HistoBankVis: Investigating Language Change via Visual Analytics

Christin Schätzle, Michael Blumenschein, and Miriam Butt
University of Konstanz, Germany

1st International Conference on Quantification in Visual Computing, University of Stuttgart, 2018

Universität Konstanz

Motivation

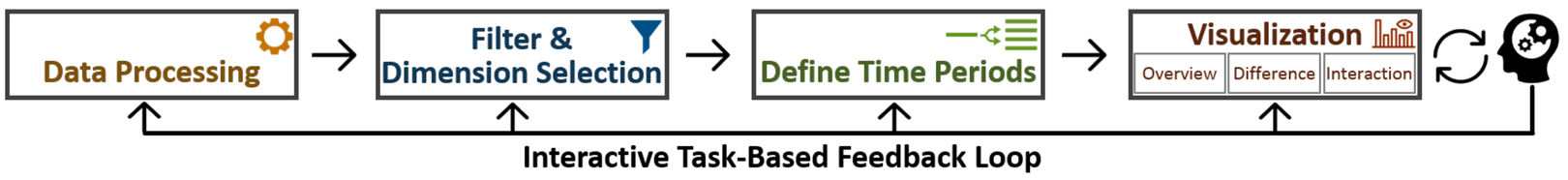
- Historical language change typically results from complex interactions between different structural language features.
- **Challenges/Problems:**
  - Highly complex interactions between various linguistic and extra-linguistic features have to be understood, while factoring in a temporal dimension.
  - The factors underlying a change are often unknown or at least highly debated among researchers.
  - Data sparsity is an inherent problem of historical corpus-based research.

→ Multiple complicated analyses including different features may have to be conducted in order to formulate clear hypotheses.

→ **Solution:**
Visual Analytics for Linguistics (LingVis)

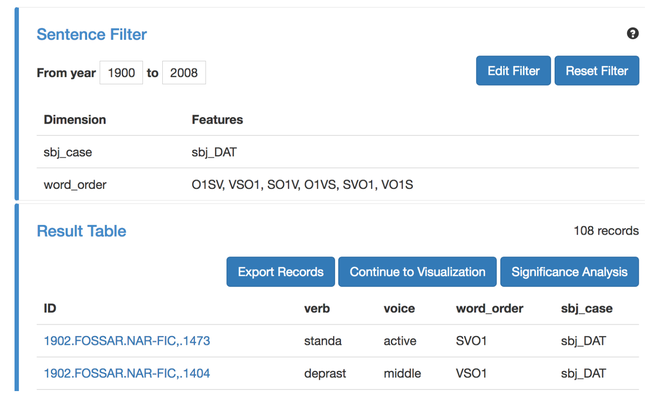
HistoBankVis: A Multilayer Visualization System

- On-line browser app: `http://histobankvis.dbvis.de`.
- Generically applicable system for historical linguistic research.
- Flexible and interactive investigation of a potentially high number of interacting linguistic features stored in a SQL database.

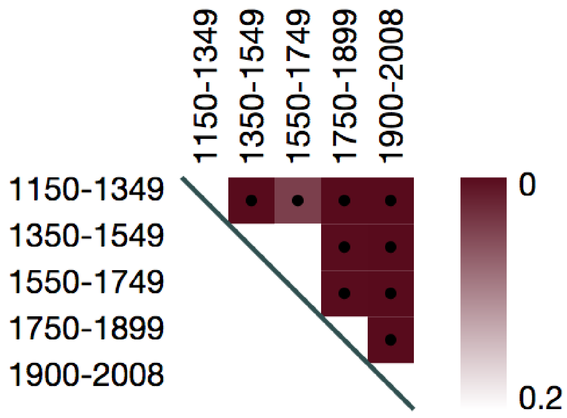


- Multiple layers of data representation at different levels of detail are combined with a structured statistical analysis process.
  - Filtering component
  - Compact Matrix Visualization
  - Difference Histograms Visualization
  - Dimension Interaction Visualization

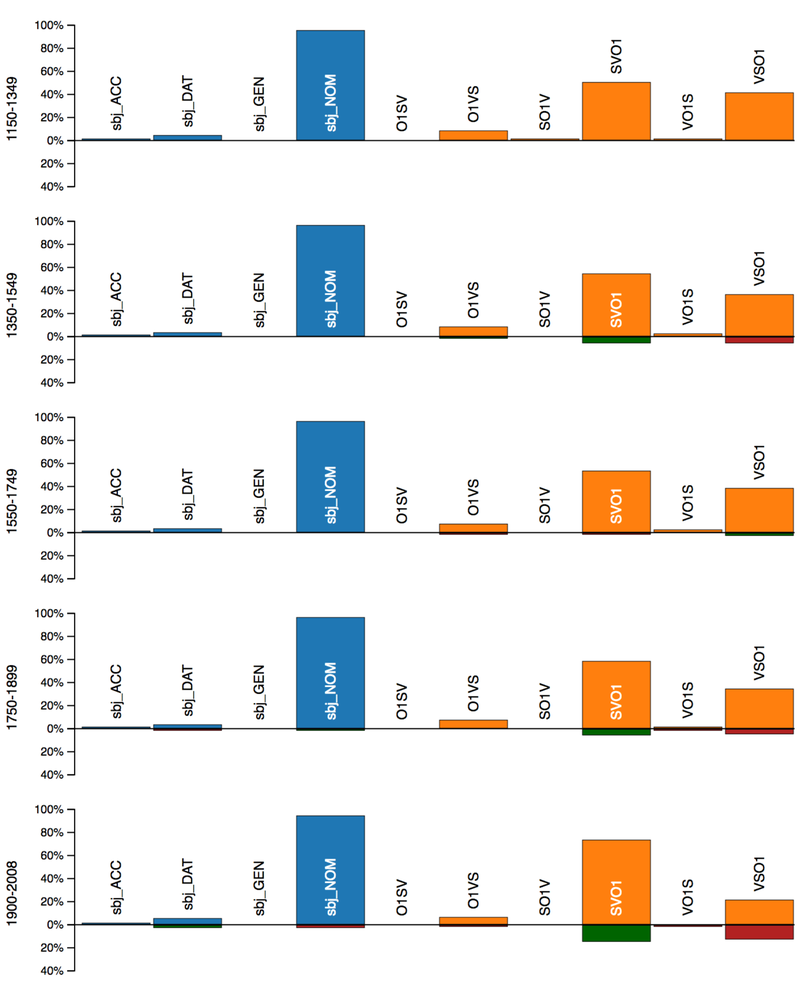
→ Iterative process of hypothesis generation and testing.



Compact Matrix

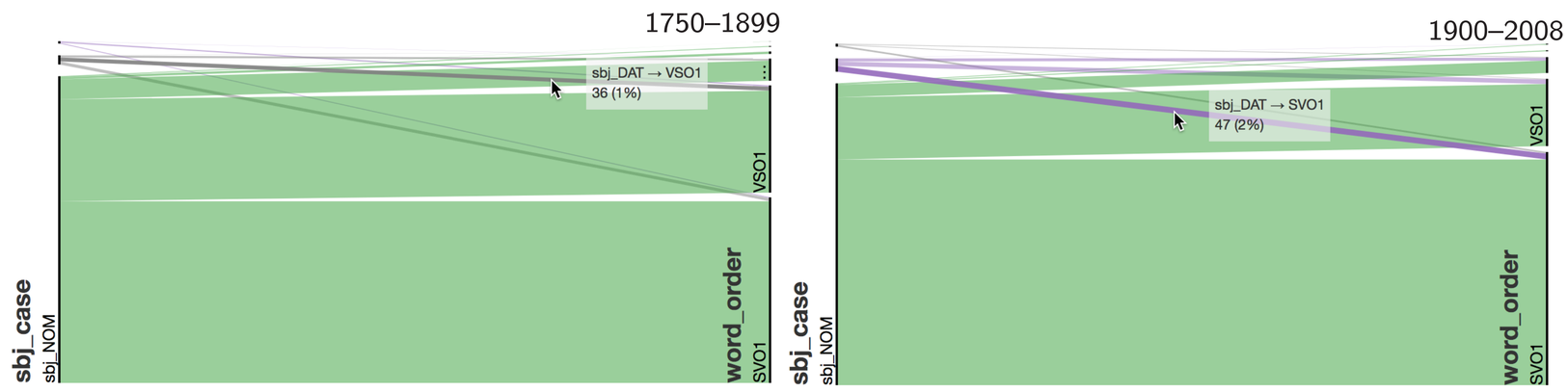


Difference Histograms



Dimension Interaction Visualization

- Visualizes dimension interactions as **Parallel Sets** (Bendix et al. 2005, Kosara et al. 2006).
  - Each feature is visualized as a proportion of an equally spaced vertical line.
  - The vertical lines represent the data dimensions.
- Flexible investigation of interactions between features from different data dimensions.
  → Particularly suitable for the analysis of historical linguistic data.



Case Study

- Investigation of the interaction between subject case and word order in the Icelandic Parsed Historical Corpus (IcePaHC, Wallenberg et al., 2011).
- **Compact Matrix:** Significant changes between the last two time periods.
- **Difference Histograms:**
  - SVO1 is the preferred word order overall (S=subject, V=verb, O1=primary object).
  - Frequency of SVO1 increases over time.
  - The use of dative subjects increases in the period post-1900.
- **Dimension Interactions:**
  - Dative subjects lag behind other subjects in being realized in a particular position.
  - Only as of 1900, dative subjects also occur preferably with SVO1.

→ Identification of a previously unknown interrelation between word order changes and subject case marking in Icelandic via HistoBankVis.

→ Dimension interactions visualization represents an effective new means for historical linguistics.

References

Bendix, F., Kosara, R., Hauser, H.: Parallel sets: Visual analysis of categorical data. In: IEEE Symposium on Information Visualization. pp. 133–140. IEEE (2005)
Kosara, R., Bendix, F., Hauser, H.: Parallel Sets: interactive exploration and visual analysis of categorical data. IEEE Transactions on Visualization and Computer Graphics 12(4), 558–568 (2006)
Wallenberg, J.C., Ingason, A.K., Sigurðsson, E.F., Rögnvaldsson, E.: Icelandic Parsed Historical Corpus (IcePaHC) (2011), `http://www.linguist.is/icelandic_treebank`, version 0.9
Schätzle, C., Hund, M., Dennig, F., Butt, M., Keim, D.: HistoBankVis: Detecting Language Change via Data Visualization. NoDaLiDa Workshop on Processing Historical Language (2017)

SFB-TRR 161

Project D02 "Evaluation Metrics for Visual Analytics in Linguistics"
Project A03 "Quantification of Visual Analytics Transformations and Mappings"