

Design Space for Diachronic Linguistic Visualizations in Theory and Practice

Christin Schätzle, Annette Hautli-Janisz, Michael Hund, Christian Rohrdantz, Miriam Butt, Daniel A. Keim

Visualisierungsprozesse in den Humanities
Universität Zürich
17th July, 2017

Methodological Challenge

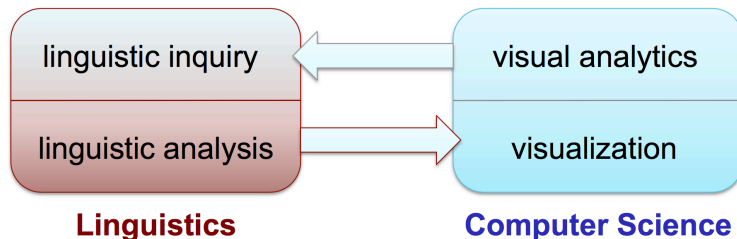
- Ever increasing availability of digitized data and annotated corpora for historical linguistic research (e.g., Newspaper Corpora, Penn Treebanks, etc.)
- Increased use of quantitative methods to analyze and evaluate data
- Programming languages specialized for text processing and statistical analysis (Python, Perl, R)

Problem: Diachronic investigations involve understanding highly complex interactions between various linguistic and extra-linguistic features and structures.

→ Meaningful patterns are difficult to see in the forest of numbers.

Opportunity: Visual Analytics for Linguistics (LingVis)

Visual Analytics for Linguistics (LingVis)



Overall interdisciplinary goal:

- Integrate methods from visual analytics into domains of linguistic inquiry
- Explore challenges based on the needs of linguistic analysis for visualization methods
- Visual Analytics Mantra (Keim et al., 2008): “Analyze first, show the important, zoom, filter and analyze further, details on demand”

⇒ Iterative and collaborative effort, merging knowledge with data-driven modeling


Visual Analytics for Linguistics (LingVis)



Emmanuelle Moureaux 'Forest of Numbers'

General Aim: turn complex data sets and their relationships into at-a-glance visualizations complemented by the possibility to work interactively with different visual perspectives of the same complex relationships

Design Space for Diachronic Visualizations

- Provide optimal solutions for mapping data values to visual representations
 - Different data dimensions: numerical, ordinal, categorical
 - Visual variables: color, position, shape, size, orientation (Bertin, 1983)
- Challenges/Problems:
 - Not every visual variable is well-suited for every kind of data dimension.
 - A good choice in one visualization might be a bad choice for another.
- Iterative design process
 - Domain experts: knowledge and hypotheses about the data
 - 
 - Visualization experts: visual components for analysis tasks and settings

⇒ Find a good design that fosters the emergence of visual patterns that point to relevant hidden patterns in the data

Time Dimension

- What does (not) change over time?
 - **Time resolution:** each data object can be considered to be a time-stamped observation of language or language use (e.g., a document or sentence)
 - Do we consider years, decades, or centuries?
 - **Distribution of observations over time:** Problems of data sparsity and overplotting may be avoided by using time sequences or aggregations into (fixed or variable) time frames instead of linearly scaled timelines
 - **Amount of observations:**
 - few data objects → individual visual representations
 - large amount of data objects → aggregate data objects by time
- ⇒ Historical linguistic data contains subspaces (e.g., interacting factors or relevant time periods) which need to be identified and understood

Data Dimensions

- Linguistic patterns under investigation
- Possibly interacting or correlating factors

	Manually edited data dimensions		Computed data dimensions	
	Manually created	Manually revised	Predefined	Open
Complexity of annotations	+++	++	+	+
Accuracy of annotations	+++	+++	++	+
Interpretability of results	+++	+++	++	+
Amount of data processed	+	++	+++	+++

Table: Overview of the different data dimensions and their characteristics

Visualizing Semantic Change (Rohrdantz et al., 2011)

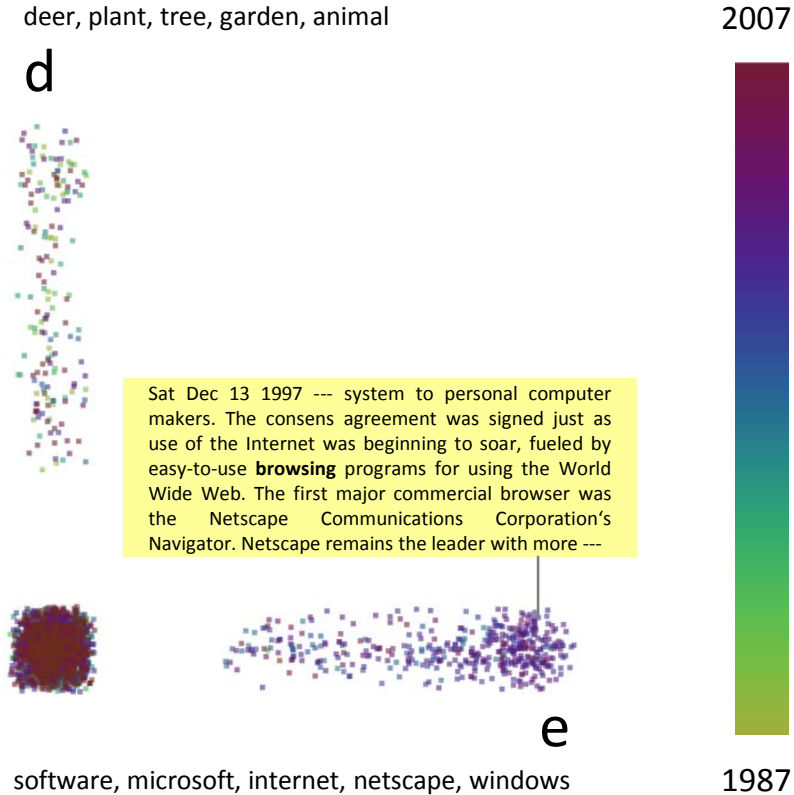
- **Goal:** investigate changes in word meaning by visually modeling and representing word contexts over time
- **Challenge:** track overall developments, identify starting points of change, comparison of prevailing senses
 - > provide an overview while allowing to delve into details of the data
- **New York Times Annotated Corpus**
 - 1.8 million newspaper articles from 1987 to 2007 (full coverage)
 - each article has a specific time stamp (year date)
- Words that have acquired a new sense due to the introduction of computing and the internet, e.g., 'to browse', 'to surf', 'bookmark'

Data Processing

- The meaning of a word is reflected by its immediate context (Firth, 1957)
- Extract context of 25 words before and after the lexical item under investigation
- Each context is complemented with the time stamp from the corpus.
- **Latent Dirichlet Allocation** (LDA; Blei et al., 2003) to statistically model word senses based on extracted word contexts → Open computed data dimensions
- Key words are represented as a numerical vector consisting of the frequencies of the LDA senses (= data dimensions)

⇒ **Aim:** Visualize the gradually overlapping senses as time progresses and make the stochastic analysis more transparent to the analyst

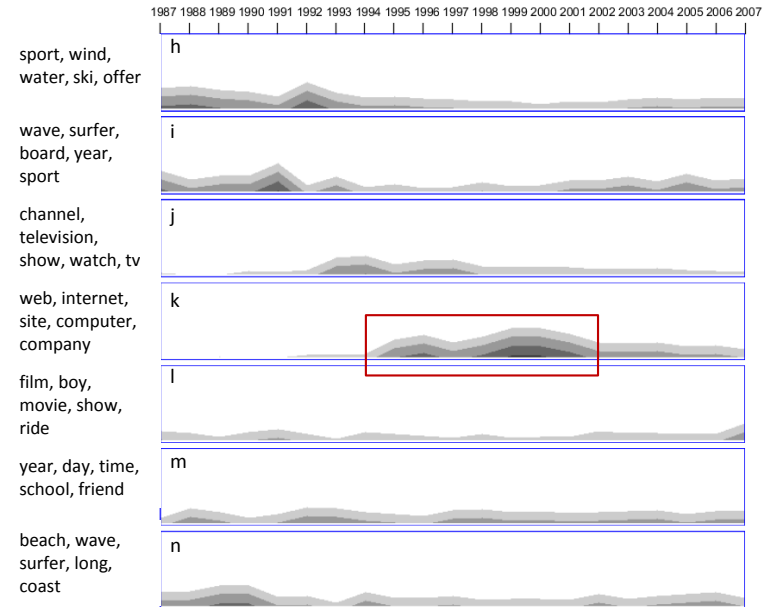
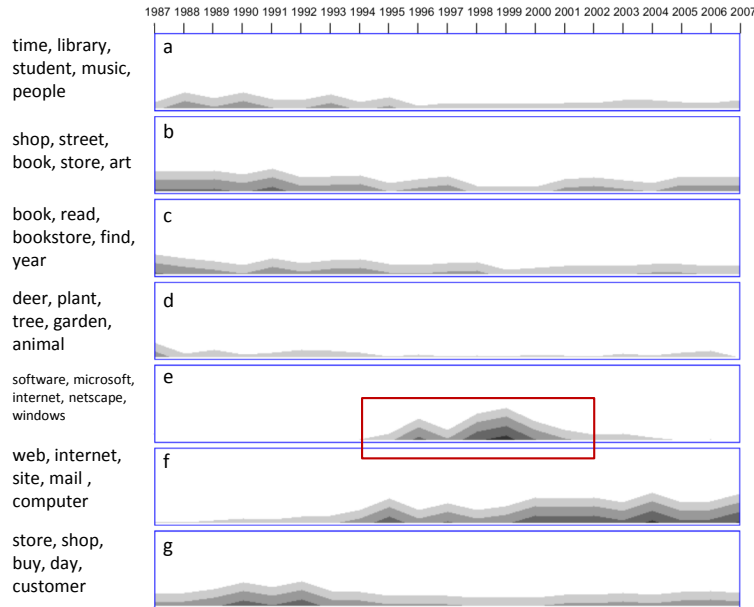
Visualization



Visualization

to browse

to surf



Visualizing Syntactic Change

- **Goal:** Investigate syntactic change in Icelandic with respect to two phenomena:
 - Dative subjects \longleftrightarrow lexical semantics and voice
 - Verb placement \longleftrightarrow topicality (null pronouns, expletives, definiteness) and verb types (unaccusatives vs. unergatives)
- **General research questions:**
 - Are subject case and word order interrelated?
 - Which strategies are used to mark grammatical relations in Icelandic?
 - Do these strategies change diachronically?

Icelandic Parsed Historical Corpus (IcePaHC)

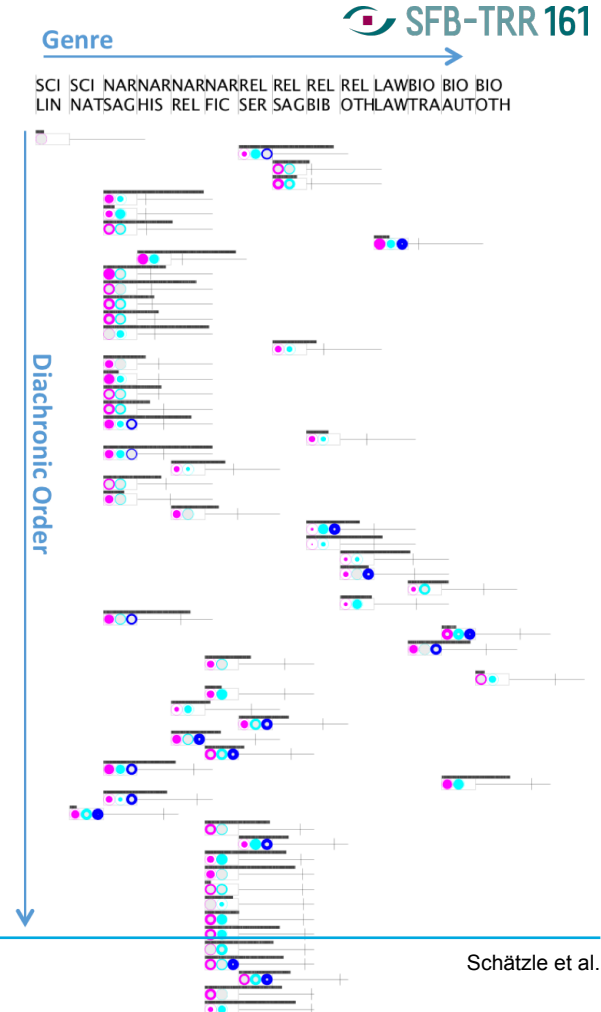
- 12th to 21st century – covers all attested stages of Icelandic
- 61 texts, 1 million words, different genres (not representative across centuries)
- annotated according to syntactic annotation of the Penn Treebank
 → Manually revised data dimensions
- provides information about sentence types, constituents, word order, grammatical relations, tense, voice, and case

```
(IP-MAT-SPE (NP-SBJ (PRO-D Mér-mér))
  (VBPI-PSY finnst-finna)
  (CP-ADV-SPE (WADVP-1 0)
    (C sem-sem)
    (IP-SUB-SPE (ADVP *T*-1)
      (NP-SBJ (PRO-N ég-ég))
      (BEPS sé-vera) (VBN sloppinn-sleppa)
      (PP (P úr-úr) (NP (NP-POS (ONE+Q-G einhvers-einhver)
        (N-G konar-konar)) (N-D fangelsi-fangelsi))))))
  (. .-.))


(ID 1882.TORFHILDUR.NAR-FIC,.603))
```

Glyph visualization

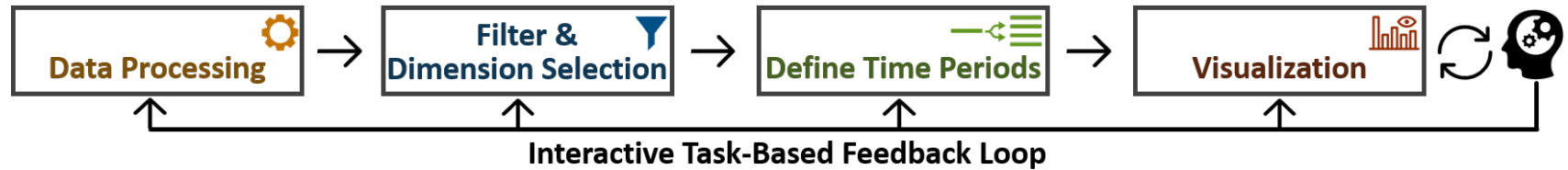
- Visualization of V1 order and dative subjects in Icelandic (Butt et al., 2014; Schaeztle et al., 2016)
- Glyph visualization allowing for a comparative diachronic evaluation
- ‘Overview first – details on demand’ (Shneiderman 1996, Keim et al. 2010)
- Compact presentation of large amounts of data
- Exploratory and confirmatory data analysis
- Generation and validation of hypotheses
- Helps to uncover and understand multidimensional interactions of factors



Drawbacks

- Glyph visualization (Butt et al., 2014; Schätzle et al., 2016):
 - Only allows for the analysis of a given amount of data dimensions (factors) and predefined interactions thereof → relies on specific assumptions about the nature of the data
 - At-a-glance identification of salient patterns merely impossible for a large number of features (categories)
 - -Complex design is difficult to interpret for the uninformed user
- **Novel Approach: HistoBankVis** 
 - Generically applicable system for historical linguistic research
 - More flexible investigation of data dimensions allowing for exploratory access to a potentially high number of factors
 - Look at each factor individually or interactions of interrelated factors on demand by breaking analysis process up into individual steps

HistoBankVis



Data Processing

- Extraction of relevant linguistic data dimensions from the annotation of IcePaHC via Perl scripts → verb type, voice, word order, case and valency
- Information is collected for each matrix declarative sentence and mapped onto its sentence ID → information about the age, name, and genre of each text

ID	VERB	VERB_TYPE	MODAL/ASP	VOICE	WORD_ORDER	VALENCY	SBJ_CASE	OBJ_CASE	OBJ2_CASE
1150.FIRSTGRAMMAR.SCI-LIN,,1	setja	VB	-	active	VSO1	trans	sbj_NOM	obj1_ACC	-
1150.FIRSTGRAMMAR.SCI-LIN,,2	setja	VB	-	active	O1VS	trans	sbj_NOM	obj1_ACC	-
1150.FIRSTGRAMMAR.SCI-LIN,,3	hafa	HV	þurfa	active	SVO1	trans	sbj_NOM	-	-
1150.FIRSTGRAMMAR.SCI-LIN,,4	rita	VB	-	active	VSO1	trans	sbj_NOM	obj1_ACC	-
1150.FIRSTGRAMMAR.SCI-LIN,,5	verða	RD	-	active	VS	intrans	sbj_GEN	-	-
1150.FIRSTGRAMMAR.SCI-LIN,,6	ganga	VB	-	active	VS	intrans	sbj_NOM	-	-
1150.FIRSTGRAMMAR.SCI-LIN,,7	rita	VB	-	active	VSO1	trans	sbj_NOM	obj1_ACC	-
1150.FIRSTGRAMMAR.SCI-LIN,,8	hafa	HV	-	active	VS	intrans	sbj_NOM	-	-
1150.FIRSTGRAMMAR.SCI-LIN,,9	taka	VB	-	active	O1VS	trans	sbj_NOM	obj1_ACC	-
1150.FIRSTGRAMMAR.SCI-LIN,,10	rita	VB	-	active	VSO2O1	ditrans	sbj_NOM	obj1_ACC	obj2_DAT
1150.FIRSTGRAMMAR.SCI-LIN,,11	taka	VB	-	passive	VS	intrans	sbj_NOM	-	-
1150.FIRSTGRAMMAR.SCI-LIN,,12	taka	VB	-	passive	VS	intrans	sbj_NOM	-	-
1150.FIRSTGRAMMAR.SCI-LIN,,13	taka	VB	-	passive	VS	intrans	sbj_NOM	-	-

Analyzing Change over Time

- Define/select time periods 



Predefined Ranges:

- 1150-1549, 1550-2008
- 1150-1349, 1350-1549, 1550-1749, 1750-1899, 1900-2008

Custom Ranges

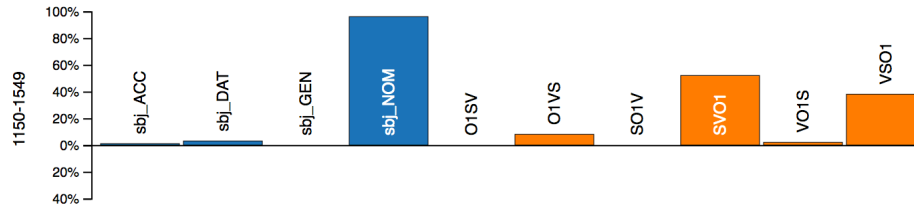
Add Range

Split in **Ranges**

- Compact Matrix Visualization 
 - Visualizes differences between selected dimensions across time
 - Measure of quality and “interestingness”
- Difference Histograms Visualization 

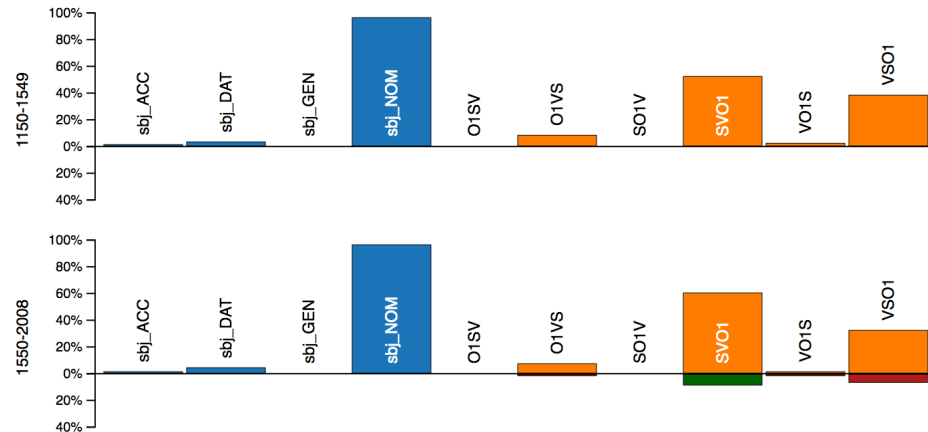
Difference Histograms Visualization

- Histograms provide detailed views on individual features and their diachrony.
- Each time period is visualized as one bar chart/histogram.
- Dimensions are encoded via different colors.
- Each bar in the histogram corresponds to an individual feature.
- The height of a bar shows the percentage of sentences containing the respective feature in the given time period.



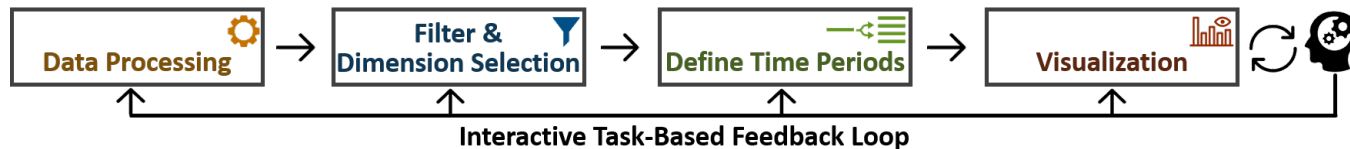
Difference Histograms Visualization

- Differences between periods are visualized as a separate bar chart below each bar:
 - green → feature increased
 - red → feature decreased
- Different comparison modes:
 - Previous period
 - First range
 - Last range
 - Average of all ranges
 - Average of previous ranges



Hypothesis Generation and Feedback ↻

- Generation and testing of new hypotheses
- Feed the knowledge gained back into the system:
 - Change filters
 - Select different dimensions
 - Use different time periods
 - Process data anew
- Iterative analysis process
- Combination of knowledge-based and data-driven modeling



Preliminary Results of the Visual Analysis

Overall move towards SVO

- Development of a fixed preverbal subject position in the history of Icelandic
 - **19th century is major key turning point**
 - Dative subjects show a slower tendency to be realized in a particular position
 - Experiencer/goal arguments are not canonical subjects
 - Have to undergo reanalysis from object to subject first
 - Functional pressure on experiencers to be realized as subjects (cf. Aissen, 1999; Dowty, 1991)
 - Decrease of V1 (Sigurdsson 1990, Butt et al. 2014) and the loss of OV (Hróarsdóttir 2000) happen around the same time
- ⇒ Evidence against the Proto Indo-European inheritance of a monolithic dative subject construction

Conclusion

- HistoBankVis is an effective and powerful visualization tool which facilitates the detection and analysis of historical linguistic data.
- The **design space** for diachronic linguistics can be realized very differently across visualization applications → depends on the type of data and the corresponding linguistic research question
- The presented visualization systems support the generation of **novel hypotheses** by granting an **interactive and exploratory access** to the data.
- Close **collaboration** between domain experts
 - Combination of knowledge-based and data-driven modeling
 - Iterative design process
 - Case studies help to refine and optimize the visualization

Thank you!
Questions?

<http://histobankvis.dbvis.de/>

Acknowledgement

This work was funded by the German Research Foundation (DFG) within projects D02 “Evaluation Metrics for Visual Analytics in Linguistics” and A03 “Quantification of Visual Analytics Transformations and Mappings” of SFB/Transregio 161 and within BU 1806/7-1 “Visual Analysis of Language Change and Use Patterns”.