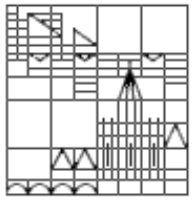
SPONSORED BY THE
Federal Ministry
of Education
and Research

Universität
Konstanz



Visual Analytics for Linguists

Miriam Butt & Chris Culy

ESSLII 2014, Introductory Course
Tübingen

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

PHILOSOPHISCHE
FAKULTÄT
Seminar für Sprachwissenschaft

Course Overview

- Day 1: LingVis
  - First Look at Possible Visualizations for Linguistics
  - Basics of Visualization (Theory)
- Day 2: LingVis II (More Use Cases and Theory)
- Days 3&4: Hands-On: Working with Visualizations
- Day 5:
  - Short tour of other tools
  - Where to go from here
  - Discussion

Day 1 – Intro to LingVis

1. Organizational Matters
2. Why use Visual Analytics for Linguistics
3. Sample Visualizations of Linguistic Information (Use Cases)
4. Visualization Basics (Theory)

Organizational Matters

- Who are we?

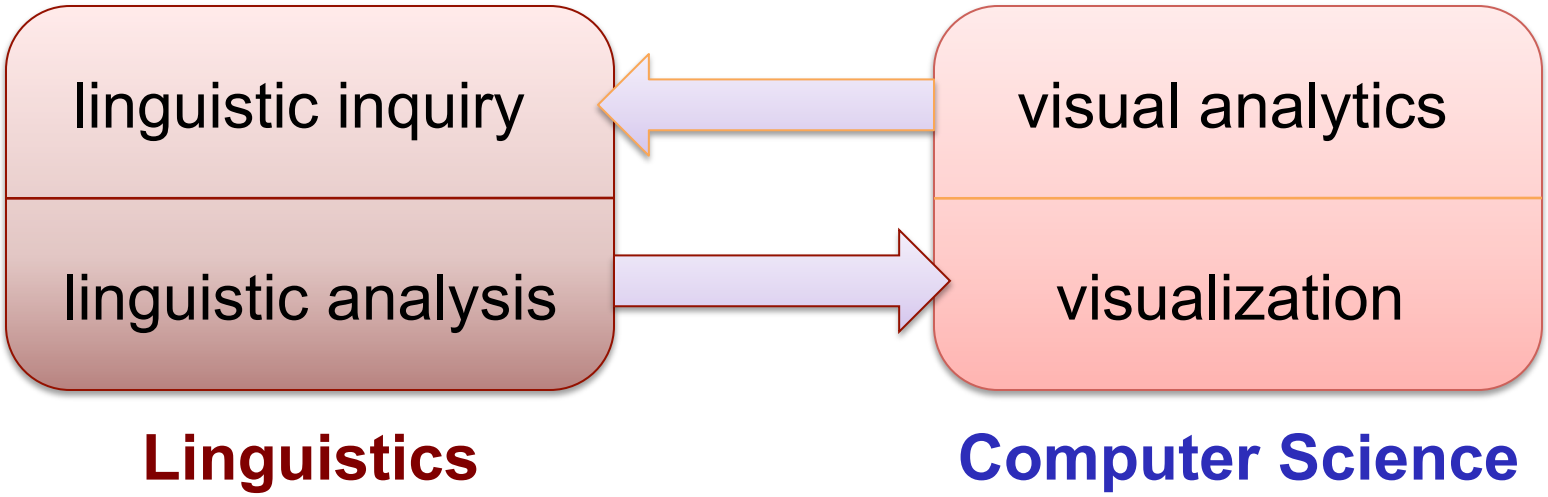
- Who are you???
  - Programming Background?
  - What types of linguistic questions interest you?
  - Do you have laptops?

LingVis

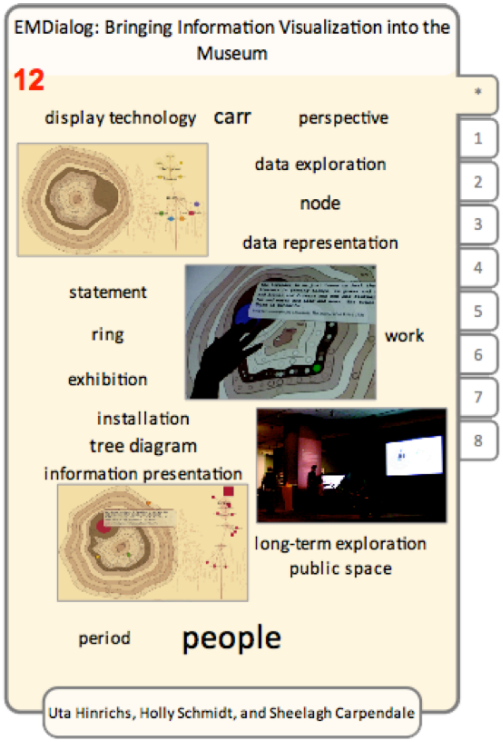
Overall Goals:

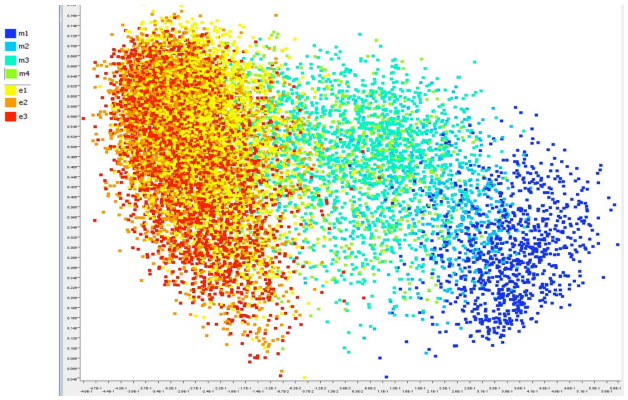
⊙ Integrate methods from **visual analytics** into domains of **linguistic inquiry**.

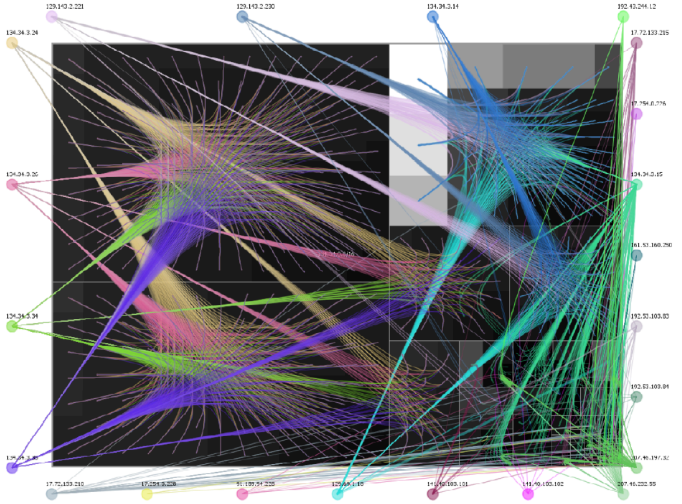
⊙ Explore challenges based on the needs of **linguistic analysis** for **visualization methods**.

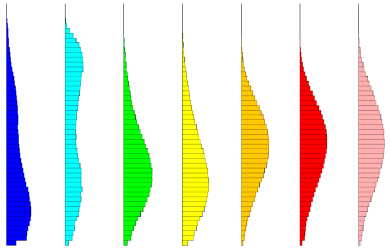


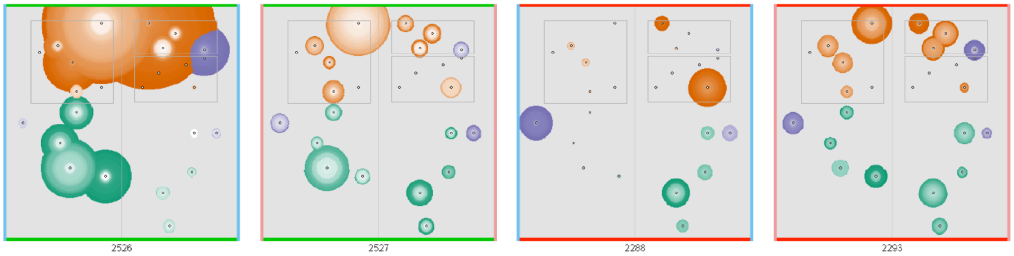
Sample Visualizations

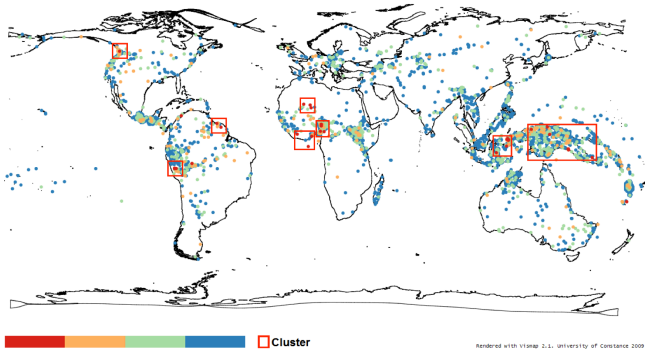






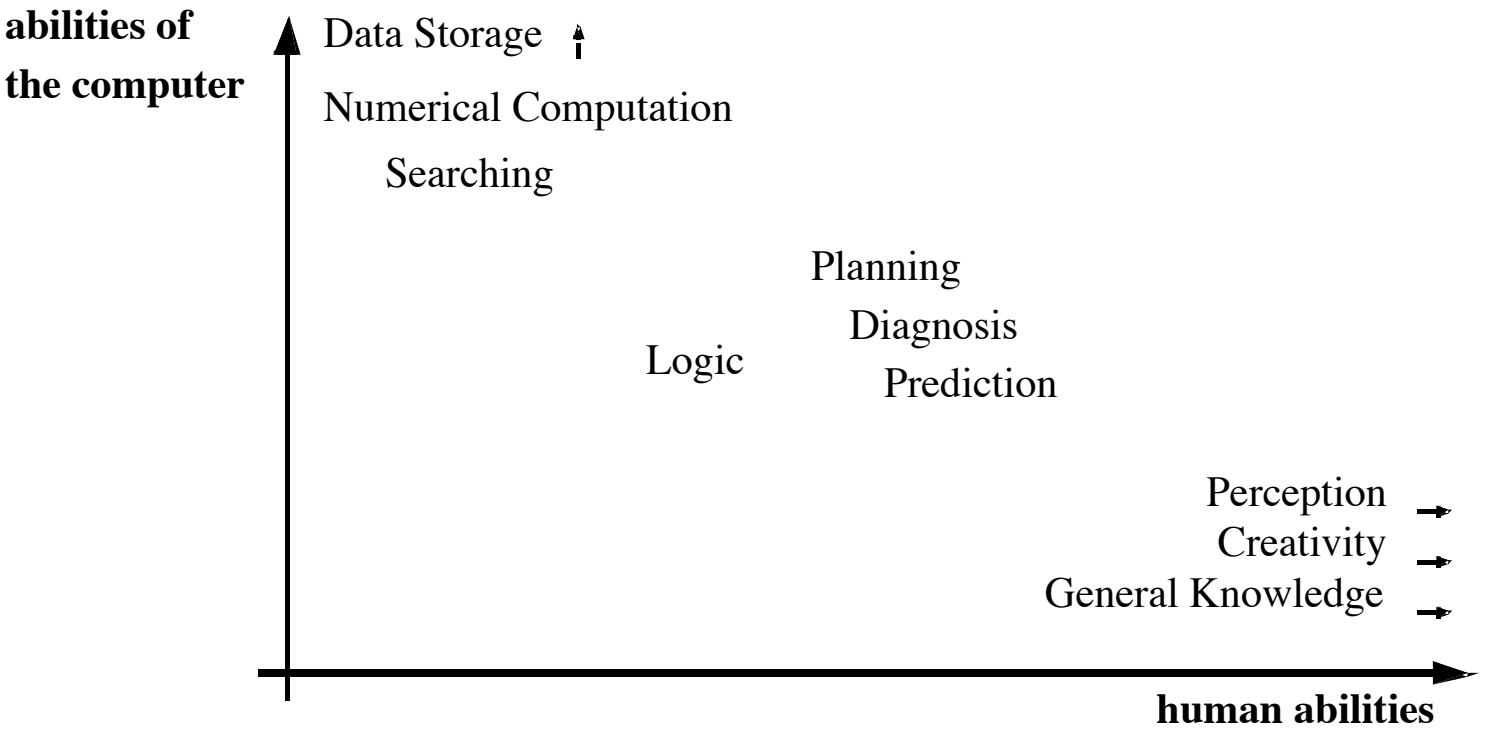






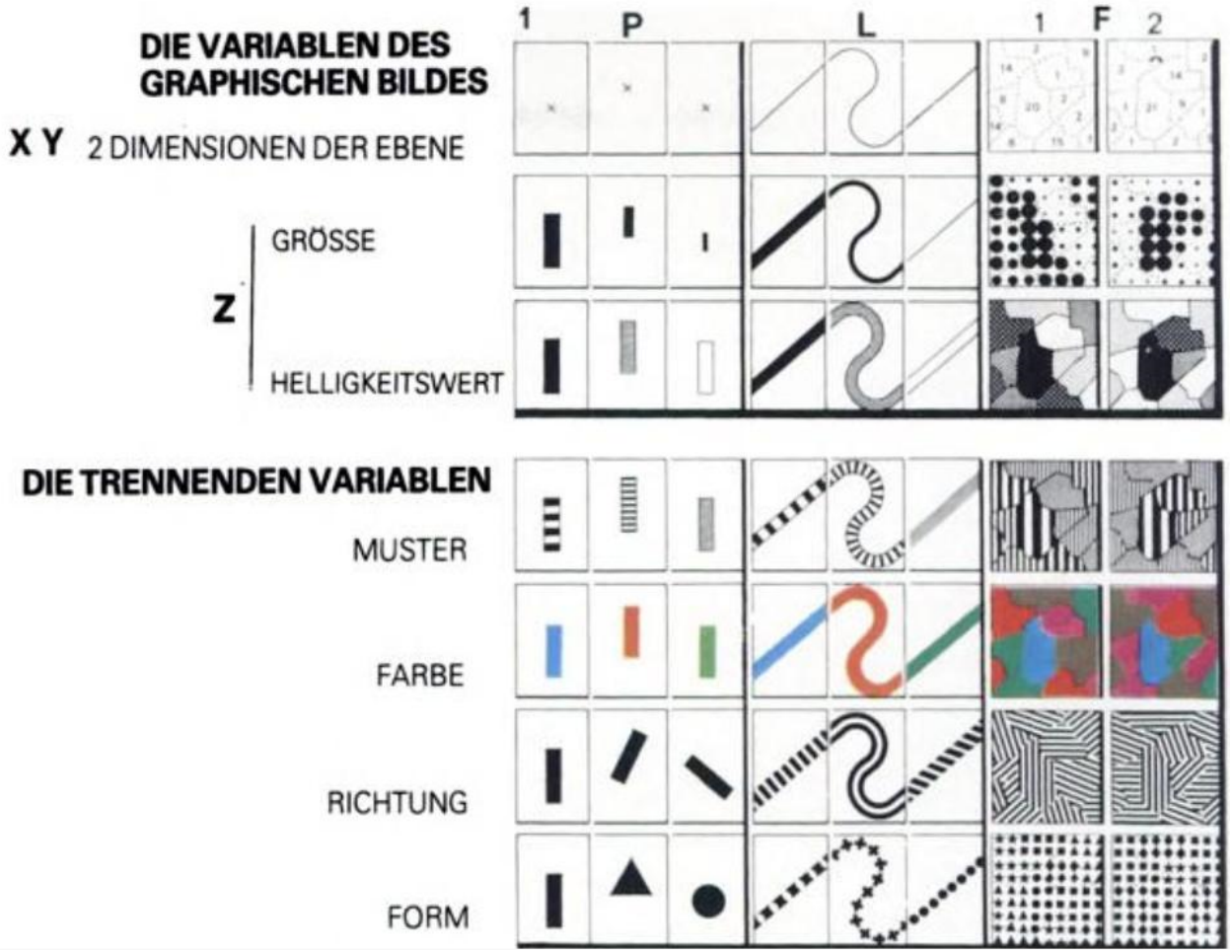
Why use Computation for Linguistic Research?

- Computer abilities complement human abilities
- **Visual Analytics**: tight integration of computation with user interactive visualizations



Why use Visualization?

- Good **interface** between computers and humans
- Triggers **pre-attentive perception**



The 8 visual variables (Bertin 1982)

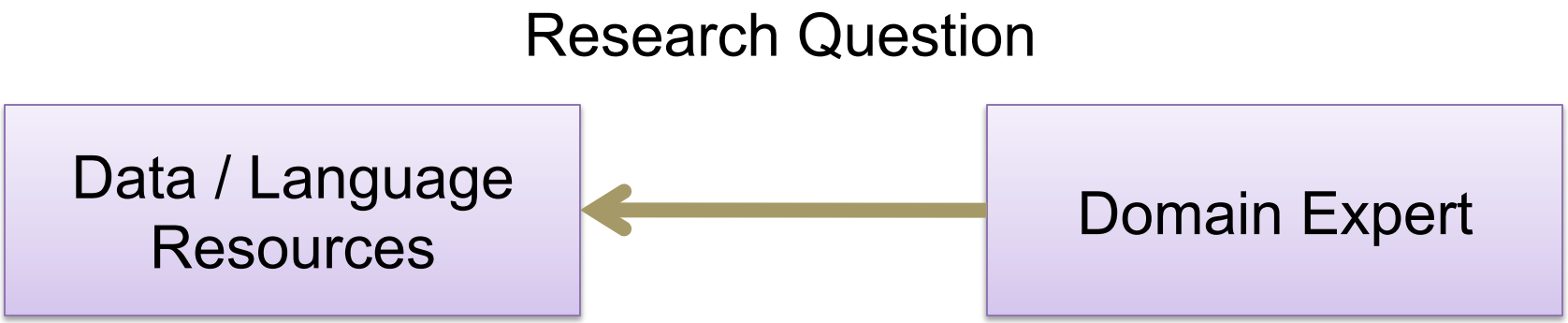
LingVis – Motivation

- 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).
- **But:** as yet only comparatively little/good use of **visualization methods**.

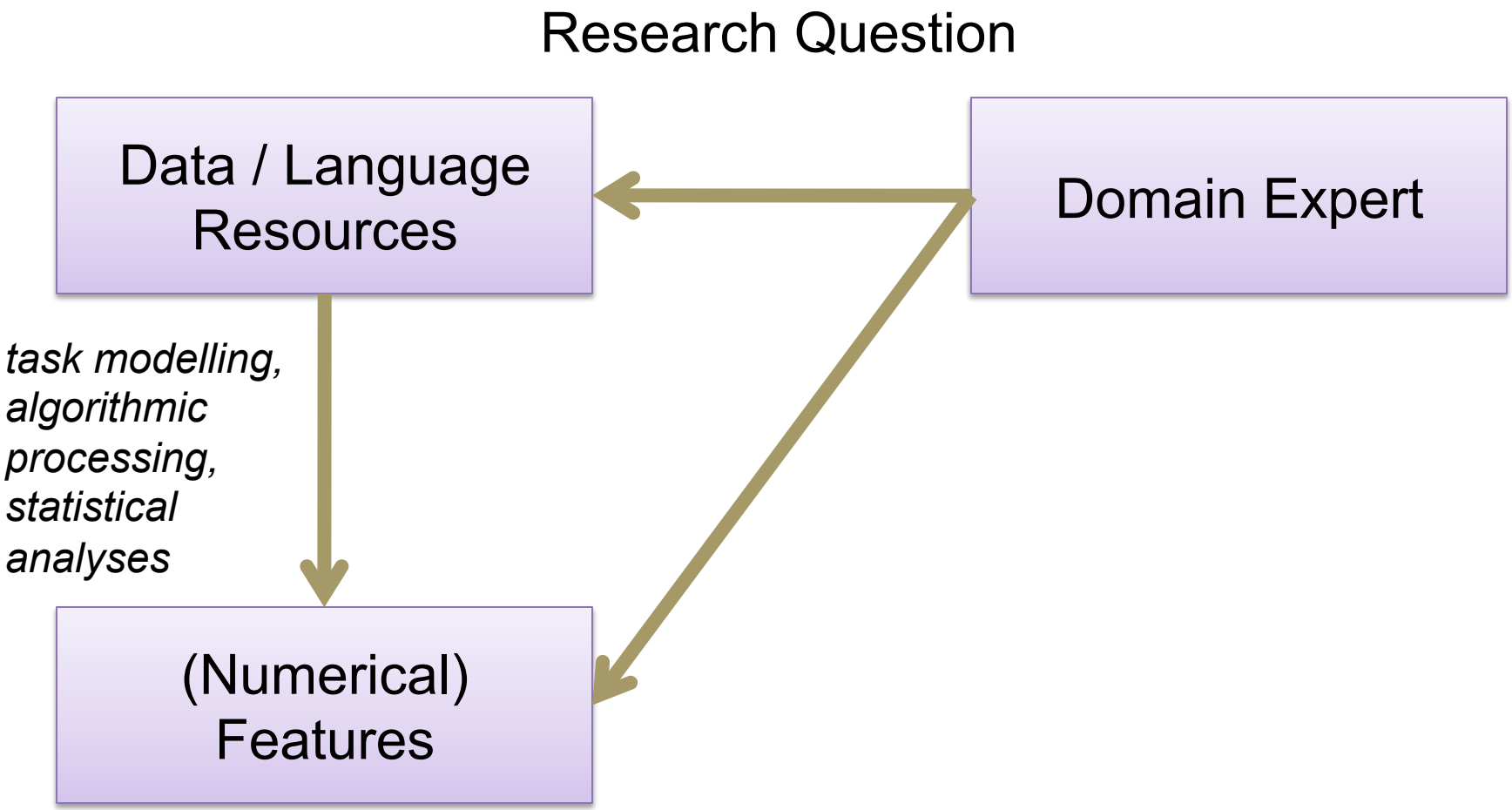
Making Sense of Numbers

- 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.

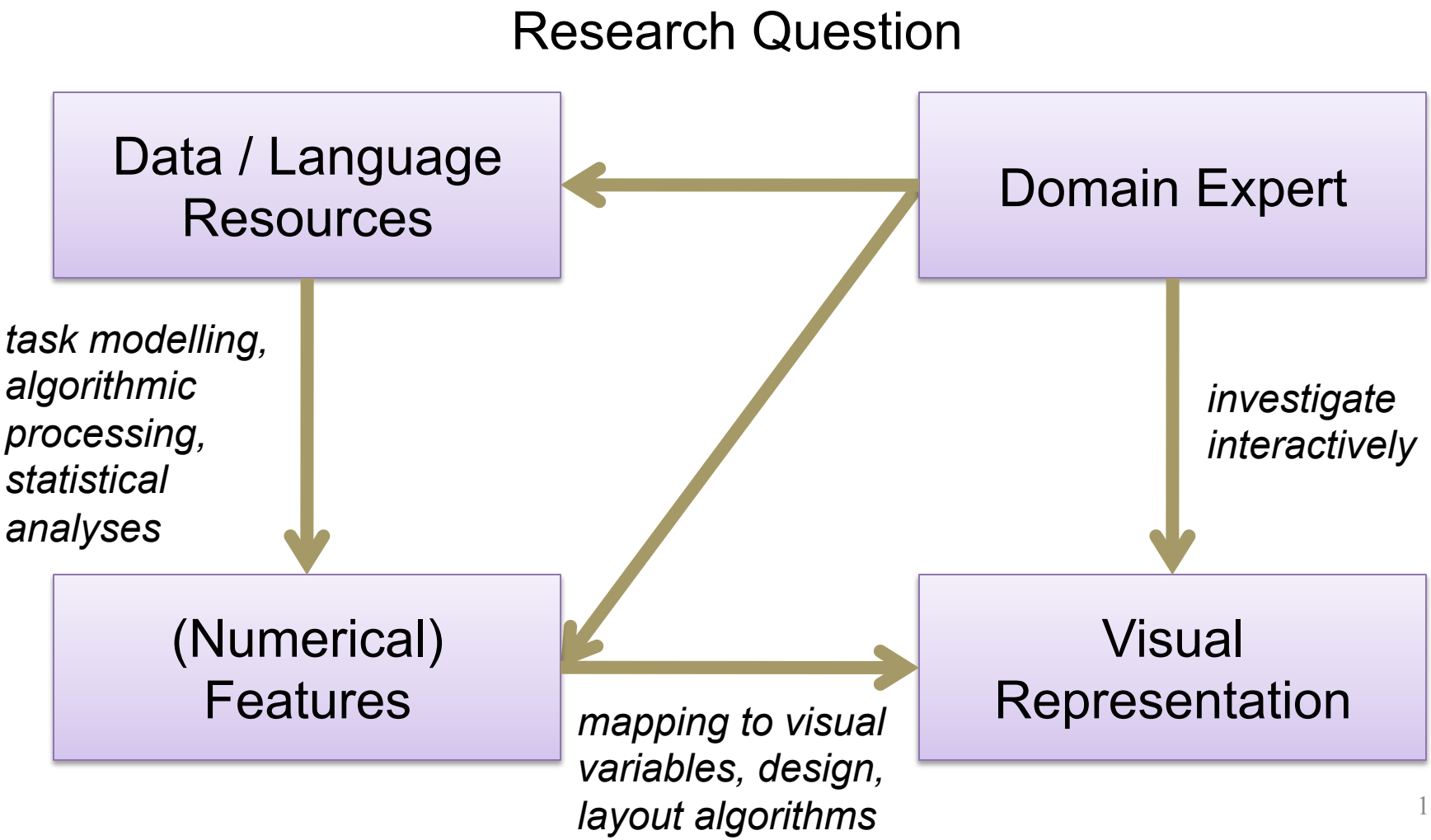
Interdisciplinary Collaboration: LingVis



Interdisciplinary Collaboration: LingVis



Interdisciplinary Collaboration: LingVis



Example: Pixel-Based Visualizations

Two Use Cases

- Vowel Harmony
- N-V Complex Predicates

Vowel Harmony (VH)

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

Turkish

Genitive suffix
deniz-in, ev-in
tütün-**ün**, çöl-**ün**
kadın-ın, adam-ın
sabun-**un**, top-**un**

Genitive suffix with plural suffix
deniz-ler-in, ev-ler-in
tütün-ler-**in**, çöl-ler-**in**
kadın-lar-ın, adam-lar-ın
sabun-lar-**ın**, top-lar-**ın**

Vowel Harmony

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 ϕ.
- **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?

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

Turkish

	a	i	o	e	u
a	-0.003	-0.075	0.094	-0.025	-0.018
i	-0.025	-0.004	0.064	-0.036	0.005
o	-0.028	-0.006	-0.075	0.098	0.026
e	-0,001	0.063	-0.073	0.016	0.021
u	0.077	0.038	-0.036	-0.057	-0.043

Spanish

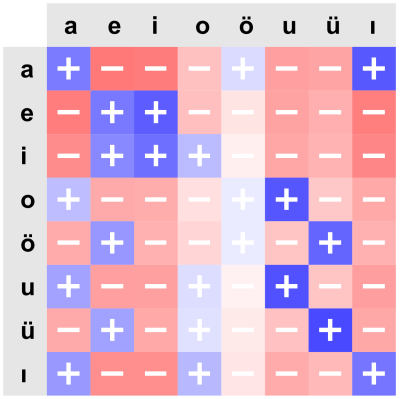
	a	o	i	ü	ö	ä	u	e
a	0.019	0.009	-0.061	-0.034	-0.008	-0.025	0.018	0.035
o	-0.023	-0.004	-0.052	-0.013	-0.020	-0.013	-0.013	0.068
i	-0.069	-0.054	-0.050	-0.039	-0.036	-0.044	-0.003	0.133
ü	-0.067	-0.045	0.070	-0.028	-0.021	-0.033	-0.021	0.050
ö	-0.049	-0.032	0.049	-0.024	-0.013	-0.021	-0.013	0.036
ä	-0.067	-0.037	0.124	-0.033	-0.018	-0.028	-0.038	0.020
u	0.012	-0.018	-0.019	0.046	-0.002	-0.013	0.004	-0.001
e	0.108	0.084	0.026	0.069	0.063	0.096	0.021	-0.195

German

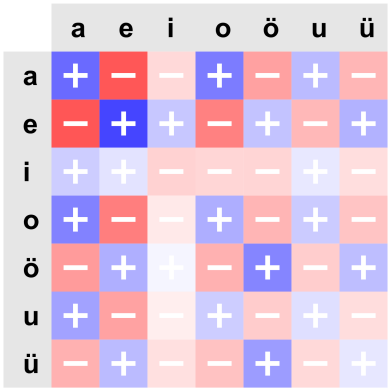
	a	o	u	i	ü	ö	e
a	0.339	0.263	0.070	-0.022	-0.081	-0.136	-0.431
o	0.239	0.099	0.041	-0.007	-0.052	-0.083	-0.253
u	0.132	0.038	0.015	-0.004	-0.017	-0.040	-0.131
i	0.037	-0.026	0.008	-0.030	-0.017	-0.027	0.011
ü	-0.093	-0.056	-0.022	-0.014	0.008	0.148	0.071
ö	-0.152	-0.093	-0.037	0.001	0.065	0.229	0.097
e	-0.435	-0.241	-0.076	0.048	0.091	0.054	0.531

Hungarian

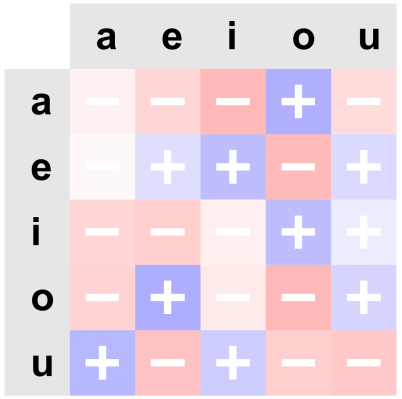
First Simplistic Visualization: Can you detect a pattern?



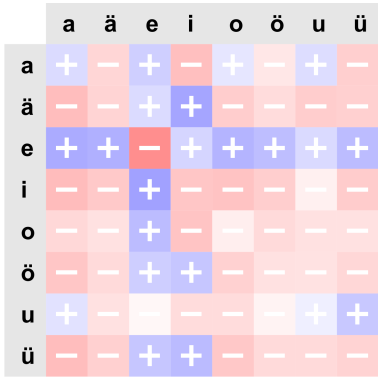
Turkish



Hungarian



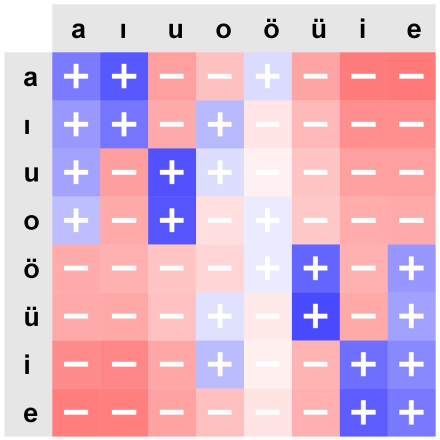
Spanish



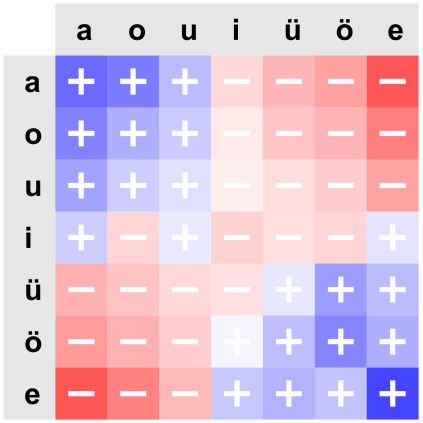
German

- 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.

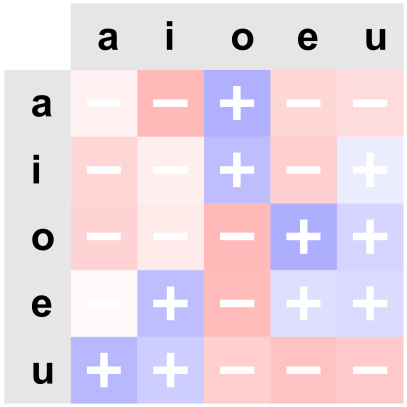
Sorted Visualization: Can you detect a pattern now?



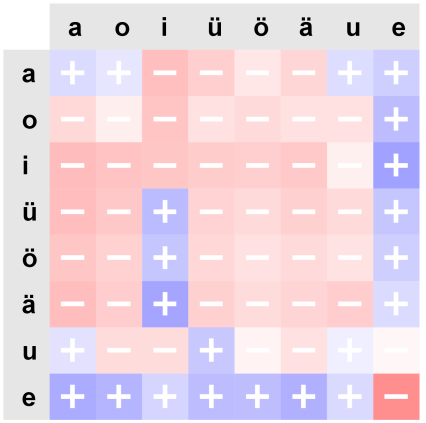
Turkish



Hungarian



Spanish



German

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

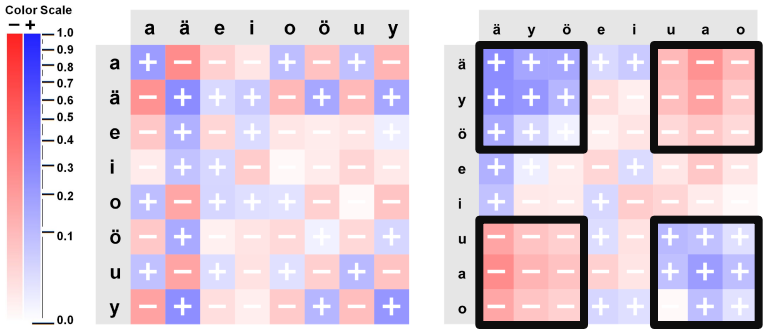
T. Mayer, C. Rohrdantz, M. Butt, F. Plank and D. A. Keim. **Visualizing Vowel Harmony**. *Linguistic Issues in Language Technology*, 4(Issue 2):1-33, 2010.

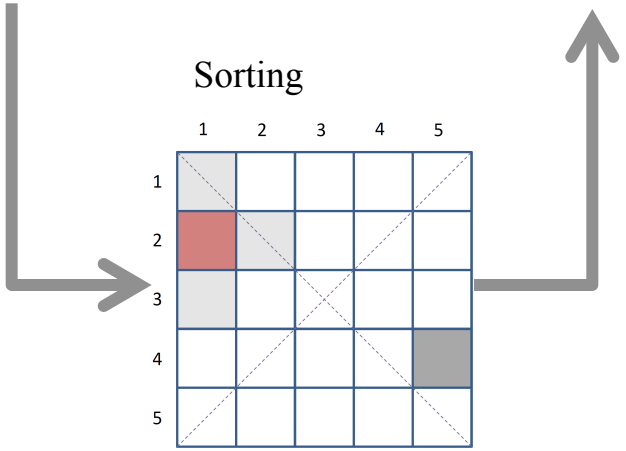
Visualizing Vowel Harmony

Counting Vowel Successions in all Bible Types
Example: Finnish

	a	ä	e	i	o	ö	u	y
a	3548	20	1940	1893	831	0	944	24
ä	35	944	806	820	10	138	33	266
e	1623	1144	1495	1608	419	56	497	187
i	1580	854	1514	1044	376	46	355	135
o	1384	7	1032	902	284	0	294	8
ö	7	125	54	39	0	3	1	18
u	1464	6	1085	850	315	1	547	8
y	39	656	368	368	35	75	4	251

Statistics & Visualization

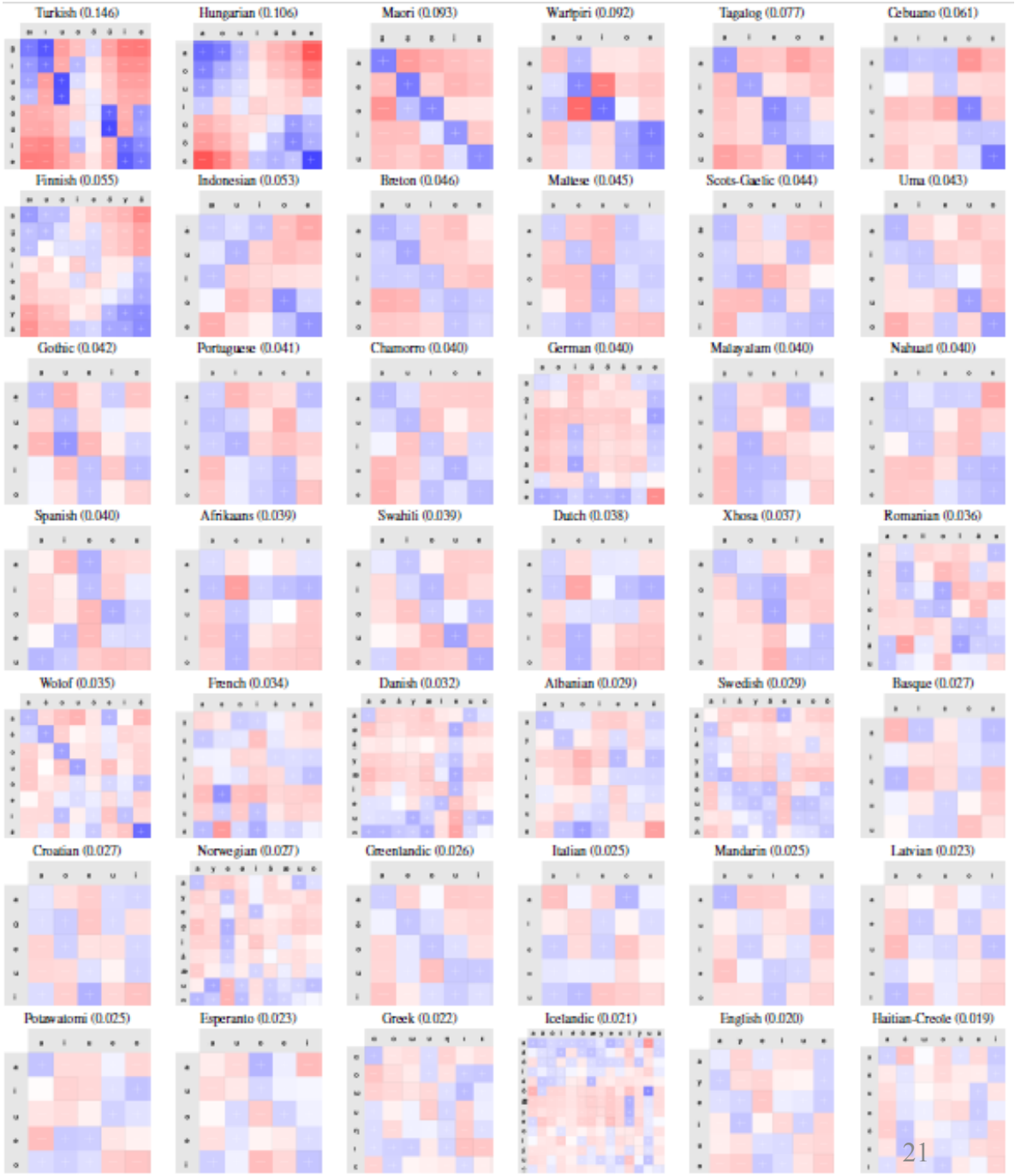




Sorting done according to feature vectors of each of the rows.

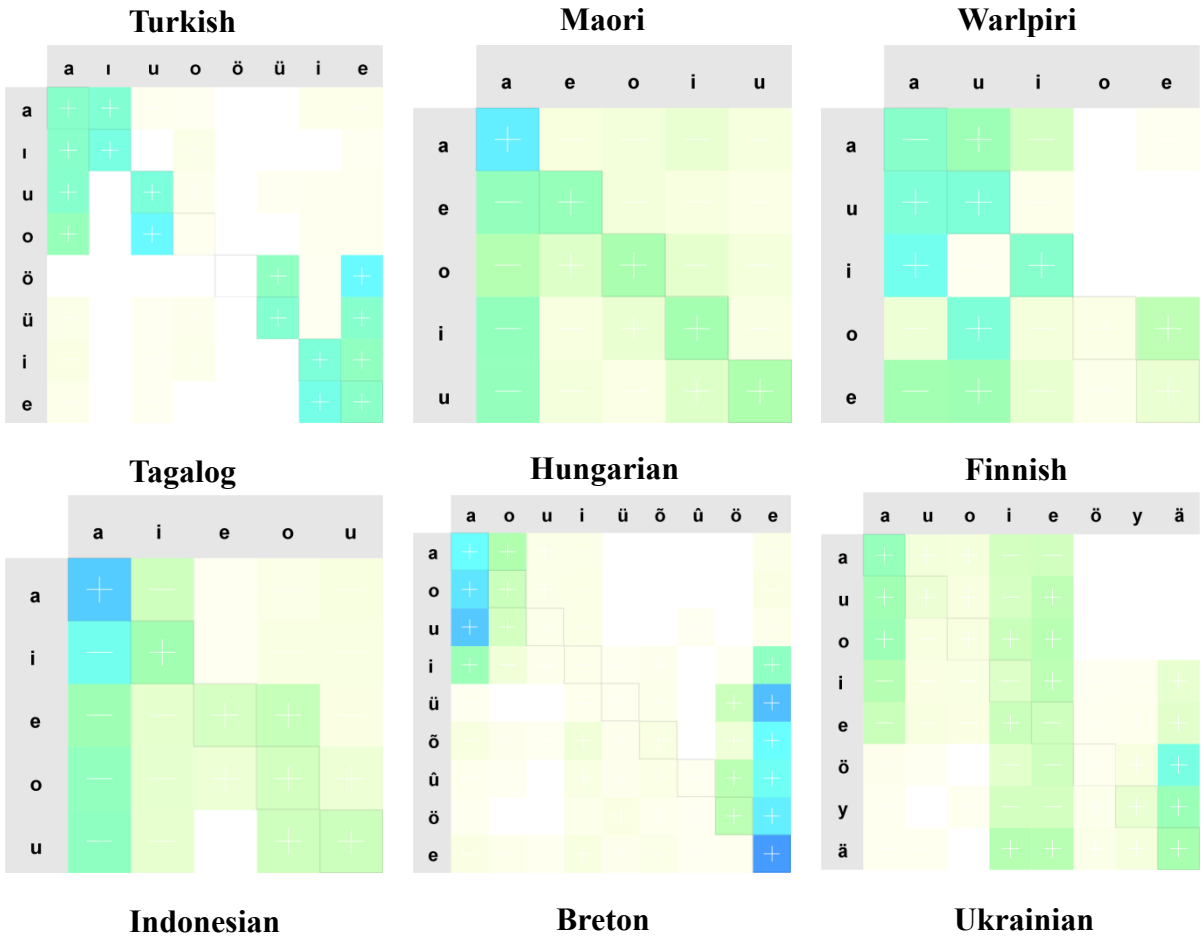
Results – Sorted Visualization:

- Automatic Visual Analysis of vowel successions for 42 languages – sorted for effect strength.



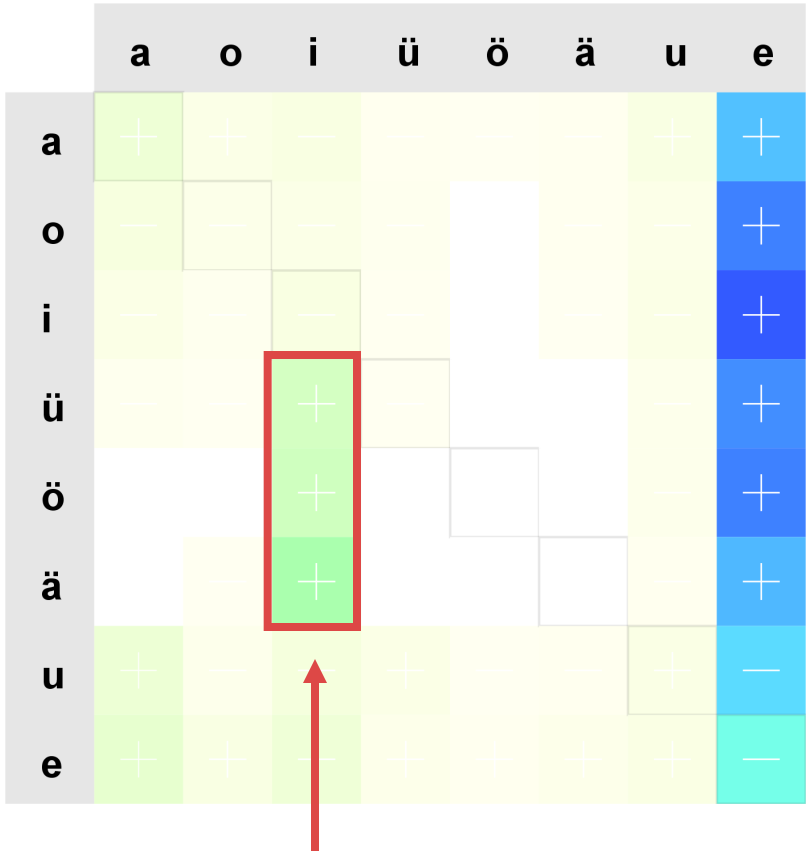
Vowel Harmony vs. Reduplication

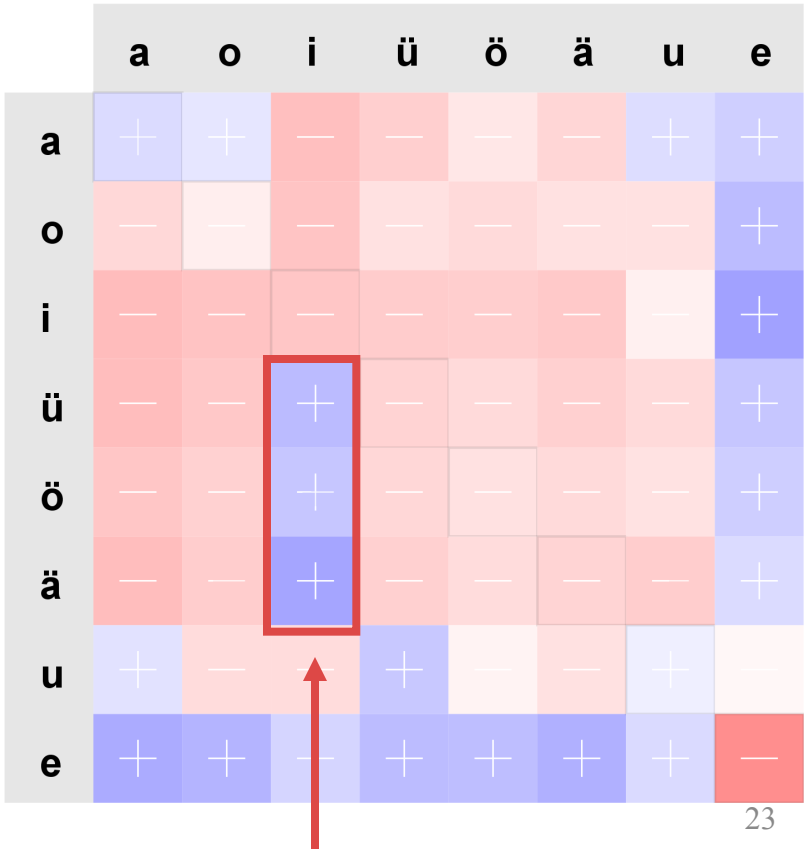
- In VH languages, crucially there are some vowels which never co-occur.
- This can be seen via a calculation of succession probabilities.
- Maori is not a VH language.



Historical Fingerprint: German Umlaut

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.

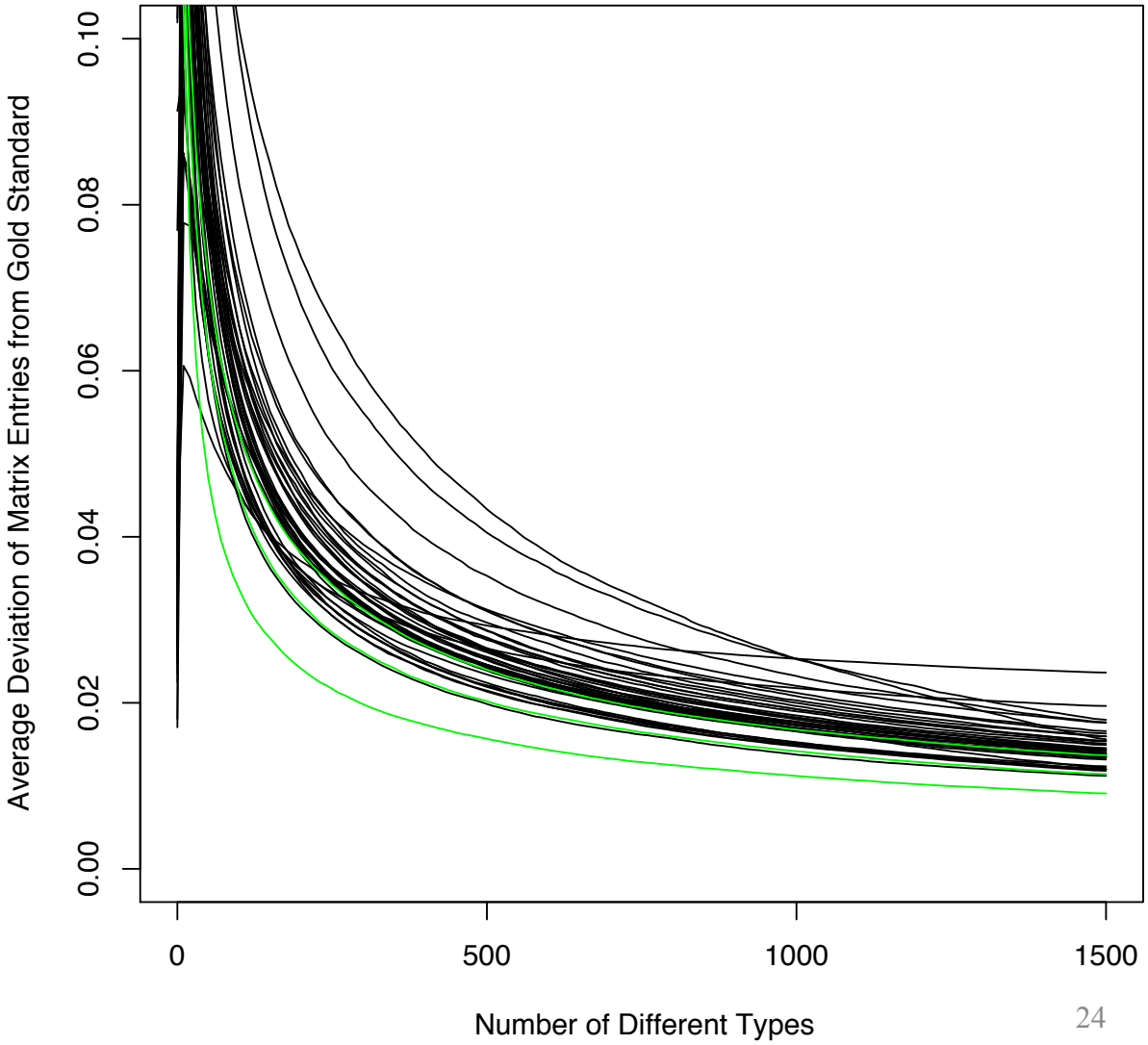




Further Nice Features

Only 2000-4000 words needed for a reliable analysis!

(The green colored lines are the VH languages.)



Further Nice Features

You can use the visualization in a new and improved form yourself on-line.

http://paralleltext.info/phonmatrix/

Main Contact Person: Thomas Mayer

Mayer, Thomas and Christian Rohrdantz. 2013. PhonMatrix: Visualizing co-occurrence constraints in sounds. In *Proceedings of the ACL 2013 System Demonstration*.

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 cooccur with which verbs.
- **Study:** took an 8 million Urdu corpus collected from BBC Urdu.

N-V Complex Predicates

- **Calculation:** counted how many times a given noun occurred with one of four (light) verbs (e.g., 75%).
- Sample data:

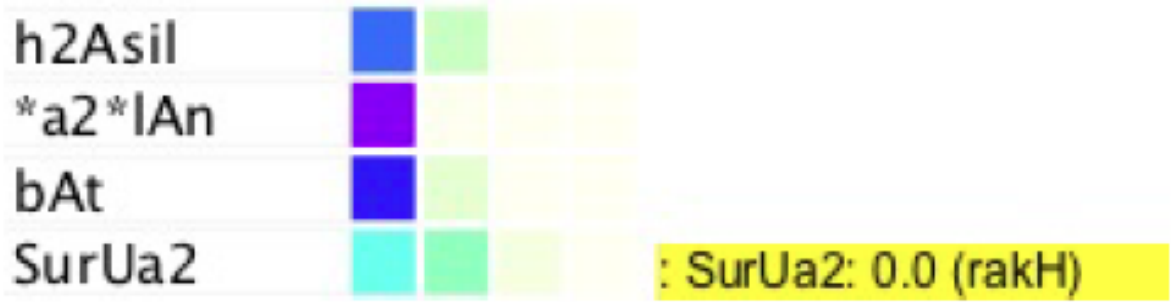
```
X,kar,ho,hu,rakh,
hAsil,0.771,0.222,0.0070,0.0
bAt,0.853,0.147,0.0,0.0
istamAl,0.873,0.121,0.0060,0.0
kOSiS,0.823,0.177,0.0,0.0
band,0.695,0.261,0.0,0.045
hamlah,0.79,0.064,0.146,0.0
zAhir,0.699,0.289,0.012,0.0
sAmnA,0.686,0.301,0.013,0.0
....
```

- Hard to evaluate in this form.

Raw data

Noun	(do) *kar*	(be) *ho*	(become) *hu*	(put) *rakH*	
h2Asil	0.771	0.222	0.007	0.000	(achievement)
*a2*lAn	0.982	0.011	0.007	0.000	(announcement)
bAt	0.853	0.147	0.000	0.000	(talk)
SurUa2	0.530	0.384	0.086	0.000	(beginning)

Visualized data



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

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).

Pixel plus Cluster Visualization

- Main patterns for nouns:

- Can mouse over to get exact values for the visualization.
- The more saturated a color, the higher the occurrence.

N-V Complex Predicates

Cluster Visualization Demo

More sophisticated version now available – will also look at that.

Andreas Lamprecht, Annette Hautli, Christian Rohrdantz, Tina Bögel. 2013. A Visual Analytics System for Cluster Exploration. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, System Demo,* 109–114, Sofia, Bulgaria.

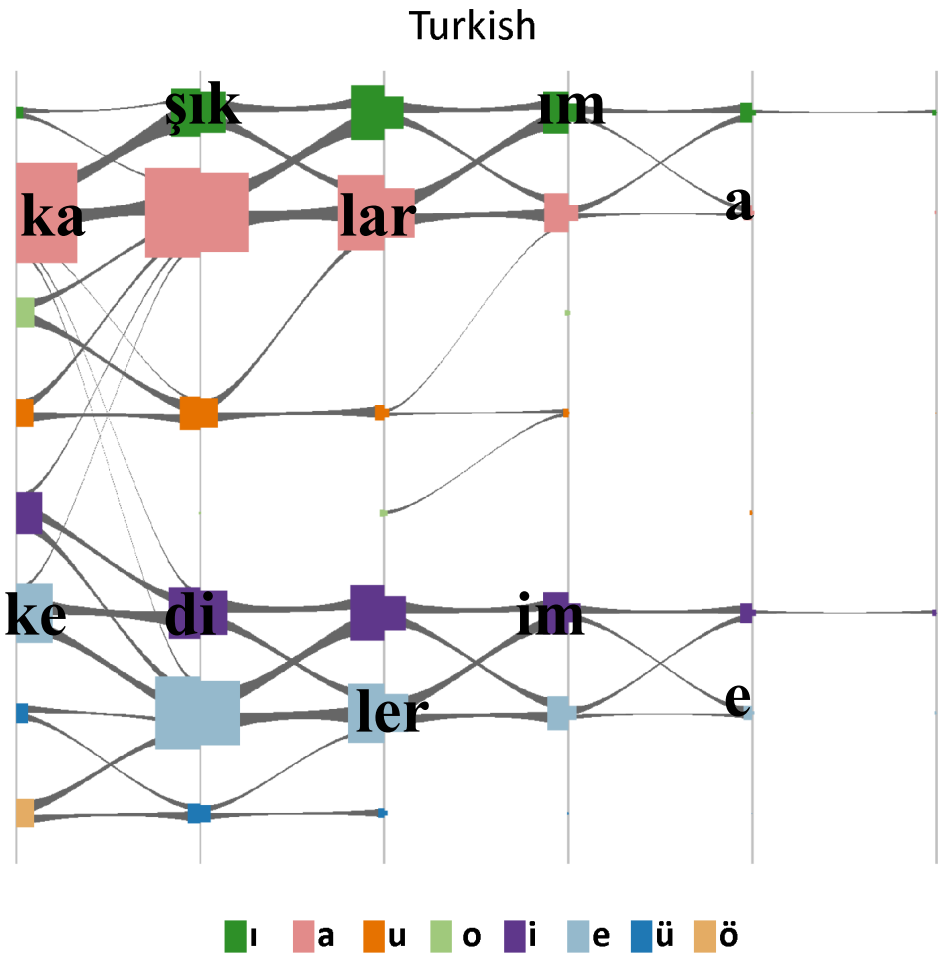
Example: Droplet Visualizations

- 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).

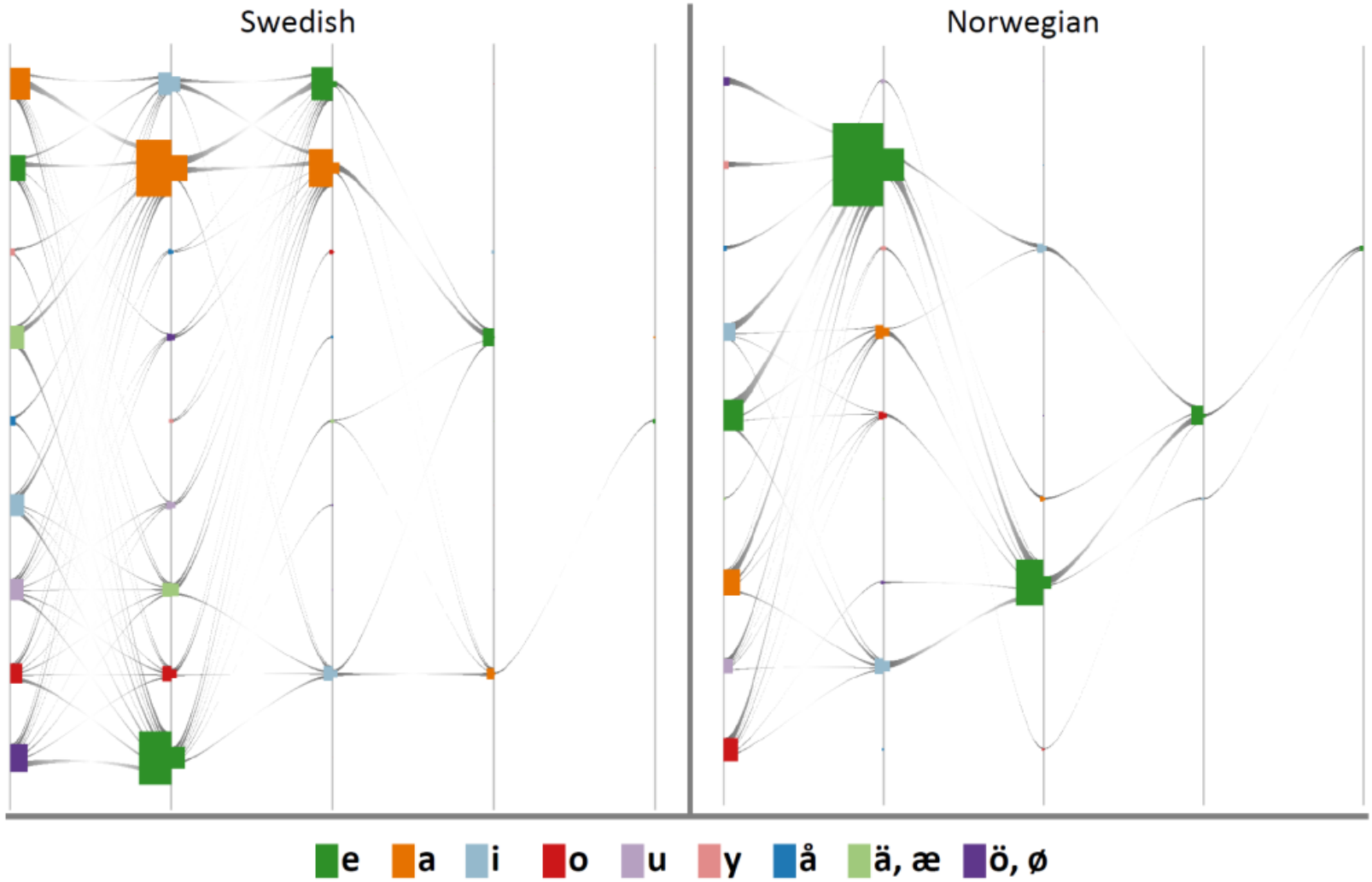
Vowel Harmony via Droplets

kaşık-lar-ım-a
spoon-Pl-1SgPoss-Dat
'my spoons'

kedi-ler-im-e
cat-Pl-1SgPoss-Dat
'my cat'



Language Comparison via Droplets



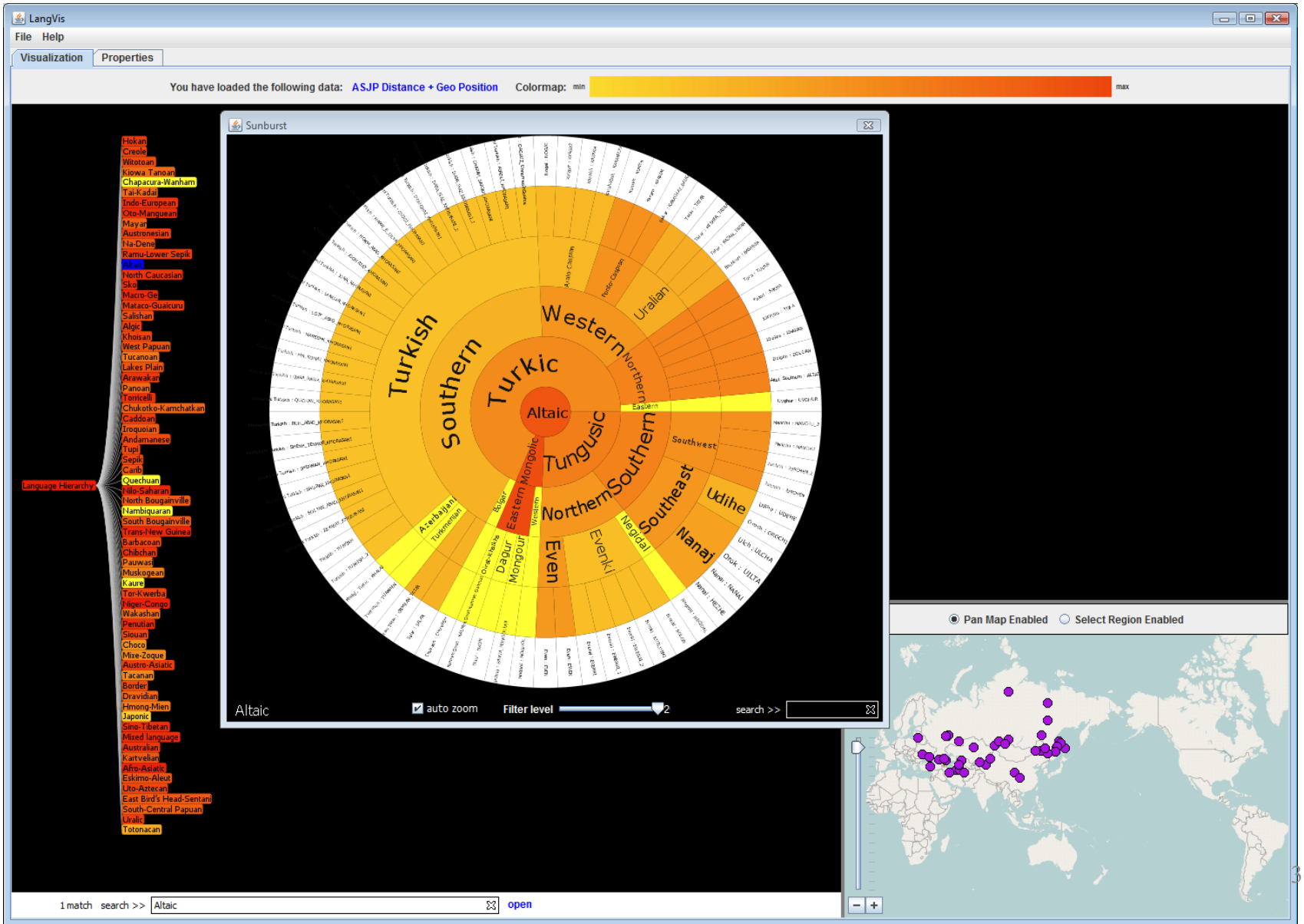
Norwegian shows language change a ➔ e in comparison to Swedish.

Example: Sunburst and maps

- 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).

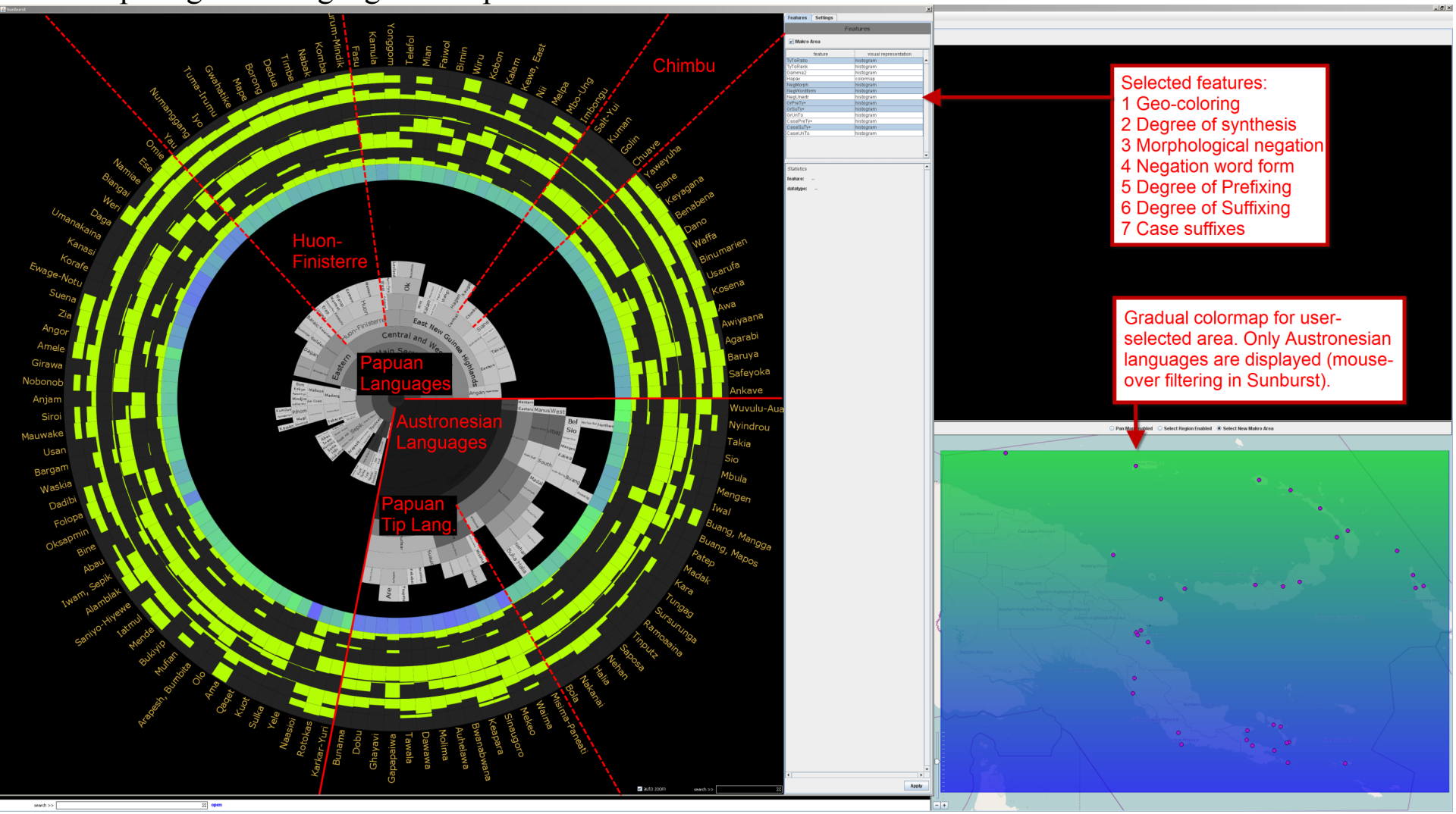
Christian Rohrdantz, Michael Hund, Thomas Mayer, Bernhard Wälchli and Daniel A. Keim. 2012. The World's Languages Explorer: Visual Analysis of Language Features in Geneaologica and Areal contexts. *Computer Graphics Forum* 31(3), 935-944.

Sunburst and Maps for Language Families



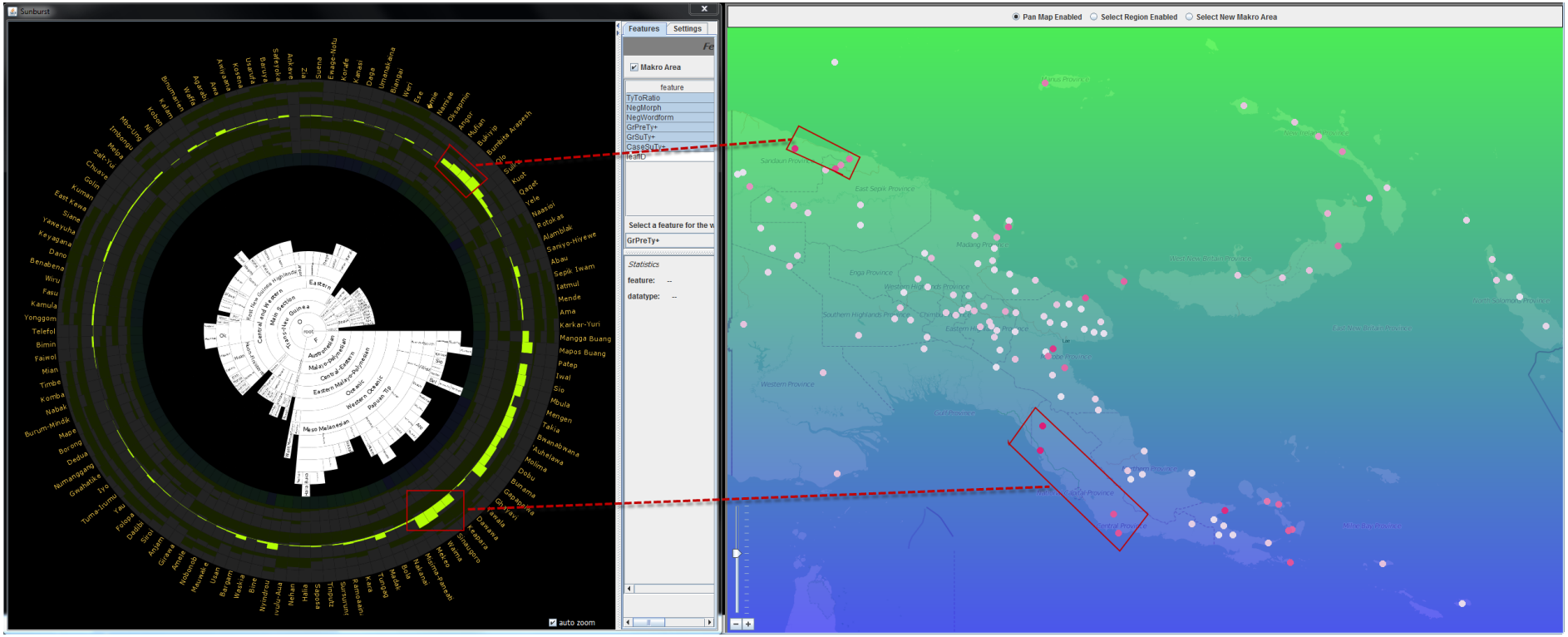
World's Languages Explorer

Comparing 126 Languages of Papua New-Guinea based on the New Testament.



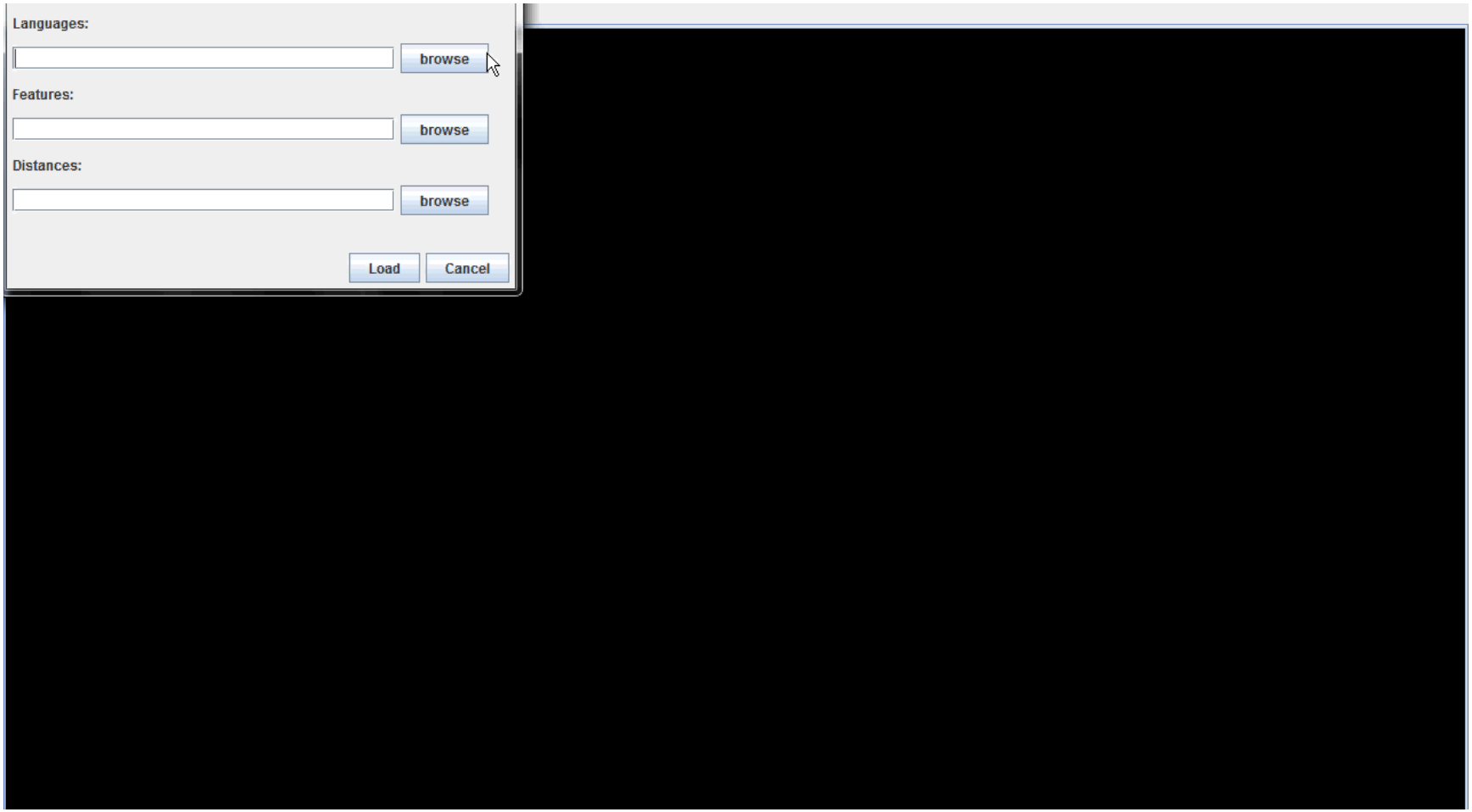
Each circle segment represents one language, each ring the values of one feature across all languages.

World's Languages 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.

Interaction



Sorting and Pattern Discovery

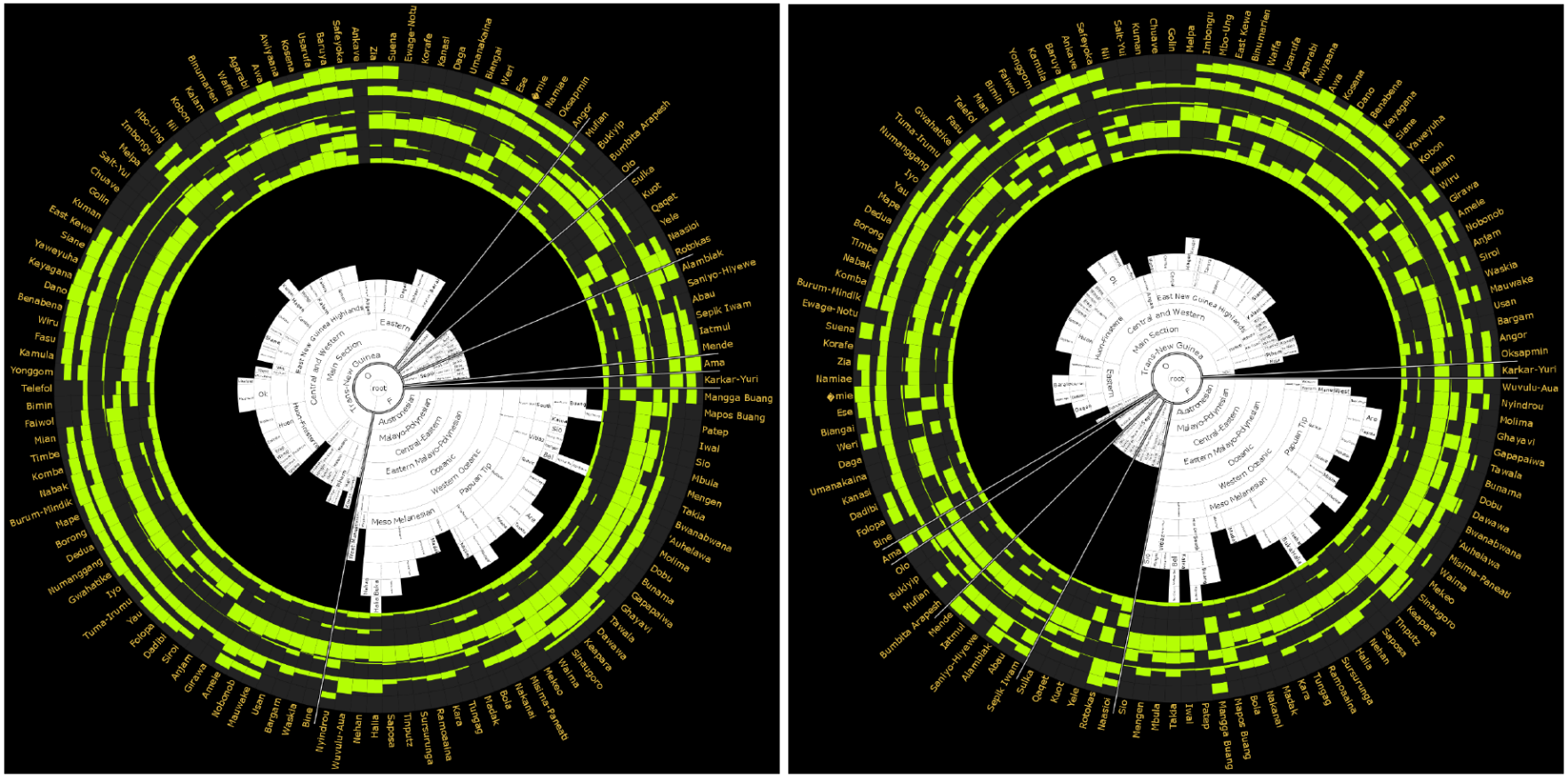
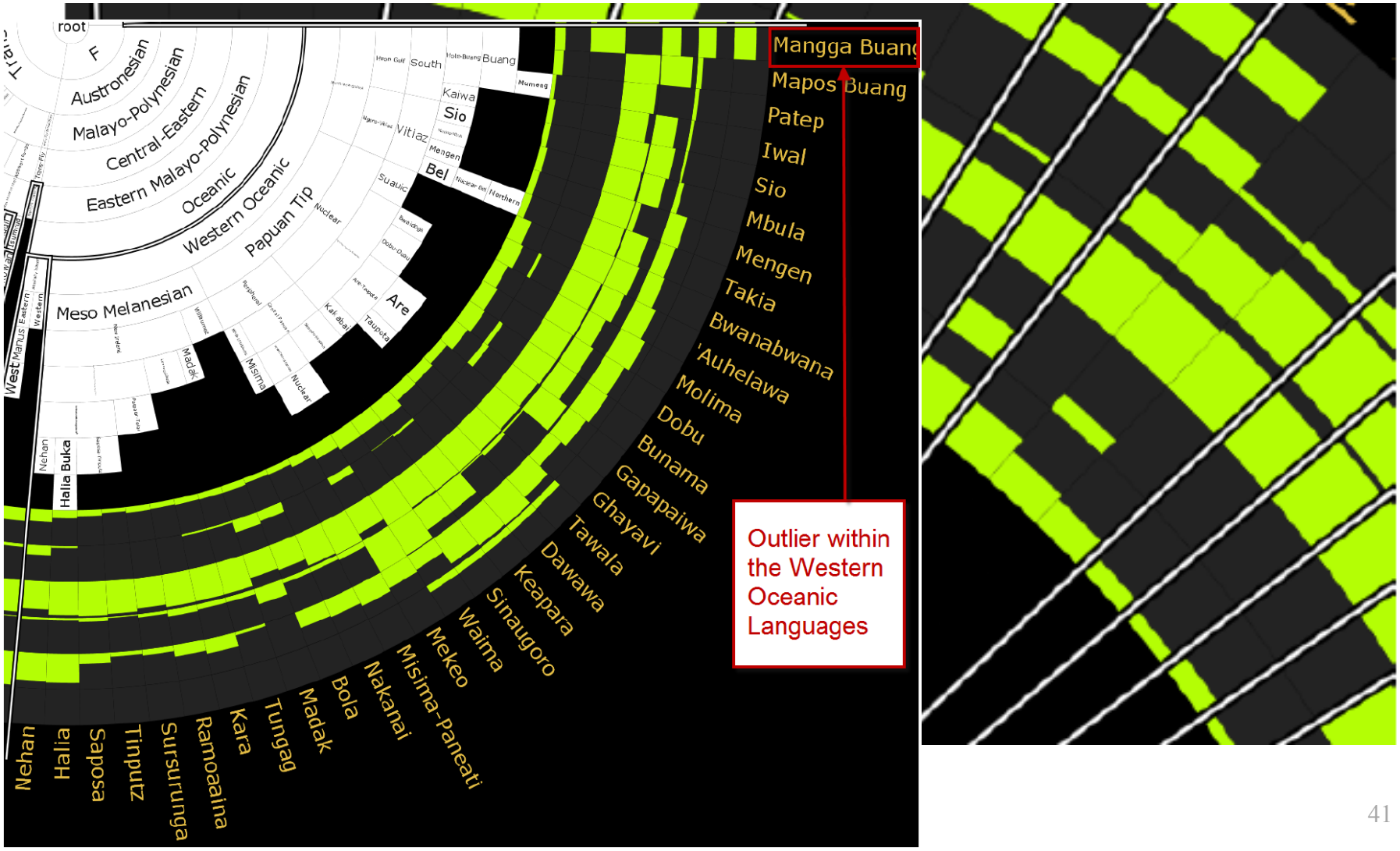


Figure 2.5: High-resolution screenshot showing automatically extracted features for languages from Papua New Guinea with leaves ordered to *maximize* (left) and *minimize* (right) the pairwise leaf similarity for neighbors.

Sorting and Pattern Discovery



WALS Explorer

- We will be working with a version that is tailored to interact with WALS.
- http://www.th-mayer.de/wals/

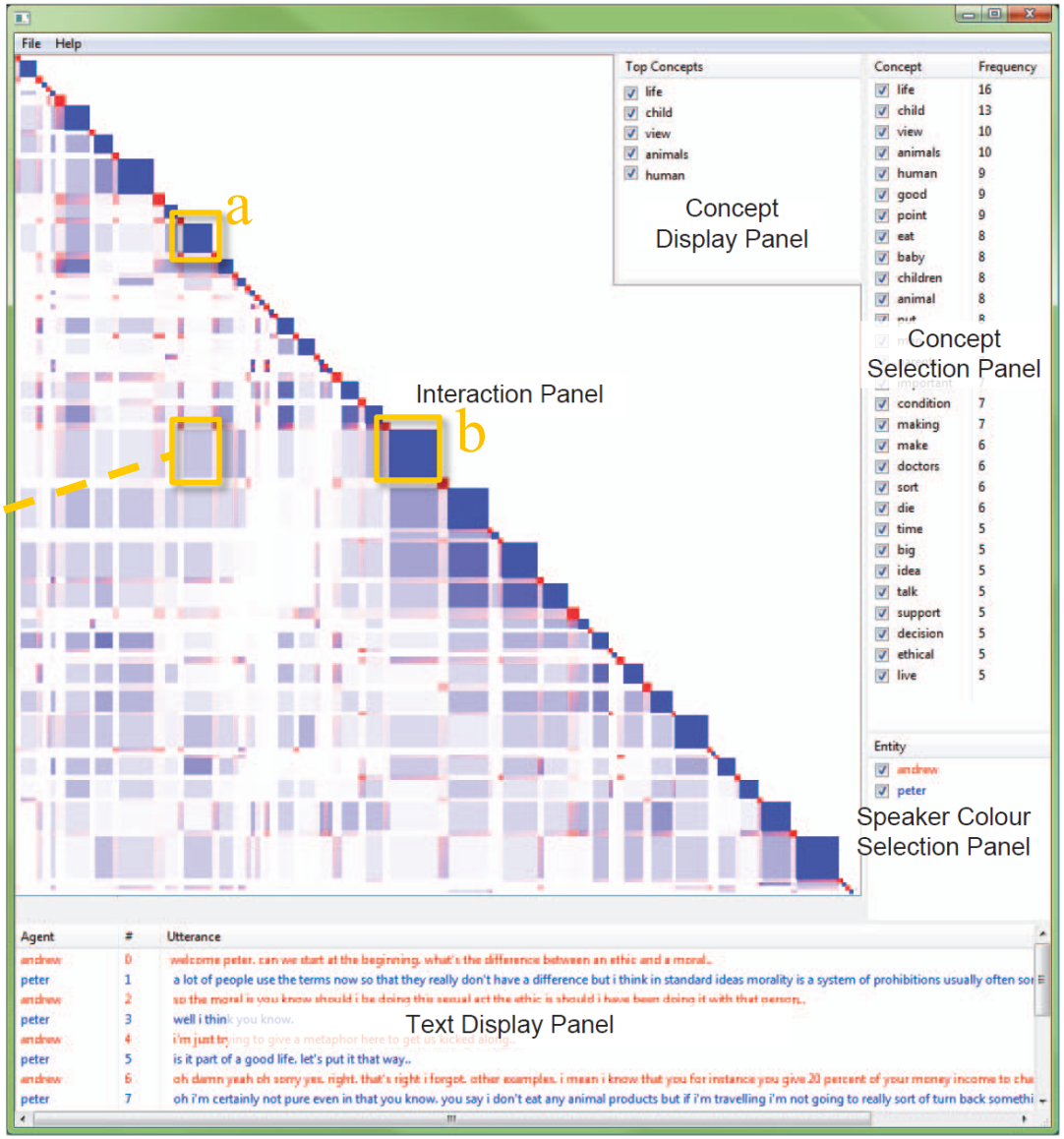
Thomas Mayer, Bernhard Wälchli, Christian Rohrdantz and Michael Hund. 2014. From the extraction of continuous features in parallel texts to visual analytics of heterogeneous areal-typological datasets. In B. Nolan and C. Periñán-Pascual (eds.), *Language Processing and Grammars: The role of functionally oriented computational models*, 13–38. John Benjamins.

Conceptual Recurrence Plots

- Another type of much studied language data: discourses.
- The context of social media (Twitter, Facebook, etc.) presents us with new opportunities but also with new challenges.
- Next up: visual analysis of a (conventional) dialog – an interview.

Daniel Angus, Andrew E. Smith, Janet Wiles: **Conceptual Recurrence Plots: Revealing Patterns in Human Discourse**. *IEEE Trans. Vis. Comput. Graph.* 18(6): 988-997 (2012)

Discursis



Saturation shows how much b relates to a (content-wise)

Discursis

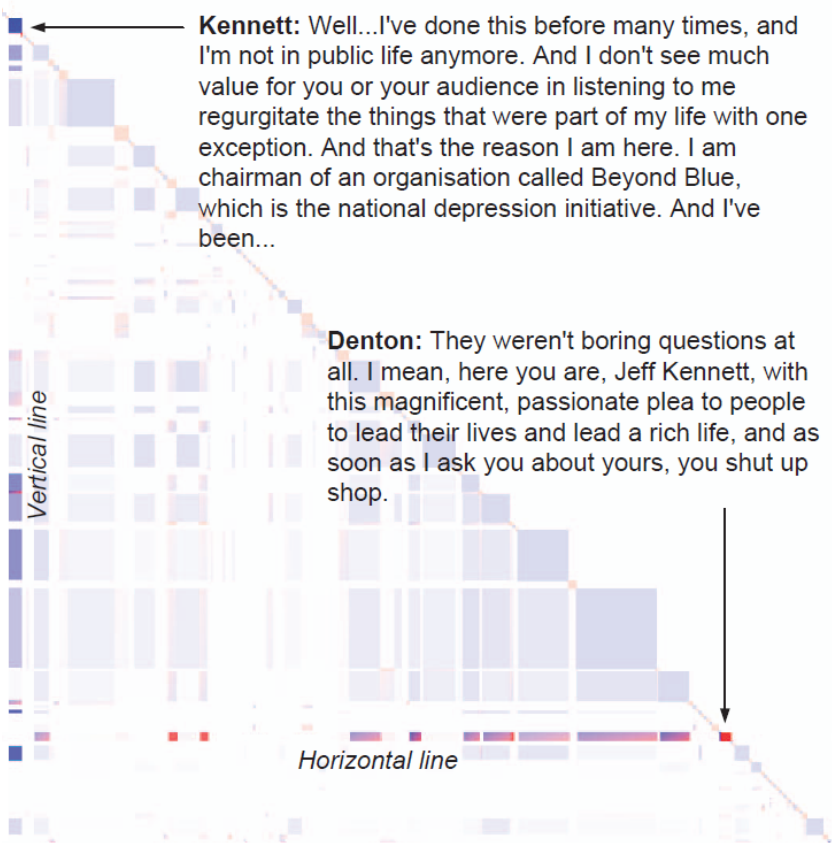


Fig. 7: Recurrence Lines (*Denton/Kennett*)

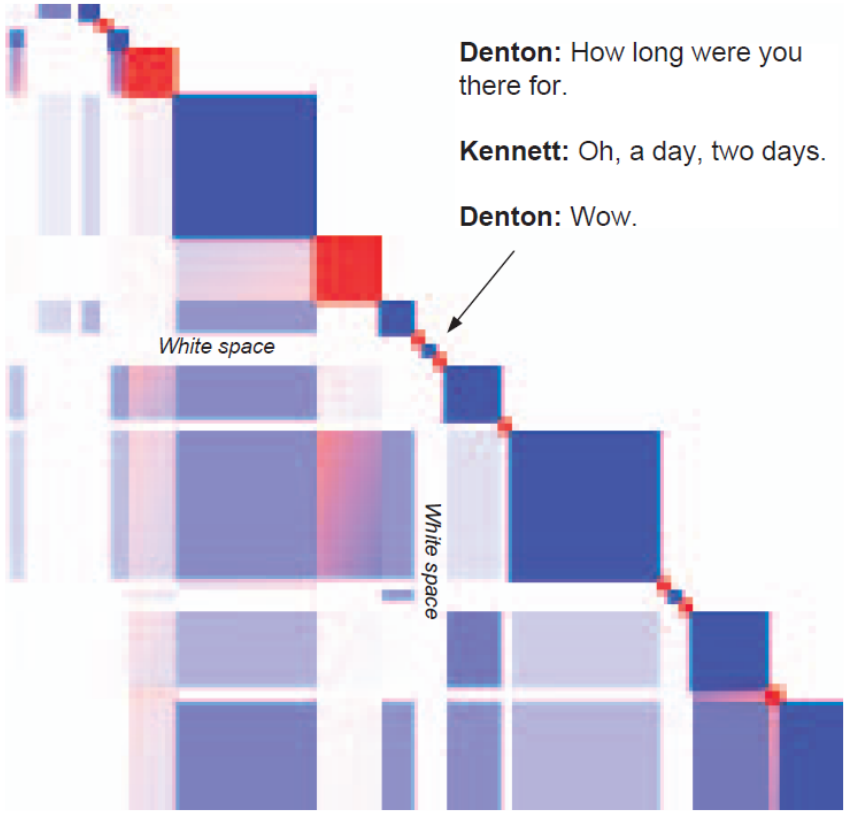


Fig. 8: White space (*Denton/Kennett*)

Discursis

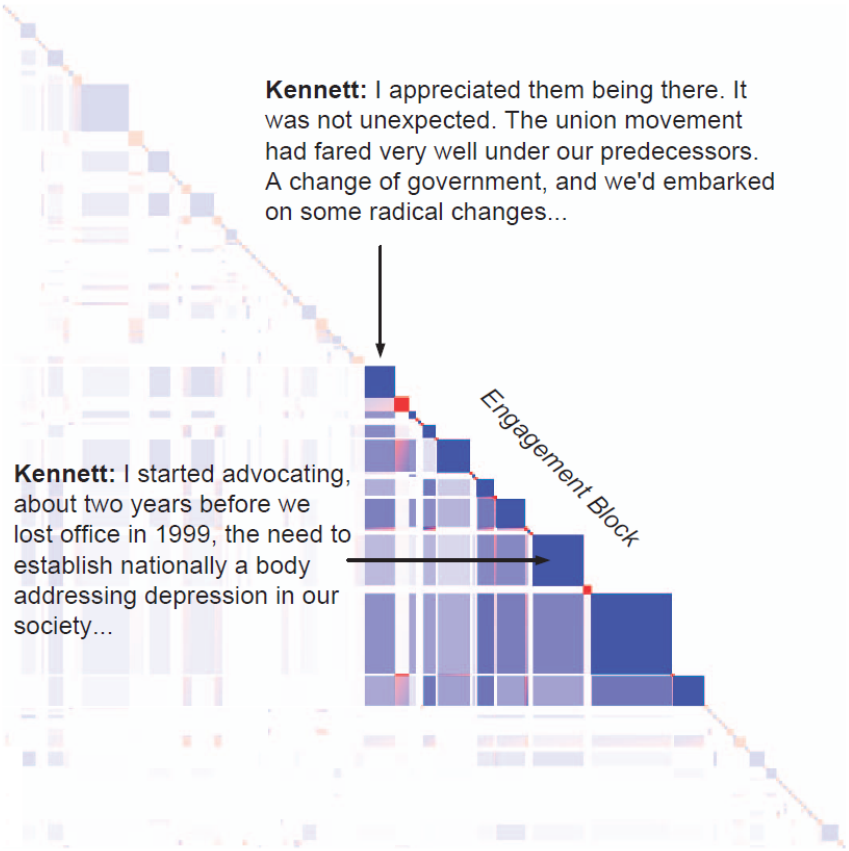


Fig. 9: Engagement Block (*Denton/Kennett*)

Fig. 10: Random Scattering (*Denton/Singer*)

Discursis

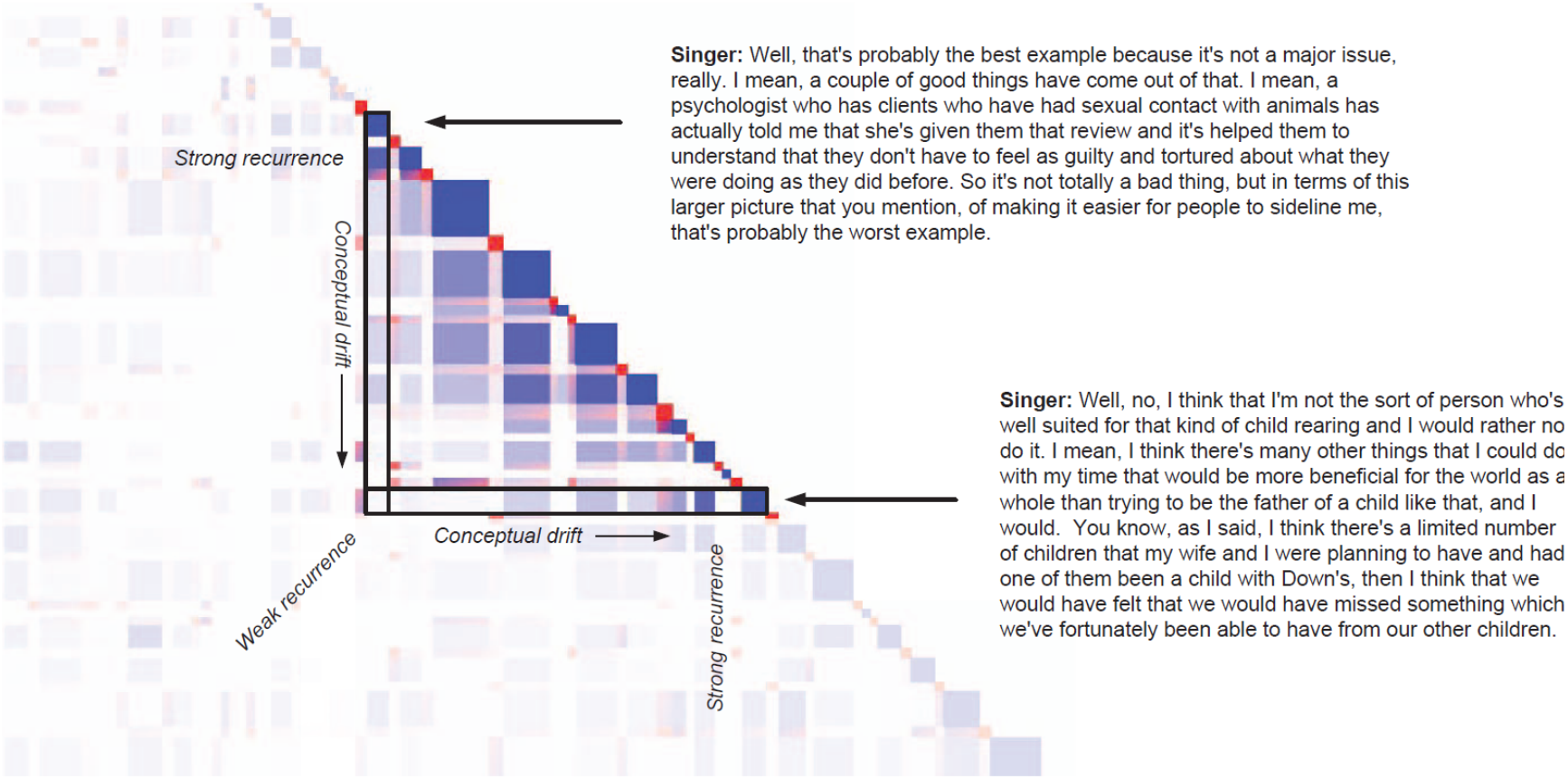


Fig. 11: Concept Drift (*Denton/Singer*)

Summary

- Have seen examples 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 tomorrow.
- Now first some design basics.