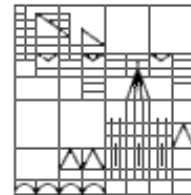


SPONSORED BY THE



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

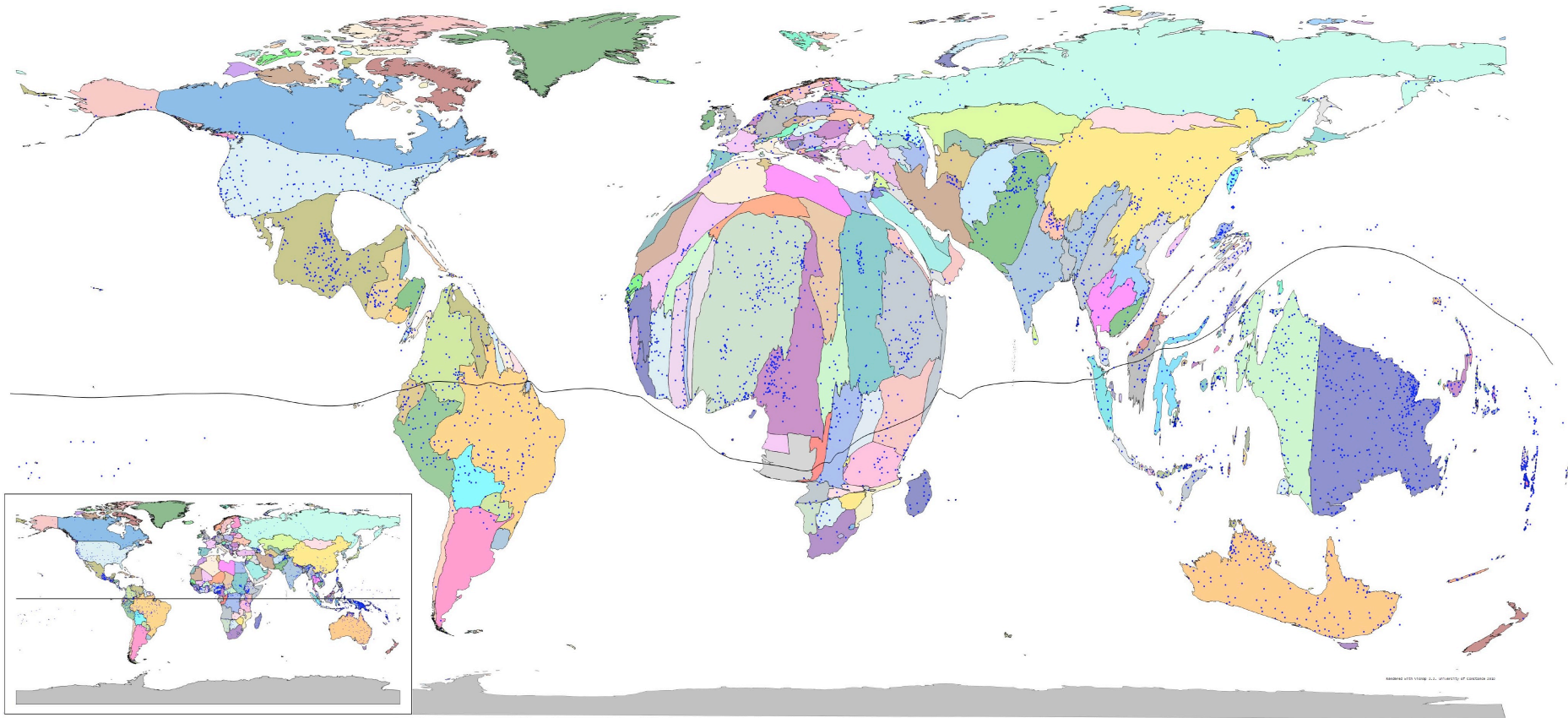


PHILOSOPHISCHE  
FAKULTÄT  
Seminar für Sprachwissenschaft

# Day 2 – More on LingVis

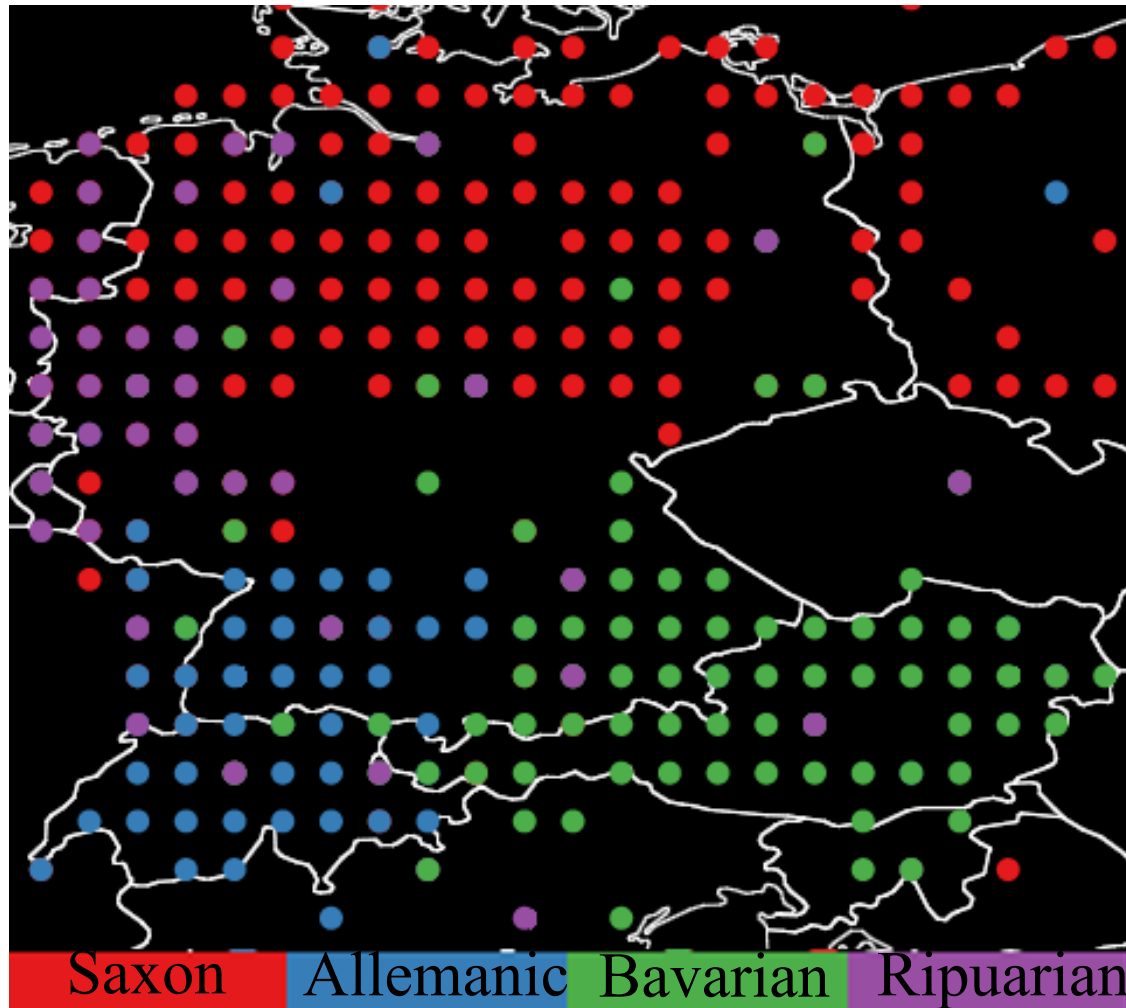
1. More Use Cases
2. Critical Discussion
  - Are the visualizations successful?
  - Are the visualizations useful?
3. What kinds of visualizations would you like?

# 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

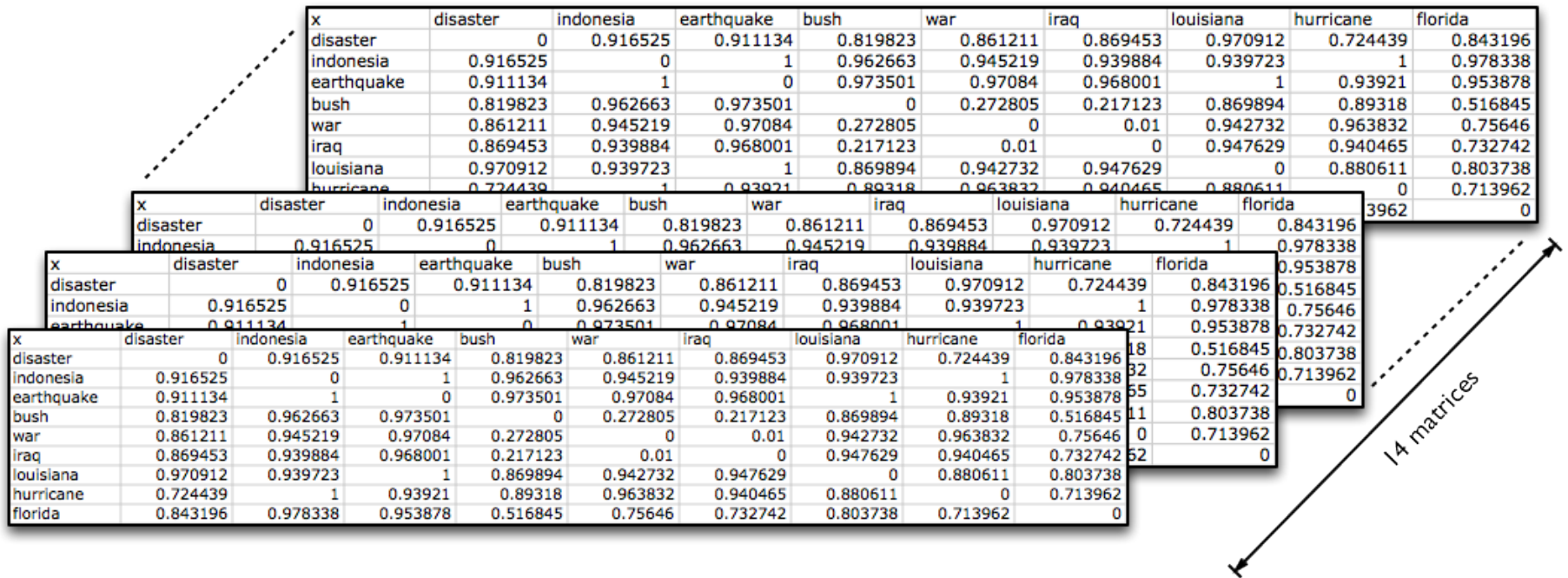
## Challenges for Visualization:

- **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



## Example: Animated Visualization

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

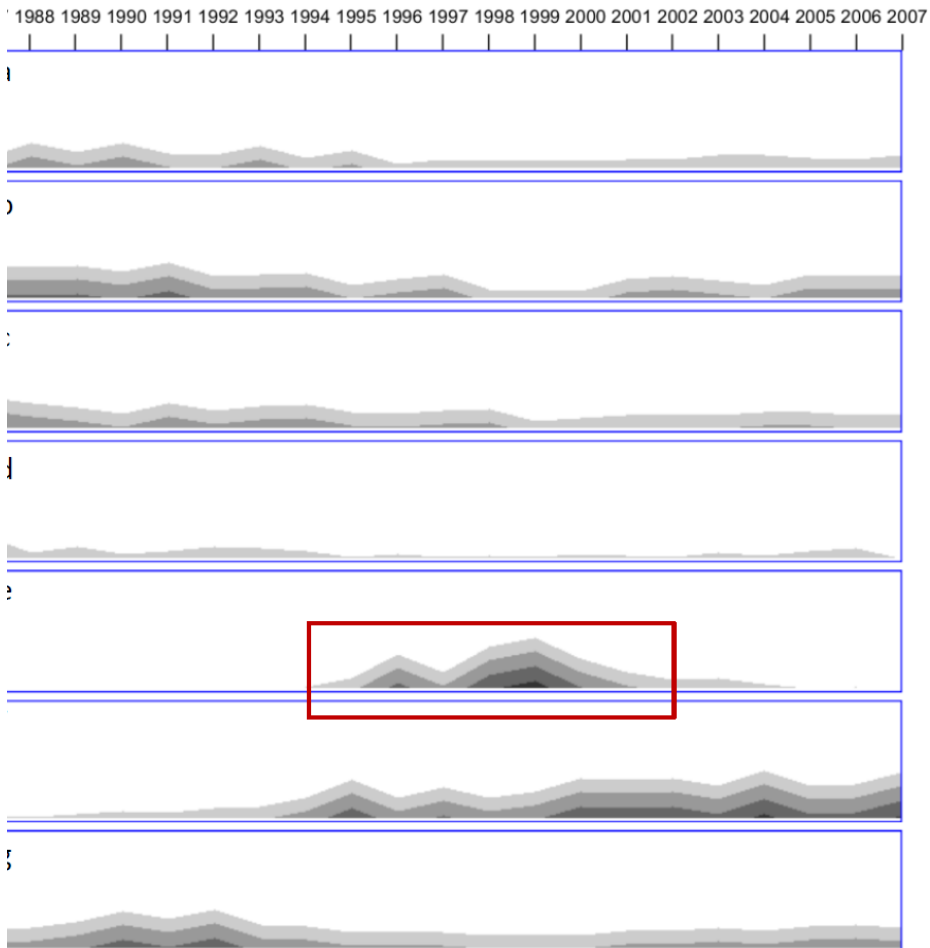


# Tracking Lexical Change

- 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

## to browse



sport, wind,  
water, ski, offer

wave, surfer,  
board, year,  
sport

channel,  
television,  
show, watch, tv

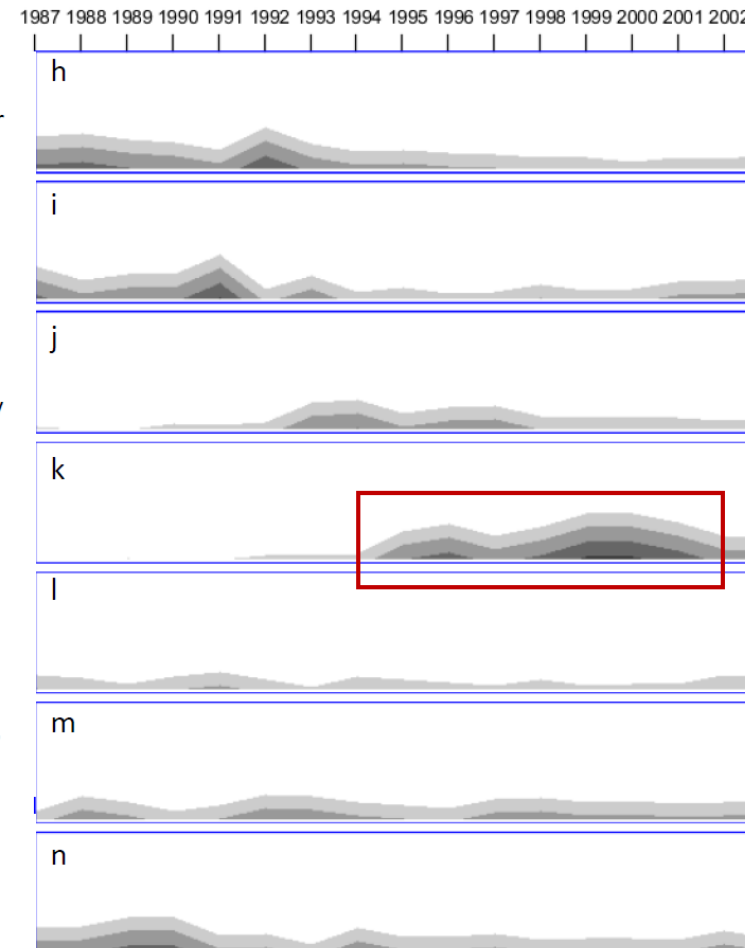
web, internet,  
site, computer,  
company

film, boy,  
movie, show,  
ride

year, day, time,  
school, friend

beach, wave,  
surfer, long,  
coast

## to surf



h

i

j

k

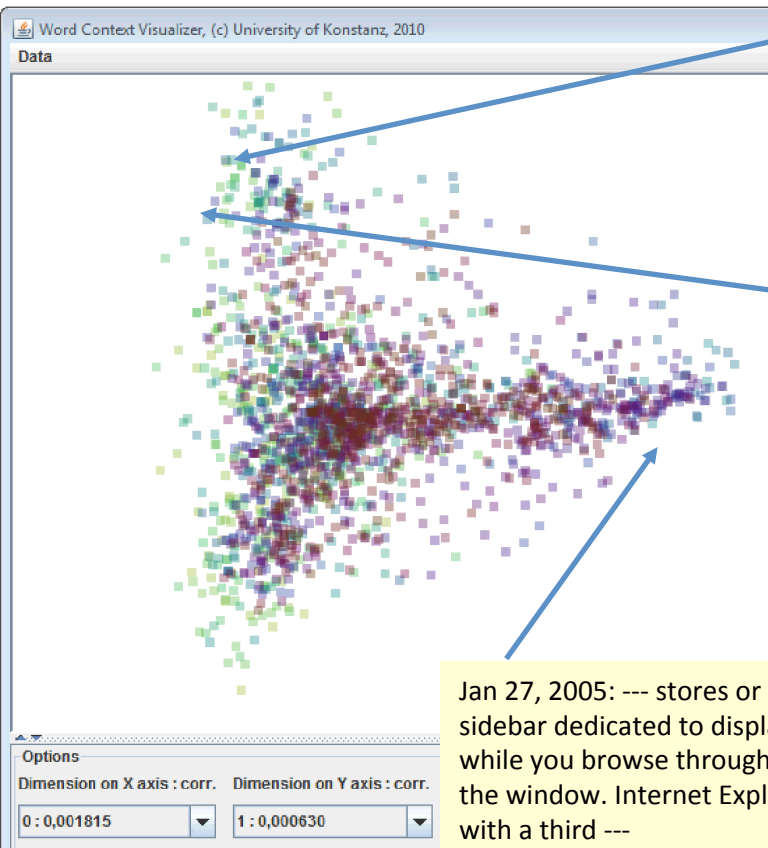
l

m

n



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

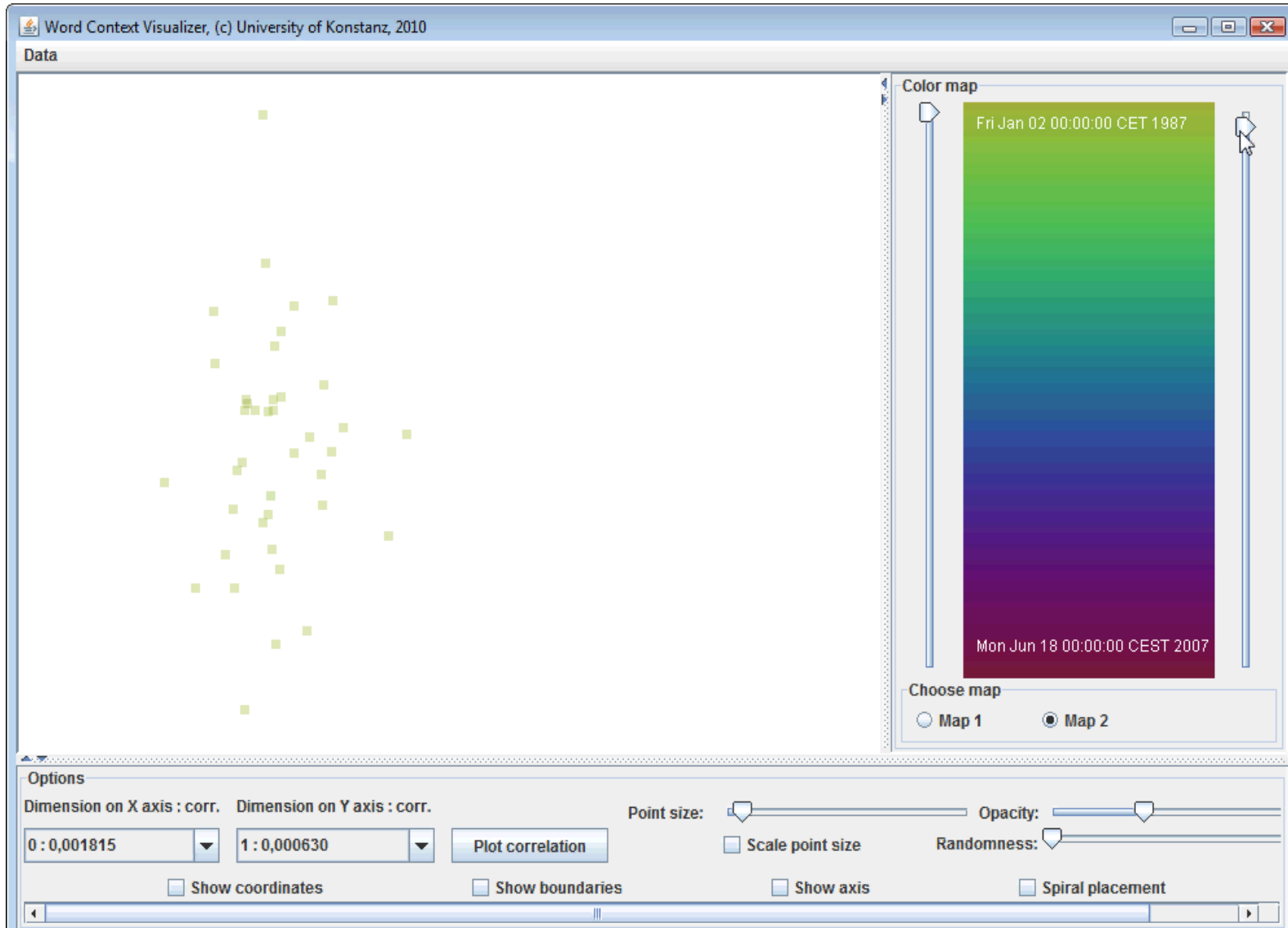


Sep 27, 1992: --- abound in Wiesbaden's pedestrian zone and around the shops within walking distance of the Casino. At the fair, the reading public is invited to browse, but not buy, as none of the books are for sale, on the weekend of Oct. 3 and 4, for a \$7 admission. Last year ---

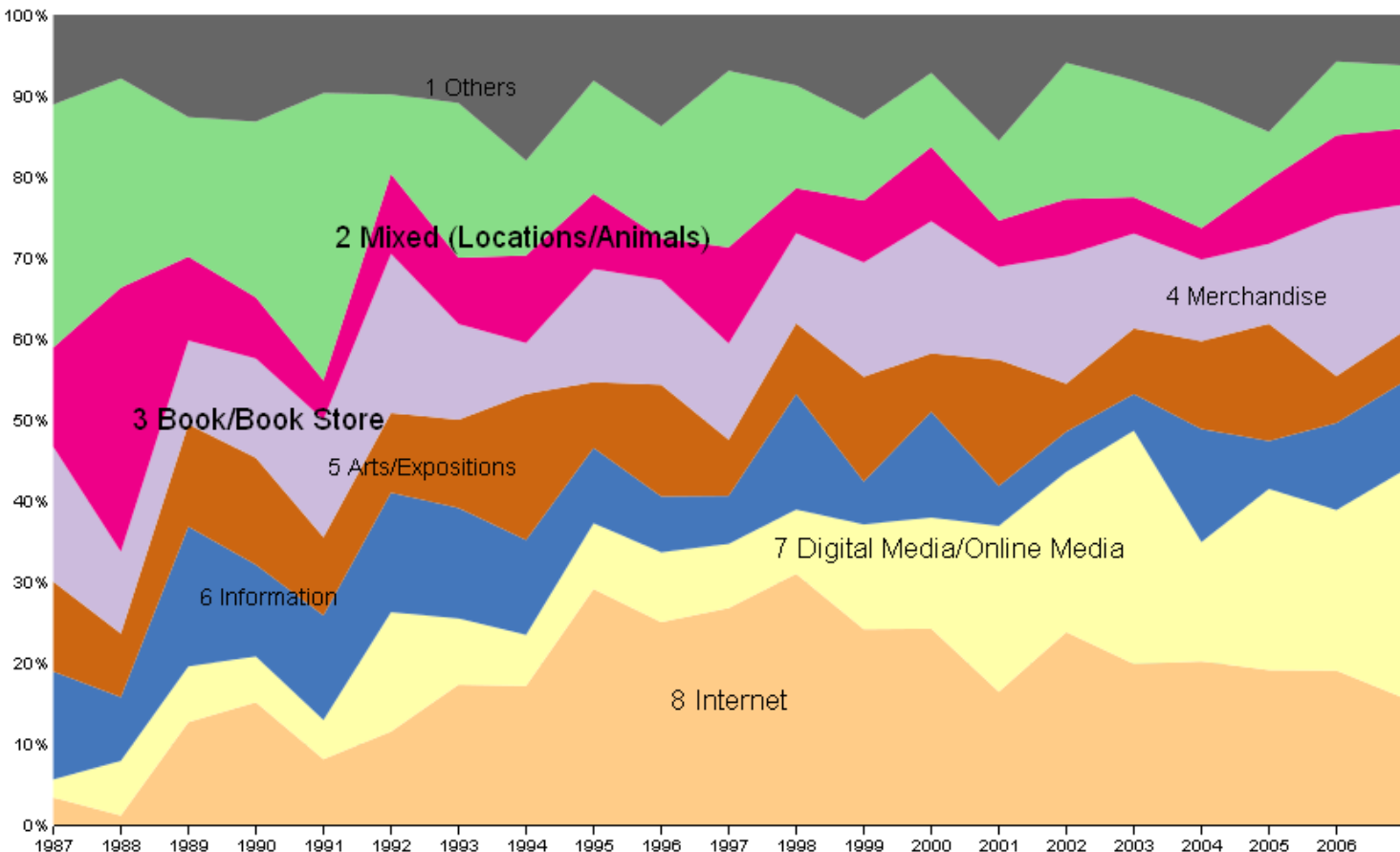
Apr 16, 1989: --- a time of failing independent bookstores. Soon after the store opened, it attracted authors, dramatists, poets and artists. Among those who came to chat, to browse and to see if their books and plays were on the shelves were Theodore Dreiser, John Dos Passos, H.L. Mencken and Eugene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

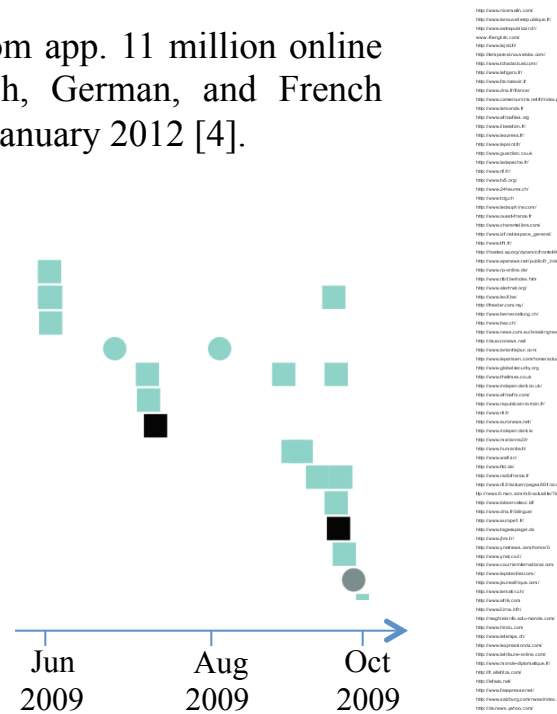
site make work page  
online list search music  
click

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].

- http://www.nicematin.com/
- http://www.lanouvellerepublique.fr/
- http://www.estrepublicain.fr/
- www.rfienglish.com/
- http://www.lejdd.fr/
- http://tempsreel.nouvelobs.com/
- http://www.tchadactuel.com/
- http://www.lefigaro.fr/
- http://www.francesoir.fr
- http://www.dna.fr/france/
- http://www.camerounlink.net/fr/index.php
- http://www.lemonde.fr
- http://www.africafiles.org



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



(C. Rohrdantz, Dissertation, Uni Konstanz)

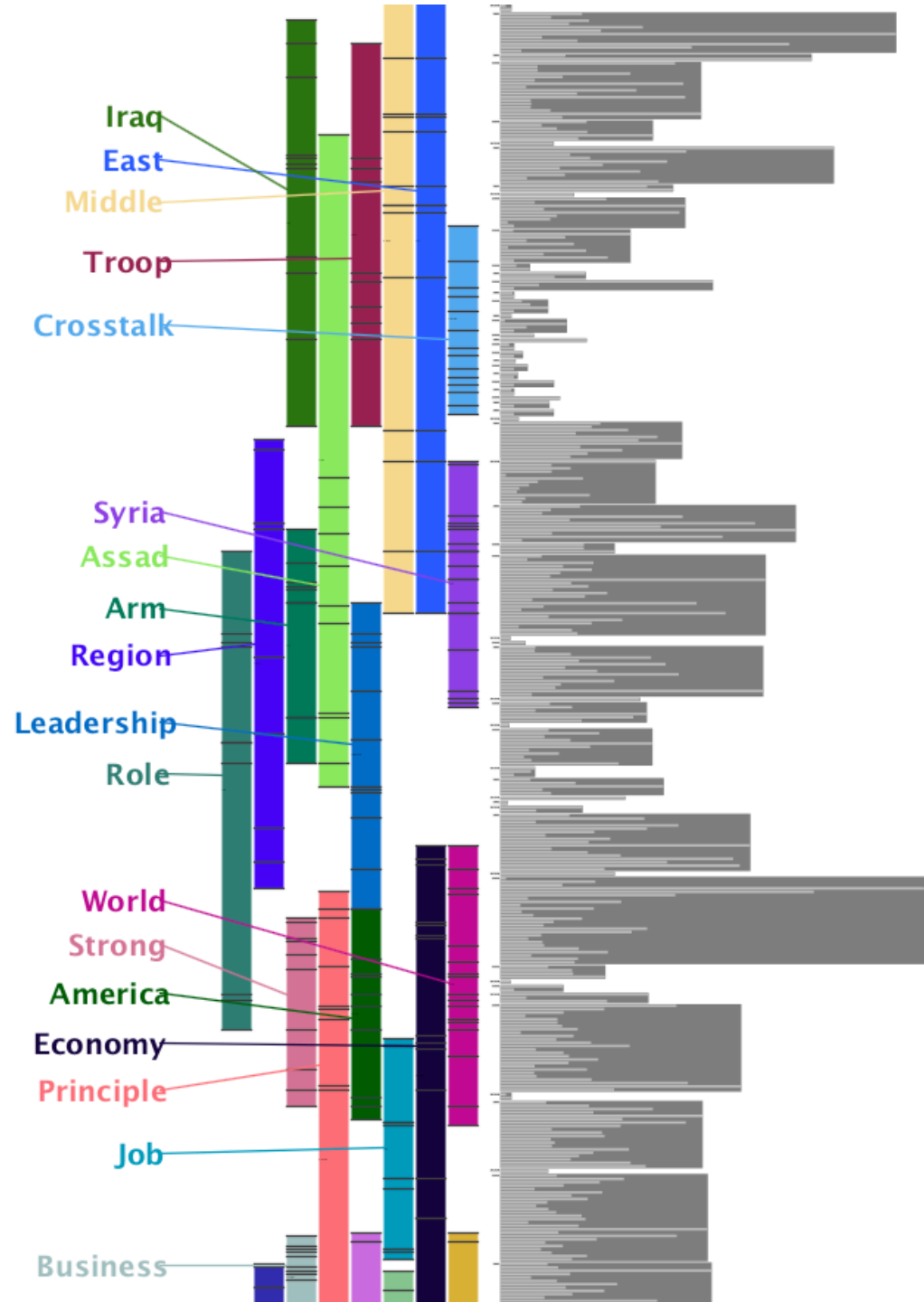
time

# 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: 3<sup>rd</sup> presidential debate between Barack Obama and Mitt Romney (Oct. 2012)

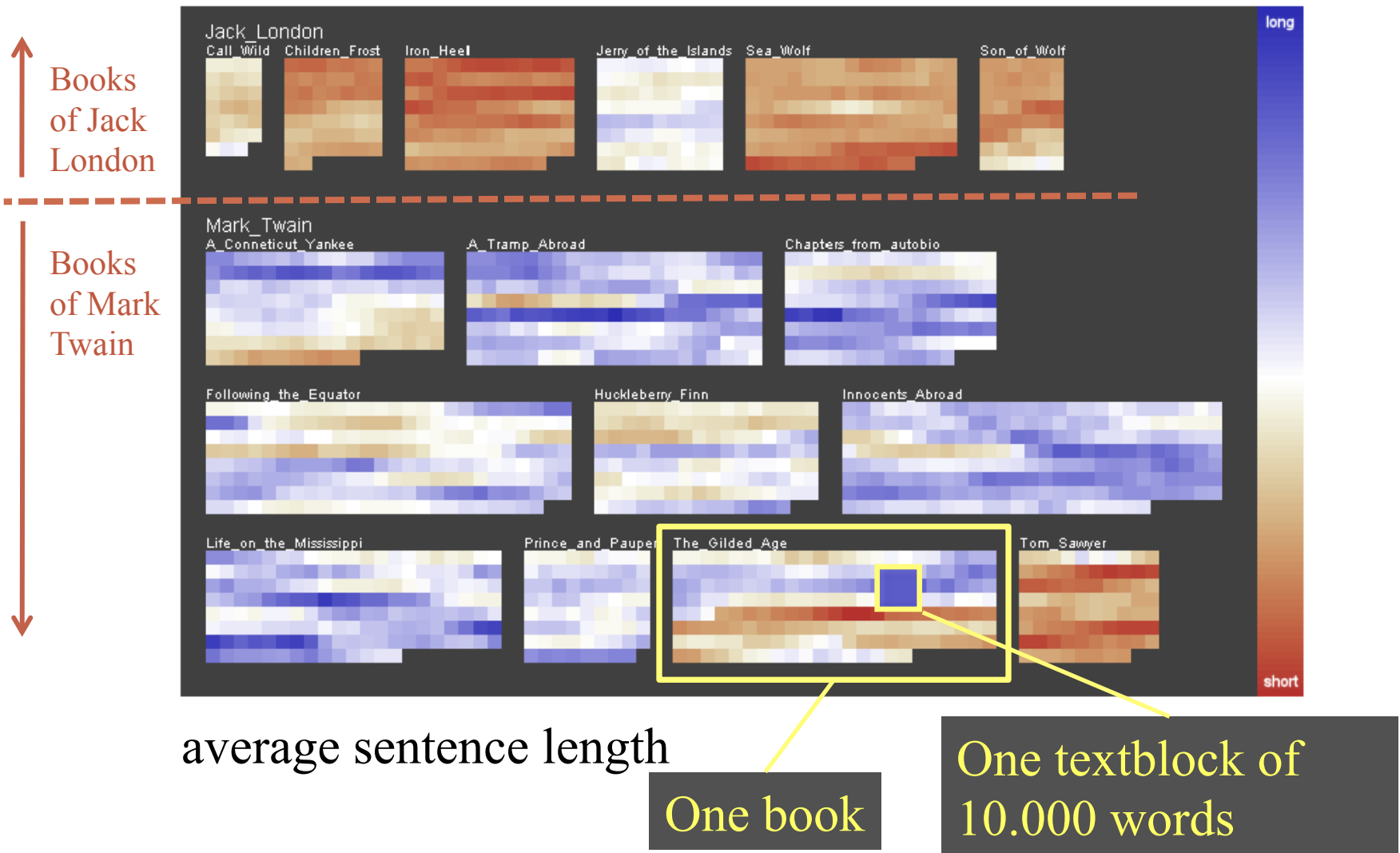
# Lexical Episodes:

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

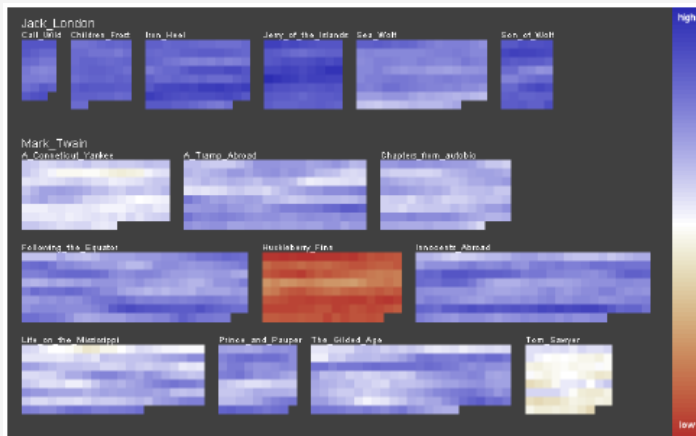


# Use Cases – Literary Analysis (More Pixel Visualizations)

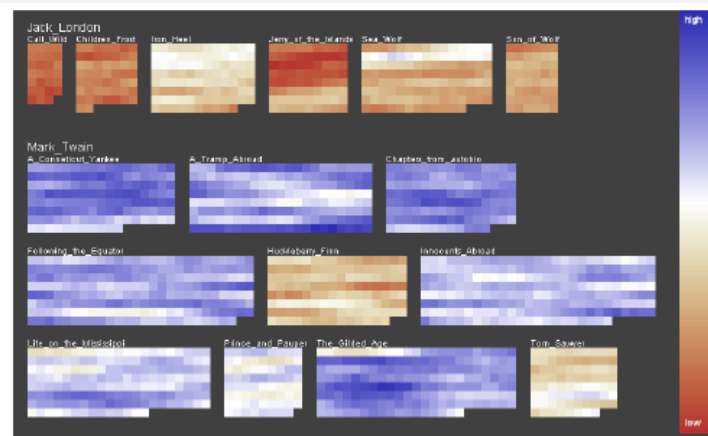
# Authorship Attribution



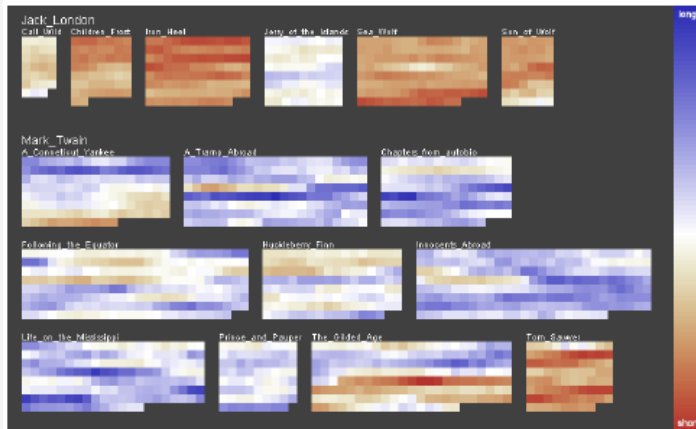




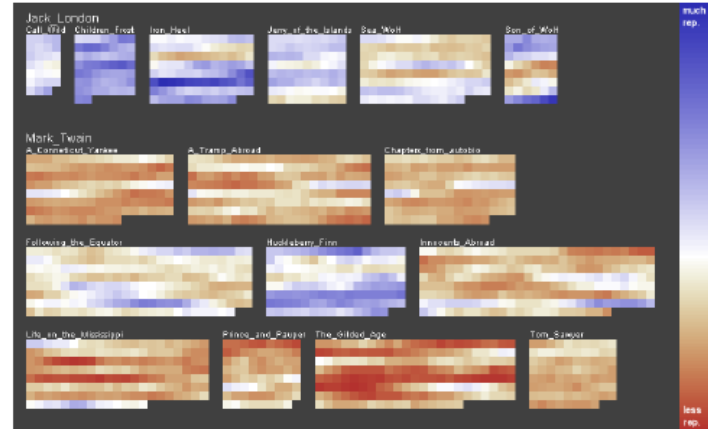
(a) Function words (First Dimension after PCA)



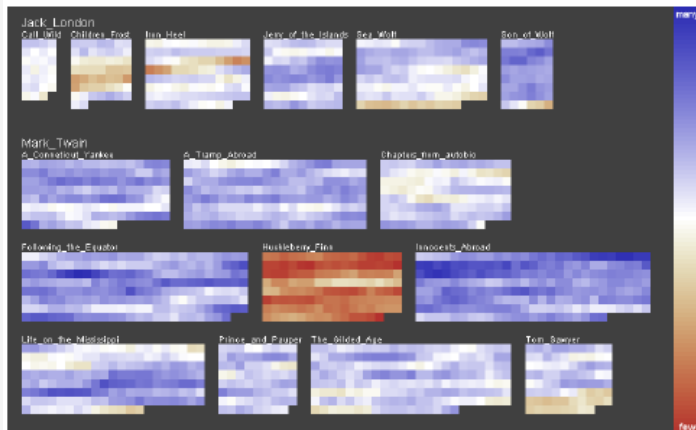
(b) Function words (Second Dimension after PCA)



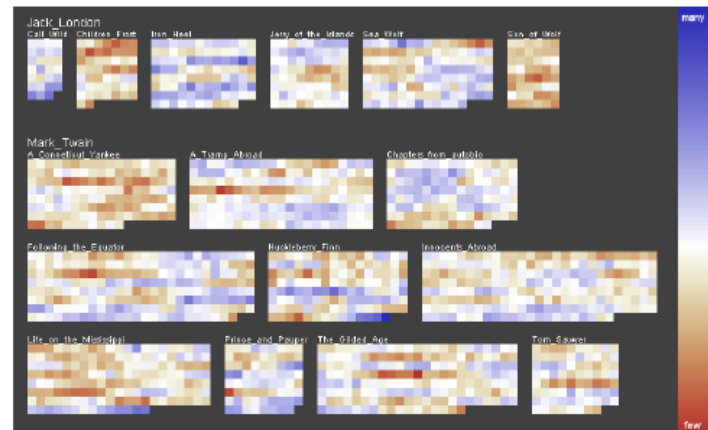
(c) Average sentence length



(d) Simpson's Index

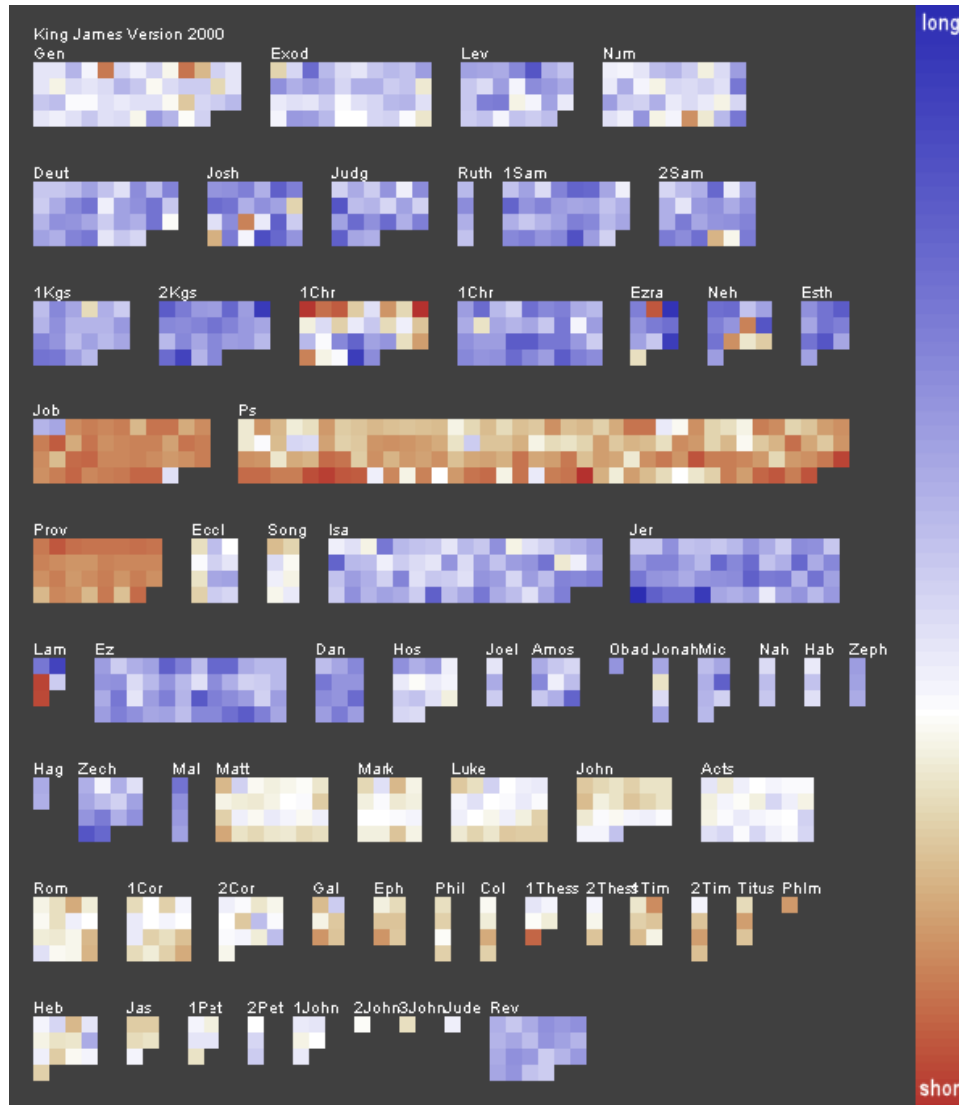


(e) Hapax Legomena

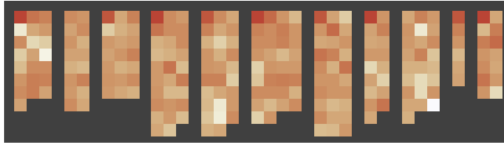


(f) Hapax Dislegomena

# Literature Fingerprinting on the Bible



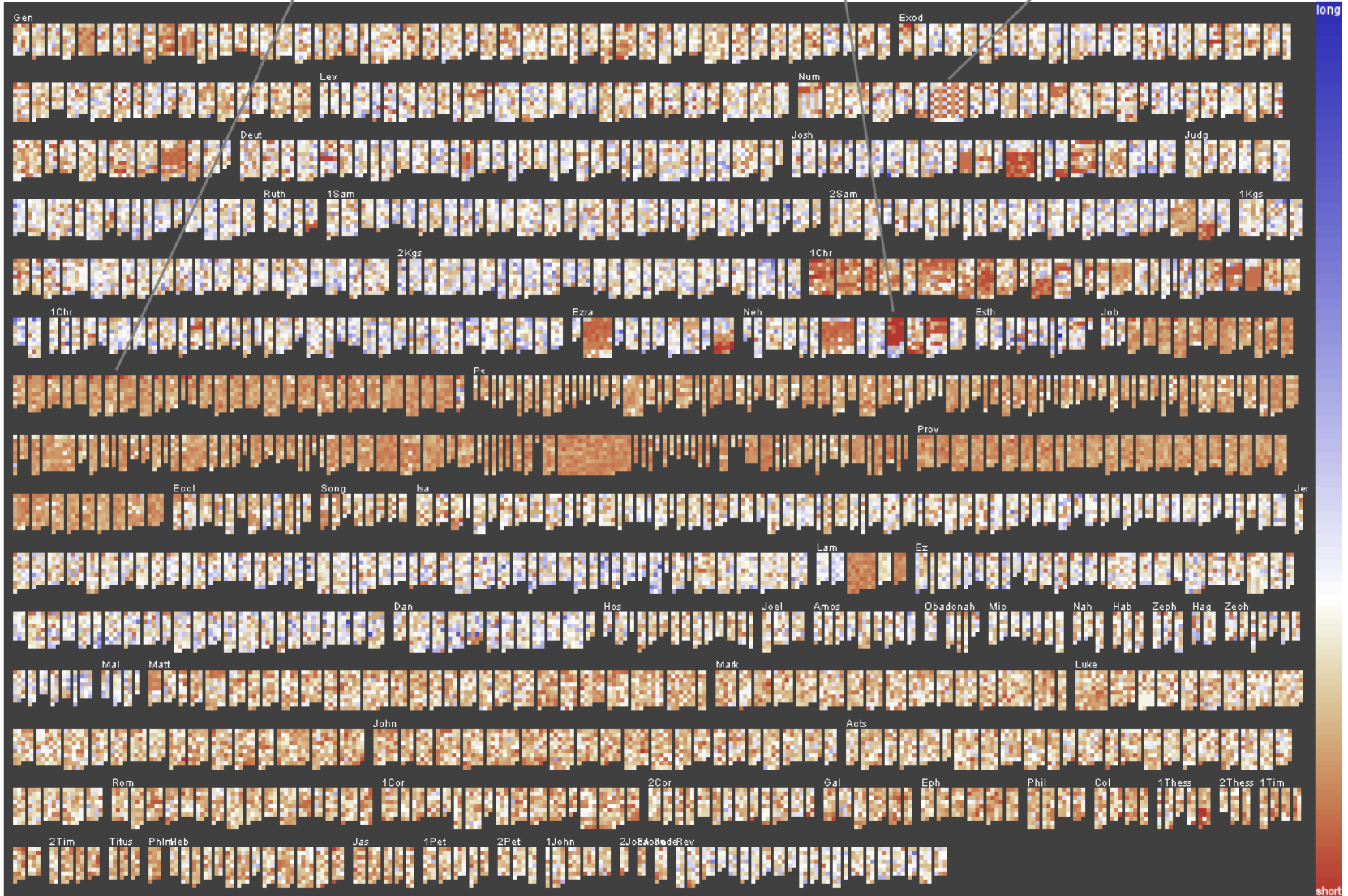
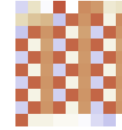
Job 16-26



Neh 10



Num 7



# Use Case – Readability Analysis

# Readability Explorer

The screenshot displays the Readability Explorer application window. The title bar reads "Readability Explorer (Build 2009.11.16.15.10)". The interface includes a "Corpus View" and "Document View" tab, with "Document View" currently selected. A dropdown menu shows "Average (non-weighted)". The main content area displays a document titled "(4) FP7 (B1) Cooperation (T3) Information and Communication Technologies - Work Programme 2009-2010". The document text is highlighted in red, indicating readability scores. The text discusses low power consumption ICT devices, innovation in health and social care, breakthroughs in ICT, and the mission of the ENIAC and ARTEMIS JTI.

Document:

## (4) FP7 (B1) Cooperation (T3) Information and Communication Technologies - Work Programme 2009-2010

low power consumption ICT devices and equipment but also through ICT solutions for better energy efficiency, lighting, virtual mobility and more efficient environmental simulation and monitoring. Support to this area is strengthened substantially and will address the various dimensions of ICT 's contribution to sustainability. In addition to the above transformations, the main mid-to-long term drivers for ICT research priorities identified for the first phase of FP7 remain valid today. These drivers include the high expectations of ' more for less ', i.e. more functionality and performance at lower cost as well as the need for better scalability, adaptability and learning capabilities of ICT systems. They also include stronger requirements for reliability and security of ICTs and the need to handle higher volumes and more complex digital content and services and to facilitate user control.

More innovation is also emerging from the use of ICT in ever more challenging applications in particular for health and social care, for transport, for lifestyle, culture and learning, energy and the environment. Achieving the best possible impact for Community support requires focusing and concentrating effort on key RTD challenges. This work program proposes a structure around seven challenges that should be addressed if Europe is to be among the world leaders in next generation ICT and their applications. The challenges are driven either by industry and technology objectives or by socio-economic goals. For each challenge precise targets and deliverables are identified in a 10 year time frame. In pursuit of the challenge targets, a set of research objectives will be called for in 2008 and 2009. These objectives are described in the next chapters of the work program and will provide the focus for the calls for proposals. For each objective, the work program defines the target outcome of the supported research and the expected impact of these outcomes on the European economy and society.

Breakthroughs in ICT increasingly come from cross-overs, combinations and convergence of technologies and disciplines at different levels, networks-services-devices. More and more, innovations come from the use of ICT in demanding application contexts. In the more technology-led challenges, research is directed towards removing roadblocks and improving the capability of generic technology components, systems and infrastructures suitable for a range of applications. In the more application-led challenges, research is focused on new technology-based systems, products and services that provide step-changes in the capabilities of the resulting application solution. The ICT work program addresses a research problem through different angles corresponding to different technological challenges. One example is the research challenge related to the ' Internet of Things ' ( IoT ). One angle is offered by Objective 1.3, concerned with the service architecture that enables the discovery of object properties and events. It is related to the governance of IoT type schemes, i.e. whether events pertaining to objects should be stored locally, be advertised systematically or not, put in a common register, access policy questions etc..

As such, it is a system-oriented Objective the mission of which is to define the service architecture within application schemes where objects can be under control of several organizations or entities over time. Another viewpoint is provided by Objective 1.1 that targets novel architectural schemes at network level. That is, it works on the fundamental networking layers, i.e. those dealing with routing and end-to-end connectivity. A third angle is given by Objective 1.4 addressing security and privacy in networks at the infrastructure level as well as the development of technologies to support security in networks of ' tiny things '. A fourth perspective is presented by Objective 3.5 that targets system level integration, including programming of possibly opportunistic collections of smart networked objects, which may further invoke higher layer services. This integration addresses both functional requirements ( e.g. reduced energy use ) and non-functional aspects ( e.g. real-time operating systems and -- possibly ad-hoc -- network protocol stacks ). JTIs are a pioneering approach to pooling public-private efforts, designed to leverage more R-D investments from Member States, Associated Countries and industry, and to reduce the tremendous fragmentation of EU R-D.

Two JTIs related to the ICT Programme have recently been launched. The focus of the ENIAC JTI in nanoelectronics will be industrial developments addressing mainly technology for the next generation of ' More Moore ' and the ' More than Moore ' domains. The ICT WP will typically cover the beyond CMOS fields and more advanced ' More than Moore ' domains preparing Europe for the design and manufacturing of the next generation components and miniaturised systems. The ARTEMIS JTI will focus on developing industrial platforms for the development and implementation of embedded systems responding to industry requirements in specific application domains ( e.g. for the automotive and aerospace sector, for smart homes and public spaces, energy efficiency, manufacturing etc. ). In the embedded systems area, the ICT WP will typically address new concepts, technologies and tools for engineering next generation systems characterised by wide distribution and interconnection and responding, in addition to timeliness and dependability, to more stringent constraints in terms of size, power consumption, modularity and interactivity.

The Ambient Assisted Living ( AAL ) joint national program will cover market-oriented R-D on concrete ICT-based solutions for ageing well with a time to market of 2-3 years, in particular with focus on involvement of SMEs and the business potential. AAL will complement the ICT WP which will focus on longer term research topics in this field which integrates emerging ICT concepts with 5-10 years time to market as well as essential research requiring larger scale projects at EU level, e.g. with strong links to standardisation. In addition to international cooperation activities addressed in the relevant objectives within the 7 Challenges and FET, horizontal international cooperation activities will be supported. By providing support to information society policy dialogues, this will contribute to increasing the participation of third country organizations in the Programme and will facilitate the widest diffusion and local exploitation of ICT research results. Complementing the research agenda, three important priorities related to policy developments and innovation have

# Readability Explorer

Readability Explorer (Build 2009.11.16.15.10)

Average (non-weighted)

Corpus View Document View

Document:  
(4) FP7 (B1) Cooperation (T3) Information and Communication Technologies - Work Programme 2009-2010

485

Sentence	Word Complexity	Sentence Length	Phrase Length	Degree of Nominalization	Vocabulary Complexity
Two JTIs related to the ICT Programme have recently been launched.					
The focus of the ENIAC JTI in nanoelectronics will be industrial developments addressing mainly technology for the next generation of `More Moore` and the `More than Moore` domains.					
The ICT WP will typically cover the beyond CMOS fields and more advanced `More than Moore` domains preparing Europe for the design and manufacturing of the next generation components and miniaturised systems.					
The ARTEMIS JTI will focus on developing industrial platforms for the development and implementation of embedded systems responding to industry requirements in specific application domains ( e.g. for the automotive and aerospace sector, for smart homes and public spaces, energy efficiency, manufacturing etc. ).					
In the embedded systems area, the ICT WP will typically address new concepts, technologies and tools for engineering next generation systems characterised by wide distribution and interconnection and responding, in addition to timeliness and dependability, to more stringent constraints in terms of size, power consumption, modularity and interactivity.					