Chapter 30

Quantitative diachronic dialectology

Jelena Prokić
Philipp University of Marburg

In this paper quantitative analysis of sound correspondences in German dialects is presented. This method combines geographical spread of sound change and regularity of sound change within a lexicon. By combining the two, language diffusion processes can be successfully modeled, enabling the researchers not only to identify main dialect groups, but also to look into the individual processes of sound change and track their geographic spread.

1 Introduction

Languages change constantly and this has already been observed by scholarly people in ancient times. While in ancient times language change was considered a 'decay', or 'corruption of a holy speech', modern linguistics sees it as a fact, i.e., an everlasting process that does not lead to 'better' or 'worse' languages, but tries to explain how and why languages change. Unlike historical linguistics, whose main goal is to determine relatedness between languages and language families, dialectology is a study of subdivisions of a particular language and deals with the variation at the micro-level. Apart from determining main dialect groups within a given language area, the focus of dialect research is determining the underlying linguistic processes that lead to the observed distributions of linguistic features. One of the central research questions in dialectology is to discover patterns of dialect interaction, e.g., if a change is spreading in a continuous or discontinuous pattern, determine *focal*, *relic* and *transition* areas of language change and main factors that lead to observed change and it's spread.

In the past three decades there has been an increasing interest in using quantitative methods to address these questions in dialect research. Computational and statistical methods have successfully been applied in tasks that, among many, include automatic detection of dialect groups (Nerbonne et al. 1996; Gooskens & Heeringa 2004; Bolognesi & Heeringa 2002), identification of linguistic features responsible for observed dialect divisions (Wieling & Nerbonne 2011; Prokić, Çöltekin & Nerbonne 2012) and analysis of social determinants of dialect variation (Wieling, Nerbonne &

Baayen 2011). Despite significant progresses in the field of dialectometry, diachronic studies are a rather recent initiative.

In this paper, I will present quantitative analysis of German dialects, in which a network model of dialect change is combined with the automatically extracted regular sound correspondences, in order to model the diffusion process of dialect change and detect centers from which language innovations started spreading from. It is a novel approach in dialectometry proposed by Prokić & Cysouw (2013), that is mainly concerned with shedding more light on the evolutionary processes in dialect formation. In the next section I will first address the importance of network models in dialect evolution, as opposed to tree models used both in traditional and quantitative historical linguistics to model language change.

2 Tree and network models of language evolution

The main assumption of both the comparative method used in traditional historical linguistics and phylogenetic methods exploited in quantitative historical linguistics, is that the underlying model of language change is tree-like. Tree-like model of evolution is based on the assumption that proto-language was completely uniform, without any internal variation, that this proto-language split suddenly into two or more daughter languages, which do not have further contact once they split, and that sound change is completely regular (Bloomfield 1973; Campbell 2004). However, historical linguists are aware that these assumptions are 'reasonable idealizations' (Campbell 2004), since abundant linguistic material shows that proto-languages were not uniform and the split of daughter languages is never sudden. A family tree model alone is not sufficient for dealing with all aspects of relationships between languages, but it is capturing the genetic relatedness between languages, the most important factor in language diversification at the macro-level.

Interaction between neighboring dialects is characterized by an intensive contact on a day-to-day basis, and easy exchange of linguistic material at all levels thanks to the mutual intelligibility of dialects. Language innovations can occur at any location and spread in various geographic directions, sometimes even opposite (Hock 1991). Regions with a long settlement history are characterized by the so-called 'diffusion model' of language change (Chambers & Trudgill 2004: Ch. 11). The diffusion model assumes that innovations spread gradually in a wave-like manner. The area in which the change originates is the so-called *focal* area, where the change is regular. The area in which the change does not happen is the *relic* area. The third area, found in between, is the *transition* area, characterized by less regular change. The consequence of the diffusion model is that dialects form continuums, rather than clearly separated groups, and their historical relationship cannot be described using a tree-model. In order to model processes involved in diachronic dialect change, a network-model is much more suitable, as shown in Prokić & Cysouw (2013). In the network model proposed by Prokić & Cysouw (2013), each site in the data set represents a node in a network and is connected only to the neighboring sites (i.e. nodes). This kind of network representation allows us to explore the spread of language change via diffu-

sion, where changes spread in a wave-like manner from one site to the neighboring sites.

In this paper, German dialect data is analyzed using the described network model. A network representation of German sites can be seen in Figure 1, in which all localities are connected in a such way that each site is connected only to the neighboring sites. In the next section I give a brief overview of the German dialect data used in this paper, including previous quantitative analysis of the data.

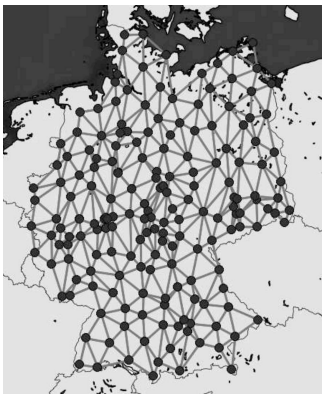

Figure 1: Network model – each site is connected to the neighboring sites.

3 Data

The data presented in this paper comes from the *Phonetischer Atlas von Deutschland* (PAD) which will be published as a part of the *Digitaler Wenker-Atlas* at the Forschungszentrum Deutscher Sprachatlas at the Philipps-Universität Marburg.[1] The PAD data consists of the pronunciations of the Wenker sentences recorded in West Germany during the 1960s and 1970s and recordings of the same sentences made in the former East Germany during the 1990s. There are in total 183 sites included in this survey and they can be seen in Figure 1.[2]

Recorded sentences were segmented into words and transcribed by the phoneticians at the Philipps-Universität Marburg. There are in total pronunciations of 202 words in the digitized version of the PAD data. The data was first digitized using the X-SAMPA phonetic alphabet during the project Visualisation of Language Variation, a joint project between the University of Groningen and the Forschungszentrum Deutscher Sprachatlas at the Philipps-Universität Marburg. X-SAMPA transcribed

[1] `http://www.diwa.info/titel.aspx`.

[2] With the exception of Docelles, which is on the territory of France, all other sites are on the territory of Germany. Data from Docelles was excluded from the present study.

data was later automatically converted into IPA at the Forschungszentrum Deutscher Sprachatlas.[3]

3.1 Previous work

Previous quantitative analysis of the PAD data set includes work by Nerbonne & Siedle (2005), in which the authors rely on the so-called aggregate approach in order to determine the most important dialect groups in Germany. Linguistic distances between each two sites in the data set are estimated using Levenshtein distance which compares pronunciation of two words by calculating the smallest number of operations (insertions, deletions and substitutions) needed to transform one pronunciation into another. Pronunciations of all corresponding words recorded in two sites are compared by means of Levenshtein distance, resulting in the aggregate distance between these two sites. The procedure is repeated for all pairs of sites in the data set, resulting in an $n \times n$ distance matrix, where n is a total number of sites. A detailed description of the Levenshtein approach in dialectometry can be found in Heeringa (2004). The obtained distance matrix is further analyzed using a weighted average clustering algorithm and multidimensional scaling. Both techniques have confirmed that the main dialect split is between High and Low German dialects, splitting the country in two main dialect areas. Using a clustering technique, the authors have detected a further split into five dialect groups, which to some extent correspond to the traditional dialects: Low German, East Central German and Upper German and a heterogeneous area in the west that is divided into two smaller groups. (Figure 2 (left)). Using multidimensional scaling, the authors have detected, next to a two-way split into High and Low German dialects, also Upper German, East Central German, East Low German, and West Low German areas (Figure 2 (right)). Nerbonne & Siedle (2005) have shown that quantitative approaches can successfully be applied on traditional dialect data to detect main dialect areas.

4 Regular sound correspondences

In order to examine diachronic processes responsible for the observed dialect divisions in Germany, in this paper geographic distribution of regular sound correspondences is examined with the help of a network model. Detection of regular sound correspondences between two or more languages is the essential part of the comparative method that aims to postulate regular sound changes, determine if examined languages are genetically related and, if so, reconstruct the proto-language. In traditional historical linguistics, extraction of regular sound correspondences is done manually by comparing lists of potential cognate words. In quantitative historical linguistics and dialectometry, this process is automatized by using algorithms to align cognate words and extract sound correspondences from the alignments (Prokić 2007; List 2012a). In order to estimate the strength of the association (i.e., if a given correspondence is regular or irregular) between any two aligned sounds, Wieling, Prokić

[3] In this paper the IPA version of the data was used.

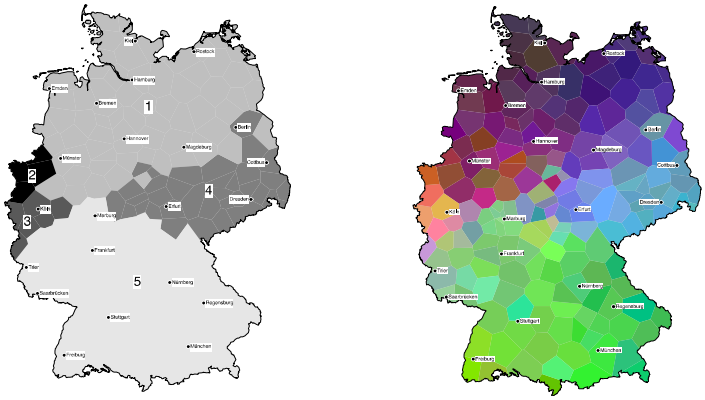

Figure 2: Results of the aggregate analysis of German dialects by Nerbonne & Siedle (2005) using clustering (left) and multidimensional scaling (right).

& Nerbonne (2009) rely on pointwise mutual information (PMI) and show that introduction of this association measure in the alignment procedure leads to improved alignments. Prokić & Cysouw (2013) employ an association measure based on Poisson distribution to detect regular sound correspondences, since this measure does not assume a normal distribution of the data and at the same time gives better estimates for the less frequent sounds. In this paper, I follow Prokić & Cysouw (2013) and use the Poisson association measure to estimate regularity of sound correspondences in the PAD data:

$$\text{Poisson Association} = \text{sign}(O - E) * \left(O * \log\left(\frac{O}{E}\right) - (O - E) \right) \tag{1}$$

where O is the observed co-occurrence and E is the expected co-occurrence of two sounds. The observed co-occurrence represents the number of times two sounds are found aligned in the data, while the expected co-occurrence is the number of times two sounds are expected to be aligned based on their frequency in the data. The expected frequencies E are calculated using the following formula:

$$E_{xy} = \frac{n_x \cdot n_y}{N} \tag{2}$$

where n_x represents the number of times sound x is aligned with any other sound, n_y represents the number of times sound y is aligned with any other sound, and N is the total number of aligned phone pairs in the data.

5 Analysis

Analysis of German dialects presented in this paper relies on a network model shown in Figure 1 to investigate geographic distribution of regular sound correspondences and their association strength. Employment of this network model enables us to compare only geographically neighboring sites and examine, for each sound correspondence individually, if it is regular or not. In traditional dialectology, areas where sound change is regular are considered to be areas where sound change originates. Less regular sound changes are characteristic for the so-called *transition areas*. The method presented in this paper detects areas with regular and irregular sound correspondences from synchronic dialect data, which allows us to infer geographic spread of sound changes. The analysis of German dialect data was performed according to the following steps.

Step 1: phonetic transcriptions were automatically multiple aligned using `lingpy` Python library (List & Moran 2013). Unlike in pairwise alignment employed by Nerbonne & Siedle (2005), in multiple alignment all strings are aligned and compared at the same time. Multiple string alignment is a good technique for discovering patterns in the aligned strings and the advantages of multiple over pairwise string alignment have relatively recently started being recognized in linguistics (Prokić, Wieling & Nerbonne 2009; List 2012b).

Step 2: all sites in the data were plotted on the map and connected to the neighboring sites, forming a network presented in Figure 1.

Step 3: for each pair of neighboring sites, all pairs of corresponding sounds were extracted from the alignments. The association strength (i.e., regularity) for all pairs of sounds was calculated using the Poisson association measure (Equation 1).

Step 4: for each site, the number of irregular correspondences between that site and all neighboring sites was calculated by counting all sound correspondences which association value is under a certain threshold. Following Prokić & Cysouw (2013), all correspondences which association value is smaller than 5 are taken to be irregular. It is a very conservative cut-off, which ensures that only highly irregular correspondences are counted.

Step 5: the number of irregular correspondences for each site was plotted on a map of Germany and analyzed using the Inverse Distance Weighting (IDW) interpolation method as implemented in the Quantum GIS software (QGIS Development Team 2012). Interpolation methods are used in spatial analyses to predict unknown values for any geographic point data based on a limited number of sample data points. They are used for the analysis of continuous spatial phenomena. The Inverse Distance Weighted method estimates unknown values by averaging the values of sample data points in the neighborhood of each processing cell.

5.1 Results

The results of the analysis can be seen on the map in Figure 3. Areas with a high number of irregular sound correspondences are colored red, while the areas with a small number of irregular sound correspondences are represented with a dark blue color. The map shows that areas in the north-west and south-west, colored dark blue, are the areas where many sound correspondences show a high degree of regularity. Two areas in the central part of the country and one in the north-east show a high number of irregular sound correspondences, which suggests that these areas are the so-called *transition areas*. The map in Figure 3 shows a historical split of dialects into northern and southern varieties, which corresponds well with the traditional scholarship that divides German dialect continuum into Low German found in the north and High German found in the south of the country. This split was also detected by method used in Nerbonne & Siedle (2005) that focuses on the synchronic dialect divisions.

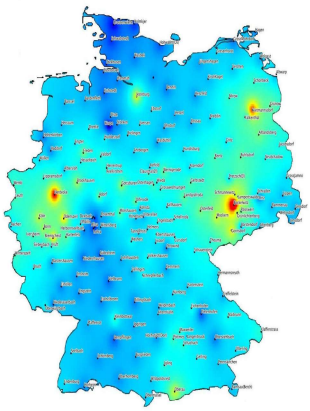

Figure 3: Map of Germany showing two areas with a high number of regular sound correspondences: in the north-west and south-west. The area in the middle of the country shows a high number of irregular sound correspondences.

6 Discussion

The method presented in this paper is the first step toward quantitative diachronic dialectology, that tries to use quantitative and statistical methods in order to examine diachronic processes that are responsible for the observed dialect divisions. The method relies on the network representation of the sites, combined with automatically extracted sound correspondences and their association strength. By analyzing regularity of all sound correspondences in the data set, it is possible to detect historically important areas when it comes to the spread of sound changes. In this paper,

all sound correspondences found in the data are analyzed, resulting in an aggregate diachronic map. Applying this method on a specific sound change would give a much sharper picture of the geographic origin of a sound change in question and it's geographic spread. This type of analysis is beyond the scope of this paper and remains future work.

References

Bloomfield, Leonard. 1973. *Language*. London: Allen & Unwin.

Bolognesi, Roberto & Wilbert Heeringa. 2002. De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten. *Gramma/TTT: tijdschrift voor taalwetenschap* 9(1). 45–84.

Campbell, Lyle. 2004. *Historical linguistics*. 2nd edn. Edinburgh: Edinburgh University Press.

Chambers, J. K. & Peter Trudgill. 2004. *Dialectology*. 2nd edn. Cambridge: Cambridge University Press.

Gooskens, Charlotte & Wilbert Heeringa. 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change* 16. 189–207.

Heeringa, Wilbert. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. University of Groningen PhD dissertation.

Hock, Hans Henrich. 1991. *Principles of historical linguistics*. 2nd edn. Berlin: Mouton de Gruyter.

List, Johann-Mattis. 2012a. LexStat: automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 117–125.

List, Johann-Mattis. 2012b. Multiple sequence alignment in historical linguistics. In Enrico Boone, Kathrin Linke & Maartje Schulpen (eds.), *Proceedings of ConSOLE XIX*, 241–260.

List, Johann-Mattis & Steven Moran. 2013. An open source toolkit for quantitative historical linguistics. In *Proceedings of the ACL 2013 System Demonstrations*, 13–18.

Nerbonne, John, Wilbert Heeringa, Erik van den Hout, Peter van de Kooi, Simone Otten & Willem van de Vis. 1996. Phonetic distance between Dutch dialects. In G. Durieux, W. Daelemans & S. Gills (eds.), *CLIN VI, papers from the sixth CLIN meeting*, 185–202. University of Antwerpen, Antwerpen.

Nerbonne, John & Christine Siedle. 2005. Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede. *Zeitschrift für Dialektologie und Linguistik* 2(72). 129–147.

Prokić, Jelena. 2007. Identifying linguistic structure in a quantitative analysis of dialect pronunciation. In *Proceedings of the ACL 2007 Student Research Workshop*, 61–66. Prague, Czech Republic: Association for Computational Linguistics.

Prokić, Jelena, Çağrı Çöltekin & John Nerbonne. 2012. Detecting shibboleths. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 72–80. Avignon: Association for Computational Linguistics.

Prokić, Jelena & Michael Cysouw. 2013. Combining regular sound correspondences and geographic spread. *Language Dynamics and Change 3.2. Special issue, "Phylogeny and Beyond: Quantitative Diachronic Approaches to Language Diversity", edited by Michael Dunn.*

Prokić, Jelena, Martijn Wieling & John Nerbonne. 2009. Multiple sequence alignments in linguistics. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELT&R 2009)*, 18–25. Association for Computational Linguistics.

QGIS Development Team. 2012. *QGIS Geographic Information System.* Open Source Geospatial Foundation. http://qgis.osgeo.org.

Wieling, Martijn & John Nerbonne. 2011. Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language* 25. 700–715.

Wieling, Martijn, John Nerbonne & R. Harald Baayen. 2011. Quantitative social dialectology: explaining linguistic variation geographically and socially. *PLoS One* 6(9).

Wieling, Martijn, Jelena Prokić & John Nerbonne. 2009. Evaluating the pairwise string alignment of pronunciations. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELT&R 2009)*, 26–34. Association for Computational Linguistics.