Visualizing Linguistic Data:
From Principles to Toolkits for Doing it Yourself

Verena Lyding, European Academy of Bozen/Bolzano
Chris Culy, University of Tübingen

AVML Conference, 5 September 2012

Outline

- Introduction InfoVis --> LInfoVis
- Visualization theory and design
- Visualization of linguistic data
  - Linguistic data types and tasks
  - Practical challenges
- Tools and toolkits
  - Demos
  - Overview on toolkits
- Hands-on

Information Visualization

Definition:

"The use of computer-supported, interactive, visual representations of abstract data to amplify cognition." (Card et al., 1999)

Linguistic Information Visualization (LInfoVis):

The application of information visualization principles to display any kind of ***information concerning language and its use****.*

Where do we stand today?

Late 1980s : Advances in computer graphics boosted visualization

1999: „The foundational period of information visualization is now ending“ (Card et al.)

→ summing-up the state-of-the-art and put vis into practice

What happened since then?

- Focus on improving the quality of graphics
- Innovative visualizations
- Wide distribution (visualizations on the web, community focus, etc.)
- Application to new domains: **language data and linguistics!**

„History“ of linguistic visualization

- Early 1990s: Sporadic visualizations of language data (e.g. TileBars (1995), SeeSoft (1992), etc.)
- Mid 1990s to 2000s: Tools for visual document analysis (e.g. InSpire, Leximancer, Document Galaxies, KeyWord visualizations etc.
- Mid 2000s: Widespread use of Word Clouds (e.g. Wordle, etc.)
- Mid-late 2000s: Appearance of a number of language visualizations on vis Websites and blogs: eagereyes.org (2006), ManyEyes (2007), hint.fm
- Late 2000s: Visualization at (Computational) Linguistics conferences
  - 2008: ACL/HLT tutorial „Interactive Visualization for Computational Linguistics“
  - 2009: ESSLII workshop „Linguistic Information Visualization“
  - 2012: EACL workshop „ Visualization of Linguistic Patterns”
  - 2012: AVML conference „Advances in Visual Methods for Linguistics“
  - 2013: DGfS workshop „Visualization of Linguistic Patterns“

The field of linguistic visualization is rapidly developing!

Lots of promises

„A picture is worth more than thousand words.“

Visualizations help:

- seeing old things in new ways
- creating and discovering ideas/knowledge
- detecting patterns, finding abstractions
- communicating ideas

Strengths of visualization

I. Visual perceptual capabilities of its human users ☺

- Attention mechanisms for monitoring, pattern detection
- Adaptivity, perceptual inference

External aid to cognitive processes

- Increasing memory and processing resources of the user
- Providing information in a manipulable medium

→ Slogan: „Using vision to think.“ (Card et al.)

II. Power of ist medium: The computer!

- Improved rendering
- Real-time interactivity
- Low processing cost
- Automatic mapping of data

Visualization theory and design

How do visualizations work?

Information is transformed into graphic representations

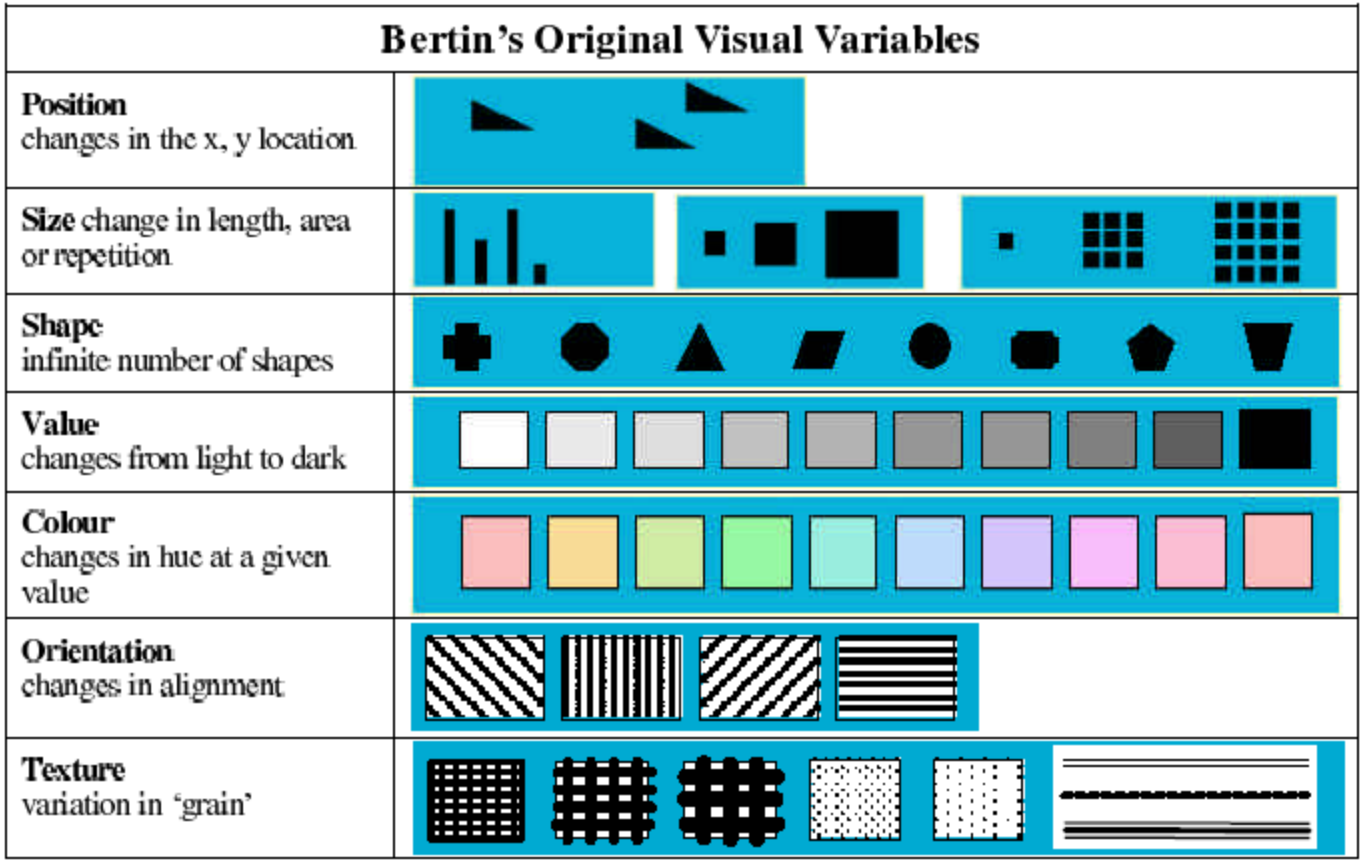
- Graphics: Marks with visual properties are placed in space
- Encoding: The mapping of information to graphics

Meaningful visualizations require:

- Fit of graphical representation and data
- Adherence to visualization principles

Building blocks: visual variables

Value = Brightness



Taken from: M. Carpendale, "Considering visual variables as a basis for information visualisation", Dept. of Computer Science, University of Calgary, Canada, Tech. Rep. 2001-693-16, 2003, Table 1.

Key characteristics of visual variables

- **Selectivity:** “Is A different from B?”
- **Associativity:** “Is A similar to B?”
  - Positioning > {size, brightness} > {color, orientation (for points)} > texture > shape
- **Order:** “Is A more/greater/bigger than B?”
  - Size and brightness; ¬ orientation, shape, texture
- **Quantity:** “How much is the difference between A and B?”
  - Position > size; ¬ other variables
- **Length:** “How many different things?”
  - Shape, Texture: infinite, but …; Brightness, hue: 7 (Assoc.) – 10 (Dist.)
  - Size: 5 (Association) - 20 (Distinction); Orientation: 4

Visual encoding

“Sameness of a visual element implies sameness of what the visual element represents” (Tufte, 2006)

- Gestalt principles: similarity, proximity
- Be consistent concerning relations of similarity, proportion and configuration.
- Adhere to conventional uses of visual variables
    - E.g. in cartography use blue color for water
- Take care of “effects without causes” (Tufte)

Visual clarity

“Clutter and confusion are failures of design, not attributes of information.” (Tufte, 1999)

- “expressive visualizations”: Encode all and only relevant information
- “let the same ink serve more than one informational purpose” (Tufte)
- Informational intent: Highlight important information

→ Good visualizations reduce the cognitive effort for understanding complex information

Data transparency and integrity

- Don’t hide information without indicating what is left out
  - Can missing information be reconstructed?
  - Can transformations/simplifications/abstractions be tracked?
- Present information in context
  - Use rulers/scales
  - Add labels and legends
  - Choose visual elements in a way that what they represent is easily memorized

Data arrangement

“Overview first, zoom and filter, then details-on-demand” (Shneiderman, 1996)

- Utilize the display space to give most room to the subject of the user’s interest (Card et al., 1999)
- User needs both overview (context) and detail (focus) simultaneously (Card et al., 1999)
  - Can be combined within a single display, like in human vision
- Layering and separation: visually stratifying various aspects of the data (Tufte, 1999)

Interactivity

“Rapid interaction fundamentally changes the process of understanding data.“ (Card et al., 1999)

- Dynamic queries
- Direct walk
- Panning across a view of the data (camera movement)
- Brushing-and-linking technique: simultaneous update of different views on the data (Hearst, 1999)
- Details-on-demand
- Direct manipulation
- Animated transitions can improve perceptions of changes between different graphical representations.

→ Many interaction techniques are essentially a form of selection

Visual processing

Visual information can be processed in two ways:

- controlled processing: detailed, serial, low capacity, slow, able to be inhibited, conscious (e.g. reading)
- automatic processing (*preattentive*): superficial, parallel, high capacity, fast, cannot be inhibited, unconscious, characterized by targets "popping-out" during search

→ Visualizations for both types are needed

Aims of visualization

- To convey information about language data
- To provide a way to interact with language data (user interface)
- To be an aid to discovery, decision making and explanation of information about language

➔ ”The purpose of visualization is insight, not pictures.“

Visualizing linguistic data

Language and linguistic data

Written language data:

- Document collections
- Running text
- Linguistic collections (like dictionnairies, etc.)

Related information:

- Extra-textual data (metadata)
- Linguistic annotations
  - hierarchical data -- e.g. compositional structure of linguistic units
  - relational data -- e.g. associations between words
  - parallel layered data -- e.g. labels for POS categories, alternative word choice, or error/correct form
  - transition states -- e.g. documentation of the text production process
  - frequency information tags, also quantitative tags
  - sequencing and distance/dispersion data -- e.g. positions of a word in text, word order

Also spoken language data with information on sound, pitch, intonation, etc.

Data types

For the purposes of visualization, it is useful to classify data (Hearst, 2009):

- **Quantitative** data: numbers, etc. that can be processed arithmetically
- **Categorical** data: everything else
  - **Interval**: ordered data with measurable distances (e.g. months)
  - **Ordinal**: ordered data without measurable distances (e.g. hot-warm-cold)
  - **Nominal**: data without organization (e.g. weather types, a collection of names)
  - **Hierarchical**: data without order, arranged into subsuming groups (e.g. {{ mammals, { bear …}, { cat { lynx…},…},…},…} , etc.)

→ Quantitative, interval, and ordered data are easier to convey visually than nominal data.

Where do textual elements fit in?

Properties (annotations) of textual elements are usually:

- Quantitative (frequencies)
- Or structured (trees, emotion scales)

→ Categorical data

Also, the actual textual items are important!

Difficulty: textual items are not *mappable*

- Too variable and too complex to be reduced

Approaches for handling textual data / providing context:

- Interactive visualizations
- Multiple data layers

Challenges for Visualization

Visualization:

- How to **render visible properties of the objects** of interest?

InfoVis:

- How to **map non-spatial abstractions** into effective visual form?

Visualization of Linguistic Information:

- How to **map non-spatial abstractions and textual elements** to visual form? Which abstractions are meaningful?

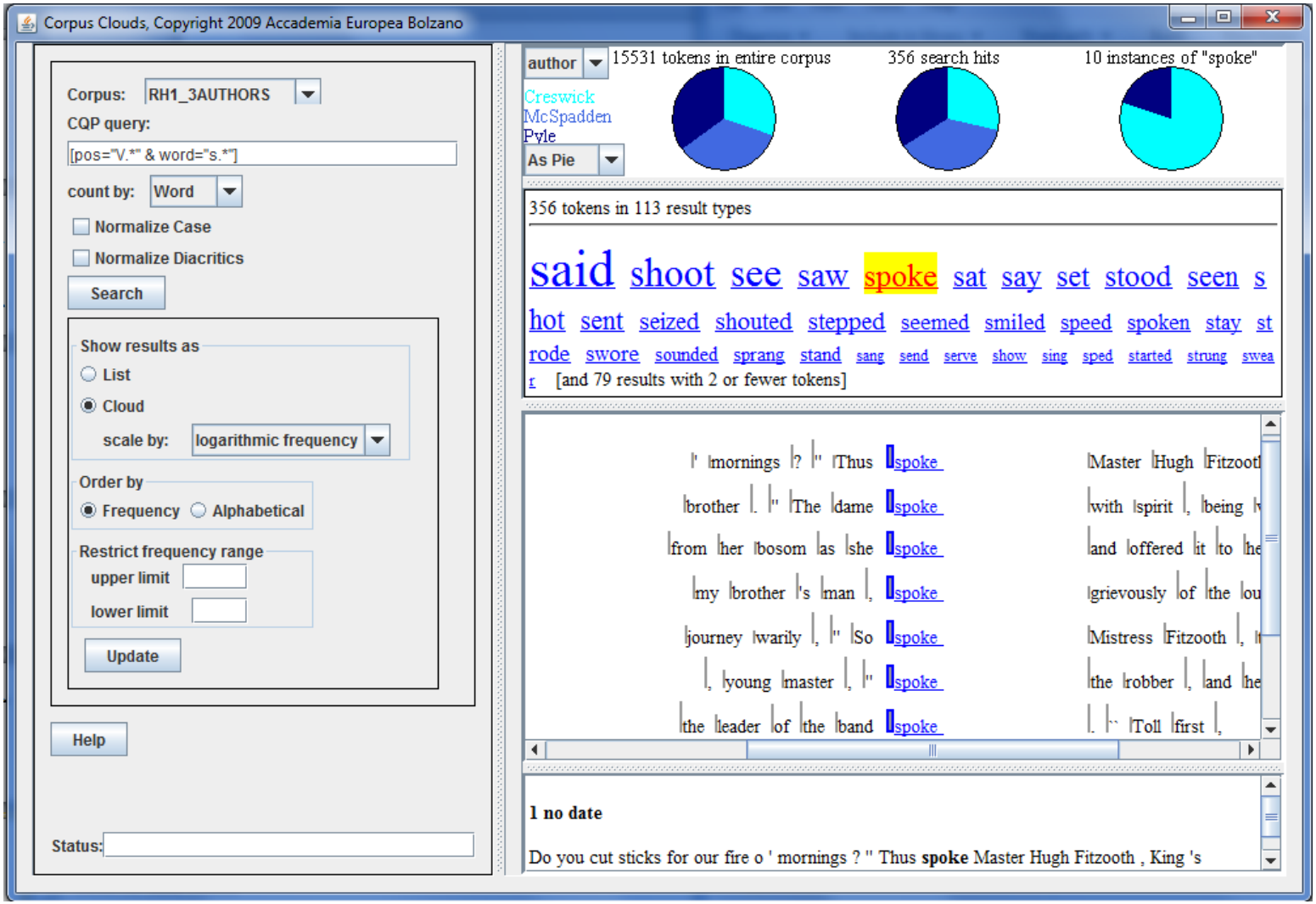
Practical challenges

- Working cross-platform / cross-browser
- Real-time interaction (client or server sided)
- Scalability of visualizations
- Data preparation
- Automated data access and aggregation
- Access to distributed language ressources
- Handling different data types and formats
- Customization vs. generic tools

Tools and toolkits

Demo Corpus Clouds

Exploration of corpus frequency data: Overview, zoom and detail



Corpus Clouds – vis principles and implementation

Visualization principles:

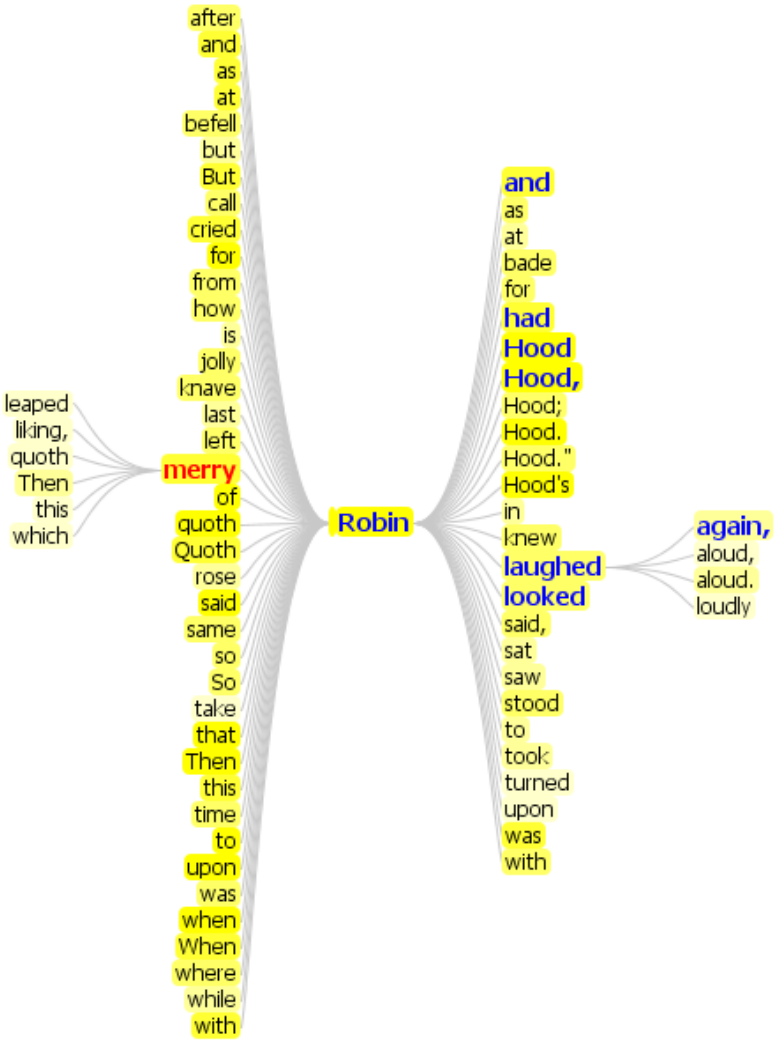
- Overview, zoom-in, details on demands
- Multiple views on the data
- Brushing and Linking
- Visual encoding of frequencies

Implementation details:

- Java implementation
- Integrated with CQP accessed through WebService
- Combination of program-external and internal, and precalculated data aggregation

Demo Double Tree

Exploration of KWIC results: interactive exploration



Double Tree – vis principles and implementation

Visualization principles:

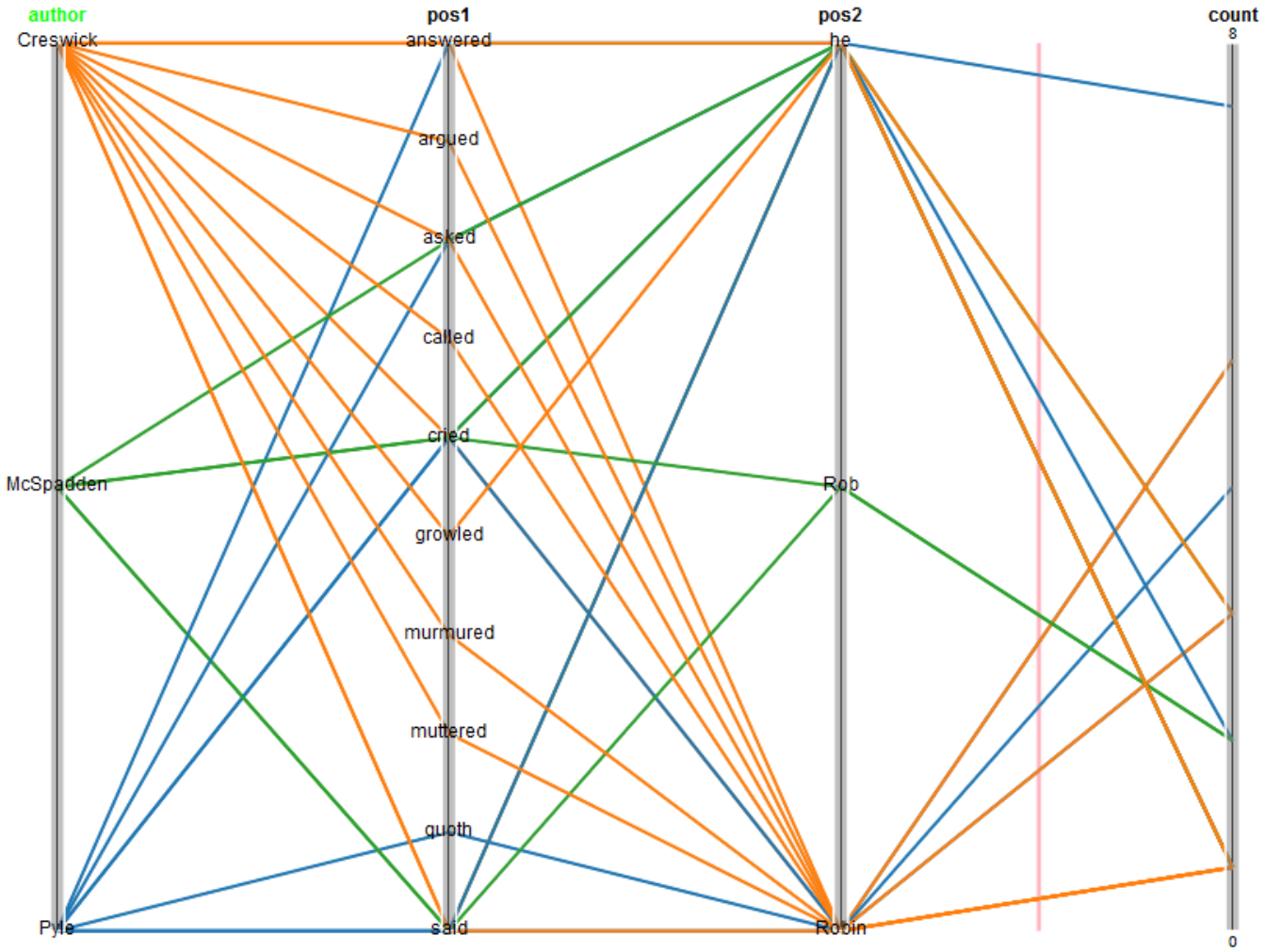
- Information-dense display
- Interaction with slow transitions
- Focus and context
- Visual encoding of frequencies

Implementation details:

- Java implementation using Prefuse
- Partly integrated with CQP accessed through WebService

Structured Parallel Coordinates

Vis for ngrams and corpus features: interactive analysis of multidimensional data



Structured Parallel Coordinates – vis principles and implementation

Visualization principles:

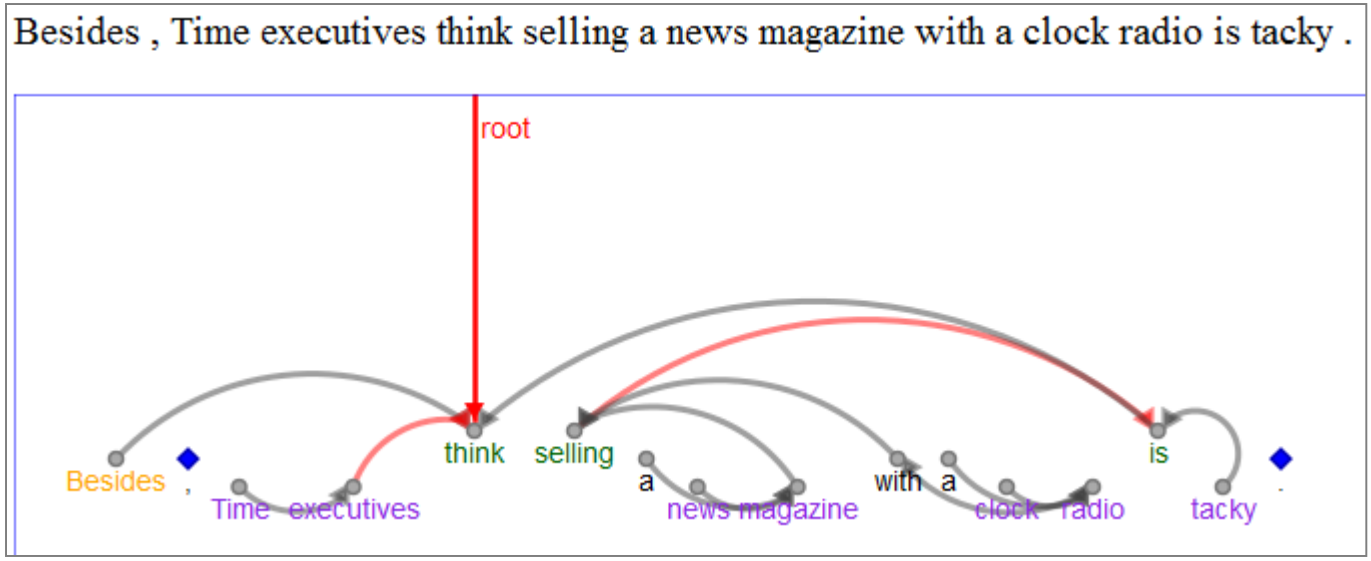
- 2D representation of multidimensional data
- Interactive data selection and filtering
- Focus and context

Implementation details:

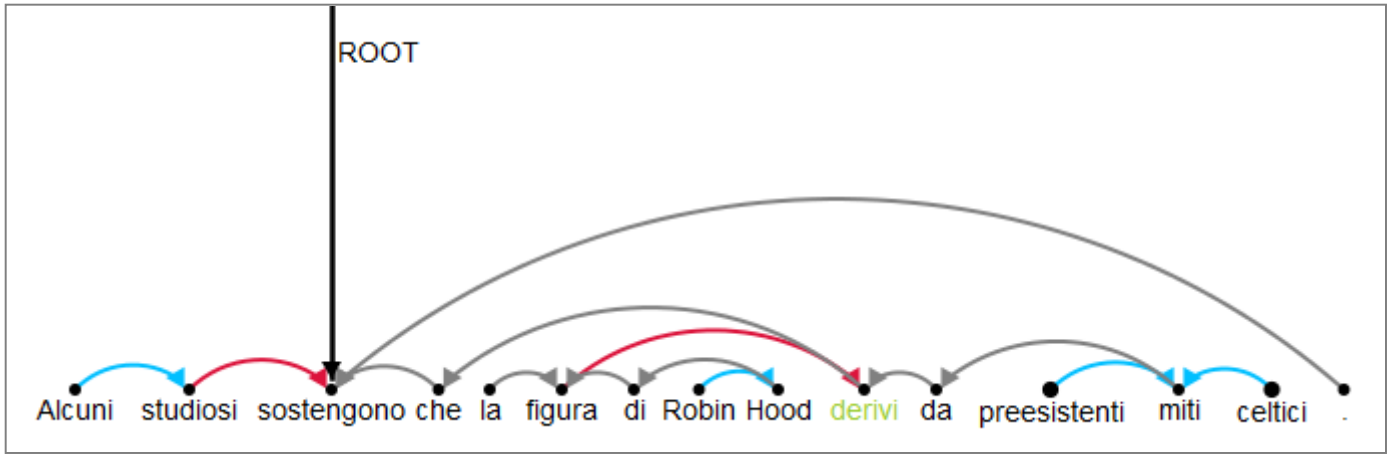
- Javascript implementation based on Protovis
- Implementation of different sample applications

Extended Linguistic Dependency Diagrams

Manipulable visualization of dependency relations



Data source: Kalashnikov 691 dependency bank (from PARC 700, as reanalyzed by By)



www.corpusitaliano.it

Extended Linguistic Dependency Diagrams – vis principles and implementation

Visualization principles:

- Manipulable visual encoding of data
- Details on demand

Implementation details:

- Javascript implementation based on Protovis
- Adjustability of different parameters like text position, word spacing, color coding, arcs curvature, etc.

Four levels of visualization tools (1)

Existing programs using a common/generic/simple data format

- e.g. spreadsheet programs, statistical packages (R), etc.
- Formats include data separated by tabs, commas, etc.

Four levels of visualization tools (2)

Existing programs using a complex/calculated format

- Specialized linguistics program
  - e.g. corpus query tools, annotation tools, etc.
- Relevant non-linguistic programs

Four levels of visualization tools (3)

A new/custom program developed using a visualization toolkit

e.g. DoubleTree, Structured Parallel Coordinates

Four levels of visualization tools (4)

A new/custom program developed (partly) without a toolkit

e.g. Corpus Clouds

Visualization tools: existing programs (levels 1+2)

Don't underestimate the power (and convenience!) of existing programs

  - e.g. spreadsheets, R

- Programs designed for other types of analysis can be used, with some imagination and effort
  - e.g. relational graphs are used in e.g. social networks and biology
- Language related, but non-linguistic tools
  - e.g. ManyEyes
- Don't forget linguistic programs like those here at AVML!

Visualization tools: Writing new programs (levels 3+4)

Before you start:

1. Who will be using the program? What level of knowledge, experience, etc?
2. What tasks will the user do?
3. What kinds of visualization would help with those tasks?
4. Is there already a program that does what you want?
   - Don't reinvent the wheel!
5. What tools are available?
   - Some visualizations are in toolkits in one programming language but not others

Visualization tools: Writing new programs (4)

An interesting combination is the NLTK toolkit in Python

- NLP tools with some visualization possibilities

There are many toolkits to create sophisticated visualizations

Often, but not always, they are complex and designed for experienced programmers

- some_links.html

A toolkit is not always necessary

- Does it provide the desired functionality?
- Is the time to learn how to use the toolkit worth the benefit?

Hands-on!

- DoubleTree (short)
  - Using and evaluating a visualization
- StructuredParallel Coordinates (longer)
  - Using variations of a visualization
  - Compiling data for the visualization
  - Evaluating a visualization
- Networks in D3 (longer, javascript programming)
  - Compiling data for the visualization
  - Extending an existing visualization
  - Evaluating a visualization

Thank you!

Verena Lyding

verena.lyding@eurac.edu

www.eurac.edu/linfovis

Chris Culy

christopher.culy@uni-tuebingen.de

www.sfs.uni-tuebingen.de/~cculy

References

Bertin, J. (1982): Graphische Darstellungen. Graphische Verarbeitung von Informationen. Berlin/New York: de Gruyter.

Card, S. K. / Mackinlay, J. D. / Shneiderman, B. (1999): Readings in Information Visualization: Using Vision to Think. San Francisco: Morgan Kaufmann Publishers

Carpendale, M. (2003): 'Considering visual variables as a basis for information visualisation', Dept. of Computer Science, University of Calgary, Canada, Tech. Rep. 2001-693-16.

Collins, C., Penn, G. and Carpendale, S. (2008). Interactive visualization for computational linguistics. ACL-08: HLT Tutorials. Retrieved from: http://www.cs.utoronto.ca/~ccollins/acl2008-vis.pdf. Access date: December 3, 2009.

Culy, C., Lyding, V., and Dittmann, H. 2011c. "Visualizing Dependency Structures" In: Proc. of the annual meeting of the Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL), Hamburg, Germany, 81-86.

Culy, C., Lyding, V., and Dittmann, H. 2011b. "xLDD: Extended Linguistic Dependency Diagrams" in Proceedings of the 15th International Conference on Information Visualisation IV2011, 12, 13 - 15 July 2011, University of London, UK. 164-169.

Culy, C., Lyding, V., and Dittmann, H. 2011a. "Structured Parallel Coordinates: a visualization for analyzing structured language data" In: Proceedings of the 3rd International Conference on Corpus Linguistics, CILC-11, April 6-9, 2011, Valencia, Spain, 485-493.

Culy, C. and V. Lyding. 2011. "Corpus Clouds - Facilitating Text Analysis by Means of Visualizations" in Human Language Technology: Challenges for Computer Science and Linguistics, Zygmunt Vetulani (ed.). Berlin:Springer. 351-360.

References

C. Culy & V. Lyding. 2010."Double Tree: An Advanced KWIC Visualization for Expert Users" In: Information Visualization, Proceedings of IV 2010, 2010 14th International Conference Information Visualization, 26-29 July 2010 London, United Kingdom, 98-103.

C. Culy & V. Lyding. 2010. "Visualizations for exploratory corpus and text analysis". In: Proceedings of the 2nd International Conference on Corpus Linguistics CILC-10, May 13-15, 2010, A Coruña, Spain, pp. 257-268.

C. Culy & V. Lyding. 2009. "Corpus Clouds - facilitating text analysis by means of visualizations". In: Proceedings of the 4th Language & Technology Conference, LTC'09 . Poznan, Poland, pp. 521-525.

Hearst, M. (2009): Search User Interfaces. Cambridge: Cambridge University Press.

Hearst, M. A. (1995): 'Tilebars: Visualization of term distribution information in full text information access', In: Proc. CHI'95, Denver, Colorado, pp. 56-66.

Lyding, V., Lapshinova-Koltunski, E., Degaetano-Ortlieb, S., Dittmann, H. and Culy, C. (2012): 'Visualising Linguistic Evolution in Academic Discourse', In: Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, April 2012, Avignon, France, Association for Computational Linguistics, pp. 44-48.

Tufte, E. (1999): Envisioning Information. Cheshire, Connecticut: Graphics Press LLC.

Tufte, E. (2006): Beautiful Evidence. Cheshire, Connecticut: Graphics Press LLC.

Todorovic, D. (2008): 'Gestalt principles'. Scholarpedia, 3(12):5345, Retrieved from: http://www.scholarpedia.org/article/Gestalt_principles. Access date: December 4, 2009.

Wattenberg, M. / Viégas, F. B. (2008): The word tree, an interactive visual concordance. In: IEEE Trans. on Visualization and Computer Graphics, vol. 14(6), pp. 1221-1228, Nov.-Dec. 2008.