HAL
open science

Visual analytics for historical linguistics: opportunities and challenges

Christin Schätzle, Miriam Butt

To cite this version:

Christin Schätzle, Miriam Butt. Visual analytics for historical linguistics: opportunities and challenges. 2020. hal-02914284v1

HAL Id: hal-02914284

https://inria.hal.science/hal-02914284v1

Preprint submitted on 11 Aug 2020 (v1), last revised 11 Dec 2020 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Journal of Data Mining and Digital Humanities

Visual analytics for historical linguistics: opportunities and challenges

Christin Schätzle[1] **and Miriam Butt**[1]

[1]University of Konstanz, Germany

Corresponding author: Christin Schätzle , christin.schaetzle@uni-konstanz.de

Abstract

In this paper we present a case study in which Visual Analytic methods for interactive data exploration are applied to the study of historical linguistics. We discuss why diachronic linguistic data poses special challenges for Visual Analytics and show how these are handled in a collaboratively developed web-based tool: HistoBankVis. HistoBankVis allows an immediate and efficient interaction with underlying diachronic data and we go through an investigation of the interplay between case marking and word order in Icelandic and Old Saxon to illustrate its features. We then discuss challenges posed by the lack of annotation standardization across different corpora as well as the problems we encountered with respect to errors, uncertainty and issues of data provenance. Overall we conclude that the integration of Visual Analytics methodology into the study of language change has an immense potential but that the full realization of its potential will depend on whether issues of data interoperability and annotation standards can be resolved.

Keywords

corpus linguistics, linguistic annotation, visual analytics, historical linguistics, syntactic change

I INTRODUCTION

In this paper we discuss the potential of methods from Visual Analytics [Thomas and Cook, 2005] for the study of language change by focusing on a web-based tool we have built in collaboration with colleagues from computer science: HistoBankVis [Schätzle et al., 2017, Schätzle et al., 2019] is a multilayer visualization system developed specifically for historical linguistic research. HistoBankVis allows for an interactive and exploratory access to a complex data set by using several interlinked visualization and filtering techniques in combination with a structured analysis process. We highlight the effectiveness of HistoBankVis by presenting a concrete test case which investigates syntactic change in Germanic, using historical corpora annotated according to the Penn Treebank format. Our tool goes a long way towards ameliorating several current methodological challenges for historical linguistics, these are discussed in section II.

As necessary background information, we provide an introduction to Visual Analytics in section III, describe the functionalities of our HistoBankVis system in section IV and show how it works with respect to investigating an interaction between dative case and word order in Icelandic. We use the IcePaHC corpus (Icelandic Parsed Historical Corpus; Wallenberg et al. [2011]) for this purpose. Like many existing corpora, IcePaHC is annotated broadly according to the Penn Treebank format [Marcus et al., 1993]. In seeking to compare our results for Icelandic with other Germanic languages, we identified further suitable corpora and experimented with these, most prominently the Penn Parsed Corpora of Historical English [Kroch and Taylor, 2000, Kroch et al., 2004, 2016] and the Heliand Parsed Database (HeliPaD; Walkden [2015, 2016]).

In seeking to extend our case studies, however, we encountered several challenges. One concerns the fact that although families of linguistic corpora are annotated according to broadly similar standards, whether this be according to the Universal Dependencies format [Nivre et al., 2016] as in PROIEL (Pragmatic Resources of Indo-European; Haug and Jøhndal [2008]) or the Penn Treebank style, the annotations are in fact not interoperable. In section V we discuss these and other challenges with respect to annotation errors, uncertainty and data provenance issues and conclude that while the integration of methodology from Visual Analytics into historical linguistic research has great potential, this potential will only be unlocked to its full extent once issues of annotation interoperability are dealt with comprehensively. This includes developing systematic methods of dealing with error correction as well as annotation uncertainty and data provenance.

II METHODOLOGICAL CHALLENGES FOR HISTORICAL LINGUISTICS

Over the past two decades, a multitude of digitized text corpora has been made available for historical linguistic research. These text corpora are often enhanced with elaborate linguistic annotations, including annotations for inflectional morphology, parts-of-speech, syntactic constituents, syntactic hierarchies and/or dependencies. Prototypical annotation standards are the Penn Treebank format [Marcus et al., 1993] and the Universal Dependencies (UD) framework [Nivre et al., 2016]. A great advantage of such corpora is that they allow for the quantitative investigation of structurally complex phenomena. However, the intricacies involved in producing high quality linguistic annotation and the difficulty of understanding highly complex interactions between various linguistic and extra-linguistic features and structures over a temporal dimension poses myriad new challenges.

Linguistically annotated corpora have usually undergone a manual annotation process in addition to an automatic preprocessing. Although the manual annotation process is time-consuming, manual annotations allow for sophisticated and high quality annotations. The annotations often reflect a deep linguistic analysis, e.g., in the form of syntactic hierarchies, dependencies between phrase structure constituents and semantic information. This kind of linguistic information is typically stored ('banked') in treebanks. Examples for historical treebanks are the Penn Parsed Corpora of Historical English [Kroch and Taylor, 2000, Kroch et al., 2004, 2016], the Icelandic Parsed Historical Corpus (IcePaHC; Wallenberg et al. [2011]), the Heliand Parsed Database (HeliPaD; Walkden [2015, 2016]), and PROIEL (Pragmatic Resources of Indo-European; Haug and Jøhndal [2008]). While the latter contains annotations in the Universal Dependencies format, the Penn Parsed Corpora of Historical English, IcePaHC, and HeliPaD are annotated in the Penn Treebank-style.

```
(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))
```

Figure 1: Sample annotation for a sentence from IcePaHC.

Figure 1 shows a sample for a Penn Treebank annotated sentence from IcePaHC with annota-

tions for clause type, constituents, noun type, grammatical relations, case marking, verb type, lemmas, tense, and voice. The sentence in Figure 1 is a matrix declarative clause (IP-MAT) with the pronominal dative subject *mér* 'me' (NP-SBJ for subject NP, PRO-D for pronoun and dative case) and the main verb *finna* 'find', which occurs in the middle form *finnst* 'think, seem' in present tense (VBPI). The availability of such elaborate syntactic annotations allows for the automatic extraction and quantitative investigation of intricate linguistic patterns over time.

Statistical methods for the quantitative analysis of extracted data has become a standard part of the methodological toolkit in historical linguistics. Typical examples include the calculation of correlations and/or dispersion statistics, multifactorial regression modeling, or the use of clustering methods (see Hilpert and Gries [2016] for an overview). Typical standard programming languages employed in historical corpus linguistics for the automatic extraction of the relevant linguistic patterns are Python and R [Bird et al., 2009, Baayen, 2008]. Additionally, off-the-shelf tools such as CorpusSearch [Randall, 2000] are available for the extraction of linguistic patterns from annotated corpora. Yet, the use of statistics, in particular of inferential statistics, is not always appropriate in historical linguistics, since data sparsity is a well-known problem of diachronic corpora.

Since multiple feature interactions have to be taken into account in the diachronic analysis of a single phenomenon, a multitude of high-dimensional data tables with different characteristics are usually generated. A prototypical historical linguistic data table is given in Table 1, with diachronic data extracted from IcePaHC showing the interaction between subject case marking (here NOM(INATIVE) vs. DAT(IVE)) and voice (active, passive, middle) across the Icelandic diachrony (see also Schätzle [2018] for similar and more detailed data).

Periods	Active			Passive			Middle		
	NOM	DAT	%DAT	NOM	DAT	%DAT	NOM	DAT	%DAT
1150-1349	8783	279	3.1%	631	71	10.1%	1009	64	6.0%
1350-1549	12954	299	2.3%	445	93	17.3%	1257	80	6.0%
1550-1749	6978	122	1.7%	603	106	15.0%	871	69	7.3%
1750-1899	7184	182	2.5%	336	67	16.6%	724	88	10.8%
1900-2008	6289	173	2.7%	269	72	21.1%	758	239	24.0%

Table 1: Distribution of subject case across voice in IcePaHC (1150-2008).

Another example of a prototypical historical linguistic data table is given in Table 2, showing data extracted by Taylor and Pintzuk [2011] from the York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE, Taylor et al., 2003) and the Penn-Helsinki Parsed Corpus of Middle English (PPCME2, Kroch and Taylor, 2000).

Finding significant patterns in such multidimensional tables is by no means easy. For one, the task is complex since identifying patterns and feature interactions across many such tables requires the pair-wise comparison of the relevant bits of information in the form of numbers and percentages, while keeping the temporal component in mind. Moreover, data sparsity is an issue and statistical significance is often calculated on the basis of only few occurrences of the actual observation. This may derogate the statistical conclusions for the data, making the comparison of results across the feature interactions extremely difficult. Meaningful patterns may also remain hidden from the researcher. In contrast, irrelevant patterns might be interpreted as significant. A further level of complexity is added by the fact that statistical calculations generally require the definition of fixed parameters, e.g., time periods. This is problematic when the

	New objects				Given objects			
Text	OV	VO	Total	%VO	OV	VO	Total	%VO
Orosius	21	9	30	30.0	68	12	80	15.0
Boethius	25	40	65	61.5	46	32	78	41.0
Cura Pastoralis	11	14	25	56.0	23	14	37	37.8
Catholic Homilies I	34	36	70	51.4	74	48	122	39.3
Catholic Homilies II	19	17	36	47.2	24	16	40	40.0
Lives of Saints	11	17	28	60.7	31	34	65	52.3
Gregory's Dialogues (C)	7	14	21	66.7	16	20	36	55.6
Trinity Homilies	23	21	44	47.7	28	25	53	47.2
Katherine Group	9	15	24	62.5	19	32	51	62.7
Ancrene Riwle	15	54	69	78.3	27	63	90	70.0

Table 2: Distribution of new and given objects across VO (Verb-Object) vs. OV (Object-Verb) order in AuxV (Auxiliary-Verb) clauses in Old and Middle English texts [Taylor and Pintzuk, 2011, 91].

selected time periods are too coarse or too fine-grained for the analysis such that transitioning periods and interesting patterns therein are absorbed by the periodization. Addressing this issue, Schätzle and Booth [2019] developed DiaHClust, a data-driven method for identifying stages of language change based on hierarchical clustering (i.e., Variability-based Neighbor Clustering; Gries and Hilpert, 2008), which groups corpus data into time periods with respect to the relevant changing linguistic features. However, the technique relies on calculating differences between features which are known to have changed over time in the language, knowledge which is often not (yet) available.

The factors involved in a change are often elusive to the researcher, either because the phenomenon has not yet been investigated by the community or because the matter is generally under dispute. Therefore, a researcher may have to investigate a multitude of different interactions in order to test existing hypotheses and to generate new ones, creating numerous high-dimensional data tables with different features and characteristics. This is a costly and time-consuming process, resulting in data which is difficult to navigate.

In the next section, we show how these methodological challenges for historical corpus linguistics can be overcome by integrating the use of Visual Analytics into diachronic investigations.

III VISUAL ANALYTICS FOR HISTORICAL LINGUISTICS

Visual Analytics (VA), "the science of analytical reasoning facilitated by interactive visual interfaces" [Thomas and Cook, 2005, 28], presents a significant methodological opportunity for historical linguistic research. VA methods combine automated algorithmic analyses with interactive visual components, integrating the human into the analysis loop (see Thomas and Cook, 2005, Keim et al., 2008). The general aim of VA is to present potentially interesting and significant correlations in a high-dimensional data set saliently so as to enable significant patterns to emerge visually. The interactive and exploratory analysis process is guided by the VA Mantra: "Analyze first, show the important, zoom, filter and analyze further, details on demand" [Keim et al., 2008]. Figure 2 illustrates the coupled process of knowledge generation in VA, where the left hand side illustrates the parts involved in a visual analytics system and the right hand side depicts the reasoning process of the human, composed of exploration, verification, and knowledge generation loops [see Sacha et al., 2014].

Since historical linguistic data is inherently multidimensional, with complex feature interactions

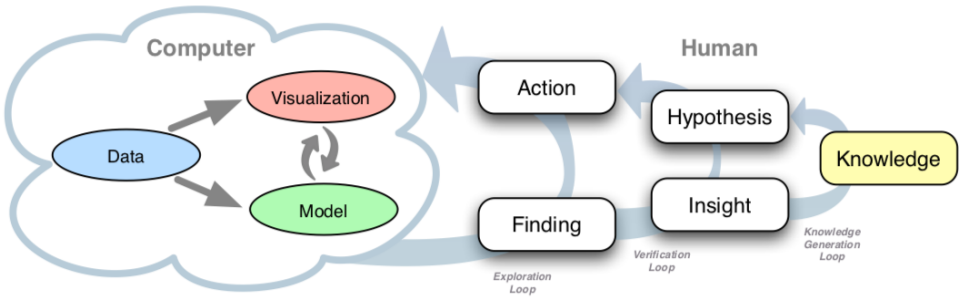

Figure 2: Knowledge generation model for Visual Analytics (taken from Sacha et al. [2014]).

being the norm rather than the exception, historical linguistics constitutes an ideal test bed for VA applications. VA tools and techniques enable an exploratory and interactive access to subspaces contained in historical linguistic data, i.e., significant correlation patterns embedded in a high-dimensional data space, which immensely facilitates the identification of language change and relevant interacting factors.

In recent years, sophisticated visualizations as developed within the field of computer science have increasingly been applied to the investigation of language change. Previous visualizations have mainly focused on the investigation of semantic change by visualizing the diachronic development of word senses. Examples are the scatterplot visualization developed by Rohrdantz et al. [2011], the pixel visualization by Rohrdantz et al. [2012] (see also Rohrdantz [2014]), and the similarity plots based on line charts by Jatowt et al. [2018]. Other approaches make use of parallel coordinate plots, e.g., Culy et al. [2011] employ Structured Parallel Coordinates for the investigation of diachronic changes in the use of modal verbs across different registers of academic discourse and Theron and Fontanillo [2015] visualize the diachrony of word meanings in historical dictionaries as parallel coordinates. There are also some visualizations which were designed for the investigation of syntactic change, focusing on the diachronic visualization of syntactic phenomena and potentially interacting factors. Examples are the glyph visualization by Butt et al. [2014] (see also Schätzle and Sacha [2016]) and the ParHistVis tool for the investigation of linguistic change in parallel corpora which employs steamgraphs and Sankey Diagrams [see Kalouli et al., 2019].

In the following, we demonstrate the efficacy of using VA for historical linguistic research by introducing our HistoBankVis system and by applying the system to a concrete case study on syntactic change in Germanic.

IV HISTOBANKVIS: VISUALIZING CHANGE AS DIMENSION INTERACTIONS

4.1 The HistoBankVis system

HistoBankVis was first presented in Schätzle et al. [2017] and further extended and improved in Schätzle et al. [2019]. We developed HistoBankVis with the overall goal of providing a generically applicable system for the flexible investigation of the type of high-dimensional and complex data typically underlying historical linguistic work. Overall, HistoBankVis combines several interlinked visualization and filtering techniques with a structured analysis process, creating the iterative workflow shown in Figure 3.

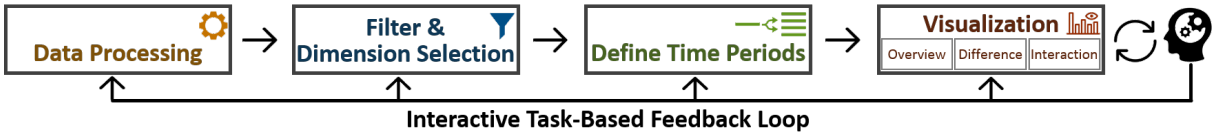

Figure 3: Iterative workflow behind HistoBankVis.

4.1.1 Data processing

HistoBankVis is implemented as an online browser-app.[1] As input, the system requires a file containing a tab-separated data set. The first column in the data set should consist of unique representation IDs for the investigated data points and the second column consists of the corresponding year dates (needed for a diachronic analysis). All other columns contain *features* associated with the different linguistic factors, i.e., data *dimensions*, under investigation. For example, in Schätzle et al. [2019], we investigated syntactic change in Icelandic using data from IcePaHC. To this end, we extracted features for data dimensions, i.e., linguistic factors, from the IcePaHC annotation which have been identified as relevant for syntactic change in the language by the existing literature. This encompassed information about verb type, subject case, object case, indirect object case, word order (constituent order), subject position, and V1 (verb-first) with respect to each matrix declarative clause in IcePaHC. The extraction was effected via our own programming scripts.[2] The extracted features were moreover mapped to the sentence IDs of the respective clauses assigned by the IcePaHC annotation (see the ID in the last line in Figure 1), creating a well-structured data set. An excerpt of this data set is shown in Figure 4. Moreover, this data set can be explored on-line with HistoBankVis.

ID	YEAR	VERB	VERB_TYPE	VOICE	WORD_ORDER	SBJ_CASE	OBJ_CASE	OBJ2_CASE	SUBJ_POSITION	V1
1150.FIRSTGRAMMAR.SCI-LIN,.1	1150	setja	VB	active	VSO1	sbj_NOM	obj1_ACC	-	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.2	1150	setja	VB	active	O1VS	sbj_NOM	obj1_ACC	-	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.3	1150	hafa	HV	active	SVO1	sbj_NOM	obj1_ACC	-	prefinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.4	1150	rita	VB	active	VSO1	sbj_NOM	obj1_ACC	-	postfinite	yes
1150.FIRSTGRAMMAR.SCI-LIN,.5	1150	verða	RD	active	VS	sbj_GEN	-	-	postfinite	no
1150.FIRSTGRAMMAR.SCI-LIN,.6	1150	ganga	VB	active	VS	sbj_NOM	-	-	postfinite	no

Figure 4: Data set extracted from IcePaHC for the diachronic investigation of subject case and word order.

A tabular data set was chosen as required input format for the system because integrating a data processing module for treebanks into HistoBankVis is difficult. The reason for this is that although corpora are in theory annotated according to mutually agreed upon standards, in practice annotations tend to diverge significantly enough to make work across several different treebanks with the same automatized tools difficult. We discuss this in more detail in Section V. Bringing the data into a tabular form conforming to the csv (Comma-Separated Values) format has the additional advantage that with respect to the backend of the system, the data is stored in a relational SQL database (supporting SQLite[3] and PostgreSQL[4]). The SQL database allows efficient and fast access to the tab-separated data.

4.1.2 Filtering component, dimension selection and time periods

Before visualization, the data set can be explored in the filtering component of the system, providing insights into the data quality. The user can build a task-specific data set by filtering

[1] http://histobankvis.dbvis.de
[2] A detailed description of the full data set is given in Schätzle [2018].
[3] https://www.sqlite.org
[4] https://www.postgresql.org

for data points with specific (task-relevant) features and/or from a specific time period, and by selecting the dimensions which are to be investigated, see Figure 5. This is done via the visual construction of SQL-like filters, based on logical AND/OR functions. For each data point, e.g., sentence in the IcePaHC data set, detailed information about all extracted features and the underlying annotations can be accessed via mouse-click.[5]

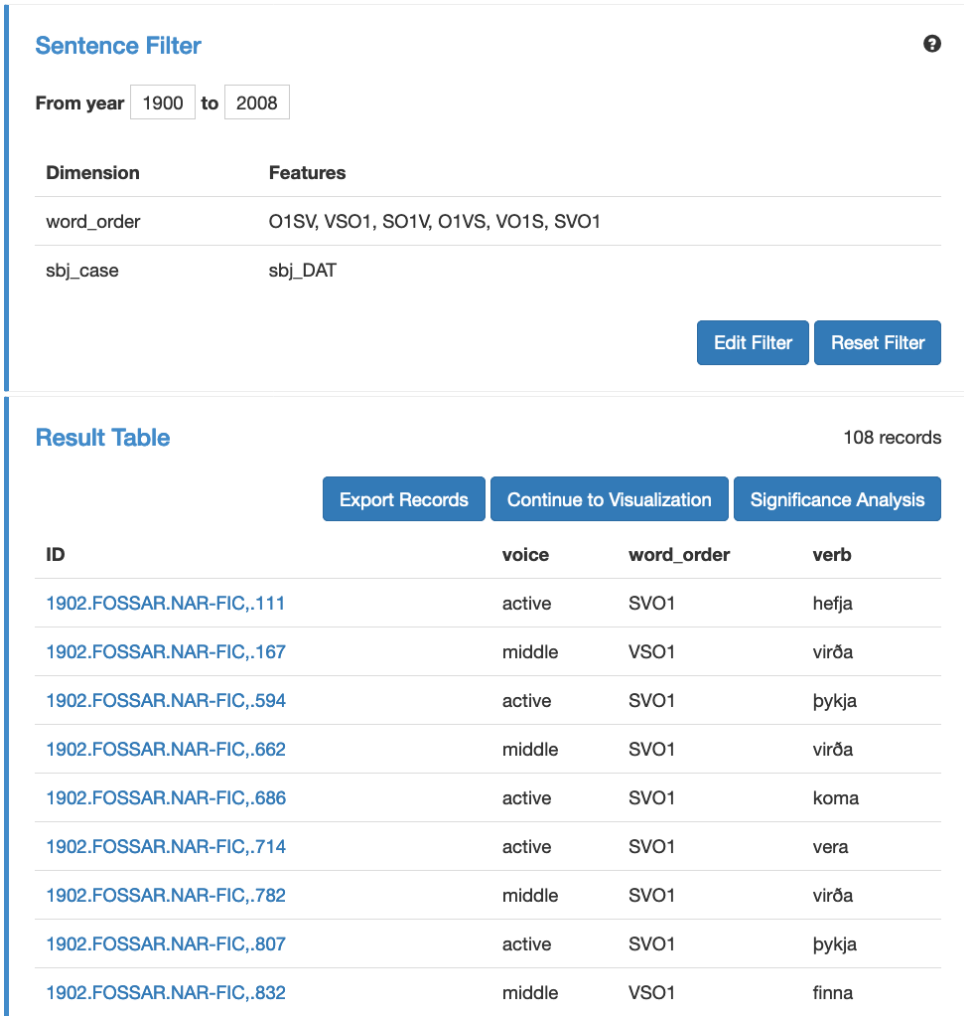

Figure 5: Sentence filter (top): The data is filtered with respect to the dimensions word order and subject case so that the resulting table only contains sentences which have the orders O1SV (Direct Object-Subject-Verb), VSO1, SO1V, O1VS, VO1S, and SVO1, and a subject which is marked with dative case. Moreover, as time range, data from the period 1900-2008 is selected. Result table (bottom): The filtered data set is displayed with respect to previously selected data dimensions, e.g., voice, word order, and verb.

For the subsequent visualization of the selected dimensions, the user has to choose time periods for the visual analysis. HistoBankVis supports a set of predefined periodization schemes for the Icelandic diachrony, but also allows for an individual definition of time periods by the user.

[5]For exploring the underlying annotations in one's own data sets, an additional meta-data file needs to be uploaded. This file can be a simple txt-file, containing, e.g., the parse trees for each sentence followed by their unique identifier, corresponding to the representation IDs. More details on data set requirements are provided at histobankvis.dbvis.de.

4.1.3 *Visualization – Overview, difference, interaction*

HistoBankVis has three visualization components which are combined in the data analysis process, providing different views of the data at different levels of detail. All visualizations are designed in D3.js[6] as Scalable Vector Graphics.

Compact matrix visualization. The first visualization component is a *compact matrix* which provides an overview of the data, see Figure 6. Each row and column of the matrix represents one time period. The matrix can be mirrored at the diagonal. This design in particular facilitates the comparison of the data in the first period to all other periods and the comparison of consecutive periods along the diagonal, letting patterns of change emerge visually. The compact matrix visualizes differences between the distributions of the selected dimensions across the individual time periods. A colormap encodes the size of the difference: red indicates a large difference, white a small one. The utility of this visualization is a first at-a-glance look at which of the data dimensions are likely to be significantly different across time periods and therefore worthy of more detailed investigation. Two modes are available for computing differences: statistical significance and distance measure. For calculating statistical significance, we employ χ^2-tests, mapping the p-values onto the colormap (red $p = 0$, white $p \geq 0$). Statistically significant differences between time periods (with $\alpha \leq 0.05$) are marked by a dot in the middle of the cell. When the necessary preconditions for χ^2 are missing, e.g., when the data is too sparse, a cross marks the corresponding cells. Alternatively, differences can be computed via Euclidean distance, whereby a high distance indicates a large difference.

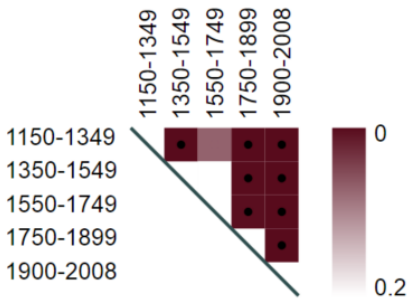

Figure 6: Compact matrix showing statistically significant differences between time periods.

Difference histograms visualization. The second visualization component is the *difference histograms* visualization. The difference histograms provide a more nuanced view on the diachrony of the investigated features and dimensions. Each time period is visualized as a composed bar chart, where the dimensions are encoded via different colors, allowing for a parallel inspection of the data from the different dimensions. Figure 7 provides an example. In this example, two dimensions are being investigated: word order and subject case. The bars representing the dimension *subject case* are blue; the dimension *word order* is orange. The distribution of these linguistic features is shown for each time period. The height of a bar corresponds to the percentage of data points (e.g., sentences), in which this feature occurs in comparison to all features from the corresponding data dimension in the given time period. In order to facilitate the temporal comparison of features, differences between the features across the periods are visualized as separate bar charts below each feature bar. A green bar indicates a feature increase (e.g. as in the last row for the feature SVO1) with respect to the previous time period. In contrast, a red bar indicates that this feature has decreased in comparison to its distribution in the previous time period. The height of the bar reflects the size of the change. For example, in Figure 7 the word order SVO1 increases over time, whereas VSO1 decreases in the period 1900–2008 compared to the previous stage (1750–1899). More detailed information in the form of numbers can be displayed via mouse over. In addition to the comparison of each time period with its previous time range, the system supports further comparison modes, e.g., the comparison of each time period with the average of all preceding time periods.

[6] https://d3js.org/

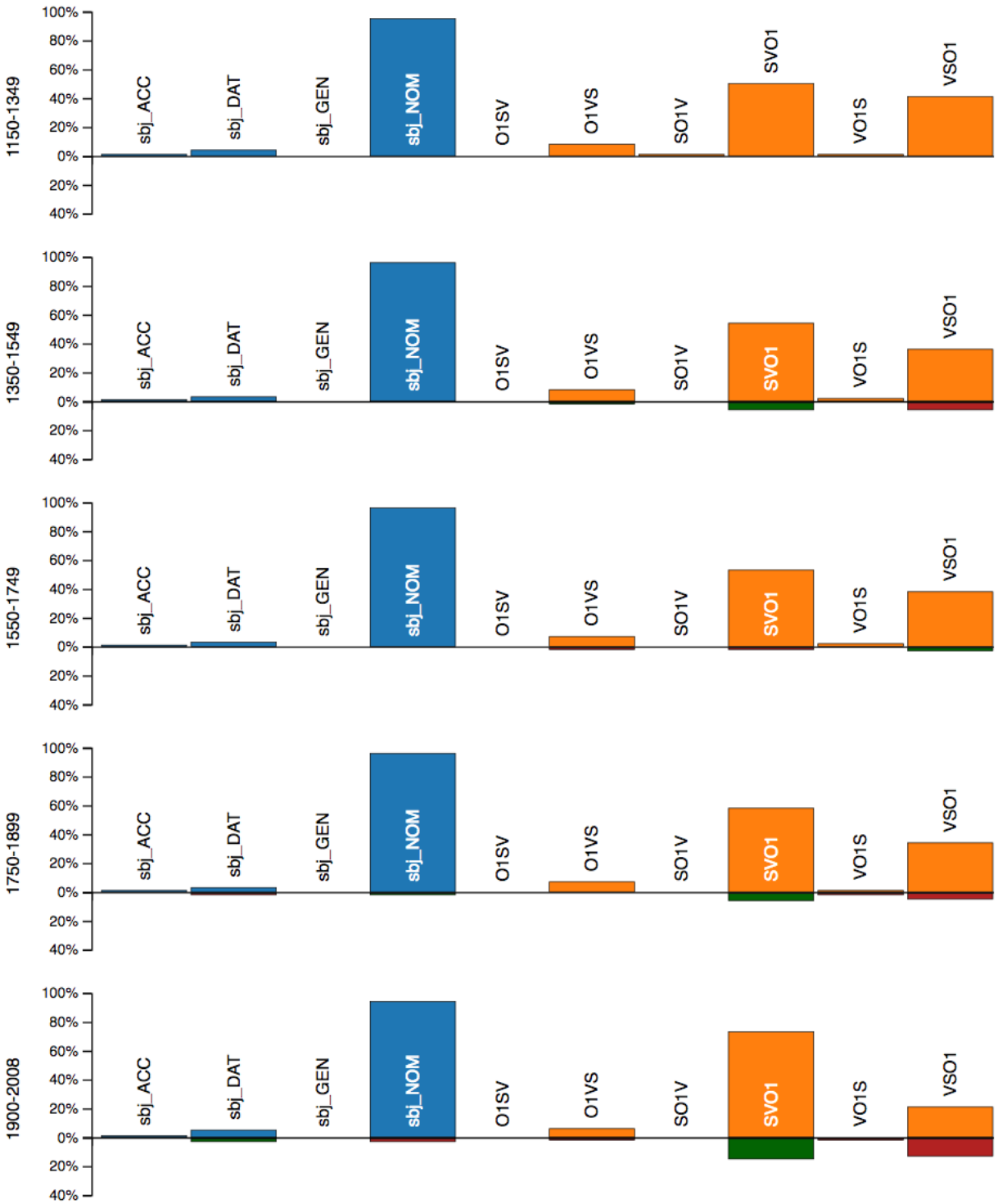

Figure 7: Difference histograms for the dimensions word order and subject case in transitive sentences from IcePaHC.

Dimension interaction visualization. Although the difference histograms provide an insight into the diachrony of individual features from different dimensions, correlations between changes occurring in the investigated data dimensions cannot be read off the composed bar charts directly. That is, whether a feature change in one dimension is indeed connected to a feature change in another dimension cannot yet be determined on the basis of the difference histograms. Therefore, HistoBankVis employs a third visualization component: the *dimension*

interaction visualization. The dimension interactions implement the Parallel Sets technique [Bendix et al., 2005, Kosara et al., 2006] for the visualization of interrelations between features from multiple data dimensions. Parallel sets are based on parallel coordinates [Inselberg, 1985, 2009], but allow for a better investigation of frequency-based categorical data.

Parallel coordinates represent each data dimension as as a vertical axis. The features are placed on the axes as coordinates. Related features between dimensions are connected by a line. Instead of connecting individual data points via polylines across different dimensions, parallel sets visualize connections between data dimensions via colored ribbons, enabling the representation of frequency-based interrelations. The size of a ribbon represents the share which a feature holds of a feature from another dimension. In the dimension interaction visualization component of HistoBankVis, each time period is visualized as a parallel sets visualization. For example, Figure 8 shows the dimension interaction between the dimension *voice* and the dimension *word order* in the period 1150–1349. The shares of the different voices (active, passive, middle) are mapped onto the shares they hold of the different possibilities of the dimension word order from left to right, allowing for a detailed investigation of interactions in the form of frequencies. Figure 8 shows that the majority of active clauses in the time period 1150–1349 have VSO1 word order, followed by SVO1. The same holds true for passives and middles, indicating that verb-initial word order was the most common in this time period, regardless of voice. So if one had hypothesized that voice played a determining factor in word order in this period of Icelandic, one would have been wrong. The advantage of the VA system is that the investigation of such a hypothesis can be effected almost as quickly as its initial formulation (if all the data have already been fed into the system, of course).

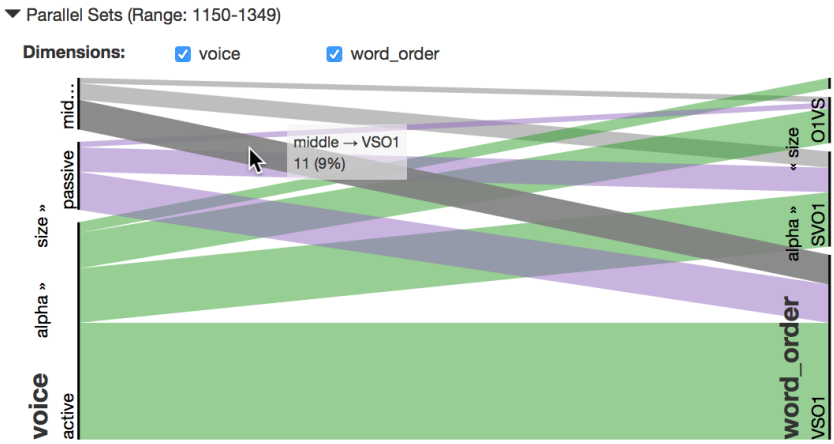

Figure 8: Dimension interaction for voice and word order in transitive sentences from 1150-1349 in IcePaHC.

The version of parallel sets implemented in HistoBankVis also provides for the flexible investigation of interrelations by allowing for the reordering of dimensions and features via drag&drop.[7] For a better overview, the features on each vertical axis can furthermore be sorted alphabetically or according to size (ascending or descending). More detailed information on frequencies and feature correspondences are available via mouse interaction techniques, see the mouse over on the middle/VSO1 ribbon in Figure 8.

[7]The code available on https://www.jasondavies.com/parallel-sets/ was used for the implementation of parallel sets in HistoBankVis.

4.1.4 Hypothesis generation and feedback loop

A researcher may have to test several different hypotheses in attempting to understand the causes and mechanisms of language change. HistoBankVis has been designed specifically to foster an easy and seamless iterative process of hypothesis testing and generation. Once the researcher has identified a change in one data dimension or detected a potentially interesting correlation between several different features across time, the researcher can react to these insights immediately by feeding the knowledge gained back into the system and then interacting directly with the different VA parts of the system. This can be done within just minutes by filtering the data anew, choosing different data dimensions and/or investigating the data with respect to a different set of time periods.

We have been interested in the correlation between case marking and word order as part of a larger project and had been investigating this issue for Icelandic. Once HistoBankVis was developed, we began to work with this system and found that the dimension interaction in combination with the overall flexibility of the system indeed facilitated our diachronic investigations immensely. More than once, we were able to identify correlations we had not been able to otherwise anticipate given the current state of the art [Schätzle, 2018]. We illustrate the general way of working with HistoBankVis in the next section via a concrete case study which examines the interrelation between subject case and word order in Icelandic (see also Schätzle et al. [2017], Schätzle [2018], Schätzle et al. [2019]).

4.2 Investigating syntactic change in Icelandic

Icelandic is generally acknowledged to be the most conservative Germanic language in terms of syntactic change. Some changes that have been observed involve word order and case marking. For one, Icelandic follows the Germanic change from OV (Object-Verb) to VO (Verb-Object) in the verb phrase [Hróarsdóttir, 1996, 2000, Rögnvaldsson, 1996]. Moreover, a decrease of V1 (verb-first) order in matrix declarative sentences [see, e.g., Butt et al., 2014], and an increasing preference for subjects to occur in the prefinite position [Booth et al., 2017] have been attested. A further change affects the case marking system of the language in that subjects are increasingly marked with dative case [Barðdal, 2011, Schätzle et al., 2015].

While changes in word order and subject case marking have been observed, interrelations between the changes have only rarely been investigated. In this paper, we show how interrelations between changes in word order and subject case marking can be identified and examined within minutes using the HistoBankVis system. For our investigation, we use the IcePaHC data set as described in Section 4.1.1. By means of just a few clicks, we were able to uncover a previously unknown link between dative subjects, word order, lexical semantics and voice in the history of Icelandic.

To begin our investigations, we chose the dimensions *subject case* (nominative, accusative, dative, or genitive) and *word order* in the filtering component. In order to avoid overly complicating the picture in the initial stages of exploration, we decided to look at transitive sentences only. Thus, we filtered for sentences which contain a subject (S), verb (V) and a direct object (O1, henceforth O) in the dimension *word order*. Moreover, we decided to investigate the diachrony of word order and subject case marking with respect to the following time periods: 1150–1349, 1350–1549, 1550–1749, 1750–1899, 1900–2008 (based on Haugen [1984]).

Moving on to the visualizations, the compact matrix, i.e., the matrix in Figure 6, showed at a glance that the distribution of word order and subject case changes over time, in particular in the last two time periods (1750-1899, 1900-2000). The difference histograms provide a

more nuanced view of these changes, see Figure 7. The difference histograms show that over time, SVO increases (green bars), while the other word order possibilities, in particular VSO, decrease (red bars). The most striking change with respect to word order occurs in the last time period (post–1900): While SVO increases significantly, VSO decreases substantially. At the same time, subject case marking changes as well: the use of dative subjects (sbj_DAT) increases slightly at the expense of nominative subjects (sbj_NOM).

Whether these changes are interlinked can now be easily investigated by means of the dimension interaction visualizations. Figure 9 shows the dimension interactions for subject case and word order in the first time period (top) and the period post–1900 (bottom). In the difference histogram for the period post–1900, most sentences occur together with the SVO order and have a subject with nominative case marking (Figure 7-bottom). The dimension interaction for this period provides more insights into how the different word order possibilities interact with the different options for subject case marking, see Figure 9-bottom. What becomes visible in the dimension interaction is a difference in how the word orders are distributed across the subject cases: While the vast majority of nominative subjects are SVO, which results in the preponderance of SVO in the difference histogram, only approximately half of the dative subjects occur together with SVO. Dative subjects also frequently appear postverbally, i.e., in the orders VSO and OSV (the green ribbon above VSO1 in Figure 9-bottom).

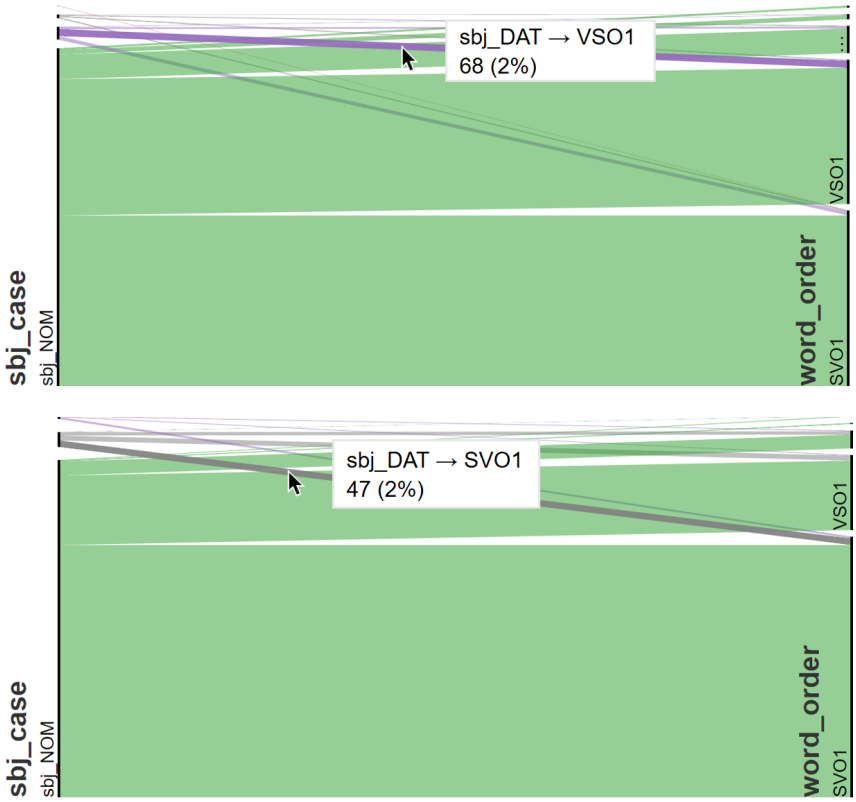

Figure 9: Dimension interactions for subject case and word order in transitive sentences from IcePaHC in the periods 1150-1349 (top) and 1900-2008 (bottom).

By comparing the dimension interactions, we found that the interrelation between subject case and word order has changed over time in IcePaHC. In the period 1150–1349, nominative subjects already preferably occur together with SVO. Yet, this preference is much smaller than in the last time period. However, dative subjects have a strong preference to occur together with the VSO order in the first time period, appearing only marginally with SVO. Although the co-

occurrence frequency of dative subjects with SVO increases diachronically, it is only in the last time period when dative subjects begin to mainly occur together with SVO order.

In order to find an explanation for why dative subjects lag behind with respect to the overall developments, we decided to take a closer look at the dative subject sentences in the last time period. In order to do this, we simply went back to the filtering component and filtered for dative subject sentences only. This has already been illustrated in Figure 5. Since voice has been determined as a conditioning factor for dative subjects in Icelandic by the existing literature [Zaenen et al., 1985, Sigurðsson, 1989], we now also included the dimension *voice* in our investigation. Moreover, we looked at the main verbs involved in the clauses to provide insights into the role of lexical semantics – a further determining factor for dative subjects in the language (see, e.g., Jónsson [2003], Barðdal [2011]). In the result table, a large amount of dative subject clauses appeared together with middle voice. In particular, the experiencer verb *finna* 'find, feel' lexicalized in its middle form *finnast* 'think, find, feel, seem' occurred most frequently. Overall, dative subjects were found most often together with experiencer predicates in the corpus, e.g., *þykja* 'think, seem' and *líka* 'like, please'.

We continued our investigation by visualizing the dimensions *word order* and *voice* with respect to the filtered data set. In the dimension interaction for the first time period, dative subjects occurred most frequently in active constructions, see Figure 8. In these constructions, as well as in the passive and middle voice, dative subjects were mainly found together with VSO. In the last time period however, dative subjects occurred most often together with middle voice, and SVO order is most frequently used together with dative subjects in active and middle constructions, see Figure 10-bottom. Thus, the change from OVS to SVO as the preferred word order in sentences with a dative subjects correlates with an increase of dative subjects together with middle voice. This increase is most striking between the last two time periods, compare Figure 10-top and Figure 10-bottom (see also Table 1). Furthermore, while dative subjects still preferably occur with VSO in active and passive constructions in the second to last time period, they already occur in the SVO order in clauses with middle voice, see Figure 10-top. Hence, the increasing realization of dative subjects in the prefinite position, i.e., together with SVO order, not only correlates with the increase of dative subjects in middle constructions, but is moreover driven by this increase.

In sum, our investigation of subject case and word order in IcePaHC by means of HistoBankVis confirmed previous corpus investigations into the Icelandic diachrony, but also led to new insights by discovering previously unknown interrelations between subject case and word order. We confirmed that subjects are increasingly realized in the prefinite position in Icelandic (cf. Booth et al., 2017). Additionally, we showed that the usage of dative subjects increases over time (see also Schätzle et al., 2015). The dimension interaction visualization as implemented in HistoBankVis allowed for a flexible and quick, but still detailed, analysis of interrelations between several different data dimensions. By means of the dimension interactions, we were able to show that dative subjects consistently lag behind with respect to the overall developments of word order. It is only around 1900 when they eventually begin to follow suit. This change is driven by an increased use of middle verbs, which are mainly experiencer predicates, together with a dative subject. The middle verbs in question mainly appear to have been subject to being lexicalized as experiencer verbs, as part of which the formerly locative oblique argument becomes reanalyzed as an experiencer subject [Schätzle, 2018]. This historical change at the lexical level coupled with a change in syntactic alignment immediately explains the slower tendency for dative subjects to occur in the prefinite position: Only once the dative experiencers

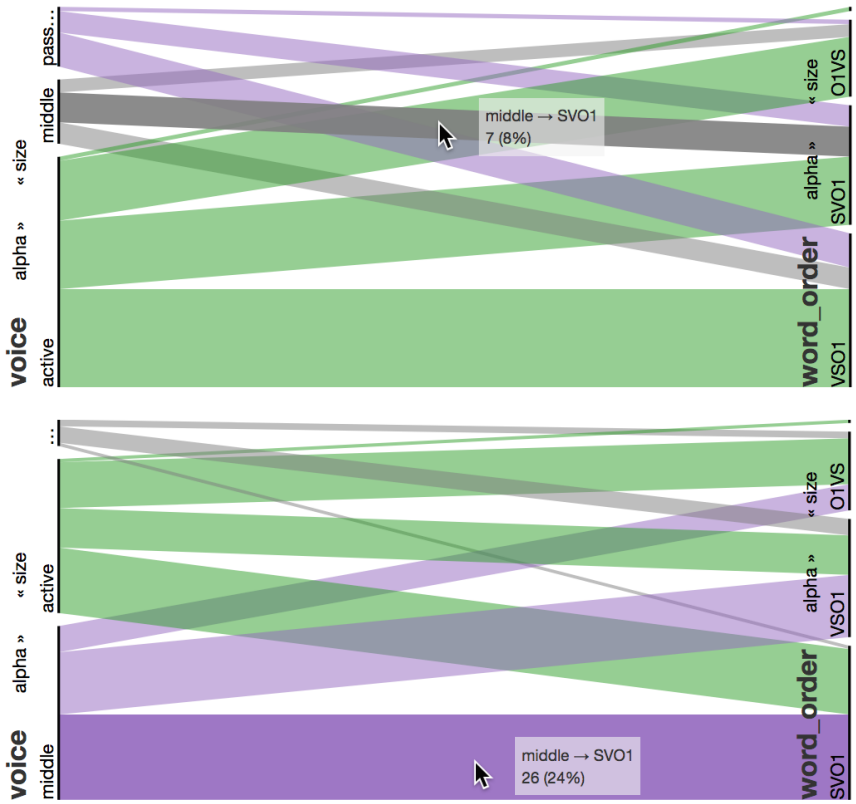

Figure 10: Dimension interactions for voice and word order in transitive sentences with a dative subject in the periods 1750-1899 (top) and 1900-2008 (bottom).

are tightly coupled with the subject role and the SVO order has been more firmly established, can the dative subjects conform to the overall word order settings.

4.3 Subject case and word order in Germanic

As a language which developed a more fixed word order while retaining a rich case morphology, Icelandic is cross-linguistically atypical. Kiparsky [1997] observed that in the history of Germanic, the development of a rigid word order generally correlates with the loss of inflectional morphology.

In order to gain a better understanding on how the Icelandic changes fit in with the historical developments of the other Germanic languages, we need to investigate correlations between subject case and word order more broadly in Germanic. This can be done fairly easily via HistoBankVis, since all we need is a well-structured tabular data set which contains the relevant data dimensions for the analysis. In terms of cross-linguistic comparison, we initially decided to make use of the family of historical Penn treebanks available for Germanic, e.g., the Penn Parsed Corpora of Historical English and HeliPaD. These have all been annotated according to the same overall guidelines so we assumed that comparability would be guaranteed. However, as discussed in section V, after having successfully worked with one new corpus (see below), we found we needed to take a step back and first address issues of data comparability.

We began our investigations by looking at the interaction between subject case and word order in HeliPaD, a corpus of Old Low German, i.e., Old Saxon. The HeliPaD annotation is similar to IcePaHC, containing sophisticated annotations for case marking and constituent order. We could thus automatically extract information about the verbs and verb types, subject and object case marking, and word order for each matrix declarative sentence on the basis of the annotation.

However, although the annotations in both HeliPaD and IcePaHC comply with the guidelines of the Penn historical corpora in general, there are differences.[8] In addition to providing annotations for case marking on nouns, HeliPaD annotates verbs and nouns for person and number, using the caret symbol (ˆ) to delimit these morphological annotations from the parts-of-speech. For example, in Figure 11, PROˆDˆ3ˆSG stands for a third person singular pronoun marked with dative case.

```
( (IP-MAT-SPE (CODE <C>)
              (ADVP-TMP (ADV than-than))
              (VBPI^3^SG thunkit-thunkian)
              (NP-SBJ (PRO^D^3^SG im-he))
              (CP-THT-SPE (C that-that)
                          (IP-SUB-SPE (NP-SBJ (PRO^N^3^SG hie-he))
                                      (NP-OB1 (PRO^A^3^SG sia-siu))
                                      (ADVP (ADV gerno-gern))
                                      (ADVP-TMP (ADV forth-forth))
                                      (CODE <R_2499>)
                                      (VB lestian-lestian)
                                      (MDPS^3^SG uuillie-willian)))
              (. ;-;))
  (ID OSHeliandC.1337.2498-2499))
```

Figure 11: Sample annotation for a sentence from HeliPaD.

The HeliPaD corpus consists of one text only, i.e., the *Heliand*, stemming from around 1100 CE. Although HeliPaD does not provide us with a diachronic perspective, we could gain insights into the interrelation between subject case and word order at this particular language stage by means of HistoBankVis. In analogy to the IcePaHC study, we filtered the data so that only transitive sentences were considered in the analysis. Then, we moved on to the visualizations of the data dimensions subject case and word order. Since HeliPaD lacks a diachronic component, the compact matrix is not suitable for presenting the data. However, the difference histogram and the dimension interaction visualization can be used for analysis. Both the difference histogram and the dimension interaction visualization indicate that the word order is rather flexible in HeliPaD, with many different possibilities, for example see the dimension interaction in Figure 12. Yet, SVO seems to be the preferred word order option, occupying the largest proportion of the word order axis. With respect to subject case marking, subjects are almost exclusively nominative. Only very few accusative and dative subjects, and no genitive subjects, were found in the transitive clauses. Suggestively, the non-nominative subjects we found did not occur together with the SVO order.

Given the small amount of dative subjects in transitive sentences in the corpus, we decided to go back to the filtering component in order to gain insights about the data from a qualitative perspective. To do this, we simply disabled the previous filter settings and filtered for sentences with a dative subject instead. In total, we found seven sentences containing a dative subject. In the result table, which is shown in Figure 13, we looked more closely at these sentences with respect to the dimensions verb and word order. The dative subjects in HeliPaD do not show a clear preference for appearing in a particular position, thus differing from nominative subjects. This is similar to our findings for the earlier stages of Icelandic based on the IcePaHC data. Moreover, like in Icelandic, the predicates occurring together with a dative subject in

[8]The annotation guidelines for the Penn historical corpora can be found here: https://www.ling.upenn.edu/~beatrice/annotation/.

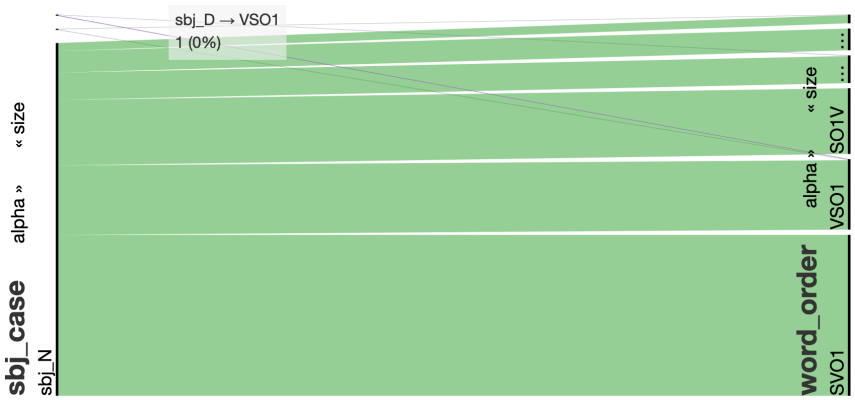

Figure 12: Dimension interaction for subject case and word order in transitive sentences from HeliPaD.

HeliPaD are mainly experiencer verbs, i.e., *thunkian* 'think, seem, feel', see, e.g., the sentence in Figure 11, and *likon* 'please'.

Result Table 7 records

Export Records | Continue to Visualization | Significance Analysis

ID	word_order	verb
OSHeliandC.106.211-214	SV	thunkian
OSHeliandC.1337.2498-2499	VS	thunkian
OSHeliandC.1758.3149-3150	SV	likon
OSHeliandC.1791.3193-3195	VO1S	likon
OSHeliandC.3047.5146-5148	VSO1	biginnan
OSHeliandC.684.1250-1254	O2SVO1	kiosan
OSHeliandC.75.157-158	SV	thunkian

Figure 13: Result table showing the dimensions word order and verb for sentences which have a dative subject in HeliPaD.

Overall, the findings for Old Saxon seem to line up with our findings for earlier Icelandic. In order to be able to add a cross-linguistic and diachronic perspective, we aimed at investigating the relationship between subject case and word order in the Penn Parsed Corpora of Historical English, which include the Penn-Helsinki Parsed Corpus of Middle English (PPCME2; Kroch and Taylor, 2000), Early Modern English (PPCEME; Kroch et al., 2004), and Modern British English (PPCMBE2; Kroch et al., 2016). Moreover, we planned to add the York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE; Taylor et al., 2003), which uses the same annotation format, in order to be able to capture a larger span of the English diachrony. However, as already noted above, we encountered serious problems during the process of data extraction. These problems are caused by a strong variation in the annotation of grammatical relations and case marking, resulting in issues of data comparability and interoperability. These issues are rooted in the annotation process and have lingustic causes, affecting areas of data uncertainty and provenance.

V UNCERTAINTY AND PROVENANCE ISSUES IN LINGUISTIC ANNOTATIONS

Historical corpora are generally annotated manually after a round of automatic pre-processing (see, e.g., Rögnvaldsson et al., 2012 on IcePaHC). While this allows for the precise annotation of quite complex structures and relations, manual annotations are still prone to errors. These errors might be caused by human unsystematicity, when the same linguistic structure is annotated differently across a single corpus. This is particular problematic with constructions that are rare, since the annotator might not be aware that, after having seen a large amount of other sentences, a certain structure has been annotated in a different way before. Moreover, problems arise when constructions are undergoing a change over time and a uniform annotation cannot be maintained across the time stages covered by the corpus. This raises issues of data uncertainty, with consequences for the reproducibility and replicability of analyses.

A further source of concern is data provenance [Buneman et al., 2000, Cui and Widom, 2003]. The field of data provenance within computer science is concerned with understanding how to model, record, and share metadata about the origin of data and the further sharing or processing that data has undergone. We find very little discussion or awareness of this topic within linguistics even though data provenance is of major concern. Issues arise already with respect to the raw data collection in terms of the origin of the data and its authenticity/reliability. Data may have been copied incorrectly in manuscripts or papers and once we proceed to the annotation of linguistic information, various steps in the annotation process may introduce errors into the annotated data. Keeping track of such potential sources for errors via a systematically organized set of meta-data seems to us to be a necessary next step within corpus linguistics.

The current state of the art has so far generally addressed issues of data soundness by conducting several rounds of annotation with different annotators, with intermediate cross-comparisons and validation of the resulting annotations, calculating, e.g., inter-annotator agreement. In this system, the creation of an annotated corpus is essentially a learning process consisting of a cycle of annotation, corpus correction, revision and reannotation.

However, manual annotations and re-checking are time-consuming and costly activities, so often a version of a corpus is published after just one or very few of such cycle iterations. Typically, after publication of the corpus more errors are found and reported by the users. Ideally, the reported errors would be corrected and a new version of the corpus released; however, due to funding and time limitations, the official versions of the corpora often remain in a faulty state for a longer period of time, while different individual sites may end up maintaining different versions of the original annotated resource.

This in turn almost inevitably leads to imperfect research results, with implications for the reproducibility and replicability of diachronic corpus studies. Often, researchers investigating a certain phenomenon 'cure' the corpus data with respect to their research endeavor to be able to deal with errors contained in the data, e.g., by excluding the erroneous cases from their investigations or by correcting the annotations for the subset of the data.

For example, during our investigation of subject case marking in IcePaHC, we encountered a range of annotation mistakes, e.g., a number of dative objects were annotated as subjects by mistake in constructions with an empty (non-overt) expletive subject. We manually corrected these mistakes. Thus, replicating our studies on the basis of the official version of IcePaHC is rather difficult. But it is possible to replicate our experiments via the 'cured' data set provided as part of HistoBankVis. Since our correction of the data immediately raises issues of data provenance (who decided when which of the dative grammatical relations are to be counted

as subjects?), we also provide the original annotation from IcePaHC for comparison as part of our 'cured' version. That is, in the result table, all extracted features for a sentence can be displayed along with the original underlying annotation. For example, Figure 14-right shows the annotation of a clause with the main verb *breyta* 'change', which takes a dative object, as provided by HistoBankVis. In the annotation, the dative argument has been erroneously annotated as a subject since there is no other overt subject candidate. The extracted features which are used in the analysis however, see Figure 14-left, have been corrected accordingly and can be inspected in comparison with the underlying annotation.

Sentence: 1480.JARLMANN.NAR-SAG,.1127

Dimension	Feature
verb	breyta
verb_type	VB
modal-aspectual	skulu
voice	active
word_order	SVO1
valency	trans
sbj_case	sbj_NOM
obj_case	obj1_DAT
obj2_case	-
sbj_type	sbj_*exp*
obj_type	obj1_PRO
obj2_type	-
subj_position	prefinite
v1	yes
genre	NAR

Metadata:

```
( (IP-MAT (CONJ og-og)
          (MDDS skyldi-skulu)
          (ADVP (ADVR svo-svo)
                (CP-CMP *ICH*-1))
          (NP-SBJ (Q-D öllu-allur))
          (VB breyta-breyta)
          (CP-ADV-1 (WADVP-2 0)
                    (C sem-sem)
                    (IP-SUB (ADVP *T*-2)
                            (NP-SBJ (PRO-N hann-hann))
                            (MDDI vildi-vilja)))
          (. .-.))
  (ID 1480.JARLMANN.NAR-SAG,.1127))
```

Figure 14: Original annotation and extracted features for a sentence from IcePaHC as provided by HistoBankVis.

The example of dative subjects illustrates a further problem with annotated data: uncertainty in linguistic annotations. Uncertainty with respect to linguistic annotations can arise when structures are inherently ambiguous and the surrounding contexts are not informative enough do decide on an interpretation. This is particularly difficult in transitioning periods, when structures are in the process of undergoing a functional change and vary between certain linguistic interpretations. As discussed above, we found that many instances of modern dative subjects arose from initial dative marked locatives/obliques in the process of a predicate gaining experiencer verb semantics. It is difficult for the annotator to decide exactly when the dative NP has transitioned to functioning as a subject rather than an object/oblique. Ambiguity and variation are part and parcel of historical change and annotating ambiguous structures is difficult. However, the standard method has been to decide on one of the possible options for analysis and to annotate that, rather than explicitly tagging/annotating instances of data uncertainty. In most cases, a stochastic decision is made and an ambiguous structure is tagged according to the option which occurs more frequently overall. However, such a priori decision making introduces unnecessary artefacts into the data.

Another issues arises with the annotation of changing structures across different stages of the language. For example, in the Penn Parsed Corpora of Historical English, verb phrases (VPs) are only annotated in cases where the boundaries are clear. In the older stages of English such boundaries are not given in most instances and the VP constituent is only annotated very rarely. In the modern stages however, the VP constituent is generally a given. Thus, the linguistic characteristics of a language stage determine the availability of annotations. The annotation of case marking in the historical English corpora poses a similar problem. While YCOE annotates for case marking, the other Penn Corpora of Historical English do not. This is due to the fact that English lost its morphological case marking system over the course of time. While data analysis must find a way of coping with annotation inconsistencies that directly reflect a changing language structure, other design decisions appear to unnecessarily complicate automatized cross-corpora historical analysis.

Even though the family of English Penn corpora generally adhere to the same guidelines, we found severe inconsistencies across the English corpora so that we were not able to apply a standardized approach to data processing via HistoBankVis unless we had invested a considerable amount of time into the historical analysis of the data. Without such a solid historical analysis it is not possible for us to clean or standardize the existing corpora, as our "cleaning" would most likely introduce errors given that we are not experts in Old and Middle English.

For example, YCOE takes nominative case marking as proxy for subjects. Thus, NPs are generally only annotated for case marking, but grammatical relations are not marked explicitly, see Figure 15. Only non-nominative subjects receive the extra subject tag -SBJ. These annotation decisions result from the fact that subjects and objects cannot be clearly demarcated in the Old English stage and nominative case marking might not always indicate the subject constituent (see, e.g., Allen, 1995). In the PPCME2 corpus for Middle English on the other hand, case marking is no longer annotated, but grammatical relations are clearly marked, see Figure 16.

```
( (IP-MAT (CONJ and)
      (NP-NOM (PRO^N he))
      (ADVP-TMP (ADV^T +ta))
      (VBDI genam)
      (NP-DAT-RFL-ADT (PRO^D him))
      (NP-ACC (N^A gemeccan)
          (ADJP-ACC (ADJ^A efenbyrde)
               (NP-DAT (PRO$ his) (N^D cynne))))
      (. ;)) (ID coeuphr,LS_7_[Euphr]:1.4))
```

Figure 15: Sample annotation for a sentence from YCOE.

```
( (IP-MAT (CONJ For)
      (NP-SBJ (PRO$ oure) (NPR Lord))
      (VBD knew)
      (NP-OB1 (D +te)
          (N waie)
          (PP (P of)
              (NP (D +te) (ADJ ry+gtful))))
      (. ,)) (ID CMEARLPS,2.21))
```

Figure 16: Sample annotation for a sentence from PPCME2.

In turn, the English Penn corpora differ from IcePaHC and HeliPaD which annotate for both grammatical relations and case marking. Yet, the way in which the information is encoded again differs between the corpora, see Figures 1 and 11 respectively. While we can see the differences,

we do not ourselves have enough language particular expertise to effect the necessary changes that would make these corpora comparable.

Overall, we conclude that although guidelines exist, there is a lack of a uniform standard for treebank creation. It is often difficult to process the annotated data in a standardized way, causing issues of data reproducibility and comparability of results. Moreover, data uncertainty is a core but only rarely addressed issue in historical linguistic work (cf. Merten and Seemann [2018]). Here again, we note that VA also has as yet unexplored potential in addressing these issues, as a promising line of research on the visualization of data uncertainty (see, e.g., Bonneau et al. [2014] for an overview) as well as data provenance (see, e.g., Stitz et al., 2016, Herschel et al., 2017, Ben Lahmar et al., 2018) exists. To our knowledge, such methods have not yet been applied to linguistic research. Integrating methods from the fields of uncertainty and provenance visualization into linguistic annotation processes and into the analysis process could be a great opportunity for mitigating issues of uncertainty, provenance, reproducibility and replicability in linguistic research.

VI CONCLUSION

In this paper we introduced a Visual Analytics system named HistoBankVis and showed how it has the potential to greatly facilitate historical linguistic research by allowing for efficient and fast interactive exploration of the underlying data. This is coupled with visual presentations of the computed correlations and statistics. The parallel sets technique provides an overview of interrelations found between various linguistic features of the corpus, allowing the researcher to formulate and test various different hypotheses with just a few clicks. HistoBankVis is furthermore good at generating at-a-glance overviews while still providing the ability to interact with the individual data points and annotations from the original corpus. We showed how the access to the underlying data does justice to one issue of data provenance in that we provide access to both our corrected version of the corpus and the original annotations of the official release.

However, in experimenting with the family of Penn-style treebanks for historical English, we also found that we could not usefully and systematically extend our investigations because of issues of annotation interoperability across corpora. We discuss specific issues with respect to data uncertainty and annotation standards that have come up in our work and note that these are general issues for any type of corpus work involving annotated data. These need to be solved in order to ensure replicability of results and analyses and we suggest that here, again, Visual Analytics provides a promising way forward and should thus become part and parcel of the methodological corpus linguistic toolkit.

ACKNOWLEDGEMENTS

We thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for funding our research within project D02 “Evaluation Metrics for Visual Analytics in Linguistics” – Project-ID 251654672 – TRR 161.

References

Cynthia L. Allen. *Case Marking and Reanalysis. Grammatical Relations from Old to Early Modern English.* Oxford University Press, Oxford, 1995.

R. Harald Baayen. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press, Cambridge, 2008.

Jóhanna Barðdal. The rise of dative substitution in the history of Icelandic: A diachronic Construction Grammar account. *Lingua*, 121(1):60–79, 2011.

Houssem Ben Lahmar, Melanie Herschel, Michael Blumenschein, and Daniel A. Keim. Provenance-based visual data exploration with EVLIN. In *Proceedings of International Conference on Extending Database Technology (EDBT)*, Vienna, Austria, 2018.

Fabian Bendix, Robert Kosara, and Helwig Hauser. Parallel sets: Visual analysis of categorical data. In *IEEE Symposium on Information Visualization*, pages 133–140. IEEE, 2005.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly, 2009.

Georges-Pierre Bonneau, Hans-Christian Hege, Chris R. Johnson, Manuel M. Oliveira, Kristin Potter, Penny Rheingans, and Thomas Schultz. Overview and state-of-the-art of uncertainty visualization. In Charles D. Hansen, Min Chen, Christopher R. Johnson, Arie E. Kaufman, and Hans Hagen, editors, *Scientific Visualization: Uncertainty, Multifield, Biomedical, and Scalable Visualization*, pages 3–27. Springer, London, 2014. ISBN 978-1-4471-6497-5. doi: 10.1007/978-1-4471-6497-5_1. URL https://doi.org/10.1007/978-1-4471-6497-5_1.

Hannah Booth, Christin Schätzle, Kersti Börjars, and Miriam Butt. Dative subjects and the rise of positional licensing in Icelandic. In M. Butt and T. H. King, editors, *Proceedings of the LFG17 Conference*, pages 104–124. CSLI Publications, 2017.

Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Data provenance: Some basic issues. In Sanjiv Kapoor and Sanjiva Prasad, editors, *FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science*, pages 87–93, Berlin, Heidelberg, 2000. Springer-Verlag.

Miriam Butt, Tina Bögel, Kristina Kotcheva, Christin Schätzle, Christian Rohrdantz, Dominik Sacha, Nicole Dehe, and Daniel Keim. V1 in Icelandic: A multifactorical visualization of historical data. In *Proceedings of VisLR: Visualization as added value in the development, use and evaluation of Language Resources*, Workshop at the 9th edition of the Language Resources and Evaluation Confernece (LREC 2014), Reykjavik, Iceland, 2014.

Yingwei Cui and Jennifer Widom. Lineage tracing for general data warehouse transformations. *VLDB*, 12(1): 41–58, 2003.

Chris Culy, Verena Lyding, and Henrik Dittmann. Structured Parallel Coordinates: a visualization for analyzing structured language data. In *Proceedings of the 3rd International Conference on Corpus Linguistics, CILC-11*, pages 485–493, April 6-9, Valencia, Spain, 2011.

Stefan Th. Gries and Martin Hilpert. The identification of stages in diachronic data: variability-based neighbour clustering. *Corpora*, 3(1):59–81, 2008.

Dag T. T. Haug and Marius L. Jøhndal. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In Caroline Sporleder and Kiril Ribarov, editors, *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34, 2008.

Einar Haugen. *Die skandinavischen Sprachen: Eine Einführung in ihre Geschichte*. Hamburg: Buske, 1984.

Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. A survey on provenance: What for? what form? what from? *The VLDB Journal*, 26(6):881–906, 2017.

Martin Hilpert and Stefan Th. Gries. Quantitative approaches to diachronic corpus linguistics. In Merja Kytö and Päivi Pahta, editors, *The Cambridge Handbook of English Historical Linguistics*, pages 36–53. Cambridge University Press, Cambridge, 2016.

Þorbjörg Hróarsdóttir. The decline of OV word order in the Icelandic VP. *Working Papers in Scandinavian Syntax*, 57:91–141, 1996.

Thorbjörg Hróarsdóttir. *Word Order Change in Icelandic. From OV to VO*. John Benjamins, Amsterdam, 2000.

Alfred Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1:69–91, 1985.

Alfred Inselberg. *Parallel Coordinates: VISUAL Multidimensional Geometry and its Applications*. Springer, New York, 2009.

Adam Jatowt, Ricardo Campos, Sourav S. Bhowmick, Nina Tahmasebi, and Antoine Doucet. Every word has its history: Interactive exploration and visualization of word sense evolution. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018*, pages 1899–1902, Torino, Italy, 2018.

Jóhannes Gísli Jónsson. Not so quirky: on subject case in Icelandic. In E. Brandner and H. Zinsmeister, editors, *New Perspectives on Case and Case Theory*, pages 129–164. CSLI Publications, Stanford, 2003.

Aikaterini-Lida Kalouli, Rebecca Kehlbeck, Rita Sevastjanova, Katharina Kaiser, Georg A. Kaiser, and Miriam Butt. ParHistVis: Visualization of parallel multilingual historical data. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 109–114, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4714.

Daniel A. Keim, Florian Mansmann, Joern Schneidewind, Jim Thomas, and Hartmut Ziegler. Visual Analytics: Scope and Challenges. In *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, Lecture Notes in Computer Science, pages 76–91. Springer, 2008.

Paul Kiparsky. The rise of positional licensing. In Ans van Kemenade and Nigel Vincent, editors, *Parameters of Morphosyntactic Change*, pages 460–494. Cambridge University Press, Cambridge, 1997.

Robert Kosara, Fabian Bendix, and Helwig Hauser. Parallel Sets: interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):558–568, July 2006. ISSN

1077-2626. doi: 10.1109/TVCG.2006.76.

Anthony Kroch and Ann Taylor. Penn-Helsinki Parsed Corpus of Middle English (PPCME2). Second edition, 2000.

Anthony Kroch, Beatrice Santorini, and Lauren Delfs. The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). First edition, 2004.

Anthony Kroch, Beatrice Santorini, and Ariel Diertani. The Penn-Helsinki Corpus of Modern British English (PPCMBE2). second edition, release 1, 2016.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Marie-Luis Merten and Nina Seemann. Analyzing constructional change: Linguistic annotation and sources of uncertainty. In *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality*, TEEM'18, page 819–825, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450365185. doi: 10.1145/3284179.3284320. URL https://doi.org/10.1145/3284179.3284320.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of LREC 2016*, pages 1659–1666, 2016.

Beth Randall. CorpusSearch: a Java program for searching syntactically annotated corpora. Dept. of Linguistics, University of Pennsylvania, Philadelphia, 2000.

Eiríkur Rögnvaldsson. Word order variation in the VP in Old Icelandic. *Working papers in Scandinavian syntax*, 58:55–86, 1996.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson, and Joel Wallenberg. The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of LREC 2012*, pages 1978–1984, 2012.

Christian Rohrdantz. *Visual Analytics of Change in Natural Language*. PhD thesis, University of Konstanz, 2014. Ph.D. Dissertation.

Christian Rohrdantz, Annette Hautli, Thomas Mayer, Miriam Butt, Daniel A. Keim, and Frans Plank. Towards tracking semantic change via visual analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, Portland: Oregon, 2011.

Christian Rohrdantz, Andreas Niekler, Annette Hautli, Miriam Butt, and Daniel A. Keim. Lexical Semantics and Distribution of Suffixes? a Visual Analysis. In *EACL 2012 Joint Workshop of LINGVIS*, Avignon, 2012.

Dominik Sacha, Andreas Stoffel, Florian Stoffel, Bum Chul Kwon, Geoffrey Ellis, and Daniel A. Keim. Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of Visual Analytics Science and Technology 2014)*, 20:1613, 2014. doi: 10.1109/TVCG.2014.2346481.

Christin Schätzle. *Dative Subjects: Historical Change Visualized*. PhD thesis, University of Konstanz, 2018.

Christin Schätzle and Hannah Booth. DiaHClust: an iterative hierarchical clustering approach for identifying stages in language change. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 126–135, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W19-4716.

Christin Schätzle and Dominik Sacha. Visualizing language change: Dative subjects in Icelandic. In *Proceedings of the LREC 2016 Workshop "VisLR II: Visualization as Added Value in the Development, Use and Evaluation of Language Resources"*, pages 8–15, Portorož, Slovenia, May 2016.

Christin Schätzle, Miriam Butt, and Kristina Kotcheva. The diachrony of dative subjects and the middle in Icelandic: A corpus study. In M. Butt and T. H. King, editors, *Proceedings of the LFG15 Conference*. CSLI Publications, 2015.

Christin Schätzle, Michael Hund, Frederik L. Dennig, Miriam Butt, and Daniel A. Keim. HistoBankVis: Detecting language change via data visualization. In Gerlof Bouma and Yvonne Asedam, editors, *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 32–39, Linköping, 2017. Linköping University Electronic Press.

Christin Schätzle, Frederik L. Dennig, Michael Blumenschein, Daniel A. Keim, and Miriam Butt. Visualizing linguistic change as dimension interactions. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 272–278, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4734.

Einar Freyr Sigurðsson. *Verbal Syntax and Case in Icelandic*. PhD thesis, University of Lund, 1989.

Holger Stitz, Stefan Luger, Marc Streit, and Nils Gehlenborg. Avocado: Visualization of workflow–derived data provenance for reproducible biomedical research. *Computer Graphics Forum*, 35(3):481–490, 2016. doi: 10.1111/cgf.12924. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12924.

Ann Taylor and Susan Pintzuk. The interaction of syntactic change and information status effects in the change from ov to vo in english. *Catalan Journal of Linguistics*, 10:71–94, 2011. ISSN 1695-6885.

Ann Taylor, Anthony Warner, Susan Pintzuk, and Frank Beths. The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE). http://www-users.york.ac.uk/ lang22/YCOE/YcoeHome.htm, 2003.

Roberto Theron and Laura Fontanillo. Diachronic-information visualization in historical dictionaries. *Information Visualization*, 14(2):111–136, 2015. doi: 10.1177/1473871613495844.

James J. Thomas and Kristin A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE-Press, 2005.

George Walkden. HeliPaD: the Heliand Parsed Database, version 0.9. http://www.chlg.ac.uk/helipad/, 2015.

George Walkden. The HeliPaD: a parsed corpus of Old Saxon. *International Journal of Corpus Linguistics*, 21(4): 559–571, 2016.

Joel C. Wallenberg, Anton Karl Ingason, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. Icelandic Parced Historical Corpus (IcePaHC). Version 0.9, 2011. URL http://www.linguist.is/icelandic_treebank.

Annie Zaenen, Joan Maling, and Höskuldur Thráinsson. Case and grammatical functions: The Icelandic passive. *Natural Language and Linguistic Theory*, 3:441–483, 1985. Reprinted in Joan Maling and Annie Zaenen (Eds.) *Syntax and Semantics 24: Modern Icelandic Syntax*, 95–164. New York: Academic Press. 1990.