Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 38 (2014) 44 – 47

Procedia Computer Science

10th Italian Research Conference on Digital Libraries, IRCDL 2014

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

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

Department of Information Engineering – University of Padua, via Gradenigo 6/B, 35131 Padova, Italy

Abstract

Research in dialectal variation allows linguists to understand the fundamental principles that underlie 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 these variations is required. In this paper, we present a Web application for annotating dialectal data; the annotated data will be adopted for investigating measures of the degree of syntactic differences between dialects.

Peer-review under responsibility of the Scientific Committee of IRCDL 2014
Keywords: Digital Geolinguistic; Synctactic Distance; Vector Space Model;

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[1,2]. 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[3]. 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[3].

* Corresponding author. Tel.: +39-049-8277929 ; fax: +39-049-8277799.
E-mail address: dibuccio@dei.unipd.it

Peer-review under responsibility of the Scientific Committee of IRCDL 2014
doi:10.1016/j.procs.2014.10.009

In this paper, we present an extension of the Synctactic Atlas of Italy (ASIt) Digital Library[4]. ASIt provides functionalities to store set (*questionnaires*) of Italian sentences, annotate sentences according to their syntactic phenomena, add sentence translations in diverse dialects, and search sentences by tags. This work proposes an extension to build a meaningful linguistic context and annotate sentence translations with tags.

2. A Web Application for Tagging Dialectal Data

Following the work by Spruit[5], the term variable (tag) is central to our 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[4] 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)[6] 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". A morpheme is the smallest meaningful unit in the grammar of a language. A clitic is a morpheme that has syntactic characteristics of a word, but shows evidence of being phonologically bound to another word.[7] A clitic cluster occurs when more than one clitic shows up within a single clause. One very interesting fact about clitic clusters is that the order in which clitics 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 – see examples in "Lectures on Clitics"[8]. 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).

Our objective is the design of a methodology that can help the linguists to identify relationships among diverse varieties in terms of syntactic phenomena; the methodology should be automatic in order to reduce the linguists effort. We propose to model each clitic cluster as a separate vector 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. A description of the framework under investigation is reported in[9].

The evaluation of the methodology effectiveness requires a dataset constituted of translations of the same sentence in the different varieties and annotations of the linguistic phenomena in these sentences. The ASIt dataset currently contains only part of this information. It is constituted of a set of questionnaires, where each questionnaire is constituted of a set of Italian sentences; diverse dialectical translations are available for each sentence. Tags at sentence level have been adopted to identify syntactic phenomena in the Italian sentences. The dataset currently lacks of the annotations on the syntactic phenomena on the sentence translations.

For this reason we designed and developed a tagging interface that helps the linguists to efficiently tag sentence translations. The interface supports the linguists in the navigation of the ASIt corpus through diverse interaction steps. The first step is the selection of the research project. Each research project aims at investigating specific research hypotheses; in order to support this investigation a set of documents is associated to each project — *questionnaires* are instances of documents. Currently two research projects are supported, but in this paper will focus on the project ASIt since the documents associated to this project are those of interest for our methodology. The second and the third steps consist respectively in the selection of the questionnaire and in the selection of a specific Italian sentence in that questionnaire. Once the Italian sentence has been selected, the linguist can access and annotate the translations of that sentence in the diverse dialects. Figure 1 reports a screen-shot of the interface for annotating dialectal translations,

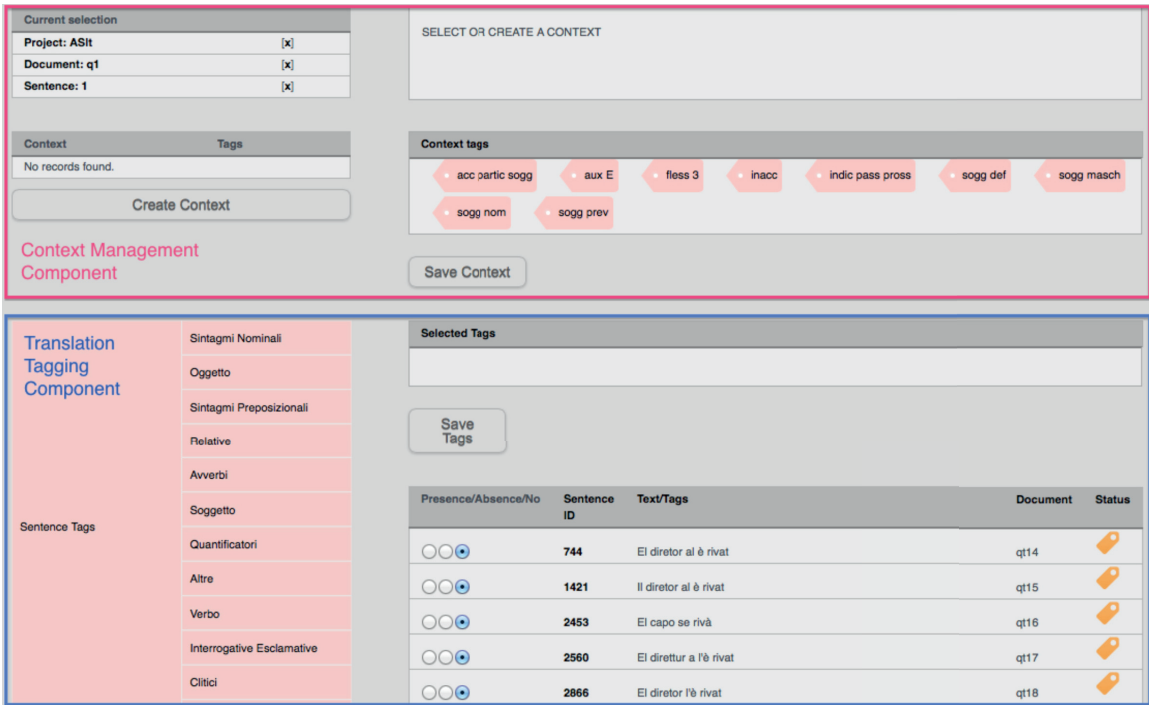

Fig. 1. Context management and translations tagging.

where the linguist selected the ASIt project, questionnaire q_1 of this project, and sentence 1 of questionnaire q_1. Two main components can be distinguished: the context management component and the translation tagging component.

The *context management* component allows the linguist to select or save the specific context where the tagging procedure is performed. A context is defined by a set of tags. When a new context is created, an initial set of tags is shown; these tags are those associated to the Italian sentence selected in the third step. The linguist can add or remove tags from the context, and save the new context by clicking on the "Save Context" button; the context list on the upper left will be automatically updated. Saving contexts could be used to identify possible combinations of tags that could be useful in subsequent analyses performed by the linguists, e.g. for searching sentences with specific syntactic phenomena. Therefore, the current search interface could be extended in order to allow the linguists to represent their information need starting from a list of contexts that result from their own previous analyses or those performed by other linguists.

The *translation tagging* component allows the selection of the tag group, e.g. "Clitici", and the selection of the tags in the selected group, e.g. "clit sogg". Once a tag has been clicked, it is shown in the "Selected Tags" box. In Figure 1 a single tag is selected, but in general the selection can involve a set of tags. Then, the radio buttons in the first column of the translation table can be adopted to specify for each translation, if the selected tag combination is present, absent, or not appropriate. Indeed, in order to investigate variations among Italian dialects linguists need not only information on the presence of a certain element, but also on the absence of an element that can be omitted supposedly only in some constructions and in conjunction with specific characteristics of the language.

3. Final Remarks

In this paper, we presented an extension of the ASIt Digital Library that allows linguists to efficiently tag sentence translations. The requirement analysis for design of the interface was carried out in collaboration with the team of linguists that are working in the ASIt project; the proposed tagging interface was designed not only to support the tagging procedure, but also to gather contextual information that could be useful in further linguistic analyses.

The data that will be collected by the new tagging interface will be the basis for the experimental evaluation of the methodology proposed in[9].

Acknowledgment

This work has been supported by the project "Un'inchiesta grammaticale sui dialetti italiani: ricerca sul campo, gestione dei dati, analisi linguistica" (Bando FIRB – Futuro in ricerca 2008).

References

1. Nerbonne, J., Wiersma, W.. A measure of aggregate syntactic distance. In: *Proceedings of the Workshop on Linguistic Distances*; LD '06. Stroudsburg, PA, USA: Association for Computational Linguistics. ISBN 1-932432-83-3; 2006, p. 82–90. URL: http://dl.acm.org/citation.cfm?id=1641976.1641987.
2. Colonna, V., Boattini, A., Guardiano, C., Dall'Ara, I., Pettener, D., Longobardi, G., et al. Long-range comparison between genes and languages based on syntactic distances. *Human Heredity* 2010;**70**(4):245–254.
3. Spruit, M.R.. *Quantitative perspectives on syntactic variation in Dutch dialects*. LOT Dissertation Series 174. Netherlands Graduate School of Linguistics / Landelijke (LOT); 2008.
4. Agosti, M., Benincà, P., Nunzio, G.M.D., Miotto, R., Pescarini, D.. A digital library effort to support the building of grammatical resources for italian dialects. In: Agosti, M., Esposito, F., Thanos, C., editors. *IRCDL*; vol. 91 of *Communications in Computer and Information Science*. Springer. ISBN 978-3-642-15849-0; 2010, p. 89–100.
5. Spruit, M.R.. Measuring syntactic variation in dutch dialects. *Literary and linguistic computing* 2006;**21**(4):493–506.
6. Dryer, M.S., Haspelmath, M., editors. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology; 2013. URL: http://wals.info/.
7. *What is a clitic?* Last modified: 5 January 2004. URL: http://www-01.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsACliticGrammar.htm.
8. *Lectures on Clitics*. 2008. URL: http://www.lel.ed.ac.uk/~packema/teaching/ling2L/L2L\%20-\%20lectures\%20on\%20clitics.pdf.
9. Di Buccio, E., Di Nunzio, G.M., Silvello, G.. A vector space model for syntactic distances between dialects. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., et al., editors. *LREC*. European Language Resources Association (ELRA); 2014, p. 2486–2489.