LREC 2016 Workshop

VisLR II:
Visualization as Added Value in the Development, Use and Evaluation of Language Resources

PROCEEDINGS

Edited by

Annette Hautli-Janisz, Verena Lyding

23 May 2016

Proceedings of the LREC 2016 Workshop
"VisLR II: Visualization as Added Value in the Development, Use and Evaluation of Language Resources"

23 May 2016 – Portorož, Slovenia

Edited by Annette Hautli-Janisz, Verena Lyding

www.eurac.edu/vislr_2016

Organising Committee

- Mennatallah El-Assady, University of Konstanz, Germany
- Annette Hautli-Janisz*, University of Konstanz, Germany
- Verena Lyding*, EURAC research Bozen/Bolzano, Italy

*: Main editors and chairs of the Organising Committee

Programme Committee

- Noah Bubenhofer, University of Zürich, Switzerland
- Miriam Butt, University of Konstanz, Germany
- Jason Chuang, Independent Researcher, USA
- Christopher Collins, University of Ontario Institute of Technology, Canada
- Chris Culy, Independent Consultant, USA
- Gerhard Heyer, University of Leipzig, Germany
- Kris Heylen, University of Leuven, Belgium
- Daniel Keim, University of Konstanz, Germany
- Steffen Koch, University of Stuttgart, Germany
- Victoria Rosén, University of Bergen, Norway

Preface

This workshop aims at providing a follow-up forum to the successful first VisLR workshop at LREC 2014, which addresses visualization designers and users from computational and linguistic domains likewise. Since the last workshop, the concern with visualizing language data has further increased, as the recurrence of specialized symposia in the linguistic and NLP contexts show (cf. e.g. ACL workshop 2014, AVML 2014, Herrenhäuser Symposium 2014, QueryVis 2015). Moreover, the application of visualization techniques to various use cases is becoming ever more agile.

As a specialized subfield of information visualization, the visualization of language continues to face particular challenges: Language data is complex, only partly structured and, as with todays language resources, comes in large quantities. Moreover, due to the variety of data types, from textual data to spoken or signed language data, the challenges for visualization are necessarily varied. The overall challenge lies in breaking down the multidimensionality into intuitive visual features that enable an at-a-glance overview of the data. The second edition of the workshop therefore aims at advancing the field of linguistic visualization by particularly focusing on more advanced visualization techniques that represent the complexity of language and that contribute to resolving them.

Annette Hautli-Janisz, Verena Lyding May 2016

Programme

09.00 – 09.05 Introduction

09.05 – 09.35 Paul Meurer, Victoria Rosén, Koenraad De Smedt
Interactive Visualizations in the INESS Treebanking Infrastructure

09.35 – 10.05 Christin Schätzle, Dominik Sacha
Visualizing Language Change: Dative Subjects in Icelandic

10.05 – 10.35 Annette Hautli-Janisz
See the Forest AND the Trees: Visual Verb Class Identification in Urdu/Hindi VerbNet

10.35 – 11.00 Coffee break

11.00 – 11.30 Thomas Wielfaert, Kris Heylen, Dirk Speelman, Dirk Geeraerts
Visual Analytics for Distributional Semantic Model Comparisons

11.30 – 12.00 Erik Tjong Kim Sang
Visualizing Literary Data

12.00 – 12.30 Andrew Caines, Christian Bentz, Dimitrios Alikaniotis, Fridah Katushemererwe, Paula Buttery
The Glottolog Data Explorer: Mapping the World's Languages

12.30 – 12.40 Closing

Table of Contents

Interactive Visualizations in the INESS Treebanking Infrastructure

Paul Meurer[1], **Victoria Rosén**[2], **Koenraad De Smedt**[2]
[1]Uni Research Computing, [2]University of Bergen
Bergen, Norway
paul.meurer@uni.no, victoria@uib.no, desmedt@uib.no

Abstract

The visualization of syntactic analyses may be challenging due to the number of readings, the size and detail of the structures, and the interrelations between levels of linguistic description. We present a range of interactive visualization techniques applied to complex syntactic analyses in INESS, an online infrastructure for parsing and the annotation and exploration of syntactically annotated corpora (treebanks). Although INESS caters to many syntactic formalisms, we focus on LFG, which allows for multiple levels of syntactic structure, in particular c-structures and f-structures. Interactive dynamic renderings of the relations between components of these structures are presented, with options on the level of detail to be displayed. Furthermore, the disambiguation of sentences with multiple possible parses needs techniques for visualizing the differences between readings. For this purpose, we present and discuss packed representations, the interactive visualization of discriminants, and the previewing of disambiguation choices. The interactive querying of treebanks benefits from appropriate ways of displaying search results. We present the highlighting of matching items in matching sentences. We also present tabular overviews with frequencies of obtained variable values, as well as the inspection of matching structures without having to navigate away from the overview.

Keywords: treebanks, syntactic structures, visualization

1. Introduction

Syntactic analyses have long been visualized as tree structures and other graphs, both on paper and on computer screens. Whereas such visualizations have often been static pictures, there are also opportunities for dynamic, interactive visualizations of syntactic structures.

In this paper, we describe various innovative features of visualization in the INESS treebanking infrastructure which is part of the CLARINO Bergen Center.[1] This infrastructure offers access to treebanks of different types, such as LFG (Lexical-Functional Grammar), HPSG (Head-driven Phrase Structure Grammar), dependency grammar and phrase structure grammar (constituency). It also provides online LFG parsing and disambiguation for several languages. INESS is accessible through federated single sign-on authentication as promoted by CLARIN.

Although various aspects of the infrastructure have been described in other publications (Rosén et al., 2009; Rosén et al., 2012), neither its approach to visualization nor its recently updated visualization components have been described. We therefore focus now on the visual inspection of complex syntactic analyses, including interactive visualizations that combine multiple dimensions and represent the dependencies between them. Although appropriate visualizations are provided for all types of treebanks, we will in the present paper concentrate on LFG analyses, which are quite complex. We also pay attention to visual aspects of the interface for annotating and searching treebanks.

Section 2 introduces the visualization of LFG structures in the infrastructure. Section 3 presents visual techniques for disambiguation. Section 4 explains the options for the presentation of search results. In Section 5 we present the implementation, in Section 6 we compare with other systems, and in Section 7 we conclude.

[1]http://clarino.uib.no/iness/

2. Interactive Visualization of LFG Structures

Syntactic data, especially those resulting from deep parsing, are among the most complex types of linguistic data and heavily rely on user-friendly visualization. LFG analyses have (at least) two separate but interrelated levels: a c-structure, which is a tree structure representing constituency imposed on a left-to-right string, and an f-structure, which is a feature-value matrix representing functional relations and features. C- and f-structure are related by projections as defined by the grammar; each c-structure node projects some (subsidiary) f-structure. For instance, an NP in the c-structure may project the value of the OBJ (object) in the f-structure. Such projection relations are important and should be appropriately visualized.

The Xerox Linguistic Environment (XLE) is a platform for developing LFG grammars (Maxwell and Kaplan, 1993; King et al., 2004). XLE offers an efficient parser and generator for LFG grammars, and it interfaces with finite-state preprocessing modules for tokenization and morphological analysis. Whereas XLE offers a visual display based on X11 and Tcl/Tk, INESS uses only the XLE parser and has developed its own visualizations. XLE-Web is an online interface to XLE where users can parse sentences in a web browser, with a number of grammars for different languages including English, French, German, and Norwegian.[2] The LFG Parsebanker (Rosén et al., 2009) uses some of the same visualizations as XLE-Web but in an environment for treebanking. Here we present mainly the aspects of visualization that are novel in INESS; a more specific comparison to XLE is provided in Section 6.

Figure 1 shows the c- and f-structures for the sentence *The boy likes the girl*. For reasons of space we have chosen a simple sentence of only five words to demonstrate that the

[2]http://clarino.uib.no/iness/xle-web

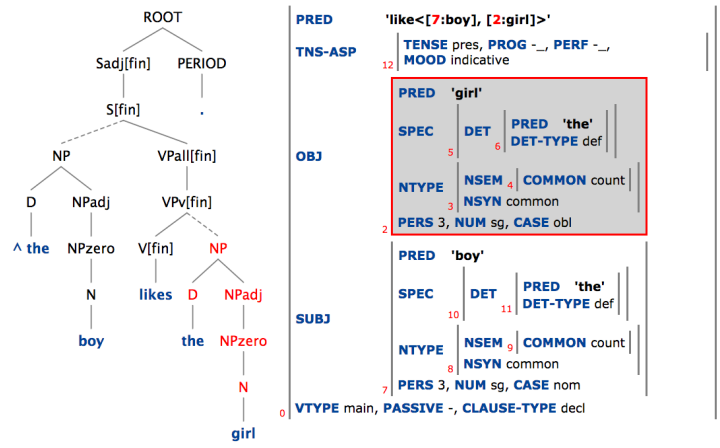

Figure 1: Mouseover visualizing the projection from the NP node in the c-structure (left) to the value of the OBJ in the f-structure (right)

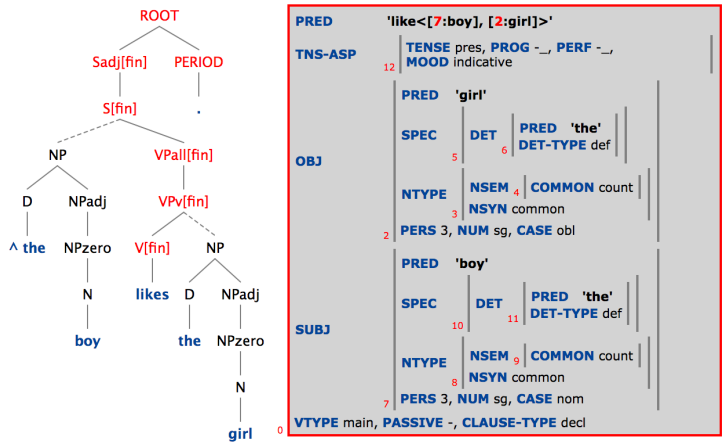

Figure 2: Mouseover visualizing the projection from the V[fin] (finite verb) node in the c-structure (left) to the f-structure of the sentence (right)

c- and f-structures are rich representations, encoding many different types of linguistic information. For longer and more complex sentences the amount of information will of course increase accordingly, and the sheer size of the representations makes good visualization techniques essential. In the c-structure, black is the basic color but terminal nodes (words) are in blue. In the f-structure, the feature names are in blue and their atomic values in black, while indices are displayed in red.

All c-structure nodes that project to the same (subsidiary) f-structure constitute a functional domain. These functional domains partition the c-structure. These partitions are indicated by the use of solid and dotted lines in the branches of the tree. Nodes connected by solid lines project to the same functional domain, whereas dotted lines connect parts of the tree which project to different functional domains.

The connection between the two representations may be seen by mousing over nodes in the c-structure. In the example in Figure 1, mousing over the NP node highlights the corresponding subsidiary f-structure, thus showing that the NP *the girl* is the OBJ of the sentence. On mouseover, all nodes that belong to the same functional domain (i.e. all nodes that project the same f-structure) are highlighted in red; holding the mouse over any of the nodes NP, D, NPadj, NPzero or N produces the same result. Figure 2 shows that the finite verb projects the main f-structure of the sentence.

One can also use mouseover from the f-structure to the c-structure. In this case, mousing over the index number of an f-structure results in all c-structure nodes projecting that f-structure being highlighted. Whether going from c- to f-structure or vice versa, the correspondence between highlighted parts of the two representations is entirely dependent on and derived from the projections defined by the particular LFG grammar used for each treebank.

It is possible to collapse or expand certain parts of the c-structure depending on what the user is interested in viewing. Clicking on a preterminal node displays the sublexical structure of the terminal node, thus making visible

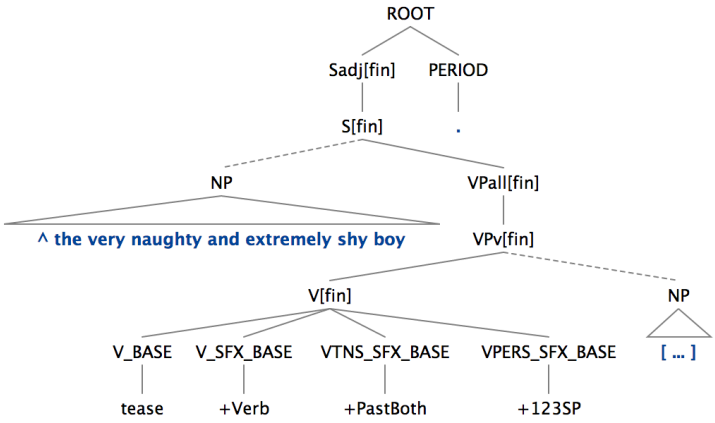

Figure 3: Expansion into sublexical nodes and collapsing of nodes

the features encoded by the morphological analyzer. Clicking again collapses the sublexical tree. Clicking on any c-structure node that is not a terminal or preterminal collapses the dominated subtree into a triangle over the entire substring. If the substring is too long, the middle of the substring may be elided. Clicking once again on the same node replaces the substring by ellipsis dots surrounded by square brackets [...]. A third click will return the full subtree. Figure 3 shows these features for the sentence *The very naughty and extremely shy boy teased the very pretty and extremely popular girl*. These visualizations can be useful for viewing very large c-structures, since parts of the c-structure that are not relevant to what the user wants to examine may be abbreviated, while other parts the user does wish to view may be expanded.

F-structures can also be made more compact. The option "Show PREDs only" suppresses attributes which do not contain a path to a PRED (predicate) value; as a result, only the functional backbone of the f-structure is shown without attributes. In this view, the f-structure resembles a dependency structure. In addition, "Suppress CHECK" is an option which suppresses auxiliary attributes which are internally used in the grammar for wellformedness checks; this option is on by default.

3. Visualizing the Effects of Discriminants on Packed Structures

Because of lexical and syntactic ambiguity, deep parsing often produces multiple analyses. When there are many possible analyses, it is difficult or practically impossible to find the intended analysis by sequentially inspecting all the visual structures. A more compact visualization is a *packed* representation in which all analyses are viewed together in one graph, with choice points on nodes. Whereas the XLE interface offers such a representation for f-structures, called an *f-structure chart* (King et al., 2004), INESS offers packed representations for c-structures as well as f-structures. Figure 4 shows the packed representations of the sentence *The boys saw the girls*, with choice points shown in green. In this figure the index *a1* is used in both structures to indicate the analysis of the verb as the present tense of *saw*, while the index *a2* indicates that the verb is the past tense of the verb *see*.

When there are multiple ambiguities in a sentence, the packed structures may become so large or complex that they are difficult to read, and they are therefore not by themselves sufficient for disambiguation. We have therefore implemented a system of discriminants (Carter, 1997; Oepen et al., 2004), which are simple properties of analyses. INESS computes discriminants and presents them to annotators of treebanks or users of XLE-Web, who can choose or reject discriminants in order to disambiguate a sentence. Usually, a small number of discriminants is sufficient to select one of potentially many possible analyses. In INESS, discriminants for LFG are computed and grouped by kind (Rosén et al., 2007), as illustrated in Figure 5.

Choosing a discriminant results in all analyses not compatible with that discriminant being removed from the parse forest. It can be useful to see the effect on the packed representation of selecting a certain discriminant, especially because there are often interdependencies between discriminants. Figure 5 illustrates the effect of choosing the discriminant *the || girl with the binoculars*.[3] The part of the c-structure that is not compatible with this discriminant is grayed out.

4. Visualization of Search Results

Exploring treebanks may benefit from not only returning sentences that match a query, but also clearly visualizing which parts of the sentence or analysis match the search expression. Overviews of search results from INESS-Search (Meurer, 2012) can be displayed in different ways: as a list of all sentences matching the query, or in tabular form. In the former mode, matching sentences are listed; clicking on a sentence displays its structure, or, in the case of LFG analyses, both c- and f-structure. In the latter mode, results are given in tabular form, where they are aggregated and sorted according to selected node variables and features. There is one table column for every selected feature, and each row

[3]The double bar in the string is a shorthand for the bracketing *[the][girl with the binoculars]*.

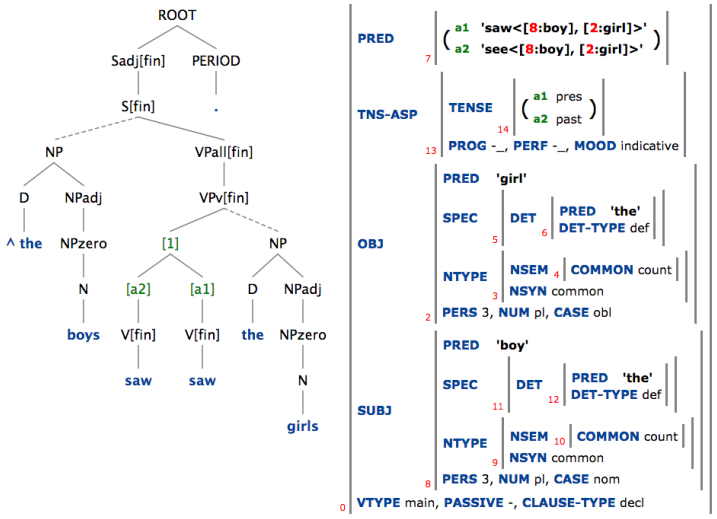

Figure 4: Packed c- and f-structures in INESS

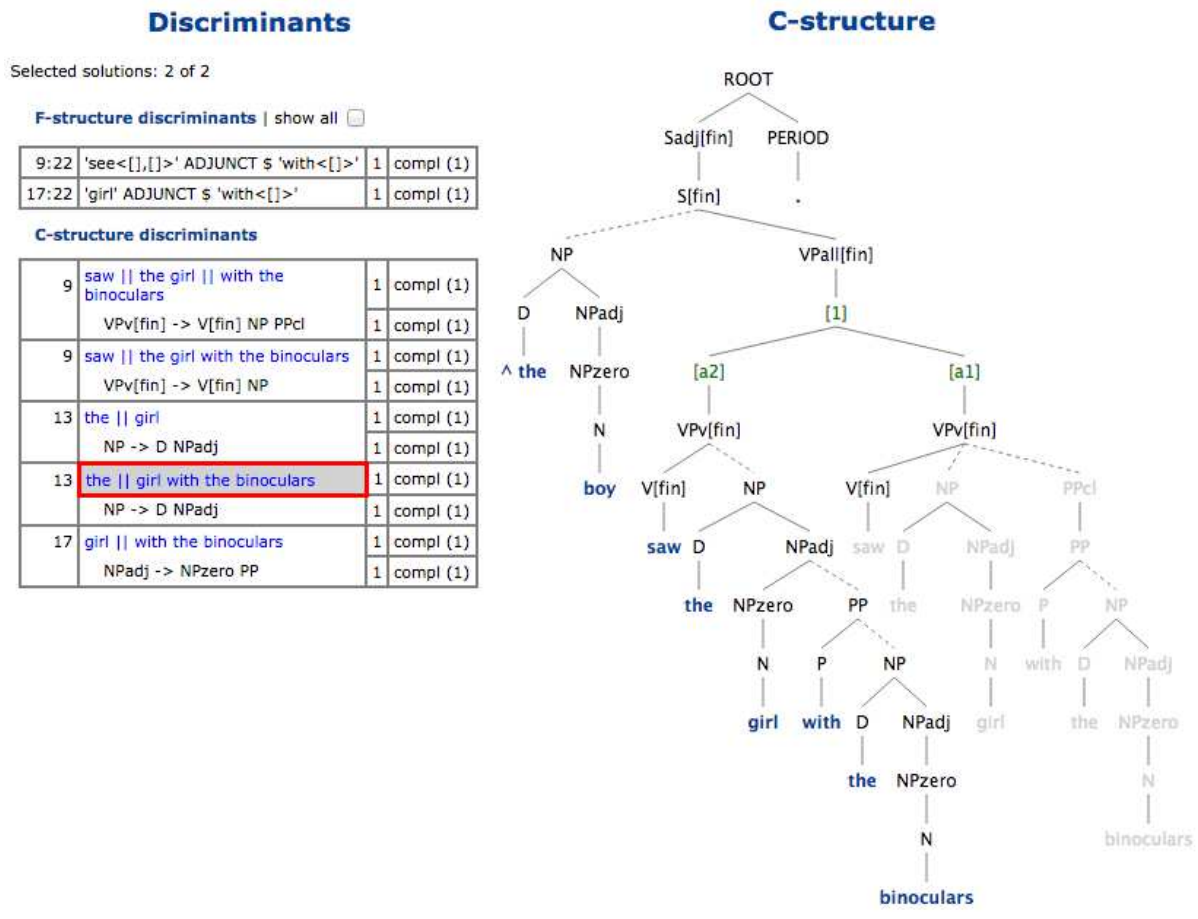

Figure 5: Previewing discriminant effect by graying out part of the packed c-structure

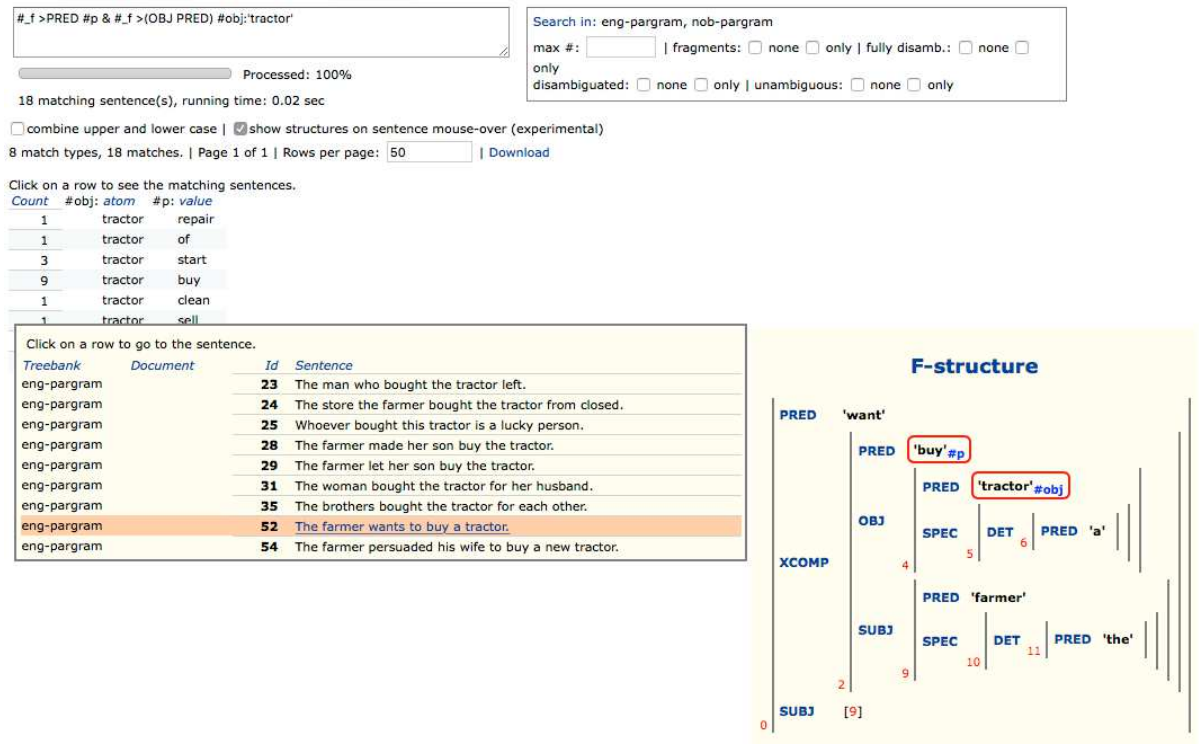

Figure 6: F-structure display on mouseover with highlighted matching nodes

of the table represents a distinct combination of values for the selected variables and features.

Clicking on a row displays all matching sentences where those feature values are assumed. This is illustrated in Figure 6: The query $\#\_f$ >PRED $\#p$ & $\#\_f$ >(OBJ PRED) $\#obj$:'tractor' matches sentences where 'tractor' is the object (OBJ, marked with variable $\#obj$) of a verb (marked with $\#p$). The fourth row of the table shows that there are nine sentences in which $\#obj$ has the value *tractor* and $\#p$ has the value *buy*. Variables containing an underscore ($\#\_f$) are not shown in the lists of results.

As shown in Figure 6, clicking on this fourth table row has opened a display window listing those nine sentences. By clicking on one of the sentences, one can open the structure display window for that sentence. It is however important to be able to easily see the matching nodes and structures in the tabular view at a glance, without having to go to a different web page and navigate back to the tabular view afterwards. Therefore, mousing over a sentence, as highlighted in orange in Figure 6, pops up a window with a preview of only the essential parts of the c- and/or f-structure. In our example, this is the f-structure displayed on the right in Figure 6. It includes all matching nodes, while parts of the structure not containing a matching node are collapsed, but can be expanded if desired. Each matching node is highlighted in red and is indexed with its corresponding variable in the search expression.

5. Implementation

The treebanking system is fully online and can be used in any modern web browser. Both the visualization code and the remainder of the treebanking framework are written in Common Lisp. Web pages are generated as XML documents that are converted into CSS-styled HTML on the server using XSLT transformations. This architecture maintains a clean separation between the representation of information to be rendered as web pages on the one hand (as XML) and the specific visual aspects of the way this information is displayed on the other hand (as CSS-styled HTML). For the interactive features, Javascript is used. The tree visualizations are implemented in SVG (Scalable Vector Graphics), whereas f-structures are coded as tables in plain HTML. The SVG code is currently generated directly from Common Lisp. Better flexibility and reuse might be achieved by generating the SVG in the browser by Javascript, or by using some features from libraries such as D3.

6. Comparison with Related Systems

Although INESS is inspired by the graphic environment in XLE, the design of the two systems is different. INESS caters to treebank constructors as well as to end users wishing to consult treebanks. XLE is mainly targeted at an audience of grammar developers, not treebank users. XLE visualizations therefore show details which are relevant for grammar development but not relevant for linguists only wishing to see parsing results. For instance, XLE allows the detailed inspection of valid syntactic structures as

well as structures that violate coherence or completeness constraints. In addition, XLE shows, for instance, certain negated and completeness constraints which are relevant for grammar debugging. XLE is an application running under X11, whereas INESS is entirely used through a browser.

An example of a packed f-structure representation in XLE is shown in Figure 7. This sentence, *They can fish*, has three analyses. One analysis has *can* as a modal auxiliary that takes an XCOMP. The other two analyses involve *can* as a main verb taking an OBJ; in addition, *fish* can either be a singular noun, or the plural of a count noun. Even with only three analyses, this visualization is difficult to read. The INESS packed f-structure for the same sentence in 8 is easier to read, partly because some less relevant information and superflous brackets are left out, and partly because of the use of colors and boldface and the placement of the indices.

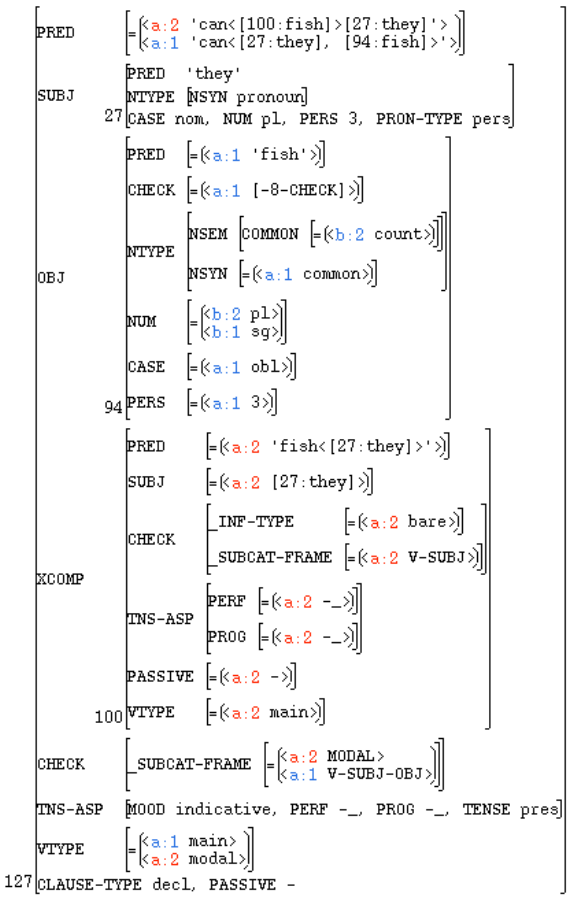

Figure 7: Packed f-structure for *They can fish* in XLE

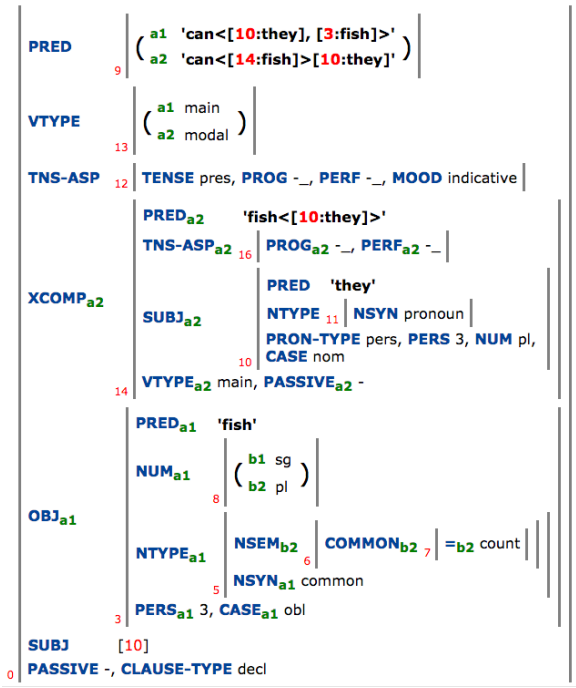

Figure 8: Packed f-structure for *They can fish* in INESS

Tündra (Martens, 2013) is an extensive application for browsing, searching and visualizing the contents of treebanks. Like INESS, it is entirely offered through a web browser and users are authenticated through federated single sign-on. TüNDRA supports both constituency and dependency treebanks, as well as mixed and hybrid treebank types, but has only limited support for directed graphs as required for LFG, and does not support discriminant disambiguation. In addition to TIGER-style (square angled) branches and more conventional phrase structures, it also offers colored labeled bracketing. For dependency structures, there is a choice between a visualization with arcs or as a tree (TrED style). TüNDRA has appropriate highlighting in red for search results.

ANNIS (Zeldes et al., 2009) offers search and visualization of multi-level corpora (including treebanks) and uses a query language similar to TIGERSearch. It is available both as a standalone application and online through a browser. Its visualization modes are rich and varied, depending on the information to be rendered. Visualizations of syntactic structures include dependency arcs, hierarchical or ordered dependencies, and Tiger-style trees with right-angled branches (also catering for right-to-left writing mode). Visualizations of token and span annotations include KWIC with interlinear glossing, grids with layered spans for information structure, colors and underscoring for coreference, and Rhetorical Structure Theory (RST) representations.

GrETEL (Augustinus et al., 2012; Augustinus et al., 2013) provides a web-based query interface for some Dutch language treebanks. It has an interface in which users can build a query based on an example sentence or phrase which is interactively parsed. From the user's options about the way the words in the example are fixed or may vary, an Xpath query is constructed. GrETEL presently only supports constituency treebanks which may have edge labels. Trees are visualized with layered nodes displaying edge label, part of speech, lemma and word.

7. Concluding Remarks

We have presented an overview of various visualizations which are useful, or even indispensable, in the INESS treebanking infrastructure. Of all treebanking systems, only INESS offers sufficient support for LFG representations. Although most treebanking systems handle constituency treebanks, and therefore can visualize c-structures, only INESS can visualize f-structures and the links between c-structures and f-structures. Moreover, INESS handles search and disambiguation of LFG structures and offers appropriate visual interfaces for these tasks.

Since syntactic structures are quite complex, in particular in LFG, smart techniques must be used to overcome the problem of the sheer size of the structures in relation to computer screens, even quite big ones. One of our guiding principles has been that visualizations of syntactic structures should be able to present different levels of detail, dependent on the context (e.g. annotation, search, or preview). Also, dependent on user choices, there should be ways to collapse or expand information, and 'hidden' information should be accessible upon request through clicking or mouseover movements.

The framework as described is fully functional and available, but it remains under active development, because more visualization features are planned as feedback from users is collected. A possible addition would consist of the ability to shrink large structures and to easily zoom into parts of them so that the user can use mouseover to zoom into parts of large structures in an intuitive interaction.

8. Acknowledgements

The work reported on in this paper is funded by the Research Council of Norway under the INESS infrastructure project and by the University of Bergen.

9. Bibliographical References

Augustinus, L., Vandeghinste, V., and Van Eynde, F. (2012). Example-based treebank querying. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Augustinus, L., Vandeghinste, V., Schuurman, I., and Van Eynde, F. (2013). Example-based treebank querying with GrETEL – now also for spoken Dutch. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, volume 16 of *NEALT Proceedings Series*, pages 423–428.

Carter, D. (1997). The TreeBanker: A tool for supervised training of parsed corpora. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pages 598–603, Providence, Rhode Island.

King, T. H., Dipper, S., Frank, A., Kuhn, J., and Maxwell III, J. T. (2004). Ambiguity management in grammar writing. *Research on Language and Computation*, 2(2):259–280.

Martens, S. (2013). TüNDRA: A web application for treebank search and visualization. In Sandra Kübler, et al., editors, *Proceedings of the Twelfth Workshop on Treebanks and Linguistic Theories (TLT12)*, pages 133–144. Bulgarian Academy of Sciences.

Maxwell, J. and Kaplan, R. M. (1993). The interface between phrasal and functional constraints. *Computational Linguistics*, 19(4):571–589.

Meurer, P. (2012). INESS-Search: A search system for LFG (and other) treebanks. In Miriam Butt et al., editors, *Proceedings of the LFG '12 Conference*, LFG Online Proceedings, pages 404–421, Stanford, CA. CSLI Publications.

Oepen, S., Flickinger, D., Toutanova, K., and Manning, C. D. (2004). LinGO Redwoods: A rich and dynamic treebank for HPSG. *Research on Language and Computation*, 2(4):575–596, December.

Rosén, V., Meurer, P., and De Smedt, K. (2007). Designing and implementing discriminants for LFG grammars. In Tracy Holloway King et al., editors, *The Proceedings of the LFG '07 Conference*, pages 397–417. CSLI Publications, Stanford.

Rosén, V., Meurer, P., and De Smedt, K. (2009). LFG Parsebanker: A toolkit for building and searching a treebank as a parsed corpus. In Frank Van Eynde, et al., editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT7)*, pages 127–133, Utrecht. LOT.

Rosén, V., De Smedt, K., Meurer, P., and Dyvik, H. (2012). An open infrastructure for advanced treebanking. In Jan Hajič, et al., editors, *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29, Istanbul, Turkey.

Zeldes, A., Ritz, J., Lüdeling, A., and Chiarcos, C. (2009). ANNIS: A search tool for multi-layer annotated corpora. In Michaela Mahlberg, et al., editors, *Proceedings of the Corpus Linguistics Conference 2009 (CL2009)*, pages 358–362.

Visualizing Language Change: Dative Subjects in Icelandic

Christin Schätzle, Dominik Sacha
University of Konstanz
78567 Konstanz, Germany
christin.schaetzle@uni-konstanz.de, dominik.sacha@uni-konstanz.de

Abstract

This paper presents a visualization tool for the analysis of diachronic multidimensional language data. Our tool was developed with respect to a corpus study of dative subjects in Icelandic based on the Icelandic Parsed Historical Corpus (Wallenberg et al., 2011) which investigates determining factors for the appearance of dative subjects in the history of Icelandic. The visualization provides an interactive access to the underlying multidimensional data and significantly facilitates the analysis of the complex diachronic interactions of factors at hand. We were able to identify various interactions of conditioning factors for dative subjects in Icelandic via the visualization tool and showed that dative subjects are increasingly associated with experiencer arguments in Icelandic across time. We also found that the rise of dative subjects with experiencer arguments is correlated with an increasing use of middle voice. This lexical semantic change argues against dative subjects as a Proto Indo-European inheritance. Moreover, the visualization helped us to draw conclusions about uncertainties and problems of our lexical semantic data annotation which will be revised for future work.

Keywords: Dative subjects, visualization, Icelandic, IcePaHC, language change

1. Introduction

In this paper, we present a multidimensional visualization for analyzing the historical development of dative subjects in Icelandic. The visualization is based on data from a concrete corpus study which investigates the diachronic interaction among lexical semantic verb classes, voice and dative subjects in the Icelandic Parsed Historical Corpus (IcePaHC) (Wallenberg et al., 2011).

The visualization follows an 'overview first – details on demand' approach (Keim et al., 2008) which allows for an interactive exploratory access to the underlying historical language data. Text glyph representations visualizing the occurrences of text features over time and genre enable the analyst to identify features and texts of interest. A previous approach by Butt et al. (2014) has been adapted for hierarchical feature dependencies and extended with further interaction possibilities.

The visualization showed that although dative subjects exist throughout all attested stages of Icelandic, their distribution has been changing diachronically. We found in particular that dative subjects of experiencer predicates begin to increase in the latter part of the 19th century and that this increase correlates with an increasing use of middle voice. Our findings demonstrate an increasing systematic relation between experiencer arguments and dative case along the history of Icelandic which argues against the Proto Indo-European inheritance of dative subjects.

Standard corpus linguistic methods generally only allow for binary comparisons of factors at a time, turning a comparison across several dimensions while maintaining a diachronic component into a complicated and tedious task. Our visualization tool represents a novel methodology which facilitates the detection of meaningful patterns in a complex multifactorial set of diachronic data and furthers the understanding of dative subjects in Icelandic across time.

2. Theoretical Background

2.1. The Diachrony of Dative Subjects

Dative subjects exist in a multitude of modern Indo-European languages. However, there is an ongoing debate about two competing narratives concerning their origin: 1.) the so-called Oblique Subject or Semantic Alignment Hypothesis (Barðdal and Eythórsson, 2009; Barðdal et al., 2012) which takes dative subjects to be a Proto Indo-European inheritance; 2.) the so-called Object-to-Subject Hypothesis which assumes that dative subjects were innovated at a later stage through reanalysis of former objects (Haspelmath, 2001).
Evidence for the latter hypothesis comes from Indo-Aryan. While there is no evidence for the existence of dative subjects in Old Indo-Aryan (Hock, 1990), dative subjects are generally found in Modern Indo-Aryan from the 11th to the 12th century onwards (Deo, 2003; Butt and Deo, 2013). There is evidence that these dative subjects emerged through the reanalysis of former objects conditioned by lexical semantic factors and the increasing systematic relation between dative case and experiencer arguments.
Barðdal and her colleagues (Barðdal and Eythórsson, 2009; Barðdal et al., 2012) on the other hand provide evidence for the Oblique Subject Hypothesis. They propose that a fixed Dative Subject Construction has been inherited through the history of Indo-European into the daughter language families by pointing towards the pervasiveness and productivity of dative subjects in the early stages of several Indo-European languages, including Icelandic.

2.2. Dative Subjects in Icelandic

Although the default case for subjects in Icelandic is nominative, non-nominative subjects, including dative subjects, are attested throughout the history of Icelandic (Barðdal and Eythórsson, 2009) with their existence being well established in Modern Icelandic (Andrews, 1976; Andrews, 2001; Zaenen et al., 1985). Dative subjects are mainly found in association with two major classes of verbs: 1.) ex-

periencer or psych predicates; 2.) happenstance predicates, including verbs of gain/success, happening, hindrance, ontological states, speaking, possession, evidentiality and modals (Barðdal, 2011). Example (1) from the oldest text in the IcePaHC ("Fyrsta málfræðiritgerðin", dated around 1150 CE) illustrates the use of the experiencer verb *líka* 'like' having a dative subject.

(1) Vel líkuðu goðrøði góð
well like.PAST.3.PL G.DAT.SG good.NOM.PL
røði, [...]
oar.NOM.PL
'Goðrøði (the good oarsman) liked good oars well, [...]'

The case marking system of Icelandic is currently undergoing a change in progress by which accusative experiencers are systematically replaced with datives, see example (2) in which the original accusative experiencer subject in (2-a) becomes a dative subject in (2-b). This change presumably began in the latter part of the 19th century and has been dubbed "Dative Substitution" or "Dative Sickness" (Barðdal, 2011; Jónsson, 2003; Smith, 1996). This change in progress suggests lexical semantics as a conditioning factor for dative subjects in Icelandic.

(2) a. Mig langar að fara.
I.ACC long.PRES to go
'I long to go.'

b. Mér langar að fara.
I.DAT long.PRES to go
'I long to go.' (Smith, 1996, 22)

Dative subject predicates may occur with all three Icelandic voices: active, passive and middle. The default case marking pattern of a transitive sentence in Icelandic is nominative on the subject and accusative on the object. Under passivization, the accusative object becomes the nominative subject of the sentence. Passive in Icelandic is formed periphrastically, i.e. via the copula *vera* 'to be' and the past participle, as shown in example (3):

(3) a. einhver barði
somebody.NOM hit.PAST.3.SG
strákana í skólanum
the.boy.ACC.PL in the.school.DAT.SG
'Somebody hit the boys in school.'

b. strákarnir voru
the.boy.NOM.PL be.PAST.3.PL
barðir í skólanum
hit.PPART.M.NOM.PL in the.school.DAT.SG
'The boys were hit in school.'
(Thráinsson, 1994, 177)

Dative and genitive objects are also possible in transitive sentences. However, these objects generally preserve their case marking under passivization, see example (4) for a dative argument.

(4) a. Skipstjórinn sökkti
the.captain.NOM.SG sink.PAST.3.SG
skipinu.
the.ship.DAT.SG
'The captain sank the ship.'

b. Skipinu var
the.ship.DAT.SG be.PAST.3.SG
sökkt af skipstjóranum.
sink.PPART.N.NOM.SG by the.captain.DAT.SG
'The ship was sunk by the captain.'
(Zaenen and Maling, 1984, 141f)

In Icelandic, middle voice on verbs is marked via the suffix *-st* (Anderson, 1990; Sigurðsson, 1989; Wood, 2015). Similar to passivization, transitive accusative objects are realized as nominative subjects under middle formation. However, dative objects show a different behavior with respect to middle formation: When the dative marks a theme/patient on the object, dative case is not preserved, see (5). When the dative object is a benefactive or goal, case is preserved, see example (6-b).

(5) a. Ég hellti mjólkinni
I.NOM spill.PAST.1.SG the.milk.DAT.SG
niður.
down
'I spilled the milk down.'

b. Mjólkin helltist niður.
the.milk.NOM.SG spil.PAST.MID down
'The milk spilled down.'
(Sigurðsson, 1989, 265)

(6) a. Pétur bauð mér vinnu.
Peter.NOM offer.PAST.3.SG I.DAT job.ACC.SG
'Peter offered me a job.'

b. Mér bauðst vinna.
I.DAT offer.PAST.MID job.NOM.SG
'I got the opportunity to get a job.'
(Sigurðsson, 1989, 260)

This case alternation is a further indicator for a lexical semantic approach to case marking in Icelandic in which voice and verbal lexical semantics seem to be conditioning factors.

3. Corpus Study

Although dative subjects are generally found in Old and Modern Icelandic, the earliest attested manuscripts of Icelandic stem from the 12th century, the time around which dative subjects emerged in Indo-Aryan. Given the Indo-Aryan developments and the empirical facts about Dative Substitution in Icelandic, we assume that the distribution of dative subject across semantic verb classes in Icelandic should change or at least vary diachronically. Moreover, we expect voice to be a conditioning factor for dative subjects in Icelandic, interacting with verbal lexical semantic factors. In order to investigate the diachronic development of dative subjects with respect to the factors lexical seman-

tic verb class and voice, we conducted a corpus linguistic study based on the IcePaHC.

The IcePaHC is ideal for the investigation of historical language change in Icelandic as it includes texts dating from the 12th to the 21st century covering the earliest attested stages of Icelandic as well as modern literature. The 60 texts in IcePaHC comprise about 1 million words and come from different genres (mainly Sagas). Moreover, the corpus provides information about case and grammatical relations and is annotated according to the syntactic annotation scheme of the Penn Treebank (Marcus et al., 1993).

As a first step, we added a secondary annotation layer containing information about the semantic verb classes of dative subject predicates to IcePaHC. These verb classes are a combination of Levin's verb classification for English (Levin, 1993) and the verb classes for dative subjects in Icelandic established by Barðdal et al. (2012).[1]

Subsequently, we extracted all dative subject predicates from the corpus via a Perl script and found over 4000 instances of dative marked subjects in combination with one of our verb classes. We then analyzed our data with respect to voice and divided our data and results into relevant time stages for Icelandic (Haugen, 1984). We used χ^2 to determine whether our observed distributions differ from what was expected (* p<0.05, ** p<0.01, *** p<0.001).

Time	active	passive	middle	χ^2
1150-1350	64.4%	18.2%	17.4%	***
1350-1550	66.8%	17.5%	15.7%	***
1550-1750	46.1%	28.8%	25.1%	***
1750-1900	53.1%	20.8%	26.2%	
1900-present	43.2%	14.3%	42.5%	***
all	55.3%	19.5%	25.5%	

Table 1: Diachronic distribution of dative subject predicates by voice

The results displayed in Table 1 show changes in the diachrony of dative subjects with respect to voice. Dative subjects most often appear in active constructions which are mainly experiencers and found in conjunction with psych predicates (mean=56.2%) such as *líka* 'like'. However, these constructions are not diachronically stable and their frequencies are decreasing over the whole time span. Passives with dative subjects appear less often than actives and also show an overall decrease. Within these constructions, dative subjects are mainly associated with verbs of communication (mean=14.8%), e.g. *tilkynna* 'announce', and change of possession (mean=14.6%), e.g. *kaupa* 'buy', but are also decreasing within these verb classes as a result of the overall decrease of passive dative subjects. While we observe an overall decrease within active and passive dative subjects, dative subjects together with middle verbs are on the increase. Middle morphology with dative subjects also occurs most often with psych predicates (mean=79.0%), e.g. *leiðast* 'be bored'.

Note that the percentages in Table 1 would seem to indicate significant changes in the third time stage (1550–1750). However, these deviations are due to a genre effect in the corpus as discussed in Butt et al. (2014): The third time stage mainly consists of religious and legal texts in the corpus, while Sagas predominate in the other time stages.

Although our data shows significant changes with respect to voice, we can only partly identify changes or respective stabilities within the lexical semantic verb classes. By analyzing 15 different verb classes across three different voices multiple data tables containing a multitude of different characteristics were generated. In order to compare and uncover significant results, one has to skip back and forth between these matrices while at the same time keeping the overall picture in mind. Meaningful patterns become difficult to identify and may stay undetected. Furthermore, some of the relative frequencies in our data tables are based on very few occurrences of the actual observation derogating the significances of the statistical conclusions. Moreover, such an analysis only allows for binary comparisons of factors at a time and only provides limited access to the actual interactions at hand. In order to be able to account for all dependencies and interactions of factors in our complex and multidimensional data set, we decided to include a visualization system into our data analysis process.

4. Diachronic Multidimensional Visualization of Historical Data

4.1. Motivation

Descriptive statistics is often based on specific assumptions and requires the definition of categories, groups, or temporal episodes/epochs for comparison. Furthermore, the final output is a number that is composed of several factors and therefore hardly interpretable. With this respect, interesting patterns may be hidden (e.g., absorbed by too large episodes/categories). In addition, analysis problems are often ill-defined and hypotheses may be very vague and hard to define. Visual interactive data analysis aims to bridge this gap by enabling the analyst to investigate the data in an exploratory fashion. Several approaches have illustrated the power of visualization as a valuable addition to well established analysis techniques, fostering the generation and validation of hypotheses that may lead to often unexpected insights.

In order to gain a better overview and access to our data set of several dependent text features, we developed and extended a visualization system (Butt et al., 2014) that enables to drill down into specific data items and to get details on demand. Note that the data (and text features) are analyzed as is and it is not necessary to specify any assumptions in advance.

Finally, the visual design is driven to support the analysts tasks best. For example, instead of visualizing the absolute number of feature occurrences, the relative numbers are calculated in order to enable an effective comparison

[1]The following categories were used in this corpus study: psych, existence, motion, sending, concealment, social interaction, permission, measure, put, involving the body, communication, verbs with predicative complements, change of possession, change of state and aspectual verbs. In future work, we want to experiment with a different set of verb classes (e.g. the verb classes proposed by Jónsson (2003)) as the classes here are not ideal for our research question and the boarders between them are not always clear.

over time. Specific glyphs are designed to visualize if the feature occurrence is more or less than expected.

4.2. Diachronic Visualization of V1

Butt et al. (2014) developed a novel visualization tool for the analysis of historical language data based on a concrete corpus study of V1 (Verb First) word order in the IcePaHC. Figure 1 shows a single text from IcePaHC which is visualized as one composed glyph. The horizontal bar on top indicates the length of the text in comparison to the longest text in the corpus which covers the whole width of the glyph. The light gray stripes in the horizontal bar correspond to the occurrences of V1 in the narrative flow of the text. The horizontal line on the right side of the colored matrix indicates the time span covered by the corpus with a vertical line indicating the age of the given text.

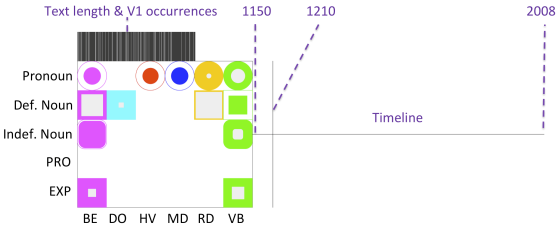

Figure 1: A V1 text glyph consists of three main parts: 1.) a horizontal bar on top displaying the text length and V1-occurrences, 2.) a matrix containing colored shapes representing the occurrence of a particular verb together with a particular subject and 3.) a timeline on the right (Figure from Butt et al. (2014)).

The colored matrix represents the interaction of verb and subject types. While the columns encode the different verb types of V1 verbs as given by the corpus, i.e. BE 'be', HV 'have', DO 'do', RD 'become', MD modals and VB main verbs, the rows indicate the type of subject involved in the V1 sentence: pronominal subject (Pronoun), definite subject (Def. Noun), indefinite subject (Indef. Noun), pro-dropped subject (PRO) and null expletive subject (EXP). Although the different interactions can be identified by position alone, a redundant coding for verb type by color and subject type by shape has been added for better visibility. If a cell is empty, the given interaction did not appear in the text. Moreover, the matrix cell shows whether a given interaction occurred more or less frequently than expected within the text by either filling the shape from outside or inside as explained in Figure 2.

All texts contained in IcePaHC are visualized as text glyphs and are arranged from oldest to newest with the oldest text at the very top and the newest at the bottom (vertical alignment). The text representations are also aligned horizontally according to their genre. The genre labels are shown on top of the visualization and the text representations have the respective horizontal positions. Our visualization in Figure 3 illustrates this layout, but the text glyphs are adapted to the research question presented in this paper. In addition, the visualization is interactive, and once a potentially interesting pattern has been identified by the analyst in the overview, one can zoom into particular glyphs and examine them in more detail. Furthermore, the visualization provides details on demand tooltips via mouse-over on different parts of the text representation.

Figure 2: The glyphs are visualized according to the feature occurrence. The glyph is filled from inside if the feature occurrence is smaller than expected, and filled from outside if the feature occurrence is larger than expected. The relative and expected occurrences are calculated based on the text length and the average occurrences in the whole corpus.

Butt et al. (2014) were able to provide meaningful findings for the diachrony of V1 in Icelandic by means of the visualization. They conclude that their tool should be applicable to any diachronic study that seeks to understand a comparable multifactorial diachronic interaction. Hence we decided to use this visualization as a starting point for visualizing the diachrony of dative subjects in Icelandic.

4.3. Design and Development Process

During our development process of adapting the system from Butt et al. (2014) to cope with the data described in this paper, we observed a set of design problems and challenges which had to be resolved in order to satisfy the analysts needs.

First, we discovered that the number of visualized text features for Icelandic dative subject verb classes is very large and therefore an aggregation to feature-classes is needed to provide an overview. In order to cope with this problem, we offer an aggregated representation for each text. The features may be extended for more detailed verb classes (horizontal) as well es for voice (active, passive and middle – vertical) on demand.

Second, in the first version of the visualization, it is hard to track specific features over time due to the genre shifted layout of the text glyphs. Therefore, we allow the analyst to switch between the genre shifted layout and the default layout that places each text representation among each other. In order to support the investigation of feature developments over time we also visualize the relative feature occurrence count via the tick marks of the time line when a specific feature is hovered.

Finally, by adding more and more functionality we observed that the variety of interactions (tool-tip, tick marks, expand/fold etc.) disturbed each other. Therefore, we needed to disable/enable the tool-tips allowing the analyst to switch between detail and comparison mode. This section illustrates that our visualization has been developed through an iterative design process for adapting the tool to cope with the data set and to support the analyst in performing analysis tasks of 1.) overview, 2.) comparison (over time), and 3.) details on demand. In the next section, we describe the final visualization design in more detail.

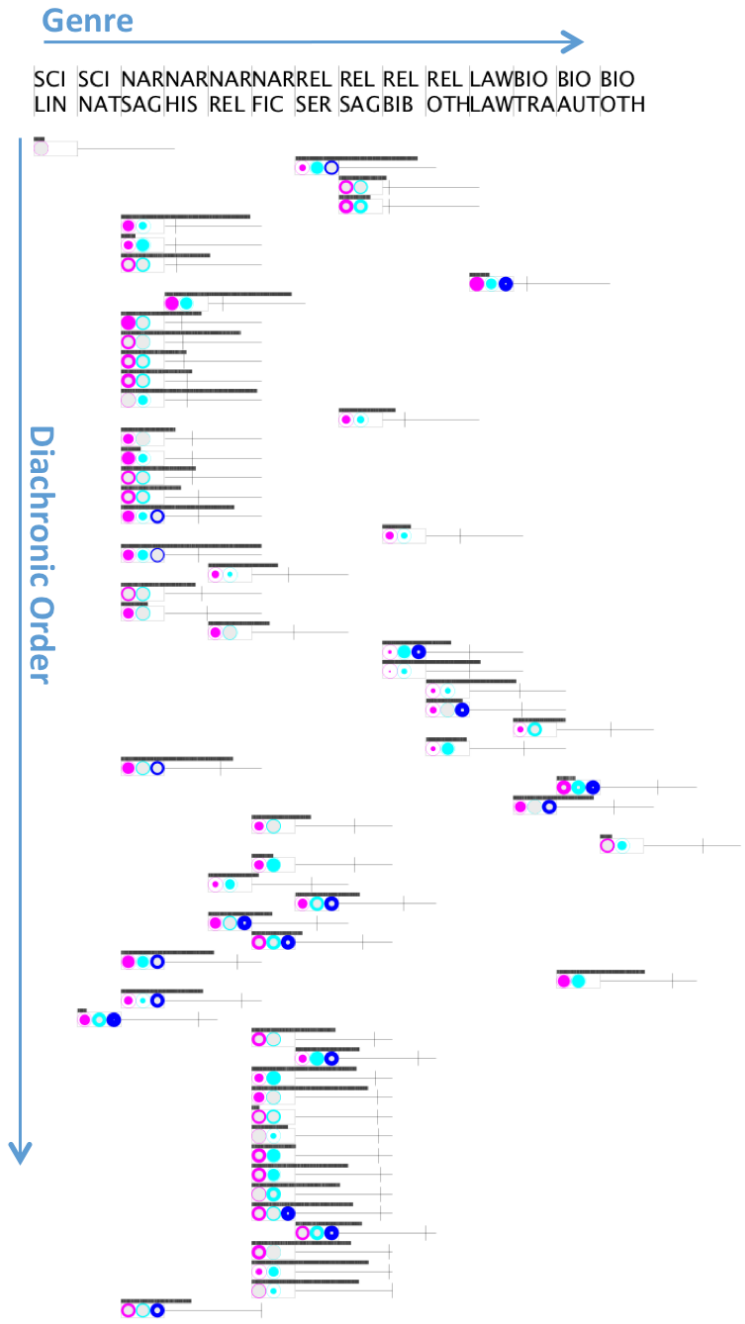

Figure 3: All texts of the corpus positioned among each other (y-axis, diachronic order) showing aggregated text feature classes including genre shift layout of texts (x-axis, genre). The genre labels are shown on top (scientific texts SCI, narratives NAR, religious texts REL, law texts LAW and biographies BIO) and can be read as columns.

4.4. Diachronic Visualization of Dative Subjects

Figure 3 shows the overall visualization of dative subject verb classes for each text in the IcePaHC which generally adopts the design of Butt et al. (2014) with the diachrony displayed from top to bottom and the genre from left to right. In Figure 3, the verb classes in each text are aggregated into a higher class categorization as given by Barðdal (2011) and Barðdal et al. (2012)[2]. In addition, only one line

[2]These categories are experience-based (experiencer) and happenstance predicates, verbs of evidentiality and verbs of modality. Verbs of evidentiality were excluded in our visualization as they consist of verbs of seeming and appearing for Barðdal and her colleagues. These verbs have been subsumed under psych verbs in our study and are hence categorized into experiencer predicates.

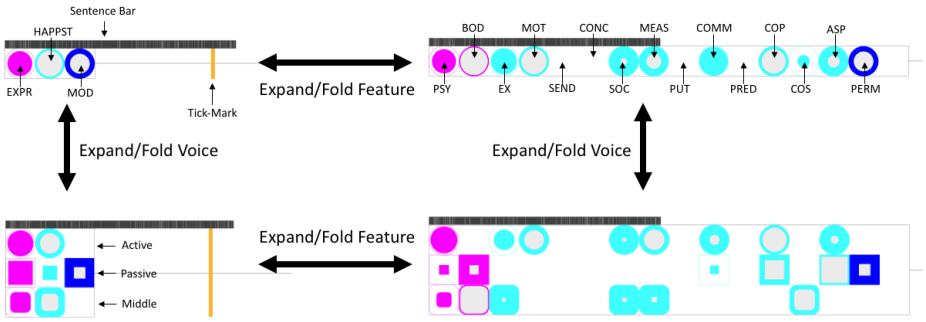

Figure 4: Text representations are composed by a sentence bar on top, the different feature glyphs, and a tick-mark indicating the selected feature value (or time if nothing is selected). Text representations may be expanded/folded on demand. Top left: aggregated feature classes, top right: detailed feature text classes, bottom left: aggregated text feature classes extended for voice, bottom right: detailed feature classes extended for voice.

of features is shown without further distinctions (instead of a matrix as in the version Butt et al. (2014)), allowing us to simplify the overview visualization. The detailed features and different types can be shown on demand.

Figure 4 illustrates the different text representation possibilities which were needed to cope with the high number of text features. Each text representation is composed of a sentence bar (indicating the length of the text as well as the positions of the features in the text as shown in Figure 1), the feature matrix and a tick-mark. The text representations may be expanded (and folded) on demand (by pressing a button, clicking on a text, or hovering over a feature glyph for temporal comparison). Each aggregated text feature glyph represents the occurence of the higher level verb classes which are coded by different colors and position, see Figure 4-top-left. Table 2 displays the labels used in Figure 4 and the respective visual encoding for each factor. The higher level verb classes EXPR, HAPPST, and MOD are redundantly coded by color and can be expanded into more fine-grained verb classes which appear at a fixed position and have the same color as the higher class under which they are subsumed, see Figure 4-top-right. Moreover, each variant may be expanded to show the interaction of verb classes with different voices which are encoded by different shapes (Figure 4-bottom). The feature glyphs are designed to visualize the relative expected occurrence of a feature by either filling the shape from inside (less than expected) or outside (more than expected), see Figure 2.

By hovering a particular position within the sentence bar, the analyst is provided with a tool-tip showing the relevant dative subject sentence. The same functionality is provided by hovering over a specific text feature glyph. In this case, the tick marks of all texts are positioned to show the relative frequency of the selected feature value (see Figure 5). This enables a comparison of feature occurrences over time. If no feature is hovered/selected the tick marks are positioned according to the age of the text.

The horizontal genre shift may be turned off and on in order to either analyze genre effects or to compare features over time. Figure 5 shows the detailed text representations without genre shift. Additionally, the PERM feature class is hovered and consequently all the tick marks align according

Label	Verb Class	Encoding
EXPR	Experiencer verbs	Magenta
PSY	Psych verbs	Magenta
BOD	Body verbs	Magenta
HAPP	Happenstance verbs	Light Blue
EX	Existence verbs	Light Blue
MOT	Motion verbs	Light Blue
SEND	Verbs of sending	Light Blue
CONC	Verbs of concealment	Light Blue
SOC	Verbs of social interaction	Light Blue
MEAS	Measure verbs	Light Blue
PUT	Put verbs	Light Blue
COMM	Communication verbs	Light Blue
PRED	Predicative complements	Light Blue
COP	Change of possession	Light Blue
COS	Change of state	Light Blue
ASP	Aspectual verbs	Light Blue
MOD	Modality verbs	Dark Blue
PERM	Permission verbs	Dark Blue

Voice	Encoding
Active	Circle
Passive	Rectangle
Middle	Rounded Rectangle

Table 2: Visual encodings of text features

to this particular feature value. We can now easily compare the feature occurrence over time and observe that e.g. the PERM feature is rarely used in the upper part compared to the middle part, see 5-B. The analyst may now enable the tool-tips and expand or fold the feature classes and voice in order to investigate further.

The visualization presented in this paper opens up new possibilities for the analysis of linguistic developments in historical language data. Our interactive and multidimensional visualization allows for an overall view of all texts in the corpus with the possibility to drill down and zoom into particular glyphs once interesting patterns have been identified in the overall picture. The glyphs contain information about the interactions between different language features occur-

Figure 5: Complete corpus visualized showing detailed feature classes. Tick marks show the occurrence of one particular selected/hovered feature (PERM) enabling comparison over time. Additionally, this representation allows to spot visual patterns that "pop out" visually.

ing in a given text. Via clicking on the text glyph or pressing a button different interaction possibilities can be accessed. Moreover, the horizontal text alignment can be modified in order to trace the diachronic developments of interactions in more detail. The visualization also provides information about genre type which can be accessed via mouse-over or text alignment. This feature of the visualization helped us to identify a genre effect in the corpus (visible in 5-A) which has been previously discussed in other work (Butt et al., 2014; Schätzle et al., 2014; Schätzle et al., 2015).

5. Results and Discussion

The visualization at hand represents a useful and powerful methodology for the analysis of multidimensional and complex language data as found in historical linguistic research. Our visualization tool provides new insights into our large data set by granting an interactive access to the underlying data.

Our visual analysis confirms that dative subjects already exist in the earliest Icelandic texts and are a common phenomenon throughout the history of Icelandic. We can identify this at a glance when considering the overall visualization displayed in Figure 3 where dative subjects are present in every text glyph. Moreover, dative subjects are mainly associated with experiencer and, to a lesser extent, with happenstance predicates throughout the corpus. Verbs of modality (which are in essence verbs of permission, see Table 2) rarely occur together with dative subjects in the first half of the corpus, but are found more often in the latter part. Furthermore, the horizontal realignment of our data helped to uncover an increase in the use of experiencer predicates starting in the end of the 19th century. The visualization tool also enabled the interactive investigation of the interaction between verb class and voice. We found that the rise of experiencer verbs is mainly caused by an increasing use of experiencer predicates with middle constructions. Overall, experience-based predicates and happenstance predicates appear with all three voices in the corpus, while verbs of modality most often occurred with passives and more rarely with middles, but never in active constructions.

These findings show that the distribution of dative subjects in Icelandic has been changing over the past millenium. These diachronic changes are mainly due to an increase of experiencer predicates with dative subjects caused by an increasing use of dative subjects with middle constructions. This points towards an increasing systematic association of dative case with experiencer arguments which in turn argues against the Oblique Subject Hypothesis.

Yet, the visual analysis of the more fine-grained verb classes represents a more tedious task. A large set of visual cues has to be distinguished and processed which makes the at a glance identification of salient patterns merely impossible. However, we could clearly identify a genre effect by means of the visualization displayed in Figure 5-A which influenced our previous statistical data analysis. Psych verbs and verbs involving the body (both in magenta), i.e. experiencer predicates, appear significantly less often than expected with dative subjects or are even absent in texts within the range of 5-A. By shifting back to the genre layout, we found that segment A mainly consists of religious and legal texts in the corpus, while the other texts are narrative in nature.

In the future, we plan to focus on the improvement of the visualization of the lower verb classes and also revise our lexical semantic annotation scheme making it more likely to find significant patterns in the visualization. Other semantic verb classifications for verbs having dative arguments as e.g. mentioned in Jónsson (2003), Jónsson (2009) and Maling (2002) will be considered hereby and we plan to include more information about the thematic roles of the subjects (e.g. experiencer, recipient, theme).

6. Conclusion

Concluding, we present a useful visualization tool for the study of historical language change. Our tool is an extension of the visualization tool developed by Butt et al. (2014) and was created with respect to a concrete corpus

study of dative subjects in Icelandic. The visualization provides a powerful method to uncover and understand multi-dimensional interactions of factors conditioning diachronic language change. Moreover, we were able to draw conclusions about flaws in our verb classification system from the visualization which will be addressed in future work.

7. Acknowledgements

This work was partially funded by the German Research Foundation (DFG) within the project "Visual Analysis of Language Change and Use Patterns" and within project D02 "Evaluation Metrics for Visual Analytics in Linguistics" of SFB/Transregio 161.

8. Bibliographical References

Anderson, S. R. (1990). The grammar of Icelandic verbs in *-st*. In J. Maling et al., editors, *Modern Icelandic Syntax*, pages 235—273. Academic Press, San Diego.

Andrews, A. D. (1976). The VP complement analysis in modern Icelandic. In *Proceedings of the North East Linguistic Society (NELS)*, volume 6, pages 1–21. Reprinted 1990 with minor revisions in *Modern Icelandic Syntax*, 165–185.

Andrews, A. D. (2001). Non-canonical A/S Marking in Icelandic. In M. Noonan, editor, *Non-Canonical Marking of Subjects and Objects*, pages 85–111. John Benjamins, Amsterdam.

Barðdal, J. and Eythórsson, T. (2009). The origin of the oblique subject construction: An indo-european comparison. In V. Bubenik, et al., editors, *Grammatical Change in Indo-European Languages*, pages 179–193. John Benjamins, Amsterdam.

Barðdal, J., Smitherman, T., Bjarnadóttir, V., Danesi, S., Jenset, G. B., and McGillivray, B. (2012). 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.

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

Butt, M. and Deo, A. (2013). A historical perspective on dative subjects in Indo-Aryan. Paper presented at the LFG13 Conference.

Butt, M., Bögel, T., Kotcheva, K., Schätzle, C., Rohrdantz, C., Sacha, D., Dehe, N., and Keim, D. (2014). 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.

Deo, A. (2003). Valency change and case marking: Marathi dative experiencers. Handout from the Pioneer Workshop on Case, Valency and Transitivity, June.

Haspelmath, M. (2001). Non-canonical marking of core arguments in European languages. In A. Y. Aikhenvald, et al., editors, *Non-Canonical Marking of Subjects and Objects*, pages 53–83. John Benjamins, Amsterdam.

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

Hock, H. H. (1990). Oblique subjects in Sanskrit. In M. Verma et al., editors, *Experiencer Subjects in South Asian Languages*, pages 119–139. CSLI Publications, Stanford.

Jónsson, J. G. (2003). Not so quirky: on subject case in Icelandic. In E. Brandner et al., editors, *New Perspectives on Case and Case Theory*, pages 129–164. CSLI Publications, Stanford.

Jónsson, J. G. (2009). Verb classes and dative objects in insular scandinavian. In J. Barðdal et al., editors, *The Role of Semantic, Pragmatic, and Discourse Factors in the Development of Case*, pages 203–224. John Benjamins Publishing.

Keim, D., Andrienko, G., Fekete, J.-D., Górg, C., Kohlhammer, J., and Melançon, G. (2008). Visual analytics: Definition, process, and challenges. In A. Kerren, et al., editors, *Information Visualization*, pages 154–175. Springer, Berlin.

Levin, B. (1993). *English Verb Classes and Alternation. A Preliminary Investigation*. Chicago: Universtiy of Chicago Press.

Maling, J. (2002). Það rignir þágufalli á Íslandi. Verbs with Dative Objects in Icelandic. *Íslenskt mál*, 24:31–105.

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

Schätzle, C., Sacha, D., and Butt, M. (2014). Diachronic visualization of dative subjects in Icelandic. Poster presentation at the Workshop on Big Data Visual Computing, *44th Annual Meeting of the Gesellschaft für Informatik*.

Schätzle, C., Butt, M., and Kotcheva, K. (2015). The diachrony of dative subjects and the middle in Icelandic: A corpus study. In M. Butt et al., editors, *Proceedings of the LFG15 Conference*. CSLI Publications.

Sigurðsson, H. A. (1989). *Verbal Syntax and Case in Icelandic: A Comparative GB Approach*. Ph.D. thesis, Lund University.

Smith, H. (1996). *Restrictiveness in Case Theory*. Cambridge University Press, Cambridge.

Thráinsson, H. (1994). Icelandic. In E. König et al., editors, *The Germanic Languages*, chapter 6, pages 142–188. London: Routledge.

Wallenberg, J. C., Ingason, A. K., Sigurðsson, E. F., and Rögnvaldsson, E. (2011). Icelandic Parced Historical Corpus (IcePaHC).

Wood, J. (2015). *Icelandic Morphosyntax and Argument Structure*. Springer, New York.

Zaenen, A. and Maling, J. (1984). Unaccusative, passive, and quirky case. In M. Cobler, et al., editors, *Proceedings of the Third West Coast Conference on Formal Linguistics (WCCFL)*, pages 317–329.

Zaenen, A., Maling, J., and Thráinsson, H. (1985). Case and Grammatical Functions. The Icelandic Passive. *Natural Language and Linguistic Theory*, 3(4):441—-483.

See the Forest AND the Trees: Visual Verb Class Identification in Urdu/Hindi VerbNet

Annette Hautli-Janisz
Department of Linguistics
University of Konstanz
annette.hautli@uni-konstanz.de

Abstract

Constructing a lexical resource like VerbNet involves the crucial task of forming syntactically motivated subclasses within larger, semantically motivated verb classes. This is particularly challenging when the syntactic behavior of verbs varies considerably within a class, making well-motivated subclasses hard to establish by hand. The present paper shows that a visual clustering approach substantially facilitates the development process: Based on a careful theoretical analysis of the syntactic properties of Urdu/Hindi motion verbs, each verb can be represented by an 8-dimensional feature vector which serves as the basis for the automatic clustering with k-Means. In order to overcome the blackbox of machine learning approaches and to make the resulting clusters interpretable, the visualization reduces the high-dimensional verb vectors to two dimensions and visualizes the clusters and their members via an interactive interface, allowing for an inspection of the underlying data. This leads to the formation of subclasses in Urdu/Hindi VerbNet that are theoretically as well as computationally well-motivated.

Keywords: Urdu/Hindi VerbNet, verb classes, visual clustering system

1. Introduction

Linguistics has a long tradition of visually representing language patterns, for instance by tree representations in syntax and spectograms in phonetics. Only rather recently, methods from the fields of Visual Data Analysis (Thomas and Cook, 2005; Ward et al., 2010) and Information Visualization (Card et al., 1999) have started to be used for the investigation of linguistic phenomena (Collins, 2010). Areas of investigation range from the cross-linguistic comparison of language features (Mayer et al., 2010a; Mayer et al., 2010b; Rohrdantz et al., 2012a) to the investigation of lexical semantic change (Rohrdantz et al., 2011; Heylen et al., 2012; Rohrdantz et al., 2012b) and discourse topic structure (Gold et al., 2015). These new ways of visually representing and analyzing large and complex data sets enables users to see overarching patterns at a glance while still maintaining a detailed view on the underlying data.

For that reason, visualization has found its way into the field of computational linguistics (CL), providing insights into methods of machine translation (Collins et al., 2007), discourse parsing (Zhao et al., 2012), discourse structure (Angus et al., 2012) and patterns in large corpora (Culy et al., 2011). Cluster visualizations are not a novel idea, as they have been applied in various other fields like finance, biology or geography (Schreck et al., 2009). However, as far as the literature is concerned, *interactive* systems are still less common, particularly for the methods employed in CL. Bringing together CL and visualization can overcome difficulties in interpreting results from machine learning algorithms – a drawback that often prevents theoretical linguists who work with computational models as tehy need to see patterns in large data sets from drawing detailed conclusions.

The present paper shows that the process of developing a linguistically well-motivated lexical resource like Urdu/Hindi VerbNet[1] can substantially benefit from visualization: Based on a thorough theoretical linguistic analysis of Urdu/Hindi motion verbs (Hautli-Janisz, 2014), a layer of visual analysis is employed to find linguistically-motivated subclasses within this verb class.

The paper proceeds as follows: Section 2 briefly provides describes the notions involved in the investigation. Section 3 presents the visual clustering system, followed by Section 4 where we present the results of the visual clustering. Section 5 illustrates yet another view on the data which allows to draw more general conclusions. Section 6 discusses the results and concludes the paper.

2. Background

In this section we briefly introduce the different concepts that are involved, namely the general structure of VerbNet (§2.1.) and Urdu/Hindi motion verbs (§2.2.).

2.1. VerbNet

English VerbNet (Kipper-Schuler, 2005; Kipper et al., 2008) is one of the most commonly used resources in English NLP applications and encodes the syntactic and semantic properties of English verbs. VerbNet is based on the work of Levin (1993), who assumes that the syntactic behavior of a verb is largely determined by its meaning. In the resource, verbs are grouped into classes according to their semantic coherence (e.g., Verbs of Motion), with their members constituting a set of syntactically synonymous words. This common syntactic behavior is manifested through the (un)grammaticality of a set of diathesis alternations, e.g. alternations like the passive, the causative or the dative shift.

A VerbNet class is characterized by a set of member verbs and a set of frames that the member verbs appear in. In turn, each frame is characterized by its syntactic structure

[1] http://ling.uni-konstanz.de/pages/home/hautli/uhvn/

and the meaning it entails: Whereas the syntactic information is encoded in terms of the parts of speech (e.g. NP, V, PP for 'John advanced into the room') that are connected to thematic roles (here: THEME, DESTINATION), the conceptual information is recorded by way of semantic predicates. For the example 'John advanced into the room', the predicates of *motion* and *path* describe the directed motion event.

Across VerbNets (for instance see Arabic VerbNet (Mousser, 2011)), the formation of subclasses within each verb class is done manually: Based on the theoretical linguistic investigation, verbs are grouped into theoretically-motivated subclasses – a laborious task that is challenging when the syntactic properties of the verbs in the class differ substantially. This paper shows that an interactive visual cluster analysis can facilitate this step in the resource development, without losing sight of the careful linguistic analysis underlying the verb class.

2.2. Urdu/Hindi Motion Verbs

Despite some recent work on Urdu/Hindi with respect to resource development, e.g. with the Hindi-Urdu Treebank (Bhatt et al., 2009) and Indo WordNet (Bhattacharyya, 2010), the language still belongs to those that are under-resourced, in particular with respect to verb resources. A first effort towards closing the gap is performed by Hautli-Janisz (2014), who investigates the syntactic and semantic properties of Urdu/Hindi motion verbs, with the aim of paving the way for an Urdu/Hindi VerbNet and establishing a set of properties that seem to be relevant when encoding Urdu/Hindi verbs in a lexical resource.

The investigation focuses on the class of motion verbs in Urdu/Hindi, a verb class that consists of 52 verbs, each characterizable by a set of syntactic and semantic properties, e.g. causativization, case marking and event structure. An initial, theoretically-motivated, subdivision of the class is based on the notion of scalarity (Levin and Rappaport Hovav, 2013): Those motion verbs that denote scalar events, those that denote nonscalar events and those that can lexicalize both depending on the context they can appear in (Hautli-Janisz, 2015). This gives us an initial set of three subclasses. However, even within these subclasses, verbs do not necessarily share properties, i.e. a further subclassification needs to be performed in order to arrive at a well-motivated class structure.

Table 1 presents an overview of the linguistic properties of Urdu/Hindi motion verbs: The first and second column provides the feature id and feature name, respectively, with the fourth and fifth column giving a brief explanation and a example verb for each feature. The third column shows the encoding of the features for the clustering step discussed in Section 3. Feature 1 describes the valency of the verb in its base form, i.e. whether it licenses only one argument (intransitive) or two arguments (transitive). Features 2 and 3 capture causativization, an alternation that is very common in Urdu/Hindi (Kachru, 1980; Saksena, 1982; Butt, 2003; Bhatt, 2003). In the direct causative (Feature 2), an external agent is added ('Ravi' in 'Ravi made her run.'), in the indirect causative, a so-called intermediate agent is added ('Amra' in 'Ravi made Amra run after the mouse.'). Features 4 and 5 deal with case marking, either the case alternation on the subject (Feature 4) or on the objects/obliques (Feature 5). Feature 6 marks whether a cognate object is available for a verb (e.g. 'run' in 'Sarah ran the tough run.'), with Feature 7 marking whether the telic path alternation is available for a verb (e.g. 'Jane ran' and 'Jane ran a mile'). Finally, Feature 8 encodes the subevental structure of the verb following First-Phase Syntax (Ramchand, 2008): A [proc] verb only licenses a process subevent, [init, proc] licenses an initiation and a process subevent and [proc, res] a process and result subevent.

3. Cluster Visualization

The automatic approach used here is based on Lamprecht et al. 2013), who show by way of a case study on Urdu N+V complex predicates (Butt et al., 2012) that cluster visualizations allows for an insightful investigation of linguistically motivated data. For the present investigation of Urdu/Hindi motion verbs, we use a slightly extended version of the system that can deal with the input format required for Urdu/Hindi motion verbs.

From a data analysis perspective, the input consists of 52 data objects, i.e. one object for each verb. Each data object is characterized by an eight-dimensional feature vector (Features 1 to 8 in Table 1), with each dimension corresponding to a verb's behavior in a specific syntactic alternation, in particular with respect to valency, the direct and the indirect causative, the case marking of subject and object, the pattern with respect to the two object alternations of cognate objects and telic paths and their event structure. In order to calculate the similarity between data objects, the system uses the Euclidean distance to measure the distances between the data object vectors. The smaller the distance between two data objects, the more similar they are in their syntactic and semantic feature structure. In order to cluster the different vectors, k-Means clustering is used. This method avoids the problem of non-reproducibility that is common to hierarchical clustering methods. However, k-Means involves an ex ante decision on the number of clusters that the data objects are allocated to. This step is a well-known issue in this type of approach, because it implies that the user has prior knowledge about the underlying data and moreover accepts a potentially less-than-optimal clustering result. From the data underlying the clustering here, it becomes clear that the patterns in the subclasses exhibit a fair amount of variation and a preliminary clustering experiment with the numbers of clusters ranging from k=3 to k=7 shows that setting k=3 and k=4, depending on the verb class, provides an appropriate approximation. Once the clustering is performed, the high-dimensional feature vectors are projected onto two dimensions using a principal component analysis (PCA) algorithm[2], which ensures that in the 2D projection, the distances between data objects in the high-dimensional space are preserved as accurately as possible.

Figure 1 provides an overview of the interface of the cluster visualization, with the configuration area on the left, the vi-

[2]http://workshop.mkobos.com/2011/java-pca-transformation-library/

ID	Feature	Value	Explanation	Example
1	VALENCY	0	Intransitive verb	*ʊbʰar-na* 'to rise'
		1	Transitive verb	*pʰand-na* 'to leap over'
2	CAUS	0	No direct causative	*a-na* 'to come'
		1	Direct causative	*ʊbʰar-na* 'to rise'
3	ICAUS	0	No indirect causative	*ja-na* 'to go'
		1	Indirect causative	*ʊbʰar-va-na* 'to rise'
4	SUBJcase	0	Nominative	*gʰʊs-na* 'to enter'
		1	Ergative	*pʰand-na* 'to leap over'
5	OBJ/OBLcase	0	N/A	*doṛ-na* 'to run'
		1	Nominative/accusative	*pʰand-na* 'to leap over'
		2	Locative	*gʰʊs-na* 'to enter'
6	CogOBJ	0	N/A	*ʈʰla-na* 'to strut'
		1	Cognate object construction	*bʰag-na* 'to run'
7	PathOBJ	0	N/A	*ter-na* 'to float'
		1	Telic path alternation	*bʰag-na* 'to run'
8	EVENT	0	[proc]	*ʊbʰar-na* 'to rise'
		1	[init, proc]	*pʰand-na* 'to leap over'
		2	[proc, res]	*gʰʊs-na* 'to enter'

Table 1: Input features and their values

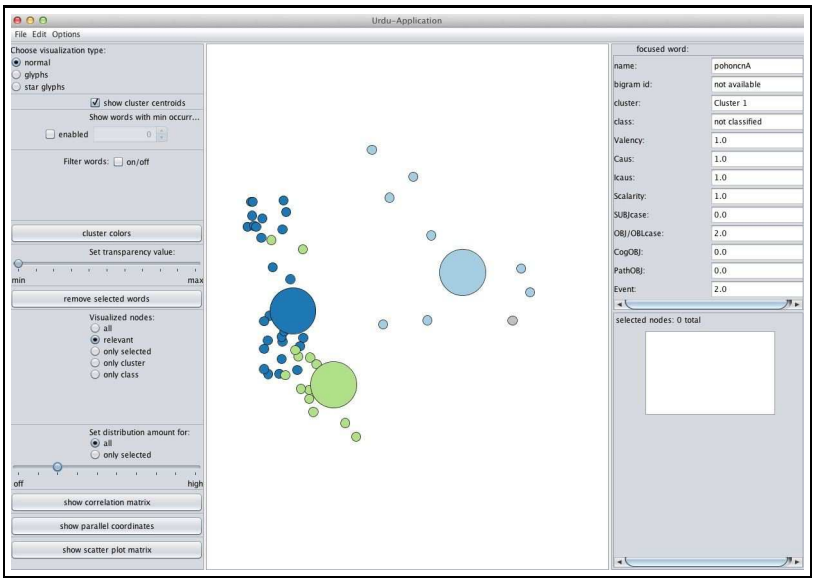

Figure 1: Interface of the cluster visualization system (Lamprecht et al. 2013)

sualization area with a sample clustering result in the middle and the description area on the right. The visualization area is mainly built with the piccolo2d library[4] and initially shows data objects as colored circles, with color indicating cluster membership (e.g. three clusters in Figure 1).
The position of the cluster centroids, i.e. the location of the prototypical cluster members, is visualized by the larger colored circle. Hovering over a dot displays information on the particular data point in the description area to the right, together with its cluster membership and the feature structure that characterizes it. By scrolling, the user can zoom in and out of the visualization canvas. The interactivity is what sets this visualization apart from those that are available in standard statistics packages: In these tools, no interaction and (consequently) verification of the clustering result is possible, a property that we believe is crucial for a linguistically well-motivated visual analysis.

Due to the nature of the data, some data objects are plotted on top of each other – this means that a number of data objects have the same feature vector. The consequence is that only the topmost data object is visible. In order to improve visual access to overplotted data objects, the system offers two interactive possibilities: On the one hand, the user can increase the transparency of the data points, making underlying data objects visible. On the other hand, the system allows for jittering, i.e. data objects can be repositioned with a small fixed deviation from their initial position. The de-

[4]http://www.piccolo2d.org/

gree of deviation can be determined interactively by using a slider in the configuration area.

Another important feature of the system for the task at hand is the possibility of selecting multiple data objects for further processing or filtering, with a list of selected data objects shown in the description area on the right. By right-clicking on a data object, the user can assign a unique class (and class color) to the item. Moreover, the user can fade in the cluster centroids (illustrated by the larger dots in the respective cluster color in Figure 1), where the overall feature distribution of the cluster can be examined in a tooltip by hovering over the respective centroid.

The following section shows how the properties of interactivity, jittering, filtering and reclustering are used for establishing a VerbNet class for Urdu/Hindi motion verbs.

4. Subclass Identification

As discussed in §2.2., the class of Urdu/Hindi motion verbs is initially subdivided into three subclasses, the set of scalar motion verbs, those verbs that lexicalize nonscalar motion and those that can lexicalize both depending on the context they appear in. Establishing syntactically synonymous groups of verbs within these larger subclasses is the task of the visualization system and will be discussed in §4.1 to §4.3 for the different subclasses of motion verbs.

4.1. Subclass 1: Scalar Verbs

As mentioned above, the lexical semantic criterion of scalarity does not imply a common syntactic and event-structural pattern of verbs that are subsumed under this group. The diversity is shown in Figure 2 with the clustering result of the 20 scalar motion verbs in Urdu/Hindi using k-Means with k=3. On the left, the result is shown without the randomization of the data objects, on the right, data points are repositioned to make overplotted objects visible. Figure 2 shows two things: On the one hand, a number of data points are overplotted, illustrated by the comparison between original clustering and repositioned data objects on the left and right of Figure 2. This means that verbs have the same feature vector, i.e. their syntactic and semantic feature structure is fully identical. Upon mousing over the data object, the tool shows that this is the case for the verbs *nɪkal-na* 'to emerge' and *gʰʊs-na* 'to enter' in cluster #1. Investigating cluster #1 further shows that the verb *pahʊnc-na* 'to arrive' differs with respect to its valency (*pahʊnc-na* 'to arrive' is intransitive instead of transitive).

The largest cluster is cluster #2, the group of light green data objects which form a largely homogeneous group: For instance, verbs like *ʊbʰar-na* 'to rise' and *mʊɽ-na* 'to turn' are intransitive in the base form, exhibit a common causativization pattern (direct and indirect causative available) and have the same event structure ([proc]). Four verbs exhibit an exceptional pattern in the cluster, represented by a slightly different position of the data objects: For one, *jʰul-na* 'to swing' (leftmost bottom data point) can license a cognate object, an alternation that is not available for any other verb in the cluster. *a-na* 'to come' and *ja-na* 'to go', represented by the two objects in the upper right corner of cluster #2, cannot have the direct and indirect causative form. *gʊzar-na* 'to cross' differs because the verb licenses a locational oblique. These differences potentially arise from the choice of k in the automatic clustering, however, linguistic information helps to untangle those instances from the larger homogeneous group of verbs.

Cluster #3 with the light blue data points on the upper right is more homogeneous and subsumes transitive motion verbs, in particular *pʰand-na* and *pʰalang-na* 'to leap over', which exhibit the exact same feature structure (indicated by an overplotted data point). *cʰor-na* 'to leave' also belongs to the cluster, but differs in the possibility for causative formation (*pʰand-na* and *pʰalang-na* 'to leap over' only have the indirect causative, whereas *cʰor-na* 'to leave' has none).

Investigating the linguistic properties in the light green and the dark blue cluster shows that these two groups of verbs differ in two major ways: They causativize differently (the verbs in the light green group do not causativize at all, the other cluster has direct and indirect causatives) and they exhibit a difference in valency (*a-na* 'to come' and *ja-na* 'to go' are intransitive whereas verbs like *gʰʊs-na* 'to enter' have a locational oblique).

Overall, the clustering result suggests that there are three subclasses of scalar motion verbs in Urdu/Hindi, separable on the basis of their syntactic and event-structural patterns. Table 2 summarizes the results.

4.2. Subclass 2: Nonscalar Verbs

In contrast to the scalar verbs discussed above, nonscalar verbs and their syntactic feature vectors seem to be best clustered with k=4, as the setting of four clusters yields the most coherent subclass identification for this group of verbs. The clustering result is presented in Figure 3 and summarized in Table 3.

Cluster #1 (light blue data points) on the top of the original and repositioned visualization (left and right of Figure 3, respectively) are representations for the two near-synonyms *kʊcal-na* and *rond-na* 'to trample'. Mousing over the data objects shows that the difference in the feature vector is that *kʊcal-na* has the direct and indirect causative, whereas *rond-na* has neither.

Cluster #2 on the bottom right (dark green cluster of verbs) shows perfect homogeneity: Represented as one data point on the left of Figure 3, with all other data points overplotted, this group of verbs is characterized by the same feature vector and therefore exhibits the same syntactic and semantic properties. The cluster includes verbs like *kud-na* 'to jump' and *mandɛla-na* 'to wander' which are intransitive and cannot undergo the causative alternation. In contrast, the verbs belonging to cluster #3 (light green data objects) are all intransitive, but form two groups: The verbs grouped to the left of cluster #3 are those that have the direct and indirect causative, whereas those grouped to the right only have the direct causative. Randomizing the position of the data objects on the right side of Figure 3 shows that the former group is larger and features 12 verbs, whereas the latter group only comprises two verbs.

Cluster #4 on the bottom left with the dark blue data points consists of three verbs, namely *ʊɽ-na* 'to fly', *doɽ-na* 'to run' and *nac-na* 'to dance'. They differ from the verbs in the light green cluster in that they can have a cognate object,

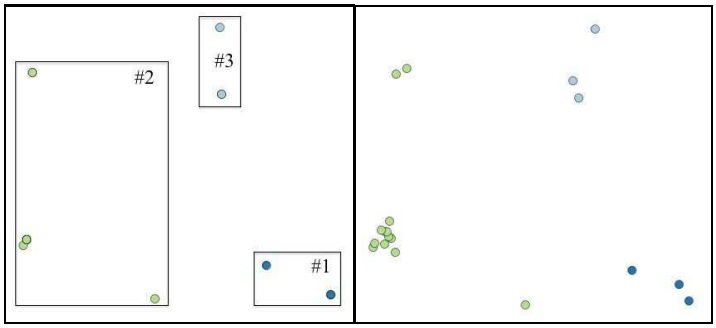

Figure 2: Subclass 1 – Original clusters (left) and repositioned data points (right).

Light green (#2)		Dark blue (#1)		Light blue (#3)	
ʊṭʰ-na	'to rise'	*gʰʊs-na*	'to enter'	*pʰalang-na*	'to leap over'
ʊbʰar-na	'to rise'	*pahʊnc-na*	'to arrive'	*pʰand-na*	'to leap over'
ʊtar-na	'to descend'	*nɪkal-na*	'to emerge'	*cʰor-na*	'to leave'
baṛʰ-na	'to advance'				
gɪr-na	'to fall'				
palaṭ-na	'to turn'				
jʰul-na	'to swing'				
loṭ-na	'to return'				
mʊṛ-na	'to turn'				
ṭapak-na	'to drop'				
gʊzar-na	'to cross'				
a-na	'to come'				
ja-na	'to go'				
Shared properties: Valency, SUBJcase		**Shared properties:** SUBJcase, OBJ/OBLcase, Event		**Shared properties:** Valency, SUBJcase, OBJ/OBLcase, Event	

Table 2: Subclasses of scalar motion verbs in Urdu/Hindi

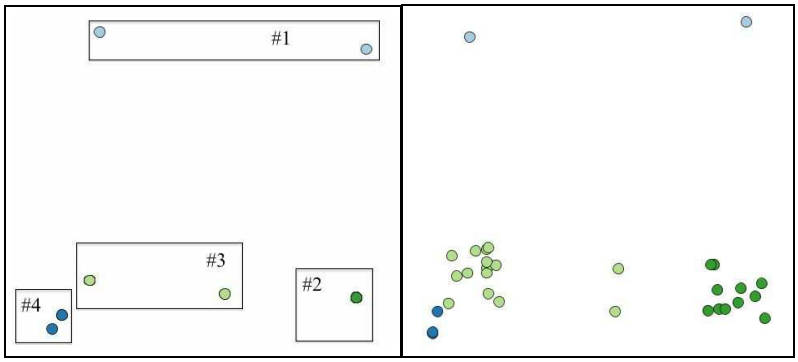

Figure 3: Subclass 2 – Original clusters (left) and repositioned data points (right).

whereas *doṛ-na* 'to run' can additionally have a path object. The behavior in the two object alternations is the only difference in the feature vectors between the two clusters, illustrated by the proximity of the clusters on the canvas.

4.3. Subclass 3: Verbs Lexicalizing Both

An investigation of complex predicates of motion in Urdu/Hindi shows that some verbs can lexicalize scalar as well as nonscalar meaning aspects, depending on the context they are used in (Hautli-Janisz, 2015). The two verbs that belong to this group, namely *cal-na* 'to walk' and *bʰag-na* 'to run', exhibit a very similar feature structure, only that *bʰag-na* 'to run' does not allow for a cognate object construction. Therefore the visualization in Figure 4 shows them as closely positioned data points (Figure 4 only shows the visualization without the randomized data points), with the two verb clusters summarized in Table 4.

Light green (#3)		Dark green (#2)	
bʰaʈak-na	'to go astray'	*tʰarak-na*	'to stomp'
ʈɛhɪl-na	'to lollop'	*ʈʰʊmak-na*	'to strut'
lapak-na	'to dash'	*kud-na*	'to jump'
sɛrak-na	'to slither'	*reng-na*	'to crawl'
pʰɪr-na	'to wander'	*rapaʈ-na*	'to slip'
pʰɪsal-na	'to slip'	*laṛkʰara-na*	'to stumble'
kʰɪsak-na	'to slide'	*ʈʰla-na*	'to strut'
matak-na	'to sashay'	*pʰʊdak-na*	'to hop'
ter-na	'to float'	*langara-na*	'to hobble'
caṛʰ-na	'to climb'	*cakara-na*	'to stagger'
bɛhɛ-na	'to run (water)'	*mandɛla-na*	'to wander'
jʰapaʈ-na	'to scram'	*lɪpaʈ-na*	'to roll'
gʰum-na	'to roll'		
lʊṛʰak-na	'to tumble'		
Shared properties: Valency, SUBJcase, Event		**Shared properties:** SUBJcase, OBJ/OBLcase, Event, Causativization	

Light blue (#1)		Dark blue (#4)	
kʊcal-na	'to trample'	*nac-na*	'to dance'
rond-na	'to trample'	*doṛ-na*	'to run'
		ʊṛ-na	'to fly'
Shared properties: Valency, SUBJcase OBJ/OBLcase, Event		**Shared properties:** SUBJcase, OBJ/OBLcase, CogOBJ, Event	

Table 3: Subclasses of nonscalar motion verbs in Urdu/Hindi

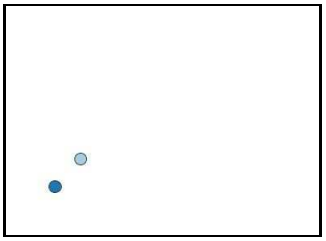

Figure 4: Clusters of motion verbs lexicalizing scalar and nonscalar aspects.

Dark blue		Light blue	
cal-na	'to walk'	*bʰag-na*	'to run'

Table 4: Subclasses of scalar+nonscalar motion verbs in Urdu/Hindi

5. Feature Correlation

Another view on the data is provided by the correlation matrix shown in Figure 5, using the complete set of motion verbs in Urdu/Hindi. This visualization shows how strong the correlation between different syntactic and semantic features is, with the size of the circles representing the correlation strength and the color indicating whether the correlations are negative (white) or positive (black).[3] Figure 5 shows that the features `Caus` and `ICaus` strongly correlate with a strength of 0.873. This means that if a verb features a direct causative, it is very likely that the verb has an indirect causative and vice versa.

Moreover, the `event` structural pattern of a verb is correlated with the case of the object (`OBJcase`) that is licensed. When investigating the relevant cases more closely, it becomes apparent that the [proc, res] verbs *gʰʊs-na* 'to enter' and *nɪkal-na* 'to emerge' have oblique case marking that marks the target and source location, respectively. This

[3] A correlation of 1.0 on the diagonal is expected, because here the features are correlated with themselves.

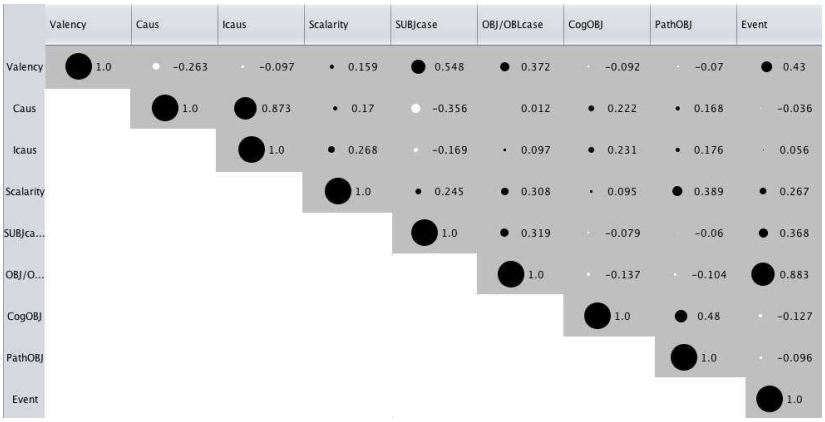

Figure 5: Correlation matrix of the motion verb data

has an impact on the event structure in that the verb licenses a result subevent.

The correlation matrix also brings to light a well-known fact of Urdu/Hindi in that the case of the subject is related to the valency of the verb. A majority of Urdu/Hindi verbs can have ergative case marking when the verb is transitive and in perfective tense, a pattern that "regular" transitive motion verbs like *rond-na* 'to trample sb' adhere to. The reason why the correlation is only 0.548 is that verbs that I have established as being transitive, for instance *gʰʊs-na* 'to enter', in fact license a locational oblique and do not allow for ergative case on the subject.

The correlation matrix shows that in general, syntactic features do not appear independently from each other, although they do not necessarily result in coherent subclasses of verbs. This means that individual features exhibit a certain degree of positive correlation, but those do not exist on a larger scale and across the whole feature set. The correlation matrix therefore complements the visualization on the internal structure of the subclasses of motion verbs in that patterns that hold across subclasses are made visible, allowing for other generalizations than the cluster visualization.

6. Discussion and Conclusion

This paper shows that combining the assumptions from theoretical linguistics with the results of an interactive method of visualizing automatically generated clusters facilitates the structuring of a diverse data set. A manual classification of the verbs based on their underlying syntactic patterns would be unfeasible and ineffective given the large variety of alternation patterns that are present in the verb class. Instead, the visual analytics system helps to automatically establish subclasses of Urdu/Hindi motion verbs.

The drawbacks of using solely automatic clustering for finding subclasses of verbs, without visually interpreting the results, are manifold: First, the results would not provide any insights into how similar data objects in one cluster are, i.e. whether the vectors of those data objects exhibit a great distance to each other or not. Moreover, it would remain unclear as to what vector values trigger differences between clusters and between individual members of one cluster. It would also be impossible to interact with the data and treat known exceptions that were worked out in the theoretical investigation as such.

Another important benefit of the system is the possibility for error detection and correction in the underlying data set. In two cases, the motion verb input file contained a wrong feature value, with the result that the verbs were clustered differently than assumed when compared to similar verbs. Consulting the visualization brought these coding errors to light, preventing an erroneous classification in the resource.

But not only the visualization component provides insights into the structure of the verb class: The information contained in the correlation matrix shows that certain syntactic features do not appear independently from one another, for example the grammaticality of the direct causative positively correlates with the grammaticality of the indirect causative and vice versa. Therefore, the system allows for a well-motivated structuring of the resource as well as the detection of patterns that hold across verb classes and allow for the deduction of further generalizations. Extending the theoretically-motivated subclasses of Urdu/Hindi motion verbs on the basis of scalarity, the visualization has generated a set of classes in each of the subclasses.

In sum, the paper shows that Visual Analytics facilitates "analytical reasoning [...] by an *interactive* visual interface" (Thomas and Cook, 2006) and helps resolving the issue of the "black box of machine learning" by offering a customizable, in-depth view on the statistically generated result of a machine learning technique. In the present case, visualization serves as the bridging element between theoretically well-motivated linguistic analyses and machine learning – one way of seeing the forest *and* the trees.

7. Bibliographical References

Angus, D., Smith, A., and Wiles, J. (2012). Conceptual recurrence plots: revealing patterns in human discourse. *IEEE Transactions on Visualization and Computer Graphics*, 18(6):988–997.

Bhatt, R., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D., and Xia, F. (2009). A Multi-Representational and Multi-Layered Treebank for Hindi/Urdu. In *Proceedings of the Third Linguistic*

Annotation Workshop, ACL-IJCNLP 2009, pages 186–189.

Bhatt, R. (2003). Causativization. Handout for Topics in the Syntax of Modern Indo-Aryan Languages, March.

Bhattacharyya, P. (2010). Indowordnet. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 3785–3792, Valletta, Malta.

Butt, M., Bögel, T., Hautli, A., Sulger, S., and Ahmed, T. (2012). Identifying urdu complex predication via bigram extraction. In *In Proceedings of COLING 2012, Technical Papers*, pages 409 – 424, Mumbai, India.

Butt, M. (2003). The morpheme that wouldn't go away. Linguistics Department Seminar Series, University of Manchester, March.

Card, S. K., Machinlay, J., and Shneiderman, B. (1999). *Readings in information visualization: Using vision to think*. San Francisco: Morgan Kaufmann Publishers.

Collins, C., Carpendale, S., and Penn, G. (2007). Visualization of uncertainty in lattices to support decision-making. In K. Museth, et al., editors, *EuroVis07: Joint Eurographics - IEEE VGTC Symposium on Visualization*, pages 51–58.

Collins, C. (2010). *Interactive Visualizations of Natural Language*. Ph.D. thesis, University of Toronto.

Culy, C., Lyding, V., and Dittmann, H. (2011). Structured parallel coordinates: A visualization for analyzing structured language data. In *Proceedings of the International Conference on Corpus Linguistics*, pages 485–493.

Gold, V., Rohrdantz, C., and El-Assady, M. (2015). Exploratory Text Analysis using Lexical Episode Plots. In E. Bertini, et al., editors, *Eurographics Conference on Visualization (EuroVis) - Short Papers*. The Eurographics Association.

Hautli-Janisz, A. (2014). *Urdu/Hindi Motion Verbs and Their Implementation in a Computational Lexical Resource*. Ph.D. thesis, Universität Konstanz.

Hautli-Janisz, A. (2015). Manner and result in urdu/hindi complex predicates of motion. *Journal of South Asian Linguistics*, 7:39–58.

Heylen, K., Speelman, D., and Geeraerts, D. (2012). Looking at word meaning. an interactive visualization of semantic vector spaces for dutch synsets. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 16–24.

Kachru, Y. (1980). *Aspects of Hindi Grammar*. New Delhi: Manohar Publications.

Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2008). A large-scale classification of english verbs. *Language Resources and Evaluation Journal*, 42(1):21–40.

Kipper-Schuler, K. (2005). *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Lamprecht, A., Hautli, A., Rohrdantz, C., and Bögel, T. (2013). A visual analytics system for cluster exploration. In *Proceedings of ACL13, System Demonstrations*, pages 109–114.

Levin, B. and Rappaport Hovav, M. (2013). Lexicalized meaning and manner/result complementarity. In B. Arsenijević, editor, *Studies in the Composition and Decomposition of Event Predicates*, volume 93 of *Studies in Linguistics and Philosophy*. Dordrecht: Springer.

Levin, B. (1993). *English Verb Classes and Alternations*. Chicago and London: The University of Chicago Press.

Mayer, T., Rohrdantz, C., Butt, M., Plank, F., and Keim, D. A. (2010a). Visualizing Vowel Harmony. *Linguistic Issues in Language Technology*, 4(2):1–33.

Mayer, T., Rohrdantz, C., Plank, F., Bak, P., Butt, M., and Keim, D. A. (2010b). Consonant co-occurrence in stems across languages: Automatic analysis and visualization of a phonotactic constraint. In *Proceedings of ACL 2010: Workshop on NLP and Linguistics: Finding the Common Ground*, pages 70–78.

Mousser, J. (2011). Classifying arabic verbs using sibling classes. In *IWCS '11 Proceedings of the Ninth International Conference on Computational Semantics*, pages 355–359.

Ramchand, G. (2008). *Verb Meaning and the Lexicon: A First-Phase Syntax*. Cambridge: Cambridge University Press.

Rohrdantz, C., Hautli, A., Mayer, T., Butt, M., Plank, F., and Keim, D. A. (2011). Towards Tracking Semantic Change by Visual Analytics. In *Proceedings of ACL 2011 (Short Papers)*, pages 305–310.

Rohrdantz, C., Hund, M., Mayer, T., Wälchli, B., and Keim, D. A. (2012a). The World's Languages Explorer: Visual Analysis of Language Features in Genealogical and Areal Contexts. *Computer Graphics Forum*, 31(3):935–944.

Rohrdantz, C., Niekler, A., Hautli, A., Butt, M., and Keim, D. A. (2012b). Lexical Semantics and Distribution of Suffixes — A Visual Analysis. In *Proceedings of EACL 2012: Joint Workshop of LINGVIS & UNCLH*, pages 7–15, April.

Saksena, A. (1982). Topics in the analysis of causatives with an account of hindi paradigms. In *University of California Publications in Linguistics*, volume 98. Berkeley and Los Angeles: University of California Press.

Schreck, T., Bernard, J., von Landesberger, T., and Kohlhammer, J. (2009). Visual cluster analysis of trajectory data with interactive kohonen maps. *Information Visualization*, 8(1):14–29.

Thomas, J. J. and Cook, K. A. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center.

Thomas, J. J. and Cook, K. A. (2006). A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13.

Ward, M. O., Grinstein, G., and Keim, D. A. (2010). *Interactive Data Visualization: Foundations, Techniques, and Application*. UK: Taylor & Francis Ltd.

Zhao, J., Chevalier, F., Collins, C., and Balakrishnan, R. (2012). Facilitating discourse analysis with interactive visualization. *IEEE Tranactions on Visualization and Computer Graphics*, 18(12):2639–2648.

Visual Analytics for Distributional Semantic Model Comparisons

Thomas Wielfaert, Kris Heylen, Dirk Speelman & Dirk Geeraerts
KU Leuven - Quantitative Lexicology and Variational Linguistics
Blijde-Inkomststraat 21 b3308, 3000 Leuven, Belgium
firstname.lastname@kuleuven.be

Abstract

Distributional semantic models have shown to be a successful technique for Word Sense Disambiguation and Word Sense Induction tasks. However, these models, and more specifically the token-level variants, are extremely parameter-rich. We are still in the dark on how the different parameters can be efficiently set and even more on how to evaluate the outcome when no gold standard is readily available. To gain a better insight, we are developing a visual analytics approach which shows these models in two ways: a scatterplot matrix for inter-model parameter comparison and zoomable individual scatter plots allowing for more details on-demand. More specifically, we first use a scatterplot matrix to compare models with different parameter settings in a single view. This enables us to track selections of tokens over different models. On top of this, we create a scatter plot for each individual model, enriched with both model dependent and model independent features. This way, we can have a more in-depth visual analysis of what is going on and visualise the distinct properties or parameters of the individual model.

Keywords: distributional semantics, corpus linguistics, visual analytics, parameter evaluation, dimension reduction

1. Introduction

Over the past decade, word space or distributional semantic models (DSMs) have become the backbone for large-scale statistical analysis of word meaning (see Turney and Pantel (2010) for an overview). The idea behind distributional semantics comes from the Firthean idea that words occurring in similar context have similar meanings (Firth, 1957). Distributional models have been successfully used for Word Sense Disambiguation, Word Sense Induction tasks and Lexical Substitution tasks, most notoriously in the context of SemEval[1], an ongoing series of evaluations of computational semantic analysis systems.

What these classification tasks have in common is an optimal classification or gold standard, provided by the organisers, which participants have to use to evaluate their models against. Model performance is subsequently reported in the form of measures like precision and recall or both combined in an F-measure or F1-score. What these measures can not tell however, is what went actually wrong. Error analysis remains a manual effort that can only be done by digging into the model, which is tricky when the model is a black box. Furthermore, when comparing different models, precision scores do not tell which part of the data was classified correctly. In theory, it is even possible that two competing models both attain a precision of 50%, without an overlap in the data they classified correctly.

Distributional semantics is nevertheless also used for purposes where an a priori categorisation is not available. Word Sense Induction (WSI), as token-level distributional models are called in Computational Linguistics, is in theory completely unsupervised. In practice however, it is often treated as a Word Sense Disambiguation task with predefined senses as a gold standard. When using unsupervised word sense induction for lexical semantic purposes, there is no other way than manually going through the occurrences of the target word to see the semantic patterns, if any, unveiled by the model. Treating these models as Word Sense Disambiguation requires that the output is additionally submitted to a clustering algorithm. This additional processing layer makes these models even more opaque.

For the visualisation of our corpus data however, we have been using a hybrid approach with manually assigned word sense labels, avoiding separate clustering algorithms. We visualise the distributional models and make use of visual analytics to get a better view of what is going on. Baroni and Lenci (2011) and Lenci argued that *[t]o gain a real insight into the abilities of DSMs to address lexical semantics, existing benchmarks must be complemented with a more intrinsically oriented approach, to perform direct tests on the specific aspects of lexical knowledge captured by the models.* We believe that visual analytics can be part of this "more intrinsically oriented approach"; we want to facilitate parameter optimisation by visual means.

As the exact type of distributional model does not matter for our approach, we will limit ourselves to a single type of token-level distributional model applied on Dutch newspaper material. However, our method is meant to be generalisable for all distributional semantic models and all languages. As long as tokens vectors are compared by similarity or distance, which these models do almost by definition, their output could be plugged in our visualisation. The goal of our tools is two-fold: first, we can use data visualisation to compare different models in a way which is both practical and attractive. This also allows us to do what we from now on will call *inter-model parameter comparison* and see how changing properties of the model changes the results visually. Second, we make interactive visualisations of individual models, giving a more zoomed-in view on the token properties, so each single model can equally be explored in a meaningful way.

The remainder of this paper is structured as following: first we briefly introduce the dataset. Next, we introduce our bag-of-words token-level distributional model. Both visualisations (inter-model comparison and individual model)

[1] http://siglex.org/

are discussed in the subsequent parts. The last section provides a more general discussion and some elements for future work.

2. Corpus and Target Noun Selection

We created small set of Dutch nouns by carefully selecting 9 polysemous and 1 homonymous noun ($koper$) from *Algemeen Nederlands Woordenboek* (General Dutch Dictionary), a fairly new free online Dutch dictionary[2] which contains information about the semantic relationship between the different senses of a word. We limited ourselves to words with 2 so-called core senses:

Dutch noun	Translation
koper	buyer / copper
match	football match / correspondence
motor	engine / motor cycle
parachute	idem.
piraterij	piracy
pony	pony (animal) / pony(tail)
prof	professional / professor
scout	scout (sport) / Boy Scout
therapeut	therapist
varken	pig

Table 1: The selected nouns from the ANW

Random samples for these target words were drawn from two large Dutch newspaper corpora: Leuven News Corpus (LeNC), 1.3 billion words from Belgian Dutch (Flemish) newspaper articles between 1999 and 2003 and the Twente News Corpus (TwNC) (Ordelman, 2002), about 500 million words, with newspaper material from the Netherlands from 1999–2004. The corpora were lemmatised and part-of-speech tagged and syntactically parsed with the Alpino parser (van Noord, 2006). Furthermore, we manually disambiguated random tokens until we reached a lower threshold of 100 tokens per sense, so that we would get close to a total of 250 tokens per noun.

3. Token-level DSM: Bag-of-words Model

Distributional Semantic Models, Word Space Models or Semantic Vector Spaces exist in many flavours and varieties. We will very briefly explain here an adapted version of one of the earlier models, namely Schütze (1998)'s bag-of-words model. This model is no longer considered state-of-the-art in distributional semantics, but its intuitive ideas make it attractive for lexical semantic purposes and to use it as a generic baseline model to experiment with. One of the main problems with token-level models is that filling token vectors with raw co-occurrence frequencies will lead to data sparsity. Suppose the corpus contains 10.000 types use a context window of 10 words, 5 left and 5 right of the target. This would mean that in a best case scenario 10 cells out of 10.000 (0.1%) of the vector would be filled with actual frequencies and the other 99.9% with zeros, and thus be very sparse making vector comparison mathematically intractable. Schütze's insight is that we can overcome this problem by moving on to so-called second-order co-occurrences. These second-order co-occurrences are the type-level context features of the (first-order) context words co-occurring with the token. This way, we still model the tokens by their "co-occurrences", but these are no longer the direct collocates. Note that the first-order model still has to be created to construct the second-order model.
Following Schütze, for each token we normalise the frequencies by the number of context words for that token:

$$o_i^{\vec{w}} = \frac{\sum_{j \in C_i^w}^{n} \vec{c_j}}{n}$$

where $o_i^{\vec{w}}$ is the token vector for the i^{th} occurrence of word w and C_i^w is the set of n type vectors $\vec{c_j}$ for the n context words in the window around that i^{th} occurrence of noun w. However, this summation means that each first-order context word has an equal weight in determining the token vector. Yet, not all first-order context words are equally informative for the meaning of a token. In a sentence like "While walking to work, the teacher saw a dog barking and chasing a cat", $bark$ and cat are much more indicative of the meaning of dog than say $teacher$ or $work$. In a second, weighted version, we therefore increased the contribution of these informative context words by using the first-order context words' PMI (Pointwise Mutual Information) or LLR (Log-Likelihood Ratio) values with the target noun. PMI and LLR can be regarded as measures for informativeness and target-noun/context-word weights were available already from our large type-level model. The PMI or LLR of a word w and a context word c_j can now be seen as a weight $pmi_{c_j}^w$ or $llr_{c_j}^w$. In constructing the token vector $o_i^{\vec{w}}$ for the ith occurrence of noun w with PMI as weight, we now multiply the type vector $\vec{c_j}$ of each context word with the PMI weight $pmi_{c_j}^w$, and then normalize by the sum of the PMI-weights:

$$o_i^{\vec{w}} = \frac{\sum_{j \in C_i^w}^{n} pmi_{c_j}^w * \vec{c_j}}{\sum_j^n pmi_{c_j}^w}$$

When this procedure has been repeated for each token, we get a token by second-order co-occurrence matrix. By computing the cosines between these token vectors, we obtain a similarity matrix[3].
Next, to be able to visualise this high-dimensional token similarity matrix we need a dimension reduction algorithm to turn this matrix into 2D coordinates. One of the traditional algorithms used for this kind of dimension reduction (similarity/distance matrix to coordinates) is non-metrical Multidimensional Scaling (MDS) (Cox and Cox, 1991)[4].

4. Inter-model Parameter Comparison: Scatterplot Matrices

When training token-level distributional models, certain parameter choices have to be made. In the previous sec-

[2] http://anw.nl

[3] For the more thorough, full explanation on how the weighted Schütze bag-of-words model works, we would like to refer the interested reader to the first sections of Heylen et al. (2015).

[4] Dimension reduction was done in R with the *MASS* package.

tion we briefly mentioned the choice between two collocational measures (PMI and LLR) to assign weights to context words. To see how distributional models behave with different parameter settings, we plugged their 2D coordinates in a so-called scatterplot matrix. Traditionally, scatterplot matrices are used to visualise high-dimensional data sets by visualising feature pairs in a separate scatter plots, generating a matrix with multiple yet parallel views. For model comparison however, we use the scatterplot matrix to compare similarity matrices from different models containing exactly the same data. The true comparative power of a scatterplot matrix, we believe, lies in addition of the so-called brushing and linking functions (Cleveland and McGill, 1988; Buja et al., 1991). This means that an area of interest (a group of tokens in this case) can be selected in one of the plots from the matrix with the brush. Consequently this selection appears in the other plots, which makes it easy to track specific tokens and verify whether patterns persist between different models. Furthermore, as Multidimensional Scaling is prone to rotations, brushing and linking helps to keep track of the selected tokens throughout the different rotations of the plot.

To demonstrate the principle, we have created 6 token models with different parameter settings and put these together in a scatterplot matrix.[5] We opted to demonstrate the weighting parameters (of collocates) for the Schütze bag-of-words model here: three measures for collocational strength: Positive Pointwise Mutual Information (PPMI) and two variants of Log Likelihood Ratio (LLR), dubbed L1LLR and L1LLRx5. Next to this, the window size for the weights of the second-order co-occurrences was varied: 4-4 and 7-7, meaning including 4 or 7 words left and right of the target.

The scatterplot matrix can be explored in two ways: on a token-level, where one can click on the individual tokens which results in the context snippet being shown next to the plot. This can give the user an idea of which kind of contexts are in which area of the plot. The second way however, which is more innate to the principle of a scatterplot matrix, is of course the above mentioned brushing and linking. The brush can be activated with a checkbox in the upper-left corner of the screen. Subsequently, one can select one or multiple tokens with a rectangular brush in one of the plots. As a result, the selected tokens are highlighted while the others turn gray. Most interestingly of course, the selection is automatically linked to the other plots in the matrix. In other words: you can select tokens and see whether they cluster together over different parameter settings.[6]

When we take a look at the scatterplot matrix for $piraterij$ (piracy) in Figure 1, we can see that the two senses (distinguished by colour) are probably best distinguished by the *LeNC77.ppmi* model. The orange coloured tokens are representing traditional piracy at sea, the blue ones refer to internet piracy. When brushing over the lower left area as shown in Figure 2, which is entirely populated by blue, we can see that these do not cluster together in any of the other plots in the matrix. This gives an indication that the selected model captures this specific sense distinction at least slightly better than the others do.

[5] All visualisations were made in Javascript, using the D3 library by Mike Bostock https://d3js.org

[6] It is impossible to do honour to interactive visualisations with still images. Therefore we would like to encourage the interested reader to explore our plots on https://tokenclouds.github.io

5. Token Properties: Individual Scatter Plots

Individual token properties can more easily be shown if we move on to a more zoomed-in, detailed level. Therefore, we also made an interactive scatter plot or "token cloud" for each model. These plots are enriched with model independent token properties derived from the corpus (i.e. newspaper title, country, year, etc.) and a manually assigned sense label, as well as parameter influenced properties (i.e. the weights assigned to the context words of a token). The plots are zoomable by double-clicking so one can see the densely populated areas better.

Looking at a scatter plot of the Dutch word $piraterij$ in Figure 3, we see 236 disambiguated tokens. Remember the orange tokens represent piracy at sea and the blue ones internet piracy. The corpus dates from a period when (illegal) file sharing services such as Napster became widely known. It is easy to spot that the internet piracy sense is definitely more frequent in our sample. This time, one can just hover over the individual tokens, after which a so-called tooltip to appears, showing the wider context of the token. Furthermore, it is noteworthy that the glyphs (tokens) in the plot can be set to differ in size. In this case, size depends on the maximum weight assigned to the token's first-order context words. This means that the model has relatively little information about the tokens have a relatively small surface. In other words: this shows us in which specific cases our model fails to distinguish our predefined senses, in this case most of these tokens are found around the origo of the plot. It also means that relatively little importance should be given to the position of tokens with low weight, as it is rather unlikely that they end up in an area of the plot where they belong, surrounded by semantically related tokens.

6. General Discussion

We have shown two distinct yet related visualisation types which allow to compare different models and zoom in on the details. Although we have used a basic bag-of-word distributional model with just a few parameter varied, it is not hard to imagine how the inter-model parameter space can quickly explode. When distributional models are used in a classification task such as WSD, the pre-existing classification (gold standard) could easily be crossed with the clustering results, visually highlighting (with the appropriate visual variable) misclassified items. On the other hand, when there is no prior classification, the scatterplot matrix is merely a visual analytics tool to compare different models and thus a way to evaluate parameter settings through visual analytics. Nevertheless, there are undeniable limitations to the visual analytics approach for inter-model comparison. With just 6 models, the scatterplot matrix is already screen-filling and we believe it is unrealistic to make the matrix larger than 3 x 3. Furthermore, it becomes unfeasible to use anything but colour to visually encode features

as it would make the scatter plots too dense to remain informative.

Two important issues remain unresolved at this point however. First, the role of the dimension reduction algorithms should definitely be further investigated. Dimension reduction inevitably leads to information loss and the introduction of noise. The latter can become really problematic when it creates artifacts that can be interpreted as meaningful patterns or structures (see also Chuang et al. (2012) for a discussion on the complex interaction between model, visual interface and user evaluation). In this paper, we have used Multidimensional Scaling (MDS) because its speed and robustness. However, MDS has its shortcomings: in our experience it can not handle more than approximately 300 data points and it is very sensitive to rotations. Nowadays a different dimension reduction technique called t-distributed Stochastic Neighborhood Estimation (t-SNE) (Van der Maaten and Hinton, 2008) is considered state-of-the-art, both in and outside the distributional semantics community. The advantage is that it can handle much larger datasets. However, when we tried the algorithm on our data, we got very unsatisfying results[7]: low weighted tokens were spread all over the plot instead of being located around the origo of the plot and the senses were no longer separated in any meaningful way. If we wish to use t-SNE for these type of visualisations, we will first need to gain more insight on how it behaves with our data. Multidimensional Scaling on the other hand has its proper evaluation function in the form of so-called Shepard Diagrams. These are scatterplots show the input and output of the dimension reduction relate to each other, for each individual point. As such, this is also a parameter that could be visually encoded in the plots. We will leave a more formal evaluation of these dimension reduction techniques for future work as it adheres to a different research question that is beyond the scope of this paper.

Second, word sense induction models are meant to work completely unsupervised, meaning that there is no predefined number of senses nor labels. Nevertheless we introduced manually obtained sense labels to visually distinguish the different senses, which is a labour-intensive task. The ultimate goal is of course to be able to skip this manual endeavour so we can tackle much larger data sets. This will only be possible however if we manage to control the process better and be able to visually (or otherwise) detect the automatically induced senses.

7. Acknowledgements

This research was funded by KU Leuven BOF grant 3H110243 for the OT project *From lexical to semantic sociolectometry: New methods for the corpus-based analysis of variation in lexical categorization*. We would like to thank Chris Culy for his input and the two anonymous reviewers for their interesting and useful comments.

8. Bibliographical References

Baroni, M. and Lenci, A. (2011). How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, GEMS '11, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Buja, A., McDonald, J. A., Michalak, J., and Stuetzle, W. (1991). Interactive data visualization using focusing and linking. In *Visualization, 1991. Visualization'91, Proceedings., IEEE Conference on*, pages 156–163. IEEE.

Chuang, J., Ramage, D., Manning, C., and Heer, J. (2012). Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 443–452. ACM.

Cleveland, W. C. and McGill, M. E. (1988). *Dynamic Graphics for Statistics*. CRC Press, Inc., Boca Raton, FL, USA, 1st edition.

Cox, T. F. and Cox, M. A. (1991). Multidimensional scaling on a sphere. *Communications in Statistics-Theory and Methods*, 20(9):2943–2953.

Firth, J. R. (1957). *Papers in linguistics, 1934-1951*. Oxford University Press.

Heylen, K., Wielfaert, T., Speelman, D., and Geeraerts, D. (2015). Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis. *Lingua*, 157:153–172.

Ordelman, R. (2002). Spoken document retrieval for historical video archives - dutch speech recognition in the echo project. Technical report, University of Twente, Parlevink Group.

Schütze, H. (1998). Automatic word sense discrimination. *Comput. Linguist.*, 24(1):97–123, March.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *CoRR*, abs/1003.1141.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85.

van Noord, G. (2006). At Last Parsing Is Now Operational. In *Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles (TALN06)*, pages 20–42, Leuven, Belgium. Presses universitaires de Louvain.

[7] An example of such a t-SNE visualisation can be found on https://tokenclouds.github.io/piraterij/piraterij.html

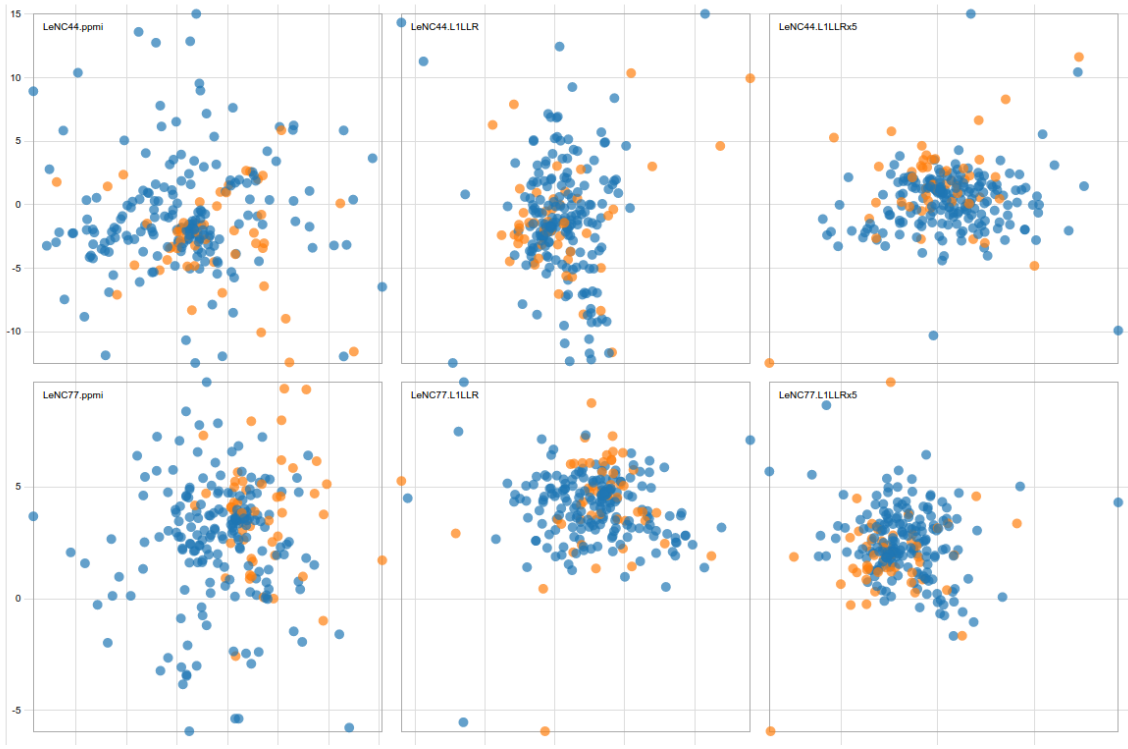

Figure 1: Scatterplot matrix comparing 6 token clouds

Figure 2: Same as above, with bushing and linking

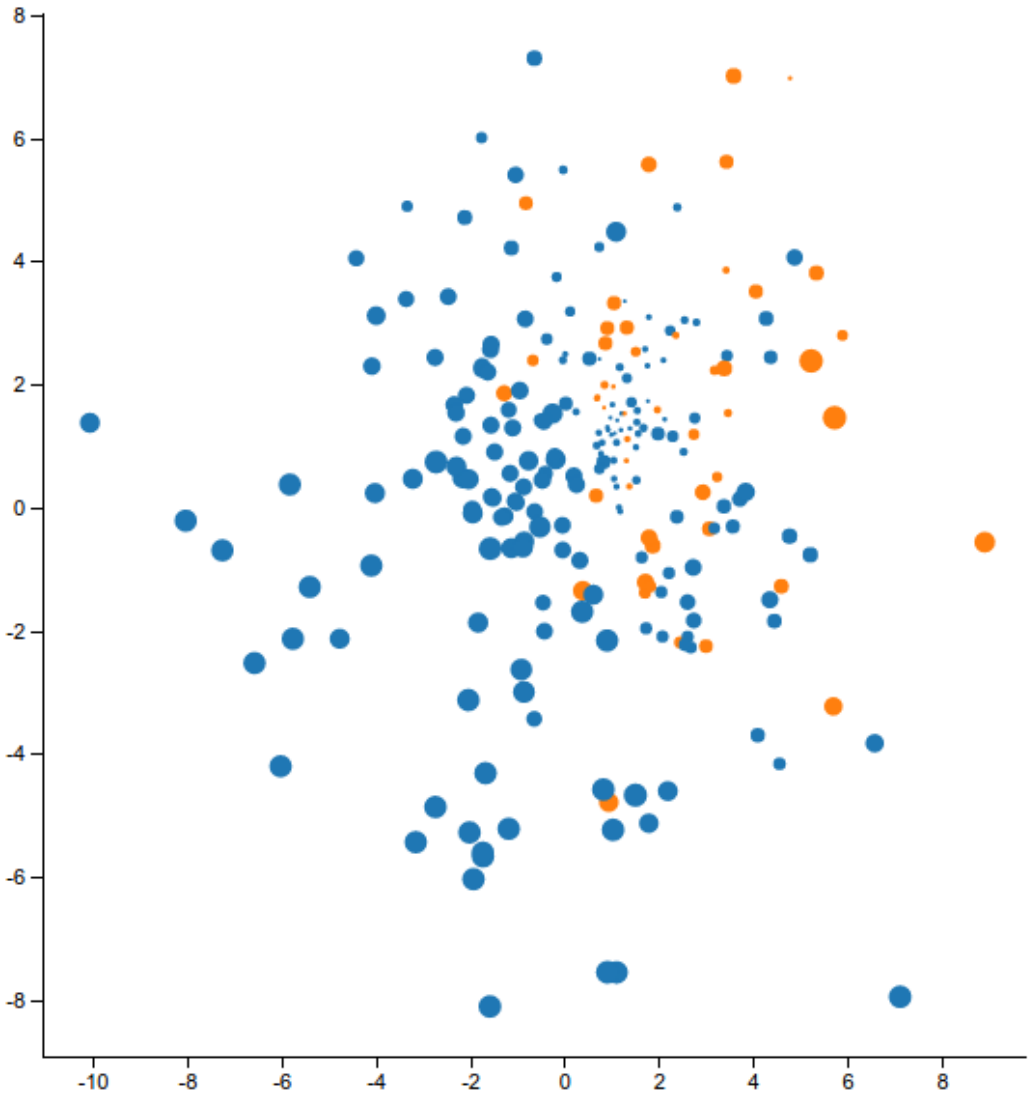

Figure 3: $piraterij$ tokens (MDS dimension reduction)

Visualizing Literary Data

Erik Tjong Kim Sang
Meertens Institute Amsterdam
Joan Muyskenweg 25, 1096 CJ Amsterdam, The Netherlands
erik.tjong.kim.sang@meertens.knaw.nl

Abstract

We look at different aspects of Dutch magazines, both from the fields of literary studies and linguistic studies. We explore the background of authors with respect to birth locations, ages and gender, and also in how language use in the magazines evolved over a period of several decades. We have created several interactive visualizations which enable researchers to browse and analyze text data and their metadata. The design of these visualizations was nontrivial: invoking questions about how to deal with missing data and documents with multiple authors. The data required for some of the visualizations useful for researchers, were infeasible for the software architecture to generate within a reasonable time-span. In a case study, we look at some of the research questions that can be answered by the data visualizations and suggest another data view that could be interesting for literary research. Interesting topics for future research rely heavily on improvements of the search architecture used and including extra annotation layers to our text corpora.

1. Introduction

Magazines are an interesting subject of study for both literary and linguistic researchers because they represent the spirit of time, both with respect to the topics covered and the language variants used. Magazine texts are even more valuable for research when they are available for long spans of time, like decades or even centuries, and when they are accompanied by metadata, covering aside from publication times information about the authors of the magazine articles, like their birth location, age and gender.

For the language Dutch, which is spoken in The Netherlands and the northern half of Belgium, a large collection of magazines is available in the Digital Library for the Dutch Literature (DBNL[1]), which contains digital versions of Dutch texts from the thirteenth century until today. The magazine collection includes the literary magazine De Gids, of which about 170 digital yearly editions are available, dating back to 1837. DBNL also offers biographical information about authors like dates and locations of birth and death.

The present (2016) websites of text collections like DBNL make it possible to search and find text segments that contain specific words or word variants. For researchers in the humanities this is not sufficient to answer their research questions. They frequently want to quantify and compare search results. For example, they would like to know how the frequency of a certain word in the seventeenth century compares with its frequency in the eighteenth century. Or they want to know if a gender and age bias which is present among a certain magazine, has disappeared in a new magazine that is a spin-off of this magazine. It is difficult, if not impossible, to answer these questions with a search interface which only returns text snippets as search results.

For such research goals, we developed interactive visualizations of text search results. The visualizations have two important goals. First, by quantifying the search results and presenting the numbers visually, they aim at improving the quality of the analyses that can be performed of this data. Second, by offering researchers the possibility to select sections of visualizations, they offer the opportunity to further zoom in into the search results and thereby increase the possibilities to browse the data.

We aim at visualizations which are as simple as possible and which give quick insights in certain aspects of the data. However some users will be interested in a more elaborate analysis of the data. Rather than making the visualizations more complex for all users, we aim at satisfying the information need of these users by enabling them to download the counts on which the visualizations were based.

After this initial introductory section, this paper contains five more sections. In the next section two, we will describe some related work. In section three we will outline the technical infrastructure required for being able to retrieve quantitative search results from our text corpus. In section four, we present the interactive visualizations we offer for exploring and analyzing the data. Section five contains a case study with these visualization. In section six, we conclude.

2. Related Work

Modern data visualizations can be traced back to the eighteenth century (Tufte, 2001). Back then, the prime data format used in visualizations were numbers, which will also be the main data format that we will work with. A useful library for numeric visualization, which we will use, can be found on the website `d3js.org`. However, our data is also suited for more text-friendly variants of visualizations, like word clouds, created by Jim Flanagan around 2002, and for geographic visualizations, like the visualizations offered by the website `openlayers.org`.

The source of the text data that we work with for this paper, is the website `dbnl.nl`. Earlier, a one-time analysis of two Dutch magazines present in DBNL was done and static comparisons and visualization of biographical author data were created (Zhang et al., 2013). We will use a vocabulary comparison method applied earlier to differences between standard Dutch and Surinamese Dutch, the variant of Dutch spoken in the South American country Suriname (Tjong Kim Sang, 2014): t-score (Church et al., 1991). We were also inspired by the Google Books Ngram Viewer (Google,

[1] dbnl.nl

```
{
 "condition": {
  "type": "equals",
  "field": "NLTitle_title",
  "value": "Amsterdam"
 },
 "response": {
  "facets": {
   "facetranges": [
    {
     "field": "NLTitle_yearOfPublicationMin",
     "start": "1600",
     "end": "2000",
     "gap": "100",
     "ex": "persons"
    }
   ]
  }
 }
}
```

Figure 1: Example of faceted search. This query looks for all documents with the word *Amsterdam* in the title and presents document counts per century in the results.

```
{
 "status": "ok",
 "facets": {
  "facetranges": {
   "NLTitle_yearOfPublicationMin": {
    "counts": [
     "1600", 9,
     "1700", 229,
     "1800", 61004,
     "1900", 127
    ],
    "gap": 100,
    "start": 1600,
    "end": 2000
   }
  }
 }
}
```

Figure 2: Output of the query presented in Figure 1: a list of document counts per century.

2010), which compares word frequencies of different years. Since we have centuries of text available, we should be able to generate similar graphs. Parallel to our work, people are working on creating visualizations of historical Dutch text data with the statistical package R (Komen, 2015).

3. Technical Infrastructure

Currently many people have experience in working with information extraction systems and therefore users have high expectations of search systems. Our user group, researchers in the humanities, expects to be able to search in large text collections and get adequate results in a fraction of a second. The technology for quickly retrieving relevant text snippets and document counts is available, for example in Apache Lucene and ElasticSearch (Gormley and Tong, 2015), but we need more. The visualizations which we aim at require faceted search, that is search in which multiple filters are combined, and segmented output, output that is divided in several sections, for example corresponding with different time segments.

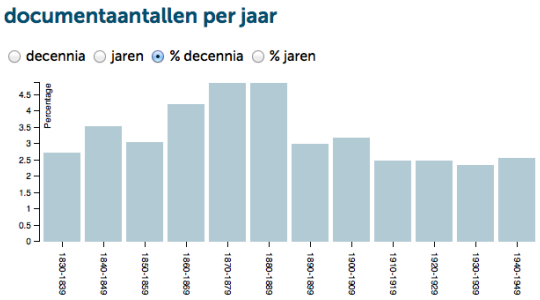

Figure 3: Visualization displaying the percentage of documents per decade containing a certain word. Users may select different output modes: absolute counts per year or decade, or relative counts (percentages) per year or decade.

An engine for faceted search for our data was created in 2013 (Brouwer et al., 2013) in the Nederlab project (Brugman et al., 2016). The data is stored in several Apache SOLR indexes. An interface layer called the broker performs the interaction between the user and the indexes. This broker translates user queries written in JSON to one or more commands for the indexes. The index responses are analyzed by the broker and sent back to the user. Examples of a user command and the broker output for this command can be found in Figures 1 and 2, respectively. The query looks for documents with the word *Amsterdam* in a document title and requires the results to be presented in matching document counts per century. A lot of processing time is saved because the combination the broker and the indexes perform aggregating different relevant counts internally.

4. Visualizations

We compiled a list of data visualizations which could be useful for humanities researchers while browsing and analyzing text collections. We used JavaScript in combination with the visualization library D3 (Bostock, 2011) for implementing these online visualizations. The communication between the JavaScript code and the data run via the broker interface described in section 3..

4.1. Documents Counts per Time Unit

Our first goal was to create a visualization of the number of times that a word or phrase is used over time. However, the broker interface does not return word counts, apparently because of limitations of the underlying information system architecture: Solr and Lucene (Brouwer et al., 2015). Therefore, the only way to obtain the relevant numbers would be to retrieve all relevant documents and perform the word counting when the user requests this. This could lead to unacceptable time delays in the processing of queries and for this reason we have refrained from creating this visualization mode. However, it is on the top of

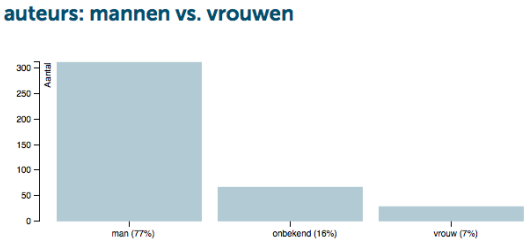

Figure 4: Bar plot showing the gender of the authors of documents that matched a query. There are three values: male (*man*), unknown (*onbekend*) and female (*vrouw*). The counts are document-based so authors that wrote multiple matching documents will be counted multiple times. Also, documents with several authors will contribute to the counts several times. Gender value unknown means either than the gender of an author is unknown or that the author of a document is unknown.

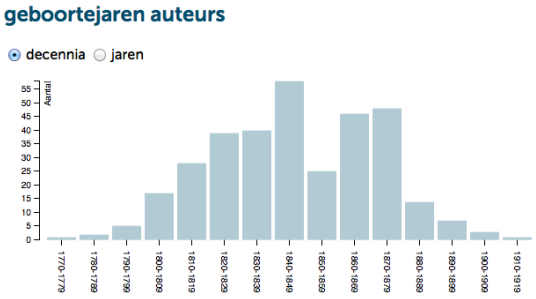

Figure 5: Bar plot displaying the number of times authors involved in documents matching a query were born in certain decades (*decennia*). Users may choose a plot with year (*jaren*) bars instead. The counts are document-based, so authors that are involved in multiple documents matching the query will be counted multiple times.

our wish list regarding extensions of our system, as many users have expressed their interest in obtaining overviews of word counts.

As an alternative for the word count visualization, we created a document count visualization, showing how many documents contain a certain word or phrase at least once, within a certain time frame. As basic time unit we chose one year but we also offered the possibility to examine the data in chunks of ten years, thus providing a smoother view for data that are available for longer time periods.

Figure 3 is an example of of a bar plot showing the percentage of documents that contain a certain word per decade in the period 1830-1949. In order to find the exact number associated with a bar, users can hover over the bar after which a small pop-up window will present the time period and the associated value. Furthermore users can choose between displaying absolute document counts per time frame and percentage counts. Since our document collection is all but balanced with respect to time, with most documents originating from recent times, the latter option often produces a fairer view.

Percentage counts in comparison with all available corpus material are not sufficient to answer all research questions. For example, researchers my want to know what percentage of articles of a magazine contained a certain word in various time periods. Or they could be interested the percentage of articles of a certain magazine written by female authors. Currently we compute percentage scores by comparing query results with overall collection counts, which works fine for studying magazine texts. However, a more general solution in which two arbitrary queries are compared, is preferable.

Clicking on a bar will impose a time-restricted filter on the query related to the time frame corresponding to the bar. For example, selecting the decade bar 1920-1929 for documents written by female authors of a certain magazine will restrict the query results to that time period. This way of interacting with the visualization enables users to quickly inspect several aspects of the corpus material.

The present visualization interaction allows users only to create time filters which correspond with a single bar in the graph. In order to select arbitrary time frames for inspection, we have experimented with two sliders under the graph: one restricting the number of years displayed in the graph and one for changing the start and the end time of the graph. However, since our users found the interaction with the sliders complicated and confusing, we have removed them from the visualization.

4.2. Gender of Authors

Next, we created a bar plot with counts of the gender of authors. Since our metadata is incomplete, we have three gender values: male, female and unknown. We needed to make a choice about how to count the genders: author-based, where each author is counted only once, or document-based, where authors of of documents that appear several times in the query result, are counted several times. We chose the second alternative because this provided the fasted query responses of the broker[2].

We also needed to choose how to deal with documents of which the authors were unknown. There were two ways to handle this: we could leave the unknown authors out of the visualization or we could assume documents with unknown authors were written by an unknown author with an unknown gender. After consulting with our users we chose the second option. This means that the bar for the value unknown in the graph represents counts for two different cases: a known author with an unknown gender or an unknown author.

Figure 4 is an example of a bar plot containing gender counts. There are three values: male (*man*), unknown (*onbekend*) and female (*vrouw*). The bars represent absolute counts but bar-count-based percentage counts have been added below the graph. Since documents can have multiple authors, the author counts may exceed the number of documents. This may confuse the user. An alternative would be to use fractions for authors of shared publications, like 0.75 and 0.25 for a document with three male and one female

[2]In more recent implementations, the broker responce speed is not influenced by the format of the gender visualization.

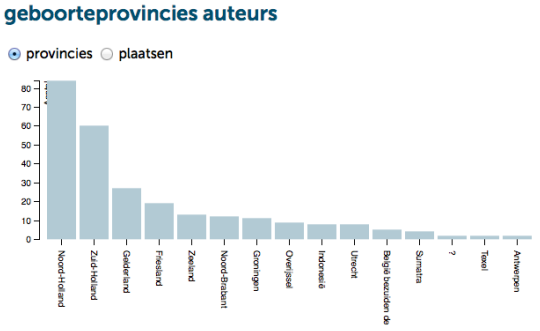

Figure 6: Birth provinces (*provincies*) of authors matching a certain query. Users can also select birth places (*plaatsen*). Provinces of birth and death include local and foreign ones. The question mark is nonstandard content of the birth location field. Provinces are sorted from the most frequent (left) to the least frequent (right). Provinces with a count of zero are not shown.

author. However, needing to deal with fractions of author could be difficult to explain to the users as well.

Like in the document count visualization, moving the mouse over a bar will display the gender and the exact count. Clicking on a bar will restrict the query results to that particular gender. However, here the fact that documents can have multiple authors causes unexpected query results. For example, a query for female authors usually generates a result including some male and unknown authors. The reason for this is that queries are document-based and documents written by females may include coauthors which are male or of which the gender is unknown.

The gender bar plot was an interesting case in which we found out that a seemingly basic view of the data, turned out to be far more complicated than was expected. One reason for this is that we had to work with a search architecture which did not always provided the numbers that we needed. The other was that our interpretation of what was the optimal way to visualize the numbers was sometimes different of the expectations of the users. The current visualization is the most simple one that our users could agree with[3].

4.3. Birth and Death Year of Authors

Bar plots for birth years or death years, can be drawn in the same way as in the visualization for document counts (see section 4.1.). Like in the visualization for gender, birth and death years of authors that are associated with more than one document in the query result, will be counted multiple times.

Figure 5 presents an example of a birth year visualization. Only absolute counts are visible in the plot. Percentage counts could be added in the same way as in the gender plot (Figure 4), by dividing the counts to the total of the

[3]The current gender visualization does not offer options for alternative data interpretations but a valid case could be made to offer choices between counting authors of multiple documents more than once or not and counting every author of a shared document once or only by a fraction.

Figure 7: Bubble map representing the birth locations of authors writing in the Dutch magazine De Gids. The number is each bubble represents how often an author of that city or region was involved as an author of an article in the magazine. Bubbles with numbers are combined when the user zooms out from the map, creating a regional overview. For example, the bubble with 226 represents the two cities The Hague (118) and Rotterdam (108). Together with Amsterdam (311), these cities are the most common birth places of authors of De Gids.

counts in the graph. Exact counts related to a bar can be inspected by moving the mouse pointer over the bar. Clicking on a bar repeats the query with an extra filter restricting the birth year/decade or the death year/decade of the author to the value corresponding to the bar. Because the search is document-based and documents may have more than one author, a birth/death time restriction may produce out-of-range results which correspond to coauthors of documents with matching authors.

4.4. Birth and Death Location of Authors

Birth and death locations can be visualized in the same way as birth years: with a bar plot. Figure 6 presents an example of a visualization of birth provinces. The provinces are sorted from the one with the highest count (left) to the one with the lowest count (right) while ignoring provinces which did not match the query. Provinces can be both local (i.e. located in The Netherlands or Belgium) or foreign. The question mark in the graph is a nonstandard way of indicating that the birth location of an author is unknown, usually the field is left blank. Users can choose to view counts for birth places rather then birth provinces. Like in the other visualizations, moving the mouse pointer over a bar will evoke a pop-up window showing the exact count corresponding to the bar. Clicking on a bar will start a new query with an extra filter restricting the birth province of the authors to the value corresponding to the bar. Again this

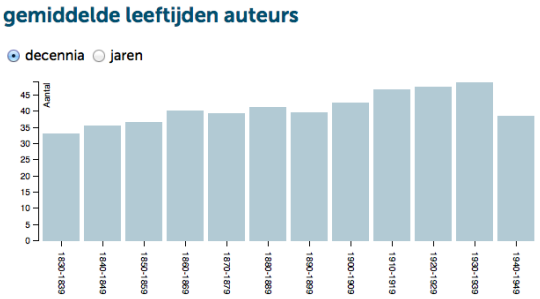

Figure 8: Visualization showing the average age of authors of a certain magazine per decade. The average age seems to move up slowly from 1830-1839 (33) to 1930-1940 (50) after which there is a big drop for the war decade 1940-1945 (38). Users can choose to view a bar plot for years (*jaren*) rather than for decades (*decennia*).

Figure 9: Word cloud for the query *Havelaar*, a novel title and a character name, in the articles of the Dutch literary magazine De Gids. The size of the font used for a word indicates its frequency in the query results. Striking words in the cloud are the names of the author of the novel: *Multatuli* and *Dekker*, as well as words common to literary works: *literatuur* (*literature*) and *roman* (*novel*). Word clouds for query results proved to be expensive to compute.

may lead to out-of-range birth province values, because the search is document-based and documents may have more than one author.

We have also experimented with a visualization based on a map. We used a standard vector data representation for maps with clustered numbers on the map which become more detailed when the user zooms in on the map (ol.Cluster of OpenLayers version 3.0). The example map in Figure 7 with birth locations of authors of the Dutch magazine De Gids makes clear that authors from Amsterdam (311) and The Hague/Rotterdam (226) contribute most to the magazine. Unfortunately, our users found this mapping visualization method confusing. Since we were unable to find a mapping visualization that suited their needs, we have stopped using them. The map was the only of the visualizations that was not (yet) interactive.

4.5. Average Age of Authors

For the publications for which we know both the publication year and the birth year of the authors, we can compute the age of the authors at publication time. The average of these numbers per time unit gives an insight in the average age of authors in different time periods . The visualization is similar to the document count visualization described in section 4.1.: a bar plot with time units years or decades. An example graph can be found in Figure 8.

The interaction with the bars is the same as in the document count graph described in section 4.1.: counts related to a bar can be inspected by moving the mouse pointer over the bar and clicking on a bar provides starts a new query with the birth year of the authors restricted to the value(s) corresponding to the bar.

4.6. Other Visualizations

Apart from the unrealized word count visualization (see section 4.1., there is one more visualization which we would like to create: one that shows which words occur more frequent in a query result than could have been expected. For example, an author may use certain words more frequently than others. By comparing the relative frequencies of the words in his work with the the frequencies of the words in all documents, we could find out his favorite words. A similar comparison could be made for documents of a certain time period.

For such a visualization, we need relative word frequencies: frequencies of words in a query result divided by their corresponding overall frequency. We only have word frequencies of query results available but not the overall frequencies. We work with a dynamic corpus, so the overall word frequencies are subject to change. Ideally they would have been computed at index time and made available, but this is not the case. This means that we need to compute these frequencies ourselves for each query. But since queries may result in tens of thousands of words, such computations will require a lot of time. Therefore we have refrained from creating this visualization.

As an alternative, we have generated a word cloud visualization of the content words in a query, that is the words minus the stop words (common words like the Dutch equivalents of *the*, *be*, *to*, ...). This offers a good summary of the words that be found in documents matching a query (see Figure 9). Users that are interested in a particular word in the cloud can restrict their query to documents containing that word by clicking on it in the word cloud.

In order to create the word clouds, all the words of all the documents in the query result need to be processed. This requires many computational resources. Queries with many results have even caused our broker search engine to crash. For this reason we have restricted the maximum number of documents used for creating the word clouds to ten. The sample of selected documents is chosen randomly. Our users did not like this fact, so they rejected this particular visualization. For that reason we have abandoned using the word cloud visualization.

5. Case Study

De Gids (*The Guide*) is a Dutch literary magazine which was first issued in the year 1837. The magazine still exists in 2016. Most of the issues are available in digital format on the website `dbnl.nl`. In 1885 a group of young Dutch authors started a new magazine De Nieuwe Gids (*The New Guide*) with a specific name reference to the older maga-

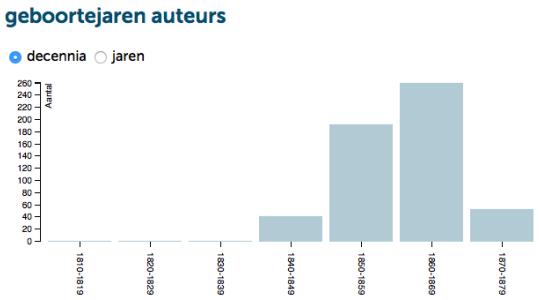

Figure 10: Overview of the decades of birth of the authors writing in the magazine De Nieuwe Gids in 1885-1894. The sequence of bars is shaped like a Poisson distribution with the maximum value in the decade 1860-1869. The author birth year distribution for the older magazine De Gids in the same time period looks similarly, but it has its peak in the decade 1840-1849.

eene haar zijne hem hare Aurelie had Guido Cecile Eduard ende was Jules den zich oom aan Ottilie welke jaren Grimm Staten Moli Bismarck Quaerts Japan hij hart Amerika Oranje Unie Molire eeuw negers Zuiden archieven brieven Jonker geschiedenis letterkunde dollars parlement oude vader Goncourts Coster godsdienst vrouw burcht Rensken dochter Dekker Tiryns re Lievendaal Flaubert Balzac in Noord Japansche Jacob Douwes gebied Holland president zoon zijnen haren Latijn grondwet Mainz der werd kon Zij vreemdeling Noorden blik zwaard mevrouw Zuid graan Amerikanen Armande Athene uitvinding Kees Europa U weerloosheid Kaakebeen middeleeuwsche Germaansche mijnheer Doopsgezinden Wilhelm sprak rijtuig Amerikaansche

Figure 11: The 100 most surprising words in the magazine De Gids in 1885-1894 in comparison with the words found in de editions of De Nieuwe Gids of the same period. The words are sorted by decreasing t-score.

zine. The editors of the new magazine were part of a new Dutch literary movement called Tachtigers (*From the Eighties*) which were organized in a group called Flanor. The new magazine lasted until 1943 but most of the original editorial staff left in 1894. Digital versions of the issues of De Nieuwe Gids can also be found on the website `dbnl.nl`.

It would be interesting to study the differences between the two magazines, especially for the rebel years period: 1885-1894. Did the authors really write about different topics in the new magazine? How did that influence their choice of words and topics? Were the authors of the new magazine younger than those of the older publication? What about the male-female ratio in the two magazines? Was there a geographical difference, for example by birth location, between the two magazines?

The visualizations presented in the previous sections make it easy to compare the two magazine section. We searched for all articles published in De Gids and De Nieuwe Gids in the years 1885-1894. We found 1034 articles for De Gids and 536 articles for De Nieuwe Gids. In the visualizations the most striking difference between the two magazines was the difference in age. The authors in De Gids were on average between 43 and 49 years old in the ten publication years while the average age of the contributors to De Nieuwe Gids was between 28 and 33. This difference is also visible in the graphs for birth years: most authors writing in De Gids in 1885-1894 were born in the decade 1840-1849 while most authors of De Nieuwe Gids in that time period were born in 1860-1869 (see Figure 10).

Another striking difference could be found in the gender visualization. 7% of the known authors in the older magazine De Gids proved to be female while only 2% of the known authors in the new De Nieuwe Gids were women. We had expected a higher female-male ratio in the new magazine but this expectation turned out to be wrong. The numbers in the gender visualizations were hard to use. The magazine De Gids had many authors with unknown gender (21%) and these made it difficult to see the actual male-female ratio: 93%-7% rather than 74%-5% as reported in the graph[4]. Having the counts for unknown gender in the graph makes it difficult to see the actual male-female ratio. However, the counts for unknown gender are in the graph because that was an explicit wish of our users. But it would have been better if displaying the numbers for unknown gender was optional.

Since most of the editorial staff of the new De Nieuwe Gids was born in and around Amsterdam, we expected that many of the authors of the magazine would be born there and in the associated province North Holland. This proved to be the case: 43% of the authors who wrote in De Nieuwe Gids in 1885-1894 were born in Amsterdam and more than half (58%) came from the province of North Holland. However, these numbers were not exceptionally higher than the corresponding numbers for the magazine De Gids: Amsterdam 33% and North Holland 45%. So one can argue that Amsterdam authors dominated De Nieuwe Gids but this was also true to a lesser extend for the magazine De Gids.

We compared the context of the two magazines by selecting ten articles of each volume and counting the words in those articles. Next we compared the word frequencies of the two magazines with the t-score (Church et al., 1991) which can be seen as a surprise factor: a high t-score indicates a surprisingly high frequency of a word in comparison with the frequency of the word in some reference corpus. In this case we calculated the surprise factor of a word in one magazine by comparing its frequency with its frequency in the other magazine.

Figures 11 and 12 contain an overview of the most surprising words in the two magazines. These include character names of stories (Aurelie in De Gids and Johannes in De Nieuwe Gids) which as expected because the two magazines publish different stories. De Gids seems to contain more articles about foreign politics: Bismarck, Japan,

[4]We assume that the distribution of unknown values is equal to that of the known values. Hence, from a 74-5-21 distribution we conclude that 74/(74+5)=93% are male and 5/(74+5)=7% are female.

Sokrates Johannes is Alkibiades ge menschen ik Heer een dingen Windekind uw dat Want mijn En politie Ktesippos wil mensch geneeskundige sentiment Thales zal inkomen Plato mooi wij kan oorzaak socialisten gij redeneering niet arbeiders dus Dit klasse begeerte deze zon zeggen doch Verstege sonnetten Winkel sonnet wezen Wistik spreken Amsterdam dit Shakespeare Marken hulp volgens ziel erg Houten Nu Hippias gemeente pct inkomstenbelasting Berthollet bestaan oorzaken socialistische Handelsblad arbeider heer en wet waarneming sociaal nous Dat brochure samenleving Robinetta lichaam stoffen kapitalistische Rochemont Agathon hebt God behandeling anders als begrip temperatuur socialisme boekje ding doelleer B elkaar woorden

Figure 12: The 100 most surprising words in the magazine De Nieuwe Gids in 1885-1894 in comparison with the words found in de editions of De Gids of the same period. The words are sorted by decreasing t-score.

Amerika, Zuiden (*South*) and grondwet (*constitution*). De Nieuwe Gids has more articles on ancient Greece (Socrates, Ktesippos, Thales, Plato and Hippias) and the Amsterdam area (Amsterdam, Marken and Houten). Both write about religion: godsdienst (*religion*) and Doopsgezinden (*Mennonites*), and God in De Nieuwe Gids. The latter magazine seems to write about left-wing politics: socialisten (*socialists*), arbeiders (*workers*) and klasse (*class*). Its word list also reveals an interest in poetry: sonnet and sonetten (*sonnets*).

The data visualization enabled us to get insights in the metadata differences between the two magazines without much effort. We believe a word cloud would be a good method to examine the vocabulary differences between the magazines, as shown with the t-score comparison in this section. Unfortunately our users did not agree. An alternative to a word cloud presentation would be to offer the top of the list of surprising words in tables with additional numeric information like t-score, absolute frequency and relative frequency. It would be interesting to be able to find topic differences in the same way but this requires a topic annotation of the corpus, which is currently unavailable.

6. Concluding Remarks

We have presented interactive visualizations which can be used to explore and analyze texts and literary metadata. These involve tracking frequencies of documents containing certain words over time and examining biographical data of authors. We used the visualizations to compare two competing Dutch magazines and found some interesting results.

There are several extensions we have in mind for future work. First, as explained in section 4.1., we would like to create graphs of word counts over time rather than document counts. However, in order for this to be possible in acceptable computing time, the software architecture which we rely on, will need to be changed. Currently people are working on this extension so we are optimistic about being able to create this visualization in the coming year.

The dependency on the architecture was one of the lessons we learned from this project. Even if interesting data is available for visualization, generating a visualization may take too much time or require too many computational resources. Ideally much of the visualized data is generated by the underlying search architecture so that little computation is required at run time. A good cooperation between the visualization team and the search team is required for achieving this.

As a topic for future work, we are also interested in building a visualization which presents summaries of the content of text collections. Since our users did not like word clouds we will need to consult them on this and find a representation of textual data that both suits them and is feasible computationally. We would also like to visualize author data on a map. A candidate for this is the bubble map in combination with a parallel coordinates plot, as shown in (Theron and Wandl-Vogt, 2014), which could interactively display birth or death location together with other biographical information in one view.

Syntactic annotations are a standard feature of our corpus which we have not yet used in our visualizations. It would be interesting to visualize aggregated information about the part-of-speech of words in the neighborhood of query words, like done in OpenSonar (Reynaert et al., 2014). With a visualization like the one in the interHist interface (Lyding et al., 2014), part-of-speech classes of several neighboring words could be inspected in one view.

One of the recurring questions of our users is about finding new and abandoned words in a certain time period. A new word can be identified from its frequency graph: it was never used before a certain year and becomes popular after that year. An abandoned word behaves in the opposite way. The question: *Which new words became popular in the nineteenth century?* can be answered but it requires a lot of computation which may need to be repeated when the text corpus changes. Still this would be an interesting search option to be offered to our users. A related but even more challenging question is *What new sense of a word became popular at a certain time?*. Methods for answering this question and visualizing the results, exist (Rohrdantz et al., 2011) and exploring this approach would be an interesting topic for future study.

7. Acknowledgments

The study described in this paper was enabled by support from the Dutch organization CLaRIAH (Common Lab Research Infrastructure for the Arts and Humanities[5]). We thank Nicoline van der Sijs for applying for funding for the project and René van Stipriaan for valuable user feedback on the visualizations. The paper was reviewed by two anonymous reviewers, who we thank for useful comments.

[5] clariah.nl

8. Bibliographical References

Bostock, M. (2011). D3 Data-Driven Documents. http://d3js.org/ Retrieved 22 March 2016.

Brouwer, M., Brugman, H., Kemps-Snijders, M., Kunst, J. P., van Peet, M., Zeeman, R., and Zhang, J. (2013). Providing searchability using broker architecture on an evolving infrastructure. unpublished manuscript.

Brouwer, M., Brugman, H., Kemps-Snijders, M., Kunst, J. P., van Peet, M., and Zeeman, R. (2015). MTAS: Extending Solr and Lucene with Scalable Searchability on Annotated Text. unpublished manuscript.

Brugman, H., Reynaert, M., van der Sijs, N., van Stipriaan, R., Tjong Kim Sang, E., and van den Bosch, A. (2016). Nederlab: Towards a single portal and research environment for diachronic dutch text corpora. In *Proceedings of LREC 2016*. ELRA, Portoroz, Slovenia.

Church, K., Gale, W., Hanks, P., and Hindle, D. (1991). Using Statistics in Lexical Analysis. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Lawrence Erlbaum Associates.

Google. (2010). Google books Ngram Viewer. https://books.google.com/ngrams Retrieved 22 March 2016.

Gormley, C. and Tong, Z. (2015). *Elasticsearch: The Definite Guide*. O'Reilly Media Inc.

Komen, E. R. (2015). *An insider's guide to the Nederlab R visualization webserver*. Technical report, Radboud University Nijmegen. Version 2.5, March 16. Retrieved March 22, 2016 from http://erwinkomen.ruhosting.nl/nedlab/2015_NederlabR webservice.pdf.

Lyding, V., Nicolas, L., and Stemle, E. (2014). 'interHist' - an interactive visual interface for corpus exploration. In *Proceedings of LREC 2014*, pages 635–641. Reykjavik, Iceland.

Reynaert, M., van de Camp, M., and van Zaanen, M. (2014). OpenSoNaR: user-driven development of the SoNaR corpus interfaces. In *Proceedings of COLING 2014*, pages 124–128. Dublin, Ireland.

Rohrdantz, C., Hautli, A., Mayer, T., Butt, M., Keim, D. A., and Plank, F. (2011). Towards Tracking Semantic Change by Visual Analytics. In *Proceedings of ACL 2011*, pages 305–310. Association for Computational Linguistics, Portland, Oregon.

Theron, R. and Wandl-Vogt, E. (2014). The Run of Exploration: How to Access a Non-Standard Language Corpus Visually. In *VisLR: Visualization as Added Value in the Development, Use and Evaluation of Language Resources*. LREC, Reykjavik, Iceland.

Tjong Kim Sang, E. (2014). Finding Syntactic Characteristics of Surinamese Dutch. Technical report, Meertens Institute, Amsterdam, The Netherlands, March.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. Graphics Press, Chesthire, Connecticut.

Zhang, J., Brouwer, M., Brugman, H., Droes, M., Kemps-Snijders, M., Kunst, J. P., van der Sijs, N., van Stipriaan, R., Sang, E. T. K., and Zeeman, R. (2013). *Visual Analytics in a Virtual Research Environment for Humanities*. Poster presented at the International UDC Seminar: Classification & Visualization, Interfaces to Knowledge, The Hague, The Netherlands.

The Glottolog Data Explorer: Mapping the World's Languages

Andrew Caines[1], **Christian Bentz**[2], **Dimitrios Alikaniotis**[1],
Fridah Katushemererwe[3], **Paula Buttery**[1]
[1]Department of Theoretical & Applied Linguistics, University of Cambridge, U.K.
[2]Institute of Linguistics, Universität Tübingen, Germany
[3]Department of Linguistics, Makerere University, Uganda
apc38@cam.ac.uk, chris@christianbentz.de, da352@cam.ac.uk, katu@chuss.mak.ac.ug, pjb48@cam.ac.uk

Abstract

We present THE GLOTTOLOG DATA EXPLORER, an interactive web application in which the world's languages are mapped using a JavaScript library in the 'Shiny' framework for R (Chang et al., 2016). The world's languages and major dialects are mapped using coordinates from the Glottolog database (Hammarström et al., 2016). The application is primarily intended to portray the endangerment status of the world's languages, and hence the default map shows the languages colour-coded for this factor. Subsequently, the user may opt to hide (or re-introduce) data subsets by endangerment status, and to resize the datapoints by speaker counts. Tooltips allow the user to view language family classification and links the user to the relevant Glottolog webpage for each entry. We provide a data table for exploration of the languages by various factors, and users may download subsets of the dataset via this table interface. The web application is freely available at `http://cainesap.shinyapps.io/langmap`

Keywords: Glottolog, world languages, endangered languages

1. Introduction

When it comes to the cartographic visualization of language resources, one resource type that lends itself well to such an exercise is the typological database, as demonstrated for example by *The World Atlas of Language Structures* (Dryer and Haspelmath, 2013). Here we present THE GLOTTOLOG DATA EXPLORER (GDE), an interactive visualization of the *Glottolog* database (Hammarström et al., 2016).

It is primarily intended to draw attention to the endangered status of many of the world's languages. Our aim is to illustrate the huge number of distinct languages around the world in the present day, and to imply that there is much to be lost, socially and culturally, if the languages classified as currently vulnerable or endangered were allowed to shrink away into extinction. Hence we allow the user to easily remove the vulnerable, endangered and extinct languages from the map and immediately visualize the projected scenario (see section 3).

The intended audience includes both linguistics specialists and the wider public. For the latter audience the GDE has to be visually appealing, interactive and easy to understand. At the same time we wish the GDE to be useful to the former group, at least to give an insight to certain research questions if not to answer them. We discuss our plans in this regard in section 5, and welcome feedback as to how we can make the GDE a useful research tool.

2. Glottolog

GLOTTOLOG is a catalogue of the world's languages, curated by members of the Max Planck Institute for Evolutionary Anthropology. It contains typological, geographic and bibliographic information for each so-called *languoid*. The creators of Glottolog chose this term to indicate that they are not just cataloguing *languages*, but any dialect, language or language family "that linguists need to be able to identify"[1]. Every languoid is curated into the catalogue with a list of any relevant linguistic works (grammars, dictionaries, *etc*), its classification in 'the Glottolog tree' (a linguistic genealogy), and – crucially for our purposes – latitude and longitude coordinates which allow its location to be mapped. Glottolog is freely available, regularly updated, and welcomes contributions from the linguistic community.

2.1. Data Collection

We obtained Glottolog languoid data using a two-step process. First, we retrieved a full list of languoids from their resource map in JSON format[2]. At the time of writing there were a total of 22,924 languoids in the Glottolog catalogue. We subsequently iterated over every languoid code, downloading the data made available by Glottolog, again in JSON format[3]. For each languoid we checked for longitude and latitude values, retaining only those languoids with these variables present – that which would allow us to map their location in the GDE. This step excluded 15,295 unsituated languoids, the bulk of which are 10,414 dialects taken from the Multitree project (The LINGUIST List, 2014), not yet systematically cleaned of errors, geo-located, and properly included in the Glottolog catalogue.

The remainder are made up of 4112 language families, which are even trickier than languages to geo-locate, and 769 languages without longitude or latitude values, many of which are so-called 'bookkeeping' languoids retained in the catalogue for the sake of completeness, even though they have for some reason been withdrawn from the 'live' language ontology. For instance ACHI′, CUBULCO was merged with ACHI in 2008, as it is "a dialect or dialect

[1]Source: Glottolog website, accessed 2016-03-24

[2]`http://glottolog.org/resourcemap.json?rsc=language`

[3]For example, the URL for LATIN, which has the Glottolog identifier `lati1261`: `http://glottolog.org/resource/languoid/id/lati1261.json`

group name, and therefore incorrect for a language designation"[4].
Thus we are left with 7629 languoids of more certain status and associated with geo-coordinates. Of these 221 are 'bookkeeping' languoids, and so we exclude them from the GDE. Our final dataset therefore contains 7407 entries, for which we present some high-level descriptive statistics regarding language type in Table 1 in the manner of the Glottolog information page[5]. The user may view the languoids for each language type listed in Table 1 by searching for that type in the GDE TABLE tab (see section 3).

Type	Count
Spoken L1 language	7183
Unattested	46
Unclassifiable	18
Pidgin	18
Mixed language	15
Artificial language	3
Speech register	3
Sign language	121
Total	*7407*

Table 1: Geo-located languoids extracted from Glottolog

For each languoid in our set of 7407, we collected the values needed for our web application: its Glottolog alphanumeric identifier, its assigned latitude and longitude coordinates, its higher-level genealogical classification, and its endangerment status on the UNESCO scale (Moseley, 2010). UNESCO's six degrees of vitality/endangerment are: extinct, critically endangered, severely endangered, definitely endangered, vulnerable, and safe. The Glottolog curators added 'unknown' for those languoids not featured in UNESCO's database, and replaced 'safe' with 'living', quite understandably given that 'safe' implies a certain stability for these languages, even those which may be on the borderline with the 'vulnerable' category.
Regarding genealogical classification, we selected at most the three highest classes for each languoid, where by 'highest' we mean the super-groupings such as Indo-European, Afro-Asiatic, Sino-Tibetan, Austronesian, and so on. Many languoids are classified to greater than three levels, maximally seventeen in fact, but the decision was taken to limit our data collection in this way so that the resulting user experience was not too unwieldy.
The information we hold for each languoid is of the following form:

```
name:  Senara Sénoufo
id:  sena1262
latitude:  10.4987
longitude:  -5.28216
family1:  Atlantic-Congo
family2:  Volta-Congo
family3:  North Volta-Congo
status:  Living
```

[4]Source: http://glottolog.org/resource/languoid/id/achi1258; accessed 2016-02-11
[5]Source: http://glottolog.org/glottolog/glottologinformation; accessed on 2016-02-11.

Our collected languoid information is presented to the user in two forms: as pop-up 'tooltips' for any selected languoid on the MAP tab, and alternatively in list format in on the TABLE tab (see section 3). Every tooltip contains a link to the languoid's Glottolog webpage, on which all associated information, including full family classification, is given. We last accessed the Glottolog catalogue on 2016-02-09, and will continue to regularly download the latest Glottolog data and ensure GDE contains up-to-date information.

3. Glottolog Data Explorer

The GDE application was written in R (R Core Team, 2015) using an interface to the LEAFLET JavaScript library (Cheng and Xie, 2015) and developed as a web application in the SHINY framework (Chang et al., 2016). It is hosted on SHINYAPPS servers and is freely available at http://cainesap.shinyapps.io/langmap.
The MAP itself is a layered widget starting with a Stamen basemap, lines and labels[6]. The decision to use Stamen is purely aesthetic: we found that our data points showed up best on these map tiles. Other basemaps we considered were OpenStreetMap, Esri's National Geographic map, and NASA's 'Earth at night' (Figure 1).

Figure 1: Considered basemap tiles, clockwise from bottom-left: NASA's 'Earth at night', OpenStreetMap, Esri National Geographic, Stamen watercolour.

We add the languoids to the basemap as geo-located markers coloured and layered according to endangerment status. Colour choices were quite straightforwardly grey for languoids of unknown status, shades of red to purple for the endangered and vulnerable languoids, and green for 'living' (Figure 2). We provide a legend to this effect, along

[6]Made available by Stamen Design under a Creative Commons Attribution (CC BY 3.0) licence, with data from OpenStreetMap under a Creative Commons Attribution-ShareAlike (CC BY-SA 3.0) licence. See http://maps.stamen.com

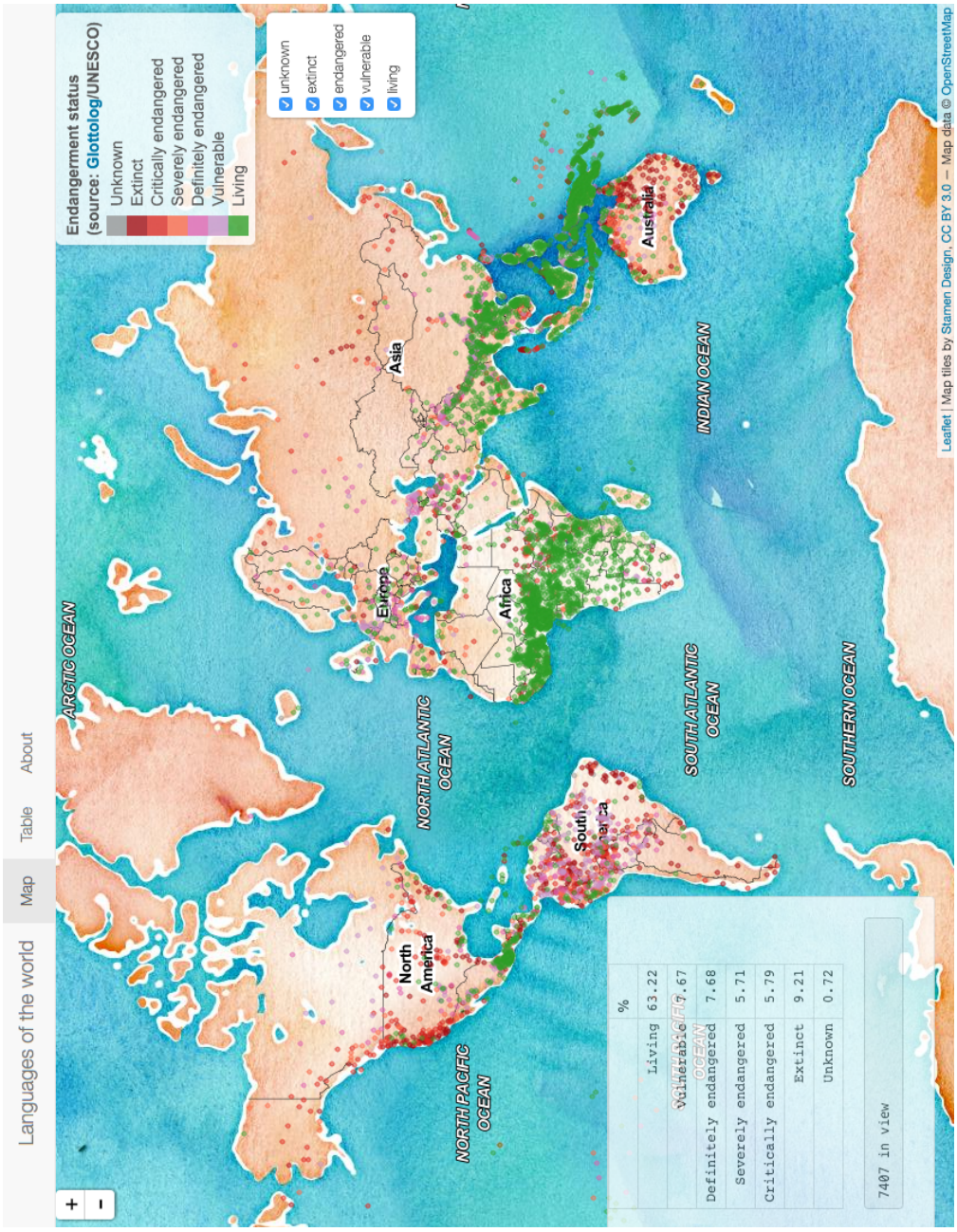

Figure 2: The GDE MAP tab, whole world view.

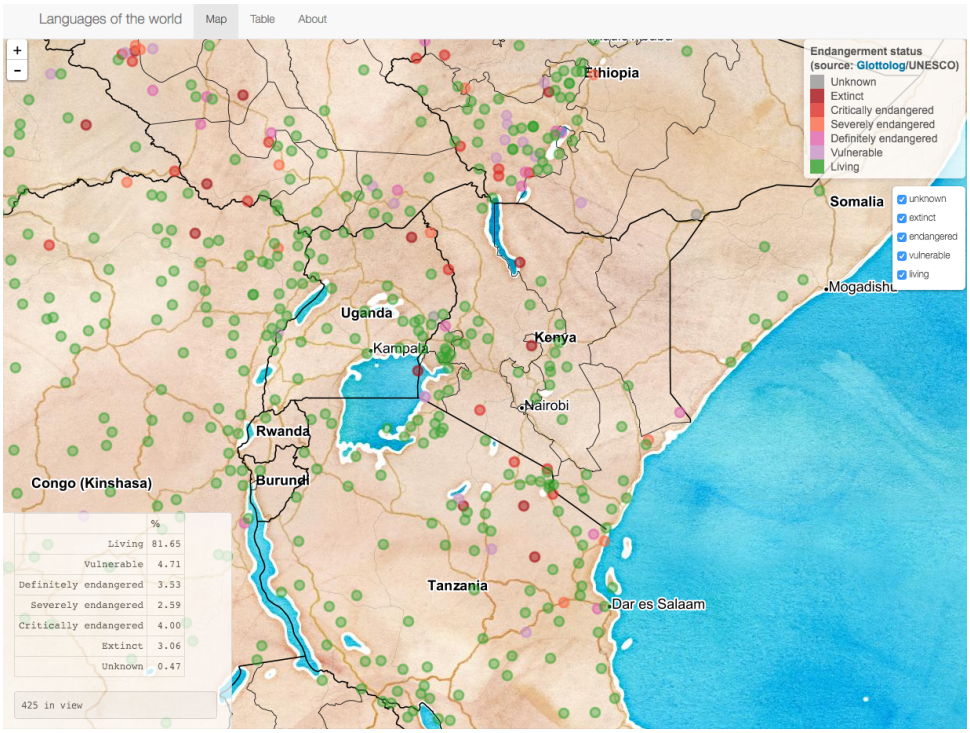

Figure 3: The GDE MAP tab, zoomed into East Africa; *n.b.* adjusted table counts, bottom-left.

Figure 4: The GDE MAP tab, tooltips example.

Figure 5: The GDE MAP tab, datapoints sized by speaker counts with, for instance, Italian the green circle at centre (55 million speakers), Judeo-Italian the small green circle to its left (200 speakers), and Sicilian the purple circle at bottom (4.7 million speakers).

with a control panel to exclude (and include) languoids according to their status (top-right of map).

We chose to layer the languoids from extinct to living to draw attention to how many living languages there are and where there are noticeably high quantities (*e.g.* West Africa and Papua New Guinea above all). To put the extinct languoids on top did not make sense to us: this would be arresting but in some sense futile. It's already too late for these languages, and what's more, many of them are distant to the present day – ancient languages such as Latin and Egyptian. However, many are not, and we have made the living languoid points less opaque than the other languoids, so that the reds and purples of extinct and endangered languages lurk ominously in the background.

On the left of the window there are zoom controls, along with a count of the number of languoids currently in view, which at the most is 7407 (Figure 2) but which, if we zoom in on the east African region for example, reduces to 293 (Figure 3). Note that the table of endangerment proportions also adjusts to the current view (*cf.* Figure 2, Figure 3), and that the table may be 'picked up' and dragged to another point on the map if so desired. Tooltips mean that if the user selects a languoid datapoint, a textbox pops up on screen with more information about that languoid: its higher level family classifications, its endangerment status, speaker count, and a link to its Glottolog webpage (*e.g.* Figure 4).

The user may also opt to display the languoid markers sized by their speaker counts from Ethnologue (Lewis et al., 2015). We use a logarithmic scale for this visualization function (Figure 5). Speaker counts were collected by the second author (Bentz, 2016); note that we make them visible as part of the tooltip but they are not available to download in the TABLE. We would need an Ethnologue licence (see also section 5) to redistribute these data. However, we envisage that it would be most valuable for other researchers to test hypotheses relating to population size and language features.

With 7407 data points at most, overplotting is evidently a danger in the GDE, and something that's hard to avoid when there are so many dense geographic clusters. Apart from the partial transparency mentioned above, we resize the data points so that they are smallest at the outermost map zoom, increase as the user zooms in, and vice versa. As the points have to be redrawn as the user moves between zoom levels this leads to a slowdown in performance. With funding, we could upgrade our Shiny Apps subscription to a paid one with its accompanying performance boost involving increased memory and multi-threading.

The TABLE tab allows the user to view the data table underlying the map. Aside from browsing page by page, there are several table-filtering methods available (highlighted in the figures with red boxes; these boxes do not appear in GDE): a search textbox, endangerment status tick-boxes,

and language family selection (Figure 6). The user may also download the table rows currently in view in one of several formats, thanks to the DT package which provides an R interface to the JavaScript library DataTables (Xie, 2015).
Finally, the ABOUT tab provides information about the GDE web application: the motivation for creating it, a brief commentary about present-day language endangerment, credits and acknowledgements as to where the data come from and the tools used to make the application (Section 6).

4. Language Endangerment

Estimates vary as to the severity of the prognosis at this point: UNESCO refers to a 'widely accepted' endangerment ratio of 50% (Moseley, 2010), whilst one heavily cited study predicts that "the coming century will see either the death or doom of 90% of mankind's languages" (Krauss, 1992), and the unattributed claim that a language dies every two weeks is in wide circulation (see the recent *Ethnos Project* blogpost by Mark Oppenneer for examples[7]). Others paint a less alarmist (though still alarming) picture (*e.g.* Campbell et al. (2013)). In any case, in the words of the linguist Lyle Campbell, the present predicament and its immediate consequences are "tragic, with its irreparable damage and loss" (Campbell et al., 2013).
As stated at the outset, language endangerment was our primary focus in creating GDE. For the 7407 geo-located languoids we extracted from the Glottolog catalogue we find that only 63% were classified by UNESCO as 'safe' (or 'living', as Glottolog relabelled it; Table 2).

Endangerment	Count	%
Living	4683	63.22
Vulnerable	568	7.67
Definitely endangered	569	7.68
Severely endangered	423	5.71
Critically endangered	429	5.79
Extinct	682	9.21
Unknown	53	0.72
Total	*7407*	*100*

Table 2: Endangerment status statistics for the languoids extracted from Glottolog

If one excludes already extinct and status-unknown languoids from the proportional calculation, then the safe percentage rises to 70% $(4683/(7407-(682+53))$.
If one assumes that the path from 'safe' to 'extinct' is monotonic and inevitable, then we are faced with the loss of three-in-ten existing languages, a prospect less alarming than the oft-repeated 90% figure that comes from Krauss Krauss (1992). Nonetheless it would be a horrendous loss of cultural heritage and diversity, especially if one considers regional endangerment. For example, by zooming the GDE map in to the approximate region of the Amazon rainforest and its surroundings in northern South America, and by considering extant languoids only, we see that the 'safe' proportion falls to just 13.5% (Figure 7). More precise regional analyses may be performed in future once we associate languoid geo-coordinates with information by country and region (section 5).
How do the endangerment counts presented here, that come from UNESCO via Glottolog, compare to others? Ethnologue's most recent report states that 63% $(4719/7480)$ of the languages identified in 1950 "are still being passed on to the next generation in a sustainable way", whilst 32% "are currently at some stage in the process of language loss", and 377 (5%) "have been identified as having lost all living speakers and ceasing to serve as a language of identity for an ethnic community in the last six decades" (Simons and Lewis, 2013).
Meanwhile the ENDANGERED LANGUAGES CATALOGUE project (ELCat) puts the endangerment statistic at 43% of 7102 existing languages, and states that 457 (9.2%) have "fewer than ten speakers and are very likely to die out soon, if no revitalization efforts are made", while 634 of all known languages have already become extinct, 141 of which in the last forty years (Campbell et al., 2013).
More worrying yet is ELCat's observation regarding language families, an issue not represented in the GDE[8]:

> We know of a hundred language families that have gone extinct over the course of history – 24% of the world's linguistic diversity. But the fact that 28 of them have gone extinct over the relatively short time span of the last 50 years is symptomatic of the accelerated rate of language loss we are experiencing in recent times.

Analogies with biodiversity loss are clear: indeed, direct parallels between language and species extinction have been drawn, with both linked to economic development (Amano et al., 2014). ELCat is an ongoing project that continues to focus on endangered languages and add to its catalogue with crowdsourced contributions. They have also produced a map visualization of language endangerment which the reader may find at `http://www.endangeredlanguages.com`, though note that it only maps endangered languages rather than all languages as in the GDE.
The above, somewhat bleak, assessments of course may be improved by government or community efforts toward language revitalization. This is no easy endeavour, but we point to notable successful efforts in recent times, such as Hebrew and Scottish Gaelic (McEwan-Fujita, 2011; Kaufman, 2005). We also draw attention to the efforts in computational linguistics to spread the kind of natural language processing technology that is so prevalent and which has so benefitted major languages (above all English) to 'low resource' languages – those languages for which the tools and databases that underpin, for instance, web search, grammar checkers and teaching apps do not yet exist. Notable examples include the HUMAN LANGUAGE PROJECT (Abney

[7] `http://www.ethnosproject.org/status-of-the-ethnosphere`

[8] Source: `http://rosettaproject.org/blog/02013/mar/28/new-estimates-on-rate-of-language-loss`; accessed 2016-12-11.

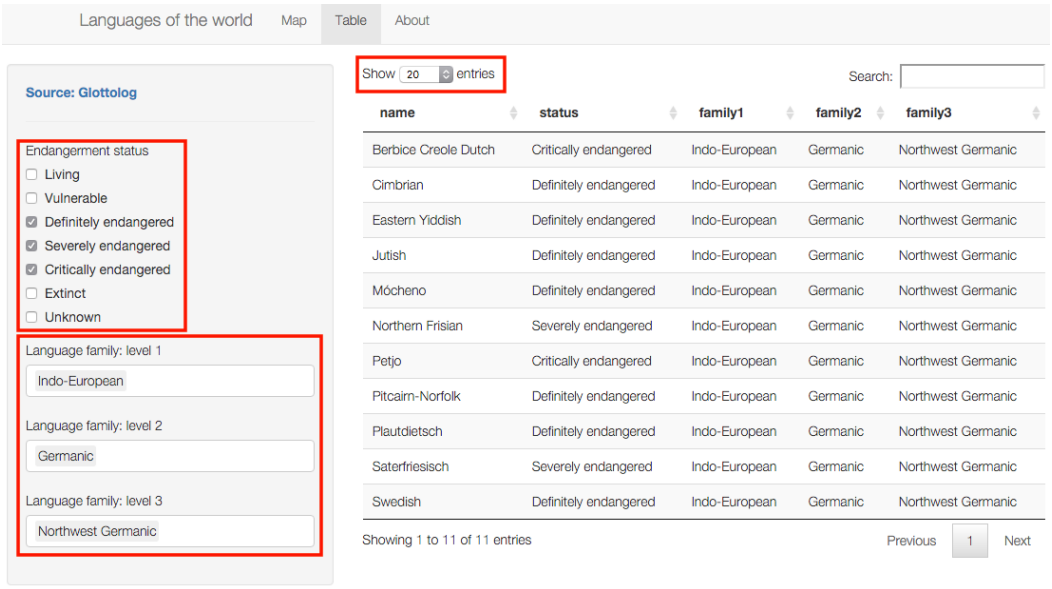

Figure 6: The GDE TABLE tab, endangered Northwest Germanic languoids, 20 per page.

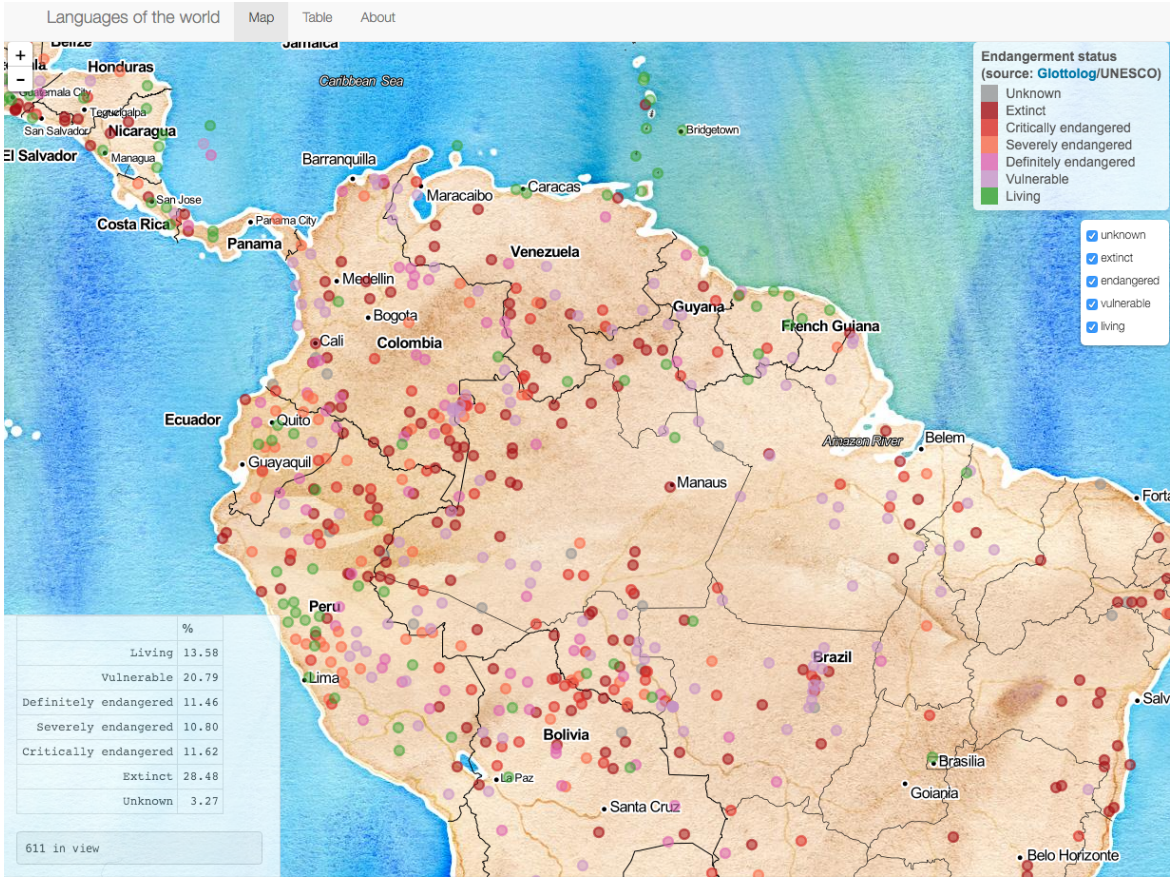

Figure 7: The GDE MAP tab, the northern region of South America including the Amazon rainforest.

and Bird, 2010; Emerson et al., 2014), LOWLANDS (*e.g.* Agic et al. (2015)) and RU_CALL – computer-assisted language learning for the revitalization of a Ugandan language, Runyakitara (Katushemererwe and Nerbonne, 2015).

5. Future Development

The GDE currently visualizes endangerment status for each languoid, indicated by datapoint colours on the map. We envisage further development to either offer endangerment data source options to the user – for instance offering a choice of UNESCO, Ethnologue, or ELCat classification – and/or to offer alternative visualizations with the Glottolog languoid set: *e.g.* number of L2 speakers, altitude and complexity measures Bentz (2016). Another improvement would be to offer the speaker counts for download via the TABLE. For this we would need funding to purchase an Ethnologue licence (Lewis et al., 2015) that allows for redistribution of these data.

As for the GDE itself, we intend to add further functionality: first, the option to minimize the data table in the MAP tab; second, the facility to download the filtered data (or, all of it) in the TABLE tab; thirdly, visualization of languoids grouped by region, country and family. We welcome further suggestions toward the improvement of GDE usability, data presentation, and content.

6. Acknowledgements

We thank Harald Hammarström of the Max Planck Institute for Psycholinguistics, and Robert Forkel of the Max Planck Institute for the Science of Human History for their help in accessing and explaining the Glottolog database. We are grateful for the many helpful comments and questions from two anonymous reviewers. We also thank Dr Anne Alexander and participants for their feedback at the 'Graphical Display: Challenges for Humanists' workshop organised by the Cambridge Digital Humanities Network in 2015, and Jane Walsh for her ongoing support of linguistics research at Cambridge as coordinator of the Language Sciences Strategic Research Initiative. The first and last authors are funded by Cambridge English Language Assessment as part of the Automated Language Teaching & Assessment Institute. The third author is supported by the Onassis Foundation. The fourth author is in receipt of a CAPREx award from the Cambridge-Africa Programme.

7. Bibliographic References

Abney, S. and Bird, S. (2010). The Human Language Project: Building a universal corpus of the world's languages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Agic, Z., Hovy, D., and Søgaard, A. (2015). If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*.

Amano, T., Sandel, B., Eager, H., Bulteau, E., Svenning, J.-C., Dalsgaard, B., Rahbek, C., Davies, R. G., and Sutherland, W. J. (2014). Global distribution and drivers of language extinction risk. *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1793).

Bentz, C. (2016). *Adaptive Languages: An Information-Theoretic Account of Linguistic Diversity*. Phd thesis, University of Cambridge.

Campbell, L., Lee, N. H., Okura, E., Simpson, S., and Ueki, K. (2013). New Knowledge: Findings from the Catalogue of Endangered Languages (ELCat). In *3rd International Conference on Language Documentation & Conservation*.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J., (2016). *Shiny: web application framework for R*. R package version 0.13.0.

Cheng, J. and Xie, Y., (2015). *Leaflet: create interactive web maps with the JavaScript 'Leaflet' library*. R package version 1.0.0.

Matthew S. Dryer et al., editors. (2013). *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Emerson, G., Tan, L., Fertmann, S., Palmer, A., and Regneri, M. (2014). SeedLing: Building and using a seed corpus for the Human Language Project. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S., (2016). *Glottolog 2.7*. Jena: Max Planck Institute for the Science of Human History.

Katushemererwe, F. and Nerbonne, J. (2015). Computer-Assisted Language Learning in support of (re-)learning native languages: The case of Runyakitara. *Computer Assisted Language Learning*, 28(2):112–129.

Kaufman, D. (2005). Acquisition, Attrition, and Revitalization of Hebrew in Immigrant Children. In Dorit Diskin Ravid et al., editors, *Perspectives on Language and Language Development: Essays in Honor of Ruth A. Berman*, pages 407–418. Springer, Boston.

Krauss, M. (1992). The world's languages in crisis. *Language*, 68(1):4–10.

M. Paul Lewis, et al., editors. (2015). *Ethnologue: Languages of the World*. SIL International, Dallas, 18th edition. Online version: http://www.ethnologue.com.

McEwan-Fujita, E. (2011). Language revitalization discourses as metaculture: Gaelic in Scotland from the 18th to 20th centuries. *Language and Communication*, 31(1):48 – 62.

Christopher Moseley, editor. (2010). *Atlas of the World's Languages in Danger, 3rd edn*. UNESCO Publishing, Paris.

R Core Team, (2015). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

Simons, G. F. and Lewis, M. P. (2013). The world's languages in crisis: a 20-year update. In Elena Mihas, et al., editors, *Responses to language endangerment. In honor of Mickey Noonan*. John Benjamins, Amsterdam.

The LINGUIST List. (2014). Multitree: A digital library of language relationships.

Xie, Y., (2015). *DT: a wrapper of the JavaScript library 'DataTables'*. R package version 0.1.48.