A Curated Database for Linguistic Research: The Test Case of Cimbrian Varieties

Maristella Agosti*, Birgit Alber †, Giorgio Maria Di Nunzio*, Marco Dussin*, Stefan Rabanus†, Alessandra Tomaselli†

*Department of Information Engineering — University of Padua
Via Gradenigo 6/a — 35131 Padova — Italy
{maristella.agosti,giorgiomaria.dinunzio,marco.dussin}@dei.unipd.it
†Department of Foreign Languages and Literatures — University of Verona, Italy
Lungadige Porta Vittoria, 31 — 37129 Verona — Italy
{birgit.alber,stefan.rabanus,alessandra.tomaselli}@univr.it

Abstract

In this paper we present the definition of a conceptual approach for the information space entailed by a multidisciplinary and collaborative project, "Cimbrian as a test case for synchronic and diachronic language variation", which provides linguists with a test bed for formal hypotheses concerning human language. Aims of the project are to collect, digitize and tag linguistic data from the German variety of Cimbrian – spoken in three areas of northern Italy: Giazza (VR), Luserna (TN), and Roana (VI) – and to make available on-line a valuable and innovative linguistic resource for the in-depth study of Cimbrian. The task is addressed by a multidisciplinary team of linguists and computer scientists who, combining their competence, aim to make available new tools for linguistic analysis.

Keywords: Linguistic Analysis, Curated Database, Germanic and Romance varieties

1. Introduction

In this contribution we present the results of an ongoing multidisciplinary collaboration which synergistically makes use of the competences of two different teams, one of linguists and one of computer scientists. Some members of the teams have previously collaborated in envisioning, designing and developing a Digital Library System (DLS) capable of managing a manually curated resource of dialectal data, which provides linguists with a crucial test bed for formal hypotheses concerning human language.

A curated database (Buneman, 2009) is a database the content of which has been collected with a great deal of human effort and which has certain characteristics: data has been edited from existing sources; raw data are annotated to enrich their interpretation and description; the database has to be updated regularly by curators who can be technicians, computer scientists, or linguists, depending on the type of maintenance task that has to be conducted. In this setting of multidisciplinary collaboration it is important to use all competences synergistically, with the aim of developing a new research approach to produce knowledge which would not otherwise be possible to obtain.

The previous collaboration was conducted in the context of the project *Atlante Sintattico d'Italia*, Syntactic Atlas of Italy (ASIt)[1] (Agosti et al., 2010) a digital library system for managing a resource of curated dialect data which provides access to grammatical data, also through an advanced user interface specifically designed to update and annotate the linguistic data.

The aim of the present linguistic project is to collect, digitize and tag linguistic data from the German variety of Cimbrian. Cimbrian, spoken in the language islands of Giazza (Veneto, province of Verona), Lusern (Trentino) and – historically – Asiago/Roana (Veneto, province of Vicenza), is of great interest to three important lines of research in linguistics:

- Romance dialectology: linguistic contact phenomena are visible especially at the lexical level,
- German dialectology: the language island varieties exhibit a high level of preservation of certain structural characteristics, and
- Historical linguistics: the diachronic development of a variety in isolation shows a particularly interesting mixture of preservation and innovation.

The interest for this linguistic situation is witnessed by the many studies on Cimbrian over the last decade (Tomaselli, 2009). Furthermore, the present project, which focuses prominently on Cimbrian syntax is consistent with similar projects at the European level in that it is creating a database of syntactic structures – which so far have been neglected in traditional dialectological work (Rabanus et al., 2008). Finally, Cimbrian is an endangered language, with only a few elderly people speaking the language fluently in Giazza.[2] This makes collection of linguistic data on this language all the more important.

The paper is organised as follows: Section 2. outlines the linguistic aspects of the project; Section 3. presents the conceptual model for the information space entailed by curated linguistic resources; Section 4. presents some conclusions and suggestions for future work.

[1] http://asit.maldura.unipd.it/

[2] The situation is much better in Lusern even though there are no children acquiring Cimbrian as their mother tongue.

2. A linguistic database of Cimbrian varieties

2.1. State of the Art

In the last two decades, several large-scale databases of linguistic material of various types have been developed worldwide. The aims of these projects are in part similar (they all aim to make linguistic data accessible to a large public) but they differ from each other with respect to the amount of collected data and the type of search for which they have been constructed.

The most well-known example is probably The World Atlas of Languages Structures (WALS) (Haspelmath et al., 2005), developed at the Leipzig Max Planck Institute of Evolutionary Anthropology and published as a print book and on the internet. The projected Atlas of Pidgin and Creole Language Structures (APICS, scheduled for 2012)[3] will map grammatical features in the same way as WALS but will additionally include an interactive electronic database on the internet.

There are also notable projects outside Europe and the United States. The "21st Century Sejong Project"[4] aims to create a Korean National Corpus which includes a grammatically tagged corpus with more than 15 million word-like units (2006). The high number is down to automatic tagging which works rather well for Korean since it is an agglutinative language. The output of the tagging process is a data structure in which grammatical labels exactly match word segments (words, endings, postpositions). The tagset reflects typological properties of Korean, the top level is constituted by word-class labels.

Academia Sinica in Taipei, Taiwan, has several large Language Archive projects.[5] These projects share with our project a concern for collecting data from endangered languages (e.g. the indigenous Austronesian Languages of Taiwan) and preserving historical language data (see their Tagged Corpus of Early Mandarin Chinese).

The Open Language Archives Community (OLAC),[6] that has recently celebrated its first 10 years of activity, is a worldwide network dedicated to collecting information on language resources (field notes, grammars, audio/video recording, descriptive papers, and so on) and developing standard protocols for interoperability, thus simplifying data extraction. This meta database, similar to a huge online library catalogue, encourages linguists everywhere to submit information, thus becoming data providers.

If the principal aim of a huge online database like OLAC is to provide information in a standard accessible way, the aim of a small, curated database like the one we are presenting here is to make it possible to compare fine-grained linguistic data. To be more specific, if a linguist is only interested in finding general studies on Cimbrian worldwide, a single search ("cimbrian") in OLAC is all that is required. However, in order to compare the pronominal paradigm (let's say all possible realizations of the first person singular and plural) in Cimbrian with the Germanic and possibly the Romance varieties from both a synchronic and diachronic perspective, a curated database is needed, one which makes use of standard tag-setting for grammatical categories, i.e., both word classes (parts of speech) and grammatical features (case, number, gender etc.).

2.2. The Cimbrian dialects

The project focuses on a geographical region in the North-East of Italy, usually referred to as the "Triveneto" area. In this region the Cimbrian dialects are in intense language contact with the Italian dialects belonging to the Lombard and Venetian dialect groups (Pellegrini and Mello, 1977). This historic language-contact situation (supplemented by the entry of spoken Standard Italian in the repertoire of the speakers in the course of the 19th century) is crucial for our idea that language variation in Cimbrian depends both on its structural possibilities as a German dialect and on the multilinguism of its speakers. Hence, it is necessary to consider the Cimbrian and the Italian dialects of the area with respect to the same grammatical categories and features (see Section 2.4.).

2.3. Cimbrian documents

In contrast to many other German dialects Cimbrian has a tradition as written language and a literature that goes back to the beginning of the 17th century. This makes it possible to reconstruct the language change for at least four empirically attested stages (1602, 1844, 1942, 2009/2010). The written documents that have been elaborated in order to form part of the database are "Christlike unt korze Dottrina" (1602 (Meid, 1985)), "Novena vun unzar liben Vraun" (1844 (Stefan, 2000)), "Taut6. Puox tze Lirnan Reidan un Scraiban iz Gareida on Lietzan" (1942 (Cappelletti and Schweizer, 1942)). These Cimbrian texts have been completely transcribed (faithfully to their graphic form) and segmented in sentences which have also been linked to their translations in Italian and Standard German. For contemporary Cimbrian, fieldwork has been conducted in Giazza (2009 and 2010). In order to be able to compare the modern Cimbrian data with data from the Italian dialects and other projects on the syntax of German varieties the questionnaire was designed as similiar as possibile to the ASIt questionnaires and has integrated questions elaborated by the SyHD project (Syntax hessischer Dialekte, University of Marburg/Frankfurt/Vienna).[7] The interviews have been digitally recorded and transcribed both according to a Cimbrian orthography developed for this purpose and phonetically. The questionnaire so far aims to elicit syntactic and morphological data, a questionnaire eliciting the phonology of the language has been developed and data has been elicited, but not yet integrated into the database.

2.4. Tags

After segmentation of the sentences, tagging of the linguistic data is carried out. We start with tagging at the word-level, determining the parts of speech of single words. Tagging of syntactic phenomena at the sentence level and tagging of syntactic constituents will take place in a second phase of the project. The starting point for developing

[3] http://lingweb.eva.mpg.de/apics/

[4] http://www.sejong.or.kr/

[5] http://ndaip.sinica.edu.tw/

[6] http://www.language-archives.org

[7] http://www.syhd.info/

a viable set of tags for Cimbrian is the tagset elaborated by the Edisyn project,[8] especially the one developed for the (Dynamic) Syntactic Atlas of the Dutch dialects (DynaSAND).[9] In collaboration with the ASIt team, we have developed a language-specific set of tags which is suitable for Cimbrian but, at the same time, allows the Cimbrian data to be linked to other databases of dialect syntax. This involves assigning the same names to same parts of speech as in the Edisyn and the ASIt databases, at most adding tags when they are needed for language-specific structures of Cimbrian, or leaving out tags which are not relevant for Cimbrian. Thus, for instance, the tag "verbal particle" has been added to identify verbal particles which can be found in German dialects (e.g. the verbal particle in the Standard German sentence, *Ich gehe weg* 'I go away'), but gender values such as "masculine" have been left out for the tag of the past participle, since past participles in analytical verb constructions never inflect for gender in German varieties. We can therefore imagine the creation of a language-specific tagset as starting from a universal core, shared by all languages, and subsequently developing a language-specific periphery, which is compatible with other databases and able to classify language-specific structures.
Another important innovation with respect to the basic rules of Edisyn is the possibility of double tagging. A first instance of double tagging is the assignment of two tags to a single lexical item, e.g., "past participle" and "adjective" to inflected past participles with attributive function inside nominal phrases, e.g. *in andere gapintate bort* 'other bound words'. Another instance of double tagging is the assignment to an item which consists of two parts not clearly separable from each other, e.g., finite verb and subject clitic clustered in a single graphical word (*vingasto* "find=you" 'you find').

3. A Conceptual Approach for the Information Space

In this section we report on the work made to define a conceptual approach to describe the information space entailed by the curated database of Cimbrian. We use a conceptual model to express the meaning of terms and concepts used by domain experts to discuss the problem, and to find the correct relationships between different concepts (Zins, 2007). To do so, we adopted a two-phase approach: at the beginning the world of interest was studied and represented at a high level of abstraction by means of the analysis of requirements, supported by the use of a website as the point of exchange of information among the people of the two teams; afterwards it was progressively refined to obtain the conceptual representation of its information space, partitioned in five modelling areas, seven main steps of advancements of the project and five actors involved.

3.1. Conceptual Schema

The component of the digital library system that manages and permanently stores the data is based on a relational database. The design and implementation of the curated database of dialectal resources followed a three phase approach:

1. The world of interest was represented at a high level by means of a conceptual representation based on the analysis of requirements,
2. Progressive refinements of the world of interest to obtain the logical model of the data of interest,
3. The relational database and the interface to access the data were implemented and verified.

The main core of the schema was developed and presented in (Agosti et al., 2011). It consists of three broad areas: i) the point of enquiry, which is the location where a given dialect is spoken; ii) the administrative area (namely, region, province), the location belongs to; iii) the linguistic area, i.e. the linguistic group the dialect belongs to. In the subsequent work, the information about tags and words has been integrated in the original schema, and the present version of the schema is now enhanced and able also to model the words of a sentence, the hierarchy of the tags, and the association between tags and words.

3.2. Linguistic Project Cycle

In the context of our linguistic project, we approach the concept of a project as a cycle that starts with a set-up of the project itself and terminates with the presentation of results through search interfaces, maps, raw results and papers.
Figure 1 represents the different steps of the linguistic enterprise, the actors involved in each step, and the information space entailed. The main steps of the project can be read on the left of the figure and can be summarized as:

- "Set-up of a new project": this consists in the creation of the linguistic project itself and on the definition of its users and resources;
- "Retrieval and preparation of written texts, conduction of fieldwork, transcription of audio data, translation of sentences, DB population": in this step the database of documents is populated and enriched with new data from different sources needed to perform the next steps;
- "Segmentation of sentences into words and constituents": documents added to the database are, in this phase, split into words and constituents to allow not only the tagging of the entire document or phrase, but a more in depth analysis (Figure 2 shows the interface for editing and splitting sentences into words);
- "Validation of editors' work": the validation of the definition of words and constituents from sentences, which is the work done in the previous step, is validated and stored in the database;
- "Tagging of words, constituents and sentences": this is the task of assigning tags and labels to the previously created words and constituents (Figure 3 shows the interface for tagging words);

[8] http://www.dialectsyntax.org/
[9] http://www.meertens.knaw.nl/sand/

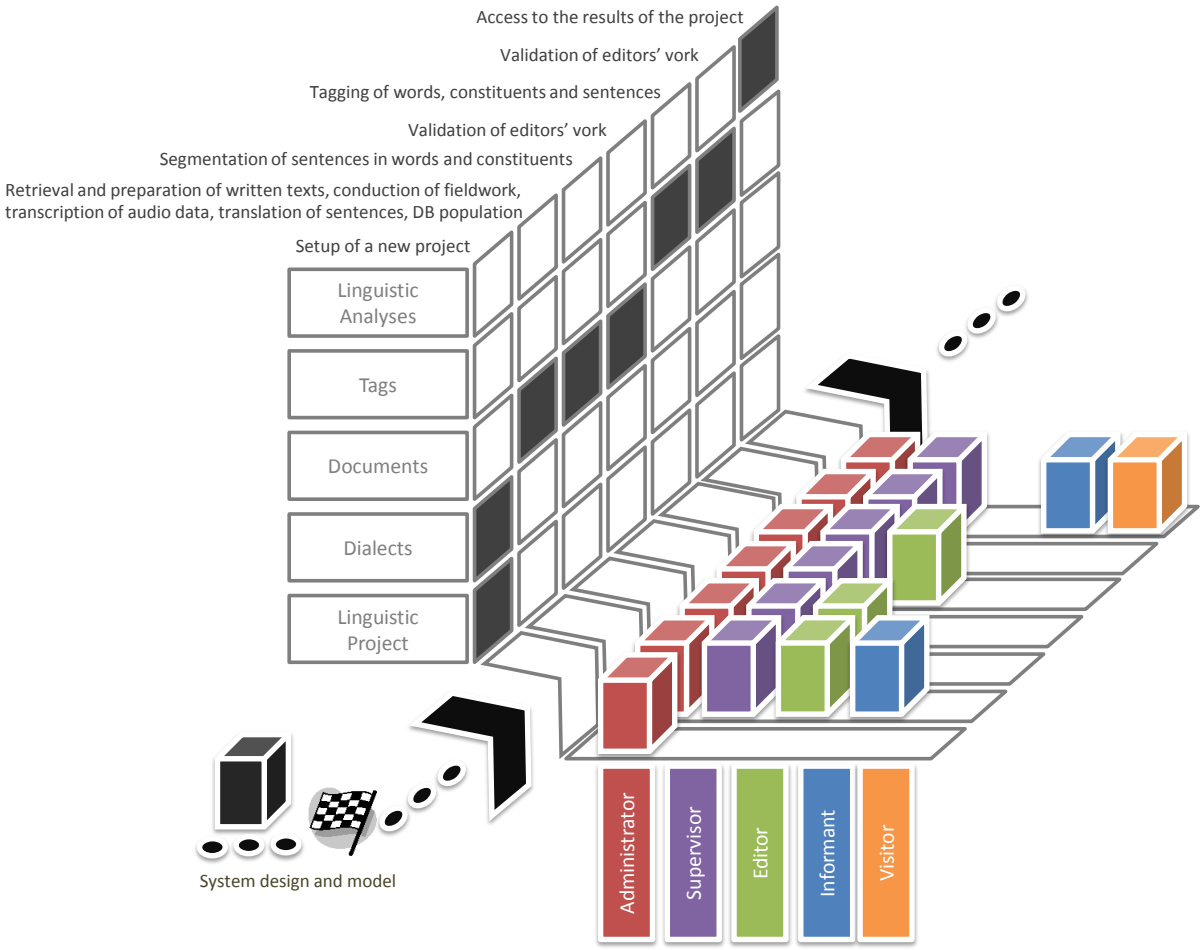
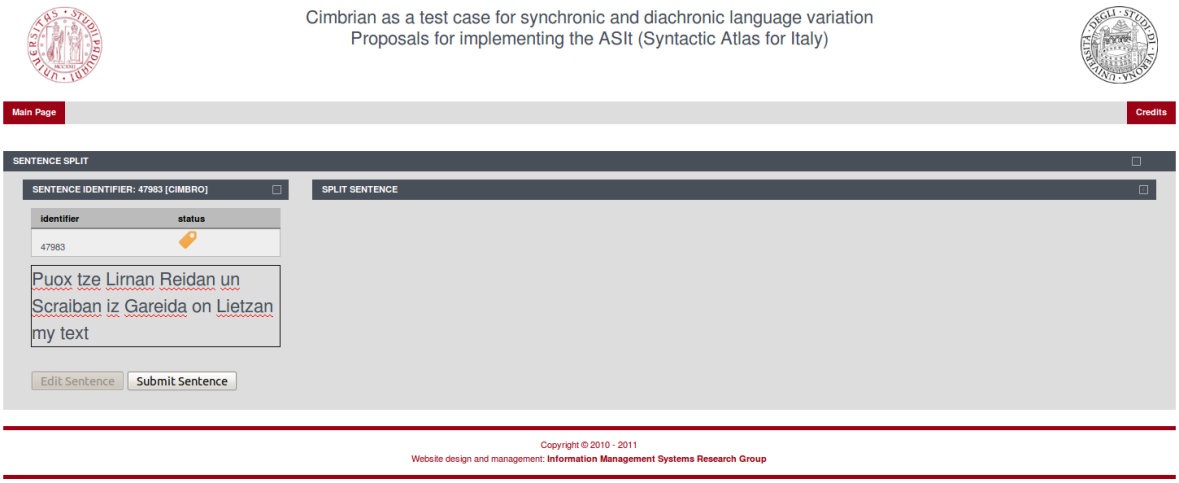


Figure 1: The different steps of the linguistic enterprise, the areas of the information space entailed, and the actors involved in each step.

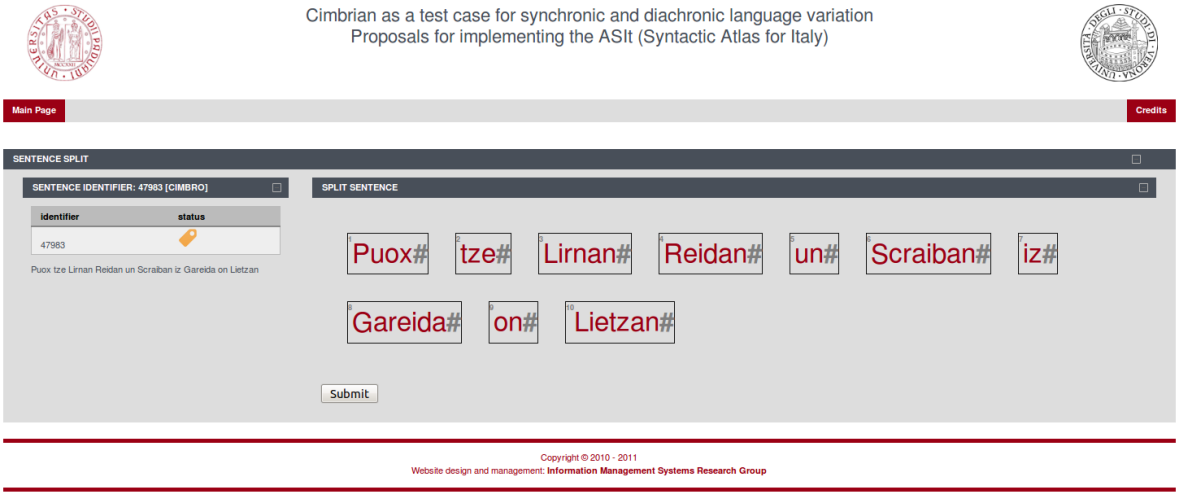
- "Validation of editors' work on tagging of words and constituents": with regard to the definition of words and constituents, their tagging also needs to be validated and stored; at present validation is achieved in regular team meetings in which problems arising during the tagging procedure are discussed and the editor's choices are confirmed or corrected by the supervisors;
- "Access to the results of the project": consultation, browsing and access to all public information resources produced during the course of the project.

The actors involved in the linguistic project, represented by differently coloured cubes in Figure 1, will interact with aspects at different levels of the five areas presented above, as summarized by the dark squares on the left side of the figure. The different types of actors modeled and their main tasks are:

- the *administrator* manages the different aspects of a project such as the setting of the project itself, the creation of the users and the administration of the system. Before the start of a project, the administrator is in charge of the design and implementation of the system itself or of its plugins and extension, and once the project is started, the administrator works in the background to support the work of the other actors;
- the *supervisor* contributes to the creation of the database of sentences by collaborating with the informant on editing interviews, finding books or providing translations, then making the transcription into the database and validating the work made by editors on sentences;
- the *editor* takes part in the project to create words and constituents from given sentences, and to provide the required tags for them. In case of doubts or errors, an editor can communicate with the supervisors (and also with the administrator, if needed) in order to receive help and support;
- the *informant* is a speaker of a dialect who is asked to produce dialect utterances or to translate one or more sentences into his or her dialect. The informant is usually interviewed and supervised by a linguistic expert;
- the *visitor* needs to consult, browse and access all the public information resources produced during the course of a project in a suitable way. He needs a simple and intuitive interface, and a set of tools to view and compare results, export and print them.



(a)



(b)

Figure 2: The interface for editing (Figure 2a) and splitting (Figure 2b) sentences.

3.3. Linguistic Analyses

The tagged corpus of Cimbrian data will be available to end users (visitors) who might be linguists interested in carrying out syntactic analyses or also informants, interested in correcting or augmenting the data. Concerning the former, it is important that the data are presented in a way which makes it usable by linguists working in different theoretical frameworks. Although it is inevitable (and, to some extent, also desirable) that the tagging of the data is influenced by theoretical considerations (in our case, the framework of generative linguistics), it is important that the database be of use not only to a small group of specialists.

With respect to the types of structures which can be analyzed in the tagged Cimbrian database, it will be possible to analyze syntactic structures and phenomena in great detail. It should also be possible to deduct morphological paradigms without too much effort, while it still remains a desideratum of further research projects to integrate a component which will make it possible to carry out phonological analyses on the database.

It is important that the structures in the database can be compared with structures present in other databases, since cross-linguistic comparison will be one of the major interests of an analysis of Cimbrian, which is in contact with Romance varieties (hence can be compared to the ASIt data) but has a Germanic base (hence can be compared e.g. to the DynaSAND data). To make just one example of what an analysis in these terms could look like, consider the case of pronouns and clitics in Cimbrian. In Cimbrian documents, sentences as the following can be found ((Bidese, 2008), p. 134):

miar	importar-z-mar	nicht	zo	sterben
me	matter=it=me	not	to	die

'I don't mind dying'

Whereas the use of the infinitive particle *zo* and the expletive pronoun *-z-* are typical of German varieties, the dou-

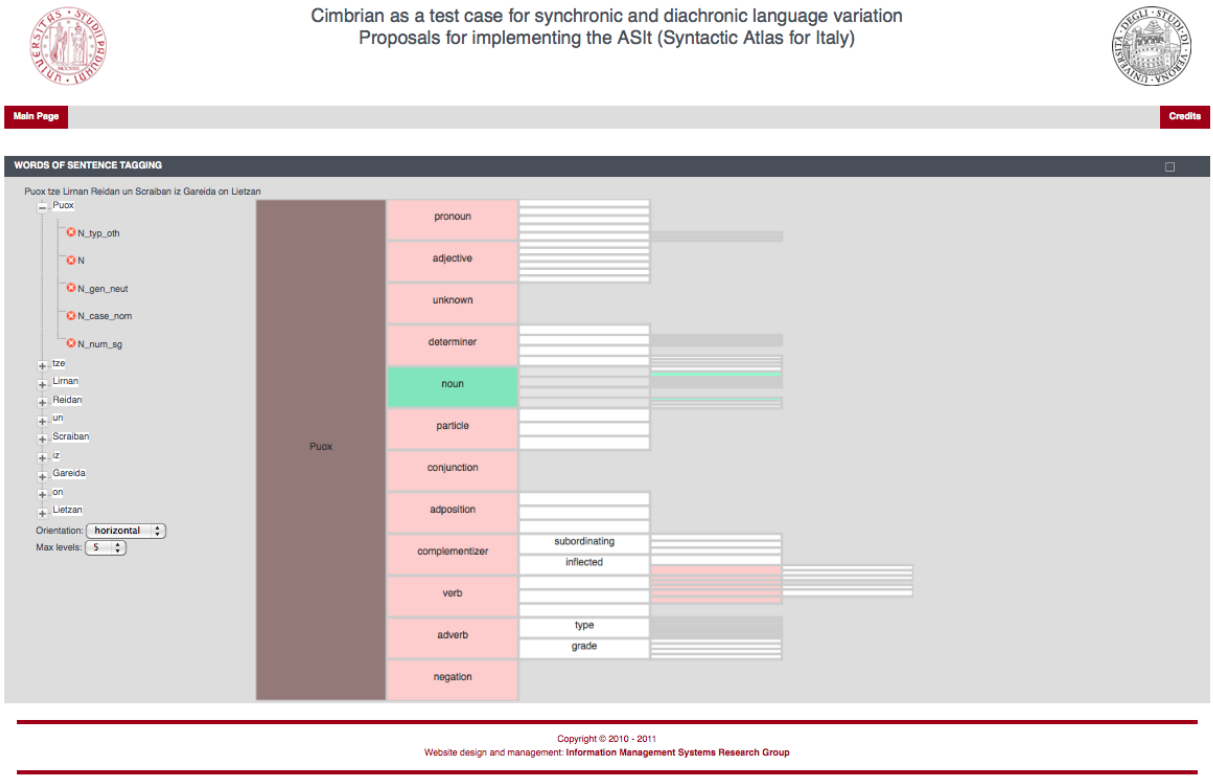


Figure 3: The interface for tagging the words of a sentence: the words of the sentence are shown on the left; the hierarchy of tags is shown on the main area of the screen.

bling of the object pronoun *miar, mar* could be evidence for the development of a Romance-like system of clitics in Cimbrian, differently from Standard German where clitics are not attested. The tagged database will make it possible to retrieve all sentences of the corpus containing potential clitics and will therefore create an empirical basis on which to test hypotheses as those of the development of a system of clitics in Cimbrian.

4. Conclusions and Future Work

In this paper we presented the results of an ongoing linguistic project which aims to collect, digitize and tag linguistic data from the German variety of Cimbrian. The project gave the opportunity to merge different fields of research and begin a multidisciplinary collaboration which synergistically makes use of the competences of two different teams, one of linguists and one of computer scientists. Since cross-linguistic comparison will be one of the major interests of an analysis of Cimbrian, the main aim is to design and implement a digital library system that enables the management of linguistic resources of curated dialect data and provides access to grammatical data.

For this purpose, a new information space implied by this new linguistic project has been framed into an appropriate conceptual model to allow us to envisage an enhanced system for the management of the new dialectal resources of interest: future work will concern the design and development of this DLS for scientific data able to properly support the course of a linguistic project and the cooperation and interaction among researchers, students, industrial partners and practicioners. Once implemented, the usability of the interface will be evaluated in two phases: firstly, by analyzing the activities of the project's members concerning the supervisioning and the editing of the data; secondly, by studying visitors' activity by means of log analysis techniques.

5. Acknowledgements

This work has been supported by the 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), the PROMISE network of excellence (contract n. 258191) and by the Project "Cimbrian as a test case for synchronic and diachronic language variation Proposals for implementing the ASIt (Syntactic Atlas of Italy)" (Fondazione Cariverona, Bando 2008).

6. References

M. Agosti, P. Benincà, G.M. Di Nunzio, R. Miotto, and D. Pescarini. 2010. A digital library effort to support the building of grammatical resources for italian dialects. In M. Agosti, F. Esposito, and C. Thanos, editors, *IRCDL*, volume 91 of *Communications in Computer and Information Science*, pages 89–100. Springer.

M. Agosti, B. Alber, G.M. Di Nunzio, M. Dussin, D. Pescarini, S. Rabanus, and A. Tomaselli. 2011. A digital library of grammatical resources for european dialects. In M. Agosti, F. Esposito, C. Meghini, and N. Orio, editors, *Digital Libraries and Archives - 7th*

Italian Research Conference, IRCDL 2011, Pisa, Italy, January 20-21, 2011. Revised Papers, number 249 in Communications in Computer and Information Science, pages 61–74. Springer.

E. Bidese. 2008. *Die diachronische Syntax des Zimbrischen*, volume 510 of *Tübinger Beiträge zur Linguistik (TBL)*. Gunter Narr Verlag, Tübingen, Germany.

P. Buneman. 2009. Curated databases. In Maristella Agosti, José Luis Borbinha, Sarantos Kapidakis, Christos Papatheodorou, and Giannis Tsakonas, editors, *ECDL*, volume 5714 of *Lecture Notes in Computer Science*, page 2. Springer.

G. Cappelletti and B. Schweizer. 1942. *Taut6. Puox tze Lirnan Reidan un Scraiban iz Gareida on Lietzan*. Ferrari-Auer, Bolzano, Italy.

M. Haspelmath, M. S. Dryer, D. Gil, and B. Comrie. 2005. *The World Atlas of Language Structures*. Oxford University Press, United Kingdom.

W. Meid. 1985. *Der erste zimbriscche Katechismus. Christlike unt korze Dottrina. Die zimbrische Version aus dem Jahre 1602 der Dottrina Christiana Breve des Kardinal Bellarmin in kritischer Ausgabe. Einleitung, italienischer und zimbrischer Text, Übersetzung, Kommentar, Reproduktionen*. Inst. für Sprachwiss, Innsbruck, Austria.

G.B. Pellegrini and A. Mello. 1977. *Carta dei dialetti d'Italia*. Pacini, Pisa, Italy.

S. Rabanus, B. Alber, and A. Tomaselli. 2008. Erster Veroneser Workshop "Neue Tendenzen in der deutschen Dialektologie: Morphologie und Syntax". In *Vorschläge für die Ausrichtung zukünftiger Dialektsyntaxprojekte. Zeitschrift für Dialektologie und Linguistik*, volume 75, pages 72–82.

B. Stefan. 2000. *Novena vun unzar liben Vraun. Die Zimbrische Mariennovene des D. Giuseppe Strazzabosco mit Übersetzung und Kommentar*. Inst. für Sprachwiss, Innsbruck, Austria.

A. Tomaselli. 2009. La grammatica cimbra di Cappelletti-Schweizer. In A. Petterlini and A. Tomaselli, editors, *L'eredità cimbra di Monsignor Giuseppe Cappelletti, supplemento a Quaderni di lingue e letterature*. Edizioni Fiorini, Verona, Italy.

C. Zins. 2007. Conceptual approaches for defining data, information, and knowledge. *JASIST*, 58(4):479–493.