Diachronic Visualization of Oblique Subjects in Icelandic

Christin Schätzle*   Dominik Sacha†   Miriam Butt‡

University of Konstanz

ABSTRACT

This paper presents a corpus linguistic study on oblique subjects in Icelandic. Our results show that dative subjects are not necessarily a Proto Indo-European inheritance. Additionally, a change in the semantics of verb classes inducing oblique subjects is revealed. We furthermore suggest a visual analysis that sheds more light on our data.

Keywords: Oblique subjects, word order, case, semantics, Icelandic, IcePaHC, language change, corpus, visualization

1 INTRODUCTION

Oblique Subjects are a well-known and common phenomenon in a multitude of Indo-European languages. This paper presents a diachronic corpus study of dative subject constructions in Icelandic based on the recently released annotated Icelandic parsed historical corpus (IcePaHC) ([10]). Our results support [2]'s finding that dative subjects are stable throughout the Icelandic language history and still productive in Modern Icelandic. However, the corpus yielded results contra [3], because verbs inducing dative subjects are not confined to inactive verb classes. Hence, they are not coercively inherited in a Proto Indo-European language stage. Nevertheless, a marginal shift from happenstance and experience-based predicates towards only experience-based predicates inducing dative subjects is visible in our data (partially according to [1]). Furthermore, we present a visualization of our data which is necessary to gain more linguistic insights into our multifactorial diachronic data set.

2 OBLIQUE SUBJECTS IN ICELANDIC

Across languages, grammatical relations like subject and object tend to marked by either word order, agreement or case (e.g., see [4]). In this paper, we focus on the occurrence of oblique subjects, in particular Icelandic dative subjects. Icelandic is a Germanic language with a rich case marking system and fairly fixed word order.

Dative subjects are common in Old-Norse Icelandic as well as Modern Icelandic (e.g., see [2]). Example 1 shows a dative subject construction in Modern Icelandic.

(1) Mér   er kalt.
     me.DAT is cold
     'I feel cold.'

Barðal and Eythórsson argue that dative subjects are a common Proto Indo-European inheritance ([2]). This approach is against traditional linguistic theory, in which it is assumed that oblique subjects have developed from objects to subjects (e.g., see [6]).

---

*e-mail: christin.schaetzle@uni-konstanz.de
†e-mail: dominik.sacha@uni-konstanz.de
‡e-mail: miriam.butt@uni-konstanz.de

They conclude that Proto Indo-European must have been a stative-active language, in which a subset of syntactic subjects of stative intransitive predicates was case-marked in the oblique.

In other recent work, Barðdal and colleagues adduce more evidence for their approach by pointing towards the pervasiveness and productivity of dative subjects in early history ([3]). In particular, they provide a set of lexical verb classes instantiating dative subject constructions across several Indo-European languages, including Old Norse-Icelandic. Particularly interesting is their use of statistical methods (Principal Component Analysis; PCA) in combination with visualization techniques to demonstrate the behavior of semantic verb classes across languages.

In another paper, Barðdal focusses on the predicate semantics of oblique subjects in Icelandic ([1]). She presents a comparative corpus study showing that the semantics of dative subjects changed from denoting happenstance and experience-based events equally in Old Norse-Icelandic to only denoting experienced-based events in Modern Icelandic. This change in the semantic scope of dative subjects is seen as the motivation for the onset of Dative Substitution (cf. [8]) in the latter part of the 19th century and may give an explanation for the productivity of dative subjects, despite a decreasing type frequency ([1]).

3 CORPUS STUDY

The IcePaHC is ideal for the study because it is syntactically annotated in the annotation scheme of the Penn Treebank ([9]) and dative subjects are distinctively coded in the corpus. The corpus consists of 60 texts in 4 different genres comprising about 2 million words and dating from the 12th to the 21st century. The different genres are not representative across the centuries.

Initially, we posed a query to the corpus via the CorpusSearch tool,[1] a specialized search tool for corpora in the Penn-Treebank format. Our search resulted in roughly 5600 sentences containing dative subjects (from a total of 73014 sentences) across nine centuries. Via a script, 4266 main verbs inducing a dative subject were extracted and manually annotated with the corresponding semantic verb class according to [2, 3]. Additionally, verbs denoting action and motion have been found and considered in the analysis. The data has been divided into time periods according to [7].

The results of the corpus study show that dative subjects are a common phenomenon throughout Icelandic language history. According to the literature, dative subjects mainly denote emotions in the corpus, but also motion and action verbs with dative subjects are a regular part of the language (contra [2]). Interestingly, the diachronic development of most verb classes exhibits a change after 1750 CE. Therefore, we checked whether the observed distribution of each verb class instantiating a dative subject in each time period differed significantly from the overall distribution via the χ^2-test. The distribution of each of the 16 verb classes over the total time span was taken as the expected factor, while the occurrences of each verb class in each time section was taken as the observed factor. We were only able to show that the distribution of dative subjects in the last time stage differs very significantly (1900–present: χ^2=33.04, p<0.005) from the overall distribution.

---

[1]http://corpussearch.sourceforge.net/CS.html

Additionally, we conducted a second χ^2-test, in which we categorized the verb classes, except modality and evidentiality, into happenstance and experience-based predicates according to [1]. The results support [1] only partly because experienced-based predicates already clearly outnumbered happenstance predicates in Old-Norse Icelandic and the narrowing to experience-based predicates is marginal. Nevertheless, the change after 1750 CE spotted in the previous alignment remains visible and is tested likewise to the first investigation. Table 1 shows the significantly deviating distributions indicated with "*".

Time	Happenstance %	Experience-based %	χ^2
upto 1350	14.9	85.1	0.27
upto 1550	19.15	80.85	0.65
upto 1750	23.58	76.42	3.9*
upto 1900	16.59	83.41	3.9*
upto present	12.17	87.83	5.46*
expected	16.84	83.16	

Table 1: Distribution of Verb Classes over Time

4 DIACHRONIC VISUALIZATION OF MULTIFACTORIAL DATA

The statistical methods used in our corpus study only show a deviance from the overall picture and merely provide a very rough picture of the data. The visualization of our data set on the other hand allows for an exploratory analysis as analysts are able to drill down and investigate each text and sentence. Unexpected patterns and outliers may be spotted by granting an interactive explorative access to the data. Feature patterns that are lost in the plain statistical analysis due to the requirement of fixed parameters (e.g. time periods) become visible and additional properties contained in the data (e.g. genre) are visualized.

The present visualization follows an overview first — details on demand approach. The verb classes can be aggregated to higher level verb categories (see Figure 1, top) that consist of detailed verb classes (see Figure 1, bottom). Each text is represented by a glyph that visualizes the text length, the dative sentence occurrence (sentence bar), the time (timeline on the right & vertical layout) and the occurrence of sentence type as colored circular glyphs that are filled from the outside or the inside indicating whether the verb class/category appears more or less often than expected based on the text length. Furthermore, the genre is visualized by horizontal layout of the texts. The visualization offers multiple interaction techniques like zoom & pan, tooltips for several visual objects (sentence bar, circular glyphs or texts) and expand/fold operations for each text which allows to analyze the verb categories and then to drill down into the detailed verb classes.
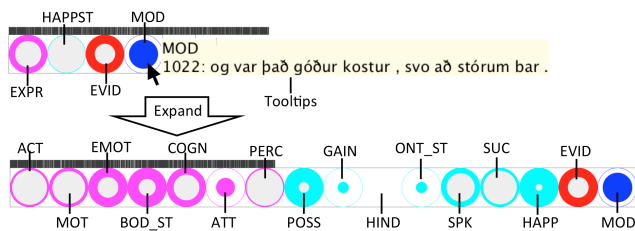

Figure 1: Visualization of dative subject occurrences in a text within IcePaHC. The aggregated verb categories (top) can interactively be expanded (bottom) and investigated by details on demand interactions. The complete visualization holds such glyphs for each text and illustrates additional properties by position.

This exploratory kind of analysis enabled an incremental develop-ment of the visualization based on continuous expert feedback. Our visualization is based on [5] which visually analyzed V1 (verb-first) word order in Icelandic and has been adjusted and extended for the exploratory analysis of dative subjects. Compared to [5], we had to adjust the occurrences of word types within the text glyphs and reduced the data matrix to a data row. A problem we had to cope with was the large amount of verb classes that had to be visualized. As a solution we decided to aggregate the verb classes to higher level categories given by linguistic literature that can be expanded on demand.

The visualization shows that the change after 1750 is in fact a genre effect. Of linguistic interest is that happenstance predicates become very rare at the end of the 19th century. Furthermore, the verb classes which are responsible for the development of the categories can be identified at a glance, allowing for future in-depth linguistic analysis.

5 CONCLUSION

This paper shows the relevance of visualization techniques in the analysis of a large diachronic data set. The data argues against the Proto Indo-European inheritance of dative subject constructions, because not only stative but also active predicates have been found in the whole corpus. A reordering of the data shows a marginal narrowing of the semantic scope of dative subjects. Our visual analysis suggests a semantic shift in the 19th century according to [1]. In the future, the visualization could be extended to illustrate more factors (e.g. transitivity, active/passive) as additional rows in the text glyph to further the linguistic understanding of the data.

REFERENCES

[1] J. Barðdal. The rise of dative substitution in the history of icelandic: A diachronic construction grammar account. *Lingua*, 121(1):60–79, 2011.

[2] J. Barðdal and T. Eythórsson. The origin of the oblique subject construction: An indo-european comparison. In V. Bubenik, J. Hewson, and S. Rose, editors, *Grammatical Change in Indo-European Languages*, pages 179–193. John Benjamins, Amsterdam, 2009.

[3] J. Barðdal, T. Smitherman, V. Bjarnadóttir, S. Danesi, G. B. Jenset, and B. McGillivray. Reconstructing constructional semantics: The dative subject construction in old norse-icelandic, latin, ancient greek, old russian and old lithuanian. *Studies in Language*, 36(3):511–547, 2012.

[4] M. Butt. *Theories of Case*. Cambridge University Press, Cambridge, 2006.

[5] M. Butt, T. Bögel, K. Kotcheva, C. Schätzle, C. Rohrdantz, D. Sacha, N. Dehe, and D. Keim. V1 in icelandic: A multifactorial 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.

[6] M. Haspelmath. Non-canonical marking of core arguments in european languages. In A. Y. Aikhenvald, R. Dixon, and M. Onishi, editors, *Non-Canonical Marking of Subjects and Objects*, pages 53–83. John Benjamins, Amsterdam, 2001.

[7] E. Haugen. *Die skandinavischen Sprachen*. Hamburg: Buske, 1984.

[8] J. Jónsson. Not so quirky: on subject case in icelandic. In E. Brandner and H. Zinsmeister, editors, *New Perspectives on Case and Case Theory*, pages 1270–164. CSLI Publications, Stanford, 2003.

[9] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[10] J. C. Wallenberg, A. K. Ingason, E. F. Sigurðsson, and E. Rögnvaldsson. Icelandic Parced Historical Corpus (IcePaHC). Version 0.9. http://www.linguist.is/icelandic_treebank, 2011.