Erschienen in: Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language / Bouma, Gerlof; Adesam, Yvonne (Hrsg.). - Linköping : Linköping University Electronic Press, 2017. - (NEALT Proceedings Series ; 32). - S. 32-39. - ISBN 9789176855034

HistoBankVis: Detecting Language Change via Data Visualization

Christin Schätzle
Department of Linguistics
University of Konstanz
christin.schaetzle@uni-konstanz.de

Michael Hund **Frederik L. Dennig**
Department of Computer Science
University of Konstanz
{michael.hund, frederik.dennig}
@uni-konstanz.de

Miriam Butt
Department of Linguistics
University of Konstanz
miriam.butt@uni-konstanz.de

Daniel A. Keim
Department of Computer Science
University of Konstanz
daniel.keim@uni-konstanz.de

Abstract

We present HistoBankVis, a novel visualization system designed for the interactive analysis of complex, multidimensional data to facilitate historical linguistic work. In this paper, we illustrate the visualization's efficacy and power by means of a concrete case study investigating the diachronic interaction of word order and subject case in Icelandic.

1 Introduction

The increasing availability of digitized data for historical linguistic research has led to an increased use of quantitative methods, with an employment of increasingly sophisticated statistical methods (Manning and Schütze, 2003; Baayen, 2008; Hilpert and Gries, 2016). However, diachronic investigations involve understanding highly complex interactions between various linguistic and extra-linguistic features and structures. Due to the complexity of this multidimensional data, significant patterns may not be uncovered or understood.

We therefore designed *HistoBankVis*, a novel visualization system which facilitates the investigation of historical change by integrating methods coming from the field of Visual Analytics (Keim et al., 2008). HistoBankVis allows a researcher to interact with the data directly and efficiently while exploring correlations between linguistic features and structures. Our system in effect consigns to history the painstaking work of finding patterns across various different tables of features, numbers and statistical significances. Rather, in our system, the researcher can first identify certain features to be investigated and within minutes can obtain an at-a-glance overview that provides information about whether interesting patterns can indeed be identified across features over time. Relevant patterns can then be further analyzed by drilling down to individual data points and new hypotheses can be generated. These hypotheses may then be tested anew with respect to a fresh look at the data. Given that historical data typically present a data sparsity problem, we also provide multiple different ways of calculating or estimating statistical significance, e.g. Euclidean distance, to deal with the small number of data points.

The efficacy of HistoBankVis is exemplified via a concrete test case, namely a syntactic investigation of the Icelandic Parsed Historical Corpus (IcePaHC, Wallenberg et al., 2011). The IcePaHC is annotated in the Penn TreeBank style (Marcus et al., 1993) and consists of 61 texts with around one million words covering all attested stages of Icelandic.

The visualization not only identifies changing syntactic features in IcePaHC ad-hoc by means of a well-structured statistical analysis process, but also supports the researcher in the generation and validation of hypotheses. Moreover, the visualization bridges the gap between annotated values, statistical analyses and the actual underlying data by providing access to the original sentences from IcePaHC during a data filter and selection process.

2 Related Work

Visualizations tailored to the analysis of historical linguistic data range from work on modal verbs within historical academic discourse (Lyding et al., 2012) to the cross-linguistic spread of new suffixes throughout mass media (Rohrdantz et al., 2012; Rohrdantz, 2014), the semantic change of word meanings (Rohrdantz et al., 2011) and the

Figure 1: The workflow of our novel visualization system: based on the analysis task, the user splits documents into sentences, extracts and filters for relevant linguistic factors (=dimensions) as well as customized or pre-defined time periods. The visualization provides different levels of detail that the user can switch back and forth between. The system crucially allows for a feed-back loop by which the user can iterate back to refilter or modify the underlying data.

evolution of meanings as represented in dictionaries (Theron and Fontanillo, 2015). With respect to Icelandic and IcePaHC, Butt et al. (2014) and Schätzle and Sacha (2016) designed a glyph visualization for the analysis of individual factors leading to syntactic change. HistoBankVis builds on the experiences gathered while working on the glyph visualization. In particular, the glyph visualization was not able to deal elegantly with the potentially large amounts of interacting data dimensions that are of interest for any kind of historical linguistic research question. The system also relied on specific assumptions about the nature of the data and the research questions to be pursued.

The goal of HistoBankVis thus is to provide both a more generically applicable system for historical linguistic research and a more flexible investigation of data dimensions, allowing for exploratory access to a potentially high number of factors. The system also either provides for the possibility of analyzing each factor at a time or to look at interactions of interrelated factors on demand.

3 The HistoBankVis System

3.1 Iterative Analysis Workflow

The idea behind HistoBankVis is an iterative workflow, displayed in Figure 1. The text data are processed by extracting linguistic factors which have been identified by the researcher as relevant for the task at hand. This is typically done by a previous careful consultation of the relevant theoretical literature. In what follows, we call these linguistic factors *dimensions* and their possible values *features*. For example, the linguistic factor *voice* is a data dimension containing the features *active, passive* and *middle*. Based on the analysis task, the user can filter for a subset of the data (e.g., only certain dimensions/features or only sentences from a specific set of genres or time periods). To visualize the historical developments of dimensions over time, the researcher needs to define time periods for the comparison. The visualization then allows the researcher to interactively compare the distribution of all selected features and dimensions of the filtered sentences across the different time periods. The visualization moreover provides details-on-demand on all views via mouse interaction techniques. Finally, the user can react to the insights collected from the visualization and test new hypotheses by interacting directly with the system. Interactions could involve changes in the data processing, adapting the filters or modifying the time periods.

3.2 Data Processing

As part of a concrete case study, we are currently working with HistoBankVis to investigate the interaction between subject case and word order. Although Icelandic is generally taken to have changed only little with respect to syntax and morphology (Thráinsson, 1996; Rögnvaldsson et al., 2011), several changes with respect to word order have been documented (e.g., Kiparsky (1996), Rögnvaldsson (1996) and Hróarsdóttir (2000) on the change from OV to VO and Franco (2008) and Sigurðsson (1990) on the decrease of V1). Some questions regarding Icelandic on the basis of the existing literature are: Which strategies are used to mark grammatical relations? Do these strategies change in the history of Icelandic?

In order to investigate these questions, we identified relevant linguistic dimensions based on information contained in the theoretical literature and automatically extracted these dimensions via Perl scripts from the annotation of IcePaHC. We included information about the type of verb, voice, word order, case and valency. These dimensions were furthermore mapped onto the sentence IDs contained in IcePaHC. These sentence IDs pro-

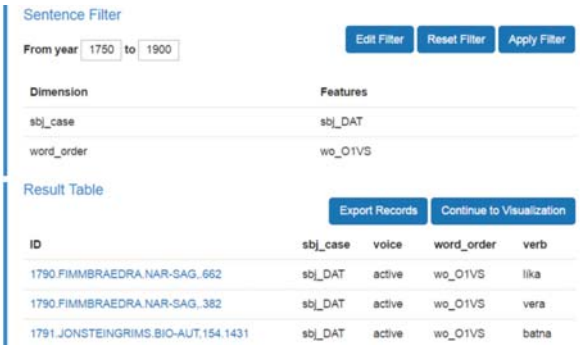


Figure 2: The filter module: The researcher can filter for data from specific years containing only specific data features before generating a data set with previously selected dimensions.

vide information about the year date, the name and the genre of the text in which the sentence occurs. As part of our preprocessing, we used this information to generate a well-structured database that HistoBankVis can operate on.

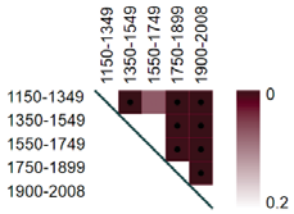
3.3 Task-based Filtering

Once the data has been processed, the researcher has the option of filtering for sentences with relevant properties. Besides filtering for data within a specific time frame, the researcher can visually construct SQL-like filters for features in the database. Based on the analysis task, the dimensions and features can be combined with logical AND- or OR-functions. For example in Figure 2, we filtered for sentences which contain the word order OVS, i.e., (direct) object, verb, subject, within texts from 1750 to 1900 CE. The researcher then further selects the dimensions for analysis, e.g., subject case, voice, word order and the verbs involved. Each sentence matching the configured filter can be analyzed by displaying it and its Penn Treebank annotation in conjunction with all available extracted features on demand. Thus, the filtering component of HistoBankVis serves as a preprocessing system on its own, providing the researcher with a more fine-grained view on the data by only selecting a certain number of dimensions and/or a subset of sentences. This not only allows the researcher to become familiar with and explore the data set at hand, but also furthers the understanding of the data quality by granting access to detailed information about each data point. Additionally, the filtered data set can be downloaded as a CSV-file to be processed in a different tool of choice.

3.4 Analyzing Change over Time

To analyze and visualize the selected dimensions over time, the researcher has to first specify relevant time periods. For Icelandic, our system automatically supports two common divisions into time periods: (1) Old and Modern Icelandic, i.e., 1150–1550 and 1550–2008 CE (e.g., see Thráinsson (1996); referred to as *Range A* in the following); (2) more fine-grained periods as defined per Haugen (1984), i.e., 1150–1350, 1350–1550, 1550–1750, 1750-1900, and 1900–2008 CE (referred to as *Range B* in what follows). The system also allows the user to enter fully customized periods.

Compact Matrix Visualization We provide a *compact matrix* representing an understanding about differences between the selected dimensions across time periods. Each row and column of the matrix corresponds to one period. This especially facilitates the comparison of the first period to all others and every period with its predecessor (entries along the diagonal of the matrix). HistoBankVis provides two comparison modes: statistical significance and distance based. In both modes the difference between two periods is mapped onto a colormap (red depicts a high and white a low significance/distance). To measure the statistical significance, HistoBankVis supports a χ^2-test. Here the p-value is mapped to the colormap: red corresponds to $p = 0$ and white to $p \geq 0.2$. ● indicates that the difference is statistically significant (with $\alpha = 0.05$) and ✕ signals the absence of necessary preconditions. Alternatively, the Euclidean distance can be used when the necessary preconditions for the χ^2-test are not met, e.g., in order to deal with problems of data sparsity. A high Euclidean distance reflects a large difference in the compared distributions and indicates high significance. The visual patterns in the matrix view serve as a measure of quality and "interestingness" as one can quickly spot combinations of periods which differ significantly and should be investigated further.



Difference Histograms Visualization While the overview matrix is a useful means to quickly gain insights, *difference histograms* provide a view

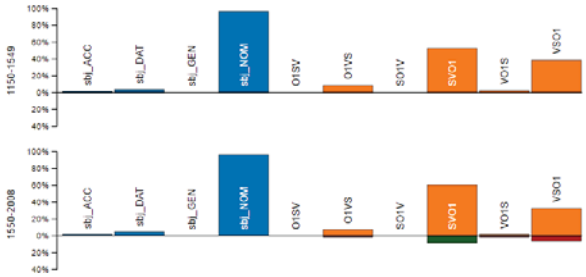


Figure 3: Difference histograms for the distribution of subject case and word order in transitive sentences in Old versus Modern Icelandic.

with more details on the diachrony of individual features. Each time period is visualized as one bar chart, see Figure 3 for Range A. Each dimension is encoded via a different color, e.g., blue for subject case and orange for word order. The height of one bar corresponds to the percentage of sentences containing the respective features. Additional information, such as the underlying sentences, the exact percentages and the relative size of the feature occurrence compared to the overall text size can be accessed via several interaction techniques.

The comparison of bar heights along different periods provides insights on which dimensions and/or combinations of features change over time. We furthermore computed the difference between two neighboring periods and visualized this as a separate bar chart below the percentages of features in the histograms. The color green indicates that a feature increased compared to the previous period and red indicates that the feature decreased, e.g., SVO increases in Figure 3, while VSO decreases. The system also allows for other comparison modes such as the option of comparing each time period with the first or last period, with the average of all periods, or with the average of all periods before the current one in order to make deviating features stand out and to observe trends.

3.5 Hypothesis Generation and Feedback ↻

Once the patterns in the data have been explored, hypotheses tested and perhaps new ones formed, the researcher can feed the knowledge gained back into each of the individual parts of the system by changing the filters, trying out different time periods or by going back to the data processing step and including different or more features. This creates an iterative analysis process in which knowledge-based and data-driven modeling are combined.

3.6 Access and Usability

HistoBankVis is implemented as an on-line browser-app and is freely available via our website.[1] The website includes a demo video which guides the user through the different analysis steps. Each analysis step performed by the user (e.g., applied filters or selected dimensions) and the current views (e.g., difference histograms) are encoded by uniquely identified URLs. The URL scheme allows a researcher to easily store and retrieve visualizations with different properties. It also allows for knowledge and data exchange between researchers supporting collaborative research projects since URLs representing a certain view on the data can be shared with other researchers locally or non-locally.

Besides the IcePaHC dataset, which HistoBankVis uses as its default data set, the system makes provision for researchers who would like to load their own data into HistoBankVis. The specifications for the new data sets are also provided. The data needs to be in a tab-separated format in which each line starts with a unique ID followed by the year date corresponding to the entry and an arbitrary number of data dimensions. Additionally, a file with meta information about the source texts (e.g., the text itself and/or the syntactically parsed sentence structure) can be uploaded as well. The mapping between the data dimensions and meta information is done via the unique ID. Further instructions and an example data set with abstract dimensions and values are available on our website, providing the user with more information on how to prepare and structure the data set.

4 Case Study

The visualizations above were obtained as part of an on-going investigation into correlations between word order and dative subjects. First, we investigated the word order distribution across all subjects in Old and Modern Icelandic by filtering for sentences containing a subject (S), a verb (V) and a direct object (O/O1). We subsequently visualized the dimensions subject case and word order. The difference histograms not only show that SVO is the dominant order for both time periods, but also that SVO is slightly increasing over time, accompanied by a concomitant decrease of VSO,

[1] http://histobankvis.dbvis.de

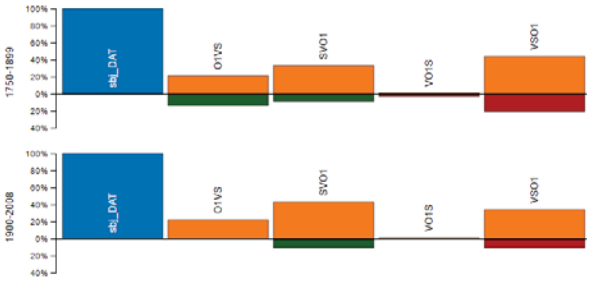


Figure 4: Word order within the past two time periods from Range B for dative subjects. See Figure 7 in the Appendix for all periods.

see Figure 3. Moreover, the subjects involved are mainly nominative and more rarely dative.

Following this initial broad look at the data, we took a more nuanced look and visualized the data with respect to Range B. Here, the distance matrix (see Section 3.4) revealed at-a-glance that there is a significant change in the last two time periods. By comparing each range with the previous one, a fairly large increase of SVO becomes visible in the last time stage (cf. the green bar under SVO1 in Figure 4), while VSO is further decreasing, as shown by the red bar underneath VSO1. Dative subjects also increase slightly in the last range (see Figure 5 in the Appendix).

Given these findings, a separate analysis of word order in dative and nominative subjects was in order. This could easily be done by configuring the filter settings to only include either dative or nominative subjects. While the word order histograms for nominative subjects (see Figure 6 in the Appendix) conform to the overall developments of word order for all subjects, dative subjects pattern differently. The difference histograms in Figure 4 show that VSO is the dominant word order for dative subject sentences until around 1900, which is when SVO surpasses VSO as dominant order following a continuous increase.

Strikingly, we found the OVS order to be standing out in the second to last time stage by deviating strongly from the average appearance in the other stages. We thus filtered the data once more for only OVS and noted that the verbs found in the relevant time period are mainly experiencer predicates, such as *líka* 'like, please', see Figure 2. We postulate that these experiencer verbs are subject to lexicalization over time and are changing from a structure in which the experiencer/goal is realized as a structural object to a structure whose sentient experiencer/goal participant is instead realized as a structural subject. I.e., something like *This pleases me*, in which the experiencer is an object is instead realized as *I like this*, where the experiencer is a subject. The general ability of experiencer/goal arguments to be realized in principle as either an experiencer subject or an undergoer/goal object has been well documented across languages (cf. Grimshaw, 1990), as have general linguistic principles by which sentient/animate participants are preferentially realized as subjects (e.g., Dowty, 1991). We postulate that the Icelandic pattern is an instance of a historical change by which experiencer participants are increasingly realized as dative subjects. Our findings are also in line with recent research on the interaction between middle morphology and dative subjects by Schätzle et al. (2015).

Recall that we also found an overall change towards SVO word order. We postulate that this points towards the development of a fixed preverbal subject position in the history of Icelandic with the 19th century as a major key turning point. Dative subjects show a slower tendency to follow this development. We explain this slower tendency by the fact that experiencer/goal arguments were not canonical subjects and that many of them underwent reanalysis from object to subject first.

Other changes with respect to Icelandic word order have been reported to happen around the same time, e.g. the decrease of V1 (Sigurðsson, 1990; Butt et al., 2014) and the loss of OV (Hróarsdóttir, 2000). These and other findings are the subject of on-going work, also with the aid of HistoBankVis. We hope to have been able to demonstrate the efficacy of HistoBankVis with this snap shot of our on-going historical work.

5 Conclusion

In conclusion, we present a powerful new visualization tool, HistoBankVis, which facilitates the detection and analysis of language change with respect to an annotated corpus. By means of just a few clicks, we were able to investigate changes in word order in interaction with subject case.

Our method combines knowledge-based and data-driven modeling. The system was developed on the IcePaHC, but has been set up in a generalized manner so that it can be applied to any Penn Treebank-style annotated corpus or indeed

any annotated corpus as the visualization builds on a database designed to process any kind of well-structured data set.

HistoBankVis can also be used as a preprocessing and filtering tool without the visualization module as it allows for the export of filtered data sets. That is, the user can simply choose to filter the data set according to some features and dimensions that they specify. The user does not need to proceed on to a visualization of the selected dimensions, but can choose to export just those filtered records. If the user does choose to proceed to the visualization, the fact that the visualization is implemented as a browser-app means that each analysis step remains accessible via a single identification URL. This not only facilitates a collaborative research structure by allowing researchers to share their analyses and perspectives on the data across machines, it also facilitates the analysis process since individual perspectives on the data can be stored and individual analyses can be (re)retrieved at any time.

Finally, we hope to have demonstrated that HistoBankVis represents a novel and effective visualization system which immensely facilitates the investigation of historical language change.

Acknowledgments

We thank the German Research Foundation (DFG) for financial support within the projects A03 and D02 of the SFB/Transregio 161.

References

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

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

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Irene Franco. 2008. V1, V2 and criterial movement in Icelandic. *Studies in Linguistics*, 2:141 – 164.

Jane Grimshaw. 1990. *Argument Structure*. The MIT Press, Cambridge.

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

Martin Hilpert and Stefan Th. Gries. 2016. 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.

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

Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. 2008. Visual analytics: Definition, process, and challenges. In *Information visualization*, pages 154–175. Springer.

Paul Kiparsky. 1996. The shift to head-initial VP in Germanic. In H. Thráinsson, J. Peter, and S. Epstein, editors, *Comparative Germanic Syntax*. Kluwer.

Verena Lyding, Stefania Degaetano-Ortlieb, Ekaterina Lapshinova-Koltunski, Henrik Dittmann, and Christopher Culy. 2012. Visualising linguistic evolution in academic discourse. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 44–48. Association for Computational Linguistics.

Christopher D. Manning and Hinrich Schütze. 2003. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, 6 edition.

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

Eiríkur Rögnvaldsson, Anton Karl Ingason, and Einar Freyr Sigurðsson. 2011. Coping with variation in the Icelandic Parsed Historical Corpus (IcePaHC). In J.B. Johannessen, editor, *Language Variation Infrastructure*, volume 3 of *Oslo Studies in Language*, pages 97–112.

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

Christian Rohrdantz, Annette Hautli, Thomas Mayer, Miriam Butt, Frans Plank, and Daniel A. Keim. 2011. Towards Tracking Semantic Change by Visual Analytics. In *Proceedings of ACL 2011 (Short Papers)*, pages 305–310.

Christian Rohrdantz, Andreas Niekler, Annette Hautli, Miriam Butt, and Daniel A. Keim. 2012. Lexical Semantics and Distribution of Suffixes - A Visual Analysis. *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 7–15, April.

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

Christin Schätzle and Dominik Sacha. 2016. Visualizing language change: Dative subjects in Icelandic. In Annette Hautli-Janisz and Verena Lyding, editors, *Proceedings of the LREC 2016 Workshop "VisLR II: Visualization as Added Value in the Development, Use and Evaluation of Language Resources"*, pages 8–15.

Christin Schätzle, Miriam Butt, and Kristina Kotcheva. 2015. 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.

Halldór Ármann Sigurðsson. 1990. V1 declaratives and verb raising in icelandic. In Joan Maling and Annie Zaenen, editors, *Modern Icelandic Syntax (Syntax and Semantics 24)*, pages 41–69. Academic Press, San Diego.

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

Höskuldur Thráinsson. 1996. Icelandic. In Ekkehard König and Johan van der Auwera, editors, *The Germanic Languages*, pages 142–189. Routledge, London.

Joel C. Wallenberg, Anton Karl Ingason, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. 2011. Icelandic Parced Historical Corpus (IcePaHC). Version 0.9.

Appendix

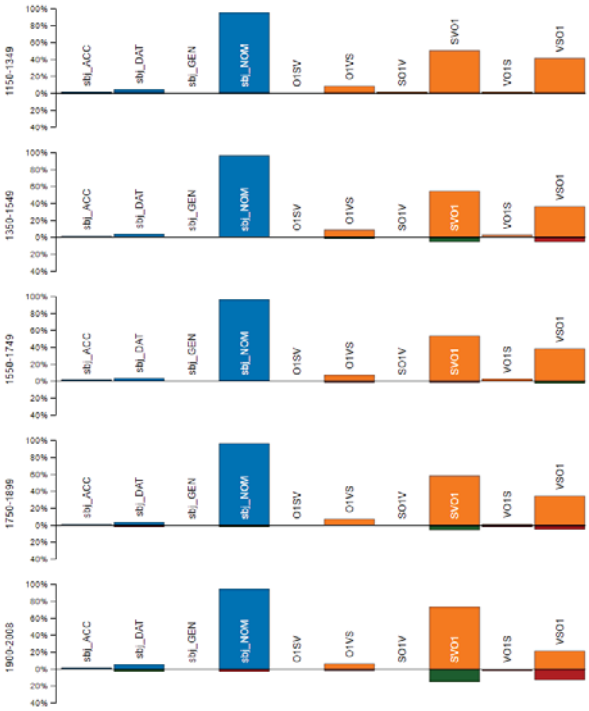


Figure 5: Word order and subject case for Range B: The blue bars represent the general distribution of subject case within the filtered data set (sentences containing a subject, a direct object and a verb). The orange bars represent the possible word order patterns occurring in the data. Over time, SVO increases consistently with respect to each previous time period (green bar). At the same time, VSO decreases (red bar). The dimension subject case remains stable until the last time period in which a slight increase of dative subjects is visible.

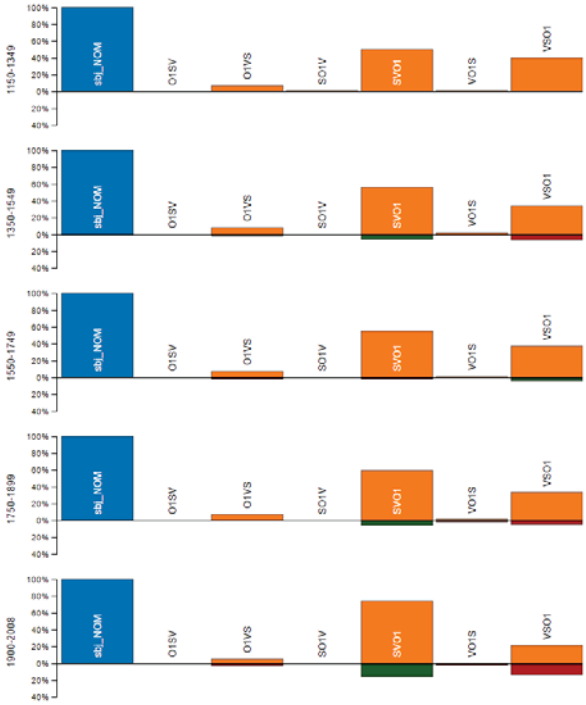


Figure 6: Word order for Range B for nominative subjects. The diachrony of the word order patterns corresponds to the one found for all subjects (as displayed in Figure 5), i.e., VSO is decreasing across the time stages, while SVO is increasing.

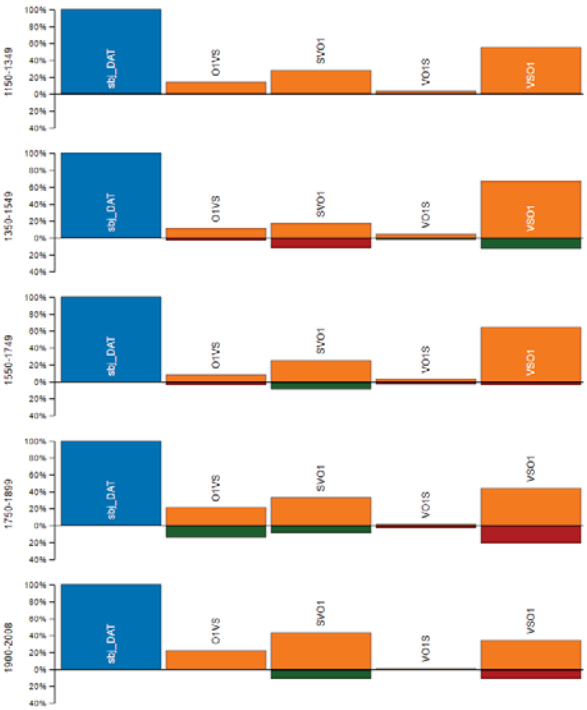


Figure 7: Word order for Range B for dative subjects. VSO is the dominant word order up until the last time stage in which SVO becomes the dominant word order after continuously increasing along the whole corpus. Moreover, OVS word order stands out in the second to last time stage.