Simonetta MONTEMAGNI & Eugenio PICCHI, Istituto di Linguistica Computazionale, CNR, Pisa

From a Computational Linguistic Atlas to Dialectal Lexical Resources

Abstract

Computers can help dialectologists to make full use of the information they have acquired: the basic dimensions of dialectal research can be enlarged and its possible outcomes can become more sophisticated. In this paper, we show how a dialectal database, DBT-ALT, containing the data collected for the Atlante Lessicale Toscano 'Lexical Atlas of Tuscany' can be used as the starting point for the production of dialectal dictionaries and other kinds of lexicographic resources provided that adequate computational tools are available to carry out the job properly. First, the architecture and functioning of DBT-ALT are described in detail. Second, we show how DBT-ALT access functionalities can be exploited to extract subsets of data which could be converted into independent lexicographic resources through the operation of a Lexicographic Workstation.

Keywords: computational dialectology, dialectal databases, construction of lexical resources

1. Introduction

The current priority in dialectal research is the identification of linguistic areas to be eventually projected onto maps. As a consequence, the use of computers in the field of dialectology has so far mainly concentrated on automatic map drawing (see the survey on Computational Dialectology in Inoue 1996a and 1996b), based either on raw linguistic data or on data preprocessed through statistical methods (as in the case, for instance, of dialectometry, see Goebl 1993). Yet, linguistic maps of either type are not the only possible outcome of dialectal research. Data collected by dialectologists in different areas from different informants are linguistic data in their own right and, as such, they are susceptible of different classifications and organisations which can, in their turn, result in different products among which dictionaries represent a crucial part. 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 dialectal data, thus making full use of the abundance and richness of acquired linguistic information (Montemagni and Zampolli 1987). This paper does not question the importance or the centrality of linguistic maps in dialectal research. Rather, it seeks to show how the computer can be used to exploit to the full collected dialectal data, in particular for what concerns the production of dialectal dictionaries. To our knowledge, this is a relatively unexplored research path in the field of dialectology.

In order to make dialectal data simultaneously exploitable from different perspectives, the preliminary step consists in organising them 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 process is rather time consuming and is excessive if only geared towards

map drawing. By contrast, this encoding 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. In such a case, computational procedures and tools are then needed which enable the semi-automatic creation of new dialectal resources.

In this paper we will concentrate on lexicographic dialectal resources which can be derived from a lexical database of dialectal data. We will first illustrate the starting point, i.e. DBT-ALT, a lexical database which has been constructed for the storage, management and interrogation of dialectal data collected for the Atlante Lessicale Toscano 'Lexical Atlas of Tuscany', henceforth ALT (Giacomelli *et al.*, forthcoming). Secondly, we will show how from this lexical database new dialectal resources can be derived through the operation of a Lexicographic Workstation (Picchi 1992, Picchi *et al.*, 1992).

2. DBT-ALT, a lexical database for ALT data

2.1. Background

DBT-ALT is a modular system for the storage, management and interrogation of the linguistic data of the Atlante Lessicale Toscano (Agostiniani *et al.*, 1992; Picchi *et al.*, 1997), a specially designed linguistic atlas in which lexical data have both a diatopic and diastratic characterisation.

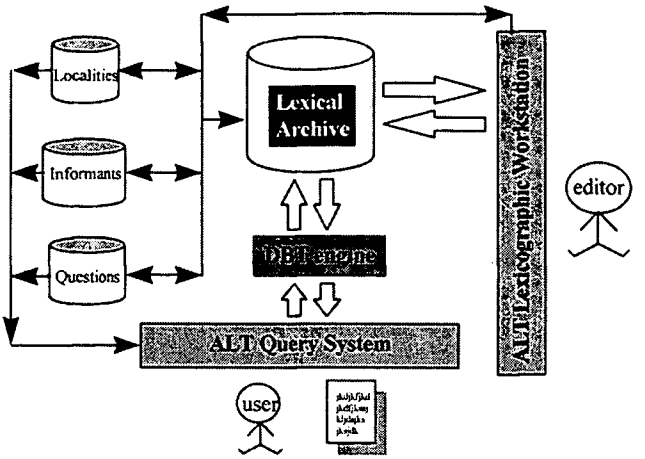
ALT lexical data bank contains the results of interviews carried out in 224 localities of Tuscany, with 2082 informants (selected with respect to a number of parameters ranging from age, socio-economic status to education and culture), on the basis of a questionnaire of 745 items: more than 350.000 responses were collected which were integrated with additional material emerged during the 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 E. 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, forthcoming), 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, i.e. the management of structured linguistic data (as in the case of dictionaries) and the processing of non-Latin alphabets. DBT-ALT is a version which includes these functionalities together with new ones in order to meet the combined needs of geolinguistic and sociolinguistic research as emerging from ALT.

2.2. Overall Architecture

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



The Lexical Archive (LA), which contains all linguistic data collected through the interviews, is linked to a system of subsidiary archives 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 have been constructed using the Lexicographic Workstation. Explorations in LA are dealt with by the Query System (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 done on the basis of information contained in the subsidiary archives (see the links between them and LA on the one hand, and between them and QS on the other hand).

2.3. DBT-ALT Lexical Entry Model

Having sketched the macro-structure of the system, let us consider now the micro-structure of data, i.e. the model according to which lexical data have been encoded. An entry model was needed sophisticated enough to represent the richness of collected linguistic information on the one hand and to enable complex information retrieval on the other hand. In order to fulfil both requirements a rather complex and articulated structure was needed: ALT entries present themselves as bundles of attribute-value pairs each of which specifies a different kind of information (for a detailed description of ALT entries see Montemagni and Paoli 1989-90). For each entry, the main coordinates LOCALITY, INFORMANT(s) and QUESTION are always specified. ALT Lexical Archive contains different entry types:

- canonical responses to questionnaire items;
- lexical items which emerged during 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'/fieldworkers' remarks on the status of words (e.g. usage, traditionality, register).

2.3.1. Encoding of phonetically transcribed data

Data in the lexical archive are either phonetically transcribed or represented according to standard Italian orthography; in this respect, the lexical archive can be seen as a kind of "bilingual" text archive. In what follows we will focus on the representation of phonetically transcribed data only.

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 (see 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 both compositional and atomic representations which, depending on the task, are automatically converted into each other.

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. This type of representation is also convenient for both sorting and retrieval phases: in fact, if basic signs only are considered, it is possible to generalise over phonetic variants. 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.

2.4. 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 the corpus of dialectal data and thus navigate through it on the basis of his/her research interests (for a detailed description of the functionalities of the DBT-ALT query system the interested reader is referred to Picchi *et al.*, forthcoming).

Lexical data can be accessed and retrieved on the basis of a wide range of parameters:

- questionnaire item to which they directly or indirectly relate;
- locality in which they were witnessed;
- semantic keywords clustering lexical items into thematically coherent groupings;
- meaning components as inferable from the definition text;
- phonetic realisation.

Each of these parameters corresponds to a specific attribute 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.

These individual parameters can be variously combined to form complex queries in which items which are looked for are linked by AND, OR and NOT operators, as in the case of i) the cooccurrence of different information types within the same record, or ii) the occurrence of one out of a set of variants.

Query results can also be filtered with respect to:

- socio-economic and/or cultural background of informant(s);
- geographic subareas either administratively or socio-economically defined;
- relevance with respect to a given semantic domain;
- socio-linguistic status of words.

With such a sophisticated query system, much information which remains normally "hidden" in printed dialectal resources (either linguistic atlases or dialectal dictionaries) can be easily retrieved.

2.4.1. Retrieving phonetically transcribed data

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. In fact, 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. 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. Within DBT-ALT, two different abstraction levels have been devised for retrieval purposes:

- level 1 which operates on basic signs only and ignores diacritic signs (e.g. at this level the distinction between the voiceless alveolar plosive /t/ and the corresponding fricative /ŧ/ is neutralised);
- level 2 which permits more powerful generalisations by clustering together different basic signs or combinations of them (e.g. /k̲j̲/, /t̲j̲/, /c͡ç/, /j/ and /t'/).

Depending on the user needs, either of the two 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.

3. The Lexicographic Workstation: from DBT-ALT to lexicographic dialectal resources

Within the general framework of the PI-System, a lexicographic workstation was implemented to assist the lexicographer in the various activities involved in the creation and revision of dictionaries (Picchi 1992, Picchi *et al.*, 1992). Underlying the lexicographic workstation there are different components, namely i) a full text retrieval system to query and analyse all kinds of texts and textual corpora and ii) a lexical database system to query structured data. The lexicographer can use these two systems to interrogate on-line text

archives and electronic dictionaries and retrieve and extract reference and citation material to be included in the lexical resource under development.

The core module of the workstation is a procedure for on-line dictionary editing which includes functions for: windowing into and copying data from the dictionary and text archives; compilation of the new entry on the basis of extracted data; formatting of the entry for printing. This module is also integrated with a structured indexing procedure that can be used to query the dictionary in compilation in order to check the regularity and consistency of the input. In this way, the newly developed dictionary, structured as an on-line database, becomes in its turn potential input for new lexical resources. This circular way of proceeding in the compilation of new lexical resources guarantees, on the one hand, a continuous checking of the previously developed ones (which can thus be revised and updated), and, on the other hand, the full exploitation of work done previously.

As already mentioned in section 2.2 above, the lexicographic workstation was used to create DBT-ALT archives; in that case, only the on-line editor was used since there were no underlying components from which data could be extracted. However, we are now in a position to use the same system operating on the DBT-ALT lexical database to extract data subsets which, after editing, revision and reorganisation, could be converted into new dialectal resources. For example:

- dialectal dictionaries of geographic subareas or even of a given locality can be constructed starting from the results of the interviews carried out in that area. In the figure below, an excerpt of the data collected in Pitigliano (Grosseto, 218), a locality in the south of Tuscany, is reported.

{Dom}004 {Forma} <ʒʒíʒʒera>
{Dom}004 {Forma} <ája>
{Dom}004 {Forma} <fréddo>
{Dom}004 {Forma} <ǧáčču>
{Dom}004 {Forma} <fréddo rígido>
{Dom}004 {Forma} <fréddo akúto>
{Dom}004 {Forma} <ǧélu>
{Dom}004 {Forma} <ʒʒíʒʒola>
{Dom}005 {Forma} <appágu>
{Dom}005 {Forma} <appágo>
{Dom}005 {Forma} <alamérja>
{Dom}005 {Forma} <all ómbra>
{Dom}006 {Forma} <alamérja>
{Dom}006 {Forma} <appágo>
{Dom}006 {Forma} <all ómbra>
{Dom}007 {Forma} <al appavénta>
{Dom}007 {Forma} <appavénta>
{Dom}007 {Forma} <appaentátu>
{Dom}007 {Forma} <appaventáto>
{Dom}007 {Forma} <riparáto>

Data are ordered here according to the questionnaire item to which they relate. The construction of a dialectal dictionary of Pitigliano will start from these data and will require reordering, revision and integration of collected data.

- dialectal dictionaries corresponding to a given socio-culturally defined linguistic variety can also be produced by enforcing the constraint that only lexical items attested by informants which have a specific socio-cultural characterisation will be included in the starting core. These socio-cultural constraints on the informants can be combined with geographic ones. Consider below the data attested in the same locality as above, namely Pitigliano, by old analphabet informants (for the same questionnaire items as above):

```
{Dom}004 {Forma} <strina>[3]
{Dom}004 {Forma} <ąąiąąa>[1]
{Dom}004 {Forma} <fręddo>[1]
{Dom}004 {Forma} <fręddo rigido>[2]
{Dom}004 {Forma} <fręddo akúto>[1]
{Dom}005 {Forma} <appágo>[2]
{Dom}005 {Forma} <ala mérja>[1]
{Dom}005 {Forma} <all ǫmbra>[2]
{Dom}006 {Forma} <ala mérja>[4]
{Dom}007 {Forma} <al appavẹnta>[2]
{Dom}007 {Forma} <appavẹnta>[2]
{Dom}007 {Forma} <riparáto>[2]
```

- a sort of dialectal thesaurus where semantically similar words from different or specific areas are grouped together can be created starting from a selection of lexical items belonging to a given semantic field in a specific geographic area. In the figure below, an excerpt of data relating to chestnuts (cultivation, alimentary traditions and all customs and beliefs relating to them) is given as an example:

```
{Punto}001 {Dom}305 {Forma} <mnūldiñ>
{Punto}001 {Dom}307 {Forma} <*bručáte>
{Punto}001 {Dom}307 {Forma} <mundina>S<mundine>P
{Punto}001 {Dom}307 {Forma} <mundina>S<mundine>P
{Punto}001 {Dom}308 {Forma} <balęt>S<balęti>P
{Punto}001 {Dom}310 {Forma} <gūśųñ kǫti>P
{Punto}001 {Dom}310 {Forma} <gūśųñ lōśá>P
{Punto}002 {Dom}087 {Forma} <bǫśko d kaśtáñi>
{Punto}002 {Dom}087 {Forma} <i kaśtáñi>
{Punto}002 {Dom}090 {Forma} <tṛáčo>
{Punto}002 {Dom}091 {Forma} <níča>
{Punto}002 {Dom}092 {Forma} <plęgra>
{Punto}002 {Dom}094 {Forma} <gráda>
{Punto}002 {Dom}138 {Forma} <púla>
{Punto}002 {Dom}303 {Forma} <farina dǫča>
```

This selection is restricted to localities in the mountains of the Tosco-Emilian Appennines, where chestnuts represent a traditional culture.

- last but not least, a linguistic atlas in the form of lists can be created, where all canonical responses are ordered according to the questionnaire items through which they were elicited. The figure below exemplifies the answers collected in the different localities of Tuscany through question 303, aimed at collecting denominations of the chestnut flour:

{Punto}001 {Dom}303 {Forma} <farina d kaśtáńa->
{Punto}001 {Dom}303 {Forma} <farina dǫ̨́sa>
{Punto}001 {Dom}303 {Forma} <farina di kaśtáńe>
{Punto}002 {Dom}303 {Forma} <farina dǫ̨́ča>
{Punto}003 {Dom}303 {Forma} <farina dǫ̨́ča>,<farina dǫ̨́za>
{Punto}004 {Dom}303 {Forma} <farina dǫ̨́ča>
{Punto}004 {Dom}303 {Forma} <farina d kastáńa>
{Punto}004 {Dom}303 {Forma} <farina d kastańáče>
{Punto}005 {Dom}303 {Forma} <farina d kastińá>
{Punto}006 {Dom}303 {Forma} <farina dǫ̨́lča>
{Punto}006 {Dom}303 {Forma} <farina da čáń>
{Punto}007 {Dom}303 {Forma} <farina da kaśtáńńa>
{Punto}007 {Dom}303 {Forma} <farina dǫ̨́lče>
{Punto}007 {Dom}303 {Forma} <farina di kastáńńe>
{Punto}008 {Dom}303 {Forma} <farina də kastińá>
{Punto}009 {Dom}303 {Forma} <farina də- čáń>
{Punto}010 {Dom}303 {Forma} <la farina də káśtińńáčča>
{Punto}011 {Dom}303 {Forma} <farina di čáččo>
{Punto}011 {Dom}303 {Forma} <farina di ǧáččo>

This output could be passed, after revision, on to automatic map drawing procedures and the resulting maps could be stored back in the linguistic atlas database through the OLE technology.

The examples above provide only some of possible lexicographic resources which can be derived from a dialectal database such as DBT-ALT. In fact, by exploiting the potentialities of the DBT-ALT query system, very specific LA subsets can be extracted and reorganised in the form of independent lexicographic resources. These newly developed dialectal resources can then be used in their turn as input for the development of further ones.

The Lexicographic Workstation can operate on different archives, either textual or structured ones. Depending on the goals of the lexicographer, evidence gathered through ALT interviews can also be combined with other kinds of linguistic data (either in the form of texts or of structured archives) to derive new resources. For instance, DBT-ALT could be used in the compilation of an Italian dictionary to provide evidence concerning the diffusion within Tuscany of Italian words or to find evidence supporting the Tuscan origin of words appearing in the dictionary; or, in literary studies, it could be used to discriminate linguistic features which are to be ascribed to the author idiolect from those which are part of his/her geographic linguistic variety.

4. Final remarks

It is a widely acknowledged fact that computers can help dialectologists to make full use of the information they have so laboriously and painstakingly acquired. Yet, this help cannot be restricted to the task of map drawing. In fact, thanks to computers the basic dimensions of dialectal research are enlarging and the variety and typology of possible outcomes is becoming more and more sophisticated. In this paper, we focussed on how a dialectal lexical database - DBT-ALT - can be used as the starting point for the production of dialectal dictionaries and other kinds of lexicographic resources provided that adequate computational tools, i.e. the Lexicographic Workstation, are available to carry out the job properly. First, DBT-ALT has been described as a fast, flexible and powerful tool for storing and querying both geolinguistic and sociolinguistic data. Its complex access functionalities, taking into account a wide range of parameters which are interactively defined by the user on the basis of his/her research interests, have been illustrated. The same access functionalities, when used through the Lexicographic Workstation, can be exploited to extract from the lexical database subsets of data which could be reorganised to form independent lexicographic dialectal resources.

5. References

L. Agostiniani, S. Montemagni, M. Paoli, E. Picchi, T. Poggi Salani, 1992, "La costruzione di un sistema integrato per il trattamento dei dati dell'Atlante Lessicale Toscano: esperienze, problemi, prospettive", in *Proceedings of the Conference of the Centro di Studi Filologici e Linguistici Siciliani on 'Atlanti Linguistici Italiani e Romanzi: esperienze a confronto'*, Palermo, 3-7 Ottobre 1990, Palermo, pp.357-393.

G. Giacomelli, L. Agostiniani, P. Bellucci, L.Giannelli, S. Montemagni, A. Nesi, M. Paoli, E. Picchi, T. Poggi Salani, (eds.), forthcoming, *Atlante Lessicale Toscano*, Regione Toscana, Accademia di Scienze e Lettere "La Colombaria".

H. Goebl, 1993, "Dialectometry. A Short Overview of the Principles and Practice of Quantitative Classification of Linguistic Atlas Data", in R. Köhler and B. Rieger (eds.), *Contributions to Quantitative Linguistics*, Dordrecht, Boston, London: Kluwer, pp. 277-315.

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

F. Inoue, 1996a, "Computational Dialectology (1)", *Area and Culture Studies*, vol. 52, pp. 67-102.

F. Inoue, 1996b, "Computational Dialectology (2)", *Area and Culture Studies*, vol. 53, pp. 115-134.

S. Montemagni, A. Zampolli, 1987, "Dialettologia e Informatica", *Rivista di Dialettologia Italiana*, XI, CLUEB, Bologna, pp.149-174.

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

E. Picchi, 1991, "DBT: a textual Database system", in *Linguistica Computazionale. Computational Lexicology and Lexicography*, VII, 2, Pisa, Giardini Editore, pp. 77-105.

E. Picchi, 1992, "Lexicographic Workstation", *ERCIM News European Research Consortium for Informatics and Mathematics*, n. 11, October 1992, GEIE-ERCIM, Le Chesnay Cedex, France, p. 21.

E. Picchi, forthcoming, *Linguistica Computazionale. Analisi Testuale e Lessicale*, Bulzoni Editore.

E. Picchi, C. Peters, E. Marinai, 1992, "The Pisa Lexicographic Workstation: the Bilingual Components", in *Proceedings of the Fifth Euralex International Congress*, Tampere University, Finland, pp.265-275.

E. Picchi, S. Montemagni, L. Biagini, forthcoming, "DBT-ALT: A System for Storing and Querying the Data of the Atlante Lessicale Toscano (ALT)", to appear in the *Proceedings of the 2nd International Congress of Dialectologists and Geolinguists*, 28/7-1/8 1997, Amsterdam.