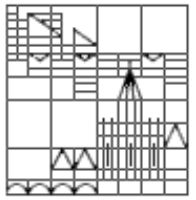
SPONSORED BY THE
Federal Ministry
of Education
and Research

Universität
Konstanz



Visual Analytics for Linguists

Miriam Butt & Chris Culy

ESSLII 2014, Introductory Course
Tübingen

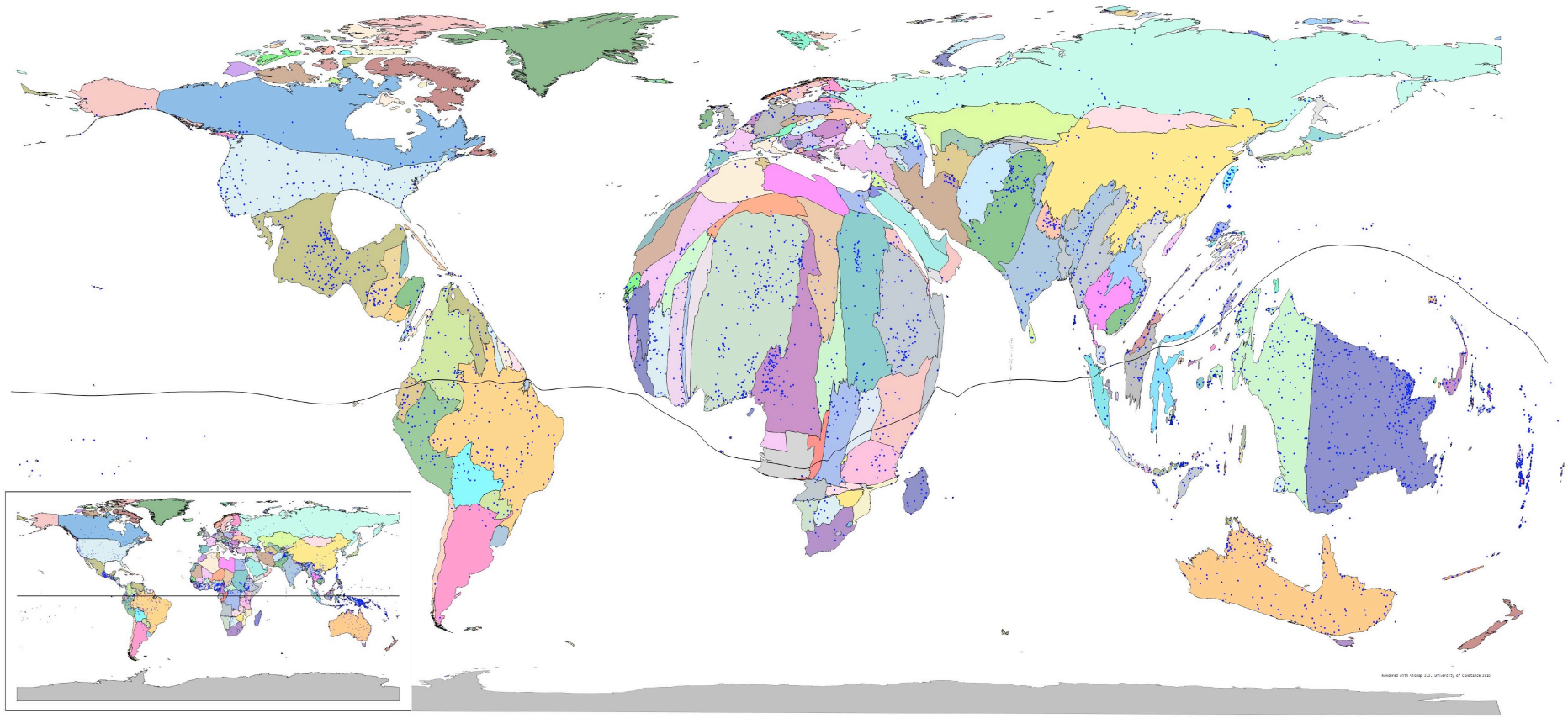
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

PHILOSOPHISCHE
FAKULTÄT
Seminar für Sprachwissenschaft

Day 2 – More on LingVis

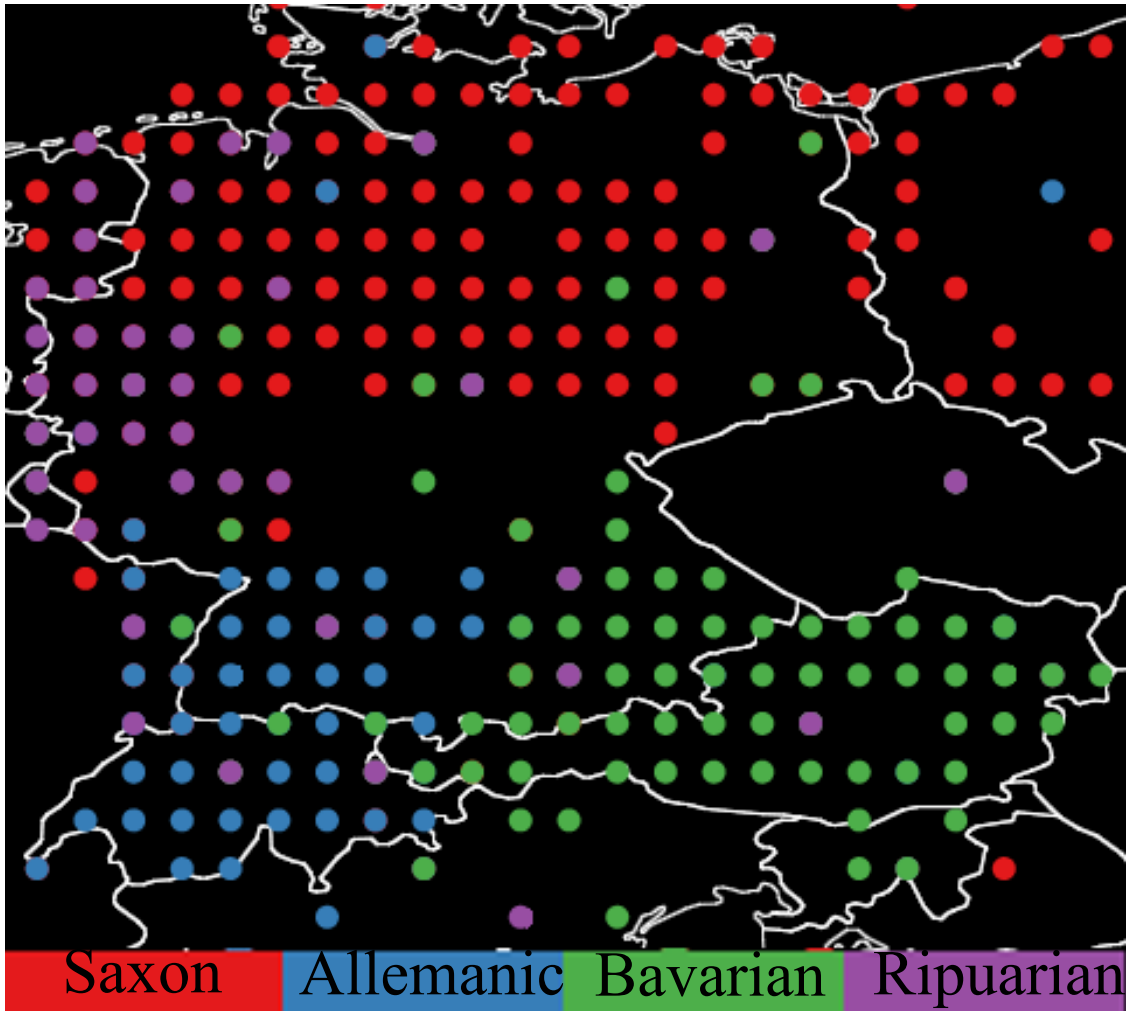
1. More Use Cases
2. Critical Discussion
   - Are the visualizations successful?
   - Are the visualizations useful?
3. What kinds of visualizations would you like?

Distorted Map according to number of languages spoken in area.



Note: visualization only as good as your data –
India massively underrepresented

Wikipoint-Analysis: Dialects



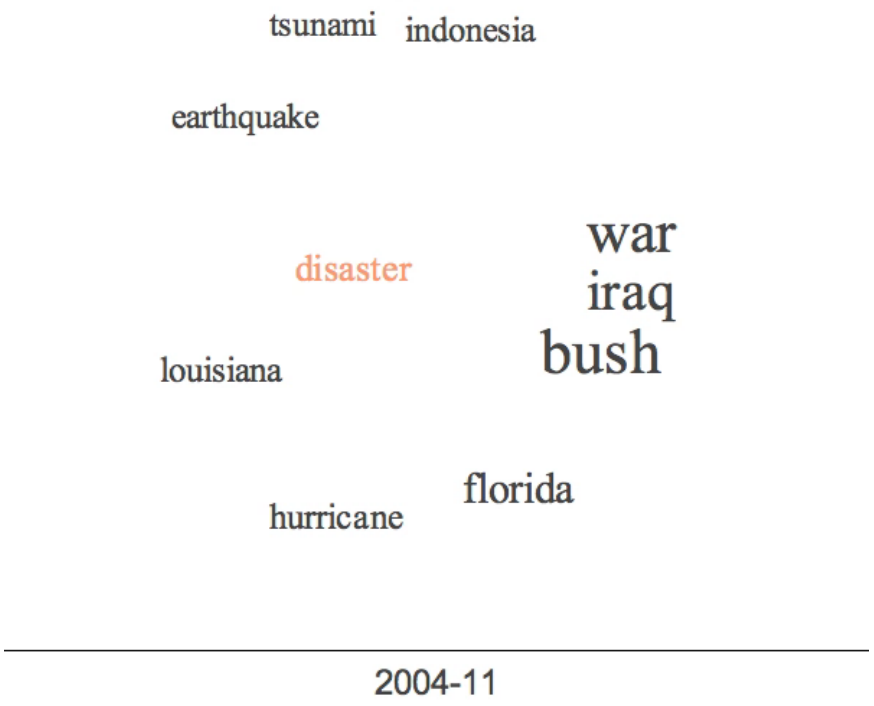
(Work Group: Daniel Keim, Uni Konstanz)

Using Motion

- highly frequent words in New York Times articles 2004-2005 and their relation to one another
- show trends/change

Challenges for Visualization:

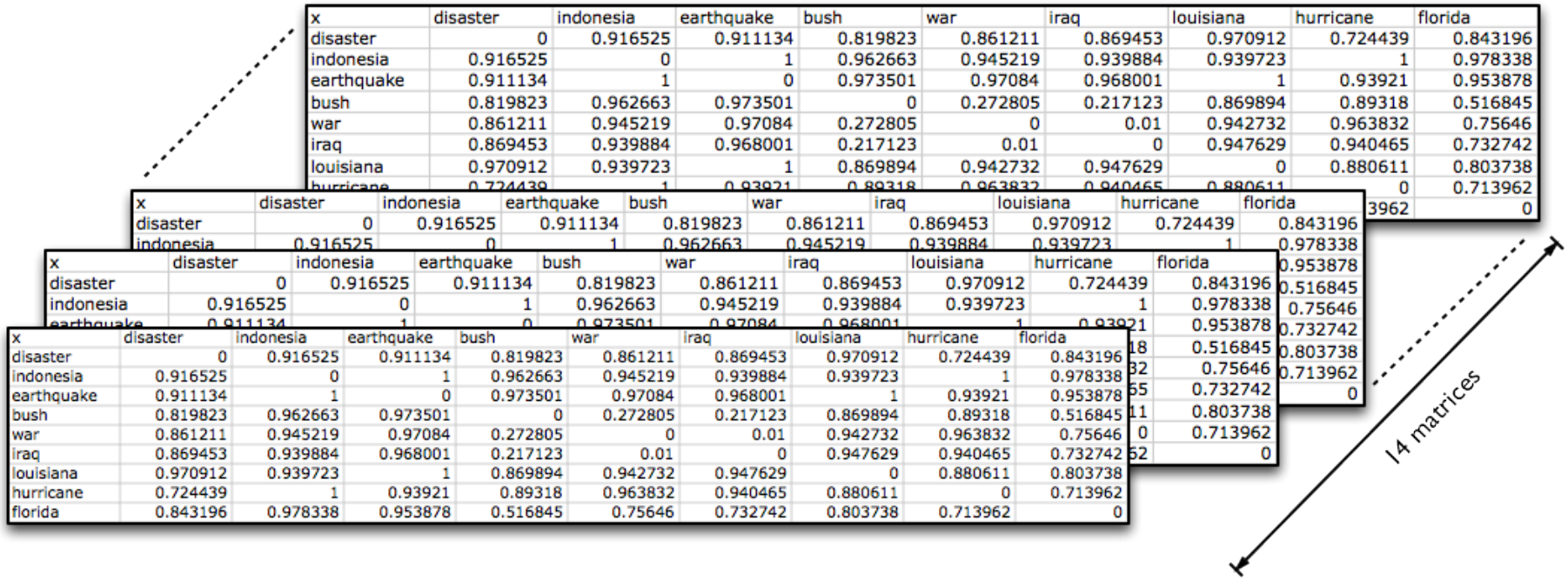
- **dimensionality reduction**: high dimensional distance matrices are shown in 2D
- **precision vs. stability**: a precise visualization for each time step would induce too much confusing movement



(Work Group: Oliver Deussen, Uni Konstanz)

Example: Animated Visualization

- the raw data without visualization:
  - 9x9 distance matrices for each of the 14 time steps



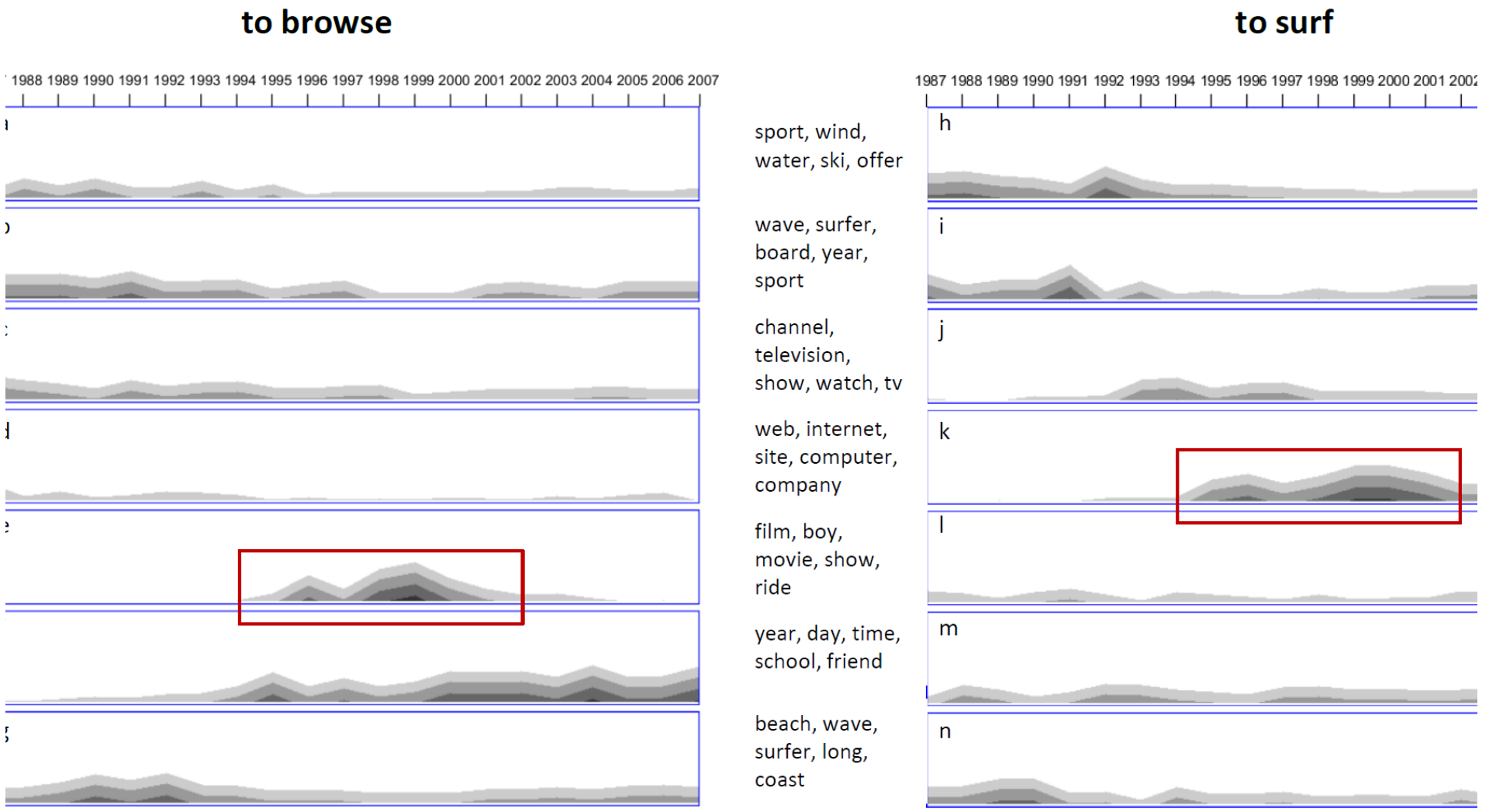
x	disaster	indonesia	earthquake	bush	war	iraq	louisiana	hurricane	florida
disaster	0	0.916525	0.911134	0.819823	0.861211	0.869453	0.970912	0.724439	0.843196
indonesia	0.916525	0	1	0.962663	0.945219	0.939884	0.939723	1	0.978338
earthquake	0.911134	1	0	0.973501	0.97084	0.968001	1	0.93921	0.953878
bush	0.819823	0.962663	0.973501	0	0.272805	0.217123	0.869894	0.89318	0.516845
war	0.861211	0.945219	0.97084	0.272805	0	0.01	0.942732	0.963832	0.75646
iraq	0.869453	0.939884	0.968001	0.217123	0.01	0	0.947629	0.940465	0.732742
louisiana	0.970912	0.939723	1	0.869894	0.942732	0.947629	0	0.880611	0.803738
hurricane	0.724439	1	0.93921	0.89318	0.963832	0.940465	0.880611	0	0.713962
florida	0.843196	0.978338	0.953878	0.516845	0.75646	0.732742	0.803738	0.713962	0

14 matrices

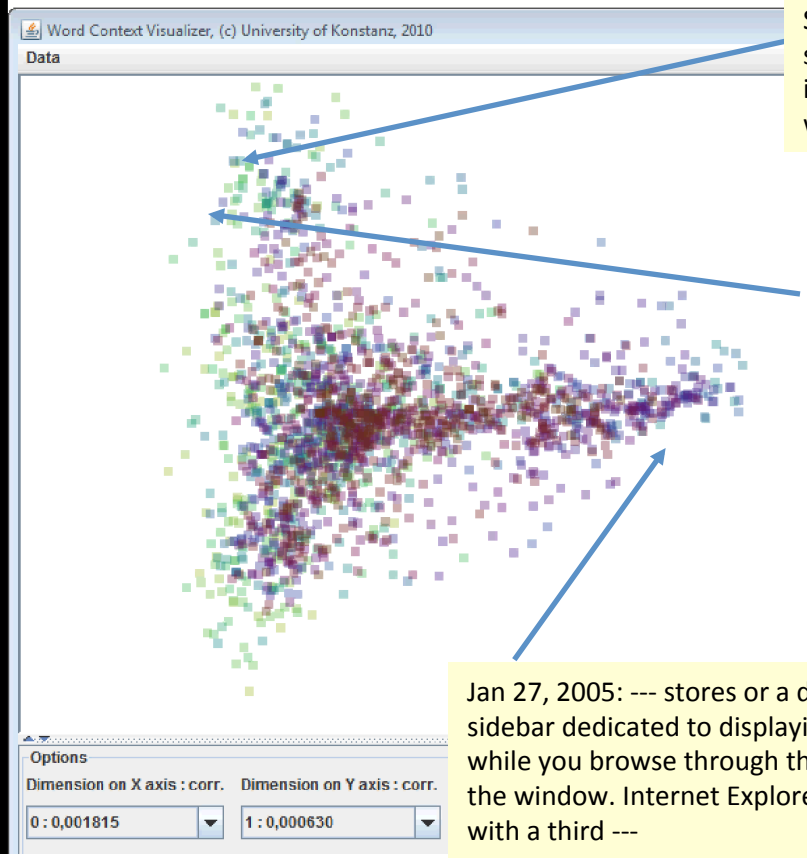
Tracking Lexical Change

- Looked at changes in usage over time via various visualization methods.
- Data:
  - New York Times Annotated Corpus
  - 1.8 million articles from daily newspaper editions 1987-2008
  - particularly: *to browse* vs. *to surf*
- Frequency development of different word senses
- Automatically induced from word contexts with standard Latent Dirichlet Allocation (LDA) topic modelling

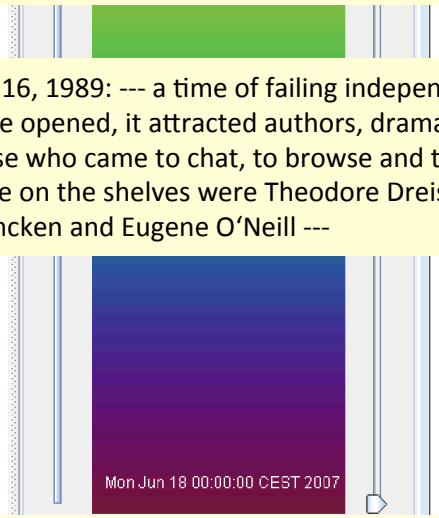
Tracking Lexical Change



Word Context Visualization for *browse* (NYT 1987-2007)



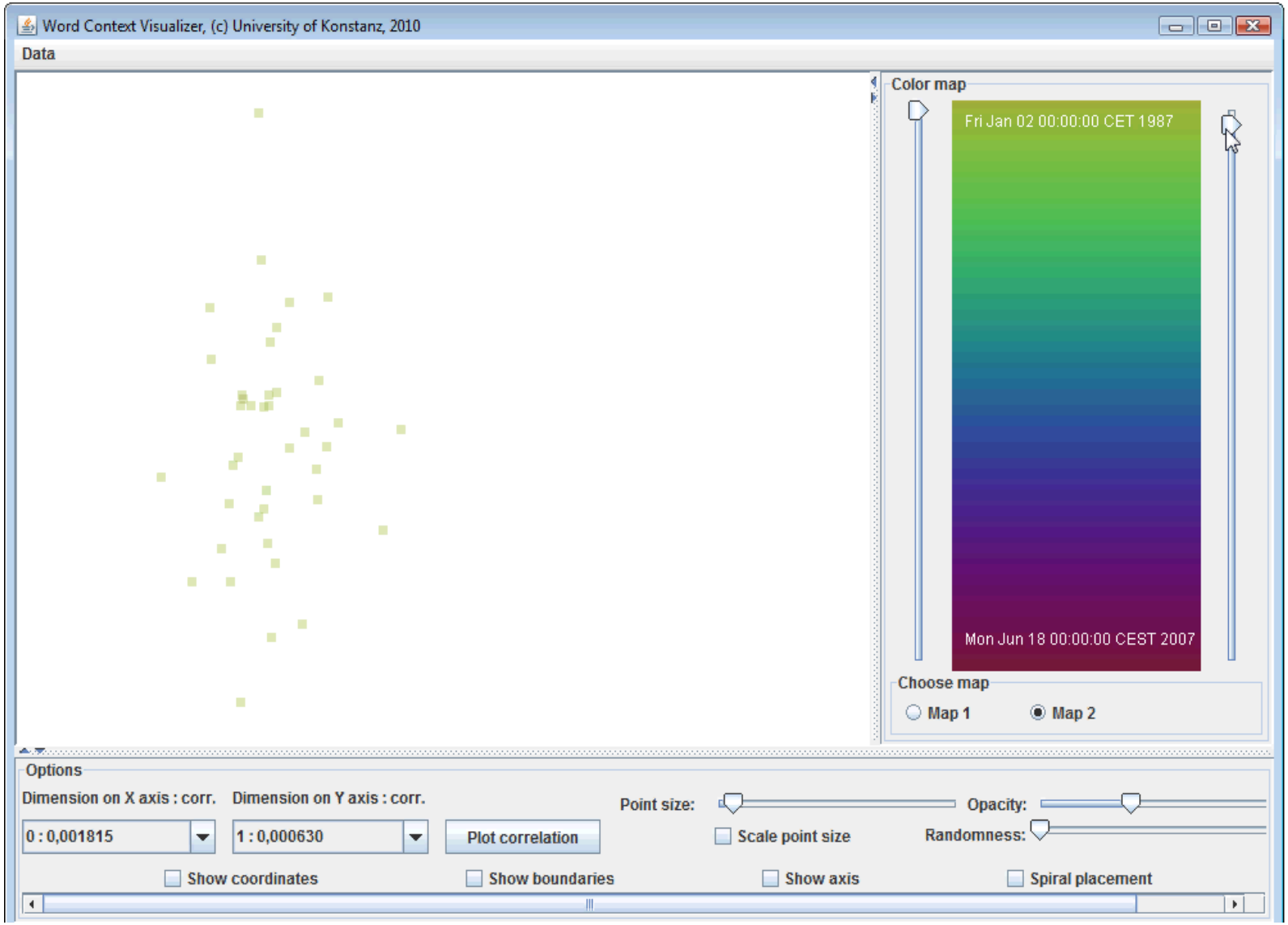
Sep 27, 1992: --- abound in Wiesbaden's pedestrian zone and around the shops within walking distance of the Casino. At the fair, the reading public is invited to browse, but not buy, as none of the books are for sale, on the weekend of Oct. 3 and 4, for a $7 admission. Last year ---



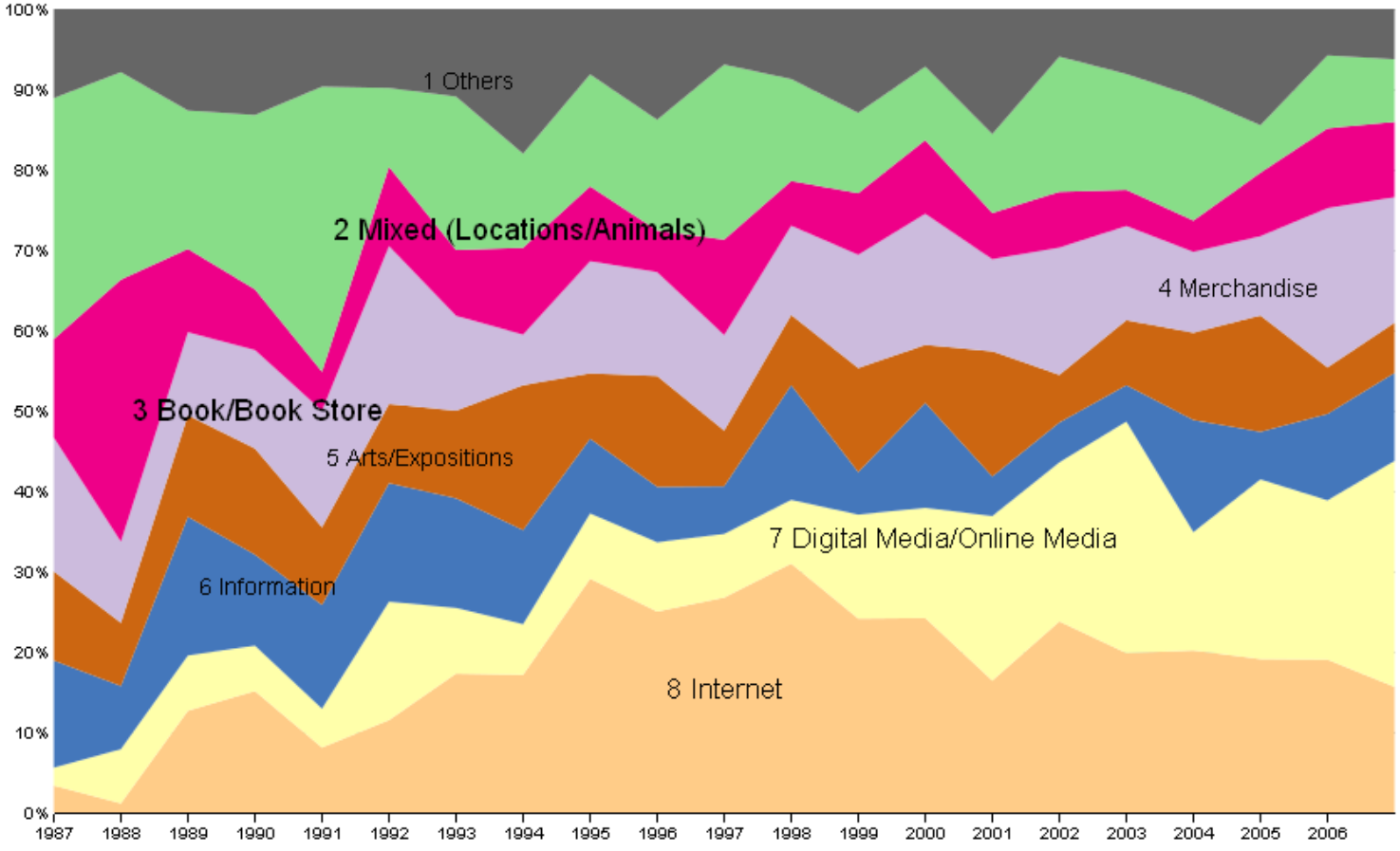
Apr 16, 1989: --- a time of failing independent bookstores. Soon after the store opened, it attracted authors, dramatists, poets and artists. Among those who came to chat, to browse and to see if their books and plays were on the shelves were Theodore Dreiser, John Dos Passos, H.L. Mencken and Eugene O'Neill ---

Jan 27, 2005: --- stores or a direct search of Amazon. Mozilla has a special sidebar dedicated to displaying the search results where you can see them while you browse through the recommended Web sites in the main part of the window. Internet Explorer users can get some of the same capabilities with a third ---

Word Context Visualization for *browse* (NYT 1987-2007)



Diachronic Development of Different Topics/Concepts in the Context of *browse* (NYT 1987-2007)



Topic/Concept Descriptors:

place city long collection **deer** antique high main hour

book buy street open include **read** small free public

mr day year **offer good customer** visit **sale** start

shop **people visitor art show line gallery display museum**

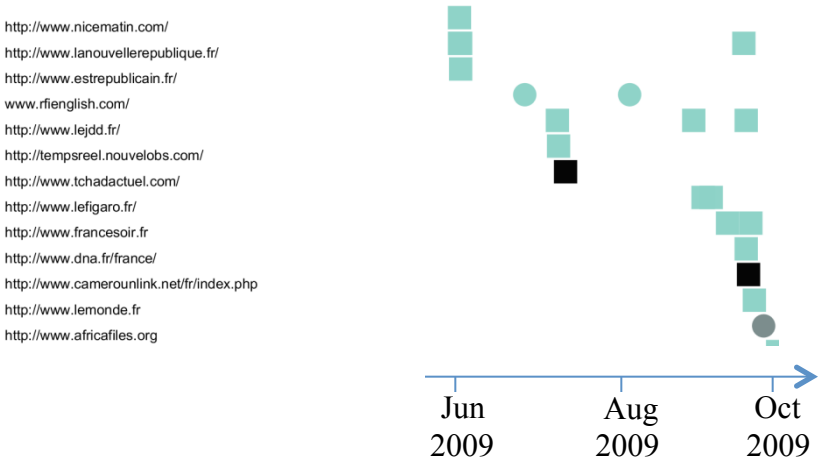
store **time find** home **information call** shopper sell **library**

site make work **page online list search music click**

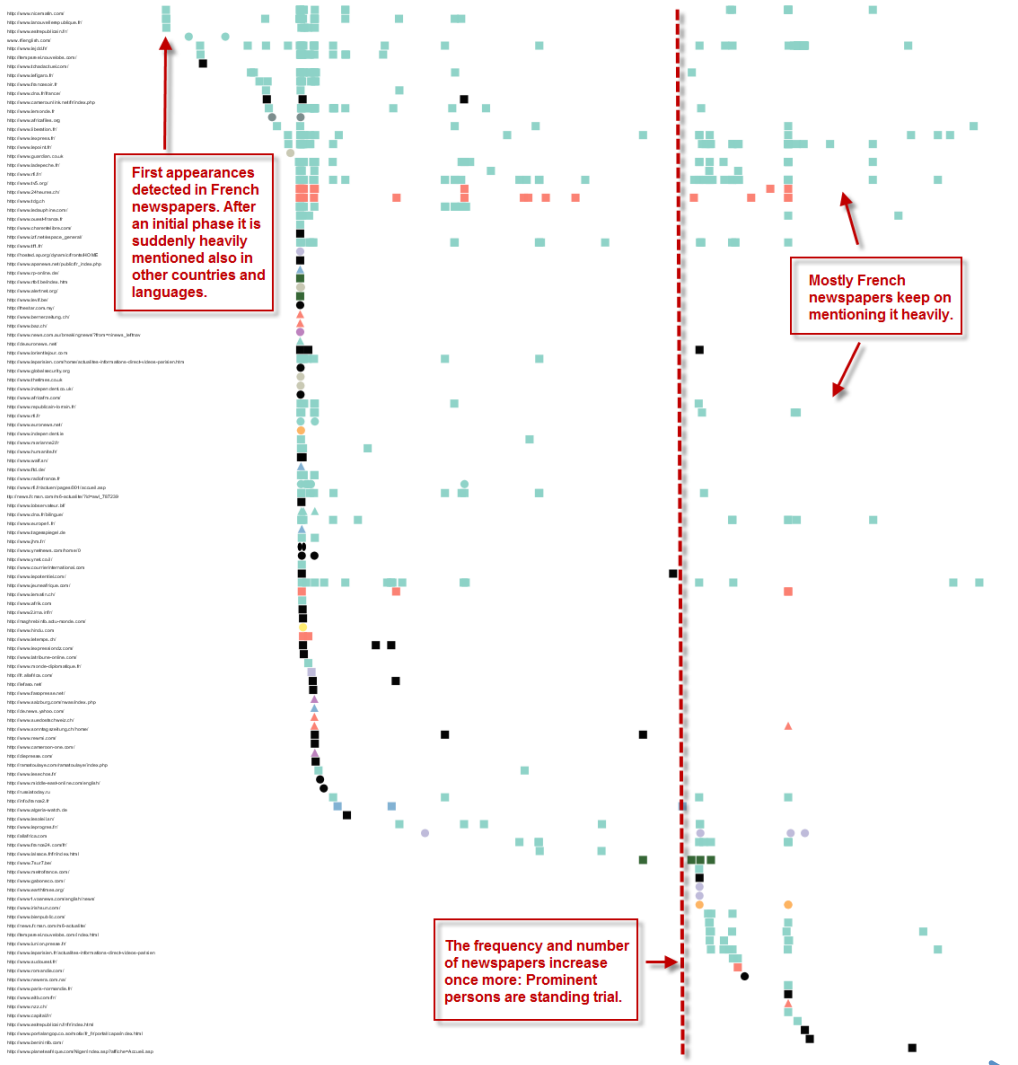
web internet computer user company **mail software service** market

Spread of a New Suffix (-gate) & International News Dynamics

Angolagate extracted from app. 11 million online news articles in English, German, and French between May 2009 and January 2012 [4].



English	Great Britain, USA, Ireland, Pakistan, India, Australia, Canada
French	France, Switzerland, Belgium
German	Germany, Switzerland, Austria



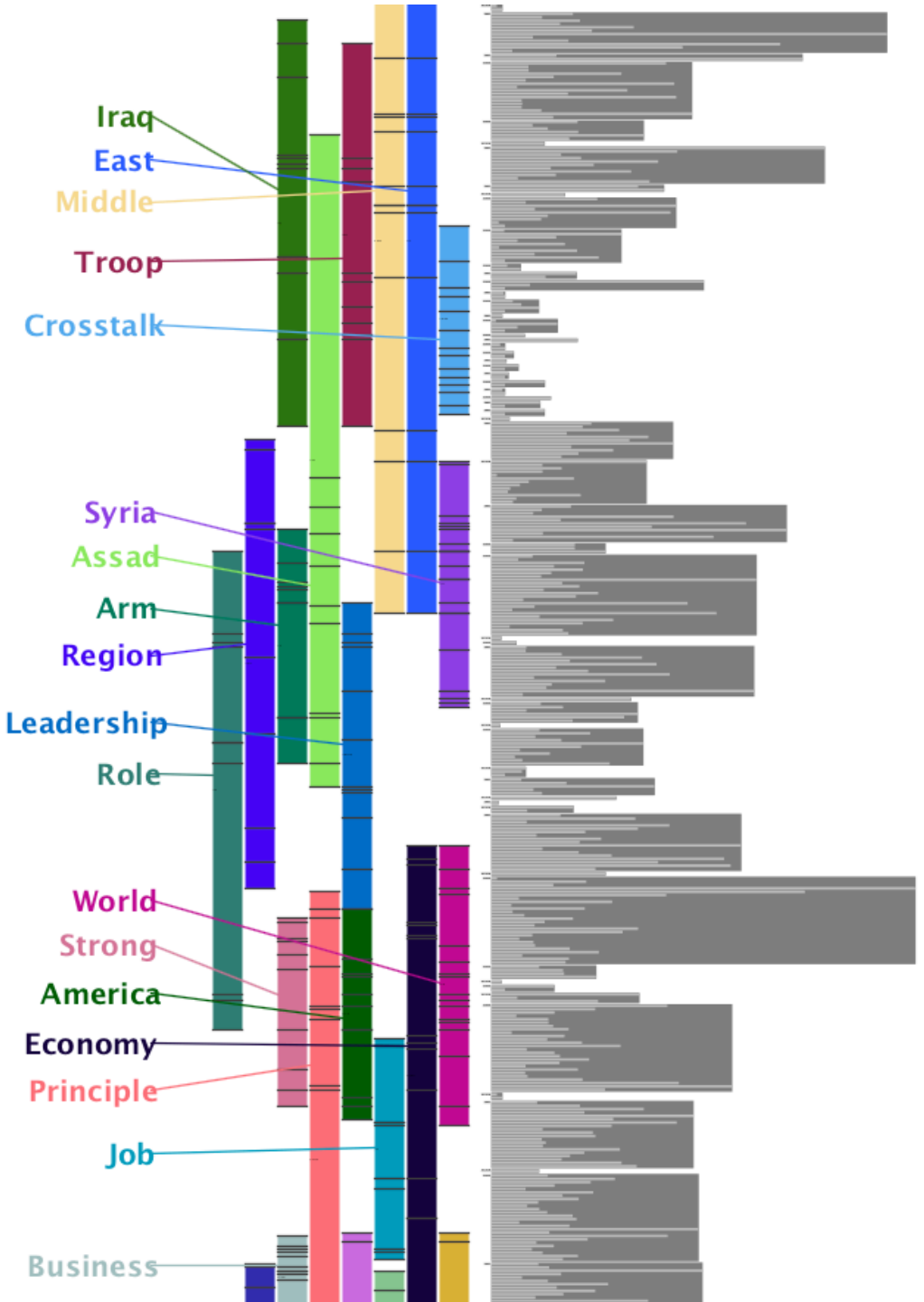
(C. Rohrdantz, Dissertation, Uni Konstanz)

Lexical Episodes

- (Pixel) Visualization of what is under discussion in a stretch of dialog
  - Words that occur more often than expected in a given stretch of text are highlighted.
  - The distance between instances of a word within an episode is smaller than the expected distance with respect to the entire corpus.
- Example: 3rd presidential debate between Barack Obama and Mitt Romney (Oct. 2012)

Lexical Episodes:

- Each grey box is a turn
- Each word has a color
- Interaction possible (mouse over, zooming)

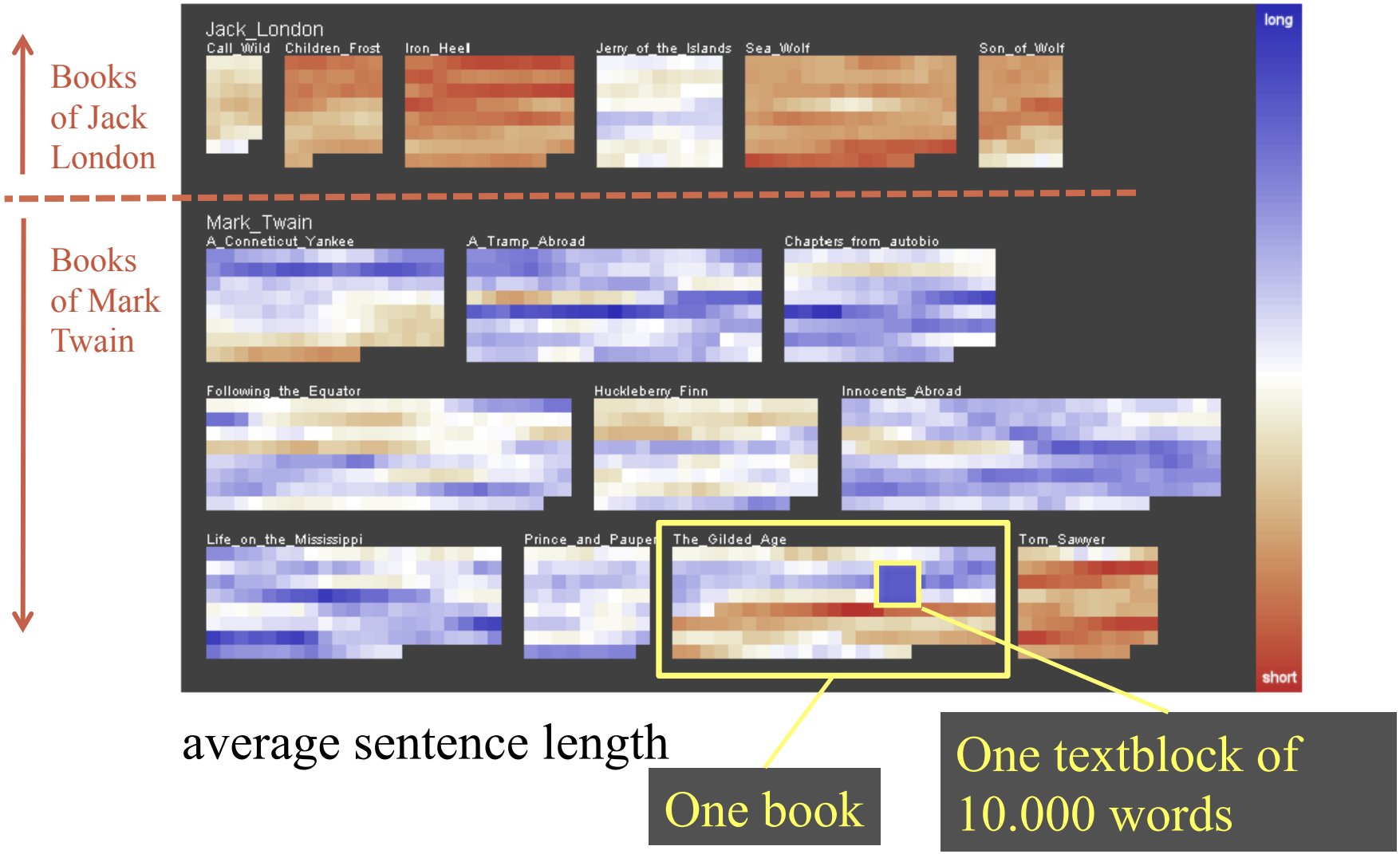


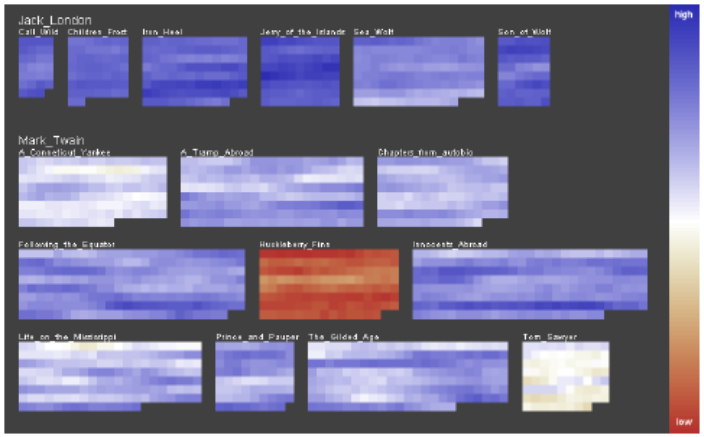
(Christian Rohrdantz. 2014)

Use Cases – Literary Analysis (More Pixel Visualizations)

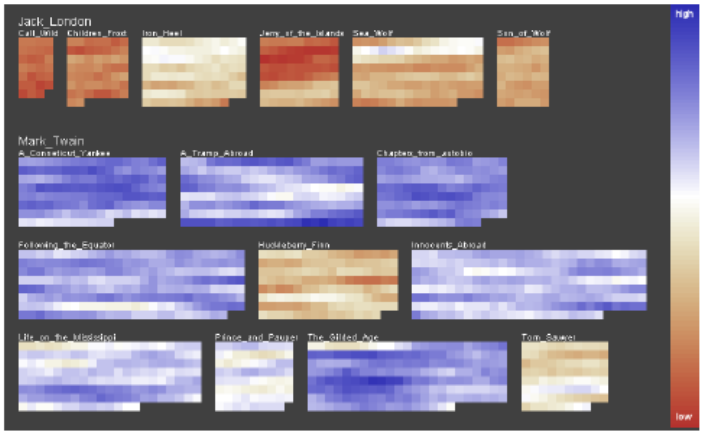
Daniel A. Keim, Daniela Oelke: **Literature Fingerprinting: A New Method for Visual Literary Analysis**. *IEEE VAST 2007*: 115-122

Authorship Attribution

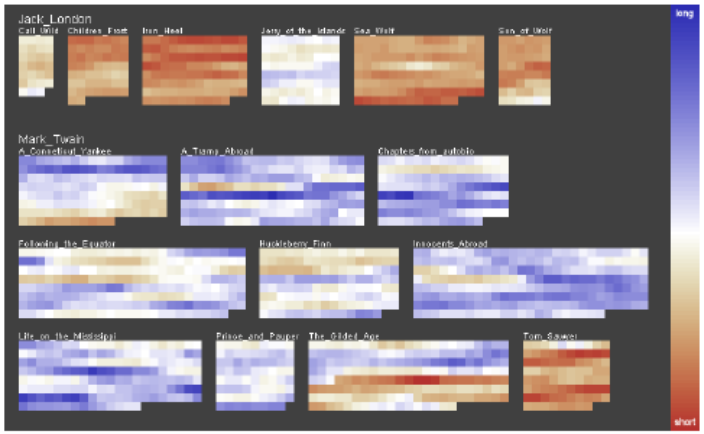




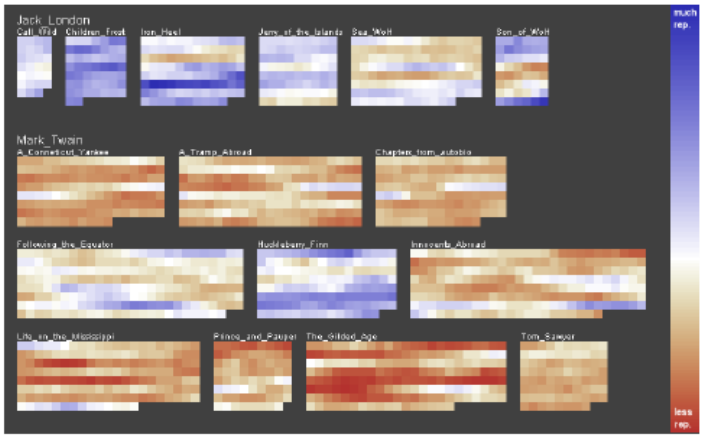
(a) Function words (First Dimension after PCA)



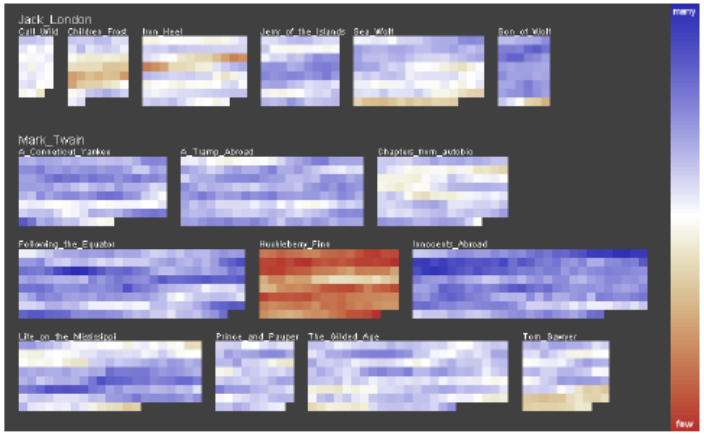
(b) Function words (Second Dimension after PCA)



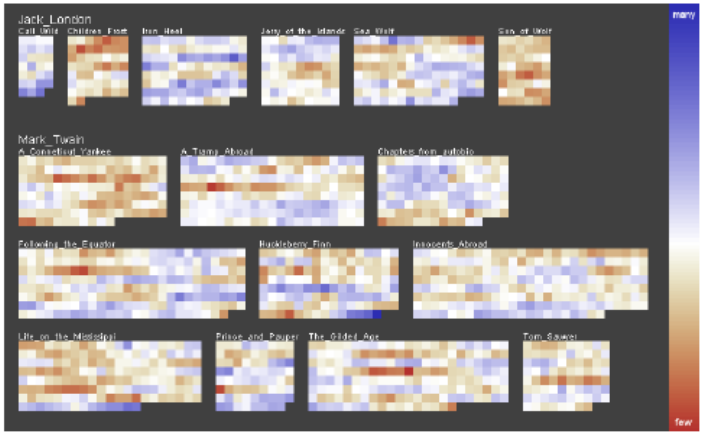
(c) Average sentence length



(d) Simpson's Index



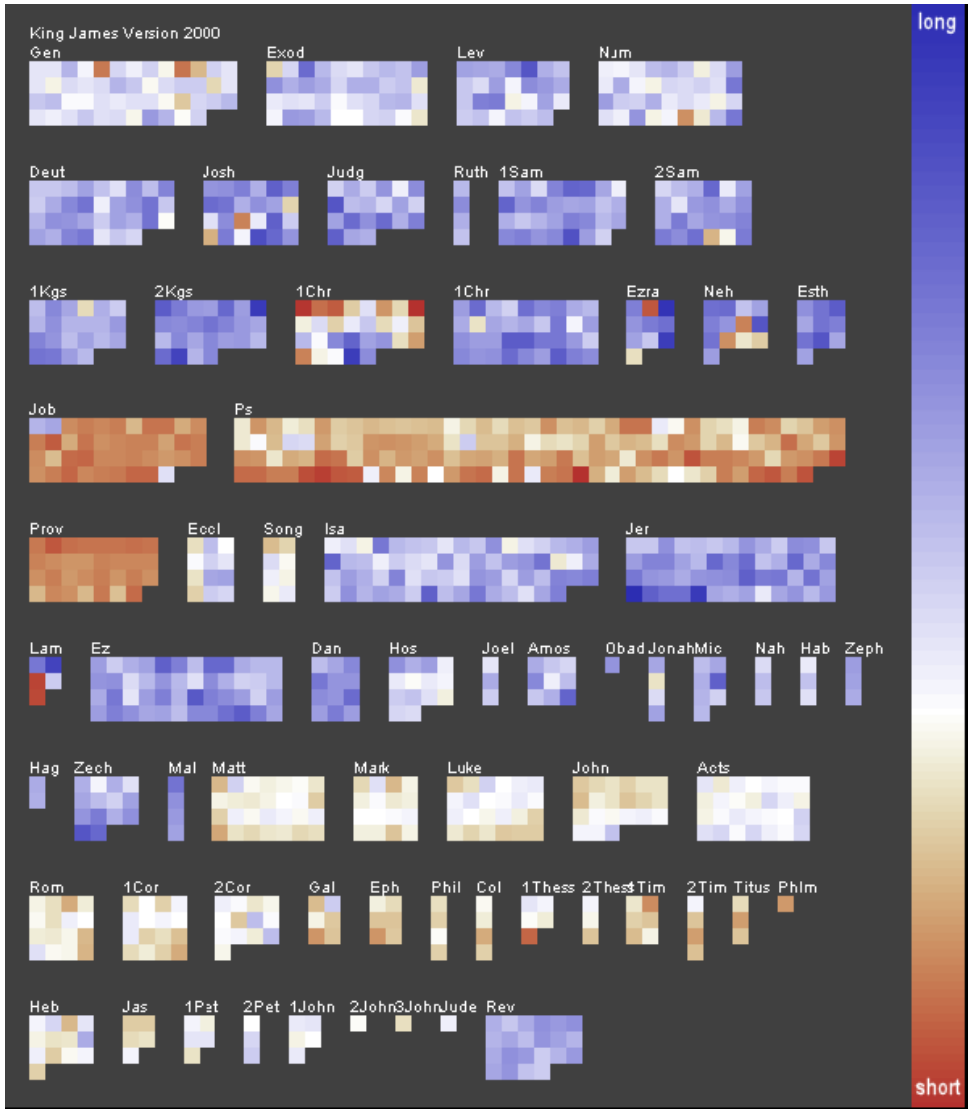
(e) Hapax Legomena



(f) Hapax Dislegomena

[1

Literature Fingerprinting on the Bible



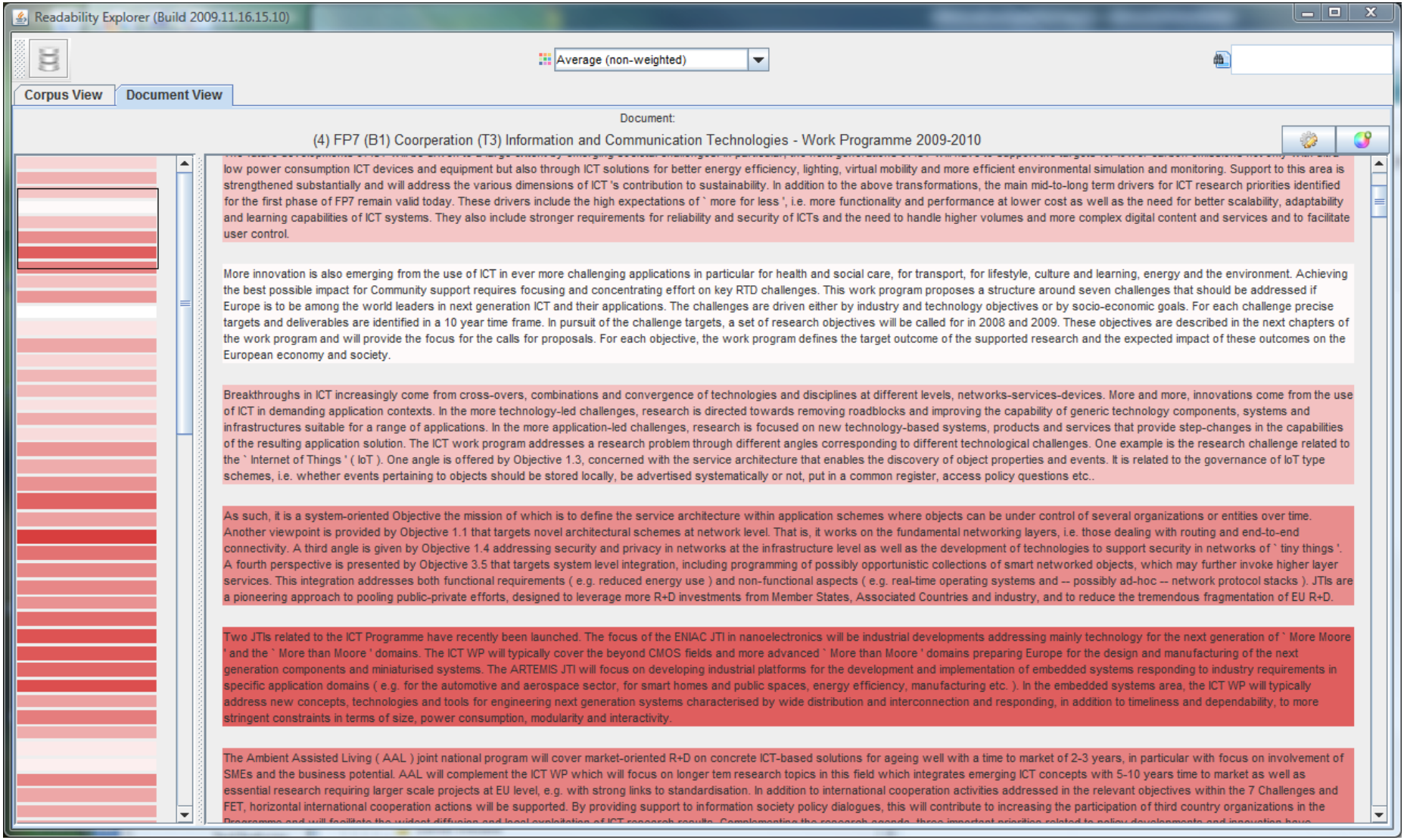
average sentence length


Job 16-26
Neh 10
Num 7
Gen
Exod
Lev
Num
Deut
Josh
Judg
Ruth
1Sam
2Sam
1Kgs
2Kgs
1Chr
1Chr
Ezra
Neh
Esth
Job
Ps
Prov
Eccl
Song
Isa
Jer
Lam
Ez
Dan
Hos
Joel
Amos
Obadonah
Mic
Nah
Hab
Zeph
Hag
Zech
Mal
Matt
Mark
Luke
John
Acts
Rom
1Cor
2Cor
Gal
Eph
Phil
Col
1Thess
2Thess
1Tim
2Tim
Titus
Phlm
Heb
Jas
1Pet
2Pet
1John
Rev
long
short

Use Case – Readability Analysis

Daniela Oelke, David Spretke, Andreas Stoffel, Daniel A. Keim: **Visual Readability Analysis: How to Make Your Writings Easier to Read**. *IEEE Trans. Vis. Comput. Graph.* 18(5): 662-674 (2012)

Readability Explorer



Readability Explorer