DBT-ALT: A SYSTEM FOR STORING AND QUERYING THE DATA OF THE ATLANTE LESSICALE TOSCANO

SIMONETTA MONTEMAGNI, EUGENIO PICCHI, LISA BIAGINI

Abstract - *Computers can help dialectologists to make full use of the information they have so laboriously and painstakingly acquired: the basic dimensions of dialectal research can be enlarged and its possible outcomes can become more sophisticated. In this paper, we describe a lexical database for dialectal data, DBT-ALT, which has been designed and constructed to contain linguistic data collected for the Atlante Lessicale Toscano (ALT), a lexical atlas of Tuscany. DBT-ALT is illustrated in detail, with particular emphasis on its search functions which allow for complex queries taking into account a wide range of parameters interactively defined by the user on the basis of his/her research interests.*

Keywords - *computational dialectology, dialectal databases, construction of lexical resources*

1. INTRODUCTION

In the field of dialectology the collection of data is the primary requirement. This entails fieldwork, the more detailed and massive the better, within the limits of practicability, 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 chosen area, as an aid to visualizing the parts of that area where alternative or competing forms are in use. The maps show the locations of features as used by native speakers: these features can be represented either by raw linguistic data (this is the case of so-called "display maps") or by more general statements (this is the case of "interpretive maps").

So far, the use of computers has mainly concentrated on the specific task of drawing linguistic maps (see the survey on

Computational Dialectology in Inoue, 1996a and 1996b). Yet, linguistic maps of either type are only one of the possible outcomes of dialectal research. 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 classifications and organisations which can eventually result in a wide range of different products (Montemagni and Picchi, 1998): e.g. dialectal dictionaries of geographic subareas or even of a given locality; dialectal dictionaries corresponding to a given socio-culturally defined linguistic variety; a sort of dialectal thesaurus where semantically similar words from different or specific areas are grouped together; last but not least, linguistic atlases. This is the reason why, in our opinion, it would be inappropriate to restrict the role of computers in the field of dialectology to the only task of map drawing.

Due to their powerful search and selection capacities on large quantities of data, computers can be used to experiment with different configurations of the same corpus of dialectal data, thus making full use of the abundance and richness of acquired linguistic information (Montemagni and Zampolli, 1987). In order to make dialectal data simultaneously accessible and exploitable from different perspectives, they need to be organised in a database structure where each linguistic item is characterised with respect to a number of different dimensions ranging over different levels of linguistic description, i.e. from phonetics, morphosyntax and syntax to semantics and pragmatics. This is a rather time consuming process which may appear too expensive if the only goal is map drawing. Such an effort becomes worthwhile if the set of maps constituting the linguistic atlas becomes only one (although the prototypical one) out of a number of possible outcomes of dialectal research.

The paper intends to shed light on this specific issue, i.e. on how the computer can be used to exploit collected dialectal data to the full. This will be illustrated through the experience of the *Atlante Lessicale Toscano*, henceforth ALT (Giacomelli *et al.*, 2000), a lexical atlas of Tuscany. In particular, we will focus on DBT-ALT, the lexical database which was constructed for the storage, management and interrogation of the corpus of ALT dialectal data.

2. BACKGROUND

DBT-ALT is a modular system for the storage, management and interrogation of 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 ALT lexical data bank contains the results of interviews 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), on the basis of a questionnaire of 745 items. More than 350,000 different responses were collected; these canonical responses were integrated with additional material emerging in the course of interviews (about 30,000 dialectal items). This entails that each lexical item in the ALT data bank is always specified both for the locality in which it was witnessed and for the informants who attested it.

DBT-ALT is a specialised version of the textual database system known as DBT (Picchi, 1991), developed by Eugenio Picchi at the Istituto di Linguistica Computazionale (ILC) of the Italian National Research Council (CNR). DBT, in its original configuration, is a textual database system for storing and querying large text archives whose basic functions include: a sophisticated query system to access the text by means of a number of different functions; generation of indices of all words occurring in the text; generation of concordances; an application tool engine. DBT is the core component of the PI-System (Picchi, this volume), a set of procedures specifically designed and developed to meet the various requirements of literary and linguistic text processing and analysis.

DBT has been implemented in different configurations to perform specific text and dictionary processing tasks. Among the specific problems tackled by DBT in its various versions, there are some which are of specific interest for ALT, namely: i) the management of structured linguistic data (as in the case of dictionaries); ii) the processing of non-Latin alphabets. DBT-ALT is a version of DBT that includes these functionalities together with new ones, aimed at meeting the combined needs of geolinguistic and sociolinguistic research as emerging from ALT. Among the added functionalities there is also the automatic generation of dialectal maps. A general description of the design of the DBT-ALT system and its underlying motivations can be found in Agostiniani *et al.* (1992).

3. Overall architecture

The modular architecture of DBT-ALT is sketched in figure 1:

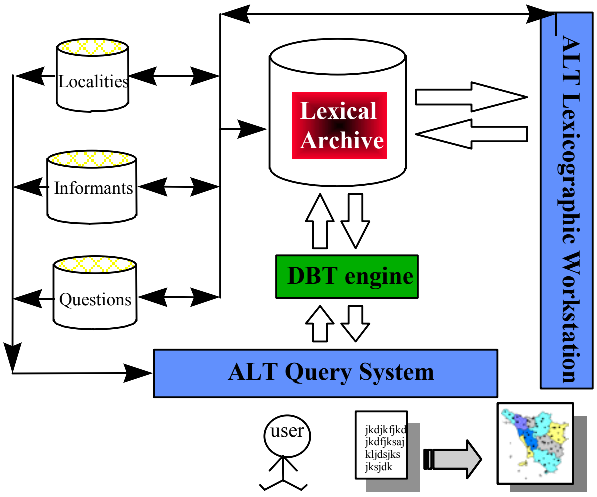


Figure 1. The modular architecture of DBT-ALT

The Lexical Archive (LA), which contains all linguistic data collected through the interviews, is linked to a system of subsidiary archives (SAs) containing information about the localities of Tuscany which were investigated, the informants who were interviewed and the questionnaire on the basis of which lexical data were elicited. These archives were created using the Lexicographic Workstation developed in the general framework of the PI-System to assist the lexicographer in the various activities involved in the creation and revision of dictionaries (Picchi, 1992; Picchi *et al.*, 1992).

Explorations in LA are dealt with by the ALT Query System (ALT-QS) which passes the user request on to the DBT core engine which, in its turn, projects it onto LA, from which the query results are extracted. Selection of lexical data can also be performed on the basis of information contained in the subsidiary archives (see the links between SAs and LA on the one hand, and between SAs and ALT-QS on the other).

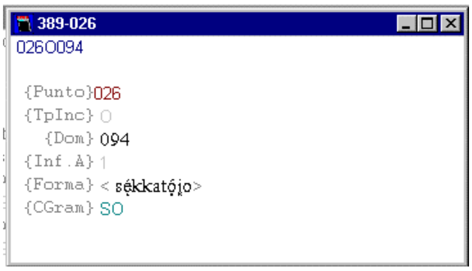
4. DBT-ALT LEXICAL ENTRY

Having sketched the macro-structure of the system, let us consider now the micro-structure of data, i.e. the model according to which collected linguistic data have been encoded in the Lexical Archive. In order to represent the richness of collected linguistic information and thus to enable complex information retrieval, a rather complex and articulated structure was needed: ALT entries present themselves as bundles of attribute-value pairs each of which specifies a specific information type (for a detailed description of ALT entries see Montemagni and Paoli, 1989-90; Montemagni *et al.*, 2000). For each entry, the main coordinates LOCALITY, INFORMANT(s) and QUESTION are always specified. The ALT Lexical Archive contains different entry types:

- canonical responses to questionnaire items;
- lexical items which emerged in the course of the interview but which are not directly related to the questionnaire (so-called additional data);
- typical contexts of use of collected lexical items (e.g. phraseology, proverbs);
- descriptions of customs and beliefs related to collected data.

Each entry type is encoded through a different configuration of attributes. All entries may also contain other kinds of specification expressed in terms of codes, for instance informants' or fieldworkers' remarks on the status of words (e.g. usage, traditionality, register).

To be more concrete, the following are some examples of ALT entries. The record reported in figure 2 below represents one of the canonical answers to question 94 ({Dom} 094), seeking for terms denoting the building used for drying chestnuts.



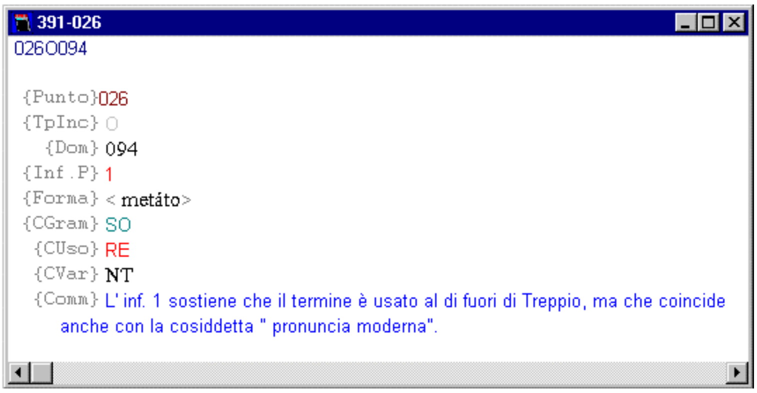
389-026

```
0260094

{Punto}026
{TpInc} 0
  {Dom} 094
{Inf.A} 1
{Forma} < sẹ́kkatọ́i̯o>
{CGram} SO
```

Figure 2. A prototypical entry of the ALT Lexical Archive

The term <sẹ́kkatọ́i̯o> (recorded as value of the attribute {Forma}) was attested in Treppio ({Punto} 036), a locality in the mountains of the Tosco-Emilian Appennines, by an old informant ({Inf.A.} 1); the grammatical category of the word in question, i.e. noun, is also specified ({CGram} SO). This is the prototypical entry of the ALT Lexical Archive. Yet, responses can be much richer in information thus giving rise to more complex entry structures, as the one shown in figure 3:

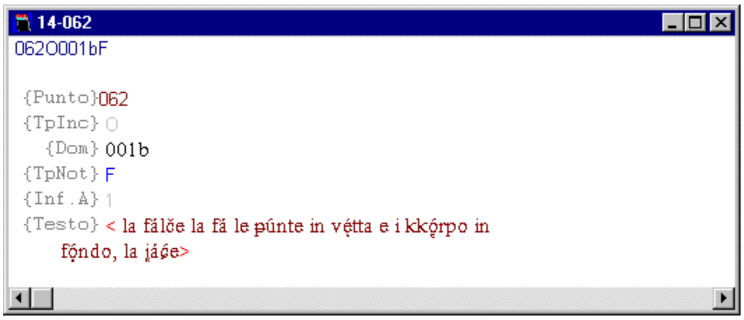


```
391-026
0260094

{Punto}026
{TpInc} 0
  {Dom} 094
{Inf.P} 1
{Forma} < metáto>
{CGram} SO
 {CUso} RE
 {CVar} NT
 {Comm} L' inf. 1 sostiene che il termine è usato al di fuori di Treppio, ma che coincide
      anche con la cosiddetta " pronuncia moderna".
```

Figure 3. A complex entry in the ALT Lexical Archive

This record contains another canonical answer by the same informant in the same locality to the same question as above. In this case, the attested term <metáto> is not actively used by the informant; this information is conveyed by the fact that the informant is specified as value of the {Inf.P.} attribute (as opposed to the {Inf.A.} attribute specifying informants who actively use the word described in the entry). The term was also qualified by the informant as recent ({CUso} RE) and not traditional ({CVar} NT). As a last example, we selected a phraseological entry with an etnotext providing a detailed description of the object of the posed question, i.e. the rising moon.



```
14-062
0620001bF

{Punto}062
{TpInc} 0
  {Dom} 001b
{TpNot} F
{Inf.A} 1
{Testo} < la fálče la fá le ᵽúnte in vẹ́tta e i kkọ́rpo in
     fọ́ndo, la j̣áǵe>
```

Figure 4. A phraseological entry in the ALT Lexical Archive

Obviously these few examples do not exhaust the wide typology of ALT entries but they are just meant to give the reader the flavour of the variety of information contained in the ALT Lexical Archive.

4.1. *Encoding of phonetically transcribed data*

As it can be noticed in the ALT sample entries above, data in the ALT Lexical Archive are either phonetically transcribed or represented according to standard Italian orthography; in this respect, LA can be seen as a kind of bilingual text archive. Phonetically transcribed data are constituted by responses to questionnaire items and additional data as well as by typical contexts of use of attested words or expressions. In what follows we will concentrate on the representation of phonetically transcribed data.

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. The phonetic alphabet used in the ALT project fieldwork was a geographically specialised version of the "Carta dei Dialetti Italiani" (CDI) transcription system (Grassi *et al.*, 1997: 373-376). In order to ensure a proper treatment of phonetically transcribed data during the different automatic analysis stages, a complex encoding schema was designed to fulfil the specific requirements of different tasks: editing, sorting, retrieval, on-screen display and printing. This encoding schema includes compositional and atomic representations which, depending on the task, are automatically converted into each other; for a detailed description of this hybrid encoding schema 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) can be 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 the term

<sẹ́kkatọ́i̯o> seen above is constituted by the string of characters <se18kkato18i4o>, where letters represent basic signs and numbers diacritics: in the case at hand, ‘1’ specifies that the preceding vocalic base is closed, ‘8’ indicates the stress and ‘4’ that the preceding base is semivocal. This type of representation is particularly convenient for both sorting and retrieval phases: 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 <sẹ́kkatọ́i̯o> and <sẹ́kkatǫ́i̯o>, which can be seen as distinct phonetic realisations of the same lexical item (differing for the quality of the vowel /o/): <se18kkato18i4o> and <se18kkato28i4o>. As it can be noticed, in both cases the sequence of basic signs is the same, i.e. <sekkatoio>; this entails that at the level of sorting both word forms will appear together; similarly, a query starting from this sequence of bases will retrieve them both (see section 5.1.3 below).

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

5. DBT-ALT QUERY SYSTEM

The DBT-ALT query system provides dynamic and flexible search procedures which permit the user to interactively define his/her access key to dialectal data and thus navigate through the corpus on the basis of his/her research interests. With such a sophisticated query system, much information which remains normally hidden in printed dialectal resources (either linguistic atlases or dialectal dictionaries) can easily be retrieved: for an introduction to alternative research paths through the corpus of ALT data see Montemagni and Paoli (1997).

In what follows, the DBT-ALT query system is illustrated through the following steps: first, in section 5.1, the typology of query parameters is described; second, in section 5.2, complex queries based on parameter combinations are exemplified; finally, in section 5.3, the filtering of query results on the basis of both linguistic and extralinguistic criteria is discussed.

5.1. *Query parameters: typology*

Lexical data can be accessed and retrieved on the basis of a wide range of parameters. The list which follows exemplifies the main ones:

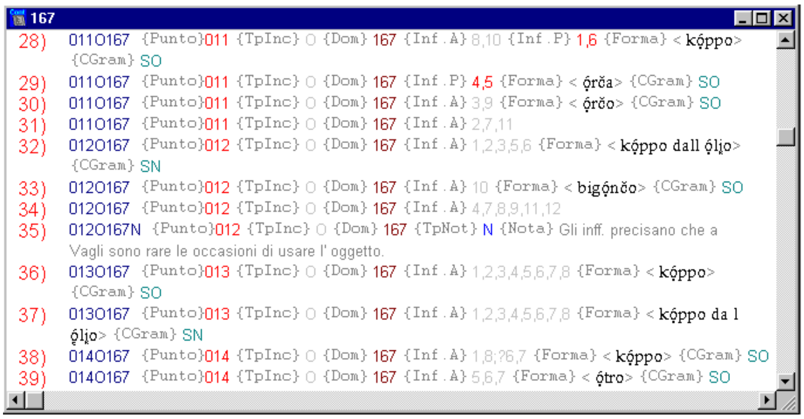
1. questionnaire item to which the lexical item relates;
2. locality in which it was witnessed;
3. phonetic realisation;
4. meaning components as inferable from the definition text.

Each of these parameters corresponds to specific attributes of the entry to which the query is addressed. Actually, they represent only the most typical ones since the range of parameters on the basis of which queries can be formulated is much wider, corresponding to the typology of attributes used to describe ALT entries.

5.1.1. *Selection based on the questionnaire*

With this type of selection, attested dialectal data which relate (either directly or indirectly) to a given questionnaire item can be extracted from the ALT Lexical Archive.

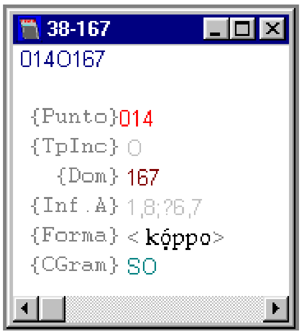
Consider as an example question n. 167, concerning the different terms denoting the oil jar. At this level, the user can choose whether (s)he wants to select the canonical answers to this question or also related additional material. Let us suppose that the user opts for canonical material only; in such a case, the result of the query can be seen as corresponding to a dialectal map in the form of list of answers. Figure 5 below reports an excerpt of the obtained result, which is the full documentation of the canonical material collected through question n. 167.



```
28)  0110167 {Punto}011 {TpInc} O {Dom} 167 {Inf.A} 8,10 {Inf.P} 1,6 {Forma} < kọ́ppo>
     {CGram} SO
29)  0110167 {Punto}011 {TpInc} O {Dom} 167 {Inf.P} 4,5 {Forma} < ọ́rča> {CGram} SO
30)  0110167 {Punto}011 {TpInc} O {Dom} 167 {Inf.A} 3,9 {Forma} < ọ́rčo> {CGram} SO
31)  0110167 {Punto}011 {TpInc} O {Dom} 167 {Inf.A} 2,7,11
32)  0120167 {Punto}012 {TpInc} O {Dom} 167 {Inf.A} 1,2,3,5,6 {Forma} < kọ́ppo dall ọ́ljo>
     {CGram} SN
33)  0120167 {Punto}012 {TpInc} O {Dom} 167 {Inf.A} 10 {Forma} < bigọ́nčo> {CGram} SO
34)  0120167 {Punto}012 {TpInc} O {Dom} 167 {Inf.A} 4,7,8,9,11,12
35)  0120167N {Punto}012 {TpInc} O {Dom} 167 {TpNot} N {Nota} Gli inff. precisano che a
     Vagli sono rare le occasioni di usare l' oggetto.
36)  0130167 {Punto}013 {TpInc} O {Dom} 167 {Inf.A} 1,2,3,4,5,6,7,8 {Forma} < kọ́ppo>
     {CGram} SO
37)  0130167 {Punto}013 {TpInc} O {Dom} 167 {Inf.A} 1,2,3,4,5,6,7,8 {Forma} < kọ́ppo da l
     ọ́ljo> {CGram} SN
38)  0140167 {Punto}014 {TpInc} O {Dom} 167 {Inf.A} 1,8;?6,7 {Forma} < kọ́ppo> {CGram} SO
39)  0140167 {Punto}014 {TpInc} O {Dom} 167 {Inf.A} 5,6,7 {Forma} < ọ́tro> {CGram} SO
```

Figure 5. A partial list of answers to question 167

A full display of each record constituting this list can easily be obtained. Figure 6 below contains the full display of record 38, which describes the answer to question 167 provided by the informants labelled as 1, 8 and 6, 7 (the latter with some doubts) in Camporgiano (locality 14).



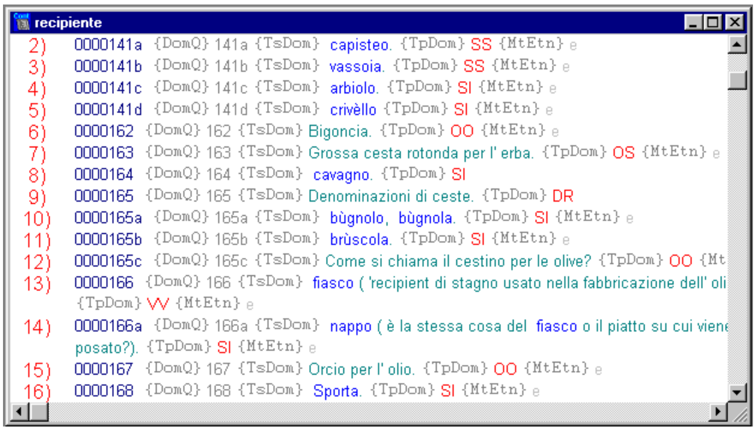
```
0140167

 {Punto}014
 {TpInc} O
   {Dom} 167
 {Inf.A} 1,8;?6,7
 {Forma} < kọ́ppo>
 {CGram} SO
```

Figure 6. Full display of record 38

For each displayed record, the user can also consult the subsidiary archives and get more detailed information about: i) the locality in which the lexical item was gathered, and ii) the informants who attested the described entry (e.g. age, sex, education and professional status).

However, the user does not always know the questionnaire on the basis of which interviews were carried out. If this is the case, a search function operating on the basis of keywords helps the user to identify the questionnaire items relevant to the subject (s)he is interested in. For example, through the keyword *recipiente* (Italian for container), the set of ALT questions dealing with this topic can be identified; it amounts to 33 items, some of which are reported in figure 7. The reader will note that question 167 is among these (n. 15); starting from the result of this query the user is in a position to select data on the basis of the questionnaire items through which they were elicited, as illustrated in the first part of this section.

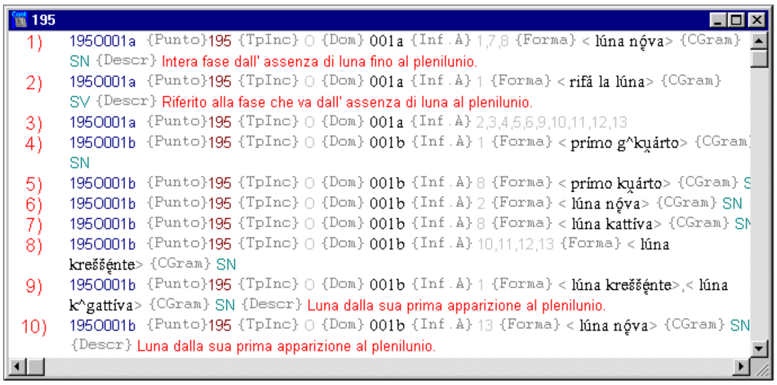


```
 2)  0000141a {DomQ} 141a {TsDom} capisteo. {TpDom} SS {MtEtn} e
 3)  0000141b {DomQ} 141b {TsDom} vassoia. {TpDom} SS {MtEtn} e
 4)  0000141c {DomQ} 141c {TsDom} arbiolo. {TpDom} SI {MtEtn} e
 5)  0000141d {DomQ} 141d {TsDom} crivèllo {TpDom} SI {MtEtn} e
 6)  0000162 {DomQ} 162 {TsDom} Bigoncia. {TpDom} OO {MtEtn} e
 7)  0000163 {DomQ} 163 {TsDom} Grossa cesta rotonda per l' erba. {TpDom} OS {MtEtn} e
 8)  0000164 {DomQ} 164 {TsDom} cavagno. {TpDom} SI
 9)  0000165 {DomQ} 165 {TsDom} Denominazioni di ceste. {TpDom} DR
10)  0000165a {DomQ} 165a {TsDom} bùgnolo, bùgnola. {TpDom} SI {MtEtn} e
11)  0000165b {DomQ} 165b {TsDom} brùscola. {TpDom} SI {MtEtn} e
12)  0000165c {DomQ} 165c {TsDom} Come si chiama il cestino per le olive? {TpDom} OO {Mt
13)  0000166 {DomQ} 166 {TsDom} fiasco ( 'recipient di stagno usato nella fabbricazione dell' oli
     {TpDom} VV {MtEtn} e
14)  0000166a {DomQ} 166a {TsDom} nappo ( è la stessa cosa del fiasco o il piatto su cui viene
     posato?). {TpDom} SI {MtEtn} e
15)  0000167 {DomQ} 167 {TsDom} Orcio per l' olio. {TpDom} OO {MtEtn} e
16)  0000168 {DomQ} 168 {TsDom} Sporta. {TpDom} SI {MtEtn} e
```

Figure 7. The set of ALT questions dealing with containers

5.1.2. *Geographic selection*

Given their diatopic characterisations, ALT lexical data can be selected on the basis of the locality in which they were witnessed. Consider as an example Santa Fiora (Grosseto), a locality on the mountains in the south part of Tuscany: figure 8 below contains the first ten records of the answers to the ALT questionnaire in this locality. In this way, the user can reconstruct the result of interviews carried out with different informants in the same locality.



```
1)  1950001a {Punto}195 {TpInc} 0 {Dom} 001a {Inf.A} 1,7,8 {Forma} < lúna nọ́va> {CGram}
    SN {Descr} Intera fase dall' assenza di luna fino al plenilunio.
2)  1950001a {Punto}195 {TpInc} 0 {Dom} 001a {Inf.A} 1 {Forma} < rifă la lúna> {CGram}
    SV {Descr} Riferito alla fase che va dall' assenza di luna al plenilunio.
3)  1950001a {Punto}195 {TpInc} 0 {Dom} 001a {Inf.A} 2,3,4,5,6,9,10,11,12,13
4)  1950001b {Punto}195 {TpInc} 0 {Dom} 001b {Inf.A} 1 {Forma} < prímo g^ku̯árto> {CGram
    SN
5)  1950001b {Punto}195 {TpInc} 0 {Dom} 001b {Inf.A} 8 {Forma} < prímo ku̯árto> {CGram} S
6)  1950001b {Punto}195 {TpInc} 0 {Dom} 001b {Inf.A} 2 {Forma} < lúna nọ́va> {CGram} SN
7)  1950001b {Punto}195 {TpInc} 0 {Dom} 001b {Inf.A} 8 {Forma} < lúna kattíva> {CGram} S
8)  1950001b {Punto}195 {TpInc} 0 {Dom} 001b {Inf.A} 10,11,12,13 {Forma} < lúna
    kreššẹ́nte> {CGram} SN
9)  1950001b {Punto}195 {TpInc} 0 {Dom} 001b {Inf.A} 1 {Forma} < lúna kreššẹ́nte>,< lúna
    k^gattíva> {CGram} SN {Descr} Luna dalla sua prima apparizione al plenilunio.
10) 1950001b {Punto}195 {TpInc} 0 {Dom} 001b {Inf.A} 13 {Forma} < lúna nọ́va> {CGram} SN
    {Descr} Luna dalla sua prima apparizione al plenilunio.
```

Figure 8. Geographic selection of ALT data: Santa Fiora (Grosseto)

5.1.3. *Selection by form*

Lexical data acquired through the interviews can also be accessed by form. This kind of selection can be compared to the lookup of an alphabetical index. In this case, there is the additional problem of retrieving phonetically transcribed data: in fact, the retrieval of phonetically transcribed data poses specific problems which require *ad hoc* solutions.

In spite of the fact that, in principle, computers facilitate access to data, narrowness of phonetic transcription may constitute a major difficulty in their recovery. 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. As illustrated in section 4.1, compositional representations are of some help to overcome this difficulty since they permit the user to formulate his/her query by abstracting away from specific phonetic features (i.e. those encoded through diacritics). However, this type of generalisation may not always be sufficient to abstract away phonetic variants of the same word over the Tuscan area. Hence, for retrieval purposes we decided to devise two different abstraction levels:

- level 1: the retrieval of word forms operates on basic signs only and ignores diacritic signs;
- level 2: more powerful generalisations are allowed by clustering together different basic signs or combinations.

Consider as an example the term <ski̯aččáta>, denoting a traditional type of bread, flat and crispy, seasoned on top with salt and oil. A search of this term operating at level 1, i.e. on basic signs only, will retrieve two different phonetic realisations, namely:

1. <ski̯aččáta>
2. <ski̯aččáŧa>

the first one with the voiceless alveolar plosive /t/ and the second one with the corresponding fricative /ŧ/. At this level, the retrieval of different word forms is possible as long as the sequence of basic signs is the same (which, in the case at hand, is <skiaččata>). With this query, 62 records containing the basic sequence <skiaččata> were found.

Yet, there may be phonetic variants which cannot be captured through this low level of abstraction. For instance, in <ski̯aččáta> the voiceless velar plosive followed by the semivowel (/ki̯/) can be alternatively realised, for instance, as a dental plosive (i.e. as /ti̯/) or as a postpalatal plosive (i.e. as /c̋/). Moreover, the palatoalveolar affricate can be either strongly or weakly articulated (i.e. as /čč/ or /č/). These variation types require a higher level of abstraction since different basic signs are involved in the encoding of the alternating phonetic realisations. A search of the same word performed by selecting level 2 of abstraction will recover more data, namely occurrences of:

1. <ski̯aččáta>
2. <ski̯aččáŧa>

3. <sti̯aččáta>
4. <sti̯aččáŧa>
5. <sc̋aččáta>
6. <sc̋aččáŧa>
7. <st′aččáta>
8. <st′aččáŧa>
9. <ski̯ačáta>
10. …

In this case, 309 contexts (as opposed to the previous 62) have been retrieved containing phonetic variants of <ski̯aččáta>.

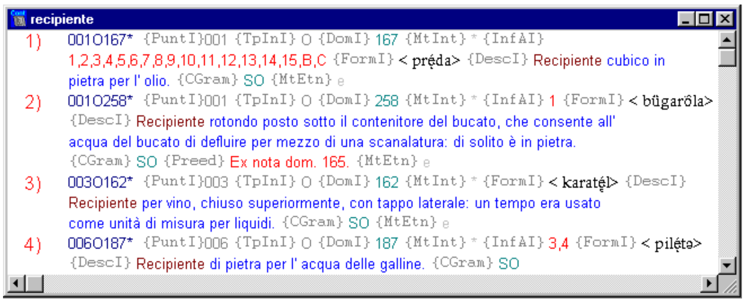
Depending on the user needs, either of the two abstraction levels is best suited. For instance, with level 2 a better recall is obtained, i.e. more data are retrieved, but precision may be lower since some noisy data could also be included in the query result. On the contrary, level 1 guarantees higher precision at the price of a lower recall. For more details on this recovery strategy of phonetically transcribed data the interested reader is referred to Agostiniani *et al*. (1998).

5.1.4. *Semantic selection*

In the field of computational lexicography, extraction of semantic information from dictionary definitions - specifically taxonomic information and other semantic relations - has nowadays become a common practice. Semantic information can be automatically extracted based on regularities that occur in definitions, both in their structure and in the recurring and systematic use of a limited set of "defining formulae". Generally, dictionary definitions adhere to a rather rigid stylistic form. 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. The extraction of the "genus" term can take advantage of the definition structure (Calzolari, 1984; Chodorow *et* *al*., 1985). As for the semantic information contained in the "differentia" part of the definition, its extraction is based on the observation that there are "defining formulae" in the definitions that systematically express conceptual categories, as well as semantic relations (Markowitz *et al*., 1986; Calzolari and Picchi, 1988).

Descriptions adopted to semantically and pragmatically characterise ALT lexical 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". As a consequence, a similar extraction procedure can be adopted to navigate in the ALT Lexical Archive. 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 what concerns the latter: in fact, as illustrated in section 5.1.1 above, canonical data can be retrieved by means of semantic keywords which classify the questions of which they represent the answer.

Let us illustrate an example of semantic selection of ALT data. Suppose that the user is interested in all kinds of containers. By searching the word *recipiente* (container) (as well as other Italian words denoting containers) within the definition text, (s)he will extract all ALT entries denoting containers. At this level, the user can circumscribe the domain of his/her research, for instance by restricting it to the set of additional data only; an excerpt of the obtained results is reported in figure 9:



```
1)  0010167* {PuntI}001 {TpInI} O {DomI} 167 {MtInt} * {InfAI}
    1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,B,C {FormI} < prẹ́da> {DescI} Recipiente cubico in
    pietra per l' olio. {CGram} SO {MtEtn} e
2)  0010258* {PuntI}001 {TpInI} O {DomI} 258 {MtInt} * {InfAI} 1 {FormI} < bŭgarôla>
    {DescI} Recipiente rotondo posto sotto il contenitore del bucato, che consente all'
    acqua del bucato di defluire per mezzo di una scanalatura: di solito è in pietra.
    {CGram} SO {Preed} Ex nota dom. 165. {MtEtn} e
3)  0030162* {PuntI}003 {TpInI} O {DomI} 162 {MtInt} * {FormI} < karatẹ́l> {DescI}
    Recipiente per vino, chiuso superiormente, con tappo laterale: un tempo era usato
    come unità di misura per liquidi. {CGram} SO {MtEtn} e
4)  0060187* {PuntI}006 {TpInI} O {DomI} 187 {MtInt} * {InfAI} 3,4 {FormI} < pilẹ́tə>
    {DescI} Recipiente di pietra per l' acqua delle galline. {CGram} SO
```

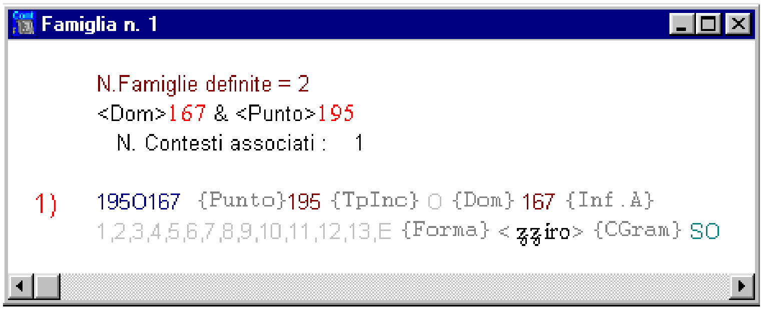
Figure 9. ALT entries denoting containers from additional data

5.2. *Querying through combinations of different parameters*

Individual parameters such as those exemplified in section 5.1 can be variously combined to form complex queries in which items which are looked for are linked by the logical operators AND, OR and NOT. In what follows, we will consider two different types of complex query: i) the co-occurrence of different information types within the same record, and ii) the occurrence of one out of a set of variants.

5.2.1. *Looking for the co-occurrence of different information types within the same entry*

For certain query types, different conditions have to be met simultaneously. In such a case, the query has to be formulated as a conjunction of the different conditions which have to be met: schematically, "cond_1 AND cond_2 … AND cond_x". This function can for instance be used to extract the answer to a given question in a specified locality, e.g. to immediately identify the answer to question 167 in locality 195 without having to scan the full list of answers (ordered either way) up to the right point. The result of this complex query is reported in figure 10:

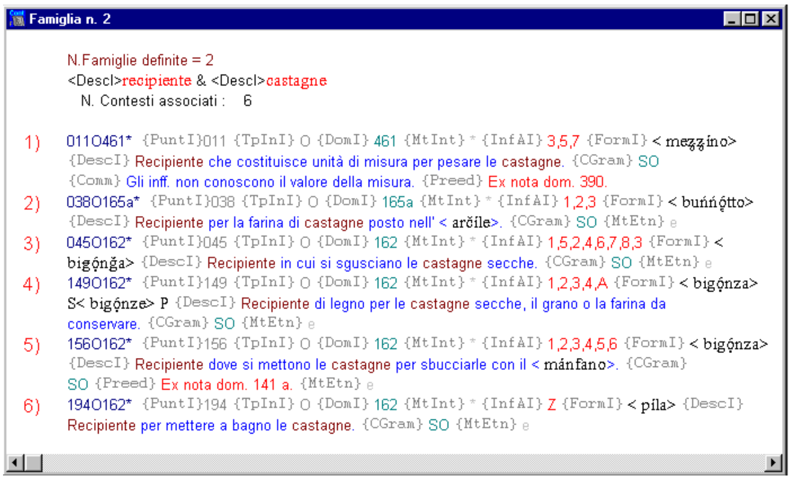


```
      N.Famiglie definite = 2
      <Dom>167 & <Punto>195
        N. Contesti associati :    1

1)    1950167  {Punto}195 {TpInc} O {Dom} 167 {Inf.A}
      1,2,3,4,5,6,7,8,9,10,11,12,13,E {Forma} < ẓẓiro> {CGram} SO
```

Figure 10. Result of a complex query: the answer to question 167 in locality 195

This kind of procedure can be applied to information recorded as value of different attributes of the same entry (as in the example above where conditions on both locality and question have been enforced) as well as to the value of the same attribute. Consider for example the case in which the user wants to extract from the collected additional material the one concerned with ‘containers for chestnuts’.



```
      N.Famiglie definite = 2
      <Descl>recipiente & <Descl>castagne
        N. Contesti associati :   6

1)    0110461* {PuntI}011 {TpInI} O {DomI} 461 {MtInt} * {InfAI} 3,5,7 {FormI} < meẓẓíno>
      {DescI} Recipiente che costituisce unità di misura per pesare le castagne. {CGram} SO
      {Comm} Gli inff. non conoscono il valore della misura. {Preed} Ex nota dom. 390.
2)    0380165a* {PuntI}038 {TpInI} O {DomI} 165a {MtInt} * {InfAI} 1,2,3 {FormI} < buńńọ́tto>
      {DescI} Recipiente per la farina di castagne posto nell' < arčíle>. {CGram} SO {MtEtn} e
3)    0450162* {PuntI}045 {TpInI} O {DomI} 162 {MtInt} * {InfAI} 1,5,2,4,6,7,8,3 {FormI} <
      bigọ́ngă> {DescI} Recipiente in cui si sgusciano le castagne secche. {CGram} SO {MtEtn} e
4)    1490162* {PuntI}149 {TpInI} O {DomI} 162 {MtInt} * {InfAI} 1,2,3,4,A {FormI} < bigọ́nza>
      S< bigọ́nze> P {DescI} Recipiente di legno per le castagne secche, il grano o la farina da
      conservare. {CGram} SO {MtEtn} e
5)    1560162* {PuntI}156 {TpInI} O {DomI} 162 {MtInt} * {InfAI} 1,2,3,4,5,6 {FormI} < bigọ́nza>
      {DescI} Recipiente dove si mettono le castagne per sbucciarle con il < mánfano>. {CGram}
      SO {Preed} Ex nota dom. 141 a. {MtEtn} e
6)    1940162* {PuntI}194 {TpInI} O {DomI} 162 {MtInt} * {InfAI} Z {FormI} < píla> {DescI}
      Recipiente per mettere a bagno le castagne. {CGram} SO {MtEtn} e
```

Figure 11. Result of a query asking for the co-occurrence of the two words within the definition text

Figure 11 illustrates the result of a query asking for the co-occurrence of the two words *recipiente* (container) and *castagne* (chestnuts) within the definition text. The search for the co-occurrence of words within the definition text is very useful for the extraction of semantic information and, as shown in the field of computational lexicography, it yields promising results.

5.2.2. *Looking for the occurrence of one out of a set of variants*

In some cases, the user looks for the presence of one out of several alternative requests. These alternative requests can be for instance different derivations of the same stem, or different ways to express the same concept (e.g. the concept of container in Italian can be expressed through different terms such as *recipiente*, *contenitore*, etc.). In such a case, the query is formulated as a disjunction of different conditions, schematically, “cond$_1$ OR cond$_2$ … OR cond$_x$”.

Going back to the <ski̯aččáta> example, the list below reports all word formations attested in the Tuscan area involving the stem <ski̯ačč>:

1. <ski̯ačča>
2. <ski̯aččáta>
3. <ski̯aččatę́lla>
4. <ski̯aččatína>
5. <ski̯aččẹ́tta>
6. <ski̯aččína>

As it can be noticed, <ski̯aččáta> is one out of a set of different word formations containing different suffixes (-ata, -at+ella, -at+ina, -etta, -ina); also the stem form, i.e. <ski̯ačča>, is attested. The user interested in all these formations, can formulate a complex query asking for the occurrence of one out of these six forms in the ALT Lexical Archive. The result will include occurrences of all of them.

More complex queries can also be formulated by combining different logical operators within the same query expression, as in “(cond$_1$ OR cond$_2$ OR cond$_3$) AND cond$_4$”.

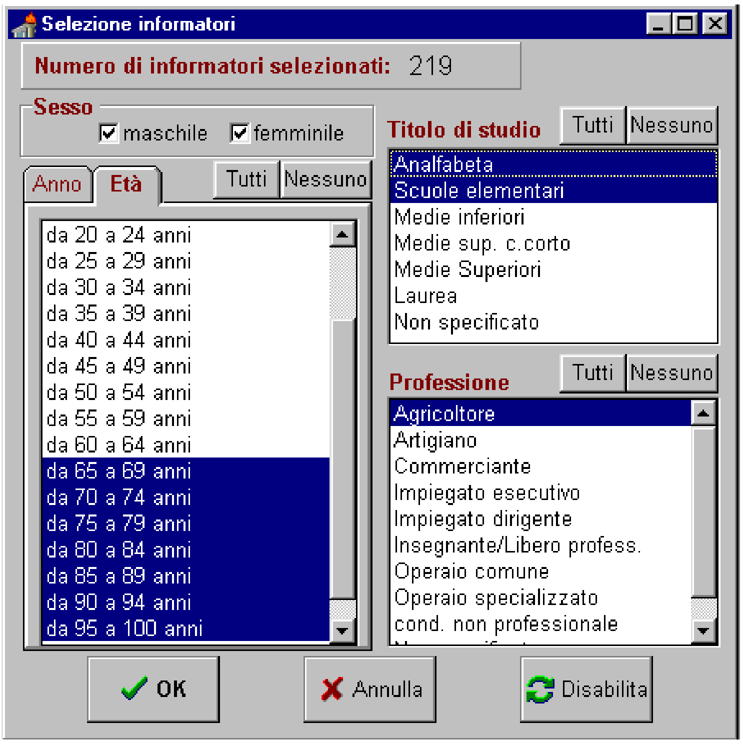
5.3. *Filtering query results*

The results of all queries illustrated so far can also be filtered with respect to a number of both extralinguistic and linguistic factors; among the most relevant ones, there are:

a) socio-economic and/or cultural background of informant(s);
b) geographic subareas either administratively or socio-economically defined;
c) relevance with respect to a given semantic domain;
d) socio-linguistic status and other features of the lexical entry.

Different factors can be combined together to form complex filters, e.g. answers by analphabet old informants of a given geographic area can be easily extracted.

Any of the queries above can be associated with a filter which has to be set before formulating the query. For instance, the user can preventively design the profile of the type of informant (s)he wants to study on the basis of a number of different parameters concerning age, sex, education and professional status; figure 12 below shows the available options for setting up this extra-linguistic filter:



Selezione informatori

Numero di informatori selezionati: 219

Sesso

☑ maschile ☑ femminile

Anno | Età

da 20 a 24 anni
da 25 a 29 anni
da 30 a 34 anni
da 35 a 39 anni
da 40 a 44 anni
da 45 a 49 anni
da 50 a 54 anni
da 55 a 59 anni
da 60 a 64 anni
da 65 a 69 anni
da 70 a 74 anni
da 75 a 79 anni
da 80 a 84 anni
da 85 a 89 anni
da 90 a 94 anni
da 95 a 100 anni

Titolo di studio

Analfabeta
Scuole elementari
Medie inferiori
Medie sup. c.corto
Medie Superiori
Laurea
Non specificato

Professione

Agricoltore
Artigiano
Commerciante
Impiegato esecutivo
Impiegato dirigente
Insegnante/Libero profess.
Operaio comune
Operaio specializzato
cond. non professionale

Figure 12. Designing the informant profile

Similarly, queries can be restricted to geographic subareas; this is done through the preventive selection of the geographic subarea or of the single localities object of the study.

Filters can also be defined on the basis of linguistic factors. For example, many words show domain-specific meanings; although they can be seen as correlated meanings (even when they belong to different semantic domains), one could prefer to isolate the meanings pertaining to a given semantic domain only. Consider for instance the word <fálda> which shows different meanings in the Tuscan area: in the weather domain, it designates snowflakes, especially big ones; it is also used to denote a sheet of pasta dough out of which *tagliatelle* or *macaroni* are made; moreover, it also means a lock of hair. If the user is studying weather terminology (s)he can exclude noisy data by projecting his/her query (<fálda> in the case at hand) onto weather terms only. Many other linguistic filters can be designed by enforcing constraints on either attribute(s) of ALT entries.

6. COMPUTER-GENERATED DIALECTAL MAPS

We started this paper claiming that dialectal data, when organised in the form of a lexical database, are susceptible of different exploitations; we showed some examples in the previous sections. Yet, projection of query results onto maps remains a central task. As a consequence, DBT-ALT also supports the automatic production of dialectal maps starting from the results of each query. All localities where a positive answer to the query is found are marked in the map, as exemplified in the figure below which represents the projection onto the map of the result of the query by form <ski̯ačča>:

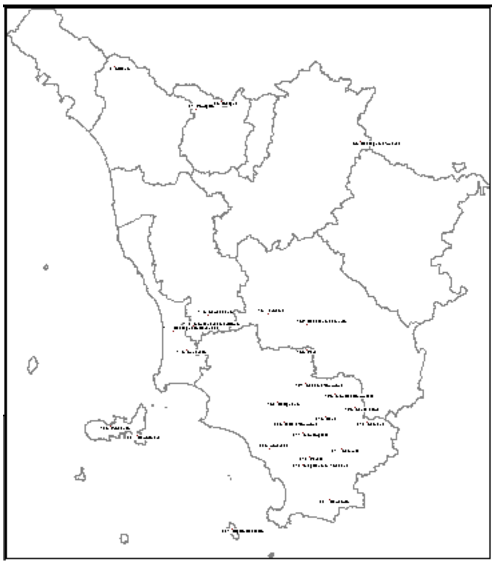


Figure 13. A computer-generated dialectal map

As it can be noticed, this word form is mainly concentrated in the southern part of Tuscany; it also appears to be sporadically known in the north of Tuscany, in the mountains, to testify frequent contacts and exchanges between northern and southern parts of Tuscany due to seasonal migrations for sheep farming.

Further developments of DBT-ALT will include multi-layered maps, combining the results of different queries (labelled through different symbols) or projecting the results of a query onto different backgrounds (e.g. a physical map of Tuscany). In this way, dialectal maps can become a useful and flexible research tool.

7. FINAL REMARKS

In this paper, we illustrated DBT-ALT, a fast, flexible and powerful tool for storing and querying both geolinguistic and sociolinguistic data collected for ALT. We showed how it supports complex queries, taking into account a wide range of parameters, which are interactively defined by the user on the basis of his/her research interests. Intelligent access procedures are also provided as far as phonetic variants are concerned. Last but not least, query results can be projected onto computer-generated maps.

REFERENCES

AGOSTINIANI L., MONTEMAGNI S., PAOLI M., PICCHI E., POGGI SALANI T., *La costruzione di un sistema integrato per il trattamento dei dati dell'Atlante Lessicale Toscano: esperienze, problemi, prospettive*, in *Proceedings of the Conference 'Atlanti Linguistici Italiani e Romanzi': esperienze a confronto*, Palermo, 3-7 October 1990, Palermo, Centro di Studi Filologici e Linguistici Siciliani, 1992, 357-393.

AGOSTINIANI L., MARINAI E., MONTEMAGNI S., PAOLI M., *Una procedura informatica di accesso intelligente a materiali in trascrizione fonetica: l'esperienza dell'Atlante Lessicale Toscano*, in *Proceedings of the V° Congresso SILFI*, Catania, 15-17 October 1998, in press.

CALZOLARI N., *Detecting Patterns in a Lexical Database*, in *Proceedings of the 10th International Conference on Computational Linguistics*, Stanford, California, 1984, 170-173.

CALZOLARI N., PICCHI E., *Acquisition of Semantic Information from an On-Line Dictionary*, in *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, 1988, 87-92.

CHODOROW M., BYRD R., HEIDORN G., *Extracting semantic hierarchies from a large on-line dictionary*, in *Proceedings of the 23rd Annual Meeting of the ACL*, 1985, 299-304.

GIACOMELLI G., AGOSTINIANI L., BELLUCCI P., GIANNELLI L., MONTEMAGNI S., NESI A., PAOLI M., PICCHI E., POGGI SALANI T. (eds.), *Atlante Lessicale Toscano*, Lexis Progetti Editoriali, Roma, 2000.

GOEBL H., *Dialectometry. A Short Oveview of the Principles and Practice of Quantitative Classification of Linguistic Atlas Data*, in R. KÖHLER, B. RIEGER (eds.), *Contributions to Quantitative Linguistics*, Dordrecht, Boston, London. Kluwer, 1993, 277-315.

GRASSI C., SOBRERO A., TELMON T., *Fondamenti di Dialettologia Italiana*, Roma-Bari, Laterza, 1997.

INOUE F., *Computational Dialectology (1)*, «Area and Culture Studies», 52, 1996a, 67-102.

INOUE F., *Computational Dialectology (2)*, «Area and Culture Studies», 53, 1996b, 115-134.

MARKOWITZ J., AHLSWEDE T., EVENS M., *Semantically significant Patterns in Dictionary Definitions*, in *Proceedings of the Association for Computational Linguistics (ACL) 24th Annual Meeting*, New York, 10-13 June 1986, 112-119.

MONTEMAGNI S., ZAMPOLLI A., *Dialettologia e Informatica*, «Rivista di Dialettologia Italiana», Bologna, CLUEB, XI, 1987, 149-174.

MONTEMAGNI S., PAOLI M., *Dalla parola al bit (e ritorno): percorsi dall'inchiesta sul campo alla banca dati dell'Atlante Lessicale Toscano*, «Quaderni dell'Atlante Lessicale Toscano», Firenze, Olschki Editore, 7/8, 1989-90, 7-52.

MONTEMAGNI S., PAOLI M., *Esplorazioni nel mondo dell'ALT: itinerari alternativi*, in A. CATAGNOTI *et al.* (eds.), *Studi Linguistici offerti a Gabriella Giacomelli dagli amici e dagli allievi*, Padova, UNIPRESS, 1997, 279-300.

MONTEMAGNI S., PAOLI M., PICCHI E., *DBT-ALT. Manuale di Riferimento*, Lexis Progetti Editoriali, Roma, 2000.

MONTEMAGNI S., PICCHI E., *From a Computational Linguistic Atlas to Dialectal Lexical Resources,* in T. FONTENELLE, P. HILIGSMANN, A. MICHIELS, A. MOULIN, S. THEISSEN (eds.), *Proceedings of the Eighth EURALEX International Congress on Lexicography*, University of Liège, Belgium, 1998, 221-230.

PICCHI E., *DBT: a textual Database system*, in «Linguistica Computazionale. Computational Lexicology and Lexicography», VII (1991), 2, 77-105.

PICCHI E., *Lexicographic Workstation*, «ERCIM News European Research Consortium for Informatics and Mathematics», GEIE-ERCIM, Le Chesnay Cedex, France, X1 (1992), 21.

PICCHI E., PETERS C., MARINAI E., *The Pisa Lexicographic Workstation: the Bilingual Components*, in *Proceedings of the Fifth Euralex International Congress*, Tampere University, Finland, 1992, 265-275.