Il database dell'Atlante Sintattico d'Italia (ASIt)

Diego Pescarini e Giorgio Maria Di Nunzio

1. Introduzione

La progettazione del database dell'Atlante Sintattico d'Italia (d'ora in poi: ASIt) si inserisce nell'ambito di un progetto più ampio che intende raccogliere, immagazzinare ed analizzare dati concernenti la microvariariazione morfo-sintattica nel dominio Italo-romanzo[1]. Il punto di partenza è il lavoro svolto per l'Atlante Sintattico dell'Italia Settentrionale (ASIS): un progetto di ricerca ventennale che costituisce la principale fonte di documentazione sulla variazione grammaticale dei dialetti dell'Italia settentrionale. Per una presentazione delle metodologie e dei risultati dell'ASIS si rimanda a Benincà (1989, 1995), Benincà & Poletto (1992, 2007).

Basandosi sulle metodologie e gli strumenti dell'ASIS, alla fine degli anni novanta il gruppo di ricerca ha iniziato ad affrontare lo studio delle varietà meridionali, determinando di fatto l'inclusione del progetto ASIS in un contenitore più ampio: l'ASIt. Per far fronte all'incremento delle base empirica, nel 2006 si è deciso di aggiornare anche l'aspetto della gestione informatizzata dei dati, dando vita ad un progetto di ricerca congiunto dei dipartimenti di Linguistica ed Ingegneria dell'informazione dell'Università di Padova[2]. In questa sede cercheremo quindi di illustrare le metodologie ed i risultati di questa collaborazione fra linguisti ed informatici, concentrandoci in particolare sull'illustrazione del nuovo database che consente di ottimizzare la gestione ed il recupero dei dati.

---

[1] Sebbene il presente lavoro sia frutto della collaborazione dei due autori, Diego Pescarini ha la responsabilità della stesura dei parr. 1, 2, 3, mentre Giorgio Maria di Nunzio si è occupato dei parr. 4, 5, 6.

[2] Tale collaborazione è stata supportata economicamente dall'Università di Padova (Progetti d'ateneo, bando 2006). Vi hanno preso parte Paola Benincà, Maristella Agosti, Cecilia Poletto, Jacopo Garzonio, Federico Damonte e Riccardo Miotto, oltre agli autori del presente contributo, che in questa sede fungono da 'portavoce' dell'intero gruppo.

Quaderni di Lavoro ASIt 10: 63-81, 2010, ISSN: 1828-2326
Atti della XV Giornata di Dialettologia
A cura di Diana Vedovato

L'articolo è organizzato come segue: nel §2 discuteremo brevemente le criticità del vecchio sistema di gestione; nel §3 cercheremo di descrivere brevemente il metodo impiegato per armonizzare il lavoro delle due unità di ricerca; nel §4 inizieremo a delineare le caratteristiche del nuovo database, illustrando brevemente la sua architettura; in §5 spiegheremo brevemente come i dati linguistici rilevanti vengano recuperati nel database sulla base di un sistema di marche morfo-sintattiche (tag); nel §6 forniremo alcune informazioni relative all'interfaccia grafica per la gestione della base di dati; infine, nelle conclusioni, cercheremo di elencare alcuni possibili sviluppi futuri del progetto.

2. I limiti del vecchio database

Il corpus dell'ASIt è il frutto delle inchieste dialettologiche realizzate tramite questionari scritti. I questionari – consultabili all'URL http://asit.maldura.unipd.it/questionnaires.html – sono insiemi di frasi italiane pensate per testare alcuni fenomeni sintattici potenzialmente rilevanti. Ad ogni questionario corrisponde una serie di traduzioni dialettali che consentono di valutare la presenza e le caratteristiche di una determinata costruzione grammaticale nei dialetti indagati.

Il recupero delle informazioni all'interno del database avviene sulla base di un sistema di marche morfosintattiche (*tag*) che specificano i tipi di fenomeni grammaticali presenti nella frase italiana (ad esempio, la marca 'acc partic' indica l'accordo fra l'oggetto ed il participio passato). Il recupero dell'informazione dialettale avviene in modo indiretto: selezionando una marca (o un piccolo insieme di marche), il sistema seleziona le frasi italiane che contengono il corrispondente fenomeno e recupera automaticamente tutte le traduzioni dialettali delle frasi selezionate. Non essendo direttamente interrogabili, i dati dialettali non sono mai stati marcati (ad eccezione di alcune frasi di prova).

Il precedente sistema ASIS/ASIt era quindi costituito da un corpus multilingue, parallelo e parzialmente annotato, organizzato sulla base di un database composto essenzialmente da due tabelle contenenti tutte le informazioni raccolte, interrogabile tramite stringhe di testo. Tale sistema mostrava però alcune criticità che, nel lungo periodo, avrebbero potuto limitare i possibili sviluppi del progetto.

In primo luogo, il database non era corredato di alcuna documentazione riguardo la progettazione concettuale né tantomento delle eventuali ristrutturazioni per la realizzazione

dello schema logico. Questa mancanza ha di fatto ostacolato la possibilità di ricostruire il progetto originale, operazione che avrebbe permesso di risalire alla comprensione dei requisiti iniziali. Inoltre l'ispezione dei dati contenuti nelle due tabelle principali ha rivelato la totale assenza di vincoli di integrità referenziale e di chiavi esterne, rendendo di fatto impossibile la ricostruzione dello schema originale nemmeno con un processo di *reverse engineering* a partire dai dati. La rigidità della definizione dello schema originale e la mancanza di documentazione non hanno permesso di riutilizzare né di estendere nessuna tabella esistente.

L'interrogazione dei dati dialettali avveniva tramite la ricerca della specifica stringa di testo (quindi, nel caso sopra, 'acc partic') e non c'erano sistemi, né in fase di immissione dei dati né in fase di ricerca, per richiamare velocemente ed efficacemente tali stringhe di testo: bastava quindi un banale errore di battitura per compromettere la memorizzazione o il recupero del dato.

Mancava inoltre completamente un sistema per consentire una corretta identificazione dei singoli dialetti: ogni varietà dialettale veniva identificata attraverso il glottonimo o, in assenza, con il nome della frazione/comune. Questo sistema disomogeneo, tuttavia, non garantiva un efficace ed immediato riconoscimento della provenienza del dato. In particolare, il fatto che non fosse disponibile alcuna informazione relativa alla provincia o alla regione di appartenenza complicava il riconoscimento della varietà esaminata e, con l'aumento dei punti d'inchiesta, aumentava il rischio di denominazioni ambigue, specialmente fra frazioni di comuni diversi.

Infine non esisteva modo di preservare la tracciabilità dei dati inseriti, ovvero che si consentisse di individuare con precisione gli informatori e tutte le persone che avevano successivamente manipolato il dato. Ogni questionario, infatti, prima di essere memorizzato nel database viene scrupolosamente esaminato dal gruppo di ricerca, che 'interpreta' i questionari compilati dagli informatori: tale interpretazione può consistere, ad esempio, nel normalizzare le scelte ortografiche – ogni informatore sceglie le proprie, solitamente modificando la grafia dell'Italiano – o nel segmentare delle sequenze ortografiche che corrispondono ad elementi funzionali distinti (complementatori, marche di negazione, pronomi clitici, ausiliari, preposizioni, articoli, ecc.). Alla fine di questa fase di revisione ('editing') il questionario ha subito quindi delle modifiche che, pur non essendo molto invasive, necessitano tuttavia di essere registrate.

Per cercare di risolvere queste criticità, si è richiesto l'intervento del gruppo di Information Management System (IMS), che aiutasse il gruppo di ricerca linguistica nella progettazione di uno strumento completo, ma, allo stesso tempo, facilmente utilizzabile da un

gruppo di ricerca privo di particolari competenze informatiche. Il gruppo IMS ha iniziato una prima fase di verifica dei requisiti del futuro sistema, sulla base della esperienza acquisita nella cura dei dati per la valutazione sperimentale e per la progettazione di risorse multi-linguistiche [Agosti, 2008; Agosti *et alii* 2006]. Tuttavia, rispetto ad alle tradizionali risorse linguistiche – ad esempio i corpora paralleli – ASIt presenta quattro peculiarità fondamentali:

ASIt tratta dialetti: se da un punto di vista glottologico è dato per scontato il fatto che non ci sia alcuna differenza fra una lingua ed un dialetto, è pur vero che delle differenze emergono nel momento in cui si cerca di costruire uno strumento per identificare univocamente un dialetto specifico, sulla base di parametri geografici (il territorio dove il dialetto è parlato) e geo-linguistici (il sotto-gruppo linguistico a cui il dialetto appartiene).

ASIt tratta aspetti grammaticali: la maggior parte delle risorse linguistiche accessibili tramite strumenti informatizzati ha come obiettivo il recupero di informazioni lessicali (es. numero e contesti di occorrenza di una parola) o semantiche (es. tutte le espressioni linguistiche che si riferiscono al medesimo significato). Al contrario, i ricercatori dell'ASIt sono alla ricerca di informazioni morfo-sintattiche, come la presenza vs assenza in un determinato campione di dialetti di un costrutto specifico (es. clitici soggetto, accusativo preposizionale, ecc.)

ASIt è un progetto in continuo sviluppo, sia quantitativo che qualitativo. In primo luogo, i dati immagazzinati nel database riguarderanno aree geografiche sempre più vaste e, in prospettiva, si possono prevedere inchieste sulle medesime aree geografiche, condotte a distanza di anni. In secondo luogo, i dati saranno impiegati per valutare ipotesi teoriche che, al momento della progettazione, non sono ancora esplicite, come la correlazione sistematica di fenomeni che allo stadio attuale sembrano non correlati.

Il progetto ASIt non prevede al momento una scadenza precisa: le inchieste dialettologiche potranno proseguire nel futuro, coinvolgendo un numero sempre crescente di collaboratori. Per questo motivo si rende necessario costruire una piattaforma tecnologica facile da utilizzare anche da utenti con poca familiarità nella gestione di questi strumenti.

Sulla base di questi requisiti si è quindi iniziato a progettare l'intervento di rifacimento del database, che ha richiesto lo sviluppo di una metodologia condivisa – frutto della fusione delle esperienze dei due gruppi di ricerca – e di un approccio innovativo, che consentisse di creare uno strumento di gestione diverso da quelli attualmente in uso. In particolare, date le caratteristiche e le finalità del progetto, il database ASIt si può definire un database 'curato' (*curated database*, Buneman 2009). Un database 'curato' è un database il cui contenuto è stato raccolto, inserito e sottoposto ad una attenta revisione da parte di 'curatori', e che

possiede alcune caratteristiche: i) i dati (tutti o una parte) vengono editati e rivisti a partire dalla loro fonte originaria; ii) tutti i dati grezzi vengono annotati e arricchiti con una descrizione che interpreta il loro significato e la loro provenienza; iii) i dati vengono aggiornati regolarmente dai curatori. L'obiettivo di un database curato è quella di ottenere una qualità di dati elevatissima a fronte di un elevato costo di impiego di personale altamente qualificato sia dal punto di vista tecnico che del particolare settore di applicazione (in questo caso la dialettologia italiana).

3. Metodo

ASIt è uno dei pochissimi database al mondo pensati per cogliere la micro-variazione linguistica concentrandosi su fenomeni morfo-sintattici. Altre esperienze nel settore sono state sviluppate per lo studio dei dialetti Olandesi (SAND, Barbiers *et alii* 2007) e Portoghesi (Cordial-Syn). A differenza di questi progetti, ASIt nasce sulla base dell'esperienza maturata nell'ambito del progetto ASIS e, di conseguenza, richiede un approccio che cerchi di conciliare l'esistente con i nuovi obiettivi. Per questo motivo, la sfida principale rappresentata dalla riprogettazione del database consisteva proprio nel tentativo di armonizzare metodi, modelli e strumenti già ampiamente testati con una piattaforma di nuova concezione. Un progetto multidisciplinare come quello qui presentato si può sviluppare secondo due metodologie distinte:

- sviluppare la progettazione all'interno del gruppo di linguisti e poi affidare la realizzazione agli informatici, i quali avranno così esclusivamente un ruolo di assistenza tecnica;
- lavorare in modo sinergico sin dalla fase di progettazione, costruendo assieme un nuovo approccio al problema basato sulla fusione delle competenze dei due gruppi di ricerca.

Date le numerose peculiarità del progetto, abbiamo optato per la seconda ipotesi, cercando così di sviluppare un progetto innovativo ed originale. All'atto pratico, gli informatici hanno partecipato a tutte le riunioni del progetto, aiutando i linguisti a rendere espliciti i requisiti del futuro sistema, anche pensando a problemi potenziali e a sviluppi futuri.

Lo sviluppo della nuova base di dati dell'ASIt nasce quindi da un lato dall'analisi dettagliata delle esigenze specifiche che emergono durante il lavoro quotidiano di gestione dei

dati e, dall'altro, dall'immaginare quali potrebbero essere gli sviluppi del progetto nel medio-lungo periodo. Per prima cosa abbiamo quindi cercato di rendere espliciti gli aspetti necessari per risolvere le criticità elencate sopra, come quelle legate ad una corretta ed efficiente memorizzazione dei dati (linguistici, geografici, anagrafici, ecc.). D'altro canto, abbiamo cercato di tenere in considerazione anche i potenziali sviluppi futuri del progetto, come la possibilità di fare delle ricerche sul lessico, di prevedere anche un archivio di registrazioni audio, di consentire una rappresentazione cartografica della distribuzione dei fenomeni grammaticali esaminati o di mutare radicalmente l'attuale classificazione geo-linguistica delle varietà dialettali, che attualmente si basa sulla Carta dei Dialetti d'Italia di Giovan Battista Pellegrini (Pellegrini 1977). Lo scopo della fase di progettazione è stato quindi quello di pensare a condizioni non ottimali o a sviluppi futuri e futuribili. In entrambi i casi era infatti necessario garantire il funzionamento efficace del sistema in modo tale che lacune o aggiunte non compromettessero la funzionalità della base di dati, ne limitassero le capacità di espansione o ne impedissero l'aggiornamento.

4. L'architettura del database

Per garantire la flessibilità necessaria allo sviluppo del progetto ASIt, si è optato per una struttura relazionale, ovvero uno schema che garantisca la massima flessibilità grazie all'indipendenza fra i dati raccolti e le relazioni fra tali dati. Illustriamo ora brevemente i tipi di dati presenti nel database cercando di dare un'idea di come sono stati messi in relazione.

4.1. Punto d'inchiesta

All'interno del database ASIt possiamo individuare due tipi di dati: linguistici (le frasi italiane e dialettali) ed extra-linguistici (informazioni sui parlanti, sui dialetti esaminati, coordinate geografiche dei punti d'inchiesta, ecc.). I dati linguistici sono organizzati in questionari (QUESTIONNAIRE), a loro volta formati dalla combinazione di frasi. Ogni questionario è collegato a due entità: AUTOR e DIALECT. La prima tabella contiene i nominativi di tutte le persone che hanno avuto a che fare con il questionario (il ricercatore che lo ha creato, l'informatore che lo ha compilato, l'*editor* che lo ha digitalizzato e ne ha controllato la resa ortografica, il collaboratore che lo ha somministrato). In questo modo

riusciamo a tracciare l'informazione linguistica, ovvero a conoscere chi sono i responsabili di ogni processo. Nell'ottica di un progetto ampio, a cui collaborano decine di persone, questo diventa un fattore cruciale per garantire la correttezza e l'affidabilità dei dati. L'entità DIALECT, invece, specifica il nome della varietà dialettale parlata dall'informatore. Abbiamo stabilito di far coincidere il nome del dialetto con il nome del comune in cui quel dialetto è parlato, consentendo poi di aggiungere delle informazioni specifiche come il glottonimo, se presente, o il nome della frazione/quartiere, se rilevante. Nel caso di più informatori per la stessa varietà – ed in assenza di informazioni aggiuntive – abbiamo provveduto ad identificarli tramite dei numeri, partendo dall'impostazione che anche parlanti della stessa area geografica possano presentare delle differenze, seppur minime. Ecco il quadro dei casi possibili:

Caso	Dialetto	Varietà
i. presenza di più informatori per lo stesso dialetto:	Padova	1
	Padova	2
ii. presenza di più informatori in zone distinte:	Venezia	Città
	Venezia	Mestre
iii. discrepanza fra il comune ed il glottonimo:	Calasetta	Tabarchino

La scelta di utilizzare il nome del comune per individuare la varietà dialettale è stata fatta con due obiettivi: da un lato consentire una ricerca più efficace poiché la classe dei nomi da ricercare è omogenea (comuni d'Italia), dall'altro individuare più facilmente le coordinate geografiche del dialetto parlato, come si vedrà nel seguente sotto-paragrafo.

4.2. Area Geografica

I dati contenuti nella porzione di database denominata 'Administrative Area' consentono di recuperare tutte le informazioni necessarie per individuare il luogo in cui la varietà dialettale è parlata. Per garantire una corretta gestione di questi dati, ci siamo affidati interamente ai codici ISTAT, evitando così di avere un sistema di catalogazione interno, che, in futuro, sarebbe destinato a diventare desueto. Sulla base dell'indicazione del comune, che – come abbiamo visto – coincide con l'identificativo dei dialetti, possiamo quindi recuperare automaticamente la provincia e la regione di appartenenza. Inoltre, sfruttando i dati ufficiali dell'ISTAT, riusciamo a conoscere automaticamente un numero notevole di informazioni

aggiuntive (superficie, abitanti, altezza s.l.m., posizione litoranea, ecc.). Inoltre, sempre sulla base dei codici ISTAT è possibile recuperare le coordinate geografiche di ogni comune, che rappresentano la chiave d'accesso per poter interfacciare l'ASIt con qualsiasi strumento di cartografia digitale.

4.3. Area Geo-linguistica

La semplice specificazione geografica di un dialetto non è però sufficiente per una corretta definizione. Il caso più problematico è infatti costituito dalle cosiddette *anfizone*: aree in cui si parlano dialetti diversi, come in molti paesi sul confine fra Veneto e Friuli in cui convivono parlanti di dialetti Veneti e Friulani. La sola indicazione relativa al comune non è quindi sufficiente a garantire una corretta identificazione di una varietà e, per questo motivo, l'area del database che specifica i dati linguistici è accompagnata da un insieme di tabelle che specificano l'appartenenza di un determinato dialetto ad un gruppo linguistico. Le entità che formano l'area Geografica e quelle che formano l'area Geolinguistica sono indipendenti fra loro: si potrà quindi avere lo stesso dialetto (es. Palmanova) associato a diversi sotto-gruppi dialettali (Veneto vs Friulano):

Dialetto:	parlato a:	tipo:
Palmanova	Palmanova	VENETO > Triestino-giuliano
Palmanova	Palmanova	FRIULIANO > Occidentale

Per il momento la definizione delle aree geo-linguistiche è stata fatta sulla base della Carta dei Dialetti d'Italia (Pellegrini 1977), ma non si esclude che proprio le ricerche dell'ASIt possano contribuire a tracciare nuove isoglosse di tipo morfo-sintattico.

5. La ricerca nel database

Come si è visto sopra, il principale obiettivo dell'ASIt è quello di comparare alcune strutture sintattiche in diversi dialetti. Assieme al database è stato quindi creato un sistema per il recupero dell'informazione (*information retrieval*) sensibile alle strutture sintattiche

presenti nelle singole frasi. Il sistema si basa su un elenco di 194 marche (*tag*) che specificano le proprietà grammaticale di ogni frase.

I corpora annotati sfruttano dei sistemi di marcatura standard come, ad esempio, CES (Corpus Encoding Standard: http://www.cs.vassar.edu/CES/). Questi tagset sono stati sviluppati per compiere l'analisi automatica di corpora molto vasti: questi tagset sono infatti solitamente impiegati in combinazione con software per il trattamento automatizzato dei dati linguistici, che, basandosi su algoritmi probabilistici, consentono di analizzare in tempi rapidi corpora molto ampi. Il vantaggio di tali sistemi si può quindi apprezzare su larga scala, poiché gli errori commessi dagli strumenti automatizzati sono ampiamente compensati dalla velocità di analisi. Tale analisi è esclusivamente *bottom up*: il software riconosce le singole unità lessicali e – sulla base delle loro proprietà – ricostruisce la struttura della frase. Per questo motivo, le marche impiegate devono specificare il maggior numero di proprietà possibile, poiché ogni unità lessicale deve essere marcata in modo esaustivo. Le caratteristiche di questi tagset sono quindi le seguenti:

- hanno finalità pratiche;
- sono esaustivi;
- servono per l'analisi automatica;
- si basano su algoritmi probabilistici;
- identificano le proprietà degli elementi lessicali;
- la struttura della frase (sintassi) dipende dalle proprietà degli elementi lessicali (analisi *bottom up*).

Nel progettare il sistema di marcatura dell'ASIt abbiamo invece preso in considerazione una serie di requisiti radicalmente differenti. In primo luogo, l'ASIt ha finalità scientifiche: serve cioè a comparare dialetti con lo scopo di isolare le proprietà basilari che determinano la variazione inter-linguistica. Per questo motivo non è necessario che l'inventario di marche utilizzato sia esaustivo: data la finalità del progetto, serviva infatti fare riferimento su un numero di marche sufficiente a cogliere le differenze interlinguistiche. In secondo luogo, il tipo di differenze ricercate sono talmente sottili da richiedere un sistema di marcatura con una granularità molto fine e, per questo motivo, abbiamo anche escluso in partenza la possibilità di ricorrere a strumenti di analisi automatizzata, procedendo così esclusivamente con il tagging manuale. Infine, il gruppo di ricerca si è concentrato su differenze sintattiche che difficilmente possono essere ricondotte a differenze di natura lessicale. Le caratteristiche finali del tagset sono le seguenti:

- ha finalità scientifiche;
- coglie esclusivamente i fenomeni potenzialmente rilevanti per la comparazione;
- serve per l'analisi manuale;
- si basa sull'analisi qualitativa di ogni singola frase;
- identifica le proprietà delle frasi;
- non è detto che tali proprietà dipendano dai tratti associati con i singoli elementi lessicali.

Le differenze con i tagset solitamente in uso sono quindi notevoli e, per questo motivo, abbiamo optato per costruire un nuovo inventario di marche, senza usare quindi alcuno standard riconosciuto. A scopo illustrativo riportiamo qui di seguito alcune marche:

	MARCHE	GLOSSE
1	acc partic clit	: accordo del participio passato col clitico complemento
2	acc partic ogg	: accordo del participio passato con l'oggetto non clitico
3	acc partic sogg	: accordo del participio passato col soggetto
4	Agg	: aggettivo
5	assenza acc partic	: mancanza di accordo tra il participio passato e il clitico complemento
6	assenza clit	: assenza di un clitico presente nella corrispondente frase italiana
7	assenza clit sogg	: assenza di un clitico soggetto
8	aux A	: ausiliare avere
9	aux E	: ausiliare essere
10	aux to C	: ausiliare in area CP
11	avere	: utilizzato come verbo pieno e non come ausiliare
12	avv altri	: forma avverbiale
13	avv aspet	: avverbi con valore aspettuale (mai, più, sempre)
14	avv loc	: forma avverbiale di luogo
15	avv mod	: forma avverbiale di modo

Schema 1 – estratto dal sistema di marcatura dell'ASIt

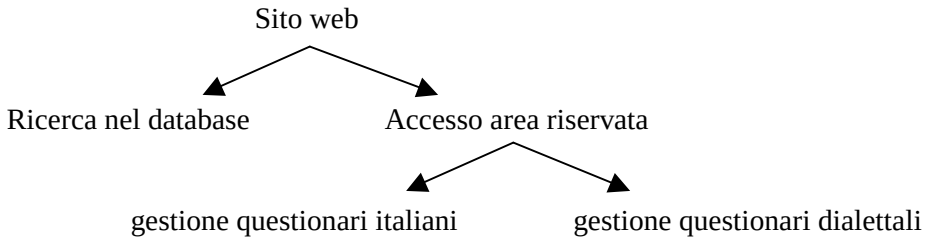
Sulla base delle marche grammaticali proposte è stato progettato e realizzato un sistema di accesso e reperimento delle informazioni geografiche/geolinguistiche consentendo quindi di ricercare non solo l'occorrenza di un dato fenomeno in un'area dialettale specifica,

ma di recuperare le sue caratteristiche specifiche, la co-occorrenza di fenomeni distinti e il contesto in cui appare ogni fenomeno specifico.

Da un punto di vista pratico, le marche sono state raggruppate in sotto-categorie (marche relative al soggetto, al verbo, all'oggetto, al sintagma preposizionale, ecc.) in modo da garantire una gestione più semplice ed intuitiva del tagset. Tale menù – che sarà illustrato nei paragrafi successivi – serve sia in fase di immissione dei dati (per associare le marche alle singole frasi), sia in fase di ricerca (per ricercare le frasi contenenti i fenomeni selezionati). Un'illustrazione più approfondita di queste due fasi è riportata nel paragrafo seguente, in cui cercheremo di mostrare le procedure essenziali per inserire/gestire/cercare i dati attraverso la nuova interfaccia grafica.

6. L'interfaccia

Tutte le interfacce per gestire il database (ricerca/immissione/gestione dati) sono accessibili on-line a partire dalla home page del progetto: http://asit.maldura.unipd.it.



Nelle prossime sezioni cercheremo di mostrare l'organizzazione ed il funzionamento di ogni interfaccia, a partire dal sito web del progetto che, con l'inaugurazione del nuovo database, è stato rinnovato.

La progettazione dell'interfaccia Web ASIt è stata realizzata in maniera da permettere all'utente di eseguire diverse funzioni a seconda del livello di autenticazione. Le funzioni, e le relative interfacce di interazione, messe a disposizione sono le seguenti:

i) creazione di nuove frasi di un questionario italiano e tagging delle frasi;
ii) traduzioni delle frasi in una delle varietà dialettali e tagging delle frasi tradotte;
iii) ricerca nel database di fenomeni e/o elementi grammaticali.

Le funzioni i) e ii) vengono fornite agli amministratori del database e sono accessibili solo mediante autenticazione. Le funzioni iii) sono accessibili per tutti gli utenti anche non autenticati. Nei paragrafi seguenti entreremo più in dettaglio su ciascuna parte del sistema.

6.1. Creazione di frasi italiane e dialettali e tagging delle frasi – funzioni i) e ii)

L'interfaccia di inserimento dati in italiano e in dialetto è stata progettata in base all'analisi dei requisiti svolta all'inizio della collaborazione con il gruppo IMS in cui sono stati evidenziati i seguenti punti:

- l'insieme di 194 tag che caratterizzano i fenomeni grammaticali sono stati raggruppati in classi grammaticali per permettere all'editor di gestire in maniera efficiente il gran numero di etichette;

- nella fase di immissione dei dati dialettali, è necessario visualizzare in parallelo la frase italiana con i suoi tag (sulla sinistra dello schermo) e la frase dialettale, anch'essa con i suoi tag. In questo modo è possibile scorrere le frasi di un questionario italiano e vedere contemporaneamente in parallelo la traduzione dialettale.

- durante la fase di tagging, la lista dei tag associati a ciascuna frase italiana può essere caricata e associata automaticamente alla frase dialettale corrispondente. L'editore a questo punto può operare velocemente delle cancellazioni o aggiunte senza dover inserire tutto l'insieme dei tag da capo, ed è inoltre permessa l'aggiunta di variazioni di diverse traduzioni della stessa frase.

Quando si crea un nuovo questionario in lingua italiana è necessario inserire ciascuna frase ed etichettarla con le proprietà grammaticali. L'associazione tra la frase e i tag che ne specificano le proprietà viene fatta da un esperto linguista con l'aiuto dell'interfaccia mostrata in Figura 1.

Figura 1. Interfaccia per la creazione di questionari in lingua italiana e l'associazione di tag alle frasi.

Analogamente al caso dell'inserimento dei dati di un questionario italiano, quando si inserisce una nuova traduzione in una varietà dialettale è necessario inserire le frasi tradotte ed associare ad esse le etichette grammaticali corrispondenti. Inoltre, è richiesta l'aggiunta di informazioni per caratterizzare il tipo di varietà dialettale in cui si sta traducendo il questionario: il nome della varietà, la località geografica del dialetto, e a quale questionario italiano si fa riferimento per la traduzione (come mostrato in Figura 2).

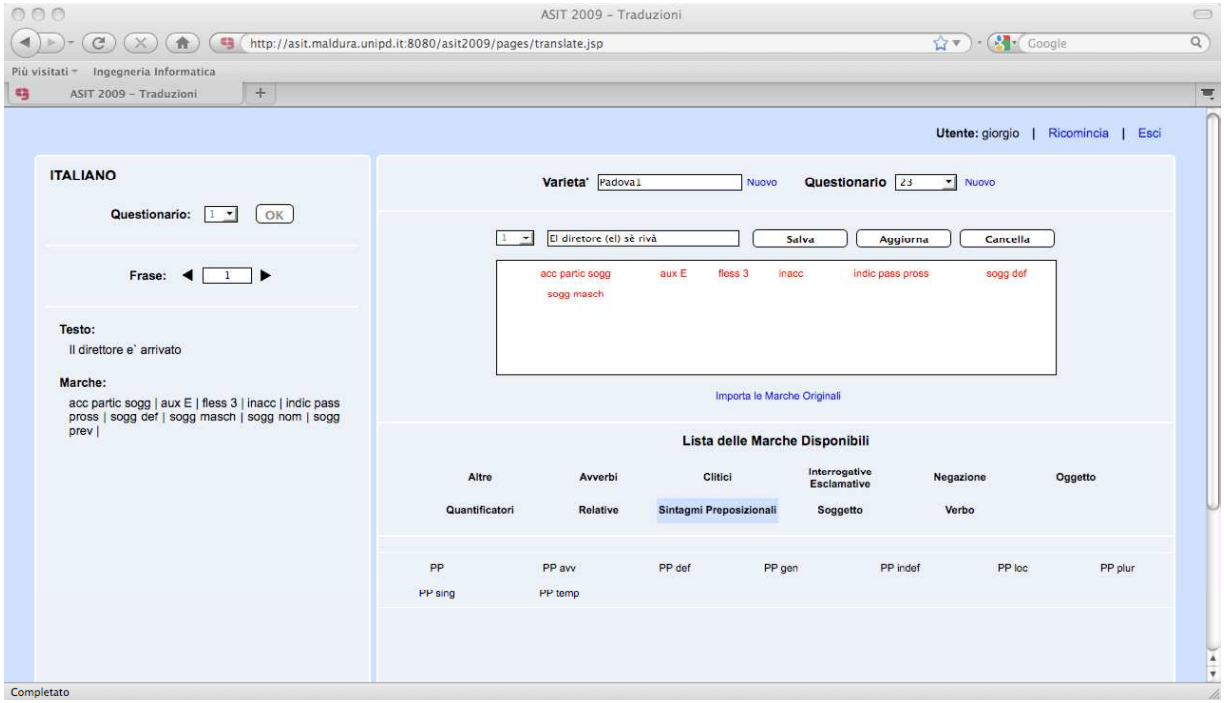


Figura 2. Interfaccia per l'inserimento delle frasi dialettali e l'associazione dei tag alle frasi.

6.2. Ricerca di dati linguistici

Il database ASIt è stato progettato per permettere agli esperti linguisti di accedere ed analizzare i dati recuperando dei fenomeni e/o elementi grammaticali di uno o più punti d'inchiesta, insieme alle caratteristiche di un fenomeno specifico, la co-occorrenza di diversi fenomeni linguistici nello stesso dialetto.

Per poter recuperare questi dati, è possibile specificare una o più proprietà grammaticali sulla base di 194 etichette pre-definite e in seguito filtrare i risultati in base a coordinate geografiche e geo-linguistiche, come mostrato in Figura 3. Le operazioni di ricerca possono essere effettuate in modi diversi, da una ricerca semplice che specifica la presenza di alcuni tag a ricerche più complesse che richiedono il filtraggio per area geografica o area geo-linguistica.

Inoltre, i risultati possono essere scaricati all'interno di un foglio di calcolo (ad esempio Excel), in modo tale che l'utente possa analizzarli off-line o importarli all'interno di file di testo.

Figura 3. Dettaglio dell'interfaccia per la ricerca in ASIt.

Oltre a filtri di natura geografica e linguistica, è possibile anche selezionare i questionari di interesse e le frasi di interesse, in maniera da non mostrare risultati provenienti da frasi e questionari non pertinenti alla ricerca in atto. La visualizzazione dei risultati è organizzata in forma tabellare con un meccanismo del tipo *mostra/nascondi* che permette di aprire solo le traduzioni delle frasi di interesse per l'utente in maniera da rendere compatta la lista di risultati restituiti. Un esempio della maschera di ricerca con le sue opzioni è mostrato in Figura 4 e i risultati della ricerca in Figura 5.

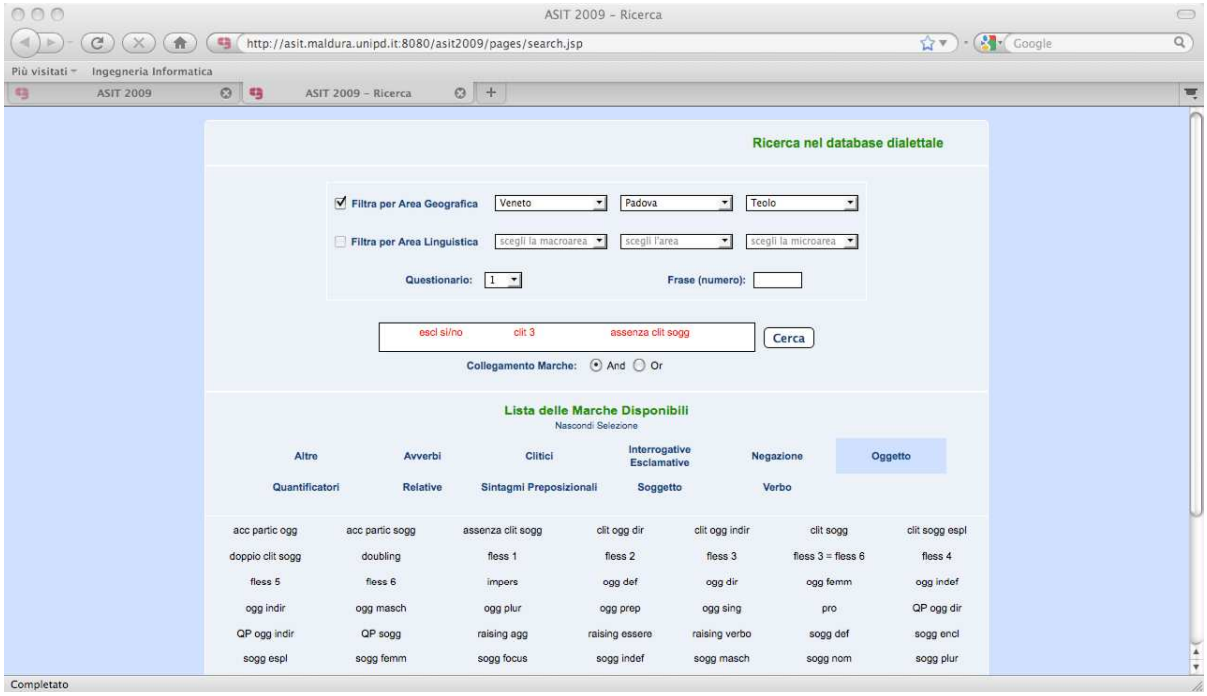


Figura 4. Maschera di ricerca per ASIt con tutte le funzionalità.

Figura 5. Tabella dei risultati di una ricerca.

Oltre alle diverse opzioni di ricerca, il sistema è stato progettato per visualizzare la distribuzione geografica dei vari fenomeni grammaticali attraverso l'utilizzo delle coordinate geografiche memorizzate nel database per ciascun comune e la creazione di file in formato Geo-tagging (tipo GeoRSS[3], Keyhole Markup Language[4], ecc.) che possono essere utilizzati per la creazione di mappe geografice. Un esempio dell'utilizzo di questa funzionalità, ancora non messa a disposizione pubblicamente, è mostrato in Figura 6 dove viene visualizzata una mappa generata dal servizio GoogleMaps per la visualizzazione dei punti di interesse in base al file fornito dal sistema ASIt, in cui ogni punto di interesse rappresenta un punto di inchiesta dialettale.

[3] http://www.georss.org/

[4] http://code.google.com/intl/en-US/apis/kml//

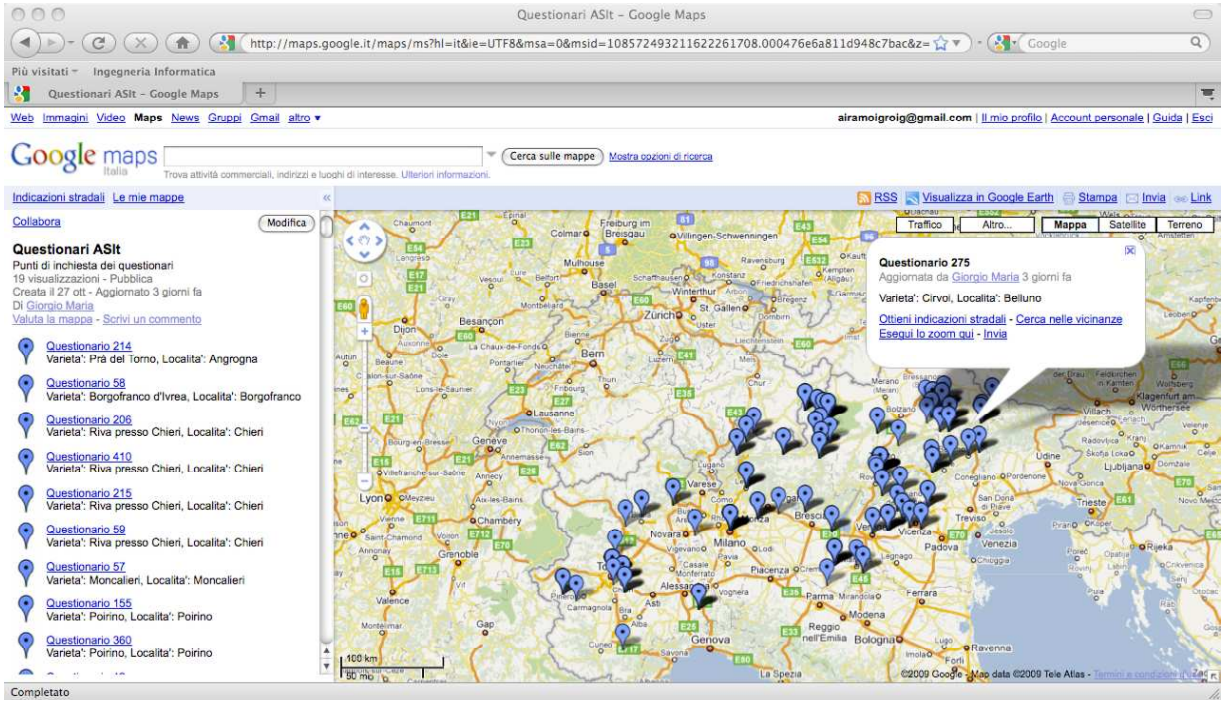


Figura 6. Visualizzazione di punti di inchiesta ASIt attraverso il servizio GoogleMaps.

7. Conclusioni

In questo lavoro abbiamo cercato di dare una breve descrizione dei motivi, metodologie e criteri che hanno portato alla progettazione e realizzazione del database ASIt. In particolare, ci siamo soffermati sulla descrizione dell'architettura del database, sul funzionamento del sistema di recupero dell'informazione e sulle interfacce grafiche che consentono di dialogare con il sistema.

Tale risultato è stato raggiunto grazie alla collaborazione interdisciplinare di due gruppi di lavoro, che hanno lavorato in modo sinergico attraverso un continuo scambio di informazioni. In questo modo è stato possibile creare un bagaglio di conoscenze condivise che può essere utile per lo sviluppo di ulteriori progetti. In particolare, è attualmente allo studio una revisione del sistema di marcatura, in modo tale da renderlo omogeneo con quello impiegato dagli altri database dialettologici europei nell'ambito del progetto europeo Edisyn

(www.dialectsyntax.org). Inoltre, il database ASIt sarà impiegato come tecnologia di supporto ad un progetto per lo studio delle varietà 'cimbre' del Triveneto[5].

Ringraziamenti

Si ringraziano tutte le persone che hanno contribuito allo sviluppo del progetto ASIt, in particolare: Paola Benincà, Cecilia Poletto, Federico Damonte, Jacopo Garzonio del Dipartimento di Discipline Linguistiche, Comunicative e dello Spettacolo, Maristella Agosti, Riccardo Miotto del Dipartimento di Ingegneria dell'Informazione dell'Università degli studi di Padova.

Bibliografia

Agosti, M. (2008). 'Information Access using the Guide of User Requirements' in M. Agosti (Ed.), *Access through Search Engines and Digital Libraries*, Heidelberg: Springer-Verlag, 1–12.

Agosti, M., Di Nunzio, G.M., Ferro, N. (2006). 'Scientific data of an evaluation campaign: Do we properly deal with them?' in C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, M. Stempfhuber (Eds.), *CLEF*, Vol. 4730 of Lecture Notes in Computer Science. Springer, 11–20.

Barbiers, S., Cornips, L., Kunst, J.P. (2007). 'The Syntactic Atlas of the Dutch Dialects (SAND): A Corpus of Elicited Speech and Text as an Online Dynamic Atlas' *CDLC1* 04(54-90).

Benincà, P. (1989). 'Note introduttive ad un atlante dialettale sintattico' in G.L. Borgato e A. Zamboni, *Dialettologia e varia linguistica per Manlio Cortelazzo*. Padova: Unipress.

Benincà, P. (1995). 'I dati dell'ASIS e la sintassi diacronica' in Emanuele Banfi, Giovanni Bonfadini, Patrizia Cordin, Maria Iliescu, *Italia settentrionale: crocevia di idiomi romanzi*. *Atti del convegno internazionale di studi Trento*, 21-23 ottobre 1993, Tubingen, Niemeyer, 1995, pp. 131–141.

[5] Titolo del progetto "Il Cimbro come laboratorio di analisi per la variazione linguistica in sincronia e diacronia - Proposte per una cartografia linguistica del Triveneto" Durata: Maggio 2009 – Maggio 2011. Coordinatore scientifico: Alessandra Tommaselli (Università di Verona). Cofinanziamento della Fondazione CariVerona. Sito web di riferimento: http://ims.dei.unipd.it/data/asit-cimbro/index.html

Benincà, P., Poletto, C. (1992). 'La dialettologia e il modello generativo' *Rivista Italiana di Dialettologia 15*: 77-97.

Benincà, P., Poletto, C. (2007). 'The ASIS enterprise: a view on the construction of a syntactic atlas for the Northern Italian Dialects' *Nordlyd 34*: 35-52. Monographic issue on Scandinavian Dialects Syntax, ed. by K. Bentzen and Ø. A. Vangsnes.

Buneman, P. (2009). 'Curated databases' in: M. Agosti, J. L. Borbinha, S. Kapidakis, C. Papatheodorou, G. Tsakonas (Eds.), ECDL, Vol. 5714 of Lecture Notes in Computer Science, Springer: 2.

Pellegrini, G.B. (1977). *Carta dei dialetti d'Italia*. Pisa: Pacini.