A Digital Library Effort to Support the Building of Grammatical Resources for Italian Dialects

Maristella Agosti[1], Paola Benincà[2], Giorgio Maria Di Nunzio[1], Riccardo Miotto[1], and Diego Pescarini[2]

[1] Department of Information Engineering, University of Padua
Via Gradenigo, 6/a, 35131 Padua, Italy
{maristella.agosti,giorgiomaria.dinunzio,riccardo.miotto}@unipd.it
[2] Department of Linguistics and Performing Arts, University of Padua
Via Beato Pellegrino, 1, 35137 Padua, Italy
{paola.beninca,diego.pescarini}@unipd.it

Abstract. In this paper we present the results of a project, named ASIt, which provides linguists with a crucial test bed for formal hypotheses concerning human language. In particular, ASIt aims to capture cross-linguistic variants of grammatical structures within a sample of about 200 Italian Dialects. Since dialects are rarely recognized as official languages, first of all linguists need a dedicated digital library system providing the tools for the unambiguous identification of each dialect on the basis of geographical, administrative and geo-linguistic parameters. Secondly, the information access component of the digital library system needs 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) rather than a specific word. Thirdly, since ASIt has been specifically geared to the needs of linguists, user-friendly graphical interfaces need to be created to give easy access to and make the building of the language resource easier and distributed. The paper reports on the ways these three main aims have been achieved.

1 Introduction

Since the 1990s, the explosion of corpus-based research and the use of automatic learning algorithms has heightened the pace of growth in language resources and as a consequence many corpora have been automatically built by means of machine learning techniques. However, this may have led to a reduction in the quality of the corpora themselves [1]. In order to make a linguistic resource usable for machines and for 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 done to produce valuable content [2]. 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

M. Agosti, F. Esposito, and C. Thanos (Eds.): IRCDL 2010, CCIS 91, pp. 89–100, 2010.
© Springer-Verlag Berlin Heidelberg 2010

point of view, curated databases [3] are a possible solution for designing, controlling and maintaining collections that are consistent, integral and high quality. A curated database is a database the content of which has been collected by a great deal of human effort and which has certain characteristics: data has been edited from existing sources and provenance; 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 the 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 building a new research approach for the production of new knowledge which would otherwise be impossible to create.

In the present contribution we show the results of a multidisciplinary collaboration which synergistically makes use of the competences of two different teams, one of linguists and one of computer scientists, which have collaborated in envisioning, designing and developing a digital library system able to manage a manually curated resource of dialectal data named ASIt[1] (*Atlante Sintattico d'Italia*, Syntactic Atlas of Italy) and which provides linguists with a crucial test bed for formal hypotheses concerning human language. From the computational point of view, the project aims to implement a digital library system that enables the management of a resource of curated dialect data and provides access to grammatical data, also through an advanced user interface specifically designed to update and annotate the primary data.

The paper is organised as follows: Section 2 outlines the peculiarities of the ASIt project and the main lines of the methodology adopted; Section 3 presents the requirements on the tagging system, which are strictly linked to both the specificity of the linguistic data and the formal theory that the system has to deal with; Section 4 presents the main characteristics of the digital library system that manages and gives access to the ASIt linguistic resource; Section 5 reports on some conclusions and future work.

2 Methodology

The manually curated resource of dialectal data stored and managed by the digital library system were collected by means of questionnaires consisting of sets of Italian sentences, each sentence having many parallel dialectal translation. However, the ASIt data resource differs from other multilingual parallel corpora in the following three aspects:

1. It contains data on about 200 Italian Dialects. Since dialects are rarely recognized as official languages, linguists need a dedicated digital library system providing the unambiguous identification of each dialect on the basis of geographical, administrative and geo-linguistic parameters.
2. It aims to capture cross-linguistic variants of grammatical structures. In other words, the information access component needs to allow users to search

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

the occurrences of a specific grammatical structure, e.g. a relative clause or a particular word order, rather than a specific word or meaning, even though a specific word such as “who” or “what” is a possible target, it appears in the data in many different dialectal forms.

3. It has been specifically geared to the needs of linguists: user-friendly graphical interfaces have therefore been created to give easy access to language resources and to make the building of the language resources easy and distributed.

Given the originality of the ASIt enterprise and the granularity of the collected data, the two teams decided to work synergistically to build a new piece of knowledge which was unattainable otherwise if the two teams were working separately or only in a support/assistance way. Moreover, a synergic approach has already been adopted in other scientific challenges, in similar but nonetheless different areas of research, and it has produced valid scientific results [4].

A number of considerations were made in terms of the approach to be used in the design of the procedures for the management, storage and maintenance of the data that were to be produced in the course of the development of the project. The main aim 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, some of which have been documented in [5,6][2]. This means that the data the ASIt project has produced is based on long-standing experience of data collection, documentation, and preservation. Therefore, we decided to design and maintain the ASIt collection of data and to bring them in line with the rules of definition and maintenance of “data curation”, as defined in the e-Science Data Curation Report [7]: “The activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purposes, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose”. As a consequence, a major target of the ASIt project has been the design and development of a “curated database” of Italian dialects. The actions we have undertaken to design this new curated database have benefited from previous experience gained in the curation of data for experimental evaluation campaigns in the field of Information Retrieval [8]. The research conducted in parallel in that field has improved the understanding of systems operating in the area of languages to produce curated databases that allow the re-use of data to generate new knowledge and the maintenance of unique observational data which is impossible to re-create [9], as is the case for data collection campaigns through questionnaires of the ASIt project.

[2] Since the early part of the project focused mainly on Northern Italian dialects, ASIt was formerly called ASIS (*Atlante Sintattico dell’Italia Settentrionale*, Syntactic Atlas of Northern Italy).

3 The ASIt Enterprise

The ASIt enterprise builds on a long standing tradition of collecting and analysing linguistic corpora, which has originated different efforts and projects over the years. The part of ASIt which is here referred has focused on the design of a new digital library system that was requested to manage and give access to a curated data resource in innovative ways for the final users together with the possibility of using the same digital library system in different scientific contexts. The design has been conducted in a way that makes possible the use of the digital library system and the data resource also as a part of other scientific efforts, making the comparison among phenomena of different dialects easier and speeding up the subsequent process of formal analysis, and making both the system and the resource scalable and expandable for other purposes. One of these further efforts has been recently undertaken using Cimbrian as a test case for synchronic and diachronic language variation, and it needs to be dealt with within the ASIt enterprise[3]. In the context of the management of the Cimbrian variation, the granularity of the data resource is going to change from the sentence level to the word level, but the digital library system is only going to be expanded, and not redesigned, because of its original approach in the design that was supporting modularity and scalability.

3.1 Corpus

Dialectal data stored in the curated resource were gathered during a twenty-year-long survey investigating the distribution of several grammatical phenomena across the dialects of Italy. These data and information were collected by means of questionnaires formed by sets of Italian sentences: dialectal speakers were asked to translate them into their dialects and write their translations in the questionnaire; therefore, each questionnaire is associated with many parallel dialectal translations. At present, there are eight different questionnaires written in Italian and almost 450 questionnaires that are corresponding to the eight questionnaires in Italian and that are written in more than 200 different dialects, for a total of more than 45,000 sentences and more than 10,000 tags stored in the data resource managed by the digital library system.

3.2 Remarks on the Annotation System

The design of a tagset for corpus annotation is normally carried out in compliance with international standards — e.g. CES (Corpus Encoding Standard)[4] — which in turn are based on the specifications of SGML (Standard Generalized Markup Language)[5] and international guidelines like EAGLE (Expert Advisory

[3] http://ims.dei.unipd.it/websites/cimbrian/project

[4] http://www.cs.vassar.edu/CES/

[5] http://www.w3.org/MarkUp/SGML/

Group on Language Engineering Standard)[6] and TEI (Text Encoding Initiative)[7] 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, word, etc. Given the objectives of the ASIt enterprise, we have focused on the tagging of sentence-level phenomena, 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.

First of all, it is worth noting that the ASIt enterprise has a different objective, being a scientific project aiming to account for minimally different variants within a sample of closely related languages. As a consequence, while other tagsets are designed to carry out a gross linguistic analysis of a vast corpus, the ASIt tagset aims to capture fine-grained grammatical differences by comparing various dialectal translations of the same sentence. Moreover, in order to pin down these subtle asymmetries, the linguistic analysis must be carried out manually.

Given its peculiarities, the ASIt team does not need a thorough POS disambiguation, since the 'trivial' identification of basic parts of speech (e.g. Nouns vs Verbs) is not enough to capture cross-linguistic differences between closely related languages. Secondly, the linguistic variants displayed by Italian Dialects cannot be reduced to lexical distinctions, i.e. syntactic differences are in general unpredictable on the basis of the properties of single lexical items. We therefore need a specific tagset designed to capture sentence-level phenomena without taking into consideration POS tags. The requirements at the basis of the ASIt tagset are finally recapitulated below:

- objective: scientific, theoretical;
- focus on grammar;
- analysis: top down;
- minimal unit of analysis: sentence;
- completeness: we lack the analysis of POS;
- accuracy: complete, but the analysis has to be carried out manually.

3.3 Granularity of the Tagged Phenomena

To explain why the needs for ASIt are so special we have to take into consideration two different aspects:

1. the nature of Italian dialects, and
2. the kind of linguistic theory the ASIt data resource aims to be related to.

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

[7] http://www.tei-c.org/index.xml

The Italian dialectal area presents a kind of variation that involves parametric choices affecting many general aspects of syntax, morphology, and phonology. If we concentrate on syntax, we find, for example, a phenomenon that cuts Italy in two, namely subject clitic pronouns: the dialects of Northern Italy have an obligatory subject clitic in at least one person of the verb (some have a subject clitic for all persons of the inflected verb), while Central and Southern Dialects never display subject clitics. Since the nature of clitics is one major topic of theoretical reflection, Italy offers an impressive range of possible variations of this phenomenon.

The kind of information we want to gather from the data resource involves for example not only the presence of a certain element, but also the absence of an element that can be omitted supposedly only in some constructions and in conjunction with specific characteristics of the language. For example, the complementiser *che* ("that") is optionally omitted in Italian varieties in subordinates with subjunctive, conditional, or future tense (mood); so we must have a tag mentioning the "absence of an element". Furthermore, in Southern Dialects there are varieties with two (or even three) specialised complementisers, sensitive to the modality of the subordinate clause and to elements moved in the periphery of the sentence. In this case too the presence is just as important as the absence of a given complementiser.

3.4 Building the Tagset

On the basis of requirements such as those outlined above, we have selected a list of tags capturing relevant phenomena, namely, grammatical properties that are expected to discriminate between dialects, i.e. between grammars. Examples of phenomena sensitive to linguistic variation are: the presence of subject clitics, the syntactic behaviour of different verbal classes (e.g. transitive, unergative and unaccusative verbs), the distribution of negative words (e.g. negation marker "not" and negative indefinites like "none" or "any"), the alternation of finite and infinite verbs in subordinate clauses. Many features that are captured by traditional POS tagsets are clearly not relevant, because they cannot identify grammatical variants within the sample of languages under investigation. For instance, a tag distinguishing concrete vs abstract nouns is relevant to disambiguate different meanings of the same string (e.g. "spirit of wine" vs "spirit of the times"), but, at the same time, concreteness is not expected to play any role in determining grammatical variants within the Italo-Romance domain. In contrast, a semantic feature of nouns like mass vs count is relevant for many phenomena, as is the case for relational or kinship or inalienable possession.

In general, we focussed only on POS features that are relevant to the analysis of sentence-level phenomena. Moreover, we included several POS tags in our tagset that allow us to distinguish classes of lexical items displaying peculiar syntactic behaviours: for instance, different classes of adverbs (like manner, aspectual) occupy different positions in the structure of the clause, kinship nouns differ from other nouns in refusing definite articles and/or requiring possessive adjectives, meteorological verbs in some dialects can obligatorily require an expletive subject. While these classes are not taken into consideration by standard tagsets, they are relevant to our aims because they are involved in syntactic phenomena which distinguish Italian Dialects from each other. Unlike standard tagsets, we also need a subgroup of tags identifying invisible phenomena, e.g. absence of an overt subject, absence of an expected complementizer, absence of a clitic pronoun within a sequence. Standard tagsets are expected to identify and specify any lexical item of the text under analysis, whereas the ASIt system must be sensitive to unpronounced/absent items. Moreover, given the scientific purposes of the ASIt enterprise, our tagset has to be open to new tags in order to capture linguistic phenomena that, according to our new findings, become crucial in distinguishing between grammars.

4 The Digital Library System

In this section we report on the design and construction of the digital library system and its information access component to deal with data curated resources of Italian dialects. A three phase approach was adopted: at the beginning 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 digital library system and the interface to access the data were implemented and verified. In order to efficiently store and manage the amount of data recorded in the questionnaires, the interviews, and the tagged sentences, the component of the digital library system that manages and stores the data is based on the relational database approach, designing and developing a specific relational schema.

In the following subsections we will briefly sketch the three main parts of the digital library system: the linguistic data resource of Italian Dialects, the component that permits the user interaction for updating and creating new linguistic annotations, and the information access component for retrieving and using the data resource.

4.1 A Data Resource of Italian Dialects

The design of the data resource schema needed careful attention on the design of the conceptual schema representing the data at an abstract level. However, prior to the conceptual schema, we carried out a thorough analysis of the requirements in order to generalise, identify, and isolate the main entities, which can be grouped into three broad areas:

- The point of inquiry, which is the location where a given dialect is spoken;
- The administrative area (namely, region, province), the location belongs to;
- The geo-linguistic area, i.e. the linguistic group the dialect belongs to.

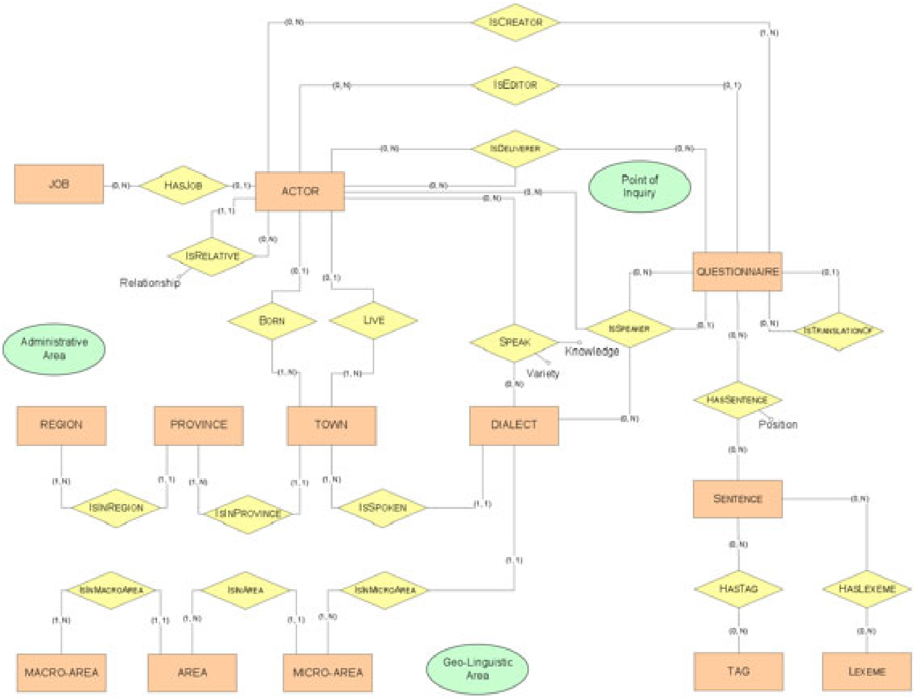


Fig. 1. The conceptual schema of the data resource of Italian dialects: the three main areas of interest are shown with ovals. Attributes of the entities have been removed for better readability.

The result of the conceptual design has generated a conceptual schema where these three areas are considered of central interest, so they were depicted in the conceptual schema that was produced at the end of the conceptual design procedure. The schema is reported in Figure 1 where the three broad areas of interest are represented by ovals.

The "point of inquiry" contains most of the data of the whole data resource. Within this area the following entities were identified and defined:

- DIALECT, the name of the dialect (which normally corresponds to the name of the town or city where the dialect is spoken);
- ACTOR, the user, who can have different roles, namely the person who prepared the questionnaire, the speaker who translated the sentences into a dialect, the editor of the questionnaire, and so on;
- QUESTIONNAIRE, the sets of sentences (either in Italian or in the dialect);
- SENTENCE, the units forming the questionnaires;
- TAG, the linguistic tags specifying the grammatical properties of each sentence.

The entity DIALECT connects all the different parts of the data resource and is therefore like a cornerstone, while the administrative and geo-linguistic areas complete the definition of each dialect, specifying the geographical area where it is spoken and the linguistic subgroup to which it belongs.

4.2 Managing the Tagset

The data resource schema was designed to allow linguists to easily access and analyse the data by retrieving phenomena and/or grammatical elements of one or more geographic locations, together with the characteristics of a specific phenomenon, the co-occurrence of different linguistic phenomena in the same dialect and the context conditioning the differences. In order to reach these objectives, the data can be retrieved on the basis of a set of 194 tags which specify the grammatical properties of each sentence.

This interface has been carefully designed by considering the information gathered during the requirements analysis phase, and in particular:

– The 194 tags exploited by the ASIt system have been grouped into grammatical classes to allow the editor to efficiently manage the whole tagset. Examples of the defined classes include tags regarding subjects, verbs, interrogative/exclamative clauses, clitic pronouns, etc.
– The list of tags associated with each Italian sentence can be automatically associated with the corresponding dialect translation. Then the editor can simply add/remove some tags in order to capture the differences between the Italian input and its dialectal translation.

Translations can be done either through an already reported dialect or by inserting new varieties. In the latter case, some additional details about the new typology of dialect are required, in particular the geographic area. This interface aims at providing users with a very easy and intuitive tool for translating sentences by following a sequence of actions. In fact, first of all, the editor is required to choose the Italian questionnaire to translate, and only once this has been done he/she will gain access to the other parts of the interface. At this point, the user has to define the dialect, and after this he will be able to choose the Italian sentence and insert its translation along with the respective tags. The interface allows for all the most common data editing operations, in particular saving, inserting, and updating.

4.3 Accessing the Linguistic Data Resource

The digital library system makes use of the PostgreSQL[8] relational database management system as component for the management of permanent data. The digital library system stores all the data of interest, among those there are questionnaires, sentences, different translations of the sentences, geographic information about the dialects, and all grammar details related to the tags.

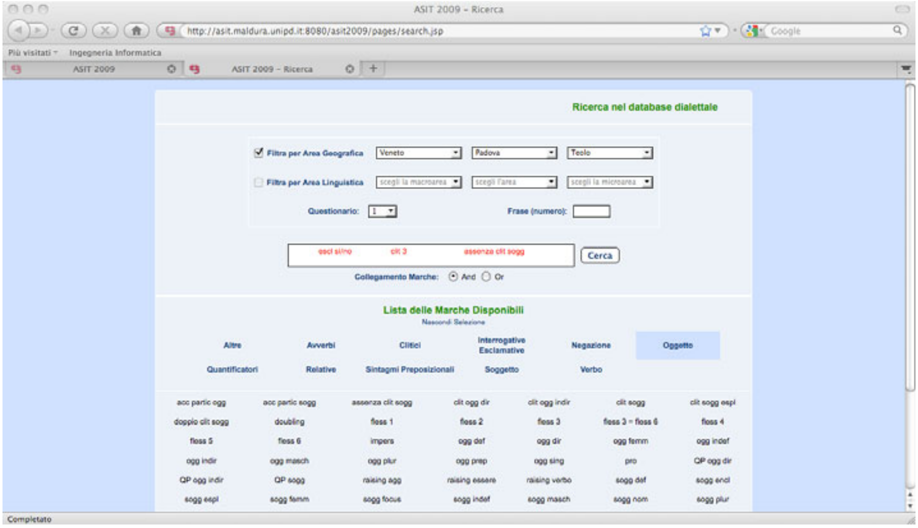
[8] http://www.postgresql.org/

The data resource can be accessed through an interface which allows for searching data and filtering of results. The information access component, which supports function that are similar to those of a search engine, is open to public use and it has been designed to be accessible online. Among the available functions, it allows the user either to visualize the results in the Web page or to download them into a spreadsheet for further analysis of possible correlations. Search operations can be performed in different ways, ranging from simple searches specified by just some tags, to more articulated ones by imposing some filters to the data retrieved. The filters are mainly related to geographical information, but the number of the questionnaire, as well as a particular sentence, can be selected.

The results are tabulated according to the retrieved Italian sentences and right behind appear sub-sections representing all the different translations, together with details about dialects and grammar tags. The visualization of the results is also characterized by a *show/hide* mechanism for the set of translations of an Italian sentence, thus providing the user with a cleaner and more compact representation of the results.

The Web-based interface was designed combining HTML and Javascript for the graphic part, while JavaServer Pages (JSP) technology was used to define the connections with the data resource and all the dynamic operations. JSP was chosen among other script languages mainly for its high portability and the possibility of easily connecting to PostgreSQL. Moreover, an approach based on Ajax (shorthand for asynchronous JavaScript and XML) was exploited to minimize the exchange of data between the clients and the main server.

A screen-shot of the search interface is shown in Figure 2.



Ricerca nel database dialettale

☑ Filtra per Area Geografica: Veneto, Padova, Teolo

□ Filtra per Area Linguistica: scegli la macroarea, scegli l'area, scegli la microarea

Questionario: 1

Frase (numero):

escl si/no clt 3 assenza clt sogg

Cerca

Collegamento Marche: ● And ○ Or

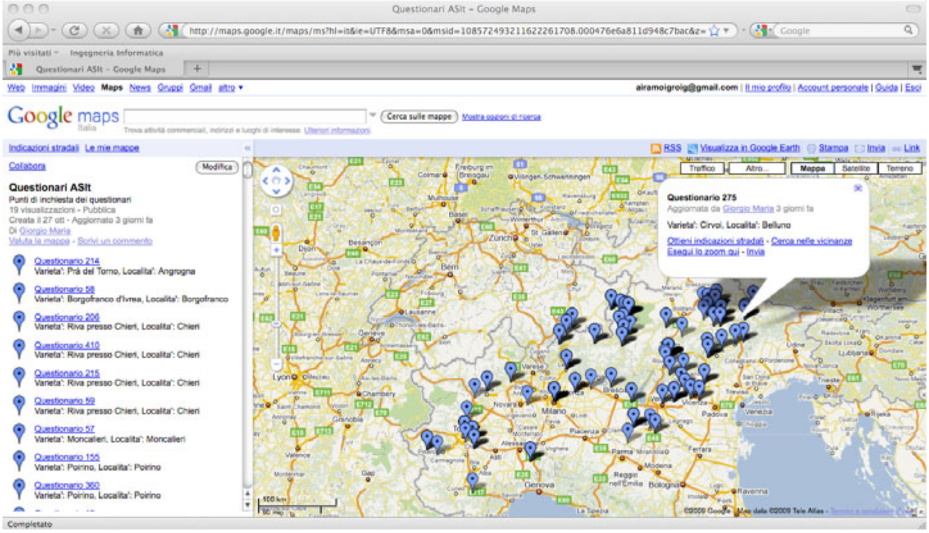
Lista delle Marche Disponibili

Nascondi Selezione

Altre	Avverbi	Clitici	Interrogative Esclamative	Negazione	Oggetto
Quantificatori	Relative	Sintagmi Preposizionali	Soggetto	Verbo	

acc partic ogg	acc partic sogg	assenza clt sogg	clt ogg dir	clt ogg indir	clt sogg	clt sogg espl
doppio clt sogg	doubling	fless 1	fless 2	fless 3	fless 3 = fless 6	fless 4
fless 5	fless 6	impers	ogg def	ogg dir	ogg femm	ogg indef
ogg indir	ogg masch	ogg plur	ogg prep	ogg sing	pro	QP ogg dir
QP ogg indir	QP sogg	raising agg	raising essere	raising verbo	sogg def	sogg encl
sogg espl	sogg femm	sogg focus	sogg indef	sogg masch	sogg nom	sogg plur

Fig. 2. The interface for searching grammatical phenomena in the database of the Italian dialects



Questionari ASIt
Punti di inchiesta dei questionari
19 visualizzazioni - Pubblica
Creata il 27 ott - Aggiornato 3 giorni fa
Di Giorgio Maria
Valuta la mappa - Scrivi un commento

- Questionario 214
  Varieta': Prà del Torno, Localita': Angrogna
- Questionario 58
  Varieta': Borgofranco d'Ivrea, Localita': Borgofranco
- Questionario 206
  Varieta': Riva presso Chieri, Localita': Chieri
- Questionario 410
  Varieta': Riva presso Chieri, Localita': Chieri
- Questionario 215
  Varieta': Riva presso Chieri, Localita': Chieri
- Questionario 59
  Varieta': Riva presso Chieri, Localita': Chieri
- Questionario 57
  Varieta': Moncalieri, Localita': Moncalieri
- Questionario 155
  Varieta': Poirino, Localita': Poirino
- Questionario 360
  Varieta': Poirino, Localita': Poirino

Questionario 275
Aggiornata da Giorgio Maria 3 giorni fa
Varieta': Cirvoi, Localita': Belluno
Ottieni indicazioni stradali - Cerca nelle vicinanze
Esegui lo zoom qui - Invia

Fig. 3. Distribution of grammatical phenomena in questionnaires using GeoRSS tagging and Google maps APIs

Beside the different search options, the digital library system also allows for the visualization of the geographical distribution 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[9], KML[10], etc.) and exploits GoogleMaps[11] APIs to visualize it. An example of the usage of this API is shown in Figure 3 which displays the distribution of a subset of the points of inquiry. This option is very important because a user can graphically view how the dialects are distributed throughout the country, and perform further analysis based on these visualizations.

5 Conclusions and Future Work

Since Summer 2009 the digital library system has been in common use by the team of linguists, whose useful feedback has enabled the refinement and improvement of user interaction with the system. The user functions that have been provided are finally consistent with the purposes of the project, aiming to speed up the comparison of syntactic structures across dialects.

The reached results have been considered of interest as a starting platform for addressing a wider set of languages and phenomena: from the end of 2009,

[9] http://www.georss.org/
[10] http://www.opengeospatial.org/standards/kml/
[11] http://maps.google.it/

the members of the ASIt team have been taking part in a new project aiming to include in the digital library data from German dialects spoken in Italy and to enrich the system with a thorough POS tagset. A Web site reporting on this project has been recently made available to the public[12] from which it will be possible to monitor the evolution of the results here reported.

References

1. Spärck Jones, K.: Computational linguistics: What about the linguistics? Computational Linguistics 33, 437–441 (2007)
2. Kilgarriff, A.: Googleology is bad science. Computational Linguistics 33, 147–151 (2007)
3. 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)
4. Agosti, M.: Information Access using the Guide of User Requirements. In: Agosti, M. (ed.) Access through Search Engines and Digital Libraries, pp. 1–12. Springer, Heidelberg (2008)
5. Benincà, P.: I dati dell'ASIS e la sintassi diacronica. In: Banfi, E., et al. (eds.) Atti del convegno internazionale di studi Trento, Tubingen, Niemeyer, Ottobre 21-23, pp. 131–141 (1995)
6. Benincà, P., Poletto, C.: The ASIS enterprise: a view on the construction of a syntactic atlas for the Northern Italian Dialects. In: Bentzen, K., Vangsnes, Ø.A. (eds.) Nordlyd. Monographic issue on Scandinavian Dialects Syntax, vol. 34, pp. 35–52 (2007)
7. Lord, P., Macdonald, A.: e-Science Curation Report. Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision. The JISC Committee for the Support of Research, JCSR (2003)
8. 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)
9. Agosti, M., Di Nunzio, G.M., Ferro, N.: The importance of scientific data curation for evaluation campaigns. In: Thanos, C., Borri, F., Candela, L. (eds.) Digital Libraries: Research and Development. LNCS, vol. 4877, pp. 157–166. Springer, Heidelberg (2007)

[12] http://ims.dei.unipd.it/websites/cimbrian/home