ASIt: A Grammatical Survey of Italian Dialects and Cimbrian: Fieldwork, Data Management, and Linguistic Analysis

Maristella Agosti[1], Birgit Alber[2], Paola Benincà[3], Giorgio Maria Di Nunzio[1], Marco Dussin[1], Riccardo Miotto[1], Diego Pescarini[3], Stefan Rabanus[2], and Alessandra Tomaselli[2]

[1] Department of Information Engineering, University of Padua
{maristella.agosti,giorgiomaria.dinunzio,marco.dussin, riccardo.miotto}@unipd.it
[2] Department of Foreign Languages and Literatures, University of Verona
{birgit.alber,stefan.rabanus,alessandra.tomaselli}@univr.it
[3] Department of Linguistics and Performing Arts, University of Padua
{paola.beninca,diego.pescarini}@unipd.it

Abstract. ASIt aims to observe, collect and analyse the linguistic variation displayed by the dialects of a language. The main theoretical hypothesis is that linguistic variation is not due to chance, but depends on the combination of a finite number of parameters. It is a first step towards the creation of a European digital library for recording and studying linguistic micro-variation.

1 Introduction and Motivation

In order to make a linguistic resource usable both for machines and humans, a number of issues need to be addressed: crawling, downloading, cleaning, normalizing, and annotating the data are only some of the steps that need to be taken in order to produce valuable content. Data quality has a cost, and human intervention is required to achieve the highest quality possible for a resource of usable scientific data. From a computer science point of view, curated databases are a possible solution for designing, controlling and maintaining collections that are consistent, integral and high quality.

The ASIt project aims to observe, collect and analyse the linguistic variation displayed by the dialects of a language [1,2,3]. The main theoretical hypothesis is that linguistic variation is not due to chance, but depends on the combination of a finite number of parameters. The study of genetically related dialects constitutes a primary field of research in order to isolate some of these parameters and reach a better understanding of the architecture of the language faculty. This project represents a significant contribution both to the field of Italo-Romance linguistics and, more widely, to formal linguistics by adopting an interdisciplinary approach that interfaces traditional dialectological study with recent developments in morphological and syntactic theory.

M. Agosti et al. (Eds.): IRCDL 2011, CCIS 249, pp. 100–103, 2011.

2 Scientific Challenges

The study of dialectal heritage is the goal of many research groups in Europe; however, a full integration of all the studies carried out by each research team is hampered by the different choices made in each project, in particular the tagging system and structure of the respective databases. Some projects devised a tagging system that index-links the whole sentence; some others are based on a tagging system that isolates and index-links every word [1,3].

One of the challenges of the ASIt project is to create a database for recording and studying linguistic micro-variation, in particular to design a database and a tagging system compatible to the Edisyn network[1] which includes linguistic research projects developed for Dutch, Portuguese, German and Scandinavian dialects. Another important objective is the inclusion of linguistic data from Cimbrian, a language spoken in German language islands of Northern Italy [2].

These two main objectives require the design and development of a "curated database" of dialects and languages and an accurate definition of the tagging system. The tags have to focus on both the sentence-level phenomena as well as he word level phenomena, which according to the EAGLE (Expert Advisory Group on Language Engineering Standard)[2] guidelines should in turn depend on two kinds of annotation: morphosyntactic annotation, part of speech (POS) tagging; syntactic annotation, annotation of the structure of sentences by means of a phrase-structure parse or dependency parse.

The choice of an automatic POS tagger to tag the dialectal sentences is not appropriate for the ASIt project because this project aims to account for minimally different variants of specific syntactic variables within a sample of closely related or geographically adjacent languages. As a consequence, even the best POS tagger with an accuracy of 98% wouldn't be sufficiently good for fine grained tagging; therefore, in order to pin down these subtle asymmetries, the linguistic analysis must be carried out manually [3].

3 Key Technologies

The ASIt project will promote the following technologies:

- comparison between closely related varieties (dialects), hence the formation of hypotheses about the nature of crosslinguistic parametrization;
- single out contact phenomena between Romance and Germanic varieties;
- find, describe and analyze syntactic phenomena of Romance and Germanic dialects to be found.

The tagged corpus of ASIt/Cimbrian data will be available to end users who might be, for example, linguists interested in carrying out syntactic analyses or informants interested in correcting or augmenting the data. Moreover, it is important that the database be of use to a wider audience than a small group of specialists alone; for this reason, the ASIt project will also:

[1] http://www.dialectsyntax.org/

[2] http://www.ilc.cnr.it/EAGLES96/home.html

- be cross-platform and easily deployable to end users;
- be as modular and extensible as possible, to properly describe the behaviour of the service by isolating specific functionalities at the proper layer;
- be intuitive and capable of providing support for different tasks and different linguistic objects;
- support different types of users who need to have access to different kinds of features and capabilities;
- support internationalization and localization allowing the application to adapt to the language of the user and his country or culturally dependent data, such as dates and currencies.

The project will also promote the exchange of information between the academic staff, the scientific collaborator and the administrative staff of the project, and also the public dissemination of works and relevant results. On the basis of their grammatical features, the sentences will be indexed according to a tag set and will be stored into a relational database that can be searched through a dedicated system of information retrieval that finds the relevant examples on the basis of the requested tags. The database will finally form an organized set of data that can be easily retrieved and compared in order to allow the extraction of sets of data and generalizations for articles and presentations.

The ASIt project will also explore new visualization tools for the analysis of the geographical distribution of grammatical phenomena. This can be done by exploiting the geographical coordinates of each location, which are stored in the database. Given these coordinates, the system automatically can create one of the Geotagging formats (GeoRSS[3], KML[4], etc.) and exploit GoogleMaps[5] APIs to visualize it. This option is very important because a user can graphically view how the dialect data are distributed through the country, and perform further analysis based on these visualizations.

4 Contribution by Italian Research Community

The ASIt project is the results of a multidisciplinary collaboration which synergistically makes use of the competences of different linguistic and computer science research teams. Some components of the teams have previously collaborated in envisioning, designing and developing a Digital Library System (DLS) able to manage a manually curated resource of dialect data in the context of the ASIt[6] project, which has collected a considerable amount of syntactic data concerning Italian dialects. This DLS provided linguists with a crucial test bed for formal hypotheses concerning human language. ASIt has demonstrated the need to abstract a specific information space of reference for the management of the linguistic resources. As a result, a new information space implied by a

[3] http://www.georss.org/
[4] http://www.opengeospatial.org/standards/kml/
[5] http://maps.google.it/
[6] http://asit.maldura.unipd.it/

new linguistic project has been framed into an appropriate conceptual model to allow us to develop an enhanced system for the management of the new dialectal resources of interest.

One of the main goals of the project is the preparation of a co-ordinated collection of Italian dialects; this co-ordinated collection can be conceived only because the present research team is building on previous and long-lasting research that has produced intermediate and basic results [2,3]. This means that the data the ASIt project has produced is based on long-standing experience of data collection, documentation, and preservation.

Another important contribution to the ASIt project is given by the research teams of the German variety of Cimbrian[7]. Cimbrian, spoken in the language island of Giazza (Veneto, province of Verona), Lusern (Trentino) and – historically – Asiago/Roana (Veneto, province of Vicenza), is of great interest to different important lines of research in linguistics - a fact which is witnessed by many studies on Cimbrian throughout the last decade.

Acknowledgements. Project FIRB "Un'inchiesta grammaticale sui dialetti italiani: ricerca sul campo, gestione dei dati, analisi linguistica" (Bando FIRB Futuro in ricerca 2008, cod. RBFR08KRA_003).

Project "Cimbrian as a test case for synchronic and diachronic language variation proposals for implementing the ASIt (Syntactic Atlas for Italy)" co-financed by the Fondazione Cariverona.

References

1. Agosti, M., Benincà, P., Di Nunzio, G.M., 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. (eds.) IRCDL 2010. CCIS, vol. 91, pp. 89–100. Springer, Heidelberg (2010)
2. Agosti, M., Alber, B., Di Nunzio, G., Dussin, M., Rabanus, S., Tomaselli, A.: Cimbrian as a test case for synchronic and diachronic language variation: a conceptual approach for the information space. In: Congresso Nazionale AICA 2010. L'Aquila città storica, città digitale, città futura. La ricostruzione dell'Aquila come laboratorio sperimentale per la comunità scientifica ed industriale nazionale ICT (2010)
3. Agosti, M., Alber, B., Di Nunzio, G.M., Dussin, M., Pescarini, D., Rabanus, S., Tomaselli, A.: A Digital Library of Grammatical Resources for European Dialects. In: Agosti, M., et al. (eds.) IRCDL 2011. CCIS, vol. 249, pp. 61–74. Springer, Heidelberg (2011)

[7] `http://ims.dei.unipd.it/websites/cimbrian/`