A Digital Library of Grammatical Resources for European Dialects

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

[1] Department of Information Engineering, University of Padua
Via Gradenigo, 6/a – 35131 Padua, Italy
{maristella.agosti,giorgiomaria.dinunzio,marco.dussin}@unipd.it
[2] Department of Foreign Languages and Literatures, University of Verona
Lungadige Porta Vittoria, 31 – 37129 Verona, Italy
{birgit.alber,stefan.rabanus,alessandra.tomaselli}@univr.it
[3] Department of Linguistics and Performing Arts, University of Padua
Via Beato Pellegrino, 1 – 35137 Padua, Italy
diego.pescarini@unipd.it

Abstract. The paper illustrates the methodology at the basis of the design of a digital library system that enables the management of linguistic resources of curated dialect data. Since dialects are rarely recognized as official languages, first of all linguists need a dedicated information management system providing the unambiguous identification of each dialect on the basis of geographical, administrative and geolinguistic parameters. Secondly, the information management system has to be designed to allow users to search the occurrences of a specific grammatical structure (e.g. a relative clause or a particular word order). Thirdly, user-friendly graphical interfaces must give easy access to language resources and make the building of the language resources easier and distributed. This work, which stems from a project named ASIt (Atlante Sintattico d'Italia), is a first step towards the creation of a European digital library for recording and studying linguistic micro-variation.

1 Motivations

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 [1]. 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 [2] are a possible solution for designing, controlling and maintaining collections that are consistent, integral and high quality.

In the present contribution we report the ongoing results of a multidisciplinary collaboration which synergistically makes use of the competences of two different teams, one of linguists and one of computer scientists. Some components of the

M. Agosti et al. (Eds.): IRCDL 2011, CCIS 249, pp. 61–74, 2011.

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 *Atlante Sintattico d'Italia*, Syntactic Atlas of Italy (ASIt)[1] 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 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.

The paper reports on this effort giving a short presentation of the European background of interest in Section 2; Section 3 introduces the objectives of the undertaking, Section 4 presents the tagging, and Section 5 introduces the conceptual approach which underlies the digital library system which manages the information space of interest for the effort under development.

2 European Background

The study of dialectal heritage is the goal of many research groups in Europe. The ASIt project is part of the Edisyn network[2] which includes similar linguistic research projects developed for Dutch, Portuguese, German and Scandinavian dialects. These projects have been developed according to slightly different goals and methods; today, a full integration is hampered by the different choices made in each project, in particular the tagging system and structure of the respective databases. For instance, the ASIt system has been devised with a tagging system that index-links the whole sentence, in order to deal with phenomena that are not directly associated with a specific morphological element. In contrast, the SAND[3](Netherlands) and Cordial-Syn[4](Portugal) projects are based on a tagging system that isolates and index-links every word.

This work has been the first step towards making ASIt compatible with the Edisyn system in order to create a European database for recording and studying linguistic micro-variation. It is also the first time linguistic data from Cimbrian, a language spoken in German language islands of Northern Italy, has been systematically digitalized and integrated into a database.

3 The Linguistic Objectives

The study aims to tag and make available data from two different sources: curated data derived from the ASIt project, which contains syntactic data on

[1] http://asit.maldura.unipd.it/
[2] http://www.dialectsyntax.org/
[3] http://www.meertens.knaw.nl/projecten/sand/sandeng.html
[4] http://www.clul.ul.pt/english/sectores/variacao/cordialsin/projecto_cordialsin.php

about 200 Italian dialects, and data from a project on the syntax of Cimbrian, a German variety spoken in the language islands of Giazza (Veneto, province of Verona), Lusern (Trentino) and – historically – Asiago/Roana (Veneto, province of Vicenza).

Research on the syntax of Italian dialects as well as on the syntax of Cimbrian is of great interest to several important lines of research in linguistics:

- it allows comparison between closely related varieties (dialects), hence the formation of hypotheses about the nature of crosslinguistic parametrization;
- it allows contact phenomena between Romance and Germanic varieties to be singled out, should they arise;
- it allows syntactic phenomena of Romance and Germanic dialects to be found, described and analyzed to a great level of detail.

Therefore a project in line with similar projects at the European level was launched to study the creation of a database of syntactic structures which so far have been neglected in traditional dialectological work [3].

4 Overview of the Tagging System

The design of a tagset for corpus annotation is normally carried out in compliance with international standards — e.g. CES (Corpus Encoding Standard)[5] — which in turn are based on the specifications of SGML (Standard Generalized Markup Language)[6] and international guidelines like EAGLE (Expert Advisory Group on Language Engineering Standard)[7] and TEI (Text Encoding Initiative)[8] guidelines.

According to these standards, each tagset is formed by several sub-systems responsible for the identification and description of different linguistic "units": text, section, paragraph, clause, and word. Given the objectives of the ASIt and the Cimbrian syntax enterprise, we have focussed on the tagging of sentence-level phenomena as well as tagging at the word level, which according to the EAGLE 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.

A tagset based on this distinction is normally designed to be used in combination with software applications processing linguistic data on the basis of probabilistic algorithms, which assign every lexical item a POS tag and, subsequently, derive the structure of the clause from the bottom up. The best automatic POS tagger can achieve an accuracy between 95% and 98% which means two to five errors

[5] http://www.cs.vassar.edu/CES/
[6] http://www.w3.org/MarkUp/SGML/
[7] http://www.ilc.cnr.it/EAGLES96/home.html
[8] http://www.tei-c.org/index.xml

on average every one hundred words. This error is acceptable in case the task is to analyze vast corpora so that practical tasks can be carried out, e.g. roughly translate a collection of texts into different languages or summarize their contents with quantitative analysis (such as frequency or contextual distribution of lexical items).

Such an error, however, is not acceptable for some tasks, like the fine-grained tagging of ASIt and Cimbrian data. First of all, it is worth noting that our enterprise has a different objective, being a scientific project aiming to account for minimally different variants of specific syntactic variables within a sample of closely related or geographically adjacent languages. As a consequence, while other tagsets are designed to carry out a gross linguistic analysis of a vast corpus, our tagset aims to capture fine-grained grammatical differences. As a consequence, in order to pin down these subtle asymmetries, the linguistic analysis must be carried out manually. In addition, the corpus of Italian dialects and Cimbrian data would presumably not be big enough to train a probabilistic algorithm. Lastly, the Romance varieties collected in the ASIt project require a different tagset from the tags employed for a German variety such as Cimbrian, since certain morphosyntactic structures are expressed in one language and not in the other. For these reasons a tagset which takes into account the specificity of both Romance and German varieties and which can be assigned manually need to be developed by linguistic experts.

4.1 Tags

The starting point for developing a viable set of tags is the tagset elaborated by the Edisyn project[9], especially for the (Dynamic) Syntactic Atlas of the Dutch dialects (DynaSAND)[10]. The ASIt team together with the Cimbrian project team have developed two language-specific sets of tags which are suitable for the Italian dialect data of the ASIt and Cimbrian, respectively, but which at the same time allow our data to be linked to other European databases of dialect syntax. This involves assigning the same denominations to the same parts of speech as in the Edisyn and the ASIt databases, at most adding tags when they are needed for language-specific structures, or leaving out tags which are not relevant for the languages of our project. 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 never inflect for gender in German varieties (compare Standard German *sie/er ist gesprungen* 'she/he jumped' and Standard Italian *(lei) è saltata* 'she jumped' vs. *(lui) è saltato* 'he jumped').

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, but which is able to classify language-specific structures.

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

Moreover, the data from ASIt and Cimbrian differ from Edysin and DynaSAND with respect to the internal organization of their tagset, which in the ASIt/Cimbrian projects follows a multidimensional and hierarchical model. First, the ASIt/Cimbrian tagset allows different types of linguistic constituents to be captured, since tags can be assigned to words, phrases, or sentences. This allows the user to retrieve data concerning word, phrase, or sentence level phenomena and formulate complex queries to test the correlation between phenomena at different levels. Second, grammatical features are organized according to a hierarchical structure wherein features are grouped into classes dominated by super-ordinate nodes, e.g. features like 'masculine', 'feminine', 'neuter' are daughters to the same node "gender". This guarantees an efficient and exhaustive tagging and, above all, it allows the values 'unassigned' and 'unknown' to be assigned to every node in case that distinction is not relevant in the dialect under observation or a specific value cannot be chosen on the basis of the known data.

4.2 Linguistic Analyses

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. 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 data tagging is influenced by theoretical considerations (in our case, the framework of generative linguistics), it is important that the database be of use to a wider audience than a small group of specialists alone.

With respect to the types of structures which can be analyzed in the tagged database, it will be possible to analyze syntactic structures and phenomena in great detail. It should also be possible to deduce 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 of the database.

Here is an example of what an analysis in these terms of dialect data can look like: pronouns and clitics in Cimbrian. In Cimbrian documents, sentences such as the following can be found ([4], p. 134):

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

'I don't mind dying'

The use of the infinitive particle *zo* and the expletive pronoun *-z-* are typical of German varieties, though the postverbal position of the latter partially corresponds to the syntax of northern Italo-Romance varieties, where subject clitic pronouns appear postverbally as well (but only in interrogative and exclamative clauses). Moreover, the doubling of the object pronoun *miar*, *mar* could be

evidence of the development of a Romance-like system of object clitics in Cimbrian, unlike Standard German, which does not exhibit pronominal clitics. But notice that the position of the object pronoun *mar* is not consistent with the position of the corresponding element in Italian dialects, in which object clitics are normally found postverbally only with non-finite verbs.

5 A Conceptual Approach for the Information Space of the Linguistic Project

In this section we report on the work made to define a conceptual approach for the information space entailed by data curated resources of Italian dialects and Cimbrian. 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, helped 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 six actors involved.

5.1 Analysis of Requirements

One of the results of the meetings between the group of computer scientists and the group of linguists has been the definition of a list of common and general requirements for the system which should:

- 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.

On the linguistic side, an accurate work has been made to agree on the tagset presented in Section 4.1 and to arrange it to be usable and automatically interpretable by software tools. Moreover, a number of requirements for the interface were defined together with computer scientists:

- the interface should show and preserve the hierarchical structure of the tags during the tagging phase;
- the interface should make it possible not only to show the entire hierarchy of the tags but also to navigate the hierarchy level by level by hiding the non-selected nodes and branches;

- it should be possible to manage uncertainty during tagging, for example the system should provide a way to store "unknown" or "not yet assigned" tags;
- the interface should alert the user when the tagging of a word is complete, that is all the mandatory features have a value.

5.2 The Conceptual Approach

As a result of the investigation of user requirements and needs, the information space implied by the linguistic project has been framed into a formal model. This model provides a conceptual approach that takes into consideration and describes all the entities involved, and defines "the rules according to which data are structured" [5]. An appropriate conceptual approach is, indeed, the necessary basis for making the produced scientific data an active part of any information enrichment, such as data provenance and citation, management, access, exchange, visualization and interpretation. In this way, data become an integral part of the process of knowledge transfer and sharing towards relevant application communities [6].

The conceptual approach is centred on five main modelling areas of linguistic interest:

- **linguistic project:** This deals with the different aspects of a linguistic research project, such as the collection and organisation of data, the management and subscription of the different types of involved actors, the validation of the work made at the different steps of the project;
- **dialects:** Since dialects are rarely recognized as official languages, linguists need a dedicated information management system providing the unambiguous identification of each dialect on the basis of geographical, administrative and geolinguistic parameters;
- **documents:** This concerns the different documents made available by a project, including questionnaires, interviews, transcriptions of parts of speech, books, and so on. Each document may be translated into different languages or dialects. The same document can be used by different projects and by different editions of the project during the years. Documents are formed by sentences, which are formed by words and eventually grouped into constituents;
- **tags:** Tags are keywords from a controlled vocabulary assigned to sentences, words, and constituents to label, identify, and recognize them;
- **linguistic analyses:** This models the different aspects concerning the linguistic analysis of the experimental results, such as the comparison of results, the statistical analyses, the cartographic representation of selected features, and so forth.

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

Each project can be viewed as a cycle, starting with a set-up of the project itself and terminating with the presentation of results through search interfaces, maps, raw results and papers, which can be used as starting information for the set-up of a new project. The main steps of the project are:

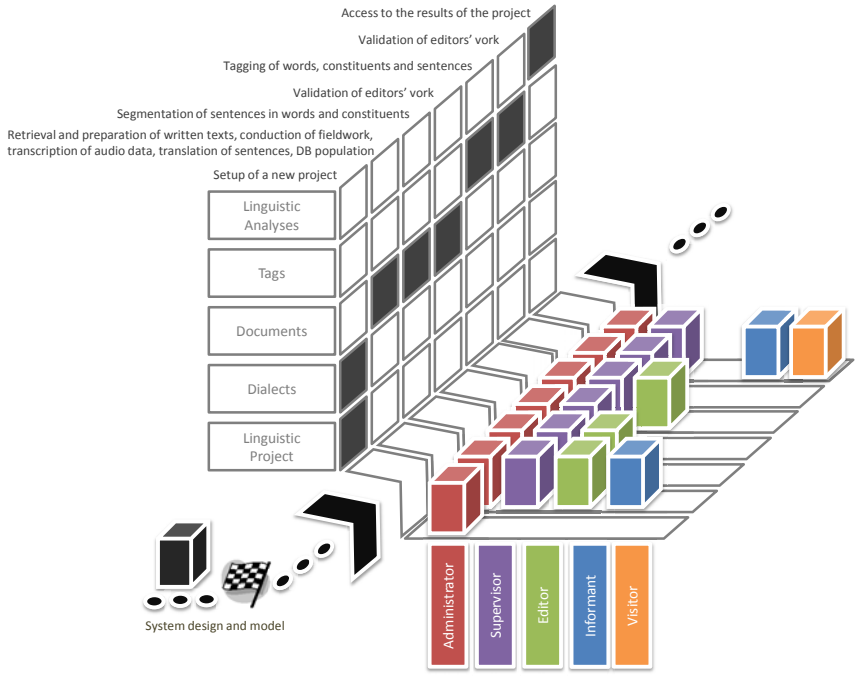


Fig. 1. The different steps of the linguistic enterprise, the areas of the information space entailed, and the actors involved in each step

- "Set-up of a new project": this consists of 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 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);

– "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;
– "Access to the results of the project": consultation, browsing and access to all the public information resources produced during the course of the project.

The actors involved in the linguistic project, represented by different 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.

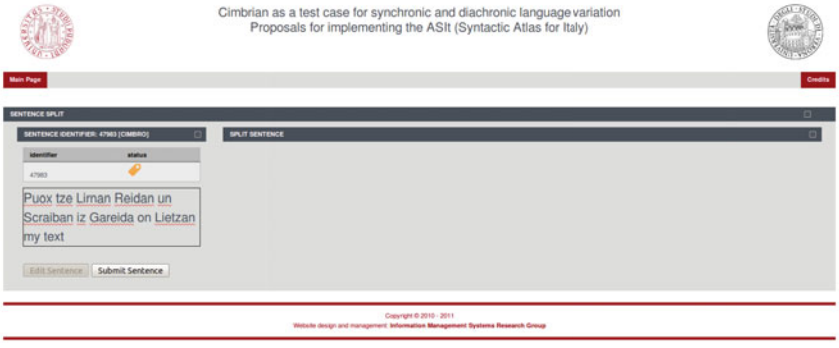
The different types of actors modeled and their main tasks are:

– the *administrator*, manages the different aspects of a project such as the setting-up 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 extensions, 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 to inform them and or to receive help and support, also from the administrator if needed;
– 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.

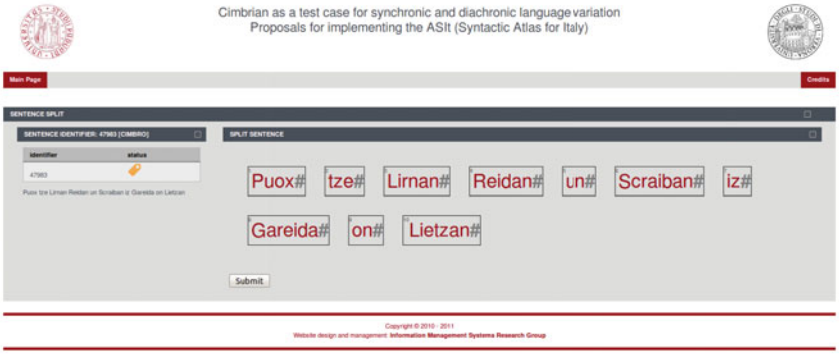
Figure 1 also shows that while the early stages are mainly devoted to the set-up of the project itself and to the preparation of the data, requiring limited interaction between the different actors; as time passes and the project comes into full swing, there is a progressive increase of interactions and supervising actions made by the actors, represented by the presence of more than one cube representing a type of user.

5.3 The Conceptual Schema

The component of the digital library system that manages and 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:



(a)



(b)

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

- the world of interest was represented at a high level by means of a conceptual representation based on the analysis of requirements,
- afterwards the world of interest was progressively refined to obtain the logical model of the data of interest, and
- lastly, the relational database and the interface to access the data were implemented and verified.

The core of the schema was developed and presented in [7]. 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 this work, the information about tags and words has been integrated in the original schema. In particular, the conceptual schema now also models: the the words of a sentence, the hierarchy of the tags, the association between tags and words.

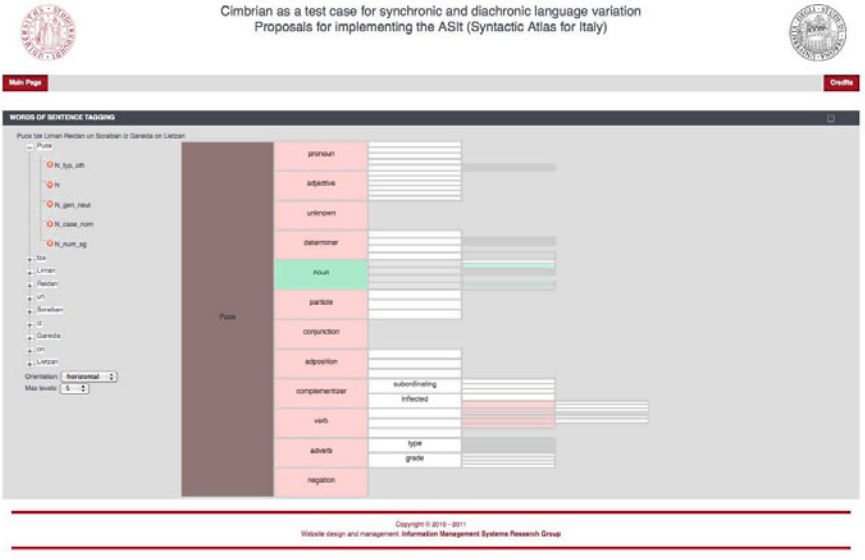


Fig. 3. The interface for tagging the words of a sentence: the words of the sentence are shown on the left; the hierarchy of tags are shown on the main area of the screen

5.4 Pilot Study

An initial run of a study with the purpose of verifying that the functionalities of the system are well-designed has been prepared. The aim is to test the two phases of the information phase presented in Section 5.2 that are currently implemented in the system: the "segmentation of sentences in words", and the "tagging of words".

A form was prepared to gather non-numeric qualitative data about user opinions, views and list of problems and observations. The form consists of seven questions about:

- the evaluation of the functionalities of orthographic correction and segmentation of sentences;
- the evaluation of the functionalities of word tagging and the hierarchical organization of tags;
- how the interface helps the user to avoid and/or correct mistakes during the two phases.

The form was distributed to six linguistic experts which are currently working in entering the data in the system: three professors, one researcher, one PhD student and one master degree student. The results can be summarized as follows:

- Positive aspects
  - the functionality of the orthographic checking of the sentences is valuable and easy to use;

  - the functionality of the segmentation of sentences into words is important although it requires to the user some extra effort to clean the text from punctuation marks;
  - the hierarchical organization of the tags is extremely positive and helps the user during the tagging phase.
  - the time spent for tagging a sentence is short and it is even shorter as the user learns the position of the tags in the hierarchy;
- Negative aspects
  - the segmentation window can become unmanageable in case of very long sentences;
  - the system should warn the user during the tagging phase whenever one or more mandatory tags are missing, and/or have a function like 'jump-to-next-tag';
  - the tags saved by the user should be ordered similarly to the hierarchical structure;
  - there is the need to speed up the phase from one sentence to another.

In general, the judgments about the functionalities and the interface are positive. The issue of speeding-up the tagging phase when the user learns the interface has been raised by most of the users. The interface helps the user to correct their mistakes but there is the need to implement the two-level checking, editor and supervisor, for further corrections during the tagging phase (as described in Section 5.2).

5.5 Data

Currently, the linguistic corpus set that has been represented and which is managed by the digital library system is characterized by:

- 468 documents;
- 48,575 sentences;
- 530 tags (sentence level and POS level);
- 16,731 tags for 1,375 sentences for an average of 12.2 tags per sentence;
- 5,411 tags for 1,501 words for an average of 3.6 tags per word.

Questionnaires, sentences and tagged sentences can be accessed via a Web interface from the ASIt Web site. The team of linguists is currently tagging sentences with a thorough POS tagset by means of a specific interface designed in order to manage both Italian and Cimbrian data[11]. When the tagging is completed, the estimated data produced will be hundreds of thousands of tags, i.e. one of the biggest multilingual POS corpus available not only to linguists for linguistic analyses but also machine learning algorithms for training automatic POS taggers.

Beside the different data management services and search options, the digital library system also allows the visualization of the geographical distribution

[11] `http://svrims2.dei.unipd.it:8080/asit/`

of grammatical phenomena. This can be done by exploiting the geographical coordinates of each location, which are kept in the data resource. Given these coordinates, the system automatically creates one of the Geotagging formats (GeoRSS[12], KML[13], etc.) and exploits GoogleMaps[14] APIs to visualize it. This option is very important because a user can graphically view how the dialects are distributed throughout the country, and perform further analyses rooted on previously presented results [7].

6 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. The project provided 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, 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 develop 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.

Acknowledgments. This work has been partially supported by the project "Cimbrian as a test case for synchronic and diachronic language variation proposals for implementing the ASIt (Syntactic Atlas for Italy)" co-financed byo the Fondazione Cariverona, and 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 system has been developed by integrating services offered by the IMS Component Integrator (ICI)[15] library develped by the Information Management System (IMS) group of the Department of Information Engineering of the University of Padova.

[12] http://www.georss.org/
[13] http://www.opengeospatial.org/standards/kml/
[14] http://maps.google.it/
[15] http://ims.dei.unipd.it/software/ici/apidoc/

References

1. Kilgarriff, A.: Googleology is bad science. Computational Linguistics 33, 147–151 (2007)
2. Buneman, P.: Curated Databases. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, p. 2. Springer, Heidelberg (2009)
3. Rabanus, S., Alber, B., Tomaselli, A.: Erster Veroneser Workshop Neue Tendenzen in der deutschen Dialektologie: Morphologie und Syntax. Zeitschrift für Dialektologie und Linguistik 75, 72–82 (2008)
4. Bidese, E.: Die diachronische Syntax des Zimbrischen. Tübinger Beiträge zur Linguistik (TBL), vol. 510. Gunter Narr Verlag, Tübingen (2008)
5. Tsichritzis, D.C., Lochovsky, F.H.: Data Models. Prentice Hall, Englewood Cliffs (N.J.) (1982)
6. Agosti, M., Di Nunzio, G.M., Ferro, N.: Scientific Data of an Evaluation Campaign: Do We Properly Deal with Them? In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 11–20. Springer, Heidelberg (2007)
7. 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)