DataBaseTestuale AtlanteLessicaleToscano

Le Banche Dati Lessicali nella ricerca dialettale e geolinguistica: l'esperienza di DBT-ALT

Simonetta Montemagni, Eugenio Picchi

Istituto di Linguistica Computazionale, CNR - Pisa

Dialettologia Computazionale e Banche Dati Dialettali

📖Linguistic geography, or dialect geography, has been said to be atomistic in that it is mainly concerned with the interpretation of geographical distribution of individual words. [...] Computational dialectology can change this situation. Computational techniques can process groups of words at a time and discover general tendencies. Thus multivariate analysis by computer enables linguistic geography to depart from "atomism" and to achieve a "synthesis". (Fumio INOUE, *Computational Dialectology*, 1996)

La strutturazione e codifica dei materiali di un atlante linguistico in banca dati con la relativa messa a punto delle procedure di accesso ed interrogazione costituiscono questioni spesso trascurate o semplicemente date per scontate.

Procedure informatiche e rendimento scientifico del dato dialettale

In che misura la strutturazione in banca dati contribuisce a migliorare il rendimento scientifico del dato dialettale raccolto?

- **velocità** e **precisione** nel recupero delle informazioni
- nuove possibilità di ricerca e di analisi di cui i materiali dialettali si rendono suscettibili, in particolare per quanto riguarda le **chiavi di accesso** al corpus dei dati la cui tipologia va ben al di là di quelle canoniche incentrate sulla domanda, il punto di inchiesta, gli informatori.

Le banche dati:

- forniscono un utile strumento di supporto alla ricerca dialettale di impianto tradizionale
- sono il punto di partenza per la creazione di visioni "sintetiche" della variazione linguistica

Dato vs Informazione

Qualsiasi dato non ha valore di per sé ma solo se inquadrato in un contesto; attraverso il confronto con altri dati si carica di significato e diventa informazione.

In un atlante linguistico con spessore sociolinguistico per ogni dato sono attivati tre tipi di legami canonici:

- ☞ quello con le risposte alla stessa **domanda**;
- ☞ quello con i materiali reperiti in risposta all'intero questionario nello stesso **punto di inchiesta**;
- ☞ quello con le risposte all'intero questionario fornite dallo stesso **informatore**.

Quindi, ogni dato è parte di tre quadri diversi, che sono rispettivamente:

- ☞ la **carta linguistica**;
- ☞ il repertorio lessicale - o **dizionario** - di una parlata locale;
- ☞ l'**idioletto** di un informatore.

Il carico informativo del dato dialettale strutturato in Banca Dati

- Il sistema di rapporti in cui ogni dato può essere situato non si esaurisce ai quadri canonici
- Lo stesso dato si può collocare in molti altri quadri con "soggetti" diversi, che variano a seconda del tipo di connessione tra i dati
- ad esempio, un dato dialettale si può legare ad altri dati per:
  - **similarità semantica**, che può collocarsi al livello della superclasse di appartenenza o di aspetti specifici del significato;
  - **similarità formale**, fonetica e/o morfologica;
  - la sua **classificazione rispetto ad un registro** di comunicazione formale o informale.
- Ma per attivare questi quadri diversi, ogni dato deve essere fornito degli agganci con dati della stessa specie mediante una codifica ed una struttura interna appropriate.

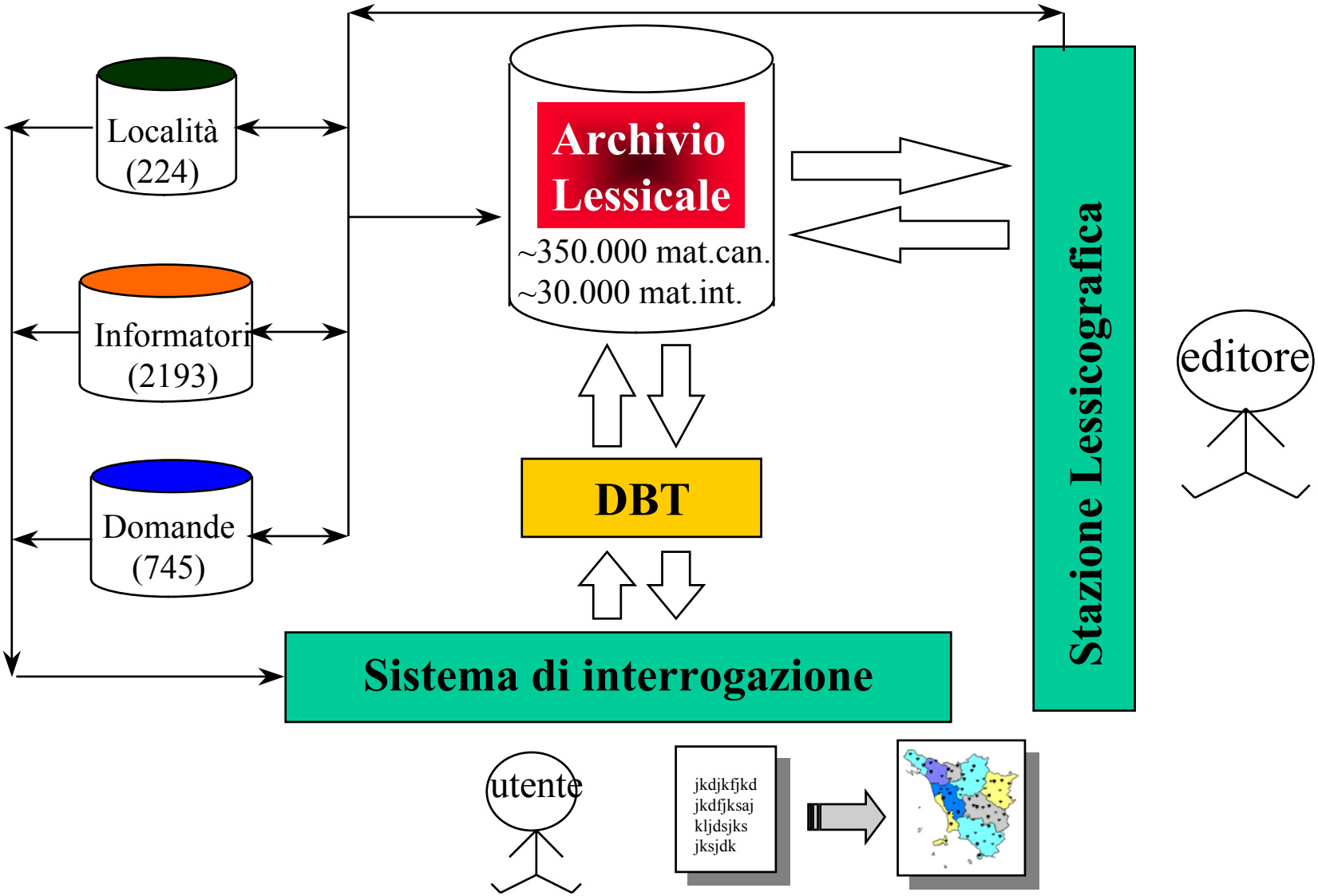
DBT-ALT: una Banca Dati per la gestione dei materiali dell'Atlante Lessicale Toscano (1)

- DBT-ALT è una versione specializzata del DBT (Eugenio Picchi, ILC-CNR), un sistema di database testuale per la memorizzazione, gestione ed interrogazione di grandi archivi di testi
- DBT è il componente centrale del ***PiSystem***, un insieme di procedure messe a punto per rispondere alle specifiche esigenze della linguistica e lessicografia computazionali. Tra queste, hanno costituito il punto di partenza di DBT-ALT i moduli per la gestione di:
  - alfabeti non latini
  - dati strutturati (ad es. database lessicali)

DBT-ALT: una Banca Dati per la gestione dei materiali dell'Atlante Lessicale Toscano (2)

- DBT-ALT gestisce una varia tipologia di dati linguistici strutturati, che contengono rappresentazioni sia in **trascrizione fonetica** sia in **ortografia italiana**
- Nella progettazione e sviluppo di DBT-ALT, si è inoltre dovuto far fronte alle specifiche esigenze della ricerca **geolinguistica** e **sociolinguistica** poste dall'ALT:
  - ☞DBT-ALT gestisce un sistema integrato di archivi sussidiari contenenti informazioni riguardo a:
    - le località indagate
    - gli informatori intervistati
    - il questionario di raccolta
  - ☞DBT-ALT offre anche la possibilità di proiettare su carta i risultati di una ricerca

DBT-ALT: architettura del sistema



Struttura e codifica del dato: il modello dell'entrata

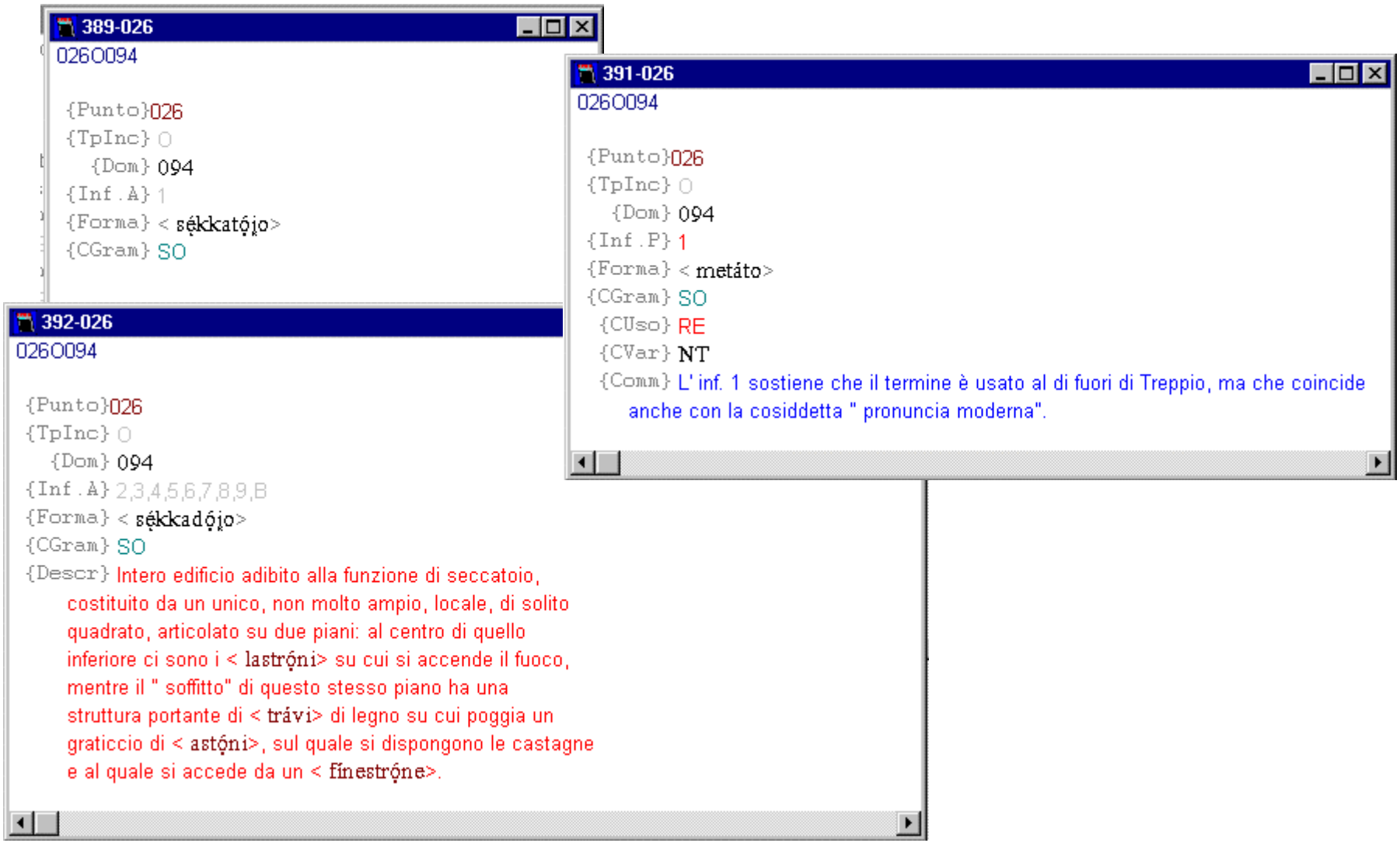
Necessità di un modello di entrata sofisticato al fine di:

- rappresentare la ricchezza dei dati raccolti nelle loro molteplici sfaccettature
- creare i presupposti per procedure di recupero di informazione operanti su molteplici dimensioni

- Le schede della BD-ALT si presentano come insiemi di tratti espressi in termini di coppie attributo-valore ciascuno dei quali convoglia uno specifico tipo di informazione
- Ogni attestazione riceve una caratterizzazione rispetto alle coordinate LOCALITA', INFORMATORI, DOMANDA
- Diverse configurazioni di tratti corrispondono a diversi tipi di attestazione:
  - risposte canoniche a domande del questionario;
  - materiali integrativi;
  - tipici contesti di uso delle attestazioni lessicali raccolte;
  - descrizione di usi e costumi o altro ancora legati ai materiali linguistici raccolti.

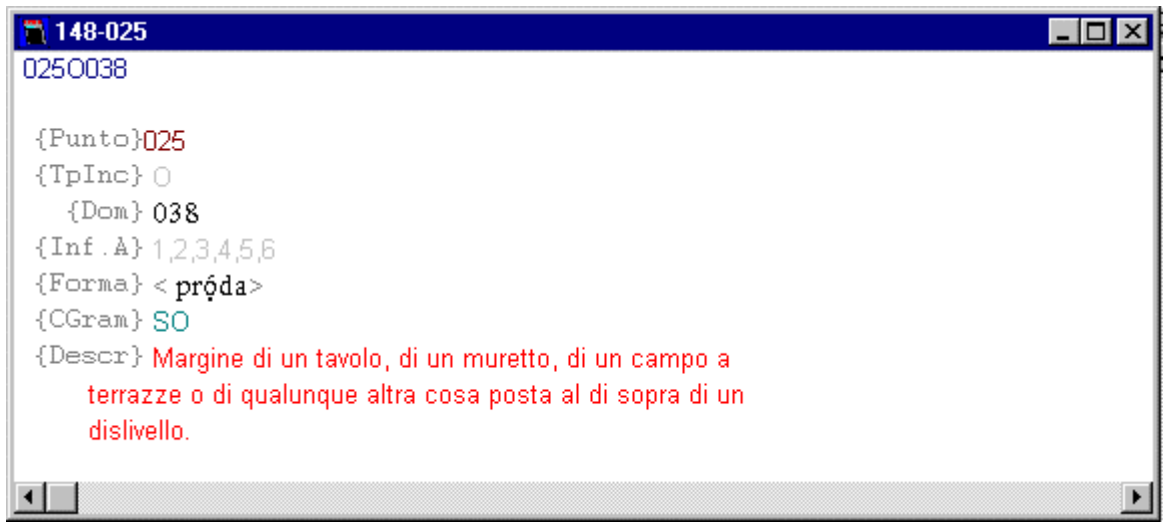
Tipologia di schede della BD-ALT

- Risposte a domande onomasiologiche



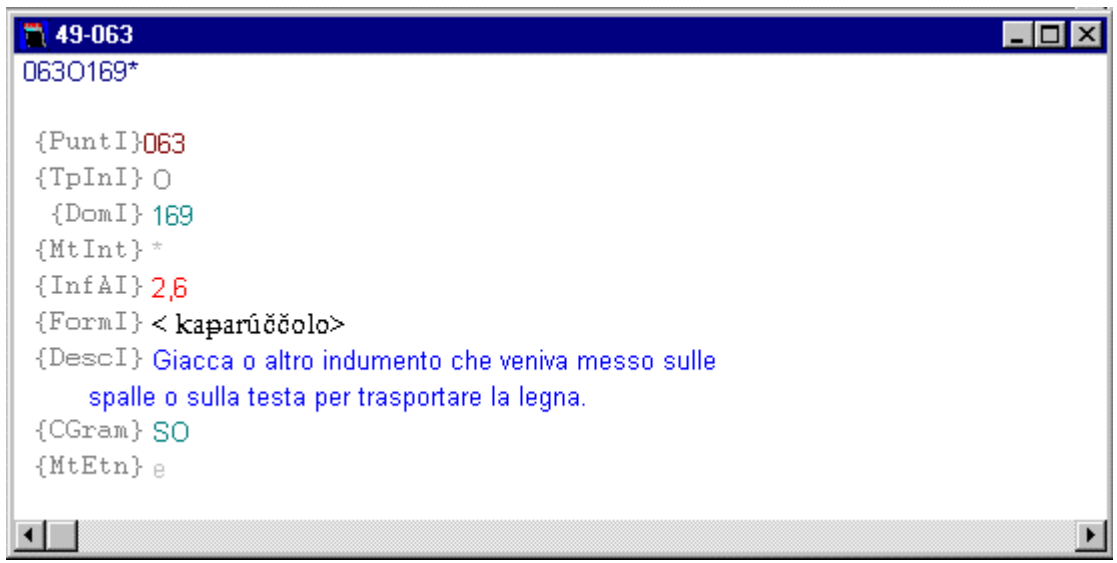
Tipologia di schede della BD-ALT

- Risposte a domande semasiologiche



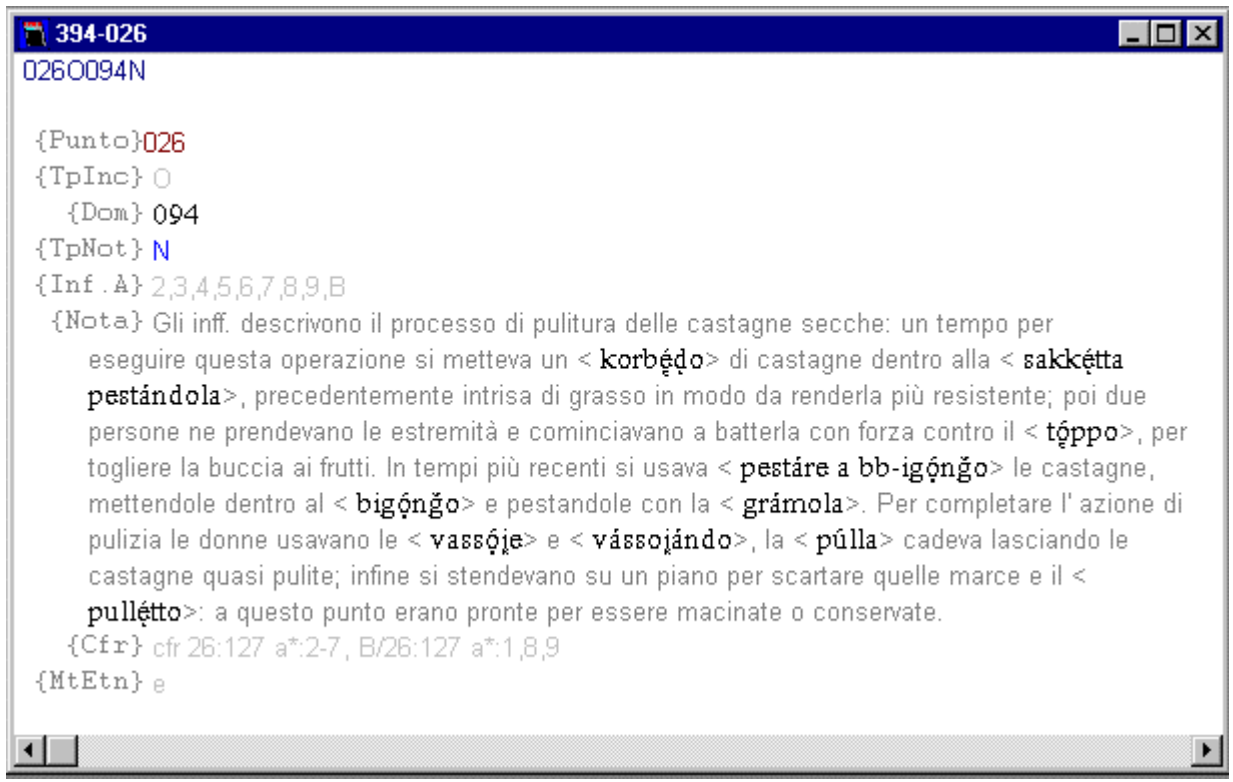
Tipologia di schede della BD-ALT

- Materiali integrativi



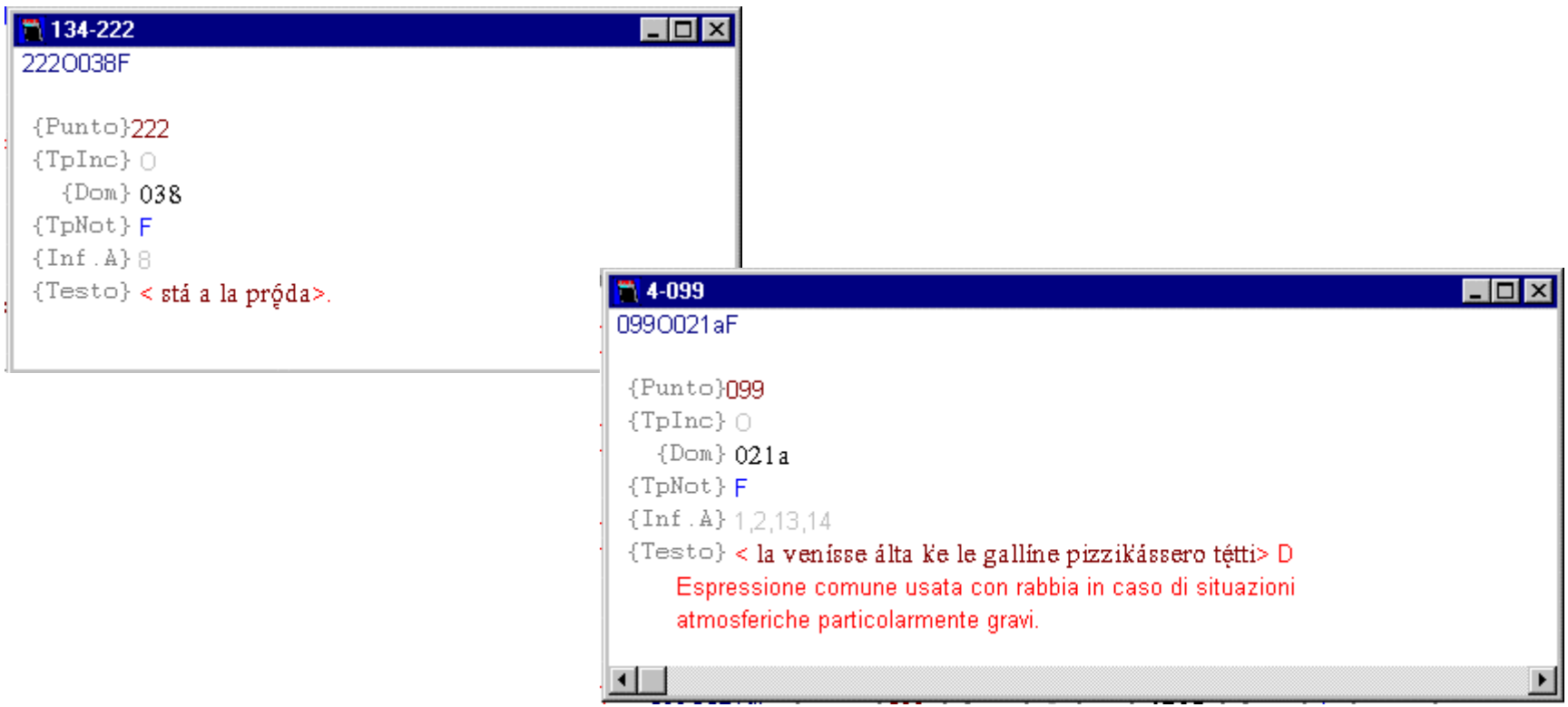
Tipologia di schede della BD-ALT

- Note di commento



Tipologia di schede della BD-ALT

- Note fraseologiche (contesti tipici, proverbi e detti)



Codifica dei materiali in trascrizione fonetica (1)

Il sistema di trascrizione fonetica adottato nell'ALT si rifà al sistema di trascrizione stabilito per la Carta dei Dialetti Italiani (CDI) in versione specializzata per la codifica di materiali relativi ai dialetti toscani

Al fine di assicurare un trattamento adeguato dei materiali registrati in trascrizione fonetica nelle varie fasi di elaborazione ed analisi è stato messo a punto un sistema di codifica complesso che soddisfacesse le richieste specifiche di compiti differenti:

- immissione e correzione;
- ordinamento;
- recupero;
- visualizzazione su schermo e in stampa.

Codifica dei materiali in trascrizione fonetica (2)

Questo schema di codifica, per uno stesso simbolo, affianca rappresentazioni di tipo diverso che, a seconda del compito, sono automaticamente convertite le une nelle altre

Rappresentazioni atomiche:

mostrano una corrispondenza 1:1 tra i simboli dell'alfabeto fonetico e i codici interni loro associati; usate per la visualizzazione su schermo e in stampa

Rappresentazioni composizionali:

codificano ogni simbolo fonetico mediante una base fonetica che può essere ulteriormente specificata da eventuali diacritici (ad es. / e̩ / > e18) ;

presentano innegabili vantaggi per:

- immissione/correzione dei dati: tutti i simboli dell'alfabeto fonetico (117) sono codificati mediante un numero ristretto di codici (36 basi e 9 diacritici)
- ordinamento e recupero: permettono di astrarre da aspetti particolari della realizzazione fonetica e dunque raggruppare varianti fonetiche

Rappresentazioni normalizzate:

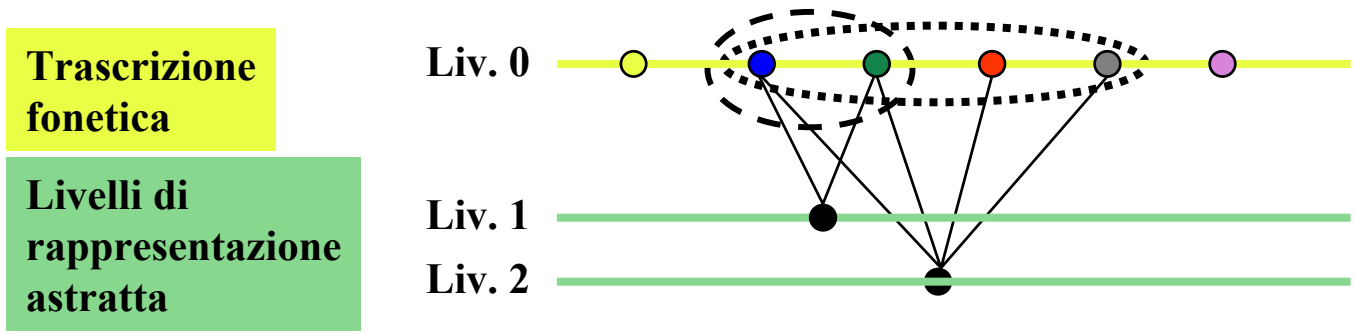
associano ad ogni simbolo o sequenza di simboli fonetici una rappresentazione normalizzata, che viene invocata per interrogazioni che vogliano fare maggiore astrazione da aspetti della realizzazione fonetica

Il sistema di interrogazione di DBT-ALT: funzionalità di base

- DBT-ALT fornisce procedure di ricerca dinamiche che permettono all'utente di definire interattivamente la chiave di accesso al corpus dei materiali
- I materiali dell'Atlante Lessicale Toscano possono essere recuperati sulla base di un ampio spettro di parametri:
  - ☑ domanda del questionario a cui si correlano, direttamente o indirettamente
    - ☑ chiavi semantiche volte a identificare le domande pertinenti rispetto all'oggetto della ricerca
  - ☑ località di inchiesta in cui sono stati raccolti
  - ☑ forma (in trascrizione fonetica)
  - ☑ componenti di significato (inferibili da elementi della descrizione loro associata)

Recupero di materiali lessicali in trascrizione fonetica

- DBT-ALT offre all'utente la possibilità di selezionare livelli differenziati (di astrazione incrementale) per l'accesso ai dati in trascrizione fonetica da selezionarsi a seconda delle finalità della propria ricerca
  - ☞ accesso di livello 0: opera sulle rappresentazioni atomiche;
  - ☞ livelli di accesso che permettono all'utente di astrarre da tratti specifici della realizzazione fonetica del dato attestato, operanti sulle rappresentazioni composizionali e normalizzate:
    - ☑ livello 1: astrae da segni diacritici quali apertura di vocale, spirantizzazione di occlusiva etc.
    - ☑ livello 2: si incentra su relazioni complesse di equivalenza tra fonie (semplici e composte); astrae da variazioni di tipo morfologico (morfologia flessiva, nomi e aggettivi)



Recupero di dati lessicali: accesso di livello 1

Tratti fonici neutralizzati:

- grado di apertura della vocale (/prǫ́da/ ≡ /prǫ́da/ ≡ /prǫ̣́da/)
- spirantizzazione di occlusiva (/abę́to/ ≡ /a~~b~~ę́~~t~~o/)
- perdita di occlusione nelle affricate (/a bačío/ ≡ /a baʗío/)
- nasalizzazione di vocale (/bǫ́mba/ ≡ /bǫ̃́mba/)
- turbamento di vocale (/lúna/ ≡ /lün̈a/)
- consonantizzazione di vocale (/vi̯ǫ̣́ttolo/ ≡ /viǫ̣́ttolo/)
- carattere velare di /l/ e /n/ (/altalę́na/ ≡ /altaɬę́na/)

- posizione dell'accento: /krǫ́ńńolo/ ≡ /kroǹńǫ́lo/
- anche in fonosintassi: ad esempio, si astrae dalla spirantizzazione di occlusiva in contesto frasale, neutralizzando la distinzione tra /prǫ́da/ e /(la) ~~p~~rǫ́da/

Recupero di dati lessicali: accesso di livello 2

CLASSE di EQUIVALENZA	ESEMPI
č, čč, ḉ, š (sse #_V // V_V)	/a bačío/ ≡ /a baččío/ ≡ /a baćío/ ≡ /a bašío/
ǧ, ǵ, ǧǧ, ʃ́ (sse #_V // V_V)	/ǧákka/ ≡ /ǧǧákka/ ≡ /ʃ́ákka/ /čeráǵa/ ≡ /čeráǧa/
k, ķ, h	/bákera/ ≡ /bákera/ ≡ /báhera/
ki̯, ķ i̯, kki̯, ti̯, ŧi̯, tti̯, č, čč, ť, ťť	/mákkia̯/ ≡ /máča/ ≡ /mačča/ ≡ /máťťa/ /pettié̯re/ ≡ /pečé̯re/
gi̯, ǥi̯, ggi̯, di̯, đi̯, ddi̯, ǧ, ǧǧ, d', d'd'	/gia̯čča̯ia/ ≡ /dia̯čča̯ia/ ≡ /ǧačča̯ia/ ≡ /d'ačča̯ia/
l, r (sse _C[≠ r]), i̯ (sse _Ci Ci), l' (sse _C)	/dó̯lko/ ≡ /dó̯rko/ ≡ /dó̯l'ko/ ≡ /dó̯ikko/
ll, ḍ, ḍḍ	/balló̯tti/ ≡ /baḍó̯tti/
ľ, ľľ, li̯, lli̯, ǧ, ǧǧ, d', d'd'	/číľľ'o/ ≡ /číľ'o/ ≡ /číli̯o/ ≡ /číǧǧo/
ń, ńń, ni̯, nni̯	/panni̯é̯re/ ≡ /pani̯é̯re/ ≡ /pańé̯re/ ≡ /pańńé̯re/
r, rr (sse V_V)	/karraré̯čča/ ≡ /karraré̯čča/
s, S, ś, š (sse _C), ʃ́ (sse _C), z (sse l_ // r_ // n_)	/roSmaríno/ ≡ /roʃ́maríno/ /ánsimo/ ≡ /ánzimo/
š, šš	/kášša/ ≡ /kaša/
z, zz, ʒ, ʒʒ	/zzázzera/ ≡ /zázzera/ ≡ /ʒʒázzera/ ≡ /ʒʒáʒʒera/

Fonosintassi:

- raddoppiamento fonosintattico: /a solatío/ ≡ /a ssolatío/; /a bačío/ ≡ /a bbačío/;
- spirantizzazione di occlusiva: /tornár di ķása/ ≡ /torná ddi hása/ ≡ /andá n kása/;
- perdita dell'elemento occlusivo delle affricate: /ai ḉipré̯ssi/ ≡ /ai čipré̯ssi/;
- affricazione delle sibilanti: /al zolatío/ ≡ /al solatío/;
- caduta della vocale finale: /ákku̯ e nnve/ ≡ /ákku̯a e nnve/; /fra llúsk e bbrúsko/ ≡ /fra llúsko e bbrúsko/.

Il sistema di interrogazione di DBT-ALT: funzionalità di base

- DBT-ALT fornisce procedure di ricerca dinamiche che permettono all'utente di definire interattivamente la chiave di accesso al corpus dei materiali
- I materiali dell'Atlante Lessicale Toscano possono essere recuperati sulla base di un ampio spettro di parametri:
  - ☑ domanda del questionario a cui si correlano, direttamente o indirettamente
    - ☑ chiavi semantiche volte a identificare le domande pertinenti rispetto all'oggetto della ricerca
  - ☑ località di inchiesta in cui sono stati raccolti
  - ☑ forma (in trascrizione fonetica)
  - ☑ componenti di significato (inferibili da elementi della descrizione loro associata)

Il sistema di interrogazione di DBT-ALT: funzionalità complesse di accesso

- Le funzionalità di accesso di base possono essere variamente combinate per la formulazione di interrogazioni complesse alla ricerca di:
  - co-occorrenza di diversi tipi di informazione all'interno della stessa scheda
  - occorrenza di una attestazione tra un insieme di varianti

- I risultati di una interrogazione possono essere filtrati sulla base di:
  - status socio-economico e culturale dell'informatore che li ha attestati
  - aree geografiche, definite su base amministrativa o geo-morfologica, in cui l'attestazione è stata raccolta
  - rilevanza rispetto ad uno specifico dominio semantico
  - status socio-linguistico dell'attestazione dialettale, ad esempio registro, connotazione, uso, etc.

Proiezione su carta dei risultati di un'interrogazione

- DBT-ALT fornisce anche la possibilità proiettare i risultati della ricerca su una mappa producendo così una carta dialettale a tutti gli effetti
- Questa mappa è una mappa di presenza/assenza, ovvero marca tutti i punti dove l'oggetto dell'interrogazione ha registrato una sua attestazione positiva.
- Carte multi-livello saranno possibili:
  - che combinano i risultati di diverse interrogazioni
  - che proiettano i risultati di una ricerca su tipologie di sfondi diversi
- In questa prospettiva, la carta dialettale si configura come uno strumento di ricerca utile e flessibile

Banche dati e ricerca dialettale

L'esperienza di DBT-ALT mostra che i materiali di un atlante linguistico, quando strutturati e codificati in modo appropriato, si configurano come una risorsa linguistica la cui utenza e uso vanno al di là di quelli previsti in partenza

Una banca dati dialettali:

- fornisce un utile strumento di supporto alla ricerca dialettale di impianto tradizionale
- costituisce il punto di partenza per la creazione di visioni "sintetiche" della variazione linguistica

Una Banca Dati volta a valorizzare ogni valenza del dato linguistico colma il vuoto che si è venuto a creare nel passaggio da depositi statici di dati a tecniche avanzate di induzione e modellizzazione della variazione linguistica

Futuro … (1)

- risorsa aperta, incrementabile
  - con risultati di ulteriori inchieste
  - con materiali di risorse preesistenti: vocabolari dialettali o altri atlanti

- punto di partenza per altre risorse ed elaborazioni
  - costruzione di repertori lessicali relativi ad una data area, oppure specializzati rispetto ad un dominio semantico (Stazione Lessicografica)
  - selezione di materiali per tecniche avanzate di identificazione e rappresentazione della variazione linguistica

Futuro … (2)

- raffinamenti dei dati e delle procedure di analisi ed elaborazione:
  - accesso ai dati a partire da:
    - tipi lessicali
    - forme in ortografia italiana
  - associazione ad ogni forma dialettale della corrispondente realizzazione fonetica
  - potenziamento accesso per chiavi semantiche, mediante collegamento delle chiavi a ontologie semantiche
  - strutturazione delle entrate in forma di tassonomia semantica, acquisita automaticamente dall'analisi sintattica delle definizioni
  - cartografazione multidimensionale

DBT-ALT:

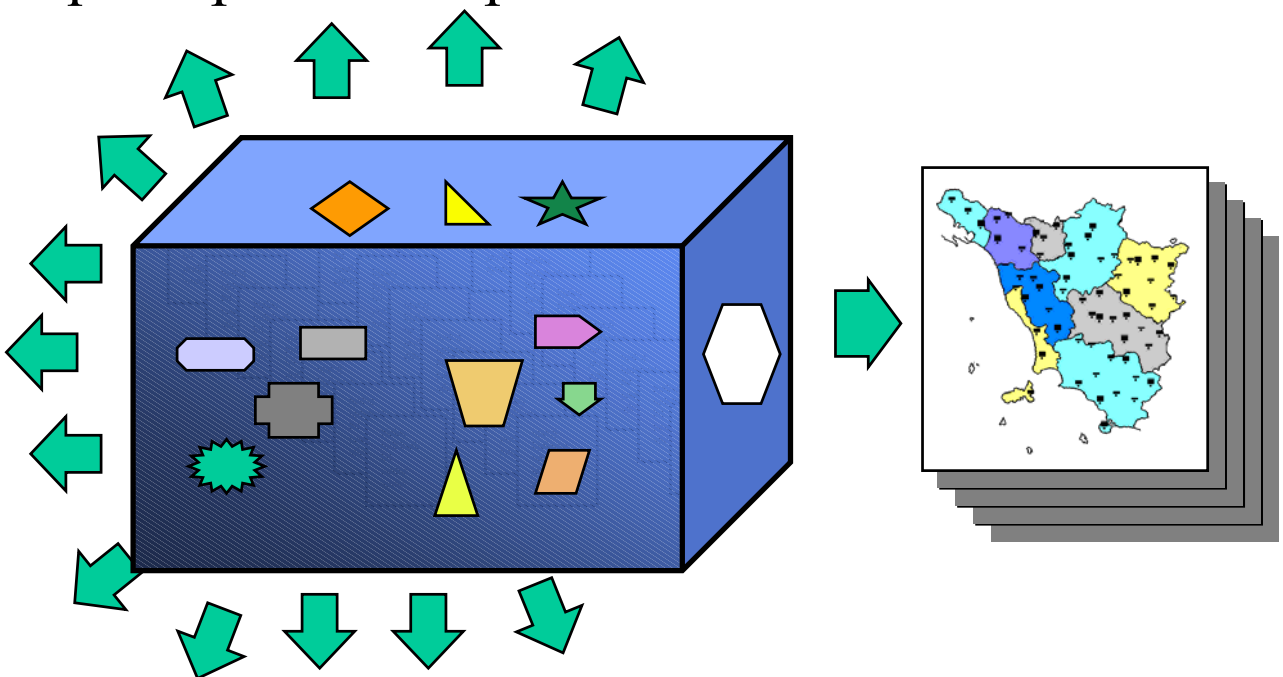
- Ritagliato sulla realtà dell'Atlante Lessicale Toscano
  - costruito a partire da moduli DBT standard
    - per la gestioni di dati strutturati
    - per la gestione di alfabeti non latini
  - integrato con componenti specifici:
    - per la gestione di archivi sussidiari contenenti informazioni relative alle località indagate, agli informatori intervistati ed al questionario
    - per la proiezione su carta dei risultati di una ricerca

Facilmente esportabile ad altre imprese una volta che siano stati attentamente considerati i seguenti fattori:

- tipologia dei dati da strutturare ed elaborare
- parametri di selezione dei dati
- presenza di materiali in trascrizione fonetica
  - tipo di alfabeto fonetico adottato
  - codifica, da definirsi anche in relazione alla tipologia dei livelli di accesso

Banche dati e ricerca dialettale

L’esperienza di DBT-ALT mostra che i materiali di un atlante linguistico, quando strutturati e codificati in modo appropriato, si configurano come una risorsa linguistica la cui utenza e uso vanno al di là di quelli previsti in partenza



Una Banca Dati volta a valorizzare ogni valenza del dato linguistico colma il vuoto che si è venuto a creare nel passaggio da depositi statici di dati a tecniche avanzate di induzione e modellizzazione della variazione linguistica