Universität Konstanz

Visualizing Linguistic Change as Dimension Interactions

Christin Schätzle, Frederik L. Dennig, Michael Blumenschein, Miriam Butt, Daniel A. Keim

1st International Workshop on Computational Approaches to Historical Language Change
ACL2019

DFG Deutsche Forschungsgemeinschaft

SFB-TRR 161
Quantitative Methods for Visual Computing
Transregional Collaborative Research Center

Methodological challenges for historical linguistics

- Ever increasing availability of digitized data and annotated corpora for historical linguistic research (e.g., Penn Treebanks, Dependency Treebanks, etc.)

```
(IP-MAT-SPE (NP-SBJ (PRO-D Mér-mér))
          (VBPI 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))
```

Annotation sample from IcePaHC (Wallenberg et al. 2011)

- Increased use of **quantitative methods** to analyze and evaluate data
- Programming languages specialized for text processing and statistical analysis (Python, R, …)

Methodological challenges for historical linguistics

- **Standard procedure:** calculation of co-occurrence frequencies and statistical significances for a multitude of different linguistic features across different time stages
- Generation of a multitude of high-dimensional data tables of varying size containing numbers computed for different linguistic features

Texts	Indefinite NPs			Definite NPs			NPs as proper names		
	OV	VO	% OV	OV	VO	% OV	OV	VO	% OV
14th century	28	33	45.9%	11	57	16.2%	3	8	27.3%
15th century	23	30	43.4%	10	25	28.6%	1	3	25.0%
16th century	15	28	34.9%	17	26	39.5%	1	5	16.7%
17th century	28	59	32.2%	18	50	26.5%	0	20	0.0%
18th century	6	28	17.6%	7	31	18.4%	1	7	12.5%
19th century	34	425	7.4%	14	351	3.8%	4	68	5.6%
	134	603	18.2%	77	540	12.5%	10	111	8.3%

Definiteness distribution of NPs across different word orders in Icelandic (Hróarsdóttir 2000, 136)

- **Aim:** **identify** the linguistic features and structures involved in a change; understand how they **interact** across the **temporal dimension**

Methodological challenges for historical linguistics

- Finding significant patterns and feature interactions is challenging:
  - Pair-wise comparison of the relevant bits of information across various tables
  - Statistical significances are often calculated on the basis of only very few occurrences of the actual observation (data sparsity)
  - Interesting patterns may stay hidden when the temporal episodes chosen for the statistical analysis are too fine or too coarse grained
  - The factors causing a change are often unknown (or at least highly debated)

- **Opportunity: Visual Analytics for Linguistics (LingVis)**

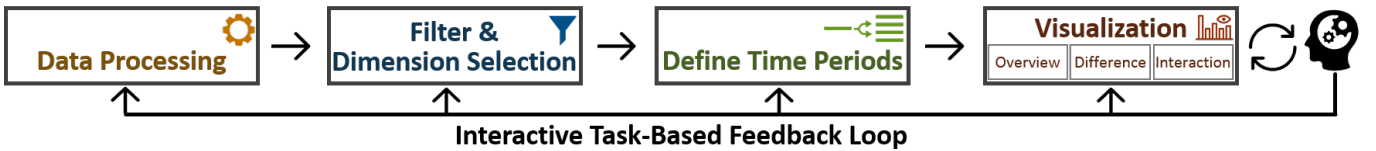
  $\longrightarrow$ turn complex data sets and their relationships into at-a-glance visualizations

  $\longrightarrow$ provide an interactive exploratory access to the data

"Analyze first, show the important, zoom, filter and analyze further, details on demand"
(Keim et al. 2008)

HistoBankVis: a multilayer visualization system

- Generically applicable system for **historical linguistic research**
- Flexible investigation of a potentially high number of interacting linguistic features stored in an SQL database



- Combination of different visualization layers and filtering techniques with a structured statistical analysis process $\longrightarrow$ exploratory access to a high-dimensional data set
  - **Overview:** Compact Matrix Visualization
  - **Difference** Histograms Visualization
  - **New component:** Dimension **Interaction** Visualization

HistoBankVis: a multilayer visualization system

- **On-going work:** investigation of **syntactic change in Icelandic** based on the Icelandic Parsed Historical Corpus (IcePaHC; Wallenberg et al. 2011)
- Automatic extraction of the relevant linguistic factors from the IcePaHC annotation
  $\longrightarrow$ verb type, voice, subject case, word order, subject position, V1 (verb-first)

ID	VERB	VERB_TYPE	MODAL/ASP	VOICE	SBJ_CASE	OBJ_CASE	OBJ2_CASE	WORD_ORDER	SUBJ_POSITION	V1
1150.FIRSTGRAMMAR.SCI-LIN,.1	setja	VB	-	active	sbj_NOM	obj1_ACC	-	VSO1	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.2	setja	VB	-	active	sbj_NOM	obj1_ACC	-	O1VS	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.3	hafa	HV	þurfa	active	sbj_NOM	obj1_ACC	-	SVO1	prefinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.4	rita	VB	-	active	sbj_NOM	obj1_ACC	-	VSO1	postfinite	yes
1150.FIRSTGRAMMAR.SCI-LIN,.5	verða	RD	-	active	sbj_GEN	-	-	VS	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.6	ganga	VB	-	active	sbj_NOM	-	-	VS	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.7	rita	VB	-	active	sbj_NOM	obj1_ACC	-	VSO1	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.8	hafa	HV	-	active	sbj_NOM	-	-	VS	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.9	taka	VB	-	active	sbj_NOM	obj1_ACC	-	O1VS	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.10	rita	VB	-	active	sbj_NOM	obj1_ACC	obj2_DAT	VSO2O1	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.11	taka	VB	-	passive	sbj_NOM	-	-	VS	postfinite	no

HistoBankVis: a multilayer visualization system

- **On-going work:** investigation of **syntactic change in Icelandic** based on the Icelandic Parsed Historical Corpus (IcePaHC; Wallenberg et al. 2011)
- Automatic extraction of the relevant linguistic factors from the IcePaHC annotation
  $\longrightarrow$ verb type, voice, subject case, word order, subject position, V1 (verb-first)

Data dimensions

ID	VERB	VERB_TYPE	MODAL/ASP	VOICE	SBJ_CASE	OBJ_CASE	OBJ2_CASE	WORD_ORDER	SUBJ_POSITION	V1
1150.FIRSTGRAMMAR.SCI-LIN,.1	setja	VB	-	active	sbj_NOM	obj1_ACC	-	VSO1	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.2	setja	VB	-	active	sbj_NOM	obj1_ACC	-	O1VS	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.3	hafa	HV	þurfa	active	sbj_NOM	obj1_ACC	-	SVO1	prefinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.4	rita	VB	-	active	sbj_NOM	obj1_ACC	-	VSO1	postfinite	yes
1150.FIRSTGRAMMAR.SCI-LIN,.5	verða	RD	-	active	sbj_GEN	-	-	VS	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.6	ganga	VB	-	active	sbj_NOM	-	-	VS	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.7	rita	VB	-	active	sbj_NOM	obj1_ACC	-	VSO1	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.8	hafa	HV	-	active	sbj_NOM	-	-	VS	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.9	taka	VB	-	active	sbj_NOM	obj1_ACC	-	O1VS	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.10	rita	VB	-	active	sbj_NOM	obj1_ACC	obj2_DAT	VSO2O1	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.11	taka	VB	-	passive	sbj_NOM	-	-	VS	postfinite	no

HistoBankVis: a multilayer visualization system

- **On-going work:** investigation of **syntactic change in Icelandic** based on the Icelandic Parsed Historical Corpus (IcePaHC; Wallenberg et al. 2011)
- Automatic extraction of the relevant linguistic factors from the IcePaHC annotation
  $\longrightarrow$ verb type, voice, subject case, word order, subject position, V1 (verb-first)

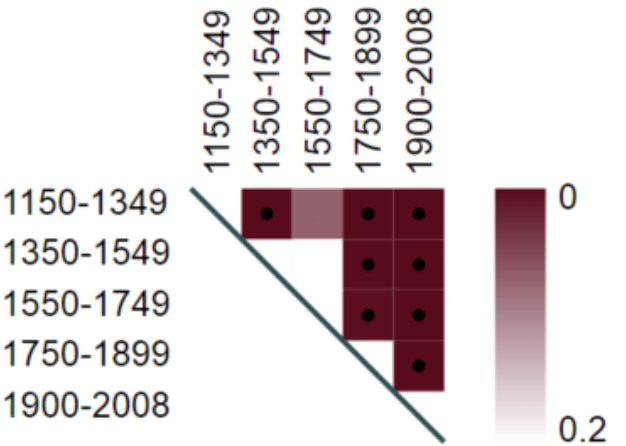
Data dimensions

ID	VERB	VERB_TYPE	MODAL/ASP	VOICE	SBJ_CASE	OBJ_CASE	OBJ2_CASE	WORD_ORDER	SUBJ_POSITION	V1
1150.FIRSTGRAMMAR.SCI-LIN,.1	setja	VB	-	active	sbj_NOM	obj1_ACC	-	VSO1	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.2	setja	VB	-	active	sbj_NOM	obj1_ACC	-	O1VS	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.3	hafa	HV	þurfa	active	sbj_NOM	obj1_ACC	-	SVO1	prefinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.4	rita	VB	-	active	sbj_NOM	obj1_ACC	-	VSO1	postfinite	yes
1150.FIRSTGRAMMAR.SCI-LIN,.5	verða	RD	-	active	sbj_GEN	-	-	VS	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.6	ganga	VB	-	active	sbj_NOM	-	-	VS	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.7	rita	VB	-	active	sbj_NOM	obj1_ACC	-	VSO1	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.8	hafa	HV	-	active	sbj_NOM	-	-	VS	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.9	taka	VB	-	active	sbj_NOM	obj1_ACC	-	O1VS	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.10	rita	VB	-	active	sbj_NOM	obj1_ACC	obj2_DAT	VSO2O1	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.11	taka	VB	-	passive	sbj_NOM	-	-	VS	postfinite	no

Features

Compact Matrix Visualization

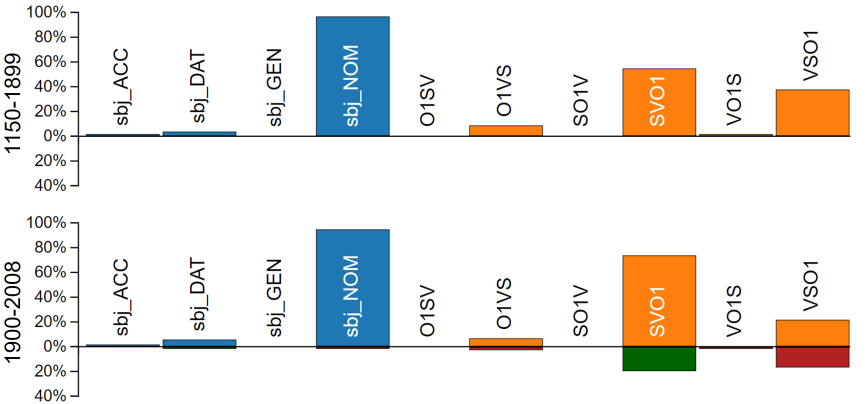
- Visualizes **differences between dimensions** across time stages
- Differences mapped onto a colormap
- Comparison of consecutive periods along the diagonal
- Two comparison modes:
  - χ^2-test
    - Statistical significance ($\alpha \leq 0.05$)
    - Absence of necessary preconditions ×
    - p-value is mapped to colormap (red $p = 0$, white $p \geq 0.2$)
  - Euclidean distance
    - Colormap indicates high (red) or low (white) distance
    - High Euclidean distance → large difference (high significance)



Difference Histograms Visualization

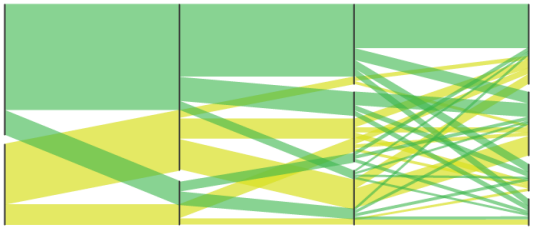
- Difference histograms provide **details on diachrony of features and dimensions**
- Each time period is visualized as one composed **bar chart/histogram**
- Dimensions are encoded via different colors for parallel inspection
- Each bar represents an individual feature of a dimension
- Bar height corresponds to the percentage of sentences containing a feature

- Differences between neighboring time periods as separate bar chart below feature bar:
  - red → feature decrease
  - green → feature increase
- Different comparison modes available



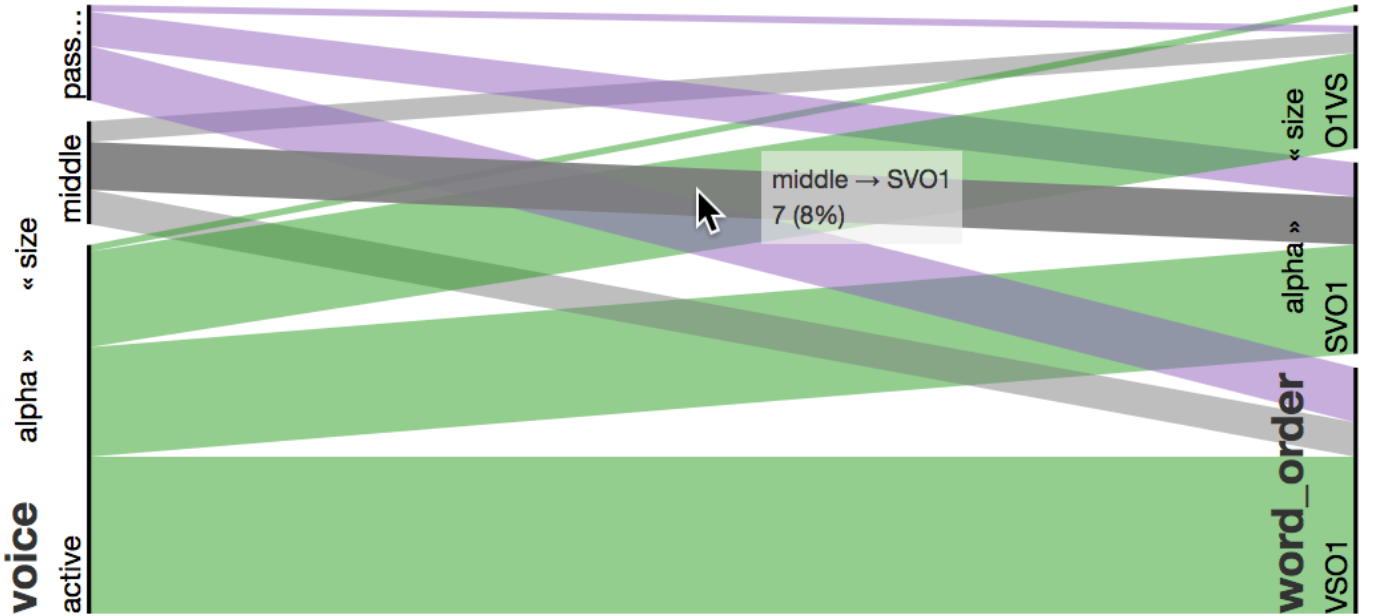
Dimension Interaction Visualization

- Dimension interactions provide insights into the **interrelation between multiple features** of different dimensions
- Application of the **Parallel Sets** technique (Bendix et al. 2005, Kosara et al. 2006)
  - Feature frequencies are visualized as proportions of equally spaced vertical lines (data dimensions)
  - Dimensions are connected by colored ribbons
  - Size of a ribbon $\rightarrow$ share which a feature holds of a feature from another dimension



- Each time period is visualized as one Parallel Sets visualization
  - Dimensions can be reordered via drag&drop
  - Features can be sorted according to their size or alphabetically
  - Details about feature correspondences can be accessed via mouse over

Dimension Interaction Visualization



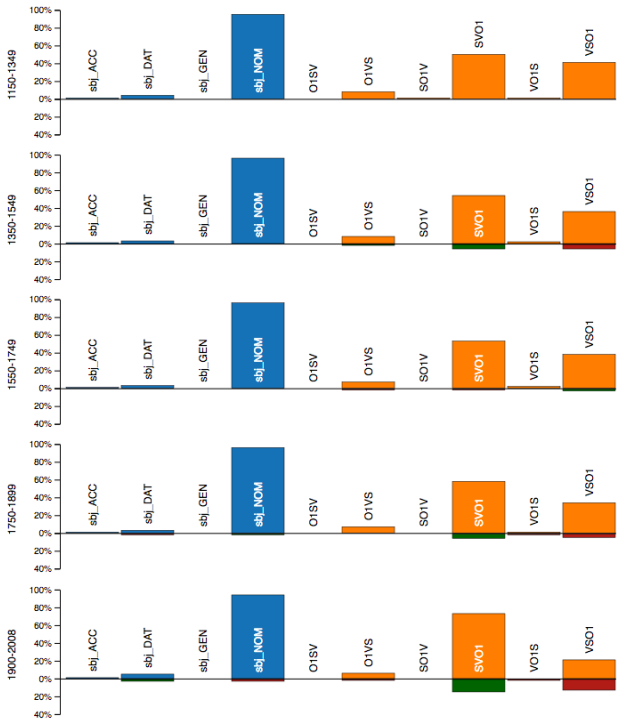
Tracking syntactic change with HistoBankVis

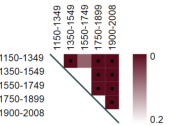
- **Case study:** interaction between **subject case** and **word order** in IcePaHC
- Previous studies on syntactic change in Icelandic:
  - Use of dative subjects increases diachronically (e.g., Barðdal 2011)
  - Word order becomes more rigid over time (e.g., Rögnvaldsson 1996)

  $\Longrightarrow$ **Are these changes interrelated?**

- **Investigation using HistoBankVis:**
  - Dimension selection: subject case & word order
  - Filtering for transitive sentences: sentences with a subject (S), verb (V), and a direct object (O1)
  - Time periods: 1150-1349, 1350-1549, 1550-1749, 1750-1899, 1900-2008

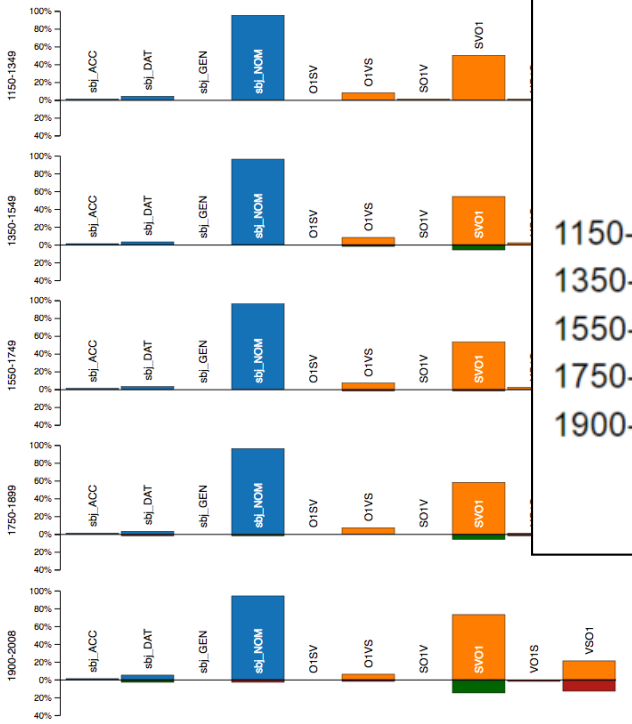
Subject case and word order in IcePaHC

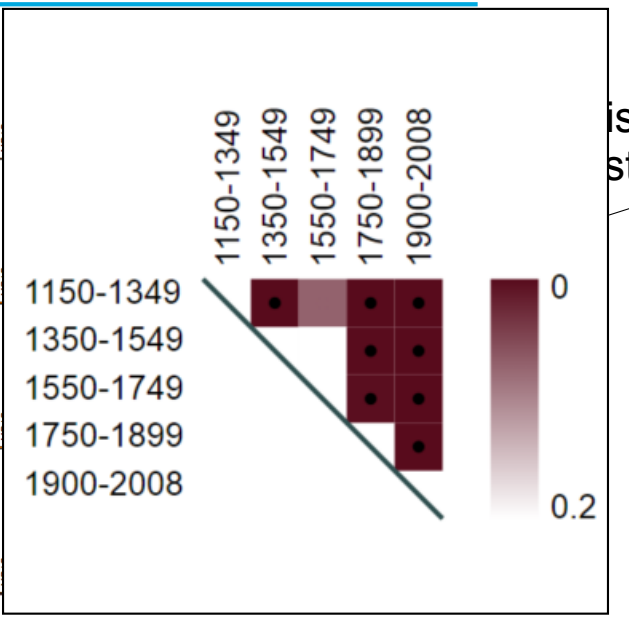




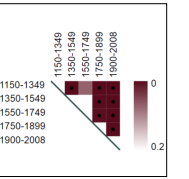
- **Compact matrix:** Distribution changes significantly in the last two periods

Subject case and word order in IcePaHC

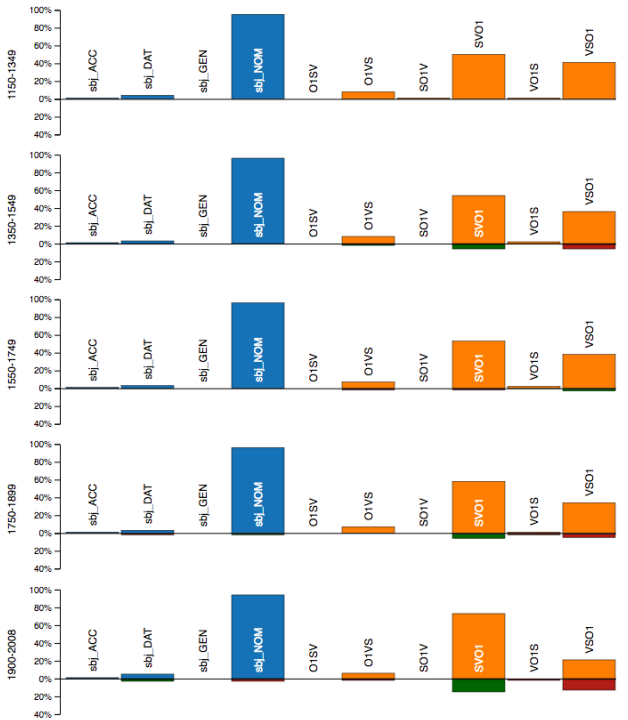


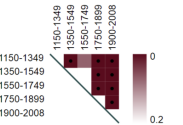


istribution changes
st two periods



Subject case and word order in IcePaHC

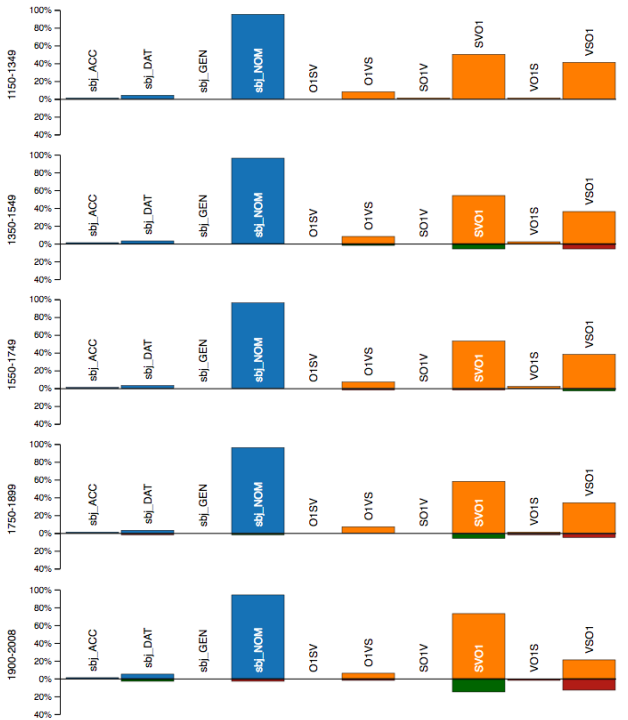


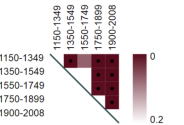


- **Compact matrix:** Distribution changes significantly in the last two periods
- **Difference histograms:**
  - SVO1 most frequent word order
  - **SVO1 increases** over time; largest increase in the period 1900-2008
  - VSO1 decreases concomitantly
  - Subjects are most often nominative
  - **Dative subjects increase** strikingly after 1900

Subject case and word order in IcePaHC

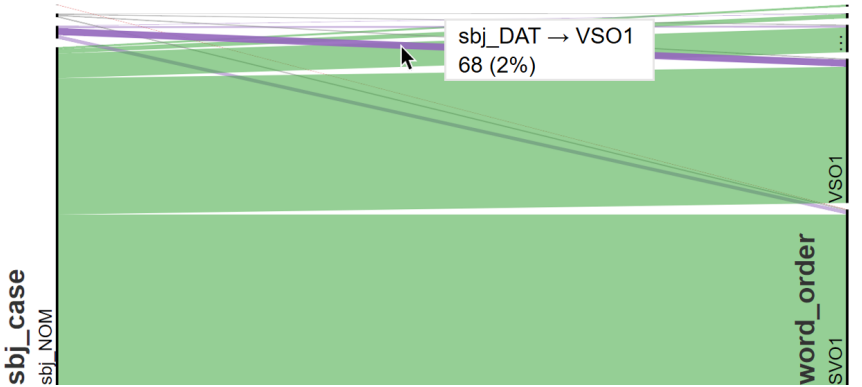
Subject case and word order in IcePaHC





- **Compact matrix:** Distribution changes significantly in the last two periods
- **Difference histograms:**
  - SVO1 most frequent word order
  - **SVO1 increases** over time; largest increase in the period 1900-2008
  - VSO1 decreases concomitantly
  - Subjects are most often nominative
  - **Dative subjects increase** strikingly after 1900
- Subject case and word order change at the same time $\Longrightarrow$ **Interrelation?**

Dimension interactions – subject case and word order

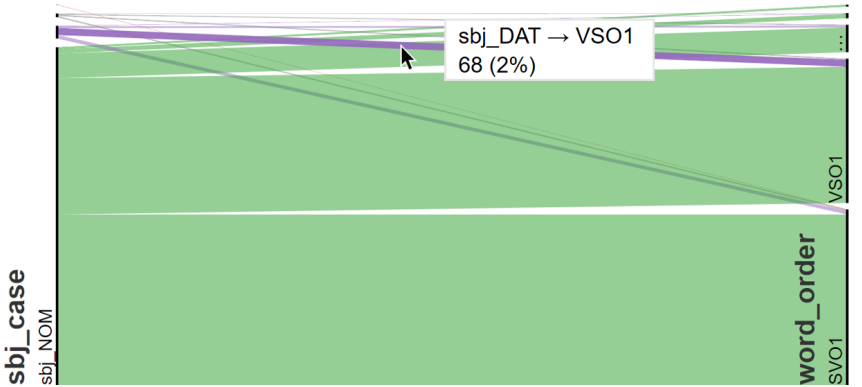


Dimension interaction subject case & word order 1150-1349

Dimension interaction 1150-1349:

- Nominative subjects: shares of SVO1 and VSO1 equal
- **Dative subjects:** large majority are **VSO1**

Dimension interactions – subject case and word order



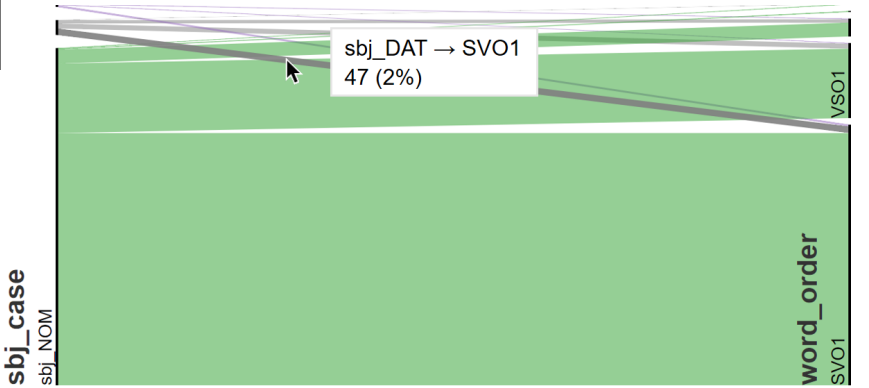
Dimension interaction subject case & word order 1150-1349

Dimension interaction 1900-2008:

- SVO1 dominant word order overall
- Share of **SVO1 with dative subjects smaller** than with nominatives

Dimension interaction 1150-1349:

- Nominative subjects: shares of SVO1 and VSO1 equal
- **Dative subjects:** large majority are **VSO1**



Dimension interaction subject case & word order 1900-2008

Dimension interactions – subject case, word order, and voice

- Dative subjects lag behind with respect to the overall word order changes
- Voice (i.e., passivization, middle formation) influences the occurrence of dative subjects in Icelandic (see Zaenen et al. 1985, Sigurðsson 1989)

Dimension interactions – subject case, word order, and voice

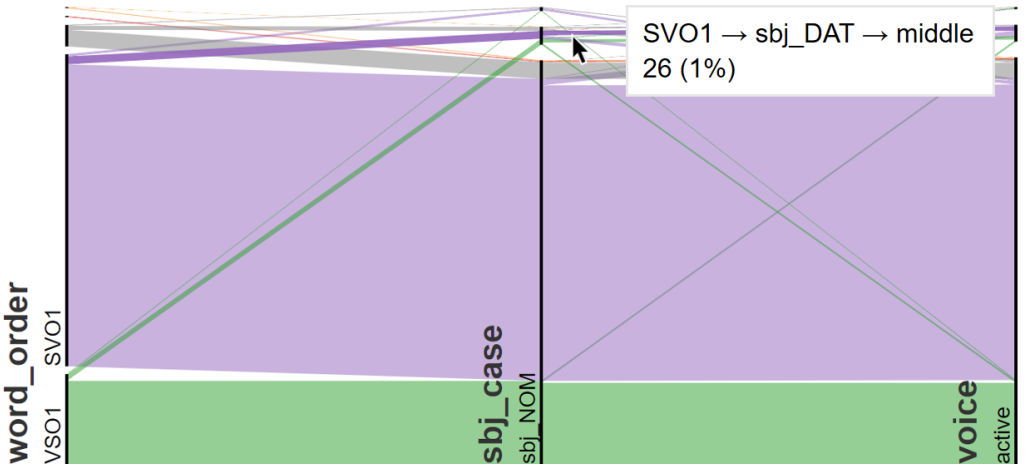
- Dative subjects lag behind with respect to the overall word order changes
- Voice (i.e., passivization, middle formation) influences the occurrence of dative subjects in Icelandic (see Zaenen et al. 1985, Sigurðsson 1989)

$\Longrightarrow$ **Correlation between voice, subject case, and word order?**

Dimension interactions – subject case, word order, and voice

- Dative subjects lag behind with respect to the overall word order changes
- Voice (i.e., passivization, middle formation) influences the occurrence of dative subjects in Icelandic (see Zaenen et al. 1985, Sigurðsson 1989)

⟹ **Correlation between voice, subject case, and word order?**



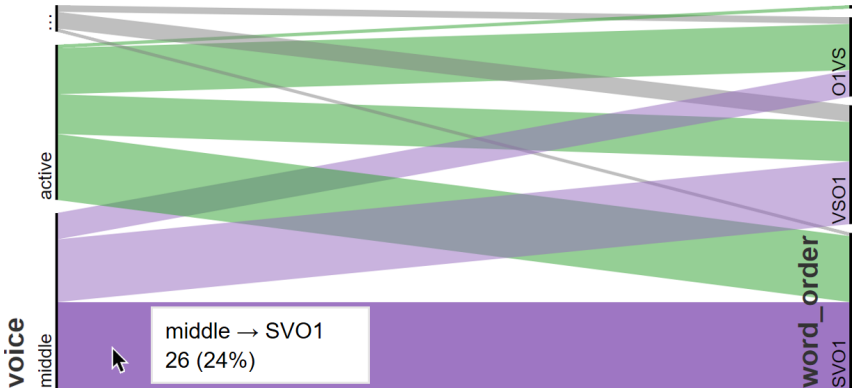
Dimension interaction word order, subject case & voice 1900-2008

- **Nominative subjects:** SVO1 most often in **active** constructions
- **Dative subjects:** SVO1 mainly with **middles**

⇓

Closer look at dative subjects and voice

Dimension interactions – Dative subjects and voice

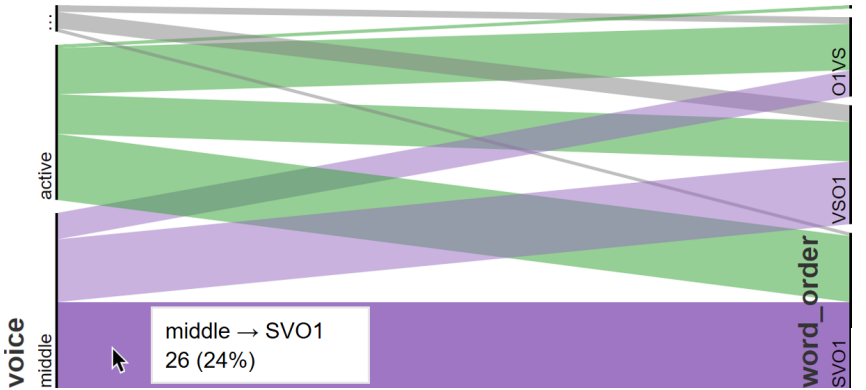


Dimension interaction voice & word order 1900-2008 (dative subjects)

Dimension interaction 1900-2008:

- Dative subjects most frequently with **middle voice**
- **SVO1** most prominent order with all voices

Dimension interactions – Dative subjects and voice



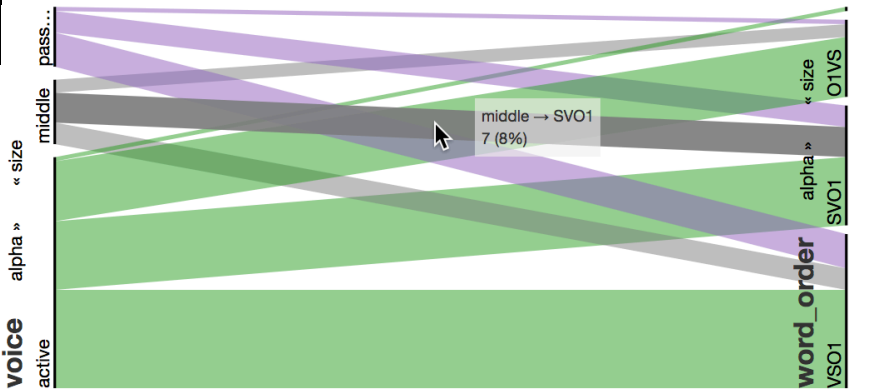
Dimension interaction voice & word order 1900-2008 (dative subjects)

Dimension interaction 1750-1899:

- Dative subjects most frequently in **active constructions**
- **VSO1** is most prominent
- **Middles** most often with **SVO1**

Dimension interaction 1900-2008:

- Dative subjects most frequently with **middle voice**
- **SVO1** most prominent order with all voices



Dimension interaction voice & word order 1700-1899 (dative subjects)

Summary and conclusion

- Dative subjects **lag behind nominative subjects** with respect to their realization in the preverbal position (SVO1)
- Increasing use of dative subjects in the preverbal position correlates with an **increase of dative subjects with middle voice**
- HistoBankVis is an efficient and powerful tool for historical linguistic investigations
  - provides multiple perspectives of the data at different levels of detail
  - fosters iterative process of hypothesis testing and generation
- **Dimension interaction visualization:**
  - Interactive visualization of complex interactions across different dimensions
  - First use of **Parallel Sets** in LingVis
  - Effective new means for historical linguistic research

  $\Longrightarrow$ Identification of previously unknown link between dative subjects, word order, and voice within minutes

Thank you!
Questions?

christin.schaetzle@uni-konstanz.de

http://histobankvis.dbvis.de/

Acknowledgement

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161 (Projects D02 “Evaluation Metrics for Visual Analytics in Linguistics” and A03 “Quantification of Visual Analytics Transformations and Mappings”).