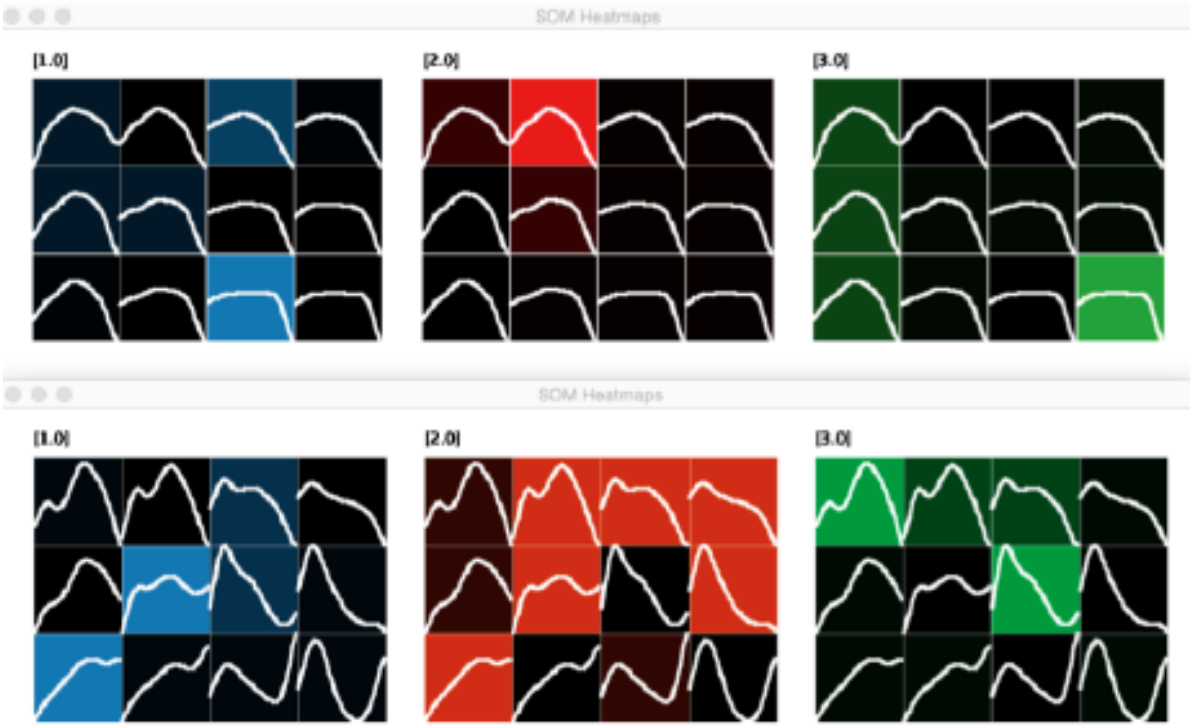
DFG Deutsche Forschungsgemeinschaft

SPONSORED BY THE

Federal Ministry of Education and Research

Universität Konstanz

Visualizing Linguistic Structure (LingVis)



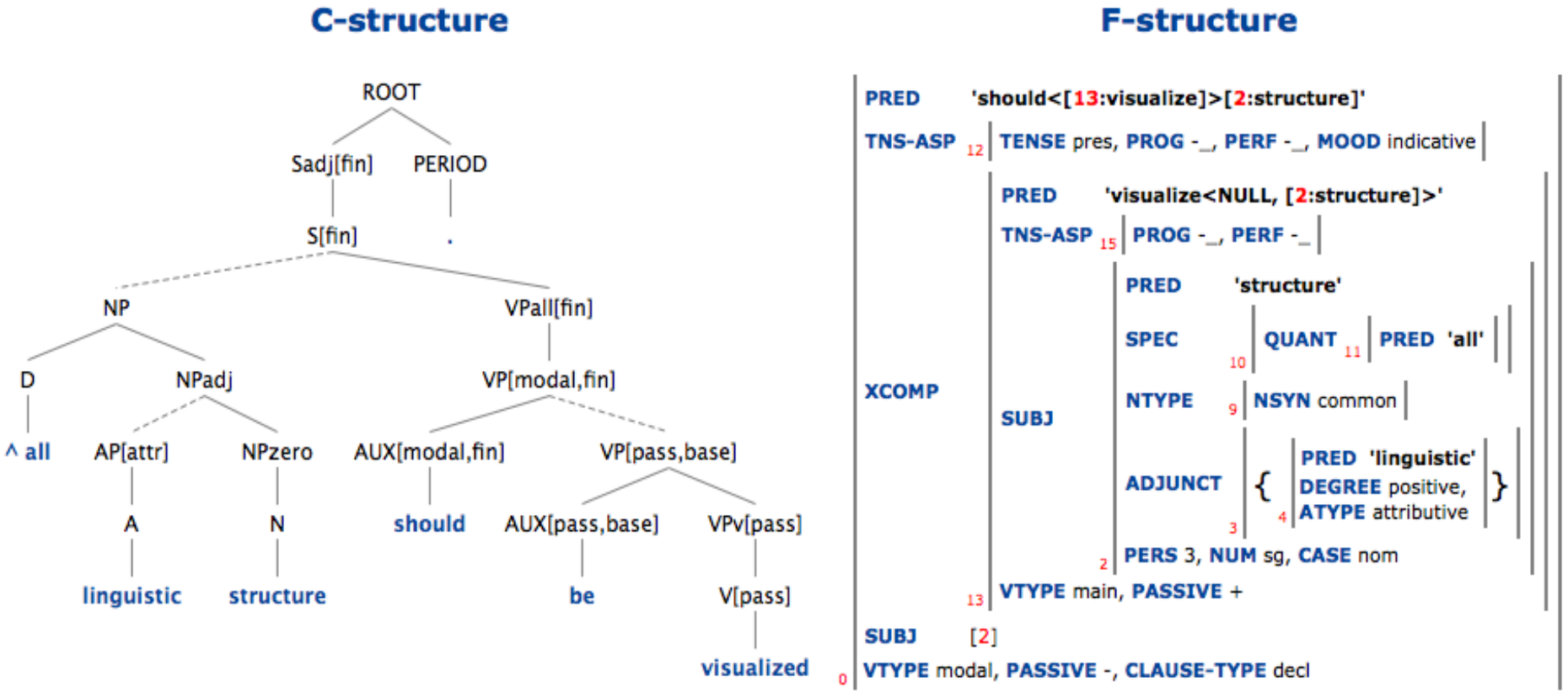
Miriam Butt

San Jose, February 15, 2015

AAAS Annual Meeting

Symposium: Visualizing Verbal Culture: Seeing Language Diversity

Standard Visualization: Syntax

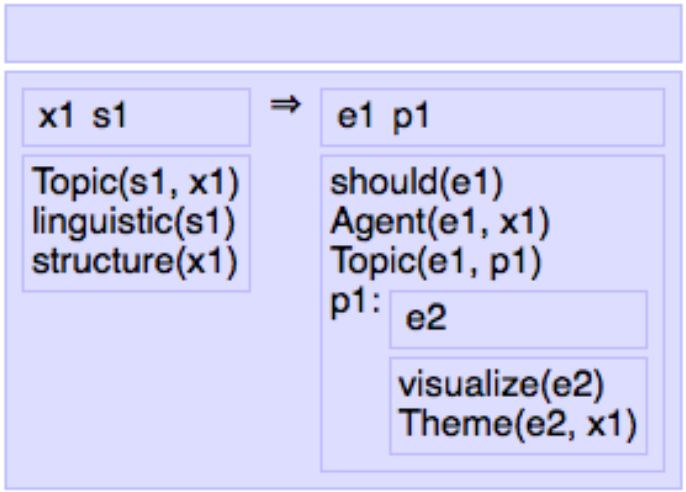


Syntactic Analysis with Lexical-Functional Grammar (LFG)
http://iness.uib.no/iness/xle-web (Web Interface for LFG Grammars)
Grammar developed at PARC

Standard 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)

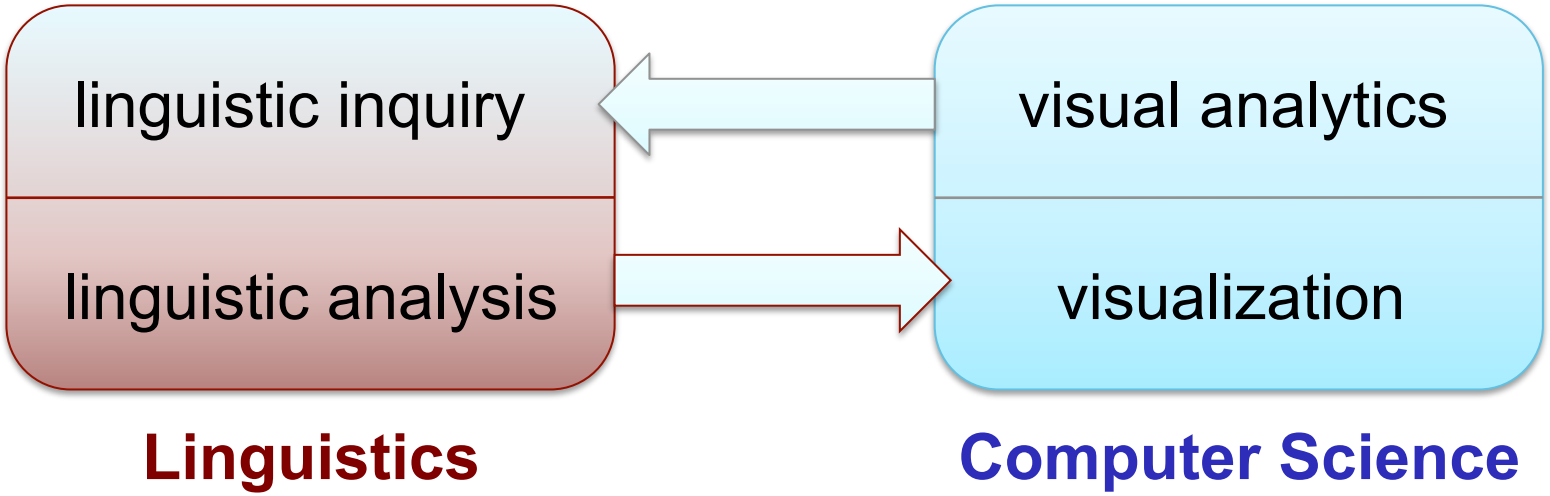
Mining Linguistic Data

Methodological Challenge/Opportunity

- Use of new technology to detect **distributional patterns** in language data.
- Ever increasing sources of digital data
  - Wikipedia, social media
  - constructed corpora (raw, annotated: morphology/syntax/semantics)
- Specialized query and search tools (KWIC, COSMAS, DWDS, ANNIS)
- Programming languages specialized for text processing and statistical analysis (Python, R)
- **Problem**: meaningful patterns difficult to see in the forest of numbers
- **Opportunity:** Visual Analytics for Linguistics (LingVis)

Overall Interdisciplinary Goal

- 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 visual access to data.
- Iterations of hypothesis-formation and hypothesis-testing.

Example: Identifying N-V complex predicates in Hindi/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)

[Butt et al., Coling 2012]

```
1  #this file lists X in X+kar, X+ho, X+hu, X+rakh sequences with corr...
   esponding occurrences in the (candidate) CP sequences
2  #X = word occurring directly to the left of LV (LV: kar, ho, hu, rakh)
3  #kar: # of occurrences of X with kar
4  #ho: # of occurrences of X with ho
5  #hu: # of occurrences of X with hu
6  #rakh: # of occurrences of X with rakh
7  X          #hu  #kar #ho  #rakh
8  اضافہ      674  466  524      0
9  شورش       378  2336 1691     0
10 معلوم      366  254  609      0
11 دھماکہ     359  135  44       0
12 حملہ       227  1232 100      0
13 رشتہ       183  178  765      0
14 نقصان      173  0    114      0
15 ایک        172  373  7027     0
16 ثبات       147  394  588      0
17 قوت        142  105  235      9
18 ادیب       103  754  956      0
19 کمال       102  1501 3609     0
20 مدد        80   210  96       0
21 فکر        74   0    263      0
22 تجربہ      62   59   1161     0
23 آغاز       59   315  75       0
24 بہ         56   0    2267     0
25 قدم        54   197  262      0
26 انکشاف     51   165  13       0
```

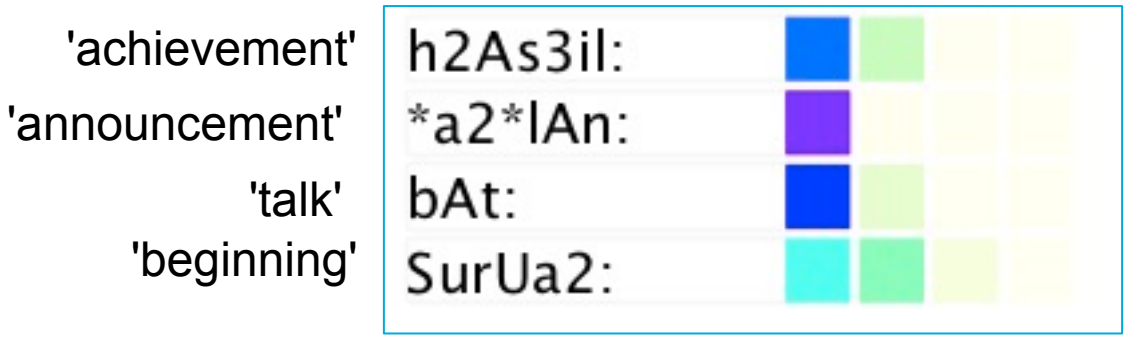
Example: Pixel Visualization

Statistical Data:

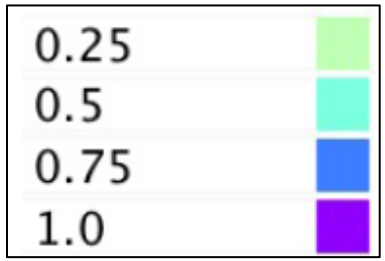
ID	Noun	Rel. freq. with *kar*	Rel. freq. with *ho*	Rel. freq. with *hu*	Rel. freq. with *rakh*
1	حاصل	0.771	0.222	0.007	0.000
2	اعلان	0.982	0.011	0.007	0.000
3	بات	0.853	0.147	0.000	0.000
4	شروع	0.530	0.384	0.086	0.000

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

do, be, become, put
kar, ho, hu, rakh

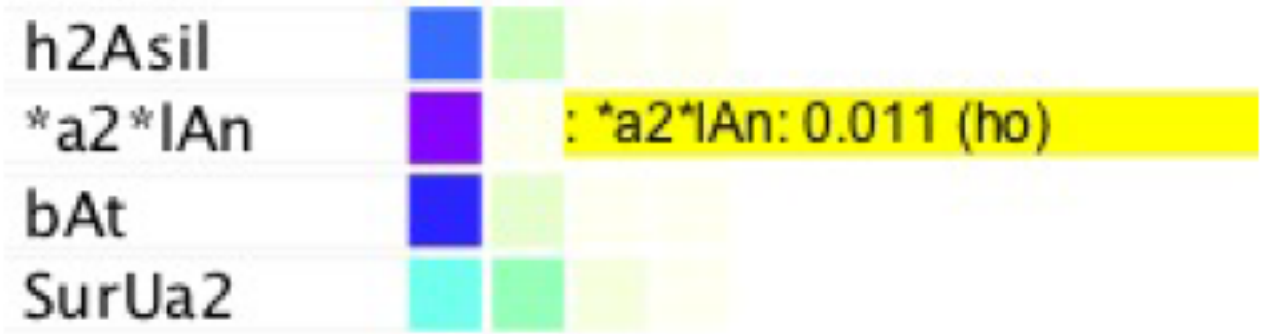


Color Scale



Example: Identifying N-V complex predicates in Hindi/lUrdu

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



Outliers/Errors are easily identified

Example: V1 in the History of 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?

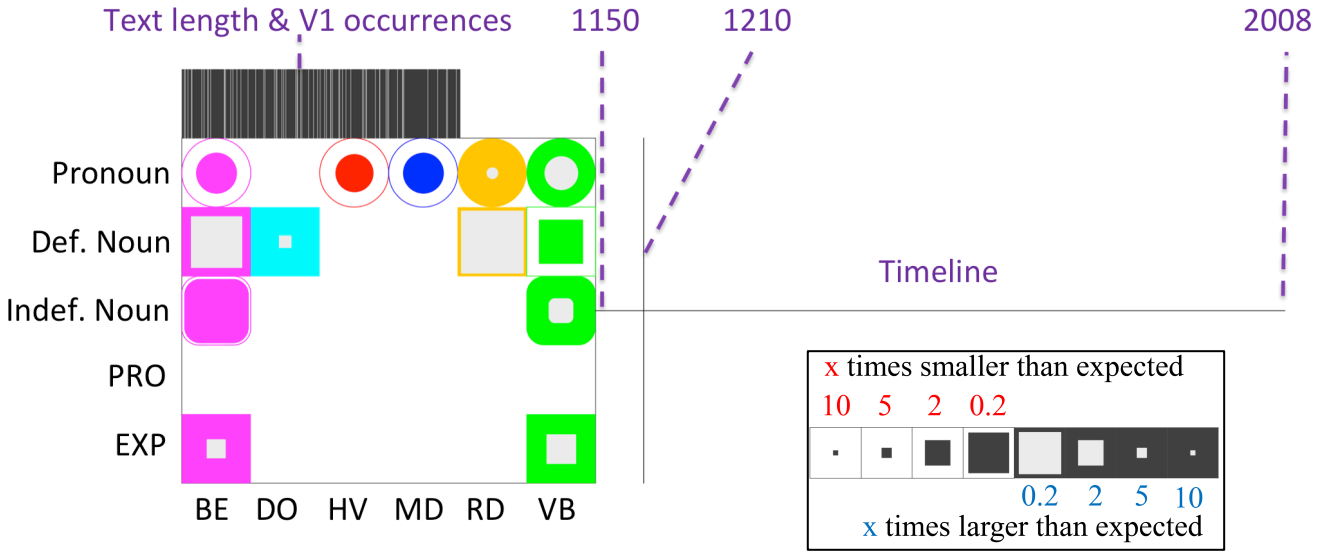
- **Case Study:** V1 in the history of Icelandic

- **Corpus:** IcePaHC
  - syntactically annotated (Penn Treebank style)
  - 60 texts
  - 12th century CE to 21st century CE

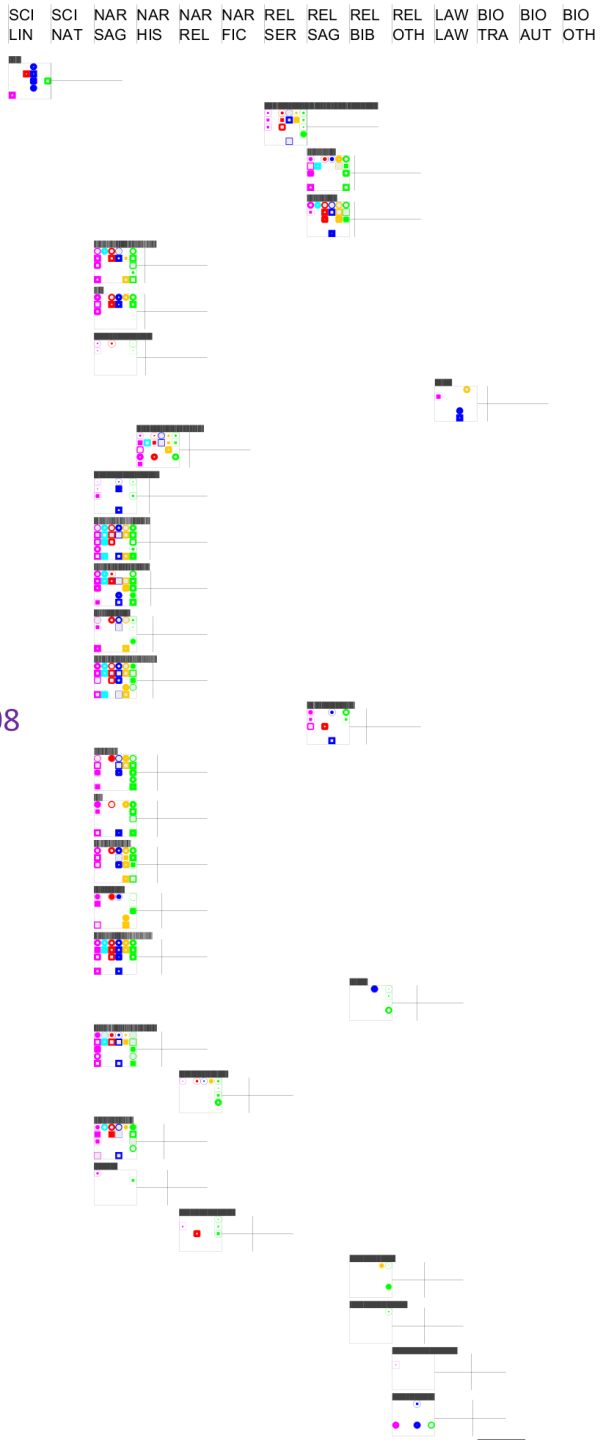
Example: V1 in Icelandic

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
- **Keim's Mantra**: Overview First – Details on Demand

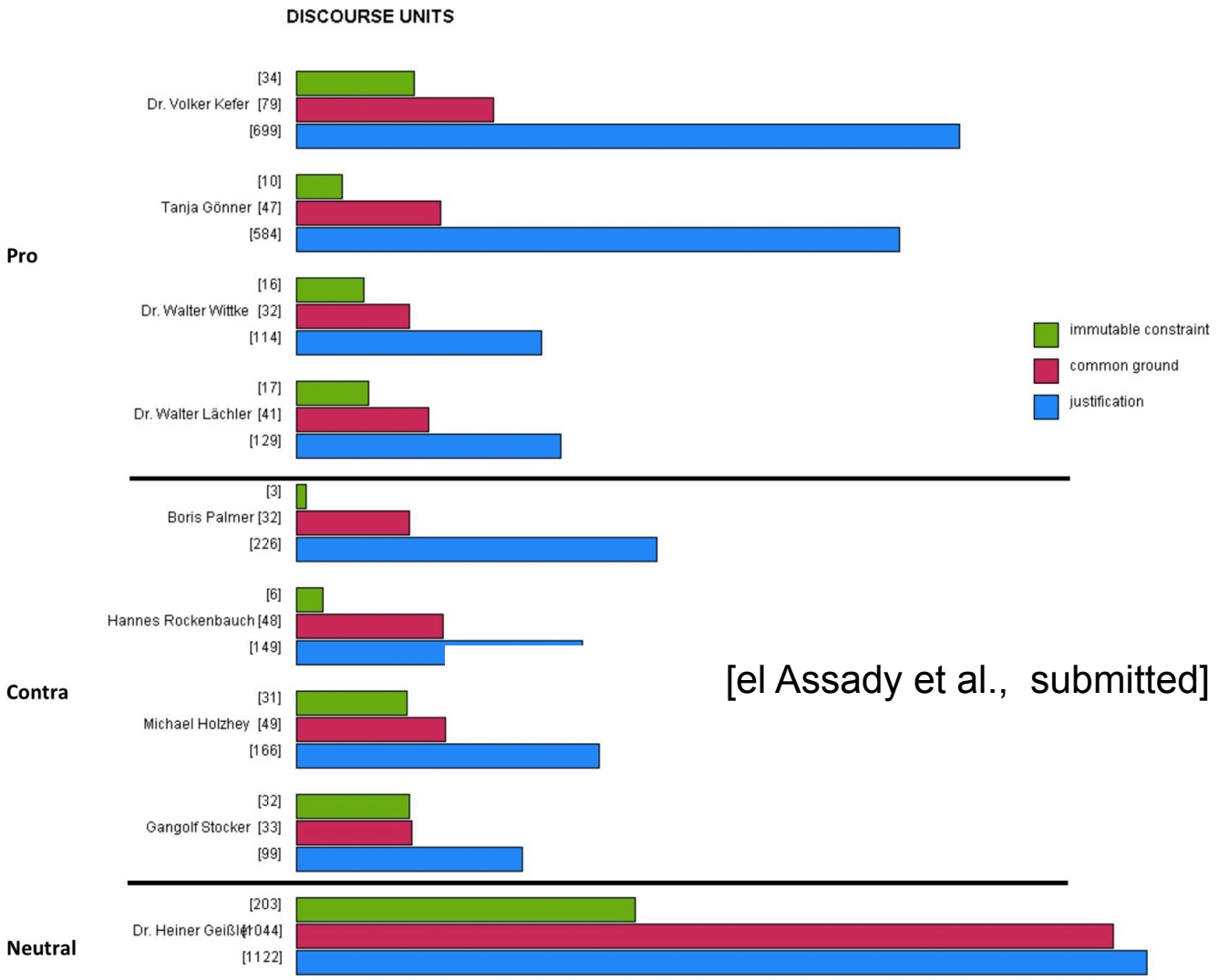


[Butt et al., LREC 2014]



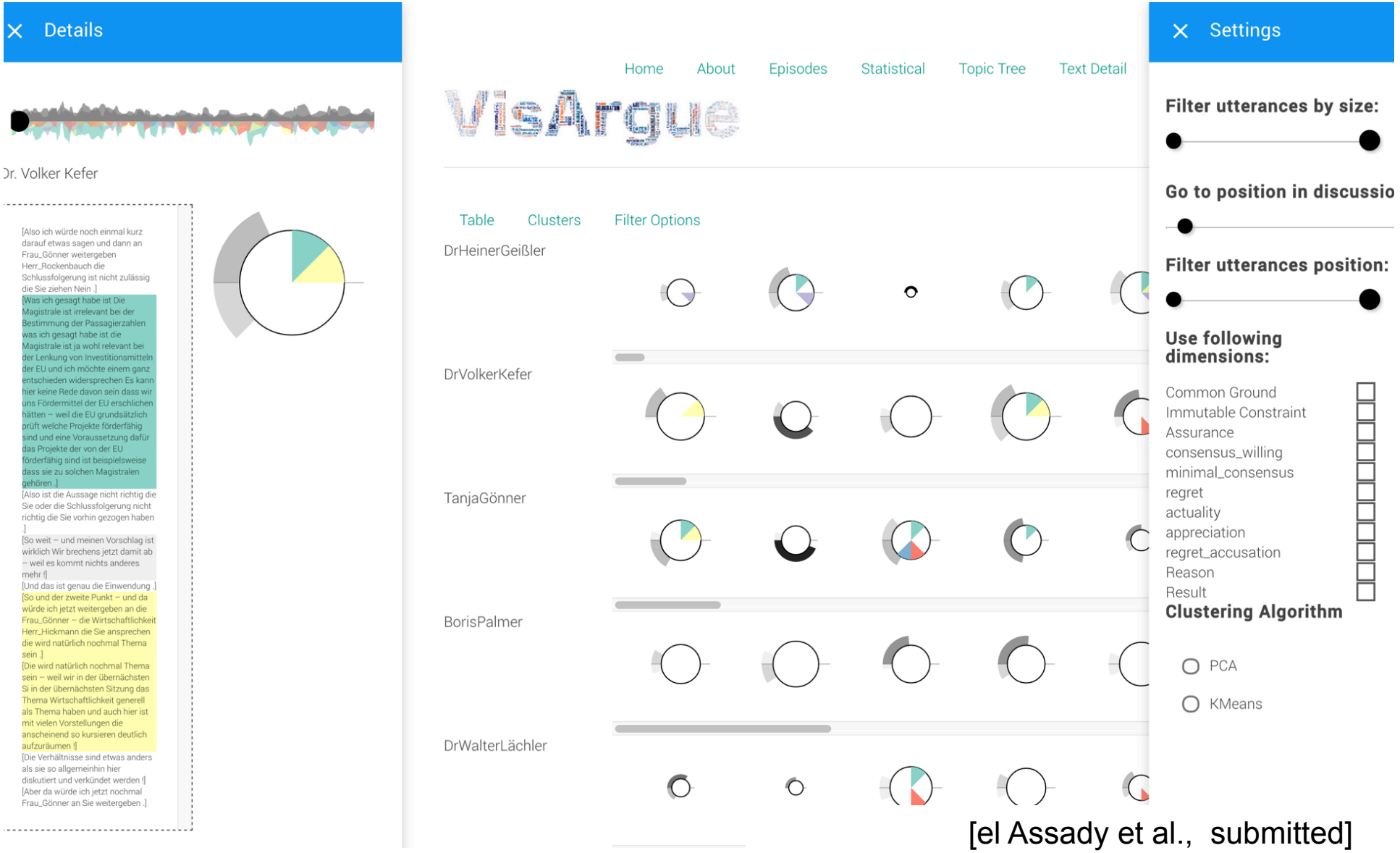
Example: Analyzing Political Argumentation (VisArgue)

- 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)



[el Assady et al., submitted]

Glyph Visualization of Utterance Content



[el Assady et al., submitted]

Example: Speech Data

- Japanese Native and German L2 Learner data (pitch contours and meta data)
- F0 contours are smoothed and normalized into pitch vectors
- The pitch vectors are visualized via self-organizing maps (SOM)

[Sacha et al., submitted]

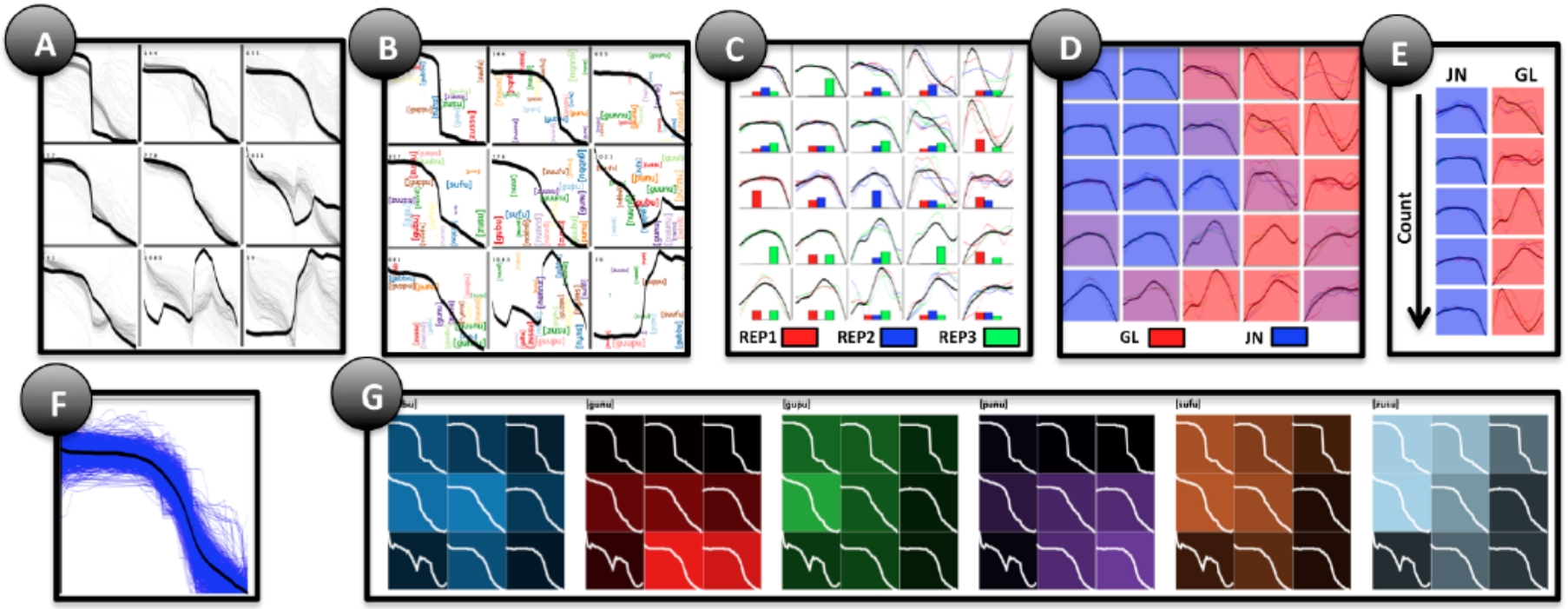


Figure 3: Different approaches to visualize SOM-results according to available meta-data. (A) Grid visualization, (B) word cloud, (C) bar charts, (D) mixed color cells, (E) ranked group clusters, (F) one single cell that visualizes contained vectors and the cluster prototype, (G) separated heatmaps for all values of a categorical attribute.

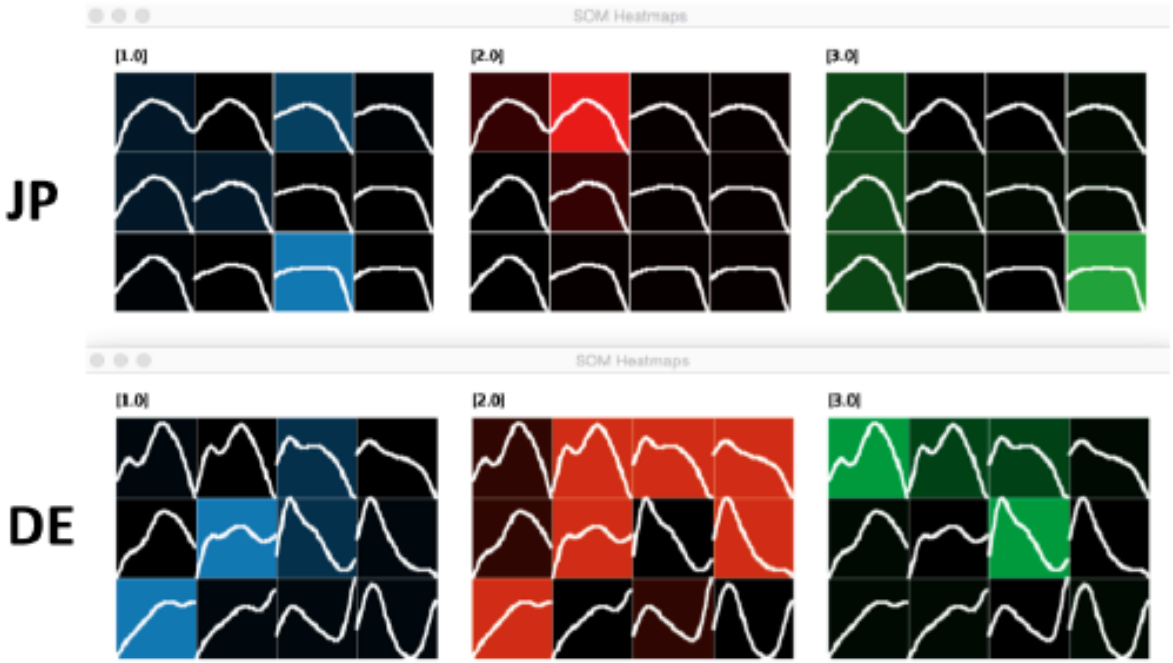
Example: Speech Data

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

- Japanese natives do not vary the pitch contour
- German learners do vary the pitch
- German beginner learners do so more

Interactive Exploration:

- individual cells can be merged
- meta data can be inspected



[Sacha et al., submitted]

Outlook

Further Exploration of Possibilities offered by Visual Analytics

- The systems illustrated here are very new.
- 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.

Thank You!

More and On-line:

- World Language Atlas Explorer: http://th-mayer.de/wals/#30A/
- PhonMatrix: http://paralleltext.info/phonmatrix/

Interdisciplinary Cooperation (University of Konstanz)

Linguistics

Tina Bögel, Annette Hautli-Janicz, **Thomas Mayer**, Maike Müller, Frans Plank, Christin Schätzle

Computer and Information Sciences

Daniel Keim, Menna el Assady, Andreas Lamprecht, Christian Rohrdantz, Dominik Sacha

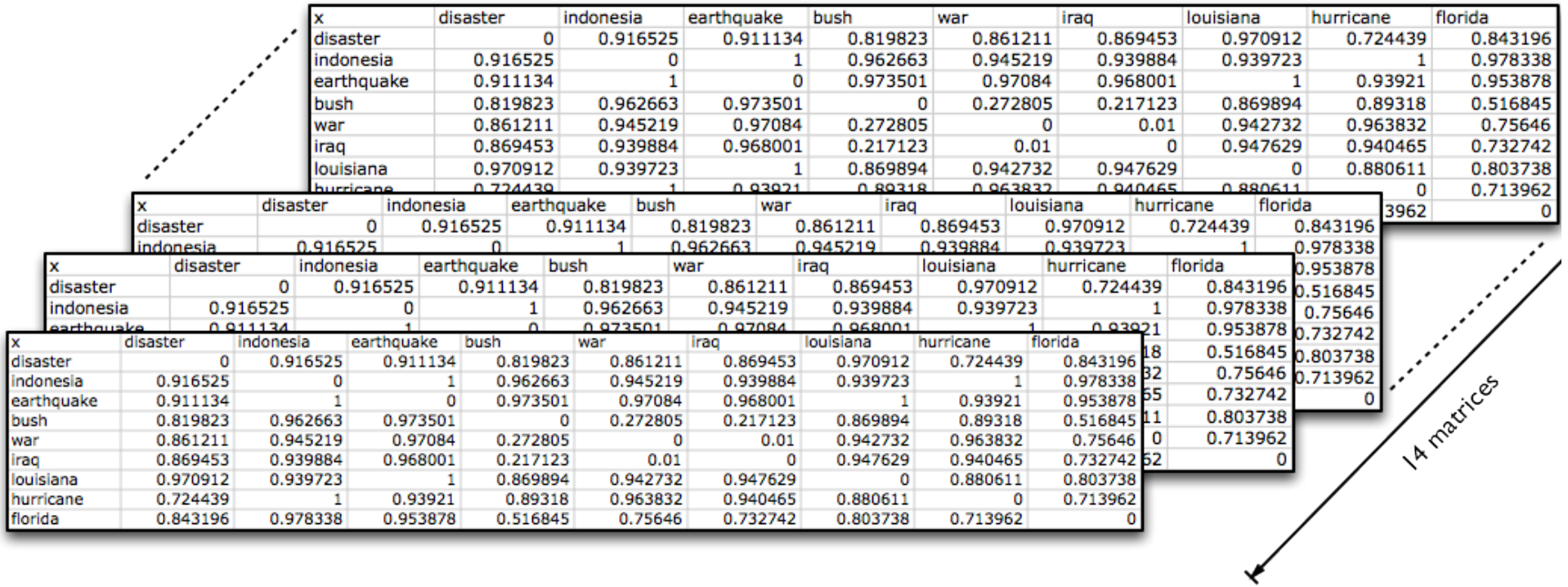
Political Science

Katharina Holzinger, Valentin Gold

Example: Correlations between highly frequent words in New York Times articles (2004-2005)

Raw Data

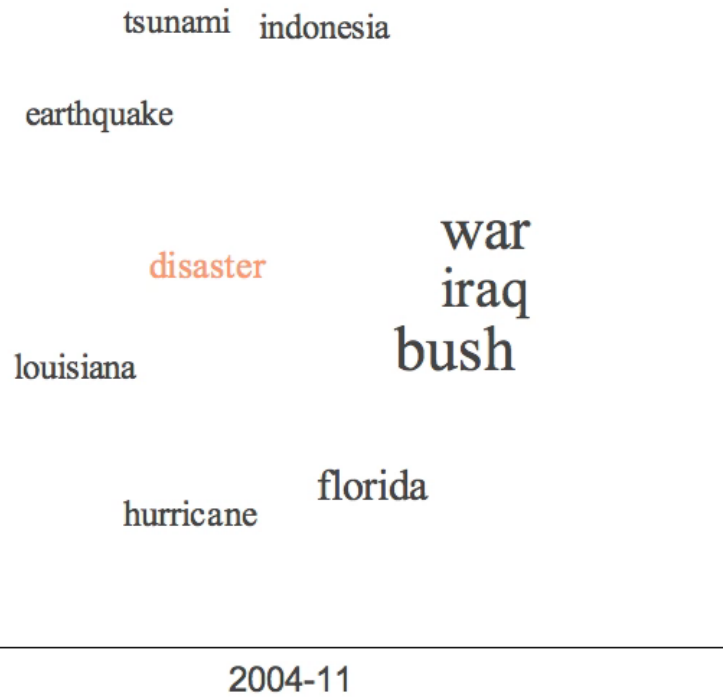
- 9x9 matrices
- 1 each for each of 14 time slices



Example: Correlations between highly frequent words in New York Times articles (2004-2005)

Animated Visualization
(Project Group Oliver Deussen, Univ. of Konstanz)

- Animation of trends/change over time
- Essentials of data easy to access via visualization
- Challenges for Visualization
  - dimensionality reduction: high dimensional distance matrices shown in 2D
  - precision vs. stability: a precise visualization for each time step would induce too much confusing movement

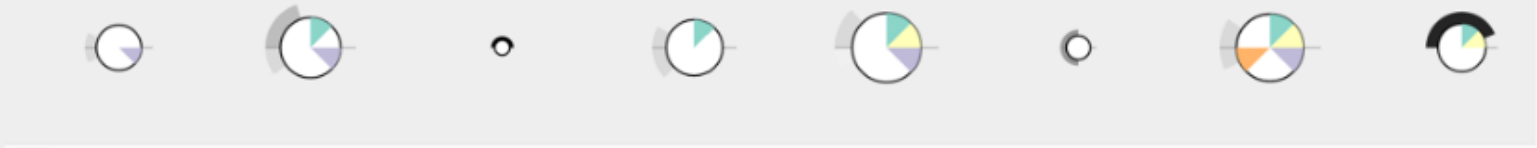


Home About Episodes Statistical Topic Tree Text Detail Deliberation Argumentation Contact

VisArgue

Table Clusters Filter Options

DrHeinerGeißler



DrVolkerKefer

TanjaGönner

BorisPalmer

DrWalterLächler