Arvid: A Computational Tool for Dialect Research

Erik Tjong Kim Sang, Lotte Hendriks

Meertens Institute
Amsterdam, The Netherlands
{erik.tjong.kim.sang,lotte.hendriks}@meertens.knaw.nl

Abstract

We present Arvid, a platform-independent computational tool for studying dialects. We demonstrate that this tool can be used for improving both quantity and quality of research work. With Arvid, researchers can visualize data, combine data and hence test hypotheses in a flexible way. The tool has helped us evaluating more complex data models than possible earlier.

Keywords: Dialect analysis, Data visualization, Tools

1. Introduction

Dialect research involves data collection, data analysis, model building and model testing. For decades, these processes required tedious manual labor but in recent years many dialect researchers have started using computational tools for analyzing data and testing models (Heeringa, 2004; Spruit, 2008; Goebl, 2010). In this paper, we describe and evaluate a new tool for visualizing and combining dialect data and for testing dialect models: Arvid. After this introduction, we present some related work. Next we describe the tool and give examples of its application. We finish the paper with some concluding remarks.

2. Related work

Heeringa (2004) compared dialects based on automatically computed distances between word pronunciations. He evaluated different methods for phonemes and computed distances between pairs of words with the Levenshtein method, which aligns the phonemes of two words and finds the smallest set of transformation rules that are required to convert one word into the other. This distance estimation method proved to be a good way for identifying dialect groups, both in Norwegian and Dutch data.

Spruit (2008) investigates syntactic dialect differences. By applying multi-dimensional scaling on value differences of syntactic features of dialects, these features can be projected on colored maps and the color codes on these maps can be used to identify dialect regions. The resulting dialect regions proved to be similar to the ones derived by Heeringa (2004) and a substantial part of the syntactic differences between dialects could be explained by geographical distance. Finally, Spruit employs the method Gewichteter Identitätswert (GIW) to show that syntax, lexicon and pronunciation of dialects are related, although the relation is not as strong as one could have expected.

Goebl (2010) presents a summary of the dialect research performed at the University of Salzburg for the last decade and more. Among others, he describes the MS Windows program Visual DialectoMetry (VDM)[1] which can be used for visualizing and studying dialect data. The paper focuses on the data of the Atlas Linguistique de la France (ALF), collected between 1902 and 1910. Five types of automatically constructed dialect maps are discussed: similarity maps, honey comb maps, beam maps, parameter maps and dendrographic maps, and examples are given of how these can be used for identifying similar regions, region boundaries and transition zones.

The tool described in paper, was tested with Dutch verb cluster interruption data collected in the study of Hendriks (2013). The study focuses on the distribution of verb cluster interruptions in Flanders. For this purpose, interviews were conducted at forty dialect locations. The interview responses were analyzed and the results were organized in 169 binary-valued linguistic features. For various reasons, 9% of the feature values are missing, which poses an additional challenge for automatic analysis.

3. Arvid

We built the tool Arvid for supporting dialect research with four main functions: data visualization, linguistic feature combination, location combination and theory testing. The next sections discuss each of these four functions.

3.1. Data visualization

We use the RUG-L04 package (Kleiweg, 2004) for visualizing dialect data on maps. The package includes a map of the Dutch language area (The Netherlands and Flanders). We removed the areas that were not used in the data of Hendriks (2013) (The Netherlands and the Flemish province Antwerp) and added a list of the forty relevant dialect locations with their latitudes and longitudes. Finally, we created a web interface[2] to the software and data because this enabled platform-independent access to the tool. So anybody with a web browser and Internet access can experiment with Arvid.

Figure 1A contains an example of a dialect feature displayed on a map. The ability to visualize any of the 169 linguistic features, improved our understanding of the data. By comparing feature maps, we can now immediately identify locations in which related features have similar or different values. The new maps are also a useful source for images which can be used in presentations for colleagues and students.

[1] http://ald.sbg.ac.at/dm/germ/VDM/

[2] http://ifarm.nl/maps/arvid

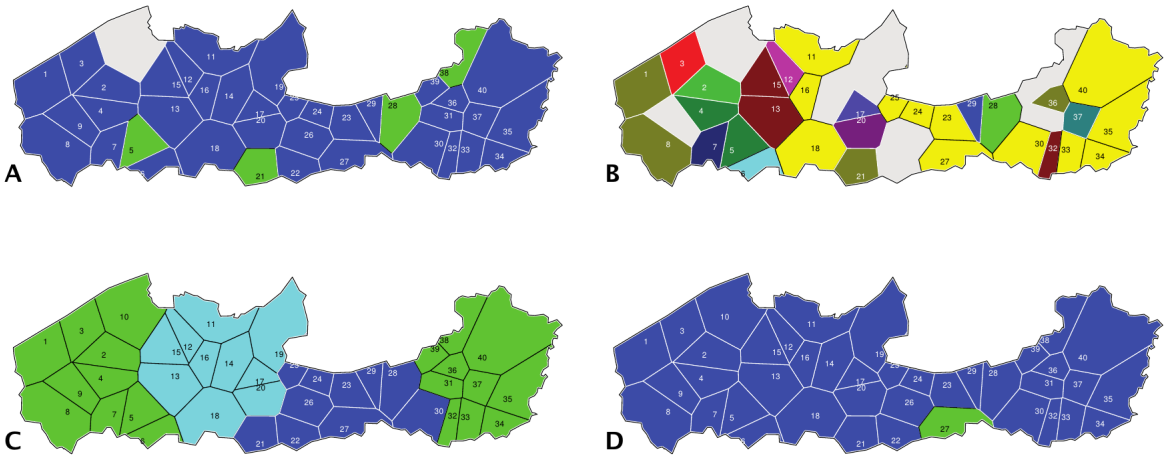

Figure 1: Four dialect maps created with Arvid: A: adverbial particle interruption (aspect): allowed (blue) and not allowed (green); B: a combination of the six features involving adpositional particle interruption: in the most frequent variant (yellow) each interruption type is allowed; C: a combination of the six features involving predicative adjective interruption examined on province level (location combination): the features have identical values in West-Flanders (left, green) and Limburg (right); D: test results for the hierarchical rule: predicate adjective interruption with the verb *moet* (*must*) requires adverbial particle interruption with the same verb: Overijse (green; 27) is the only counterexample.

3.2. Combining linguistic features

Within our set of 169 linguistic features, several groups of related features can be found. For example, there are six features related to adpositional particle interruption and four features related to an interruption by the word *mogelijk* (*perhaps*). Therefore we have included the option to visualize groups of features. Groups of combined features have values which consist of lists of feature values. This means that the number of values of such a group is usually much larger than the three values that an individual feature can take (present, not present and unknown). Thus the maps of combined features usually require more than three colors.

Figure 1B contains an example of combined dialect features displayed on a map. The main insight we had from inspecting maps of combined features, was that feature values were spread over the map heterogeneously. So unlike we hoped, similar feature values were not clustered on the map. While this was an interesting observation, it was not new: Hendriks had already observed that it was hard to identify geographical patterns in the feature values.

3.3. Combining locations

An important problem with dialect data is data sparseness. Because of the costs of collecting the data, only a small amount of information is available per location. For example, in the case of the data of Hendriks (2013), each location is only associated with dialect data originating from a single speaker. While this condition is common for dialect corpora, it will cause noise to have a large influence on data analysis.

We can decrease the influence of noise by examining regions containing groups of locations rather than individual locations. In that case, the question is what values we will assign to the regions. We do not want to use feature sets, since this would make a comparison of regions with different numbers of locations impossible. If all locations of a group have the same feature values assigned to them, then combining feature values is trivial. But what should we do when two locations have different values for a linguistic feature?

When combining locations with different linguistic feature values, we have chosen to use the value that occurs most often (majority). In case there was a tie between two or more values, we selected the feature value that was highest in the ranking *present* > *not present* > *unknown*.

Figure 1B contains an example of a dialect feature displayed on a map while combining locations by province. Maps of feature values of combined locations give quick insights in the differences between regions. We most commonly use this tool property for comparing dialect usage in provinces. It is now easy to find out which linguistic features are shared among provinces. We also use this tool property in combining the feature grouping described in the previous section in order to find out feature group similarities among provinces.

3.4. Evaluating dialect theories

An important finding of the Hendriks (2013) study, is a hierarchy of verb cluster interruptions. This hierarchy was based on the frequency of the verb cluster interruptions. For example particle interruptions were acceptable for 93% of all dialect informants while adverb interruptions were acceptable only for 8% of the group. For this reason particles were placed higher in the interruption hierarchy than adverbs.

While the resulting hierarchy is interesting, a much stronger hierarchical theory could have been proposed if the location information in the data had been used. Such a theory would predict that certain linguistic features are only also present in a dialect if other linguistic features are present. However, the size of the data set and its lack of geographically clustered feature values made it difficult to evaluate such a

theory in the time frame of the study.

Testing if features values satisfy a hierarchical constraint, can be performed automatically. If we want to evaluate a rule stating that a dialect can only have linguistic feature B if it has feature A, then we should test if feature B is present while feature A is absent. Only in that case, the rule is rejected. Since some of the feature values are missing, we also need to decide how to deal with unknown features (see Table 1).

Feature A	Feature B	Assessment
present	present	accepted
present	not present	accepted
present	unknown	accepted
not present	present	rejected
not present	not present	accepted
not present	unknown	unknown
unknown	present	unknown
unknown	not present	accepted
unknown	unknown	unknown

Table 1: Feature pairs and their impact on testing the rule *feature B is only present if feature A is present*.

After including the values of Table 1 we can test any pair of features or feature combinations for a hierarchical relation and examine the results of the test on a map. For example, we found that every dialect which allows predicative adjective interruptions, also allowed adpositional particle interruptions. However, the same strict relation was not present between adverbial particle interruptions and predicative adjective interruption: one dialect (Overijse) proved to be a counterexample. The map of the second relation test can be found in Figure 1D.

4. Concluding remarks and future work

We have presented a new platform-independent tool for dialect research, Arvid, which enables researchers to improve the quantity and quality of their research. We have demonstrated different properties of the tool for visualizing data, combining linguistic features, combining locations and evaluating dialect theories. We have also shown that Arvid offers interesting insights in dialect data.

The tool offers map views on feature values in different locations which allow for quick detection of similarities and differences. However, finding interesting observations, like hierarchy violations, may require inspecting many different maps. An automatic function which evaluates large numbers of feature sets and looks for aspects which could be interesting for dialect research, is an important topic of our future work.

5. References

Hans Goebl. 2010. Dialectometry: theoretical prerequisites, practical problems, and concrete applications (mainly with examples drawn from the 'Atlas linguistique de la France', 1902-1910). *Dialectologia*, special issue I:63–77.

Wilbert Heeringa. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*, volume 46 of *Groningen Dissertations in Linguistics*. University of Groningen, The Netherlands.

Lotte Hendriks. 2013. *Verbal breakups and the interaction between syntactic structure and processing*. MSc thesis, Utrecht Institute of Linguistics OTS, Utrecht University, The Netherlands.

Peter Kleiweg. 2004. *RuG/L04 - Software for dialectometrics and cartography*. http://www.let.rug.nl/ kleiweg/L04/.

Marco René Spruit. 2008. *Quantitative perspectives on syntactic variation in Dutch dialects*, volume 174 of *LOT Dissertation Series*. LOT, Utrecht, The Netherlands.

Number of words: 1989 (estimated with `wc`)