Dialectal resources on-line: the ALT-Web experience

Nella Cucurullo, Simonetta Montemagni, Matilde Paoli, Eugenio Picchi, Eva Sassolini
Istituto di Linguistica Computazionale – CNR
Via Moruzzi 1 – PISA 56124 – ITALY
{nella.cucurullo,simonetta.montemagni,picchi,eva.sassolini@ilc.cnr.it, tildepaoli@tiscalinet.it}

Abstract

The paper presents an on-line dialectal resource, ALT-Web, which gives access to the linguistic data of the *Atlante Lessicale Toscano*, a specially designed linguistic atlas in which lexical data have both a diatopic and diastratic characterisation. The paper focuses on: the dialectal data representation model; the access modalities to the ALT dialectal corpus; ontology-based search.

1. Introduction

In the field of dialectology the collection of data is the primary requirement. Dialectologists have been assiduous in collecting and archiving a great deal of data, especially involving pronunciation, morphology, syntax and lexical differences. This entails fieldwork, the more detailed and massive the better, and its presentation in different forms. A typical outcome of dialectal research is represented by a linguistic atlas, namely a book of maps which show the distribution of language features over a geographic area. The maps show the locations of features as used by native speakers: these features can be represented either by raw linguistic data, e.g. the responses for a particular questionnaire item (this is the case of so-called "display maps") or by more general statements (this is the case of "interpretive maps").

Yet, data collected by dialectologists in different areas from different informants are linguistic data in their own right and they are susceptible, as such, of different types of analyses and classifications, not only by dialectologists but also by linguists, ethnographers etc. But for these data to be used as a valuable source of linguistic and cultural information for different purposes and from different perspectives, their exclusive representation in terms of linguistic maps cannot always be the optimal solution. As a matter of facts, linguistic maps should rather be seen as one of the possible outcomes of dialectal research. In order to make full use of the information dialectologists have laboriously and painstakingly acquired through fieldwork, dialectal data need to be organized and structured in such a way that each linguistic item is characterized with respect to a number of different dimensions ranging over different levels of linguistic description, i.e. from phonetics, morpho-syntax and syntax to semantics and pragmatics. It goes without saying that such an effort is too expensive if the final goal is map drawing; it becomes worthwhile when map visualization becomes only one (although the prototypical one) out of a number of possible outcomes of dialectal research. In this way, the basic dimensions of research can be enlarged and the possible outcomes can become more sophisticated.

These observations underlie the publication in year 2000 of the *Atlante Lessicale Toscano* 'Lexical Atlas of Tuscany' (henceforth ALT, Giacomelli *et al*. 2000) as a CD-Rom where dialectal data can be retrieved through complex queries taking into account a wide range of parameters interactively defined by the user on the basis of his/her research interests. To give a few examples, ALT dialectal data can be retrieved on the basis of the question through which they were elicited, or of the locality (or more generally the geographic area) in which they were witnessed, or they can be looked up like in a dictionary; they can also be filtered on the basis of informants' features (e.g. age, education, etc.). With the advent of Internet, the CD-Rom version of the Lexical Atlas of Tuscany is being replaced by ALT-Web[1], an on-line dialectal resource which gives access to the entire corpus of linguistic data gathered for the *Atlante Lessicale Toscano* to a widened target audience ranging from professionals to citizens who want to know more about their culture and history.[2] At the time of writing, this is the first resource of this type in Italy, and one of the few at the international level.

The paper illustrates ALT-Web with particular emphasis on: 1) the dialectal data representation model; 2) the access modalities to the ALT dialectal corpus designed to produce an output tailored to the specific needs of the different classes of users (both professionals and common citizens); 3) ontology-based search. These represent three main features which differentiate ALT-Web both from the previous digitalised ALT version (DBT-ALT, Picchi *et al.* 2001) and, most interestingly, from other on-line dialectal resources. The paper is organised as follows. After a brief overview of the ALT project (section 2), ALT-Web is presented as an augmented version of DBT-ALT in terms of both conveyed information (section 3) and access functionalities (sections 4 and 5).

2. The ALT project

The *Atlante Lessicale Toscano* is a specially designed linguistic atlas in which dialectal data have both a diatopic and diastratic characterization. ALT is an *atlas* in the name but it is a much richer resource if we consider the abundance and richness of linguistic information acquired through fieldwork. It goes without saying that socio-cultural (diastratic) variation cannot be easily projected onto a map; the same holds for the rich corpus of additional data recorded during the interviews. ALT is *lexical* in the sense that its main focus is on lexical variation but this does not exclude that it contains valuable information for what concerns e.g. phonetic or morphological variation. Last but not least, it is a *regional* atlas focusing on dialectal variation within Tuscany.

[1] http://serverdbt.ilc.cnr.it/altweb/

[2] The ALT-Web project was financially supported by the Regione Toscana (U.O.C. "Musei, Paesaggio e Attività Culturali").

ALT interviews were carried out in 224 localities of Tuscany, with 2,193 informants selected with respect to a number of parameters ranging from age, socio-economic status to education and culture. Field workers employed a questionnaire of 745 target items, designed to elicit variation mainly in vocabulary, semantics and pronunciation. In particular, informants were asked two main types of questions: onomasiological questions starting from concepts and looking for the lexical items designating them (a typical onomasiological question is "How is this concept designated or named?"), and semasiological questions starting from word forms and asking for their meanings (a typical semasiological question is "Which meanings does this word have?").

Interviewers took down responses to all types of questions in detailed phonetic transcription, indicated any special circumstances of responses and captured informants' comments as well as recorded any other type of attested linguistic evidence even if not directly relevant with respect to a specific questionnaire item. The data were collected between 1974 and 1986, resulting in millions of responses from the 2,193 speakers who were each asked 745 questions (Giacomelli 1987/1988).

In 1985, the digitalisation of the huge corpus of dialectal data collected through fieldwork started. The entire ALT corpus was compacted into about 380,000 database entries partitioned as follows: about 350,000 entries containing different canonical responses to the questionnaire items attested in different locations (inclusive of typical contexts of use and informants' comments) and about 30,000 entries recording dialectal items which have been collected as additional material emerged in the course of interviews. All these entries formed the ALT lexical archive, which was linked to subsidiary archives containing information about the localities of Tuscany which were investigated and the informants who were interviewed. In this way, the prerequisites were created for the selection of data from the lexical archive on the basis of information contained in the subsidiary archives.

In order to account for the richness of collected linguistic information and to make the ALT corpus simultaneously accessible and exploitable from different perspectives, a rather complex and articulated entry model was needed. ALT entries were encoded as bundles of attribute-value pairs each conveying a specific information type (for a detailed description of ALT entries see Montemagni *et al*. 2000). For each entry, the main coordinates LOCALITY, INFORMANT(s) and QUESTION are always specified.

In the ALT lexical archive different entry types can be distinguished, each encoded through a different configuration of attributes:

- canonical responses to questionnaire items, be they onomasiological questions or semasiological ones;
- lexical items which emerged in the course of interviews not directly related to the questionnaire (so-called additional data);
- typical contexts of use of collected lexical items (e.g. phraseology, proverbs as well as short ethnotexts);
- descriptions of customs and beliefs connected with witnessed dialectal data.

All entries may also contain other kinds of specification, for instance informants' or fieldworkers' remarks on the status of attested words (e.g. usage, traditionality, register). In order to enable complex information retrieval, original data recorded as natural language texts were annotated with different levels of linguistic information, ranging from phonetics, morphosyntax and syntax to semantics and pragmatics (for more details see Montemagni *et al*. 2000).

3. The representation of dialectal data

In ALT all dialectal responses, be they individual lexical items or short ethnotexts, were phonetically transcribed. The phonetic alphabet used in the ALT project was a geographically specialized version of the "Carta dei Dialetti Italiani" (CDI) transcription system (Grassi *et al.* 1997: 373-376).

The encoding of phonetically transcribed data is one of the major problems that has to be faced in the construction of computational dialectal resources based on oral interviews. Solutions may differ, depending on the types of analyses phonetically transcribed data should be subjected to. On the one hand, there is the need to ensure a proper treatment of phonetically transcribed data during different automatic analysis stages including editing, sorting, retrieval, on-screen display and printing. On the other hand, there are the specific problems of retrieving phonetically transcribed data: in spite of the fact that, in principle, computers facilitate access to data, narrowness of phonetic transcription may constitute a major difficulty for what concerns their recovery. Here, we are in front of the paradoxical situation in which the user should know in advance the exact phonetic realisation of the word(s) (s)he is looking for, and this may not always be the case. From this, it follows that in the encoding of phonetically transcribed data we are in front of different and sometimes contrasting needs. To overcome the problems sketched above a complex and articulated encoding schema was designed in ALT-Web to fulfill the specific requirements of the different processing tasks.

In the ALT-Web data bank, all dialectal responses are assigned different levels of representation: a first level rendering the original phonetic transcription; other levels containing normalized representations of the original form encoded in standard Italian orthography. In this multi-level representation model, dialectal data are encoded in layers of progressively decreasing detail going from phonetic transcription to different levels of normalized representations abstracting away from details of speakers' pronunciation. In the current representation model the most abstract level neutralizes vital phonetic variation phenomena; more abstract normalization levels (e.g. lemmatization) are envisaged for future developments. With this model the user can select the representation level which best suits her/his needs: dialectal data can be retrieved from lowest (this is the case of phonetic transcription) to highest levels of aggregation (in the case of normalized representations). Section 3.1 describes the encoding of phonetically transcribed data as inherited from DBT-ALT, whereas section 3.2 illustrates the normalization levels associated with original data in ALT-Web.

3.1. Encoding phonetic transcription

In ALT phonetically transcribed data are represented through a hybrid encoding schema including both compositional and atomic representations which,

depending on the task, are automatically converted into each other (see Montemagni and Paoli 1989-90: 36-43).

Compositional representations encode each phonetic symbol with a basic sign which may be further specified through one or more diacritics (conveying information, for instance, about stress or nasality of vowels). This representation type is particularly convenient for inputting and editing ALT data since all different phonetic symbols (about 110) are encoded by means of a restricted number of codes (36 basic signs and 9 diacritics) which can be directly accessed through the computer keyboard. To be more concrete, the compositional representation of a word like /sɛ́kkatɔ́i̯o/, denoting the building used for drying chestnuts, is <se18kkato18i4o> where letters represent basic signs and numbers diacritics: in the case at hand, '1' is a mark for close vowels, '8' indicates the stress and '4' represents a semivowel sound. This type of representation is particularly convenient for both sorting and retrieval tasks: in fact, if basic signs only are considered, it is possible to generalise over phonetic variants. Consider as an example the compositional representation of the word forms /sɛ́kkatɔ́i̯o/ and /sɛ́kkatɛ́i̯o/, which can be seen as distinct phonetic realisations of the same lexical item differing for the quality of the vowel /o/: <se18kkato18i4o> and <se18kkato28i4o>. In both cases, the sequence of basic signs is the same, i.e. <sekkatoio>; this entails that a query starting from this sequence of bases will retrieve both of them.

Atomic representations, on the other hand, show a 1:1 correspondence between ALT phonetic symbols and computer codes; they are typically used for on-screen display and printing. So, to keep with the <se18kkato18i4o> example, the combination of each base together with its diacritics is encoded through a symbol which uniquely identifies it (e.g. /e18/>/ɛ́/).

3.2. Normalization of attested dialectal data

In section 3.1 we pointed out that compositional representations can be of some help to overcome the difficulty of querying a corpus of phonetically transcribed data since they create the prerequisites for queries abstracting away from specific phonetic features (namely those encoded through diacritics). However, this may not always be sufficient to abstract away from phonetic variants of the same word over the Tuscan area. Hence, for the ALT corpus of dialectal data a two-level orthographic transcription system was devised in order to make the dialectal data easily understandable by users not familiar with phonetic notation on the one hand, and to allow for more abstract queries on the other hand. Currently, each phonetically transcribed dialectal item is assigned two different types of orthographically transcribed forms, henceforth referred to as *basic orthographic transcription* and *normalized representation* respectively.

At the first level, attested dialectal data are encoded according to standard Italian orthography: this level of representation is designed to help the non-expert user to understand the phonetically transcribed form. From this it follows that this level of representation seeks to account for the variety of phonetic realizations attested by informants. Yet, Italian orthographical conventions imposed some unavoidable neutralizations, due to the unavailability of the corresponding graphemes. For instance, /skɟa͡ʧʧáta/, /skɟa͡ʧʧáʦa/ and /sʧa͡ʧʧáʦa/ (denoting a traditional type of bread, flat and crispy, seasoned on top with salt and oil) are all assigned the same word *schiacciata* as the corresponding orthographically transcribed form in spite of the fricative realization of /t/ in the second case or of the postpalatal plosive realization of /kɟ/ in the last case. The orthographic transcription of phonetically transcribed data was carried out semi-automatically on the basis of 289 mapping rules (~200 context-sensitive rules and ~90 context-free rules). To show how close the orthographic transcription was with respect to the originally attested form, a normalization factor was calculated as the ratio between the number of different phonetically transcribed forms and the number of different orthographically transcribed forms: the result is 1.13, showing that neutralized representations are resorted to in a quite reduced number of cases.

The second normalization level abstracts away from within-Tuscany vital phonetic variation. Keeping with the *schiacciata* example, this means that the group of variants which are assigned the same normalized form grows to include also words such as /stɟa͡ʧʧáta/, /stɟa͡ʧʧáʦa/, /skɟa͡ʧáta/, /skɟa͡ʧʧáda/, /skɟa͡ʧʧáđa/, /sʧa͡ʧʧɛ́da/, /sʧassɛ́da/, etc. Note that this representation level does not deal with morphological variation: from this it follows that words such as /skɟa͡ʧʧáta/ and /skɟa͡ʧʧáte/ are assigned different normalized forms. This is expected to have a low impact on the retrieval of normalized words since the ALT data, mainly consisting of answers to questionnaire items, show a quite limited range of morphological variation.

Even in this case, normalization was carried out semi-automatically on the basis of a more extended set of mapping rules (i.e. 414); revision of automatically normalized forms was carried out manually with the help of the DBT lemmatization procedure (Picchi 2003).

4. ALT-Web access functionalities

ALT-Web provides flexible and dynamic search procedures which permit the user to interactively define the access key to the corpus of dialectal data and to navigate through it on the basis of his/her research interests. Information can be accessed and retrieved on the basis of a wide range of parameters which can be variously combined; for example, lexical data can be searched on the basis of the location in which they were witnessed and/or the socio-economic features of the informant(s), or of their relevance with respect to a given semantic field or register. ALT-Web also supports the automatic production of dialectal maps starting from query results.

The ALT-Web main navigation page proposes the user two different query types to be selected according to whether (s)he wants a guided trip through the ALT corpus (described in section 4.1) or a personalized search path across the data (reported in section 4.2). In what follows we will refer to the first as *basic access functionality* and to the latter as *advanced access functionality*.

4.1. Basic access functionalities

ALT-Web information needs to be profitable for a large public ranging from the specialized scientist to teachers or anybody who may be interested by Tuscan dialectology related topics. This widened target audience asks for easy access functionalities. To this end, an easy

query interface was devised to make the user to familiarize with the ALT data before asking more advanced queries (if needed). Under this basic access functionality, the number of choices the ALT-Web user needs to make is quite limited, i.e. (s)he can select among two "obliged" search paths corresponding to the typical access keys to the data of a linguistic atlas: namely, the questionnaire item through which the dialectal word was elicited; the locality in which it was witnessed.

4.1.1. Selection based on the questionnaire

With this type of selection, attested dialectal data which directly relate to a given questionnaire item (or more) can be retrieved from the ALT lexical archive. The user can identify the ALT question(s) corresponding to his/her research interests in two different ways: by consulting the entire questionnaire, ordered alphabetically or by question number (this is particularly useful to the user already familiar with the project); or by navigating through a kind of conceptual hierarchy going from very broad classes (so-called *settori*) to intermediate ones (referred to in the query interface as *chiavi*) up to the individual questions (for more details see section 5). This selection can involve an individual question or a group of them interactively defined by the user and can be combined with geographic selection as described below.

4.1.2. Geographic selection

The diatopic dimension of ALT lexical data makes them suitable for geographic selection, i.e. selection based on the locality (or area) in which they were witnessed. This can be done by selecting the locality or the set of localities corresponding to a given area in an alphabetically ordered list or directly on a sensitive map of Tuscany. As in the previous case, geographic selection can be combined with the selection of one or more questionnaire items.

4.1.3. Query results

In both types of selection, the final result is a list of expanded entries satisfying the user requests. Note that for what concerns phonetically transcribed data, the user can choose the representation type(s) (s)he wants to visualize: phonetic transcription, basic orthographic transcription and normalized representation (see section 3).

Figure 1

Independently of the starting point, user queries concerning the results of an individual question produce also a synthesis of canonical responses gathered through fieldwork. For each question, three different lists of answers are given corresponding to the different representation levels as shown in Figure 1.Each list can be ordered alphabetically or by decreasing frequency: this latter ordering is particularly useful to quickly identify the most frequent responses to a given questionnaire item. For instance, the example in Figure 1 reports the most frequent answers (in normalized form) to the onomasiological question 290 *schiacciata* 'flat and crispy bread, seasoned with salt and oil'. For each dialectal item in the lists it is possible either to look up the corresponding entries in the ALT-Web databank or to project the result onto a map.

Figure 2

If projection on the map is selected, the result appears as in Figure 2 which shows the geographic distribution of the dialectal term *ciaccia* for 'schiacciata'. By selecting the other option it is possible to look up the list of entries describing the selected term in the databank. Figure 3 exemplifies one of them describing the dialectal term *ciaccia* (whose different representation types are recorded as value of the attribute "Forma") as attested in Ca' Raffaello (n. 67), by seven informants (1-7); the grammatical category of the word, i.e. noun, is also provided (SO) together with a sketchy description of its referent (which is seasoned with salt, oil and rosemary).

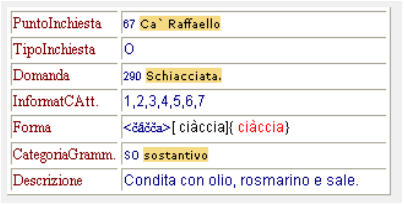

Figure 3

4.2. Advanced access functionalities

With advanced access functionalities the user can ask more complex queries defining personalised search paths across the ALT corpus. In this case, ALT data can be accessed and retrieved on the basis of a wider range of parameters: besides the questionnaire item to which the dialectal word relates and the locality in which it was witnessed, the dialectal corpus as well as the corpus of descriptions provided by informants and fieldworkers can also be queried. Under this access modality, the ALT-Web user can thus choose among four rather than two different search domains.

Advanced access functionalities also include the possibility of filtering query results with respect to a number of extralinguistic and linguistic factors: among them, it is worth mentioning here age, socio-economic and/or cultural background of informants, as well as the

socio-linguistic status and other features associated with the dialectal item.

As in the previous case, for all queries it is possible to select the most appropriate representation level for the visualisation of dialectal data. If the user does not make any choice in this respect, the basic orthographic transcription of dialectal words is given by default, which corresponds to the type of representation understandable by the wider audience.

This access functionality also supports dynamic queries: during the query formulation process, the ALT-Web user is provided continuous feedback in terms of new relevant choices which can be made at that specific point. For instance, projection of the query results onto a map is only allowed when the query involves individual dialectal items; this follows from the fact that in ALT-Web dialectal maps are boolean maps, i.e. they mark all localities where a positive answer to the query is found (in fact, a boolean map of all answers to a given question makes no sense). Another example involves the selection of the sub-corpus on which the query operates: when the user is asking a query about the results gathered with respect to a questionnalire item and/or in a given locality, (s)he is then asked whether the results should be circumscribed a) to the dialectal words given as canonical answers to questionnaire items, or b) to additional data emerged during interviews, or c) they should include both.

In what follows, we will briefly illustrate the query types which are peculiar of this access modality. For the access keys common to the basic access functionality we refer to sections 4.1.1 and 4.1.2 above. The only difference is that these query types are augmented here with new selection parameters and can also be combined with extra-linguistic filters.

4.2.1. Selection by dialectal form

Dialectal data acquired through the interviews can also be accessed by form. In this kind of selection queries are projected onto the corpus of dialectal data, typically represented by lexical items (i.e. the answers to questionnaire items) but also including contexts of use or short ethnotexts.

In this case, the user has first to select the most appropriate representation level with respect to his/her research interests. For instance if (s)he is interested in a specific phonetic realisation of a given word, e.g. /baḍǫ́tti/ for boiled chestnuts with the voiced retroflex plosive /ḍ/, the query should be projected onto the phonetic representation level. On the other hand, if (s)he is looking for all occurrences of the abstract lexical type *ballòtti* then the most appropriate representation level is the normalized one through which 49 items are retrieved with different phonetic realizations (e.g. /balŏt/, /balǫ́to/, /balǫ́tti/, /baḍǫ́tti/, /ballǫ́tti/, etc.). In order to query phonetically transcribed data, the user can type his/her request by using an on-screen virtual keyboard (see Figure 4) containing the symbols of the ALT phonetic alphabet. Note that for each selected phonetic symbol the corresponding base is typed in resulting in an "abstract" query formulated in terms of basic signs only. In this way, different phonetic realizations sharing the same sequence of bases are given back to the user who can thus select the dialectal items (s)he is interested in. If the queried representation level was not among the previously selected visualization levels, the system automatically updates the typology of levels to be visualized by also including it.

Figure 4

4.2.2. Querying the corpus of descriptions

In ALT-Web, queries can also be addressed to the corpus of descriptions and comments by informants or fieldworkers. Queries of this type can be used, for instance, to access the semantic information contained in definitions of dialectal items. In fact, descriptions adopted to semantically and pragmatically characterise ALT dialectal data are similar to dictionary definitions, both from the structural point of view and for the recurring use of a limited and recurring set of "defining formulae". For instance, noun definitions are typically realised as a noun phrase whose syntactic head represents the "genus", which expresses the class to which the "designatum" of the "definiendum" belongs, and whose modifiers represent the "differentia" part of the definition, which reports the properties discriminating the "definiendum" with respect to other members of the same class. Therefore, another parameter on the basis of which the corpus of ALT data can be accessed is represented by meaning components as inferable from the definition text. This parameter can be used to access both canonical and additional data, although it is particularly crucial for the latter whose semantic classification deriving from the questionnaire item they relate to is quite loose and imprecise.

4.2.3. Query results

In all cases, the final result is a list of compact entries matching the request made by the user as exemplified in Figure 5:

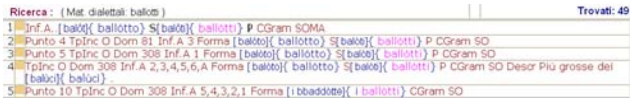

Figure 5

The user can then expand each individual entry in the list to visualise the full entry as reported in Figure 3. This type of visualization makes the query results more readable especially in the case of very productive requests generating a long list of ALT entries. As mentioned before, results can also be projected onto boolean maps to check for the geographical distribution of dialectal data.

5. Ontology-based search

Among the advanced searching capabilities of ALT-Web it is worth mentioning ontology-based search which can be resorted to from both access modalities described above. Through this functionality, access to ALT-Web dialectal data is easier and more effective.

As pointed out above, ALT-Web query parameters include the questionnaire item to which the dialectal word

relates. Yet, it is not always the case that the user knows the questionnaire on the basis of which interviews were carried out. In order to help the user not familiar with the ALT questionnaire, DBT-ALT included keyword-based search, i.e. the different questionnaire items were recoverable through a list of more that 300 keywords (against the 745 questionnaire items). For example, starting from the keyword *recipiente* 'container', the set of ALT questions (namely, 33) dealing with this topic could be identified. To overcome the well known drawbacks of keyword searching, in ALT-Web we turned to ontology-based search which gives the opportunity to the users to take advantage of the ontological data structure.

To this specific end, an ontology organizing the concepts covered by the ALT questionnaire was developed starting from the original classification of the questionnaire into 13 sections corresponding to the investigated semantic domains (e.g. agriculture, food, wild animals, wheather, housing, etc.). These broad semantic classes were used to define the top-level categories of the ALT ontology. Each top level node of the ontology was then structured into finer-grained semantic groupings corresponding to intermediate concepts (on average, each top level node is partitioned into 29 conceptual classes). These intermediate semantic classes were in their turn linked to more specific concepts, expressed in terms of italian lexical items (either simple words or multi-word expressions) roughly corresponding to the onomasiological questions of the ALT questionnaire. Through this hierarchical structure each of the individual concepts investigated by the ALT questionnaire is linked to its lexical variants (or instances) registered all over Tuscany. The very same ontological structure can also be used to explore the different semantic dimensions of dialectal words investigated through semasiological questions (note that the relevance of a given semasiological question with respect to a given concept was established on the basis of a careful analysis of its results). Figure 6 shows how the same ontological structure can be used to explore both the different lexicalization patters of a given concept and the different meanings associated with the same dialectal word.

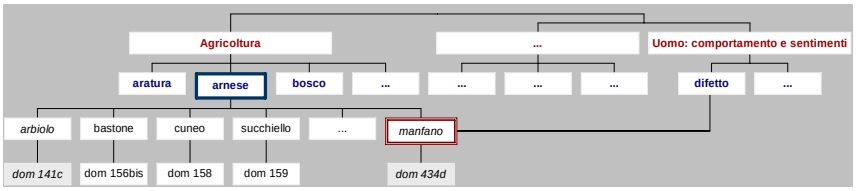

Figure 6

6. Conclusions

We presented ALT-Web, an on-line dialectal resource giving access to the corpus of linguistic data of the *Atlante Lessicale Toscano*. We focused on three main features which make it unique with respect to state-of-the-art on-line dialectal resources. First, the complex and articulated dialectal data representation model allowing for queries concerning the specific pronunciation of a given word as well as queries abstracting away from all the fine points of phonetic transcription. Such a type of flexibility is not offered by other on-line linguistic atlases (see section "Links Utili" in the ALT-Web site) which typically report the dialectal data in phonetic transcription, sometimes accompanied by its recording (the latter is the case of so-called "speaking" linguistic atlases). The only exception is represented by LAMSAS (Linguistic Atlas of the Middle and South Atlantic States) which allows users to view "simplified" phonetic data, i.e. data where diacritical marks of the transcriptions are filtered out (Kretzschmar).

Second, ALT-Web provides flexible and dynamic search procedures which allow the user to define his/her access key to the corpus of dialectal data: in this way, much information which remains normally "hidden" in standard dialectal atlases can be easily retrieved. ALT linguistic data can be accessed and retrieved on the basis of a wide range of parameters (going beyond the canonical access keys of questionnaire and locality) which can be variously combined; ALT data can also be searched on the basis of the socio-economic features of the informant(s). This is another feature characterising ALT-Web with respect to other on-line atlases whose access keys are circumscribed to the questionnaire items and the locality: this is the case, for instance, of ALD (Linguistic Atlas of Dolomitic Ladinian and neighbouring dialects), LAMSAS and ALPI (Linguistic Atlas of the Iberian Peninsula (ALPI). Last but not least, ALT-Web is the only dialectal resource supporting ontology-based search, helping the users to formulate "semantic" queries and to retrieve exactly the information they are interested in.

7. References

Giacomelli G. (1987/1988). *Storia, criteri, metodi, prospettive dell'Atlante Lessicale Toscano*, «Quaderni dell'Atlante Lessicale Toscano», 5/6, pp. 7-25.

Giacomelli G., Agostiniani L., Bellucci P., Giannelli L., Montemagni S., Nesi A., Paoli M., Picchi E., Poggi Salani T. (eds.) (2000). *Atlante Lessicale Toscano*, Lexis Progetti Editoriali, Roma.

Grassi C., Sobrero A., Telmon T. (1997). *Fondamenti di Dialettologia Italiana*, Roma-Bari, Laterza.

Kretzschmar W.A. *Linguistic Databases of the American Linguistic Atlas Project (ALAP)*, available at citeseer.ist.psu.edu/478606.html.

Montemagni S., Paoli M. (1989/1990). *Dalla parola al bit (e ritorno): percorsi dall'inchiesta sul campo alla banca dati dell'ALT*, in AA.VV., «Quaderni dell'Atlante Lessicale Toscano» VI/VIII, pp. 7-52.

Montemagni S., Paoli M., Picchi E. (2000). *DBT-ALT. Manuale di Riferimento*, Lexis Progetti Editoriali, Roma, 2000.

Picchi E., Montemagni S., Bigini L. (2001). *DBT-ALT: A System for Storing and Querying the Data of the Atlante Lessicale Toscano (ALT)*, in «Dialectologia et Geolinguistica (DiG) », vol. 9, pp. 85-103.

Picchi E. (2003). *PiSystem: sistemi integrati per l'analisi testuale*, in A. Zampolli *et al.* (eds.), *Computational Linguistics in Pisa. Linguistica Computazionale*, Special Issue, XVIII-XIX, Pisa-Roma, IEPI, 2003, pp. 597-627.