

HistoBankVis: Detecting Language Change via Data Visualization

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

Processing Historical Language
NoDaLiDa 2017

Methodological Challenge

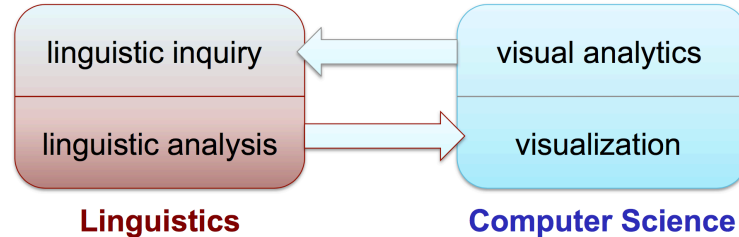
- Ever increasing availability of digitized data and annotated corpora for historical linguistic research (e.g., Newspaper Corpora, Penn Treebanks, etc.)
- Increased use of quantitative methods to analyze and evaluate data
- Programming languages specialized for text processing and statistical analysis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

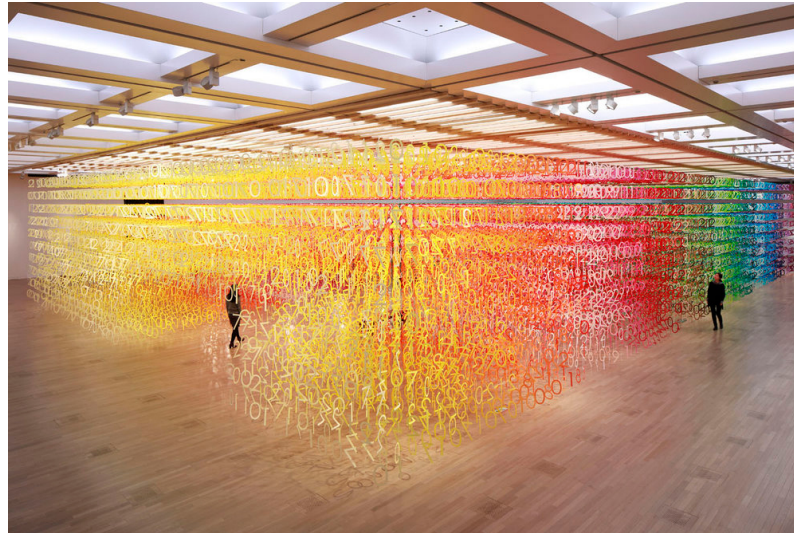
Visual Analytics for Linguistics (LingVis)



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”
- Exploratory and interactive access to data
- Iterative process of hypothesis formation and hypothesis testing

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

Research Context

- Interdisciplinary collaboration within SFB-TRR 161 “Quantitative Methods for Visual Computing”
- Project D02 “Evaluation Metrics for Visual Analytics in Linguistics”
 - Language change in Germanic and Indo-Aryan
 - Research involves working with raw texts as well as annotated corpora (this talk: Penn Treebanks)
 - Experiment with different VA possibilities for different types of data
- Project A03 “Quantification of Visual Analytics Transformations and Mappings”
 - Quantification of data transformations → subspace analysis
 - Identification of subspaces in larger amounts of high-dimensional data

⇒ Historical linguistic data contains subspaces (e.g., interacting factors or relevant time periods) which need to be identified and understood

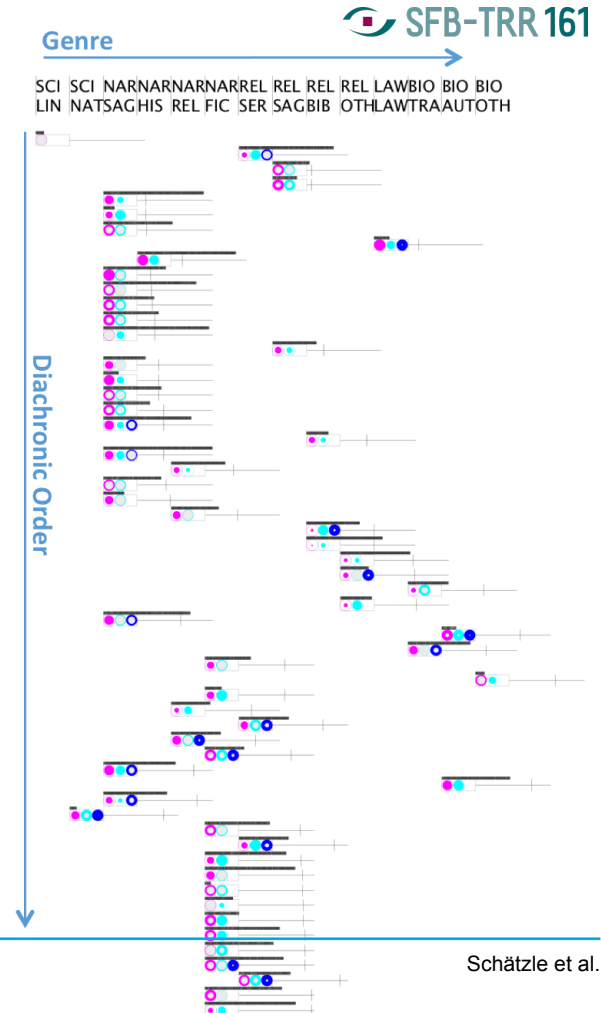
Icelandic Parsed Historical Corpus (IcePaHC)

- 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 (Marcus et al. 1993)
- provides information about sentence types, constituents, word order, grammatical relations, tense, voice, and case


```
(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))
```

Visualizing Syntactic Change in IcePaHC

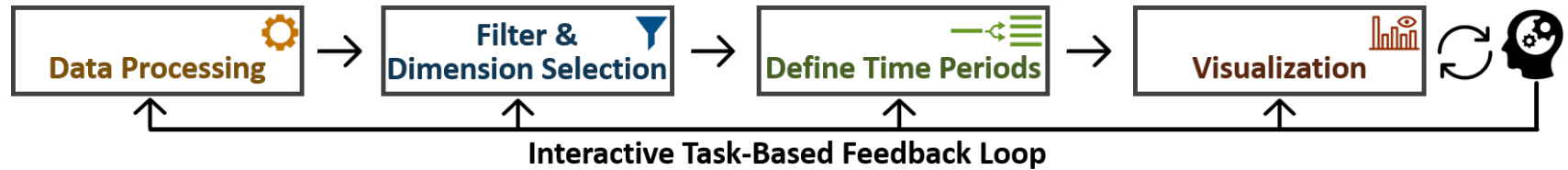
- Study of the diachrony of dative subjects in Icelandic (Schätzle et al., 2016)
- Glyph visualization allowing for a comparative diachronic evaluation
- ‘Overview first – details on demand’ (Shneiderman 1996, Keim et al. 2010)
- Compact presentation of large amounts of data
- Exploratory and confirmatory 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
 - At-a-glance identification of salient patterns merely impossible for a large number of features (categories)
 - 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

HistoBankVis



Data Processing

- Concrete case study: interaction between subject case and word order in the history of Icelandic
- Reported word order changes in Icelandic:
 - change from OV to VO (Kiparsky 1996, Rögnvaldsson 1996, Hróarsdóttir 2000)
 - decrease of V1 (Franco 2008, Sigurðsson 1990, Butt et al. 2014)
- Research questions:
 - Which strategies are used to mark grammatical relations in Icelandic?
 - Do these strategies change diachronically?

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_ORD	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	-	-

Task-based Filtering

Sentence Filter

From year to

[Edit Filter](#) [Reset Filter](#) [Apply Filter](#)

Dimension	Features
sbj_case	sbj_DAT
word_order	wo_O1VS

Result Table

[Export Records](#) [Continue to Visualization](#)

ID	sbj_case	voice	word_order	verb
1790.FIMMBRAEDRA.NAR-SAG,.662	sbj_DAT	active	wo_O1VS	lika
1790.FIMMBRAEDRA.NAR-SAG,.382	sbj_DAT	active	wo_O1VS	vera
1791.JONSTEINGRIMS.BIO-AUT,154.1431	sbj_DAT	active	wo_O1VS	batna

- Explore data set before visualization
- Construction of a task-specific data set
- Filter for sentences with relevant properties
 - Specific time frame
 - Specific features (i.e., entries in cells)
 - SQL-like filter construction (AND- or OR-functions)
- Dimension selection
 - Dimensions (i.e., columns) to be displayed in result table
 - Dimensions to be analyzed in visualization

Task-based Filtering

Sentence Filter

From year to

[Edit Filter](#) [Reset Filter](#) [Apply Filter](#)

Dimension	Features
sbj_case	sbj_DAT
word_order	wo_O1VS

Result Table

[Export Records](#) [Continue to Visualization](#)

ID	sbj_case	voice	word_order	verb
1790.FIMMBRAEDRA.NAR-SAG..662	sbj_DAT	active	wo_O1VS	lika
1790.FIMMBRAEDRA.NAR-SAG..382	sbj_DAT	active	wo_O1VS	vera
1791.JONSTEINGRIMS.BIO-AUT,154.1431	sbj_DAT	active	wo_O1VS	batna

- Download filtered data as CSV-file
- Process data with a different tool of choice

- Explore data set before visualization
- Construction of a task-specific data set
- Filter for sentences with relevant properties
 - Specific time frame
 - Specific features (i.e., entries in cells)
 - SQL-like filter construction (AND- or OR-functions)
- Dimension selection
 - Dimensions (i.e., columns) to be displayed in result table
 - Dimensions to be analyzed in visualization

Task-based Filtering

- Access to detailed information about each data point
- Furthers understanding of data quality
- Comparison of annotated values and extracted features

Result Table

Export Records Continue to Visualization

ID	verb	word_order
1790.FIMMBRAEDRA.NAR-SAG,.662	líka	O1VS
1790.FIMMBRAEDRA.NAR-SAG,.382	vera	O1VS
1791.JONSTEINGRIMS.BIO-AUT,154.1431	batna	O1VS
1791.JONSTEINGRIMS.BIO-AUT,126.736	gleyma	O1VS

Dimension	Feature
verb	líka
verb_type	VB
modal-aspectual	-
voice	active
word_order	O1VS
valency	trans
sbj_case	sbj_DAT
obj_case	obj1_NOM
obj2_case	-
sbj_type	sbj_Q
obj_type	obj1_N
obj2_type	-
genre	NAR

Metadata:

```
( (IP-MAT (NP-0B1 (D-N þetta-þessi) (N-N ráð-ráð))
  (VBDI líkaði-líka)
  (NP-SBJ (Q-D öllum-allur))
  (ADVP (ADV vel-vel)))
  (ID 1790.FIMMBRAEDRA.NAR-SAG,.662))
```

Analyzing Change over Time



- Define/select time periods 

Predefined Ranges:


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

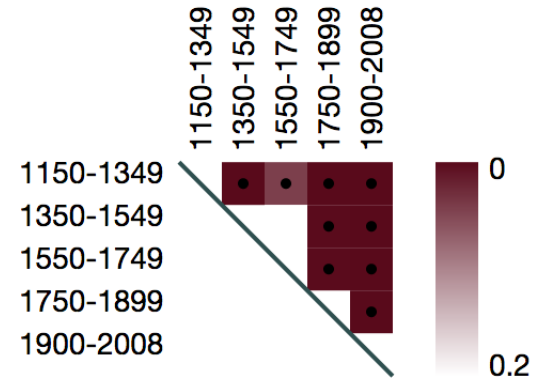
Custom Ranges

Split in Ranges

- Compact Matrix Visualization 
- Difference Histograms Visualization 

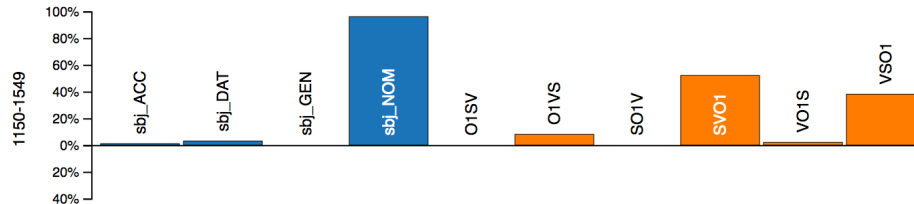
Compact Matrix Visualization

- Visualizes differences between selected dimensions across time
- Comparison of periods along the diagonal
- Differences mapped onto a colormap
- Two comparison modes:
 - χ^2 -test
 - Statistical significance ($\alpha \leq 0.05$) 
 - Absence of necessary preconditions \times
 - 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 \rightarrow large difference (high significance)
- Measure of quality and “interestingness”



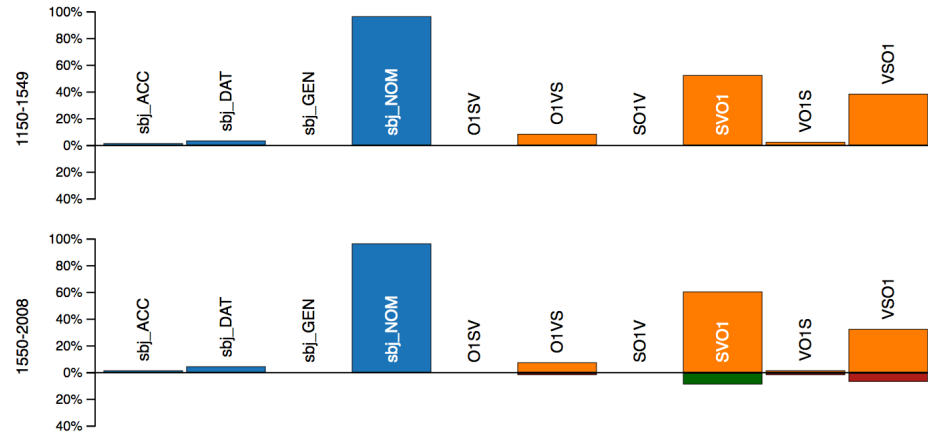
Difference Histograms Visualization

- 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.



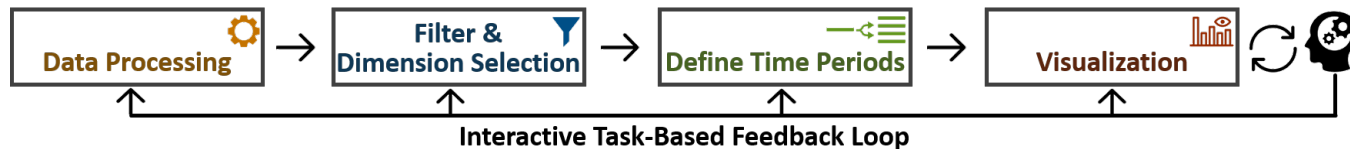
Difference Histograms Visualization

- 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



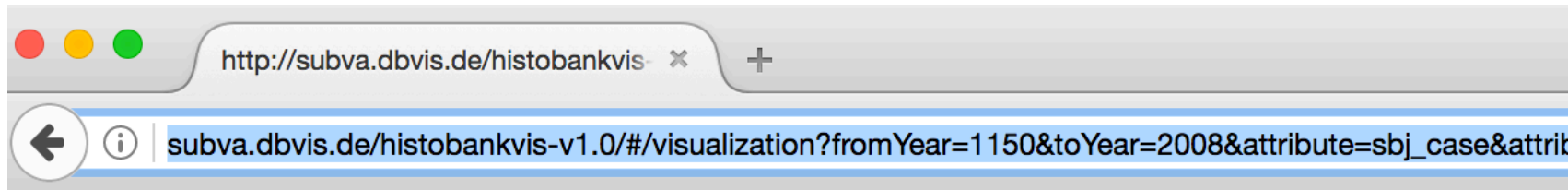
Hypothesis Generation and Feedback ↻

- 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



Access and Usability

- On-line browser app: <http://histobankvis.dbvis.de/>
- Analysis steps and current views are encoded by unique identification URLs



- Store and retrieve visualizations/analyses
- Share data and knowledge with other researchers
- Supports research collaborations

Access and Usability

- IcePaHC data set implemented as default
- Upload of own data
 - Tab-separated files
 - Must start with unique ID followed by a year date
 - Meta information, e.g., the corresponding full texts or parse trees, can be uploaded as well → unique IDs map between the files

ID	YEAR	ATT_1	ATT_2	ATT_3
id_1	2000	no	a	num
id_2	2001	no	b	text
id_3	2002	no	b	text
id_4	2003	yes	c	num
id_5	2004	yes	c	text

⇒ Further instructions and example data sets are provided on-line!

Case Study

- Part of on-going research on the diachrony of V1 word order and dative subjects in Icelandic
- Concrete case study on the interaction of word order in transitive sentences and subject case in IcePaHC
- HistoBankVis allows for the investigation of correlations between word order changes and dative subjects in Icelandic by means of just a few clicks.
- In particular, we were able to identify a correlation between the loss of V1 and an increase of dative subjects over time via HistoBankVis.

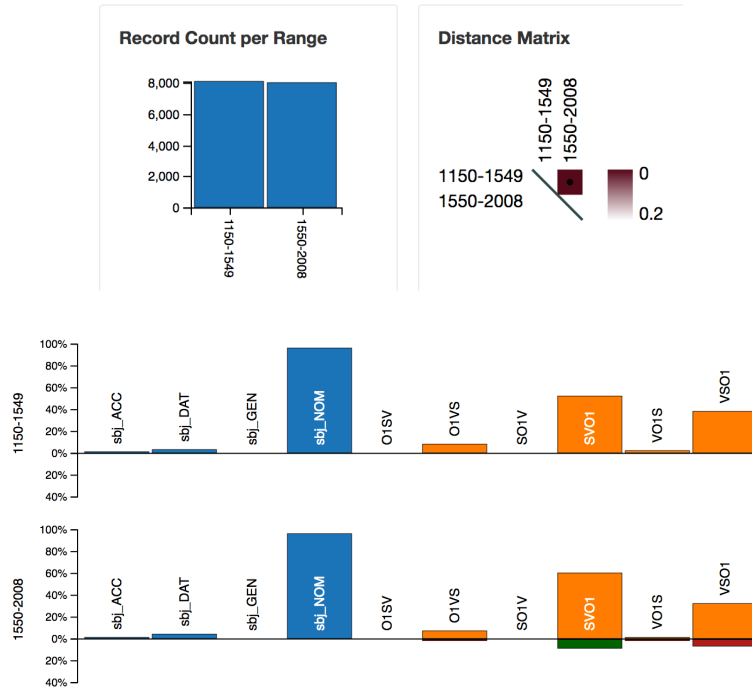
⇒ HistoBankVis is a powerful and effective tool which immensely facilitates historical linguistic research.

Case Study

Live Demo

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

Subject Case and Word Order I



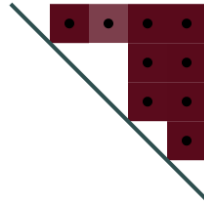
Settings

- Filter for sentences containing a subject (S), a verb (V) and a direct object (O/O1)
- Select dimensions subject case and word order
- Old vs. Modern Icelandic (Range A)
- 'Previous range' comparison mode

Findings

- SVO is the dominant word order in both time periods.
- SVO is slightly increasing, while VSO is decreasing.
- Mainly nominative subjects and to a lesser extent dative subjects

Subject Case and Word Order II

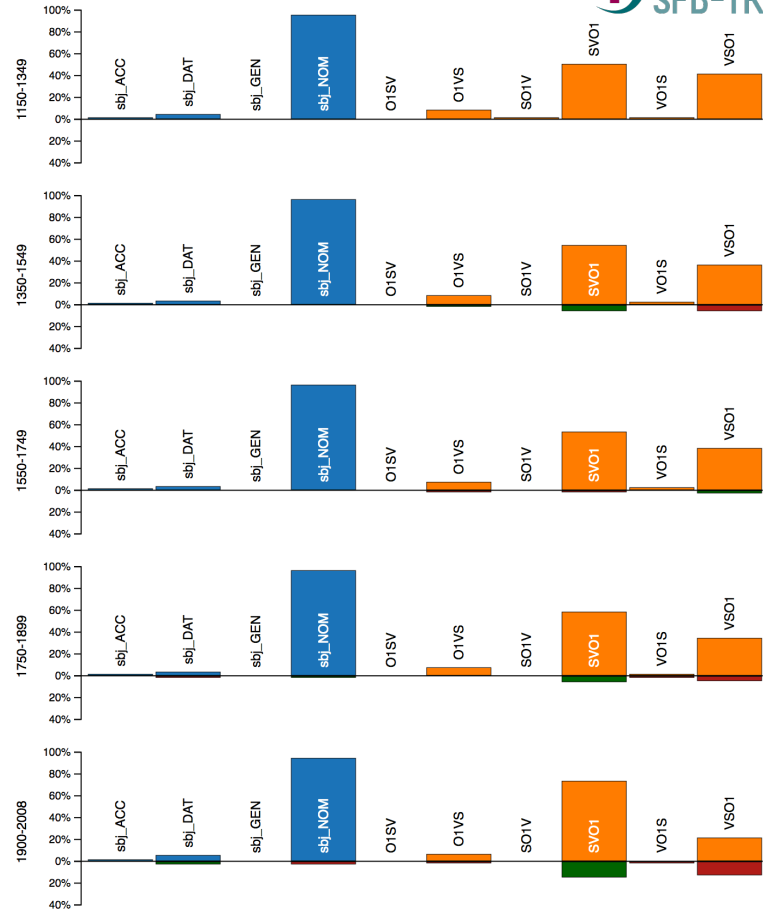


Settings

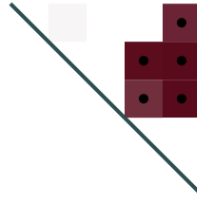
- Same settings as before
- More fine-grained time periods (Range B)

Findings

- Significant change within the last two time stages
- Fairly large increase of SVO in the last time stage, VSO is further decreasing.
- Dative subjects increase slightly in the last time stage.



Dative Subjects

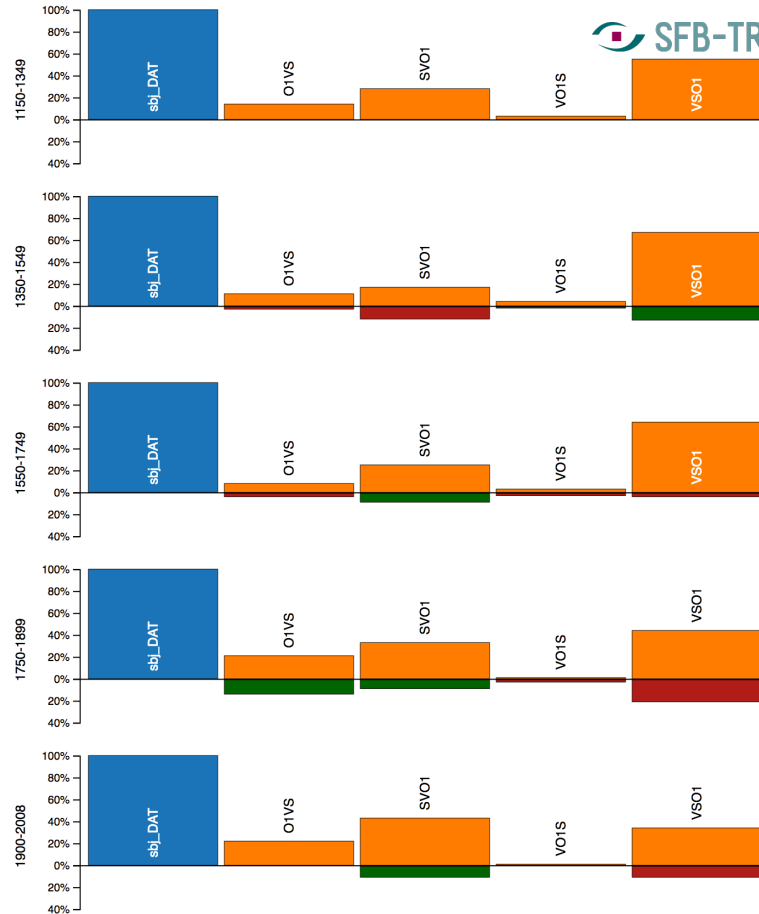


Settings

- Same settings
- Additional filter for dative subject sentences

Findings

- Still: SVO is increasing, while VSO is decreasing.
- **But, VSO is the dominant word order for dative subjects until 1900!**



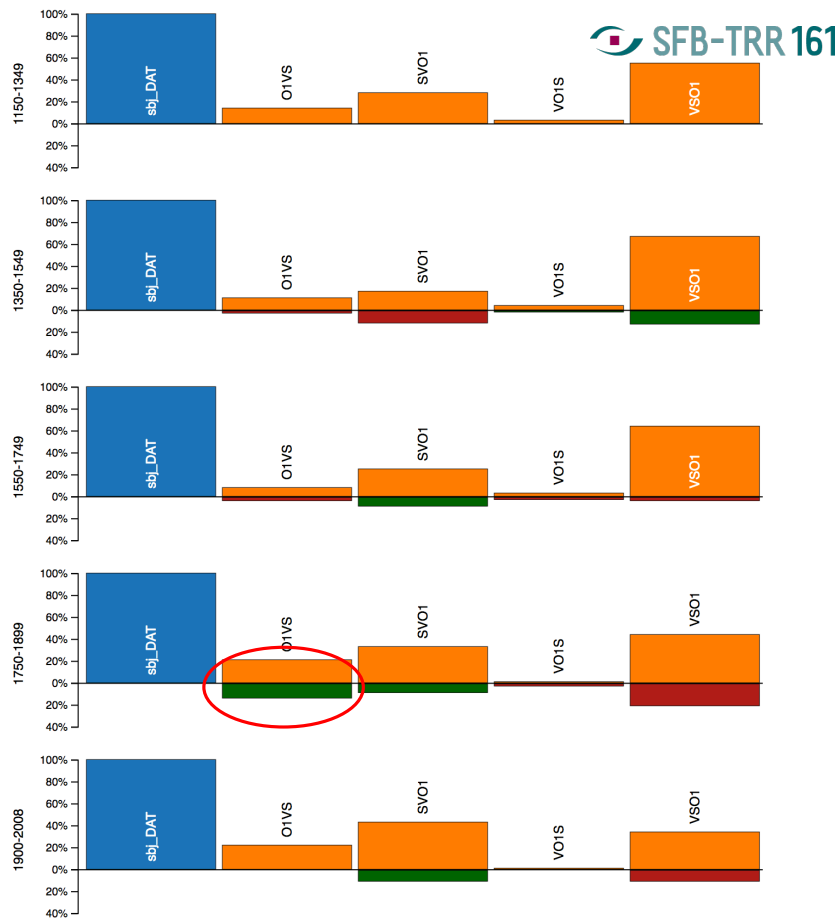
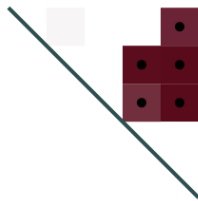
Dative Subjects

Settings

- Same settings
- Additional filter for dative subject sentences

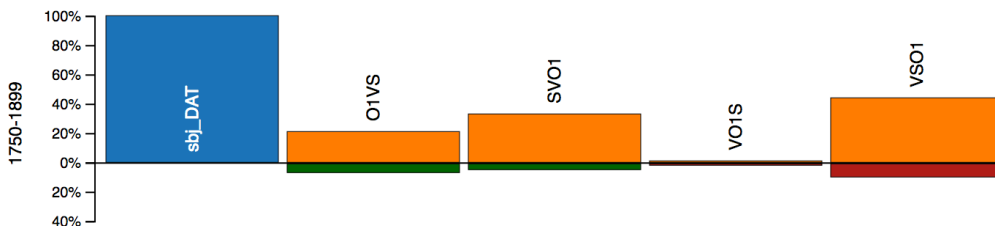
Findings

- Still: SVO is increasing, while VSO is decreasing.
- **But, VSO is the dominant word order for dative subjects until 1900!**



Dative subjects and OVS

- OVS word order stands out in second to last time stage
 - 1. 'Average of all ranges' comparison mode
 - 2. Filter for data from 1750-1900 containing dative subjects and OVS
 - 3. Look at the data
- Mainly experiencer predicates, e.g. *líka* 'like, please'



Sentence Filter

From year to

Edit Filter

Reset Filter

Apply Filter

Dimension	Features
sbj_case	sbj_DAT
word_order	wo_O1VS

Result Table

Export Records

Continue to Visualization

ID	sbj_case	voice	word_order	verb
1790.FIMMBRAEDRA.NAR-SAG,.662	sbj_DAT	active	wo_O1VS	líka
1790.FIMMBRAEDRA.NAR-SAG,.382	sbj_DAT	active	wo_O1VS	vera
1791.JONSTEINGRIMS.BIO-AUT,154.1431	sbj_DAT	active	wo_O1VS	batna

Results I

Dative Subjects and OVS

- Experiencer verbs are subject to lexicalization over time.
- Change from experiencer/goal objects to sentient experiencer/goal subjects

This pleases me → *I like this*

- Experiencer/goals can be realized as both subject and object (cf. Grimshaw, 1990)
 - Dative arguments may occur in more than one structural position
 - Sentient/animate participants are preferentially realized as subjects (Dowty, 1991)
- ⇒ Experiencer participants are increasingly realized as dative subjects in the history of Icelandic
- ⇒ In line with research on the interaction between middle morphology and dative subjects (Schätzle et al., 2015)

Results II

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.
- Combination of knowledge-based and data-driven modeling
- The visualization bridges the gap between annotated values, statistical analyses and the actual underlying data.
- The system can in general be applied to any Penn Treebank-style annotated corpus or any kind of well-structured data set.
- Each analysis step is accessible via a single identification URL
 - Storage of multiple perspectives on the data which can be retrieved at any time
 - Support of collaborative research
- HistoBankVis can also be used as a preprocessing and filtering tool as it allows for the export of filtered data sets.

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.

