    - What is/are the relevant measures of strength?
    - How would the strength be encoded in the different visualizations?