Visualizing Linguistic Change as Dimension Interactions

Christin Schätzle, Frederik L. Dennig, Michael Blumenschein,
Daniel A. Keim and **Miriam Butt**
University of Konstanz
firstname.lastname@uni-konstanz.de

Abstract

Historical change typically is the result of complex interactions between several linguistic factors. Identifying the relevant factors and understanding how they interact across the temporal dimension is the core remit of historical linguistics. With respect to corpus work, this entails a separate annotation, extraction and painstaking pair-wise comparison of the relevant bits of information. This paper presents a significant extension of HistoBankVis, a multilayer visualization system which allows a fast and interactive exploration of complex linguistic data. Linguistic factors can be understood as data dimensions which show complex interrelationships. We model these relationships with the Parallel Sets technique. We demonstrate the powerful potential of this technique by applying the system to understanding the interaction of case, grammatical relations and word order in the history of Icelandic.

1 Introduction

Historical linguistic research is corpus-based by nature. In recent years, a large amount of digitized and linguistically well-annotated corpora have been made available and the historical linguistic research community is increasingly employing quantitative and statistical methods for their analysis. This includes the calculation of co-occurrence frequencies, correlations, dispersion statistics, and more sophisticated methods such as clustering (see, e.g., Hilpert and Gries, 2016). Statistical measurements are well-established for the analysis of linguistic change with respect to the quantification of individual structures. However, these methods are not per se suitable for the uncovering and understanding of the complex interactions between various linguistic structures typically involved in a change.

This paper extends our HistoBankVis system (Schätzle et al. 2017) by a powerful visualization to analyze and explore the interrelationship between multidimensional linguistic factors. HistoBankVis was specifically developed for the analysis of historical linguistic data. The system allows for an interactive exploratory access to a complex data set by using several interlinked visualization and filtering techniques. The extension presented in this paper integrates a *Dimension Interaction* visualization, based on the Parallel Sets technique (Bendix et al., 2005; Kosara et al., 2006), into the HistoBankVis system. Parallel Sets support the flexible analysis, visual presentation, and exploration of correlations between a large number of features from different *data dimensions*, i.e., linguistic factors, which immensely facilitates the analysis of interactions between features from changing dimensions.

We demonstrate the efficacy of the Dimension Interaction technique for historical linguistic research using a concrete case study which investigates interrelations between word order changes and subject case in Icelandic. The visualization not only proved to be an extremely valuable tool for the analysis of complex interactions across different data dimensions, but also facilitated the uncovering of previously unknown interdependencies in the data.

2 Challenges for Diachronic Linguistics

More and more digitized text corpora have been made available for historical linguistic research in recent years. These comprise large linguistically unannotated collections of historical texts, e.g., the Bibliotheca Augustana,[1] TITUS[2] and GRETIL,[3] but also increasingly include annotated corpora.

Annotated corpora are usually smaller in size and have undergone a manual annotation process in addition to an automatic preprocessing. The

[1] https://www.hs-augsburg.de/~harsch/augustana.html
[2] http://titus.uni-frankfurt.de/indexd.htm
[3] http://gretil.sub.uni-goettingen.de/

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 272–278
Florence, Italy, August 2, 2019.

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

Table 1: Definiteness distribution of NPs across different word orders in Icelandic (Hróarsdóttir, 2000, 136).

manual annotation procedure allows for a linguistically sophisticated annotation which often includes a deep syntactic analysis of hierarchies and dependencies between phrase structure constituents. Prototypically, such structural information is annotated in the Penn Treebank-style (Marcus et al., 1993). Examples are the Penn Parsed Corpora of Historical English (Kroch and Taylor, 2000; Kroch et al., 2004, 2010), the Icelandic Parsed Historical Corpus (IcePaHC, Wallenberg et al., 2011), the Heliand Parsed Database (Walkden, 2015), the Latin Dependency Treebank (Bamman and Cane, 2006), the Prague Dependency Treebank (Hajič, 1998), and PROIEL (Haug and Jøhndal, 2008).

The standard procedure within diachronic corpus linguistics incorporates the use of programming languages for text processing and statistical analysis, e.g., Python, Perl, and R (Baayen, 2008; Bird et al., 2009; Christiansen et al., 2012), to extract the relevant patterns on the basis of the annotation and to calculate co-occurrence frequencies and statistical significances across different time stages. A multitude of high-dimensional data tables containing different features and data characteristics are generated. For example, Table 1 represents a prototypical historical linguistic data set.

Finding significant patterns and feature interactions across such tables is by no means a trivial task, as a temporal component not only has to be factored in, but numbers computed for several features belonging to different data dimensions need to be compared across many data tables of varying size. Moreover, statistical significances are difficult to interpret and often calculated on the basis of only very few occurrences of the actual observation, derogating the significance measures and statistical conclusions. Thus, meaningful patterns may not be identified, whereas irrelevant patterns are likely to surface as significant. Interesting patterns may furthermore stay hidden when an analyst chooses temporal episodes that are either too coarse or too fine grained for the statistical analysis. The factors causing a language to change are often unknown or at least highly debated among researchers. Therefore, a researcher may have to conduct several different analyses, experimenting with different combinations of data dimensions. This is time-consuming and the resulting data is difficult to navigate.

HistoBankVis addresses these challenges by providing an exploratory access to a high-dimensional data set by means of different visualization layers combined with a structured statistical analysis. The system is part of on-going work which investigates visualization possibilities and the needs of historical linguistic data stored in treebanks.

3 HistoBankVis: a multilayer visualization system

As part of our on-going work, we developed HistoBankVis, a visualization system originally designed for the investigation of syntactic change in Icelandic based on IcePaHC (see Schätzle et al. 2017 for details). The tool is an online browser app and publicly available.[4] HistoBankVis requires well-structured, tabular datasets in the csv-format as input. Thus, corpus data needs to be processed by extracting linguistic factors relevant for the research task, usually identified by consulting the theoretical literature. HistoBankVis stores these factors as data dimensions in an SQL database, with the corresponding values referred to as features.

The user can filter for a subset of the data, specifically for dimensions and features from particular time periods. Before visualizing the historical developments of the selected data dimensions, the researcher has to define time periods for the temporal comparison, either by specifying them manually or by selecting predefined periods.

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

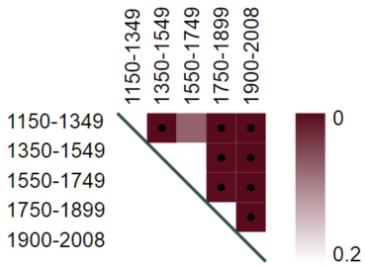

Figure 1: Compact matrix visualization showing statistically significant differences between data distributions from different time periods.

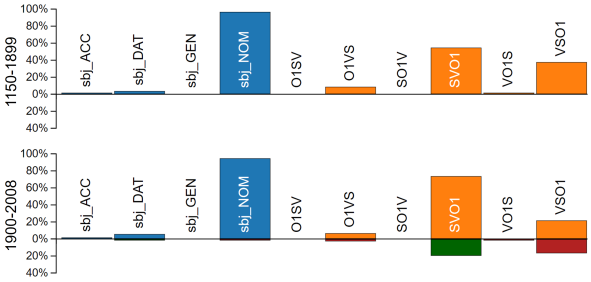

Figure 2: Difference histograms for the distribution of subject case and word order pre- and post-1900.

The original HistoBankVis version (Schätzle et al., 2017) has two main visualization components, the *Compact Matrix* and the *Difference Histograms Visualization*. Both visualizations allow researchers to interactively compare the distributions of selected features and dimensions of the filtered sentences across time periods on different granularity levels. The compact matrix visualizes differences between the selected data dimensions across time stages. Each row and column represents one period as shown in Figure 1. The differences are measured by χ^2-tests or Euclidean distances and represented by color. The matrix is a useful means to show differences among all time period combinations.

Difference histograms provide a more nuanced view on the diachrony of individual features and dimensions. The difference histograms visualize each time period as one composed bar chart, see Figure 2. For each time period, the dimensions are encoded via different colors and can be inspected in parallel. The bar height corresponds to the percentage of sentences containing a given feature in the respective time period. To facilitate the comparison of periods, we show the difference between the distributions of two neighbouring time periods with a separate bar chart below each feature bar. A green bar indicates that a feature increased compared to the previous period and red indicates that the feature decreased. For example, in Figure 2, SVO1 (Subject-Verb-Direct Object) word order increases, while VSO1 (Verb-Subject-Direct Object) decreases.

While the matrix and the histograms allow for the exploration of differences between linguistic factors across different time periods, the representations lack a perpendicular comparison of interactions between different factors to correlate, e.g., the occurrence of a particular type of subject case with the observed word order variation. That is, while it is clear that most of the subjects have nominative case (sbj_NOM) in Figure 2, one cannot correlate this information directly with word order: the question of which attested word order possibilities the subjects appear in must be tackled in a different way. To this end, we extended HistoBankVis with a visualization of dimension interactions.

4 Dimension Interaction Visualization

To provide insights into the interrelation between multiple features of different dimensions, we extended the HistoBankVis system by a *Dimension Interaction* visualization, based on the Parallel Sets technique (Bendix et al., 2005; Kosara et al., 2006). Parallel Sets extend the idea of Parallel Coordinates (PC; Inselberg 1985, 2009) to a frequency-based representation of categorical data dimensions.

PC represent relations between individual data points from a multidimensional data set on a 2D plane by visualizing each dimension along a vertical axis with the related features of the dimensions being connected by a polyline. This allows to identify both relationships between data points and neighboring dimensions. Structured Parallel Coordinates (Culy et al., 2011), a specialized version of PC for the analysis of linguistic data, have been used to analyze word co-occurrences (Culy et al., 2011) and to investigate meanings of modal verbs within historical academic discourse (Lyding et al., 2012). Moreover, the diachronlex diagrams by Theron and Fontanillo (2015) which track the evolution of meanings as represented in historical dictionaries make use of PC.

Parallel Sets visualize the frequency of each feature as proportions of equally spaced vertical lines (data dimensions). In this way, Parallel Sets allow for the sophisticated investigation of interactions between features from different data dimensions, whereas PC only allow for the analysis of co-occurrence frequencies of specific features. For ex-

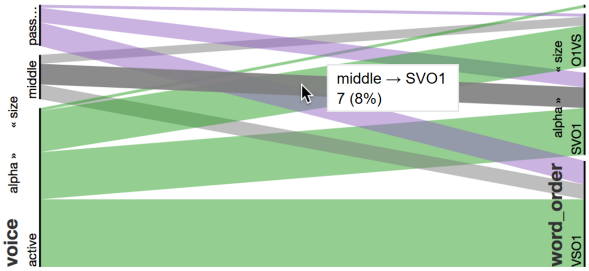

Figure 3: Dimension interaction for voice and word order in dative subject sentences from 1750–1899.

ample in Figure 3, the interactions between the dimensions voice and word order are visualized. The dimensions are connected by colored ribbons. The size of a ribbon indicates the share which a feature holds of a feature from another dimension from left to right, allowing for the investigation of interactions between the features. In Figure 3, active constructions occur most often with VSO1, while middles are mostly SVO1.

In our Parallel Sets implementation, dimensions can be reordered via drag&drop, allowing for a flexible investigation of different types of interactions.[5] To provide a better overview, the features on each dimension axis can be sorted according to their size or alphabetically. Additionally, details of a feature correspondence can be accessed via mouse interaction techniques, see Figure 3.

To our knowledge, Parallel Sets have not yet been used in the context of linguistic research. In this paper, we show that our implementation of Parallel Sets, i.e., the dimension interaction visualization, is an extremely effective and powerful device for historical linguistic research as it fosters the identification and understanding of interactions between a variety of features contained in a multi-dimensional data set.

5 Tracking Syntactic Change

In the visualization community, the general practice is to use case studies to evaluate the usefulness of a visualization with regard to whether significant and novel insights about the data could be yielded (Carpendale, 2008; Isenberg et al., 2013). This section presents a case study which shows how HistoBankVis can be employed for the flexible investigation of syntactic change in Icelandic, focusing on the interaction between subject case and word order. Previous studies (e.g., Rögnvaldsson, 1996; Barðdal, 2011) that investigate changes with respect to these phenomena do not factor in potential interactions between the changes. By visualizing the data, we found that the two phenomena are closely interlinked.

Overview and Differences. We first looked at the diachronic development of word order in transitive sentences, i.e., sentences containing a subject (S), a finite verb (V), and a direct object (O1), vis-á-vis subject case (nominative, accusative, dative, or genitive) via the difference histograms. The compact matrix visualization showed at-a-glance that the distribution of word order and subject case changes significantly as of 1900, see Figure 1. Figure 2 provides the difference histogram distributions for subject case and word order in the periods before and after 1900. The most striking change with respect to word order is that SVO1 is increasing in the period from 1900–2008 (green bar), whereas VSO1 is decreasing concomitantly (red bar). At the same time, dative subjects increase slightly. The question is whether these two developments are linked to one another.

Dimension Interactions. The dimension interaction visualization allows for a detailed view of correlations between the features of each selected data dimension in order to investigate potential interactions. Figure 4 shows the dimension interaction for subject case and word order in the period 1900–2008 in the upper right corner. Both dimensions have been sorted according to the size of their features, with the largest feature displayed at the bottom. The shares of the subject cases on the left are mapped onto the shares they hold of the word orders on the right. The dimension interaction shows that SVO1 is the most prominent word order overall. The large majority of nominatives occur together with SVO1, while the share of SVO1 of the dative subjects is considerably smaller.

The patterns observed in the period from 1900 to 2008 differ from the interactions in an earlier time period (1150–1350), compare the top right with the top left of Figure 4. In contrast to the period post-1900, the shares of SVO1 and VSO1 are about equal for nominative subjects. Additionally, dative subjects occur most frequently with VSO1. Thus, word order develops differently with respect to subject case over time. The difference histograms in Figure 2 indicated that subjects are increasingly realized preverbally, the dimension interaction shows

[5] This is based on Jason Davies' work: https://www.jasondavies.com/parallel-sets/.

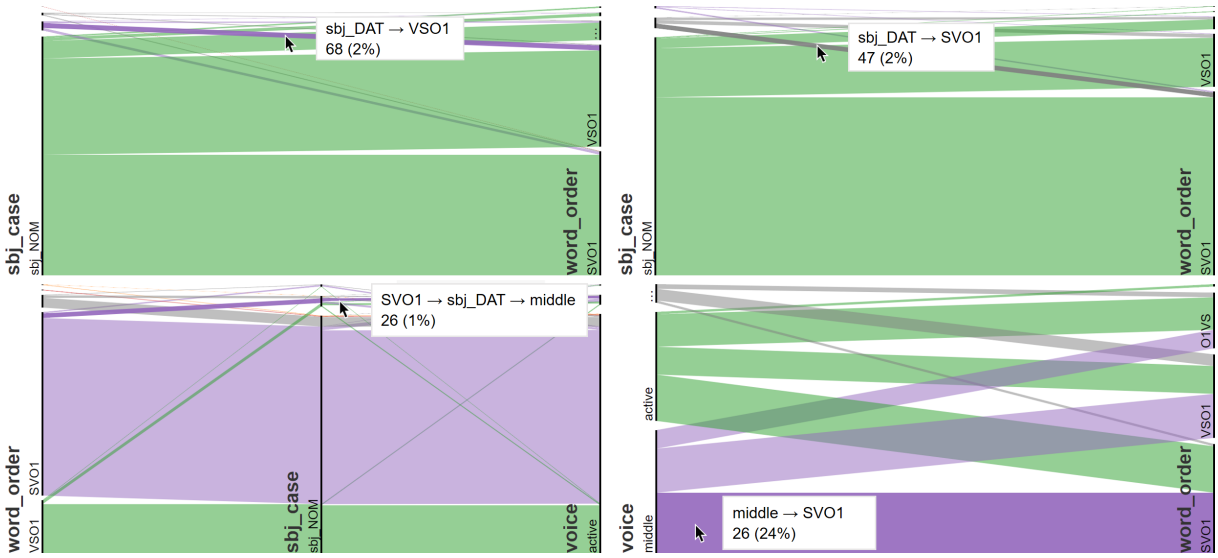

Figure 4: Top: Dimension interactions for subject case and word order from 1150–1350 (left) and 1900–2008 (right). Bottom: Dimension interactions from 1900–2008 for word order, subject case and voice (left) and voice and word order in dative subject sentences (right).

that dative subjects lag behind this development.

It is well-known that voice influences the occurrence of dative subjects in Icelandic (e.g., Zaenen et al., 1985; Sigurðsson, 1989). However, whether there is an actual correlation between voice, subject case and word order has not yet been investigated. This can be accomplished easily with the help of HistoBankVis since we can simply integrate the dimension voice for an analysis of the dimension interactions between subject case, word order and voice, cf. Figure 4-bottom-left for the period 1900–2008. The dimension interactions show that SVO1 occurs most often with nominative subjects in active constructions. With dative subjects though, SVO1 order mainly occurs in middle constructions. A separate analysis of the interaction between voice and word order for dative subjects allows for a more nuanced look at interactions, see Figure 4-bottom-right (1900–2008). Dative subjects occur most frequently with middle voice and SVO1 is the most prominent word order for both, active and middle constructions. However, in earlier stages of the language, word order and voice pattern differently, see Figure 3 for the dimension interaction from 1750 to 1899. First, dative subjects occurred most often in active clauses and not with middles. Moreover, SVO1 is already the dominant word order for middle forms, but not for the active constructions in which VSO1 prevails.

Concluding, these findings indicate that the increasing realization of dative subjects in before the verb correlates with an increasing use of dative subjects together with middle voice. With the aid of HistoBankVis, in particular the dimension interactions, we were able to easily identify a previously unknown link between word order, dative subjects and voice in a matter of minutes.

6 Conclusion

HistoBankVis serves as an efficient and powerful tool for historical linguistic investigations as it provides multiple perspectives of the data at different levels of detail on demand, fostering an iterative process of hypothesis testing and generation. In particular, we introduced the use of Parallel Sets to provide an interactive visualization of complex interactions across different dimensions of data. To our knowledge, this is the first use of Parallel Sets in a linguistic visualization.

We illustrated the flexibility and strength of HistoBankVis on the basis of a concrete case study which investigated changing linguistic features in Icelandic. We demonstrated that our system can yield new insights and we have shown that the analysis of dimension interactions as provided by the extended system represents an effective new means for historical linguistic research.

Acknowledgements

We thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 251654672 – TRR 161 (Projects A03 and D02) for their financial support.

References

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

David Bamman and Gregory Cane. 2006. The design and use of a Latin dependency treebank. In Jan Hajič and Joakim Nivre, editors, *Proceedings of the Fifth International Treebanks and Linguistic Theories*. pages 67–78.

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

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

S. Bird, E. Klein, and E. Loper. 2009. *Natural Language Processing with Python*. O'Reilly.

Sheelagh Carpendale. 2008. Evaluating information visualizations. In *Information Visualization - Human-Centered Issues and Perspectives*, pages 19–45. https://doi.org/10.1007/978-3-540-70956-5_2.

Tom Christiansen, Jon Orwant, Larry Wall, and Brian Foy. 2012. *Programming Perl*. O'Reilly, 4 edition.

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

Jan Hajič. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In E. Hajičová, editor, *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*, Karolinum, Charles University Press, Prague, Czech Republic, pages 106–132.

Dag T. T. Haug and Marius L. Jøhndal. 2008. 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.

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*, Cambridge University Press, Cambridge, pages 36–53.

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

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

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

Tobias Isenberg, Petra Isenberg, Jian Chen, Michael Sedlmair, and Torsten Möller. 2013. A systematic review on the practice of evaluating visualization. *IEEE Trans. Vis. Comput. Graph.* 19(12):2818–2827. https://doi.org/10.1109/TVCG.2013.126.

R. Kosara, F. Bendix, and H. Hauser. 2006. Parallel Sets: interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics* 12(4):558–568. https://doi.org/10.1109/TVCG.2006.76.

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

Anthony Kroch, Beatrice Santorini, and Ariel Diertani. 2010. The Penn-Helsinki Corpus of Modern British English (PPCMBE). First edition.

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

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*. Association for Computational Linguistics, pages 44–48.

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

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

Halldór Á. Sigurðsson. 1989. *Verbal Syntax and Case in Icelandic. In a Comparative GB Approach*. Institute of Linguistics.

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

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

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

Annie Zaenen, Joan Maling, and Höskuldur Thráinsson. 1985. Case and grammatical functions: the Icelandic passive. *Natural Language and Linguistic Theory* 3(4):441–483.