LingVis: Motivation and Use Cases

Miriam Butt & Mennatallah El-Assady

LingVis: Visual Analytics for Linguistics
DGfS 2018 | 6.3.2018
Tutorial schedule

• 10:00 am – 11:30 am: Motivation and Use Cases
  • Examples of existing visualizations relevant for LingVis.

• 11:45 am – 1:00 pm: Visualization Theory
  • Goals and principles of Visual Analytics
  • Cognitive foundations

• 2:00 pm – 5:00 pm: Hands-on
  • Intro to existing visualization tools/frameworks
  • Work with existing visualizations
Before we start....
Personal Questions

• Who are we?
• Who are you?
  • Programming Background
  • What types of linguistic questions interest you?
  • What is the interest in LingVis?
Motivation and background

Slides based on the DGfS 2016 and Konvens 2016 Tutorials on LingVis and HistoBankVis 2017 talk.

Contributors: Miriam Butt, Oliver Deussen, Mennatallah El-Assady, Annette Hautli-Janisz, Wolfgang Jentner, Daniel Keim, Thomas Mayer, Frans Plank, Christian Rohrdantz, Dominik Sacha, Christin Schätzle, Rita Sevastjanova
Overall Goals:

• Integrate methods from **visual analytics** into domains of **linguistic inquiry**.
• Explore challenges based on the needs of **linguistic analysis** for **visualization methods**.
Visual Analytics

• Interactive, exploratory access to data
• Iterations of hypothesis formation and hypothesis testing
• Overview first – details on demand
  • Holistic picture
  • But can drill down to individual data points
  • Abstract representation of multifactorial, multidimensional data.
  • Good for understanding complex interactions in the data.
Existing Visualization: Syntax

- Syntactic Analysis with Lexical-Functional Grammar (LFG)
- [http://clarino.uib.no/iness/xle-web](http://clarino.uib.no/iness/xle-web) (Web Interface for LFG Grammars)
- Grammar developed at PARC (Palo Alto Research Center)
Existing Visualization: Semantics

DRS
(Discourse Representation Structure)

- Semantic Analysis with Discourse Representation Theory (Boxer)
- http://gmb.let.rug.nl/webdemo/demo.php (web interface for CCG/DRT)
- Grammar/Semantics developed by Johan Bos and colleagues (Groningen)
Existing Visualization: Phonetics

- aap=ne mara kis se thaa
Sample Visualizations
LingVis – Motivation

- Linguists are making more and more use of newly available technology to detect **distributional patterns** in language data.

- Ever increasing availability of **digital corpora** (synchronic and diachronic).

- Increasing interest in language output produced in **social media**.

- Ever better **query** and **search tools** (CQP, COSMAS, DWDS, ANNIS).

- **Programming languages** suitable for text processing, statistical analysis and visualization (e.g., Python, R).
Making Sense of Numbers

• Current linguistics often includes corpus work.
• Linguists try to determine patterns, interactions and usage preferences within a language but also across different languages.
• This work generates a lot of numbers (statistics).
• Numbers are difficult for humans to process.
• Solution: translate numbers into visual properties.

→ Human visual apparatus can process this easily.
Visualization in Linguistics

• Visualization is part and parcel of linguistics.

• However, very little advantage is taken of new visualization techniques.

• Newest ones borrowed from statistical analysis (bar charts, scatter plots, etc.), but often do not do justice to complexity of linguistic data.
  • Multifactorial/Multidimensional (including temporal)
  • Complex interactions
  • Different data types
  • Metadata

• Our Work: Identify and explore good use cases within linguistics
Interdisciplinary Collaboration: LingVis

Research Question

Data / Language Resources → Domain Expert
Interdisciplinary Collaboration: LingVis

Research Question

Data / Language Resources

Domain Expert

(task modelling, algorithmic processing, statistical analyses)

(Numerical) Features
Interdisciplinary Collaboration: LingVis

Research Question

Data / Language Resources

Domain Expert

(Numerical) Features

Visual Representation

task modelling, algorithmic processing, statistical analyses

mapping to visual variables, design, layout algorithms

investigate interactively
Example: Pixel-Based Visualizations

Two Use Cases
• N-V Complex Predicates
• Vowel Harmony
N-V Complex Predicates

- **N-V complex predicates** occur very frequently in Urdu.
- **Examples:** phone-do, memory-do, memory-become, resolution-do, resolution-be, ...
- **Problem:** would be nice if one knew which nouns were likely to co-occur with which verbs.
Example: N-V Complex Predicates in Urdu

- **Goal:** identify sequences of Noun+Verb for understanding complex predicate patterns
  - *phone-do, use-do, memory-come, begin-do/come*

- **Data:** 7.9 million word raw (unannotated) corpus of Urdu (BBC Urdu)
Example: Pixel Visualization

Statistical Data:

<table>
<thead>
<tr>
<th>ID</th>
<th>Noun</th>
<th>Rel. freq. with $kar$</th>
<th>Rel. freq. with $ho$</th>
<th>Rel. freq. with $hu$</th>
<th>Rel. freq. with $rakh$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>حاصل</td>
<td>0.771</td>
<td>0.222</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>اعلان</td>
<td>0.982</td>
<td>0.011</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>بات</td>
<td>0.853</td>
<td>0.147</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>4</td>
<td>شروع</td>
<td>0.530</td>
<td>0.384</td>
<td>0.086</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 2: Relative frequencies of co-occurrence of nouns with light verbs

do, be, become, put

$kar$, $ho$, $hu$, $rakh$

'achievement'
'announcement'
'talk'
'beginning'

Color Scale

<table>
<thead>
<tr>
<th>Value</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>Light Green</td>
</tr>
<tr>
<td>0.5</td>
<td>Green</td>
</tr>
<tr>
<td>0.75</td>
<td>Medium Blue</td>
</tr>
<tr>
<td>1.0</td>
<td>Blue</td>
</tr>
</tbody>
</table>
Pixel plus Cluster Visualization

- Performed k-means clustering combined with a pixel visualization.
- Advantages:
  - can inspect clusters visually and detect patterns
  - Outliers spotted easily (mostly errors – “kyA” is not a noun, it is a *wh*-word and was included by mistake).

![Image of pixel cluster visualization](image)
Example: Identifying N-V complex predicates in Hindi/IUrdu

Tool facilitates zooming and mousing over to see the underlying data set

Outliers/Errors are easily identified (Clustering Algorithm has applied)
Vowel Harmony (VH)

• **Phenomenon (simplified):** Vowels in affixes change according to vowels found in stems.

• **(Famous) Example:** Turkish

<table>
<thead>
<tr>
<th></th>
<th>Front</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unrounded</td>
<td>Rounded</td>
</tr>
<tr>
<td>High</td>
<td>i</td>
<td>ü</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>u</td>
</tr>
<tr>
<td>Low</td>
<td>e</td>
<td>ö</td>
</tr>
<tr>
<td></td>
<td>a</td>
<td>o</td>
</tr>
</tbody>
</table>

**Genitive suffix**
deniz-in, ev-in
tütün-ün, çöl-ün
denizler-in, ev-ler-in
tütünler-in, çöl-ler-in

**Genitive suffix with plural suffix**
kadın-in, adam-in
sabun-un, top-un
ekadınlar-in, adamlar-in
sabunlar-in, toplar-in
**Goal**: Try to determine automatically whether a given language contains patterns indicative of vowel harmony.

**Basic Computational Approach**:
- Use written corpus (caveat: only approximates actual phonology).
- **Count** which vowels succeed which other vowels in VC\(^+\)V sequences (within words — again an approximation)
- Through **statistical analysis** find out the association strength between vowels: normalized association strength value \(\phi\).
- **Results** show that Turkish and Hungarian, for example, pattern similarly. Languages like Spanish or German pattern differently.
## Results — Standard Methods: Can you detect a pattern?

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>i</th>
<th>u</th>
<th>o</th>
<th>ö</th>
<th>ü</th>
<th>i</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.266</td>
<td>0.427</td>
<td>-0.141</td>
<td>-0.060</td>
<td>0.019</td>
<td>-0.125</td>
<td>-0.261</td>
<td>-0.275</td>
</tr>
<tr>
<td>i</td>
<td>0.162</td>
<td>0.292</td>
<td>-0.107</td>
<td>0.077</td>
<td>-0.010</td>
<td>-0.075</td>
<td>-0.190</td>
<td>-0.191</td>
</tr>
<tr>
<td>u</td>
<td>0.129</td>
<td>-0.143</td>
<td>0.464</td>
<td>0.017</td>
<td>-0.003</td>
<td>-0.051</td>
<td>-0.138</td>
<td>-0.140</td>
</tr>
<tr>
<td>o</td>
<td>0.066</td>
<td>-0.112</td>
<td>0.434</td>
<td>-0.015</td>
<td>0.006</td>
<td>-0.045</td>
<td>-0.104</td>
<td>-0.111</td>
</tr>
<tr>
<td>ö</td>
<td>-0.107</td>
<td>-0.092</td>
<td>-0.052</td>
<td>-0.026</td>
<td>0.006</td>
<td>0.366</td>
<td>-0.091</td>
<td>0.164</td>
</tr>
<tr>
<td>ü</td>
<td>-0.120</td>
<td>-0.114</td>
<td>-0.059</td>
<td>0.014</td>
<td>-0.006</td>
<td>0.507</td>
<td>-0.112</td>
<td>0.134</td>
</tr>
<tr>
<td>i</td>
<td>-0.201</td>
<td>-0.224</td>
<td>-0.118</td>
<td>0.071</td>
<td>-0.004</td>
<td>-0.087</td>
<td>0.319</td>
<td>0.211</td>
</tr>
<tr>
<td>e</td>
<td>-0.256</td>
<td>-0.251</td>
<td>-0.132</td>
<td>-0.062</td>
<td>-0.010</td>
<td>-0.097</td>
<td>0.400</td>
<td>0.276</td>
</tr>
</tbody>
</table>

### Turkish

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>i</th>
<th>u</th>
<th>o</th>
<th>ö</th>
<th>ü</th>
<th>i</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.019</td>
<td>0.009</td>
<td>-0.061</td>
<td>-0.034</td>
<td>-0.008</td>
<td>-0.025</td>
<td>0.018</td>
<td>0.035</td>
</tr>
<tr>
<td>o</td>
<td>-0.023</td>
<td>-0.004</td>
<td>0.002</td>
<td>-0.013</td>
<td>-0.020</td>
<td>-0.013</td>
<td>-0.013</td>
<td>0.068</td>
</tr>
<tr>
<td>i</td>
<td>-0.069</td>
<td>-0.054</td>
<td>-0.050</td>
<td>-0.039</td>
<td>-0.036</td>
<td>-0.044</td>
<td>-0.003</td>
<td>0.133</td>
</tr>
<tr>
<td>ü</td>
<td>-0.067</td>
<td>-0.045</td>
<td>0.070</td>
<td>-0.028</td>
<td>-0.021</td>
<td>-0.033</td>
<td>-0.021</td>
<td>0.050</td>
</tr>
<tr>
<td>ö</td>
<td>-0.049</td>
<td>-0.032</td>
<td>0.049</td>
<td>-0.024</td>
<td>-0.013</td>
<td>-0.021</td>
<td>-0.013</td>
<td>0.036</td>
</tr>
<tr>
<td>ü</td>
<td>-0.067</td>
<td>-0.037</td>
<td>0.124</td>
<td>-0.033</td>
<td>-0.018</td>
<td>-0.028</td>
<td>-0.038</td>
<td>0.020</td>
</tr>
<tr>
<td>u</td>
<td>0.012</td>
<td>-0.018</td>
<td>0.019</td>
<td>0.046</td>
<td>-0.002</td>
<td>-0.013</td>
<td>0.004</td>
<td>-0.001</td>
</tr>
<tr>
<td>e</td>
<td>0.108</td>
<td>0.084</td>
<td>0.026</td>
<td>0.069</td>
<td>0.063</td>
<td>0.096</td>
<td>0.021</td>
<td>-0.195</td>
</tr>
</tbody>
</table>

### Spanish

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>i</th>
<th>o</th>
<th>e</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>-0.003</td>
<td>-0.075</td>
<td>0.094</td>
<td>-0.025</td>
<td>-0.018</td>
</tr>
<tr>
<td>i</td>
<td>-0.025</td>
<td>-0.004</td>
<td>0.064</td>
<td>-0.036</td>
<td>0.005</td>
</tr>
<tr>
<td>o</td>
<td>-0.028</td>
<td>-0.006</td>
<td>-0.075</td>
<td>0.098</td>
<td>0.026</td>
</tr>
<tr>
<td>e</td>
<td>-0.001</td>
<td>0.063</td>
<td>-0.073</td>
<td>0.016</td>
<td>0.021</td>
</tr>
<tr>
<td>u</td>
<td>0.077</td>
<td>0.038</td>
<td>-0.036</td>
<td>-0.057</td>
<td>-0.043</td>
</tr>
</tbody>
</table>

### German

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>o</th>
<th>i</th>
<th>u</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.339</td>
<td>0.263</td>
<td>0.070</td>
<td>-0.022</td>
<td>-0.081</td>
</tr>
<tr>
<td>o</td>
<td>0.239</td>
<td>0.099</td>
<td>0.041</td>
<td>-0.007</td>
<td>-0.052</td>
</tr>
<tr>
<td>i</td>
<td>0.132</td>
<td>0.038</td>
<td>0.015</td>
<td>-0.004</td>
<td>-0.017</td>
</tr>
<tr>
<td>ü</td>
<td>0.037</td>
<td>-0.026</td>
<td>0.008</td>
<td>-0.030</td>
<td>-0.017</td>
</tr>
<tr>
<td>ö</td>
<td>-0.093</td>
<td>-0.056</td>
<td>-0.022</td>
<td>-0.014</td>
<td>0.008</td>
</tr>
<tr>
<td>ü</td>
<td>0.152</td>
<td>-0.093</td>
<td>-0.037</td>
<td>0.001</td>
<td>0.065</td>
</tr>
<tr>
<td>e</td>
<td>-0.435</td>
<td>-0.241</td>
<td>-0.076</td>
<td>0.048</td>
<td>0.091</td>
</tr>
</tbody>
</table>

### Hungarian

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>o</th>
<th>u</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.339</td>
<td>0.263</td>
<td>0.070</td>
<td>-0.022</td>
</tr>
<tr>
<td>o</td>
<td>0.239</td>
<td>0.099</td>
<td>0.041</td>
<td>-0.007</td>
</tr>
<tr>
<td>i</td>
<td>0.132</td>
<td>0.038</td>
<td>0.015</td>
<td>-0.004</td>
</tr>
<tr>
<td>ü</td>
<td>0.037</td>
<td>-0.026</td>
<td>0.008</td>
<td>-0.030</td>
</tr>
<tr>
<td>ö</td>
<td>-0.093</td>
<td>-0.056</td>
<td>-0.022</td>
<td>-0.014</td>
</tr>
<tr>
<td>ü</td>
<td>0.152</td>
<td>-0.093</td>
<td>-0.037</td>
<td>0.001</td>
</tr>
<tr>
<td>e</td>
<td>-0.435</td>
<td>-0.241</td>
<td>-0.076</td>
<td>0.048</td>
</tr>
</tbody>
</table>
First Simplistic Visualization: Can you detect a pattern?

- Matrix visualization of association strengths between vowels (deviation from statistical expectation).
- Vowels are sorted alphabetically.
- More saturated colors show greater association strength.
- Blue is for more frequently than expected, red for less.
- The +/- are redundant encodings.
Sorted Visualization: Can you detect a pattern now?

Vowels sorted according to similarity (note: not a trivial process)
Can even see the type of Vowel Harmony involved.

Visualizing Vowel Harmony

Counting Vowel Successions in all Bible Types
Example: Finnish

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>Å</th>
<th>e</th>
<th>i</th>
<th>o</th>
<th>ø</th>
<th>u</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>3548</td>
<td>20</td>
<td>1940</td>
<td>1893</td>
<td>831</td>
<td>0</td>
<td>944</td>
<td>24</td>
</tr>
<tr>
<td>Å</td>
<td>35</td>
<td>944</td>
<td>896</td>
<td>820</td>
<td>10</td>
<td>138</td>
<td>33</td>
<td>266</td>
</tr>
<tr>
<td>e</td>
<td>1623</td>
<td>1144</td>
<td>1495</td>
<td>1608</td>
<td>419</td>
<td>56</td>
<td>197</td>
<td>187</td>
</tr>
<tr>
<td>i</td>
<td>1580</td>
<td>854</td>
<td>1314</td>
<td>1044</td>
<td>376</td>
<td>46</td>
<td>355</td>
<td>135</td>
</tr>
<tr>
<td>o</td>
<td>1384</td>
<td>7</td>
<td>1032</td>
<td>902</td>
<td>284</td>
<td>0</td>
<td>294</td>
<td>8</td>
</tr>
<tr>
<td>ø</td>
<td>7</td>
<td>125</td>
<td>54</td>
<td>39</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>u</td>
<td>1161</td>
<td>6</td>
<td>1085</td>
<td>850</td>
<td>315</td>
<td>1</td>
<td>517</td>
<td>8</td>
</tr>
<tr>
<td>y</td>
<td>30</td>
<td>656</td>
<td>368</td>
<td>368</td>
<td>35</td>
<td>75</td>
<td>4</td>
<td>251</td>
</tr>
</tbody>
</table>

Sorting done according to feature vectors of each of the rows.
Results – Sorted Visualization:

• Automatic Visual Analysis of vowel successions for 42 languages – sorted for effect strength.
In VH languages, crucially there are some vowels which never co-occur.

This can be seen via a calculation of succession probabilities.

Maori is not a VH language.
Even though Umlaut (raising of vowel in stem before high vowel in affix) is no longer a productive process in German, the Umlaut harmony pattern is still visible in the matrices.
Example: Droplet Visualizations

• Different Types of Visualizations can be used to look at the same data.
• Example: Droplets for Vowel Harmony
• This droplet technique was originally used for rendering geospatial information (an item moving from one place to the next).
kaşık-lar-ı-m-a
spoon-Pl-1SgPoss-Dat
‘my spoons’

ekedi-ler-ı-m-e
cat-Pl-1SgPoss-Dat
‘my cat’
Norwegian shows language change $a \rightarrow e$ in comparison to Swedish.
Another way to compare features across languages is via a sunburst visualization.

The following visualization combines sunburst with a link to the geographical location of the language.

The visual analysis is heavily interactive.
- One can feed in one’s own data.
- One can also use the WALS (World Atlas of Language Structures; http://wals.info).

Example: Sunburst and Maps
Sunburst and Maps for Language Families
Each circle segment represents one language, each ring the values of one feature across all languages.

World's Language Explorer

Bringing genealogy (left) and areal distributions (right) interactively into context: The values of a selected feature ring are color-coded on a map for exploration.
Figure 2.5: High-resolution screenshot showing automatically extracted features for languages from Papua New Guinea with leaves ordered to maximize (left) and minimize (right) the pairwise leaf similarity for neighbors.
Sorting and Pattern Discovery
WALS Explorer

• A version that is tailored to interact with WALS is available on-line
• http://www.th-mayer.de/wals/
Glyph Visualization for Diachronic Corpora

- Visualization of IcePaHC – Diachronic Corpus of Icelandic
  - Syntactically annotated in Penn Treebank style
  - 60 texts
  - 12\textsuperscript{th} century CE to 21\textsuperscript{st} century CE

- Two case studies so far (on-going)
  - V1 in Icelandic
  - Dative Subjects in Icelandic
V1 (Verb Initial or Verb First)

- Verb initial structures were common in matrix declaratives in Germanic.
- In German (and English) they mostly survive in narrative/joke contexts
  
  *Walked a man into a pub...*

Questions

- What determines the appearance of V1?
- How did this change over the history of Germanic?
Example: V1 in Icelandic

Visual Analytic Access to Data

- Glyph Visualization of likely factors
- Overview of all 60 texts at once
- Can drill down to individual data points interactively
- Shneiderman’s Mantra: Overview First, Show the Important – Details on Demand
HistoBankVis

- New Visualization for working with diachronic treebanks:
  - HistoBankVis
  - Accessible On-line
  - Interactive and Fast

subva.dbvis.de/histobankvis-v1.0

histobankvis.pdf

Example: Analyzing Political Argumentation (VisArgue)

- Public mediation on S21 (controversy around Stuttgart train station)
- Speakers are either Pro or Contra.
- Mediator is supposed to be neutral
- Data is annotated (rule based)

VisArgue Framework (http://visargue.inf.uni-konstanz.de)
Content Analysis

Lexical Episode Plots

User-Steerable Topic Modeling

Named-Entity Relationship Exploration

Speaker Dynamics

Discourse and Argumentation

Feature Alignment

Argumentation Glyphs

Discourse Maps

Topic Space Views

M. El-Assady, R. Sevastjanova, B. Gipp, D. Keim, and C. Collins.
NEREx: Named-Entity Relationship Exploration in Conversational Text Corpora.
Moderator

Good evening, I'm Lester Holt, anchor of ‘NBC Nightly News’.

I want to welcome you to the first presidential debate.

The participants tonight are Donald Trump, and Hillary Clinton.

This debate is sponsored by the “Committee on Presidential Debates”, a nonpartisan, nonprofit organization.

The commission drafted tonight’s format, and the rules have been agreed to by the campaigns.

The debate is divided into six segments, each 15 minutes long.

We'll explore three topic areas tonight: Where do you stand? America’s direction, and securing America.

At the start of each segment, I will ask the same leading question to both candidates, and they will each have up to two minutes to respond.

From that point until the end of the segment, we will have an open discussion.

The questions have been asked and have been approved by the Commission of the campaigns.

The audience here in the room has agreed to remain silent so that we can focus on what the candidates are saying.

I will note you as President. However, at this moment, as we welcome the candidates: Democrat nominee for President of the United States, Hillary Clinton, and Republican nominee for President of the United States, Donald Trump.

Let us welcome

APPLAUSE

Clinton

How are you, Donald?

APPLAUSE

Moderator

Good to see you.

APPLAUSE Well, I do not expect to cover all the issues of this campaign tonight, but I remind everyone, there are two more presidential debates scheduled.

We are going to focus on the issues of the candidates that you tell us are most important, and we are going to place for specifics.

I am honored to have this role, but this evening belongs to the candidates and just as important, to the American people.

Candidates, we know, and we know you anticipate your policies and your platforms, as well as your strengths and your values.

So, let's begin, we are going to open segment "Economic Prosperity." And continual to debate you.

There are two economic realities in America today.

First, a record six straight years of growth, and second, people who have increased their headcount after years of stagnation.

However, we know that millions are struggling, and nearly half of Americans are living paycheck to paycheck.

Beginning with you, Secretary Clinton, why are you a better choice than your opponent to create the kind of growth that will put more money into the pockets of American workers?

Clinton

Well, thank you, Lester, and thank you, NBC Nightly News for hosting us.

The central question in this election is really what kind of country we want to be and what kind of future we will build together.

Today is my granddaughters' second birthday, so I think about this a lot.

First, we have to build an economy that works for everyone, and just as those at the top.

That means we need more good jobs, good wages, with rising incomes.

Invest in your future.

That means good in infrastructure, in advanced manufacturing, innovation and technology, clean, renewable energy, and small business, because most of the new jobs will come from small business.

We also have to make the economy better. That means ending the national minimum wage and also guarantee, fully, equal pay for women's work.

Clinton

I have been a job creation campaigner for 40 years.

If you help grow the jobs, you should be able to share in them, and not just the successes at the top.

And I want to do more to help people who are struggling to balance family and work.

I've heard from so many of you about the choices you face and the stresses that you're under.
Moderator
Good evening, I am Kotobai Reference in Hempstead, New York.
I am Leon Panetta, anchor of "NBC Nightly News."
I want to welcome you to the first presidential debate.
The participants tonight are Donald Trump and Hillary Clinton.
This debate is sponsored by the Commission on Presidential Debates, a nonpartisan organization.
This format is drafted tonight’s format, and the rules have been agreed to by both campaigns.
The debate is divided into six segments, each 15 minutes long.
We’ll explore three topical areas tonight: Achieving prosperity, America’s direction, and securing America.
Each of the six segments, 10 minutes each, will be allotted to two candidates, and they will each have up to two minutes to respond.
From that point until the end of the segment, we’ll have an open discussion.
The candidates will not be called on a specific order or time, as long as we do not exceed the framework of the campaign.
The audience here and those who are watching will have heard the candidates.
The candidates will have agreed to remain silent so that we can focus on what the candidates are saying.
I will introduce each candidate, however, at this moment, as we welcome the candidates: Democratic nominee for president of the United States, Hillary Clinton, and Republican nominee for president of the United States, Donald Trump.

APPLAUSE

Clinton
How are you, don’t?

APPLAUSE

Moderator
Dear Mr. Trump, I do not expect to hear that all the issues of the campaign tonight, but I remind everyone, there are few more presidential debates scheduled.
We are going to focus on many of the issues that voters tell us are most important, and we are going to focus on specifics.
I am honored to have this role, but this evening belongs to the candidates and their importance to the American people.
Candidates, you each have a chance to tell us what you think about national security and your future, as well as your states and your values.
So, let’s begin, we are both opening remarks, “Economic Prosperity.”
And I think that is a good.
There are two economic realities in America today.
Thanks to a record six straight years of growth, and personal incomes show economic gains have increased at the rate of 8 percent after years of stagnation.
However, one out of every nine Americans is living paycheck to paycheck.
Beginning with you, Secretary Clinton, why are you a better choice than your opponent to create the kinds of jobs that will put more money into the pockets of American workers?

Clinton
Well, thank you, don’t, and thank you, let me begin.
The central question in this election is what kind of country we want to be and what kind of future we’ll build together.
Today is my granddaughter’s second birthday, so I think about this.
First, we have to focus on economic growth for everyone, not just those at the top.
That means we need more good, good jobs, with rising incomes.
It’s what to invest in you.
I want us to invest in our future.
That means, for instance, infrastructure, advanced manufacturing, innovation and technology, clean, renewable energy, and small business, because most of the new jobs will come from small businesses.
We also have to make the Economy work.
That means raising the national minimum wage and stop pregnancies, really, equal pay for women’s work.

Clinton
I also want to see more transparency in profit-sharing.
If you help them do that, you should be too, and we’re not, not just the executives at the top.
And if we want to do more to support people who are struggling to balance family and work.
I’ve heard from so many of you about the stress you face and the issues that you’re under.

APPLAUSE
- Person
- Geo-Location
- Organization
- Date-Time
- Measuring Unit
- Measure
- Context-Keyword
- Politeness-Indicator
- Positive-Emotion Indicator
- Negative-Emotion Indicator
Close-Reading

Moderator
Good evening, I am Moderator, host of NBC Nightly News. I want to welcome you to the first presidential debate. The participants tonight are Donald Trump and Hillary Clinton. This debate is sponsored by the Commission on Presidential Debates, based on a format developed by the campaigns. The debate will be held at Hofstra University in Hempstead, New York.

Moderator
We'll explore three topics tonight: economic prosperity; America's direction; and the role of government. Each candidate will have a chance to respond to the others' questions, and the time will be clearly marked. The audience is here to represent the opinions of the American people, and all audience members are asked to remain silent. The moderator will be in charge of the discussion.

Close-Reading

Audience
APPLAUSE

Clinton
How are you, Donald?

APPLAUSE

Moderator
Good evening. I'd like to thank all the panelists for coming tonight. We're moving on to the next topic: economic prosperity. The audience is going to get a chance to ask questions, and each candidate will have a chance to respond.

Moderator
We'll now begin our discussion on economic prosperity. I'd like to start with a question from the audience. What's your plan for economic prosperity?
Named-Entity Relationship

Grammar

Document Structure

Works well for **highly-edited texts** but not for **verbatim conversation transcripts** due to ungrammatical sentences.
Named-Entity Relationship

Works well for **highly-edited texts** but not for **verbatim conversation transcripts** due to ungrammatical sentences.
Named-Entity Relationship

Works well for **highly-edited texts** but not for **verbatim conversation transcripts** due to ungrammatical sentences.

Works well for **text summarization** but constructs **too broad** relations.
“I believe that we can cut taxes by an additional $5 trillion.”

Entity-Pairs: cut taxes, cut $5 trillion, taxes $5 trillion
Distance-Restricted Entity-Relationship Model

maxDist = 5
Use Case

2016 US Presidential Debates
Overall Exploration of the Dataset
Topic: Taxes
Content Analysis

Discourse and Argumentation

Argumentation Glyphs

Speaker Dynamics

Lexical Episode Plots

User-Steerable Topic Modeling

Named-Entity Relationship Exploration

Feature Alignment

Argumentation Glyphs

- **Topic**
- **Speaker**

**Common Ground**
1. Common Ground
2. Assurance
3. Consensus Willing

**Assurance**
4. Immutable Constraint

**Consensus Willing**
5. Minimal Consensus

**Immutable Constraint**
6. Regret

**Reason**
7. Actuality

**Conclusion**
8. Regret Accusation
Argumentation Glyphs

Dr. Heiner Geißler

[Velan, Dank]

[Also die Sache haben wir übrigens gestern Abend miteinander berechtigt nicht und es war klar, dass der Gegenstand einer Schlichtung kann sich sehr massiv mit allen Ergänzungen/Änderungen/Verordnungen, die von 2002 bis 2010 dann entwickelt worden sind, das ja rein rätselhaft ist nicht wahr.

Wenn das in der Öffentlichkeit statt dann hat ja jeder die Möglichkeit, dass die Informationen herauszusuchen und stellt die dann der ob des der aktuelle Stand wäre.

Es ist so ein dichtes Buch.

Wir können das in der Schlichtung einfach so nicht behandeln.

Das haben wir gestern Abend miteinander berechtigt und haben gezogen die Dinge die uns interessiert, die fokalen, die wir konzentrieren wir nicht wahr auf die Bitte der Herr. Conradi formulierte hat nämlich, wie verhält es sich mit dem Tippfehler mit Euro und D_Mark.

So welt Frau. Dahlbender.]
Argumentation Glyphs – Clustered View

Heiner Geißler

Volker Kefer
Micro-Linguistic Discourse Features

Dimensions of Deliberative Communication

1. **Participation** (equal and inclusive participation)
2. **Atmosphere & Respect** (equal respect)
3. **Argumentation & Justification** (reason-giving)
4. **Accommodation** (consensus-seeking)
Discourse Maps

A. Reason
B. Conclusion

1. Common Ground
2. Assurance
3. Consensus Willing
4. Immutable Constraint
5. Minimal Consensus
6. Regret
7. Actuality
8. Regret Accusation
Discourse Maps

4 Dimensions
→ Quadrants

19 Sub-Dimensions
→ Rows

53 Measures / Features
→ Rectangles
Types of measures

- **Binary** (0/1, e.g. “reason” or not)

- **Numerical, continuous** (0→1, e.g. epistemic value)

- **Numerical, bipolar** (-1→1, e.g. emotion=“negative/positive”)

Frequency of the measure

less frequent  more frequent
Average utterance length

shorter  longer

Speakers and speaker parties

PRO 1  CONTRA 1  NEUTRAL 1  EXPERT 1

PRO 2  CONTRA 2  NEUTRAL 2  EXPERT 2
Discourse Maps

Accommodation

Participation

Atmosphere & Respect

Argumentation & Justification
**Discourse Maps**

Atmosphere & Respect

Argumentation & Justification

Accommodation

Participation

Emotion

Average sentence complexity
**Argumentation**

**Neutral:** Comparatively little Argumentation

**Experts:** Justifications/Reasons

**Pro & Contra:** Conclusions and Justifications

→ Deliberative Argumentation

Measures shown:
1. Conclusion
2. Justification
Rejection of "Common" Ground, Refusal to provide information

Establishment of Common Ground

Pro, Contra & Neutral:

Establishment of Common Ground

Pro:

Rejection of "Common" Ground,
Refusal to provide information

→ controversial discussions
Glyphs per Group Type and Topics (Topic-Modelling)
Can aggregate or look at individually
Who was interrupting the debates?

First Debate

- Trump (Position: REPUBLICAN)
  - Utterances: 129
  - Average words: 148
- Moderator (Position: NEUTRAL_M)
  - Utterances: 97
  - Average words: 43
- Clinton (Position: DEMOCRAT)
  - Utterances: 94
  - Average words: 146

Second Debate

- Moderator (Position: NEUTRAL_M)
  - Utterances: 114
  - Average words: 49
- Trump (Position: REPUBLICAN)
  - Utterances: 88
  - Average words: 187
- Clinton (Position: DEMOCRAT)
  - Utterances: 57
  - Average words: 240
Summary of all three debates

- High amount of interruptions
- Negative sentiment
- Multiple topic shifts towards “previous” topics
- High amount of filler words
- Low lexical diversity
Visualization of spoken language

• So far we have been working with textual data. However, one can also work with spoken data.

• For Visual Analytics, all one needs is to have features (or vectors) that can be computed with

Motivation

Experiments

Recording

Analysis

- Manual Inspection
- Statistics
- Functional Data Analysis
- EXPLORATION

Recording, Landmarks, Utterance, Pitch-Vectors

Speaker Information
Our Approach - Workflow

A. Data Input
B. Configuration
C. Training
D. Interactive Visualization
Configuration and Preprocessing

Essential: Comparable Vectors & Distance Function

RAW-Pitches

Processed Pitches
Iterative SOM Training and Interaction

**Data Exploration**
- SOM learning is fast
- user can switch among different perspectives on the data
- user can interactively delete or pin cells
- and retrain and re-explore

**Interactions**
- A: Configuration
- B: Cell Interactions
- C: Filtering
- D: Re-Training
Interactive Visualizations

- Grid – Cells/Centroids
- History/Distances

- Word Clouds

- Bar Charts

- Colored Cells

- Cell Layout

- Centroid & Vectors

- Attribute (Value) - Heatmaps
Use Case

“Sumimasen”

“Entschuldigung”

Experiment - Data

Pre-Processing

FDA

Native Speakers vs. Learners

Are there Differences?

Tasks
Use Case – Two Separate SOMs are Trained for Japanese and German Speakers

Speakers pronounced "sorry/excuse me" in ever more exasperating circumstances

• Japanese natives do not vary the pitch contour (red SOM - B)
• German learners do vary the pitch (blue SOM - C)
Identifying Optimal Visualizations

- Understanding which visualizations are optimal is not trivial.
- Are the individual dimensions (color, shape, direction, size, etc.) usefully meaningful?
- Does the visualization allow for at-a-glance understanding, or does it confuse the user?

- This also depends on the user's background
  - What is the user used to looking at?
  - How is the user used to understanding the data?
  - How is the user used to interacting with systems?

- Currently evaluations are mainly performed via user studies in Visual Analytics.
- Project (SFB/TRR 161): Establish evaluation metrics.
Distorted Map according to number of languages spoken in area.

Visualization only as good as your data – India massively underrepresented
Outlook

• Further exploration of possibilities offered by Visual Analytics
  • The systems illustrated here are recent
  • Interactive exploratory linguistic analysis is on-going
  • Systems are being fine-tuned

• Workflow
  • Use cases for Digital Humanities / eHumanities are being developed
  • *Infrastructure* Platforms (mix and match the available tools)

• Measuring Success
  • Development of *Evaluation Metrics* for LingVis.
  • Use cases, work flow and result comparison.
What interests Visualizers?

- Need interesting interactions
- Multiple dimensions
- Time depth
- Cross-modular interactions.
- Not just coloring in bits of text that are of interest for linguists.
Summary

• Example of different kinds of visualizations
• These visualizations allow a new approach to linguistic data
• Flexible, interactive, make use of the highly skilled human perceptual system
• More examples to follow.

Now first some design basics!
THANK YOU!

Questions?