Exploring Language Variation Across Europe
A Web-based Tool for Computational Sociolinguistics

Dirk Hovy and Anders Johannsen
Center for Language Technology, University of Copenhagen
{dirk.hovy,ajohannsen}@hum.ku.dk

Abstract

Language varies not only between countries, but also along regional and socio-demographic lines. This variation is one of the driving factors behind language change. However, investigating language variation is a complex undertaking: the more factors we want to consider, the more data we need. Traditional qualitative methods are not well-suited to do this, and therefore restricted to isolated factors. This reduction limits the potential insights, and risks attributing undue importance to easily observed factors. While there is a large interest in linguistics to increase the quantitative aspect of such studies, it requires training in *both* variational linguistics and computational methods, a combination that is still not common. We take a first step here to alleviating the problem by providing an interface, www.languagevariation.com, to explore large-scale language variation along multiple socio-demographic factors – without programming knowledge. It makes use of large amounts of data and provides statistical analyses, maps, and interactive features that will enable scholars to explore language variation in a data-driven way.

Keywords: language variation, social media, web explorer, computational sociolinguistics

1. Introduction

Language varies. Not only over time, but also between people who live at the same time, and within the same person between different situations. We do not expect a teenager to speak the same way as a pensioner, and we would not talk the same way at a dinner party as we would at a scientific conference, or at a six-year-old's birthday party. We use language, sometimes actively, more often subconsciously, to mark our membership in a group, defined by socio-demographic factors: *age* (Barke, 2000; Barbieri, 2008; Rickford and Price, 2013), *gender* (Holmes, 1997; Rickford and Price, 2013), *regional origin* (Schmidt and Herrgen, 2001; Nerbonne, 2003; Wieling et al., 2011), *social class* (Labov, 1964; Milroy and Milroy, 1992; Macaulay, 2001; Macaulay, 2002), *ethnicity* (Carter, 2013; Rickford and Price, 2013), and many more. At the same time, we use language to distinguish ourselves from other groups within the same socio-demographic category (Silverstein, 2003; Agha, 2005): young vs. old, men vs. women, town vs. country. Since everyone is at the intersection of multiple such socio-demographic groups, language is affected by the interaction of many of these competing factors.

Interactions between demographic factors can help explain similarities between otherwise distinct groups, say between age and gender, if we observe the same phenomenon in young women and elder men. So while interactions provide us with more explanatory power, they come at a cost: the more factors we consider, the more ways to interact there are. For two factors (say, age and gender) there is only one possible interaction, but this number grows exponentially, as a binomial coefficient of the number of factors. Tripling the number of factors to six increases the number of potential pairwise interactions to 15.

And so, while language variation is an active research area in linguistics, most research to date has only studied isolated socio-demographic factors, or at most interactions of two. This reduction is due to the combinatorial complexity and the methodological challenges consequently faced by traditional, qualitative analysis methods. This complex interaction has a direct impact on methodology: as the example above shows, considering more than two factors quickly becomes unfeasible for manual analysis, both because of the number of interactions to consider, and the amount of data needed.

However, reducing the complexity of the problem to individual factors risks missing important interactions, and attributing higher importance to easily observed factors. Robust analysis of interactions, though, is crucial in answering central research questions of language variation, such as "*Who are really the drivers of language change, and under which conditions?*"

In order to address this question, we need to take more and more data into account. Traditional, qualitative methods are often not designed to handle more than a handful of data points (albeit in depth). In contrast, computational, quantitative methods offer the possibility to explore language variation at an unprecedented scale. However, most computational methods require formal training in computer science or related fields, a training that is not yet a standard part of linguistics curricula. This absence is not for lack of interest, but because of the time required to master many of these techniques. Luckily, large-scale processing can be facilitated by computational tools which do not require in-depth knowledge of programming. On the opposite side of the field, computer scientists, who possess the necessary programming skills, are often neither aware nor particularly interested in many linguistic research questions.

The goal of the tool we present, Humboldt,[1] is therefore to enable linguists to explore language variation at a large scale.

We provide a web interface that allows the user to query for lexical phenomena in several languages, and to get both

[1] Named after the Humboldt brothers, because it, too, combines their interests: linguistic knowledge and scientific exploration.

statistical analysis and map representations of the results along several demographic factors.

The interface gives linguists capabilities that would otherwise require programming experience, training in statistical methods, knowledge of geographical information systems, and expertise in natural language processing.

The main contributions of this new interface, publicly available at `www.languagevariation.com`, are:

- Automated comparison of variants along multiple demographic dimensions using proper statistical methods.
- Spelling variations, pervasive in internet media, may be automatically grouped when spelling is not the focus of the investigation.
- Fast and responsible interface, enabling an open exploration cycle, in which the results of one search engenders new hypotheses, leading to new searches.

2. Backend

We use data collected from a web source, namely online reviews of companies. The data contains both text and meta-information about the authors, as described in (Hovy et al., 2015), and is representative with respect to age, gender and spatial distribution. The reviews cover a wide range of European languages. The data was additionally augmented with gender-information based on first names, and geolocation data.

We store the data in an inverted-indexing database, Apache Solr[2], where documents are indexed by the words they contain. The database allows for fast keyword search, but also enables aggregation over age, gender, and location. Since we are interested in the influence of those factors on language, but use a search over the words, an inverted indexing scheme as offered by Solr makes sense. Under this scheme, each word is a associated with a list of all documents they occur in. Similarly, a demographic attribute (e.g. county) is represented as a list of documents authored by a person with that attribute. Identifying documents with both traits can be done efficiently as it amounts to list intersection. Grouping by demographic attributes may therefore be performed in real time for user-provided queries.

2.1. Spelling variation and soft matches

Any user-generated content usually contains a lot of spelling variation, either due to neologisms, or simply because of orthographic errors. When searching for deliberate neologisms (such as "regional" spelling variants), we want hard matching: we know which variants we expect, and want to get an estimate of how frequently they are used by different demographics.

However, when comparing two expressions with each other, we might not care about spelling variations. Instead, we would like to get an overall picture of how often each of them is used, including any spelling variants. For this case, we need to introduce a soft matching. Our initial solution to the problem is to use n-gram indexes over characters. We index bi- and tri-grams, and – at query time – search for any word that matches the initial tri-gram and final bigram (since these positions are less likely to be targeted by accidental misspellings), and a number of bigrams in between, which is a parameter that can be set by the user. This allows, for example, to find all of the more than 300 variants of "definitely", including *definately*, *deffinitely*, or *defanately* (see Section 4.1. for the example).

3. Frontend

The goal of the interface is to enable user without programming experience to explore language variation at scale. As for now, the interface allows the exploration along the three main demographic axes: age, gender, and location. In order to facilitate open exploration of ideas, we opted for a strategy where the user has to specify only minimal input, but receives all potentially interesting analyses.

Users can enter one or two sequences of key words, and select a country and language. For now, we support Danish in Denmark and English in the United Kingdom, and more language-country pairs are in preparation (German-Germany, French-France, Dutch-Netherlands).

If only one term is entered, the display will show the distribution of that one term. If two keywords are supplied, the interface will show a split comparison of the terms. Each keyword can be a single term, a multiword expression, or a list of alternatives. The latter allows the comparison of entire word groups or spelling variations of the same word. The interface will then query the data base table for the keywords, and retrieve all relevant entries. The data is aggregated and displayed according to age, gender, and region.

For age and gender, there are potentially confounding factors at play: the same phenomenon might be observed for both men and women, but could differ along age: say it is used by young women and older men. Analyzing only one of the factors would hide this fact, so we display the results together in one graph. Separate graphs are used for the different keywords.

Regional variation is still a large factor, and the geo-coded information in our database allows us to explore this dimension as well. Since regional coverage can be sparse, we aggregate at the county level, and use relative rather than absolute counts (absolute counts would skew the picture towards populous counties). The relative frequencies are then displayed as a heat map of the respective country, with mouse-over information of the frequency and the country name. For counties we use the European Union *Nomenclature of territorial units for statistics* (NUTS) classification scheme.

4. Example Case

We illustrate our interface showing a simple example for a common English spelling variation, *definately* instead of *definitely*.

4.1. The case of *definitely*

One common spelling mistake in British English is *definately* instead of *definitely*. One might wonder whether this is simply a random mistake, or whether there are any underlying variable governing it. A typing error seems unlikely,

[2] `http://lucene.apache.org/solr/`

given that the differing keys `i` and `a` are separated on the keyboard.

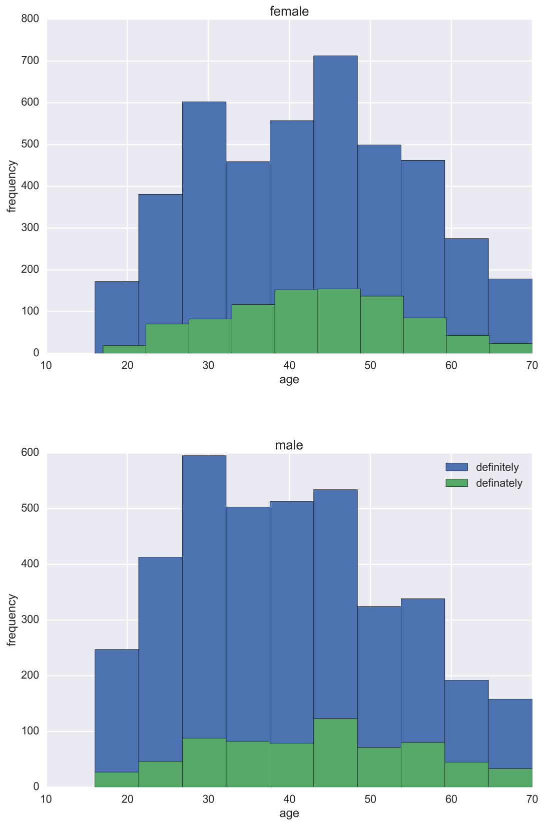

Figure 1: Frequency of *definitely/definately* spelling variations in English data by age and gender

Looking for differences in age and gender (see Figure 1) does not show any substantial differences in use.

However, when plotting the regional distribution on a map, we can see that the occurrences are centered around the regions of the Midlands (see Figure 2). This makes a dialect-phonological feature a likely source for the variant.

5. Discussion

Analysis is so far restricted to three demographic factors, due to availability in the source data. However, as we have shown, these three factors already allow for a number of variational studies.

The current source of the data – online reviews – is certainly a biasing factor: people discuss only a limited range of topics, and they potentially use a special register for this text genre (a "*reviewese*"). However, our setup is extensible, and could incorporate further demographically annotated sources (e.g., social media such as Twitter) in the future, if we have access.

We also plan to enable syntactically based search, both based on word classes, as well as syntactic constructions (Johannsen et al., 2015). In order to produce reliable results, though, we require a reliable way of processing non-standard language, namely POS taggers and parsers. As recently shown by (Hovy and Søgaard, 2015; Jørgensen et al., 2015), reliability of these tools for non-canonical data is still uneven.

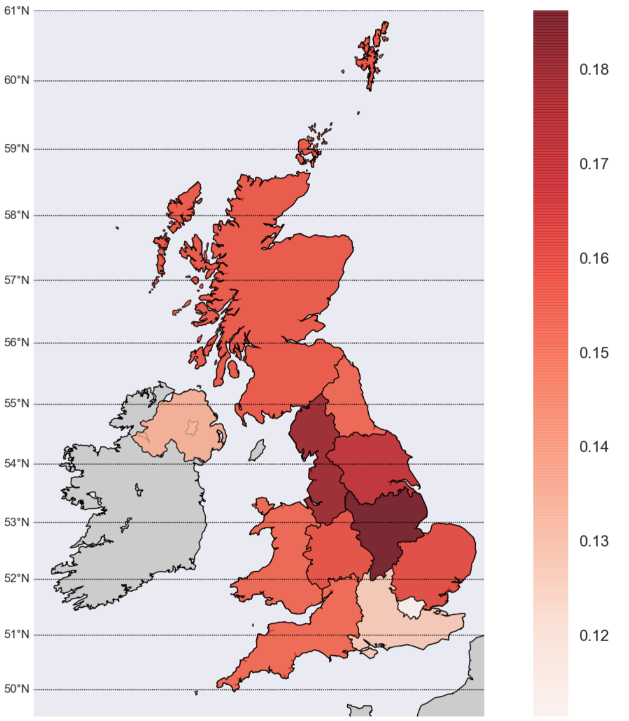

Figure 2: Ratio of *definately* spelling in English data by NUTS regions

In the meantime, we hope for active feedback from our users and plan to incorporate their suggestions and wishes. We believe that the main drivers of such a tool's functionality should be the practitioners who use it.

6. Related Work

A number of similar online projects exist. The most well-known one is certainly the Google Ngram corpus (Michel et al., 2011), which enables lexical search over enormous amounts of text. However, it does not include demographic factors, only time of publication, and has recently been criticized for inherent biases (Pechenick et al., 2015). The German Language Atlas (DSA) allows online browsing of mostly historical data on maps (Schmidt and Herrgen, 2001). It does not contain a demographic search option, although information about subjects can be obtained manually. Lastly, an explorer for lexical variation across Swedish regions exists,[3] but we were not able to find a more detailed description.

7. Conclusion

We have presented Humboldt, a web interface to explore large-scale language variation in several languages, available at `www.languagevariation.com`. It provides statistical analyses and mapping capabilities, without requiring any programming knowledge.

In the future, we plan to extend the search options to include syntactic phenomena as well. Additionally, the interface

[3] `http://mumin.ling.su.se/cgi-bin/dialects.py`

will play an even more active role in assisting the user in analyzing the data, taking the provided query terms as *seeds* and suggesting new queries based on that.

8. Acknowledgments

We would like to thank Klaes Rasmussen for his help with the prototype system, and the anonymous reviewers for valuable comments and feedback. This research is funded in part by the ERC Starting Grant LOWLANDS No. 313695, and by Carlsberg infrastructure grant No. CF14-0694.

9. Bibliographical References

Agha, A. (2005). Voice, footing, enregisterment. *Journal of linguistic anthropology*, pages 38–59.

Barbieri, F. (2008). Patterns of age-based linguistic variation in American English. *Journal of sociolinguistics*, 12(1):58–88.

Barke, A. J. (2000). The Effect of Age on the Style of Discourse among Japanese Women. pages 23–34.

Carter, P. M. (2013). Shared spaces, shared structures: Latino social formation and african american english in the us south. *Journal of Sociolinguistics*, 17(1):66–92.

Holmes, J. (1997). Women, language and identity. *Journal of Sociolinguistics*, 1(2):195–223.

Hovy, D. and Søgaard, A. (2015). Tagging performance correlates with author age. In *ACL*.

Hovy, D., Johannsen, A., and Søgaard, A. (2015). User review-sites as a source for large-scale sociolinguistic studies. In *Proceedings of WWW*.

Johannsen, A., Hovy, D., and Søgaard, A. (2015). Cross-lingual syntactic variation over age and gender. In *Proceedings of CoNLL*.

Jørgensen, A., Hovy, D., and Søgaard, A. (2015). Challenges of studying and processing dialects in social media. In *Workshop on Noisy User-generated Text (W-NUT)*.

Labov, W. (1964). *The social stratification of English in New York City*. Ph.D. thesis, Columbia university.

Macaulay, R. (2001). You're like 'why not?' the quotative expressions of glasgow adolescents. *Journal of Sociolinguistics*, 5(1):3–21.

Macaulay, R. (2002). Extremely interesting, very interesting, or only quite interesting? adverbs and social class. *Journal of Sociolinguistics*, 6(3):398–417.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Milroy, L. and Milroy, J. (1992). Social network and social class: Toward an integrated sociolinguistic model. *Language in society*, 21(01):1–26.

Nerbonne, J. (2003). Linguistic variation and computation. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 3–10. Association for Computational Linguistics.

Pechenick, E. A., Danforth, C. M., and Dodds, P. S. (2015). Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLOS One*.

Rickford, J. and Price, M. (2013). Girlz ii women: Age-grading, language change and stylistic variation. *Journal of Sociolinguistics*, 17(2):143–179.

Schmidt, J. E. and Herrgen, J. (2001). Digitaler Wenker-Atlas (DiWA). Bearbeitet von Alfred Lameli, Tanja Giessler, Roland Kehrein, Alexandra Lenz, Karl-Heinz Müller, Jost Nickel, Christoph Purschke und Stefan Rabanus. Erste vollständige Ausgabe von Georg Wenkers "Sprachatlas des Deutschen Reichs".

Silverstein, M. (2003). Indexical order and the dialectics of sociolinguistic life. *Language & Communication*, 23(3):193–229.

Wieling, M., Nerbonne, J., and Baayen, R. H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PloS one*, 6(9):e23613.