Visual Corpus Interface – Putting Text Visualizations at Use

Verena Lyding
verena.lyding@eurac.edu
Institute for Specialised Communication and Multilingualism, EURAC research, Bolzano/Bozen, Italy

Michel Généreux
genereum@mcmaster.ca
Department of Linguistics and Language, McMaster University, Hamilton, Ontario, Canada

21/07/2016

Overview

- Aims and usage setting
- Data and use case
- The visual corpus interface
  - Macrostructure
  - Visualization components
  - Technical details
  - Interaction of visualizations
- Challenges and future work

Aim of the visual corpus interface

Objective:

Provide a corpus search interface with different text visualizations to support the linguistic analysis of corpora.

Motivation:

- Increased use of text collections as empirical research base (in the Digitial Humanities and in linguistics)
- Increased demand for tools that support the analysis of data
- Availability of visualizations for language data, but often disconnected from concrete usage applications and real-life data sources (text collections).

Usage setting

- OPATCH project - *Open Platform for Access to and Analysis of Textual Documents of Cultural Heritage*
- Cooperation between public library (text data provider) and computational linguistic research group (tools for text processing)
- Search and analysis interface for linguists, combining corpus query facilities (search, using linguistic features and statistics) and visualizations (data presentation)
- For three South Tyrolean German text collections

Data to visualize: text corpora

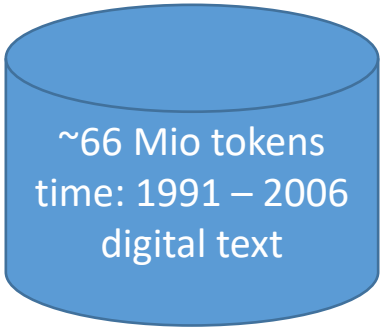
Linguistically processed text collections, which contain:

- The text itself
- For each word: information on its base form and word class
- Annotation of Named Entities
- Metadata on each text, e.g. author and year of publication

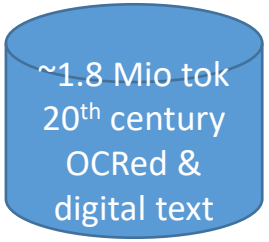
Three corpora of South Tyrolean German:

Historical newspapers

Contemporary newspapers



Balanced collection of fiction, informative, functional and newspaper texts



Use case: linguistic analysis

Typical linguistic research tasks focus on analyzing:

- Meanings of words or complex linguistic units
- Structural patterns
- ➢ *Data exploration* and *in-depth analyses*

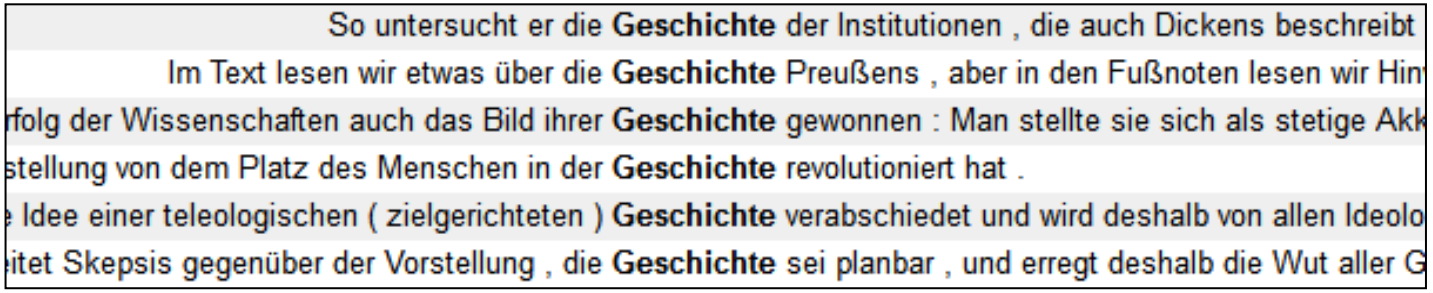
Basic techniques of linguistic analysis, with special reference to corpus concordance data (cf. Wynne, 2008), examine:

- Patterns of similarity or contrast around search terms
- Repeated patterns across search results
- Detailed linear context of results and its semantics
- Collocates indicating statistical significance of features

Visual corpus interface

It combines corpus search and visualizations.

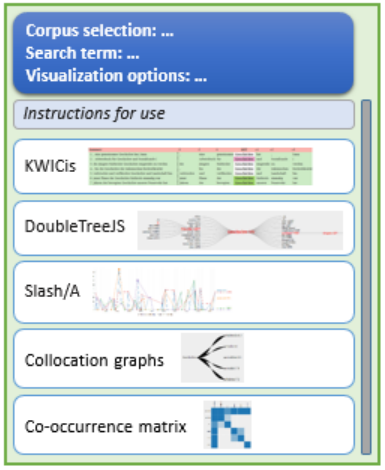
Historically, the presentation of corpus data has been limited to KWIC (KeyWord In Context) displays, i.e. plain text aligned on the search term.



The visual corpus interface combines:

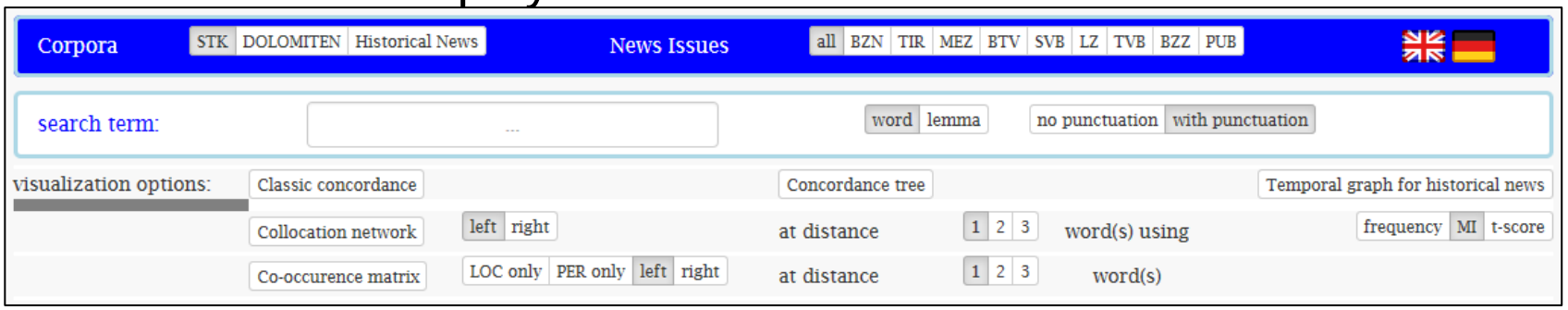
- Standard concordance displays (KWIC), and
- Recent visualizations of concordance data, which incorporate frequency and statistical information

Interface macrostructure



The interface is composed of a fixed search area (on top), usage instructions, and a series of five scrollable visualization areas (one below the other).

The search menu allows to set various linguistic parameters and choose the visualization for display.



Visualization: KWICis (classic concordance)

|< < > >| Showing results 1 to 20 of 431 results: || Cluster distribution: 0 1 2

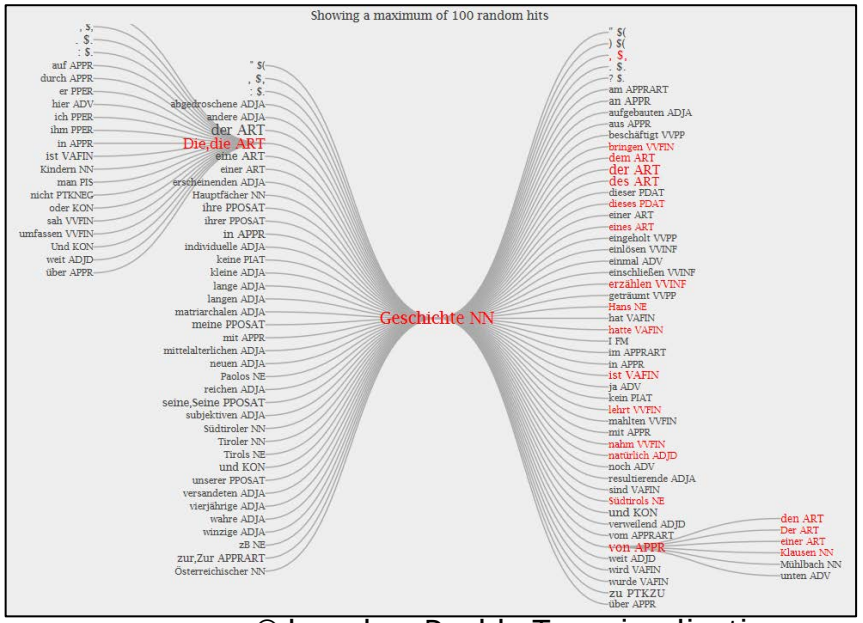
Results	-3	-2	-1	HIT	+1	+2	+3
1. , eine gemeinsame Geschichte hat, kann	,	eine	gemeinsame	Geschichte	hat	,	kann
2. . Arbeitsbuch für Geschichte und Sozialkunde (	.	Arbeitsbuch	für	Geschichte	und	Sozialkunde	(
3. die jüngere Südtiroler Geschichte eingereiht zu werden	die	jüngere	Südtiroler	Geschichte	eingereiht	zu	werden
4. , bis die Geschichte der italienischen Rechtsfakultät	,	bis	die	Geschichte	der	italienischen	Rechtsfakultät
5. verkitschte und verfälschte Geschichte und Landschaft bin	verkitschte	und	verfälschte	Geschichte	und	Landschaft	bin
6. jener Phase der Geschichte Südtirols einmalig war	jener	Phase	der	Geschichte	Südtirols	einmalig	war
7. Jahren der bewegten Geschichte unserer Feuerwehr hat	Jahren	der	bewegten	Geschichte	unserer	Feuerwehr	hat
8. Kommandant Aus der Geschichte der Freiwilligen Feuerwehr	Kommandant	Aus	der	Geschichte	der	Freiwilligen	Feuerwehr
9. und verlas die Geschichte vom Werdegang und	und	verlas	die	Geschichte	vom	Werdegang	und
10. Markstein in der Geschichte unserer Landwirtschaft.	Markstein	in	der	Geschichte	unserer	Landwirtschaft	.
11. Kapitel in der Geschichte der SBJerweist sich	Kapitel	in	der	Geschichte	der	SBJerweist	sich

© based on KWICis visualization: http://linguistics.chrisculy.net/lx/software/KWICis

- KWICis shows results in context, structured into table columns
- Visual mark-up according to the automatic clustering

➔ Supports the close inspection of search results

Visualization: Double Tree (concordance tree)

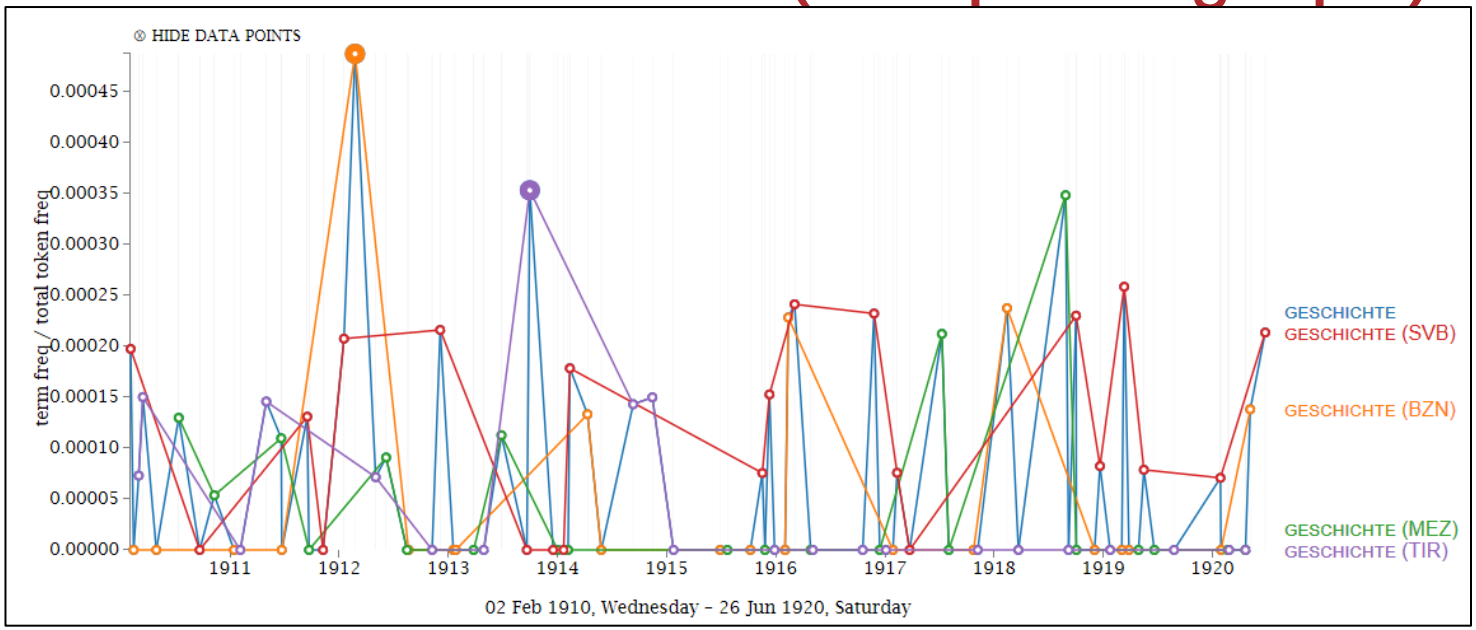


© based on Double Tree visualization:
http://linguistics.chrisculy.net/lx/software/DoubleTreeJS/index.html

- Contexts of a search term are collapsed into a double-sided tree
- Font size indicates frequency; frequency details on mouseover
- Interactive expansion of contexts; valid continuations are colored

➔ Serves to gain an overview of the data

Visualization: Slash/A (temporal graph)

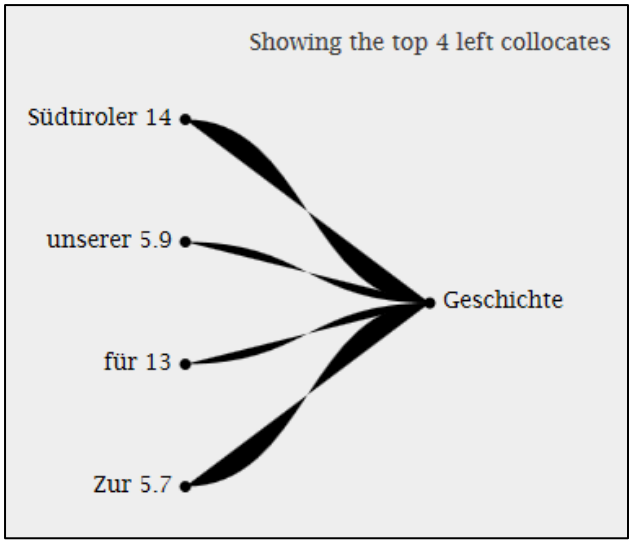


© based on Slash/A visualization: http://linguistics.chrisculy.net/lx/vistola/tools/slasha.html

- Word occurrence frequencies are plotted as a line graph over time
- Significant differences between occurrences are calculated based on Pearson's chi squared test, or in case of too little data Fisher's test

➔ Serves for the analysis of language evolution over time and the comparison of terms between different collections of texts

Visualization: collocation network

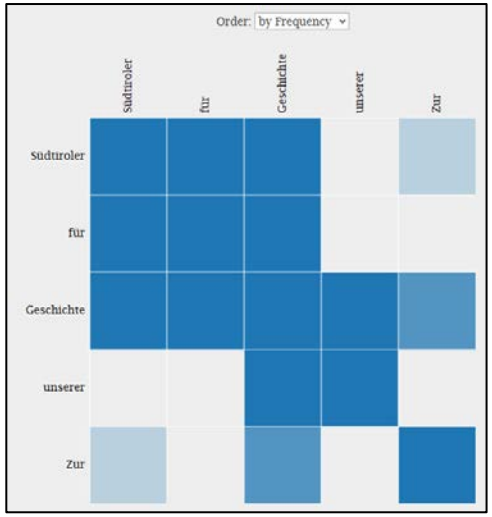


Build with JavaScript library d3 (https://d3js.org/)

- Network visualization showing most relevant left or right collocates
- Calculation of collocates by frequency, Mutual Information or t-score
- Adjustable distance to search term of 1 to 3 words

➔ Serves to get a quick overview on the most relevant left and right collocates of a search words, indicates statistical significance

Visualization: co-occurrence matrix



Build with JavaScript library d3 (https://d3js.org/)

- Cross-tabulation of frequencies of two words occurring as neighbors of one another
- To the left or right, and at a specified distance of 1 to 3 words
- Coloring of rectangles indicates co-occurrence strength of pairs

➔ Support the exploration of relations between collocates of a search word among each other, as well as in relation to the search word

Implementation: technical details

Visualizations are integrated with a *server-based corpus infrastructure* and adapted *data transformation scripts*:

- IMS Open Corpus Workbench as core infrastructure for the management, storage and retrieval of corpus data
- Communication through query API based on CQP language
- Transformations through client-side JavaScript and utilities
- Partly pre-computed data model stored as a JSON file

- Language visualization tools and JavaScript library d3

Interaction of different visualizations

The visual corpus interface provides five different visualizations in one connected and scrollable display.

- Inspection of coordinated visualizations on the same data and search queries, or comparison of different data and search queries
- Dynamic update of individual visualizations through user interaction, e.g. filtering of concordance data and temporal graph, update of collocation network
- Partial interlinking: clicking on a co-occurrence matrix field starts classic concordance search.

Open challenges and future work

The visual interface needs to address :

- Efficient data retrieval and processing
  - Scalability questions regarding data retrieval, pre-loading vs online
  - Processing times for recalculating and updating data sets
  - Handling of data access and related liabilities concerning copyright
- Scalability of visualizations to large data sets
  - Research on relevant visual adaptations or filtering options
- Consistent interlinking and interactivity of visualization components
  - To allow for dynamic updates of visualizations through user actions
  - To improve transparency of data flows
- Overall system usability
  - User study related to linguistic analyses

Thank you for your attention.

URL: http://commul.eurac.edu/opatch/

Contact: Verena Lyding (verena.lyding@eurac.edu), Michel Généreux (genereum@mcmaster.ca)

Partnership:
EURAC research, Bolzano
Dr. Friedrich Teßmann library, Bolzano
Institute for Corpus Linguistics and Text Technology (ICLTT), Vienna

Funding:
The OPATCH project is financed by the 'Provincia Autonoma di Bolzano-Alto Adige, Diritto allo studio, università e ricerca scientifica, Legge provinciale 13 dic. 2006, n. 14'

AUTONOME PROVINZ BOZEN SÜDTIROL — PROVINCIA AUTONOMA DI BOLZANO ALTO ADIGE

Special thanks go to Chris Culy and Velislava Todorova for technical support in adapting their visualization tools.

Bibliography

C. Culy and V. Lyding, “Double Tree: An advanced KWIC visualization for expert users,” Proc. of 14th International Conference on Information Visualization, IV 2010, pp. 98-103. 2010.

M. Généreux, "Correcting OCR results with computational linguistic methods". Fachtagung ‘Historischen Zeitungen im digitalen Zeitaler - I giornali storici nell'era ditale.’, Bolzano, 27.10.2014.

M. Généreux, E. Stemle, L. Nicolas and V. Lyding “Correcting OCR errors for German in Fraktur font”. Proc. of the 1st Italian Conference on Computational Linguistics, CLiC-it 2014, Pisa 09-10 Dec. 2014.

M. Généreux, “NLP challenges in dealing with OCR-ed documents of derogated quality”. Workshop on Replicability and Reproducibility in Natural Language Processing: adaptive methods, resources and software at IJCAI 2015, July 25-27, 2015, Buenos Aires, Argentina.

S. Hechensteiner, “Ein Klick in die Vergangenheit”. ACADEMIA #68, EURAC, p. 24-25.

V. Lyding, M. Généreux, K. Szabò, and J. Andresen, “The OPATCH corpus platform - facing heterogeneous groups of texts and users,” Proc. of the 2nd Italian Conference on Computational Linguistics, CLiC-it 2015, C. Bosco, F.M. Zanzotto, and S. Tonelli, Eds. Trento: Accademia University Press. December 2015.

V. Todorova and M. Chinkina, “Slash/A n-gram tendency viewer - Visual exploration of n-gram frequencies in correspondence corpora,” Proc. of the ESSLLI 2014 Student Session, R. de Haan, Ed. Tübingen, Germany, pp. 229-239, 2014.

M. Wynne, “Searching and concordancing,” in Corpus Linguistics: An International Handbook, vol. I, A. Lüdeling and M. Kytö, Eds. Berlin: de Gruyter, pp. 706-737, 2008.