Measuring Syntactic Distances Between Dialects: A Web Application for Annotating Dialectal Data

Emanuele Di Buccio, Giorgio Maria Di Nunzio, and Gianmaria Silvello

Dept. of Information Engineering – University of Padua
[dibuccio,dinunzio,silvello]@dei.unipd.it

Abstract. Research in dialectal variation allows linguists to understand the fundamental principles underlying language systems and grammatical changes in time and space. Since different dialectal variants do not occur randomly on the territory and geographical patterns of variation are recognizable for an individual syntactic form, we believe that a systematic approach for studying this variations is required. In this paper, we present a Web application for annotating dialectal data, in particular with the aim of measuring the degree of syntactic differences between dialects.

1 Motivation and Background

Syntactic comparison across languages is essential in the research field of linguistics. In fact, the study of closely-related varieties has proven to be extremely useful in finding relations between cross-linguistic syntactic differences that might otherwise appear unrelated, and in analysing the linguistic structures in the task of historical reconstruction [7, 6]. More precisely, syntactic variation studies the ways in which linguistic elements, i.e. words and clitics, are put together to form constituents, that are phrases or clauses. In this context, the analyses of dialectal variation patterns may result in more fine-grained linguistic theories, and empirical dialect data may also help improve the validation process of linguistic theories. Therefore, dialectal variation research may contribute to a better understanding of the inner workings of the human language system [9]. Different dialectal variants do not occur randomly on the territory and geographical patterns of variation are recognizable for an individual syntactic form. In other words, the geographical distribution of an individual syntactic phenomenon is often geographically coherent to a certain extent. This indicates that there might be a relationship between syntactic variation and geographical distance. However, when several distribution patterns of syntactic phenomena are combined for joint analysis, the interpretation of geographical distributions is less clear [9].

This paper builds on previous work by the authors [2, 3, 5, 4] and present an extension of the Web application for annotating dialectal data in Synctactic Atlas of Italy (ASIt), which allows linguists to build a meaningful linguistic context and annotate documents with tags.

2 A Web Application for Tagging Dialectal Data

Following the work of [8], the term variable (tag) is central to this work. Generally speaking, a variable may be defined as a linguistic unit in which two language varieties can vary. We define a syntactic variable as a form or word order in a syntactic context where two dialects may differ. Several types of variables can be distinguished; for instance, they can be distinguished according to the linguistic unit to which they refer. The ASIt [1] tag set was defined to support the study on Italian dialects; it includes two different types of tags to capture word-level and sentence-level phenomena. Another example is the set of 192 features made available by The World Atlas of Language Structures (WALS)[1] in which each feature describes one aspect of cross-linguistic diversity.

The main linguistic idea behind this work is built on the concept of "clitic clusters" which happens when more than one clitic shows up within a single clause. One very interesting fact about clitic clusters is that the order in which they are in a cluster appears to be random; that is, it is not normally the same order as the corresponding order of full noun phrases, and there is what appears to be random variation between languages as to which ordering restrictions they impose. For example, a third person dative clitic must follow a third person accusative clitic in French, whereas the order must be the other way around in Italian, Spanish and Romanian. [2] For example, the sentence "Martine sends it to him" is translated in:

- Martine le lui envoie (French) (accusative-dative)
- Martina glielo spedisce (Italian) (dative-accusative)
- Martina i-l trimite (Romanian) (dative-accusative)

A first person dative clitic, however, must precede a third person accusative clitic in French (as in the other Romance languages). For example, "Martine sends it to me" becomes:

- Martine me l'envoie (French) (dative-accusative)

Therefore, we can study each clitic cluster as a separate as a separate space. Each space forms a context in which some linguistic phenomena should characterise a variety, that is the vectors of varieties that are similar should be closer in this space.

References

1. M. Agosti, P. Benincà, G. M. Di Nunzio, R. Miotto, and D. Pescarini. A Digital Library Effort to Support the Building of Grammatical Resources for Italian Dialects. pages 89–100, 2010.

[1] `http://wals.info`

[2] See examples in "Lectures on Clitics" `http://www.lel.ed.ac.uk/~packema/teaching/ling2L/index.htm`

2. E. Di Buccio, G. M. Di Nunzio, and G. Silvello. A system for exposing linguistic linked open data. In P. Zaphiris, G. Buchanan, E. Rasmussen, and F. Loizides, editors, *TPDL*, volume 7489 of *Lecture Notes in Computer Science*, pages 173–178. Springer, 2012.
3. E. Di Buccio, G. M. Di Nunzio, and G. Silvello. A curated and evolving linguistic linked dataset. *Semantic Web*, 4(3):265–270, 2013.
4. E. Di Buccio, G. M. Di Nunzio, and G. Silvello. A geolinguistic web application based on linked open data. In G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, and T. Sakai, editors, *SIGIR*, pages 1101–1102. ACM, 2013.
5. E. Di Buccio, G. M. Di Nunzio, and G. Silvello. An open source system architecture for digital geolinguistic linked open data. In Trond Aalberg, C. Papatheodorou, M. Dobreva, G. Tsakonas, and C. J. Farrugia, editors, *TPDL*, volume 8092 of *Lecture Notes in Computer Science*, pages 438–441. Springer, 2013.
6. V. Colonna, A. Boattini, C. Guardiano, I. Dall'Ara, D. Pettener, G. Longobardi, and G. Barbujani. Long-range comparison between genes and languages based on syntactic distances. *Human Heredity*, 70(4):245–254, 2010.
7. J. Nerbonne and W. Wiersma. A Measure of Aggregate Syntactic Distance. In *Proceedings of the Workshop on Linguistic Distances*, LD '06, pages 82–90. Association for Computational Linguistics, 2006.
8. M. R. Spruit. Measuring syntactic variation in dutch dialects. *Literary and Linguistic Computing*, 21(4):493–506, 2006.
9. M. R. Spruit. *Quantitative perspectives on syntactic variation in Dutch dialects*. LOT Dissertation Series 174. Netherlands Graduate School of Linguistics / Landelijke (LOT), 2008.