Chapter 37

Identifying dialect regions from syntactic data

Erik Tjong Kim Sang
Meertens Institute Amsterdam

The *Syntactic atlas of Dutch dialects* (SAND) is a database of syntactic features observed in the language spoken by people from different dialect regions in The Netherlands and Flanders. We would like to know how specific syntactic features are for the different dialects. For this purpose we try to generate dialect boundaries from the syntactic data only. We show that a plausible binary division of the dialect can successfully be derived but that is more difficult to divide the data in three or more regions. We build on earlier work by Nerbonne, Heeringa & Kleiweg (1999), who performed this task for phonetic data, and on work by Spruit (2008), who also attempted to identify dialect regions from syntactic data.

1 Introduction

One of my early memories of the research work of John Nerbonne goes back to 1999, to the yearly meeting of the European Chapter of the Association of Computational Linguistics, held in Bergen, Norway. At that meeting John presented a paper, joint work with colleagues Wilbert Heeringa and Peter Kleiweg, in which they outlined how gradual differences between dialects could be derived and visualized from dialect data using computational methods (Nerbonne, Heeringa & Kleiweg 1999). Their visualization, a beautiful color map displaying gradual transitions between dialects as diverse as West Low German, Limburgish and West-Flemish, was the first color page ever in a proceedings of a major computational linguistics conference.

The paper and its followup work has inspired much other work, such as the Gabmap website in Groningen (Nerbonne et al. 2011), which enables researchers to visualize their own dialect data, and the project Maps and Grammar in Amsterdam (Barbiers et al. 2008), which examines the relation between linguistic properties and their geographic distribution. It is in the latter project that the current paper is based.

One of the research questions of the Maps and Grammar project is central to this paper. It is inspired by the data set we have available in the project: the *Syntac-*

tic atlas of the Dutch dialects. Unlike John Nerbonne and his colleagues, who used pronunciation data, we exclusively want to use this syntactic data set for deriving dialect boundaries. Therefore our research question is: *can reasonable dialect boundaries automatically be derived from syntactic linguistic data only?* In order to answer this question, we analyze our research data, compute linguistic distances between geographic locations and divide the locations in groups based on the distances.

After this introduction, we will discuss related work in section two. In the third section we will present our data while section four outlines our methods for analyzing and clustering the data. In the same section we present and discuss some results. We conclude in section five.

2 Related work

Nerbonne, Heeringa & Kleiweg (1999) showed how mathematical methods such as the Levenshtein distance, clustering and multidimensional scaling can be used for comparing word pronunciations, computing Dutch dialect distances and visualizing the dialects on a map. Similar work has been done by Goebl (2010) for French dialects. Spruit (2008) was the first to apply clustering and multidimensional scaling to syntactic dialect data, from the Dutch SAND database (Barbiers et al. 2008).

Like with data sets from other fields, a challenge in working with linguistic data is how to deal with missing data. Spruit (2008) replaced all missing data of the SAND with the zero values but as a result of this underspecified dialects may appear more similar than they really are. van Craenenbroeck (2014) adopted a similar strategy for locations visited by the SAND team and used mathematical methods for estimating data values in locations skipped by the team. Tjong Kim Sang (2015) proposed to keep all unknown values in the data and to use a new distance function for comparing the values. We will adopt this method when dealing with missing data (see section 4).

3 Data

We use linguistic data from the Syntactic Atlas for Dutch Dialects (SAND) (Barbiers et al. 2008). This data is based on interviews with people from 267 locations in The Netherlands and Flanders, the Dutch-speaking part of Belgium. The interviews have been transcribed and have been checked for the presence of predefined syntactic variables. We use the subset of 220 syntactic variables which are available from the digital version of the atlas: DynaSAND (Barbiers 2006).

A problem of the data set is the limited availability of negative values. The data was developed as a source for dialect maps, which means that the focus in the collection process was on finding locations where syntactic features were present and not on marking locations where they were absent. Therefore only 17% of the maps in the atlas contain explicit negative data (Tjong Kim Sang 2014).

For data modeling, different data values are required in order to be able to distinguish separate groups. For this purpose, earlier studies that used the SAND data

have used the evidence of absence assumption: if a syntactic feature or value was not observed at a location then it was assumed that the variable did not occur in the local dialect (Spruit 2008; van Craenenbroeck 2014). While this assumption might be perfectly valid in other contexts, its application here can be challenged. The interviews were held by different interviewers which asked different questions. So data feature absence might also be caused by the interviewers rather than only by the local dialect. Therefore we did not use the evidence of absence assumption and kept all unknown feature values in our data (marked as *?*).

In the SAND data, most (81%) of the syntactic features are not binary but contain many values (Tjong Kim Sang 2014). An example of this is verb order in phrases with three verbs (A, B, C), which may contain up to six different values (ABC, ACB, BAC, BCA, CAB, CBA). Previous studies that used the data have represented such features as sets of binary values (Spruit 2008; van Craenenbroeck 2014). For example, six word order patterns are represented as six binary values. This causes such features to be represented in the model several times which increases their importance to the model. While the best weight of each feature is unknown, we believe it is unfair to give certain syntactic features more weight because of their shape. Therefore we use each feature only once in the model and use combined feature values when local dialects allow more a feature to have more than one value.

4 Method and results

We used k-means clustering (Manning & Schutze 1999) for dividing the 267 locations in separate regions. For this purpose, it was necessary to define what the distance between two feature values was. This was nontrivial for our data, since they contain basic feature values, sets of basic feature values and unknown values. We define the distance between two basic values as 0 when they are equal and 1 when they are different. The distance between sets of values is 0 when the sets share a common value and 1 otherwise. The distance between an unknown value and any other value (including unknown) is defined as 0.5.

Next, we define the distance between two locations as the sum of the distances between their syntactic feature values. We used these settings for the *k*-means algorithm to compute the best two regions starting from two randomly chosen region centers. This experiment was repeated 100 times and each time the algorithm generated the same two regions with the locations Erica in Drenthe and Bevere in East Flanders as centers. The resulting map can be found in Figure 1 (Left).

The clustering algorithm has divided the Dutch-speaking community along the national border: The Netherlands vs. Flanders (left part of Figure 1). Only in the southwest and in south-east the linguistic boundary did not follow the national border. In the southwest, three Dutch locations were classified as Flemish: Oostburg, Hoek and Hulst (see the bottom right of Figure 1). All three are part of the region Zeelandic Flanders which is separated by the rest of The Netherlands by water. Until 2003, when a tunnel connection was opened, the only way to reach the region was either by boat or by a detour over land via Belgium. For this reason, the region has strong

ties with the neighboring Flanders. Dialect researchers Daan and Blok, classified Oostburg and Hoek among the Dutch dialects and Hulst among the Flemish dialects (Daan & Blok 1969)

In the southeast, 13 Flemish cities Bree, Eigenbilzen, Genk, Grote-Spouwen, Hamont, Hasselt, Houthalen, Jeuk, Lauw, Maaseik, Opglabbeek, Stokkem and Tongeren were classified as Dutch. This is not surprising: the dialect spoken in these cities is more common to the dialect of the neighboring Dutch province than to the neighboring Flemish region. The dialect map of Daan & Blok (1969) classified all of Limburgish dialects as one big group, including four locations which assigned to the Flemish group by *k*-means: Borgloon, Eksel, Lummen and Sint-Truiden (see the top right part of Figure 1).

Next we applied the *k*-means algorithm for dividing the Dutch language in three regions. We ran the algorithm 100 times from sets of three randomly chosen initial locations as centers of the regions, but this time 21 different region divisions were suggested. Most of them had one (43%) or two (54%) of the region centers in common with the two-region division of Figure 1. However, this time the proposed regions seemed less sensible. For example, the most frequent set (19%) with region centers Bevere, Boskoop and Erica, split Flanders in two parts but allocated part of The Netherlands to one of the parts (see Figure 2, left part). Only three of the 21 suggestions corresponded with known dialect borders to some extent: one which identified Frisian, be it with some unrelated locations (Figure 2, center), one which recognized French Flanders, with one additional location (Figure 2, right) and one in which East Flanders was separated from West Flanders.

k-means did not work well for dividing the linguistic space in three regions. The algorithm did not converge to a single configuration and few of the proposed sets were reasonable. This did not improve when we attempted to divide the space in four regions. In 100 runs, the algorithm generated 56 different region sets. The most frequent of these (11%) was an extension of the most frequent set of three regions (Figure 2, left). None of the regions were complete and Belgian Limburg contained a section of four adjacent nodes which belonged to four different regions. None of the four-region sets that we inspected was very good and we conclude from this that for our data set *k*-means does not work very well in generating more than two regions.

5 Concluding remarks

We have applied *k*-means clustering to a set of syntactic dialect data in order to find out if it was possible to derive reasonable dialect boundaries. For the case of dividing the data set in two regions, this worked well. In 100 runs, the clustering algorithm always converged to the same two regions without any gaps or separate islands. This is impressive because no geometric data was used during the classification process. The syntactic space which was used for classification, is different from geometric space. The location of the region centers is evidence of this: they are by no means the geographic mean points of their region (Figure 1, left, the blank locations Bevere for Flanders in the green area and Erica for The Netherlands in the red area).

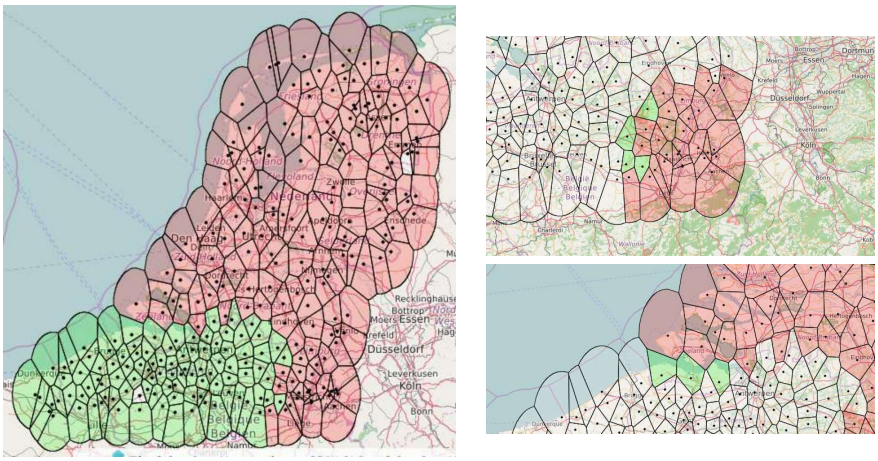

Figure 1: **Left**: a map with the two main Dutch-speaking regions identified by the *k*-means algorithm from the syntactic SAND data. **Right top**: the transnational area where Limburgish is spoken according to Daan & Blok (1969). **Right bottom**: the locations Oostburg, Hoek and Hulst in Zeelandic Flanders (green) are the only three Dutch locations classified as Flemish by *k*-means.

However, *k*-means performed less well for generating more than two regions. In the three-region case, it suggested 21 different region configurations, a few of which made some sense but these were not the most frequently proposed ones. None of the proposed sets of four regions that we checked, was reasonable. However, the few plausible cases for the three-region case (Figure 2, center and right) show that the data set contains useful information for automatic dialect boundary detection. We just might not have found the best algorithms to use this information.

There are different directions for future study. First, we would like to explore an idea of Heeringa (2004) for dealing with missing data: to use average scores as distances so that missing data values can safely be ignored. Next we would like to experiment with alternative hard clustering algorithms. We have already performed initial experiments with Expectation Maximization (Fraley et al. 2012), which generate similar results for the two-region case and suggest that there is even more useful information for this task in the data than we found with *k*-means clustering. We also would like to go back from the detected regions to the syntactic variables to check if certain dialect boundaries are predicted by individual syntactic variables. Some dialects may not be separated by a sharp boundary but by a fuzzy boundary, like on the map of Daan & Blok (1969). It would be interesting to see if we can identify such transition zones with computational measures. And finally, it would be interesting to apply these clustering methods to mixed data, for example a data set which included both syntactic and phonological data.

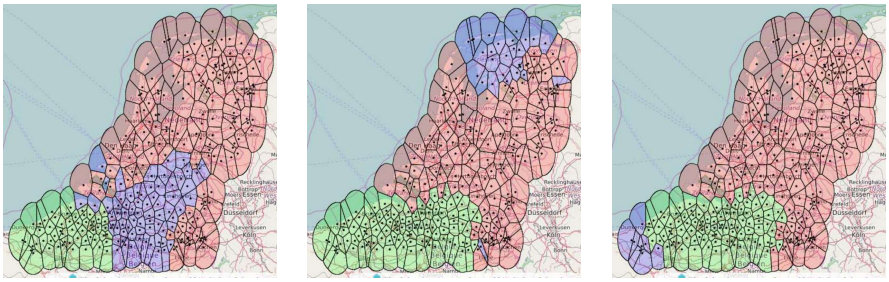

Figure 2: *k*-means suggested 21 different sets of three regions in 100 runs. **Left**: the most frequent set (19%) which split Flanders and assigned several locations from the Dutch South to one of the new regions. **Center**: a division in which the province Friesland has been identified, plus six unrelated locations (set frequency: 6%). **Right**: a configuration in which French Flanders, with one extra location, has been separated from the rest of Flanders (set frequency: 2%).

The work described in this paper was inspired by earlier work by John Nerbonne, in particular Nerbonne, Heeringa & Kleiweg (1999). Like John Nerbonne, we aim at creating models and visualizations for large sets of dialect data. Even though John is leaving his group in Groningen, his work will keep to have an influence in the field of dialectology in the coming decade. We wish him well in his future activities.

References

Barbiers, Sjef. 2006. *Dynamische Syntactische Atlas van de Nederlandse Dialecten (DynaSAND)*. http://www.meertens.knaw.nl/sand/ Accessed 26 February 2015.

Barbiers, Sjef, Johan van de Auwera, Hans Bennis, Eefje Boef, Gunther Vogelaer & Margreet van der Ham. 2008. *Syntactic atlas of the Dutch dialects*. Amsterdam University Press.

Daan, Jo & D. P. Blok. 1969. *Van Randstad tot Landrand*. Noord-Hollandsche Uitgevers Maatschappij, Amsterdam, The Netherlands.

Fraley, Chris, Adrian E. Raftery, T. Brendan Murphy & Luca Scrucca. 2012. *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*.

Goebl, Hans. 2010. Dialectometry: theoretical prerequisites, practical problems and concrete applications. *Dialectologia* Special Issue(I). 63–77.

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

Manning, Christopher D. & Hinrich Schutze. 1999. *Foundations statistical natural language processing*. MIT Press.

Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg & Therese Leinonen. 2011. Gabmap – a web application for dialectology. *Dialectologia: revista electrónica* (Special Issue II).

Nerbonne, John, Wilbert Heeringa & Peter Kleiweg. 1999. Comparison and classification of dialects. In *Proceedings of EACL99*. ACL, Bergen, Norway.

Spruit, Marco René. 2008. *Quantitative perspectives on syntactic variation in Dutch dialects*. PhD thesis, LOT, Utrecht, The Netherlands.

Tjong Kim Sang, Erik. 2014. *SAND: relation between the database and printed maps*. Tech. rep. Meertens Institute, Amsterdam, The Netherlands.

Tjong Kim Sang, Erik. 2015. *Discovering Dialect Regions in Syntactic Dialect Data*.

van Craenenbroeck, Jeroen. 2014. *The signal and noise in Dutch verb clusters – A quantitative search for parameters*. Manuscript, http://jeroenvancraenenbroeck.net/s/paper_signal_noise.pdf , version 18 December 2014, Retrieved 26 February 2015.