DBT-ALT: a System for Storying and Querying the Data of the Atlante Lessicale Toscano (ALT)

Eugenio Picchi
Simonetta Montemagni
Lisa Biagini

Istituto di Linguistica Computazionale - CNR
Pisa (Italy)

Atlante Lessicale Toscano

- ALT is a specially designed linguistic atlas in which lexical data have both a diatopic and diastratic characterisation
- ALT Data Bank contains the results of interviews carried out in 224 localities of Tuscany, with 2082 informants on the basis of a questionnaire of 745 items
- We expected at least

2.082 * 745 = 1.551.090 individual responses

equivalent to

224 * 745 = 166.880 areal responses

- We collected more than 350.000 areal responses which were integrated with additional material emerged during the interviews (about 30.000 dialectal items)

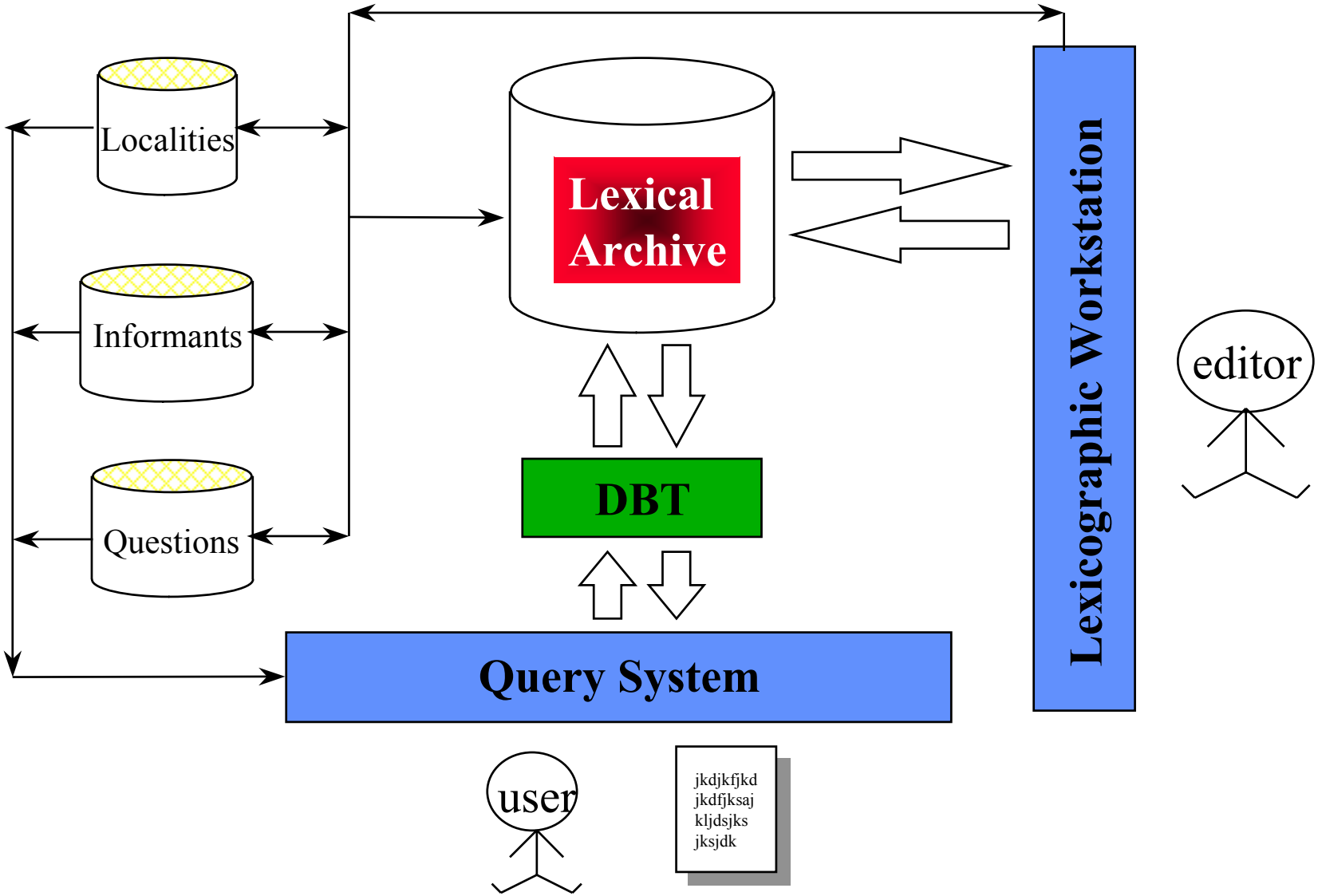
DBT

- DBT is a textual database system for the storage, management and interrogation of large text archives whose basic functions include:
  - a query system
  - generation of indices (ordered either alphabetically or by frequency)
  - generation of concordances
- DBT is the core component of the PI-System (Picchi), a set of procedures specifically designed to tackle specific problems in the area of computational linguistics and lexicography. Among them, of specific interest for ALT,
  - processing of non-Latin alphabets
  - management of structured data

DBT-ALT

- DBT-ALT is a specialised version of DBT tailored to meet the combined needs of geolinguistic and sociolinguistic research as emerging from ALT
- DBT-ALT is designed to handle structured linguistic data either phonetically trascribed or represented according to standard Italian orthography
- DBT-ALT handles an integrated system of subsidiary archives containing information about
  - localities which have been investigated
  - informants who have been interviewed
  - the questionnaire
- DBT-ALT supports the automatic production of dialectal maps

DBT-ALT: overall architecture



ALT Entry Model (1)

- An entry model was needed sophisticated enough
  - ☞to represent the richness of collected linguistic information
  - ☞to enable complex information retrieval
- ALT entries present themselves as bundles of attribute-value pairs each of which specifies a different kind of information
- 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 with the questionnaire
  - typical contexts of use (e.g. phraseology, proverbs)
- All entries may also contain informants'/fieldworkers' remarks on the status of words (e.g. usage, traditionality, register)

ALT Entry Model (2)

{Punto}026
{TpInc}O
{Dom}094
{Inf.A}1
{Forma} <sẹ́kkatǫ̣o>
{CGram} SO

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

{Punto}026
{TpInc}O
{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'.

Encoding phonetically transcribed data (1)

- The phonetic alphabet used in the project fieldwork is a geographically specialised version of the Carta dei Dialetti Italiani (CDI) transcription system
- In order to ensure a proper treatment of phonetically transcribed data, a complex encoding schema was designed to fulfil the specific requirements of different tasks:
  - editing
  - sorting
  - retrieval
  - on-screen display
  - printing

Encoding phonetically transcribed data (2)

- 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
  - are particularly suitable for editing since all different phonetic symbols can be encoded by means of a restricted number of codes (36 basic signs and 9 diacritics)
  - permit to generalise over phonetic variants during both sorting and retrieval phases
- **Atomic representations:**
  - show a 1:1 correspondence between ALT phonetic symbols and computer codes; used for on-screen display and printing

DBT-ALT Query System: main functionalities (1)

- The DBT-ALT query system provides dynamic 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
- 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
  - ☑ semantic keywords clustering questionnaire items into thematically coherent groupings
  - ☑ locality in which they were witnessed
  - ☑ phonetic realisation
  - ☑ meaning components as inferable from the definition text

Retrieving phonetically transcribed data

- Computers should facilitate access to data but narrowness of phonetic transcription may constitute a major difficulty
- Compositional representations permit the user to overcome this difficulty by abstracting away from specific phonetic realisations
- Two different abstraction levels have been devised for retrieval purposes:
  - **level 1** operates on basic signs only and ignores diacritic signs (e.g. /t/ and /t/)
  - **level 2** clusters together different basic signs or combinations of them (e.g. /ki/, /ti/, /c/, /j/ and /t'/)

DBT-ALT Query System: main functionalities (1)

- The DBT-ALT query system provides dynamic 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
- 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
  - ☒ semantic keywords clustering questionnaire items into thematically coherent groupings
  - ☒ locality in which they were witnessed
  - ☒ phonetic realisation
  - ☑ meaning components as inferable from the definition text

DBT-ALT Query System: main functionalities (2)

- These parameters can be variously combined to form complex queries looking for:
  - cooccurrence of different information types within the same record
  - occurrence of one out of a set of variants
- Query results can 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

Computer-generated dialectal maps

- 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 was found are marked in the map
- Symbolisation conveys information about the frequency of occurrence of the response(s) within the informants' group
- Multi-layered maps will be possible
  - combining the results of different queries
  - projecting the results of a query onto different backgrounds
- In this way, dialectal maps become a useful and flexible research intrument

Conclusions

- DBT-ALT is a fast, flexible and powerful tool for storing and querying both geolinguistic and sociolinguistic data
- 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
- Query results can be projected onto computer-generated maps
- “Intelligent” access procedures are provided as far as phonetic variants are concerned

DBT-ALT and multidimensional dialectal data