Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources

Miriam Butt[1] **Jelena Prokić**[2] **Thomas Mayer**[2] **Michael Cysouw**[3]
[1]Department of Linguistics, University of Konstanz
[2]Research Unit Quantitative Language Comparison, LMU Munich
[3]Research Center Deutscher Sprachatlas, Philipp University of Marburg

1 Introduction

The LINGVIS and UNCLH (Visualization of Linguistic Patterns & Uncovering Language History from Multilingual Resources) were originally conceived of as two separate workshops. Due to perceived similarities in content, the two workshops were combined and organized jointly.

The overal aim of the joint workshop was to explore how methods developed in computational linguistics, statistics and computer science can help linguists in exploring various language phenomena. The workshop focused particularly on two topics: 1) visualization of linguistic patterns (LINGVIS); 2) usage of multilingual resources in computational historical linguistics (UNCLH).

2 LINGVIS

The overall goal of the first half of the workshop was to bring together researchers working within the emerging subfield of computational linguistics — using methods established within Computer Science in the fields of Information Visualization (InfoVis) and Visual Analytics in conjunction with methodology and analyses from theoretical and computational linguistics. Despite the fact that statistical methods for language analysis have proliferated in the last two decades, computational linguistics has so far only marginally availed itself of techniques from InfoVis and Visual Analytics (e.g., Honkela et al. (1995); Neumann et al. (2007); Collins et al. (2009); Collins (2010); Mayer et al. (2010a); Mayer et al. (2010b); Rohrdantz et al. (2011)). The need to integrate methods from InfoVis and Visual Analytics arises particularly with respect to situations in which the amount of data to be analyzed is huge and the interactions between relevant features are complex. Both of these situations hold for much of current (computational) linguistic analysis. The usual methods of statistical analysis do not allow for quick and easy grasp and interpretation of the patterns discovered through statistical processing and an integration of innovative visualization techniques has become imperative.

The overall aim of the first half of the workshop was thus to draw attention to this need and to the newly emerging type of work that is beginning to respond to the need. The workshop succeeded in bringing together researchers interesting in combining techniques and methodology from theoretical and computational linguistics with InfoVis and Visual Analytics.

Three of the papers in the workshop focused on the investigation and visualization of lexical semantics. Rohrdantz et al. present a diachronic study of fairly recently coined derivational suffixes (*-gate, -geddon, -athon*) as used in newspaper corpora across several languages. Their analysis is able to pin-point systematic differences in contextual use as well as some first clues as to how and why certain new coinages spread better than others. Heylen et al. point out that methods such as those used in Rohrdantz et al., while producing interesting results, are essentially black boxes for the researchers — it is not clear exactly what is being calculated. Their paper presents some first steps towards making the black box more transparent. In particular, they take a close look at individual tokens and their semantic use with respect to Dutch synsets. Crucially, they anticipate an interactive visualization that will allow linguistically informed lexicogra-

Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, pages 1–6,
Avignon, France, April 23 - 24 2012. ©2012 Association for Computational Linguistics

phers to work with the available data and patterns. A slightly different take on synset relations is presented by Lohk et al., who use visualization methods to help identify errors in WordNets across different languages.

Understanding differences and relatedness between languages or types of a language is the subject of another three papers. Littauer et al. use data from the WALS (World Atlas of Language Structures; Dryer and Haspelmath (2011)) to model language relatedness via heat maps. They overcome two difficulties: one is the sparseness of the WALS data; another is that WALS does not directly contain information about possible effects of language contact. Littauer et al. attempt to model the latter by taking geographical information about languages into account (neighboring languages and their structure). A different kind of language relatedness is investigated by Yannakoudakis et al., who look at learner corpora and develop tools that allow an assessment of learner competence with respect to various linguistic features found in the corpora. The number of relevant features is large and many of them are interdependent or interact. Thus, the amount and complexity of the data present a classic case of complex data sets that are virtually impossible to analyze well without the application of visualization methods. Finally, Lyding et al. take academic texts and investigate the use of modality across academic registers and across time in order to identify whether the academic language used in different subfields (or adjacent fields) of an academic field has an effect on the language use of that field.

3 UNCLH

The second half of the workshop focused on the usage of multilingual resources in computational historical linguistics. In the past 20 years, the application of quantitative methods in historical linguistics has received increasing attention among linguists (Dunn et al., 2005; Heggarty et al., 2010; McMahon and McMahon, 2006), computational linguists (Kondrak, 2001; Hall and Klein, 2010) and evolutionary anthropologists (Gray and Atkinson, 2003). Due to the application of these quantitative methods, the field of historical linguistics is undergoing a renaissance. One of the main problems that researchers face is the limited amount of suitable *comparative* data, often falling back on relatively restricted 'Swadesh type' wordlists. One solution is to use synchronic data, like dictionaries or texts, which are available for many languages. For example, in Kondrak (2001), vocabularies of four Algonquian languages were used in the task of automatic cognate identification. Another solution employed by Snyder et al. (2010) is to apply a non-parametric Bayesian framework to two non-parallel texts in the task of text deciphering. Although very promising, these approaches have so far only received modest attention. Thus, many questions and challenges in the automatization of language resources in computational historical linguistics remain open and ripe for investigation.

In dialectological studies, there is a long tradition, starting with Séguy (1971), in which language varieties are grouped together on the basis of their similarity with respect to certain properties. Later work in this area has incorporated methods of string alignment for a quantitative comparison of individual words to obtain an average measure of the similarity of languages. This line of research became known as dialectometry. Unlike traditional dialectology which is based on the analysis of individual items, dialectometry shifts focus on the aggregate level of differences. Most of the work done so far in dialectometry is based on the carefully selected wordlists and problems with the limited amount of suitable data (i.e. computer readable and comparable across dialects) are also present in this field.

This workshop brings together researchers interested in computational approaches that uncover sound correspondences and sound changes, automatic identification of cognates across languages and language comparison based both on wordlists and parallel texts. First, Wettig et al. investigate the sound correspondences in cognate sets in a sample of Uralic languages. Then, List's contribution to the volume introduces a novel method for automatic cognate detection in multilingual wordlists which combines various previous approaches for string comparison. The paper by Mayer & Cysouw presents a first step to use parallel texts for a quantitative comparison of languages. The papers by Scherrer and Prokić et al. both are in the spirit of the dialectometric line of research. Further, Jäger reports on quantifying language similarity via phonetic alignment of core vocabulary items. Finally, some of the pa-

pers presented in this workshop deal with further topics in quantitative language comparison, like the application of phylogenetic methods in creole research in the paper by Daval-Markussen & Bakker, and the study of the evolution of the Australian kinship terms reported on in the paper by McConvell & Dousset.

In the next section, we give a brief introduction into the papers presented in this workshop, ordered according to the program of the oral presentations at the workshop.

4 Papers

Christian Rohrdantz, Andreas Niekler, Annette Hautli, Miriam Butt and Daniel A. Keim ('Lexical Semantics and Distribution of Suffixes — A Visual Analysis) present a quantitative cross-linguistic investigation of the lexical semantic content expressed by three suffixes originating in English: *-gate, -geddon* and *-athon*. Using data from newspapers, they look at the distribution and lexical semantic usage of these morphemes across several languages and also across time, with a time-depth of 20 years for English. Using techniques from InfoVis and Visual Analytics is crucial for the analysis as the occurrence of these suffixes in the available corpora is comparatively rare and it is only by dint of processing and visualizing huge amounts of data that a clear pattern can begin to emerge.

Kris Heylen, Dirk Speelman and Dirk Geeraerts ('Looking at Word Meaning. An Interactive Visualization of Semantic Vector Spaces for Dutch synsets') focus on the pervasive use of Semantic Vector Spaces (SVS) in statistical NLP as a standard technique for the automatic modeling of lexical semantics. They take on the fact that while the method appears to work fairly well (though they criticize the standardly available evaluation measures via some created gold standard), it is in fact quite unclear how it captures word meaning. That is, the standard technology can be seen as a black box. In order to find a way of providing some transparency to the method, they explore the way an SVS structures the individual occurrences of words with respect to the occurrences of 476 Dutch nouns. These were grouped into 214 synsets in previous work. This paper looks at a token-by-token similarity matrix in conjunction with a visualization that uses the Google Chart Tools and compares the results with previous work, especially in light of different uses in different versions of Dutch.

Ahti Lohk, Kadri Vare and Leo Võhandu ('First Steps in Checking and Comparing Princeton WordNet and Estonian WordNet') use visualization methods to compare two existing WordNets (English and Estonian) in order to identify errors and semantic inconsistencies that are a result of the manual coding. Their method opens up a potentially interesting way of automatically checking for inconsistencies and errors not only at a fairly basic and surface level, but by working with the lexical semantic classification of the words in question.

Richard Littauer, Rory Turnbull and Alexis Palmer ('Visualizing Typological Relationships: Plotting WALS with Heat Maps') present a novel way of visualizing relationships between languages. The paper is based on data extracted from the World Atlas of Language Structures (WALS), which is the most complete set of typological and digitized data available to date, but which presents two challenges: 1) it actually has very low coverage both in terms of languages represented and in terms of feature description for each language; 2) areal effects are not coded for. While the authors find a way to overcome the first challenge, the paper's real contribution lies in proposing a method for overcoming the second challenge. In particular, the typological data is filtered by geographical proximity and then displayed by means of heat maps, which reflect the strength of similarity between languages for different linguistic features. Thus, the data should allow one to be able to ascertain areal typological effects via a single integrated visualization.

Helen Yannakoudakis, Ted Briscoe and Theodora Alexopoulou ('Automatic Second Language Acquisition Research: Integrating Information Visualisation and Machine Learning') look at yet another domain of application. They show how data-driven approaches to learner corpora can support Second Language Acquisition (SLA) research when integrated with visualization tools. Learner corpora are interesting because their analysis requires a good understanding of a complex set of interacting linguistic features across corpora with different distributional patterns (since each corpus potentially diverges from the standard form of the language by a different set of features). The paper

presents a visual user interface which supports the investigation of a set of linguistic features discriminating between pass and fail exam scripts. The system displays directed graphs to model interactions between features and supports exploratory search over a set of learner texts. A very useful result for SLA is the proposal of a new method for empirically quantifying the linguistic abilities that characterize different levels of language learning.

Verena Lyding, Ekaterina Lapshinova-Koltunski, Stefania Degaetano-Ortlieb, Henrik Dittmann and Chris Culy ('Visualizing Linguistic Evolution in Academic Discourse') describe methods for visualizing diachronic language changes in academic writing. In particular, they look at the use of modality across different academic subfields and investigate whether adjacent subfields affect the use of language in a given academic subfield. Their findings potentially provide crucial information for further NLP tasks such as automatic text classification.

Grzegorz Kondrak's invited contribution ('Similarity Patterns in Words') sketches a number of the author's research projects on diachronic linguistics. He first discusses computational techniques for implementing several steps of the comparative method. These techniques include algorithms that deal with a wide range of problems: pairwise and multiple string alignment, calculation of phonetic similarity between two strings, automatic extraction of recurrent sound correspondences, quantification of semantic similarity between two words, identification of sets of cognates and building of phylogenetic trees. In the second part, Kondrak sketches several NLP projects that directly benefitted from his research on diachronic linguistics: statistical machine translation, word alignment, identification of confusable drug names, transliteration, grapheme-to-phoneme conversion, letter-phoneme alignment and mapping of annotations.

Thomas Mayer and Michael Cysouw ('Language Comparison through Sparse Multilingual Word Alignment') present a novel approach on how to calculate similarities among languages with the help of massively parallel texts. Instead of comparing languages pairwise they suggest a simultaneous analysis of languages with respect to their co-occurrence statistics for individual words on the sentence level. These statistics are then used to group words into clusters which are considered to be partial (or 'sparse') alignments. These alignments then serve as the basis for the similarity count where languages are taken to be more similar the more words they share in the various alignments, regardless of the actual form of the words. In order to cope with the computationally demanding multilingual analysis they introduce a sparse matrix representation of the co-occurrence statistics.

Yves Scherrer ('Recovering Dialect Geography from an Unaligned Comparable Corpus') proposes a simple metric of dialect distance, based on the ratio between identical word pairs and cognate word pairs occurring in two texts. Scherrer proceeds from a multidialectal corpus and applies techniques from machine translation in order to extract identical words and cognate words. The dialect distance is defined as as function of the number of cognate word pairs and identical word pairs. Different variations of this metric are tested on a corpus containing comparable texts from different Swiss German dialects and evaluated on the basis of spatial autocorrelation measures.

Jelena Prokić, Çağrı Cöltekin and John Nerbonne ('Detecting Shibboleths') propose a generalization of the well-known precision and recall scores to deal with the case of detecting distinctive, characteristic variants in dialect groups, in case the analysis is based on numerical difference scores. This method starts from the data that has already been divided into groups using cluster analyses, correspondence analysis or any other technique that can identify groups of language varieties based on linguistic or extra-linguistic factors (e.g. geography or social properties). The method seeks items that differ minimally within a group but differ a great deal with respect to elements outside it. They demonstrate the effectiveness of their approach using Dutch and German dialect data, identifying those words that show low variation within a given dialect area, and high variation outside a given area.

Gerhard Jäger ('Estimating and Visualizing Language Similarities Using Weighted Alignment and Force-Directed Graph Layout') reports several studies to quantify language similarity via phonetic alignment of core vocabulary items (taken from the Automated Similarity Judgement Program data base). Jäger compares several string

comparison measures based on Levenshtein distance and based on Needleman-Wunsch similarity score. He also tests two normalization functions, one based on the average score and the other based on the informatic theoretic similarity measure. The pairwise similarity between all languages are analyzed and visualized using the CLANS software, a force directed graph layout that does not assume an underlying tree structure of the data.

Aymeric Daval-Markussen and Peter Bakker ('Explorations in Creole Research with Phylogenetic Tools') employ phylogenetic tools to investigate and visualize the relationship of creole languages to other (non-)creole languages on the basis of structural features. Using the morphosyntactic features described in the monograph on Comparative Creole Syntax (Holm and Patrick, 2007), they create phylogenetic trees and networks for the languages in the sample, which show the similarity between the various languages with respect to the grammatical features investigated. Their results lend support to the universalist approach which assumes that creoles show creole-specific characteristics, possibly due to restructuring universals. They also apply their methodology to the comparison of creole languages to other languages, on the basis of typological features from the *World Atlas of Language Structures*. Their findings confirm the hypothesis that creole languages form a synchronically distinguishable subgroup among the world's languages.

Patrick McConvell and Laurent Dousset ('Tracking the Dynamics of Kinship and Social Category Terms with AustKin II') give an overview of their ongoing work on kinship and social category terms in Australian languages. They describe the AustKin I database which allows for the reconstruction of older kinship systems as well as the visualization of patterns and changes. In particular, their method reconstructs so-called 'Kariera' kinship systems for the proto-languages in Australia. This supports earlier hypotheses about the primordial world social organization from which Dravidian-Kariera systems are considered to have evolved. They also report on more recent work within the AustKin II project which is devoted to the co-evoluation of marriage and social category systems.

Hannes Wettig, Kirill Reshetnikov and Roman Yangarber ('Using Context and Phonetic Features in Models of Etymological Sound Change') present a novel method for a context-sensitive alignment of cognate words, which relies on the information theoretic concept of Minimum Description Length to decide on the most compact representation of the data given the model. Starting with an initial random alignment for each word pair, their algorithm iteratively rebuilds decision trees for each feature and realigns the corpus while monotonically decreasing the cost function until convergence. They also introduce a novel test for the quality of the models where one word pair is omitted from the training phase. The rules that have been learned are then used to guess one word from the other in the pair. The Levenshtein distance of the correct and the guessed word is then computed to give an idea of how good the model actually learned the regularities in the sound correspondences.

Johann-Mattis List ('LexStat: Automatic Detection of Cognates in Multilingual Wordlists') presents a new method for automatic cognate detection in multilingual wordlists. He combines different approaches to sequence comparison in historical linguistics and evolutionary biology into a new framework which closely models central aspects of the comparative method. The input sequences, i.e. words, are converted to sound classes and their sonority profiles are determined. In step 2, a permutation method is used to create language specific scoring schemes. In step 3, the pairwise distances between all word pairs, based on the language-specific scoring schemes, are computed. In step 4, the sequences are clustered into cognate sets whose average distance is beyond a certain threshold. The method is tested on 9 multilingual wordlists.

5 Final remarks

The breadth and depth of the research collected in this workshop more than testify to the scope and possibilities for applying new methods that combine quantitative methods with not only a sophisticated linguistic understanding of language phenomena, but also with visualization methods coming out of the Computer Science fields of InfoVis and Visual Analytics. The papers in the workshop addressed how the emerging new body of work can provide advances and new insights for questions pertaining to theoretical linguistics

(lexical semantics, derivational morphology, historical linguistics, dialectology and typology) and applied linguistic fields such as second language acquisition and statistical NLP.

6 Acknowledgments

We are indebted to the members of the program committee of the workshop for their effort in thoroughly reviewing the papers: Quentin Atkinson, Christopher Collins, Chris Culy, Dan Dediu, Michael Dunn, Sheila Embleton, Simon Greenhill, Harald Hammarström, Annette Hautli, Wilbert Heeringa, Gerhard Heyer, Eric Holman, Gerhard Jäger, Daniel Keim, Tibor Kiss, Jonas Kuhn, Anke Lüdeling, Steven Moran, John Nerbonne, Gerald Penn, Don Ringe, Christian Rohrdantz, Tandy Warnow, Søren Wichmann.

We also thank the organizers of the EACL 2012 conference for their help in setting up the joint workshop.

References

Christopher Collins, Sheelagh Carpendale, and Gerald Penn. 2009. Docuburst: Visualizing document content using language structure. *Computer Graphics Forum (Proceedings of Eurographics/IEEE-VGTC Symposium on Visualization (EuroVis '09))*, 28(3):1039–1046.

Christopher Collins. 2010. *Interactive Visualizations of Natural Language*. Ph.D. thesis, University of Toronto.

Matthew S. Dryer and Martin Haspelmath, editors. 2011. *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich, 2011 edition.

Michael Dunn, Angela Terrill, Ger Resnik, Robert A. Foley, and Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309(5743):2072–2075.

Russell Gray and Quentin Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origins. *Nature*, 426:435–439.

David LW Hall and Dan Klein. 2010. Finding cognate groups using phylogenies. In *Proceedings of the Association for Computational Linguistics*.

Paul Heggarty, Warren Maguire, and April McMahon. 2010. Splits or waves? trees or webs? how divergence measures and network analysis can unravel language histories. In *Philosophical Transactions of the Royal Society (B)*, volume 365, pages 3829–3843.

John Holm and Peter L. Patrick, editors. 2007. *Comparative Creole Syntax*. London: Battlebridge.

Timo Honkela, Ville Pulkki, and Teuvo Kohonen. 1995. Contextual relations of words in grimm tales, analyzed by self-organizing map. In *Proceedings of International Conference on Artificial Neural Networks (ICANN-95)*, pages 3–7.

Grzegorz Kondrak. 2001. Identifying cognates by phonetic and semantic similarity. In *Proceedings of the North American Chapter of the Association of Computational Linguistics*.

Thomas Mayer, Christian Rohrdantz, Miriam Butt, Frans Plank, and Daniel A. Keim. 2010a. Visualizing vowel harmony. *Linguistic Issues in Language Technology (LiLT)*, 2(4).

Thomas Mayer, Christian Rohrdantz, Frans Plank, Peter Bak, Miriam Butt, and Daniel A. Keim. 2010b. Consonant co-occurrence in stems across languages: Automatic analysis and visualization of a phonotactic constraint. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground, ACL 2010*, pages 70–78.

April McMahon and Robert McMahon. 2006. *Language Classification by Numbers*. OUP.

Petra Neumann, Annie Tat, Torre Zuk, and Sheelagh Carpendale. 2007. Keystrokes: Personalizing typed text with visualization. In *Proceedings of Eurographics IEEE VGTC Symposium on Visualization*.

Christian Rohrdantz, Annette Hautli, Thomas Mayer, Miriam Butt, Daniel A. Keim, and Frans Plank. 2011. Towards tracking semantic change by visual analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 305–310. Portland, Oregon.

Jean Séguy. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane*, 35(138):335–357.

Benjamin Snyder, Regina Barzilay, and Kevin Knight. 2010. A statistical model for lost language decipherment. In *Proceedings of the Association for Computational Linguistics*.