Carla Marello

ALLA RICERCA DELLA FINESTRA D'INTERROGAZIONE AMICHEVOLE MA COMPLETA

1. L'interfaccia di interrogazione di un corpus: osservazioni preliminari

La creazione di un corpus prevede, tra le molteplici operazioni che la accompagnano, un'attenzione particolare alla gestione della finestra di interrogazione dei dati: più che finestra, una vera e propria "porta" d'ingresso di un corpus, di cui rappresenta un aspetto centrale rispetto all'uso che se ne farà. Si tratta, quasi sempre, dell'unica modalità di interazione dell'utente con lo strumento e le informazioni che esso contiene; lo stile che caratterizza la finestra è una spia del tipo di utenza che i creatori del corpus hanno in mente. Il vincolo dell'inserimento di un'espressione regolare nella finestra di interrogazione è già un modo, per esempio, di selezionare la tipologia di fruitori del corpus: l'utente medio della rete non conosce infatti la sintassi specifica con cui costruire le sequenze di simboli che utilizzerebbe invece un linguista computazionale.

A corpus with a view, potremmo dire: una vista che spazia sugli affascinanti orizzonti della lingua, ma solo a condizione che la nostra "finestra" sia realmente un'interfaccia funzionale alla successiva analisi.

In questo contributo si presenta l'evoluzione dell'interfaccia del corpus VALICO (*Varietà Apprendimento Lingua Italiana Corpus Online*, sviluppato presso l'Università di Torino)[1] e le ragioni che hanno portato alla sua attuale finestra di interrogazione, anzi alle due finestre: una per ricerche linguistiche semplici, complesse, miste su corpus e base di dati sociolinguistici e l'altra attraverso una mappa dei paesi di provenienza degli autori dei testi.

[1] Cfr. www.valico.org

2. dai newsgroup all'intera suite di corpora.unito.it

Il gruppo di ricerca *corpora.unito.it* ha scelto per le proprie risorse, sviluppate a partire dal 2002, gli strumenti informatici elaborati dall'IMS (*Institut für Maschinelle Sprachverarbeitung*) di Stoccarda: il *Corpus Query Processor* (CQP) per la codifica del corpus[2] ed il TreeTagger per l'etichettatura delle parti del discorso (PoS), appoggiandosi all'architettura del *Corpus Query Workbench* (CWB)[3].

Siamo partiti dal linguaggio utilizzato all'interno dei newsgroups, basandoci su testi di UseNet e puntando ad una collezione multilingue di corpora di lingua contemporanea, i NUNC (*Newsgroups UseNet Corpora*), sia generici sia specialistici, costruiti per italiano, spagnolo, inglese, francese, tedesco, liberamente interrogabili in rete[4]. La scelta di una collezione di dati a partire da newsgroup si è dimostrata riuscita: la "produttività media" del gruppo di ricerca è stata di 119.825.164 token all'anno, ossia complessivamente 328.288 token al giorno. In termini quantitativi l'aver scelto i newsgroup si è rivelata una decisione adeguata per i nostri obiettivi di ricerca linguistica perché ci ha consentito di elaborare approssimativamente un terzo di milione di parole al giorno per otto anni.

Dopo questa imponente operazione di elaborazione iniziale del linguaggio dei gruppi di discussione in rete[5], si è puntato ad ottenere gli stessi strumenti per l'implementazione di tutti i corpora legati a www.corpora.unito.it, con un grado di specializzazione tale da poter manipolare CQP e TreeTagger in modo da farli aderire ai vari progetti come un vestito tagliato su misura, in un'ottica sia di fidelizzazione dell'utenza sia, evidentemente, di maggiore efficacia e funzionalità.

Tale decisione si è rivelata un vantaggio non solo per coloro che hanno

[2] Si veda Ulrich Heid, *Il "Corpus Workbench" come strumento per la linguistica dei corpora. Principi ed applicazioni*, in M. Barbera - E. Corino - C. Onesti (a cura di), *Corpora e linguistica in rete*, Perugia, Guerra, 2007, pp. 89-108; Id., *Metadata for Learner Corpora: a case study on VALICO*, in E. Corino - C. Marello (a cura di), *VALICO. Studi di linguistica e didattica*, Perugia, Guerra, 2009, pp. 151-65.

[3] Cfr. Oliver Christ - Bruno Maximilian Schulze, *CWB. Corpus Work Bench, Ein flexibles und modulares Anfragesystem für Textcorpora*, in H. Feldweg - E.W. Hinrichs, (edd.), *Lexikon und Text. Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen*, Tübingen, Niemeyer, 1996, pp. 121-33; *online* alla pagina http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/Papers/christ+schulze:tuebingen.94.ps.gz.

[4] Per un elenco completo dei NUNC si veda http://www.bmanuel.org/projects/ng-HOME.html.

[5] Cfr. Manuel Barbera, *Une introduction aux NUNC. Histoire de la création d'un corpus*, in A. Ferrari - L. Lala (a cura di), *Variétés syntaxiques dans la variété des textes* online *en italien: aspects micro- et macrostructuraux* (= «Verbum. Revue de linguistique», XXXIII, 2011, 1-2), pp. 9-36.

lavorato "dietro le quinte" alla progettazione, ma anche per i fruitori stessi, che hanno attualmente la possibilità di operare su un sistema omogeneo.

Il corpus di apprendenti di lingua italiana VALICO ha però beneficiato particolarmente dell'esperienza fatta con il Corpus Taurinense su varietà di italiano delle origini. Per annotare efficacemente testi in italiano non standard è infatti necessaria una pesante opera di tokenizzazione[6].

Illustro qui il passaggio da token a type a lemma in VALICO, esemplificandolo con un enunciato autentico tratto dal corpus:

	la	*prima*	*giornato*	*nela*		*scola*	*non*	*ero*	*ancore*	*capace*
Token	la	prima	giornato	ne	÷la	scola	non	ero	ancore	capace
Type	la	prima	giornata	nella		scuola	non	ero	ancora	capace
Lemma	il	primo	giornata	in	il	scuola	non	essere	ancora	capace
POS	DET	ADJ	NOM	PREP	DET	NOM	ADV	VVFIN	ADV	ADJ

Tab. 1

Come si evince dalla tabella, in particolare dalla terza riga, si è lavorato alla riconduzione del token prodotto dall'apprendente non nativo al type dell'italiano standard (cfr. *giornato* → *giornata*, *nela* → *nella*, *scola* → *scuola*, *ancore* → *ancora*). Questa operazione è indispensabile se si desidera che, impostando la ricerca di una parola, il corpus restituisca non solo i contesti in cui la parola è scritta correttamente, ma anche quelli in cui presenta una forma non standard. È un'operazione che precede il disegno dell'interfaccia di interrogazione, ma che fa parte dei servizi resi all'utente.

Segue l'abbinamento del type al lemma, che possiamo seguire nella quarta riga: *la* ricondotto alla forma base del singolare maschile *il*; *ero* associato al verbo *essere*; l'aggettivo femminile *prima* alla forma maschile *primo*; la preposizione articolata *nella* ricondotta a *in* + *il*; se ci fossero stati dei plurali sarebbero stati ricondotti al lemma al singolare.

3. Evoluzione dell'interfaccia di VALICO

Per quanto riguarda l'interfaccia grafica di VALICO, inizialmente è stata adottata l'interfaccia dei corpora interrogabili con lo standard europeo CQP, che è la stessa attualmente in uso nei NUNC. In particolare, seguendo il

[6] Cfr. Manuel Barbera - Carla Marello, *Corpo a corpo con l'inglese della* corpus linguistics*, anzi, della linguistica dei* corpora, in A. Nesi - D. De Martino (a cura di), *Lingua italiana e scienze*, Firenze 6-8 febbraio 2003, Firenze, Accademia della Crusca, 2012, pp. 357-70.

modello base fornito dall'IMS, il corpus consentiva l'interrogazione linguistica secondo la sintassi CQP o l'interrogazione semplificata per la semplice ricerca di parole all'interno dei testi.

Nell'ottobre 2004 il corpus VALICO è stato reso disponibile online con tale modalità – un'interfaccia tuttora visibile per i vari corpora presenti nel sito www.corpora.unito.it (cfr. Fig. 1).

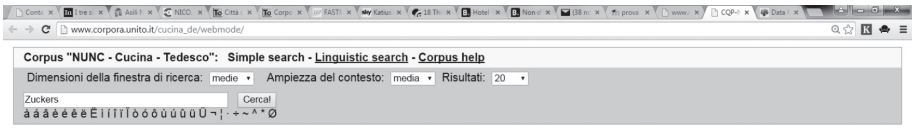

Fig. 1

Questa interfaccia consente la possibilità di variare le dimensioni della finestra di ricerca da piccole, a medie, a grandi; di selezionare un'ampiezza del contesto piccola, media o grande; di visualizzare sulla stessa pagina un numero di risultati variabile tra 20, 100, 250 oppure 1000. Nella finestra vera e propria se non si fa una ricerca di parola singola, come illustrato nella Fig. 1, bisogna impostare delle espressioni regolari. Unico altro aiuto è la riga di caratteri con accenti e pochi altri simboli, riportati sotto la finestra perché l'utente li possa copiare e incollare al bisogno, senza dover andare a cercare il loro numero nel sistema di codifica Unicode.

Tuttavia la tipologia di utente interessato alla fruizione dei testi di VALICO (non solo linguisti, ma docenti, ed eventualmente studenti, di lingua italiana) presuppone la necessità di strutturare un'interfaccia intuitiva ed il meno tecnica possibile, che permetta il raggiungimento rapido ed efficace dei risultati, senza richiedere una competenza informatica strutturata.

In questa direzione Adriano Allora ha sviluppato una maschera di interrogazione che, attraverso bottoni "parlanti" e quindi autoesplicativi, guida l'utente nella gestione di una richiesta (*query*), facendo creare di fatto l'espressione regolare corrispondente alla ricerca desiderata, ma utilizzando un linguaggio non informatico, che si avvale di nozioni intuitive, come "inizio parola" o "parol*" per indicare una sequenza di lettere date seguita da non importa quali e quante lettere[7], e di una terminologia legata alla tradizionale suddivisione in parti del discorso, padroneggiata con sicurezza dai docenti potenzialmente interessati a VALICO.

Di seguito, nella Fig. 2, è possibile seguire il percorso per generare una

[7] Opzione di ricerca molto utile per individuare le famiglie lessicali di derivati suffissali in italiano. La ricerca per "*arola" è invece quella che permette di trovare i prefissati o i derivati con prefisso o primo elemento di non importa quali e quante lettere.

query del tipo *[pos='VER:futu'][pos='ADV'] [pos='ADJ']*, che restituisca cioè forme verbali al futuro seguite da un avverbio e subito dopo da un aggettivo (*pos* sta per "part-of-speech"):

Operazioni utente sull'interfaccia	**Sintassi CQP generata in automatico**
inizio parola	[
un futuro	[pos='VER:futu'
fine parola	[pos='VER:futu']
inizio parola	[pos='VER:futu'][
un avverbio	[pos='VER:futu'][pos='ADV'
fine parola	[pos='VER:futu'][pos='ADV']
inizio parola	[pos='VER:futu'][pos='ADV'][
un aggettivo	[pos='VER:futu'][pos='ADV'][pos='ADJ'
fine parola	[pos='VER:futu'][pos='ADV'][pos='ADJ']

Fig. 2

Dalla *query* [pos='VER:futu'][pos='ADV'] [pos='ADJ'] si ottengono in VALICO risultati come questi:

1	3796	, quando lui si sveglia , sarà proprio furioso ! IERI AL PARCO HO ASSIS
2	4584	ia bellissima . Certo che saranno molto felici . Mi piacciono loro molt

E schiacciando il tasto che dice "Vuoi fare la stessa ricerca in VINCA?" si ricavano altri contesti[8]:

1	118	sembra quel signore , avrà forse sessant ' anni , in bocca una sigaretta
2	397	rovesciamento del tavolino sarà immediatamente indaffarata a rimettere insieme
3	637	abbigliamento stile boy scout , farà anche qualche spattacolino magari.

Poiché ogni testo raccolto per i corpora VALICO e VINCA è accompagnato da un breve profilo sociolinguistico dell'autore/autrice e da dati

[8] VINCA (*Varietà Italiano di Nativi Corpus Appaiato*) è un corpus formato da testi di italofoni nativi; tali scritti sono elicitati a partire dagli stessi stimoli iconici usati per VALICO.

relativi al luogo e data di raccolta, trascrittore, vignetta-stimolo, ci è parso importante poter accedere al corpus anche attraverso questa base di dati. Elisa Corino ha negli anni formato decine e decine di studenti dell'Università di Torino affinché trasformassero le schede e i testi che ci arrivavano scritti a mano da varie parti del mondo in documenti annotati per CQP e ha poi collaborato con Simona Colombo a selezionare i campi di interrogazione più proficui collegati ai metadati previsti nella fase di elaborazione. Dopo un tentativo fatto per trovare un modo di interrogare interno a CQP[9], si è optato per collegare il corpus a una base di dati e a questo punto il disegno della finestra di interrogazione è diventato più complesso e affollato.

Si è deciso (si veda Fig. 3) di mettere nella parte centrale l'annualità di studio della lingua italiana, verso sinistra la consegna (cioè la storia disegnata servita come stimolo), la lingua madre dell'apprendente, le altre lingue da lui/lei conosciute, in quanto "filtri" fra i più importanti sia per i docenti che per i linguisti.

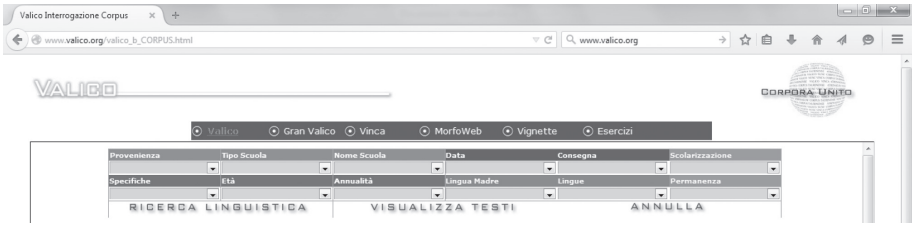

Fig. 3

Selezionando queste caratteristiche una per volta, o anche in combinazione, chi consulta il corpus può ritagliarsi un sottocorpus, o sottocorpora via via più omogenei, come ha fatto ad esempio Elisa Corino per studiare i testi degli apprendenti germanofoni[10].

3.1. *Interfacce con finestra di interrogazione versatile*

Questa finestra di interrogazione combina le potenzialità di interrogazione del CQP con quelle di una base dati in cui sono state caricate le intestazioni,

[9] Si veda U. Heid, *Metadata for Learner Corpora: a case study on VALICO*, cit. e Annette Schaupp, *Entwicklung und Anwendung eines Metadatenmodells für das italienische Lernerkorpus VALICO mit Fokussierung auf den Lernerhintergrund*, Stuttgart, Institut für Maschinelle Sprachverarbeitung, 2006 (Diplomarbeit Nr. 51, Prüfer HD Dr. Ulrich Heid, Zweitprüfer Dr. Helmut Schmid).

[10] Cfr. Elisa Corino, *Italiano di tedeschi. Una ricerca* corpus-based, Perugia, Guerra, 2012.

o *header*, sociolinguistiche associate a ciascun testo. È il frutto del lavoro di Simona Colombo ed è un'interfaccia complessa e versatile perché

a. tramite la selezione del tasto "Ricerca linguistica", che porta all'interfaccia sviluppata da Adriano Allora illustrata sopra nel § 3 Fig. 2, permette di accedere al corpus nella sua interezza senza filtri di ricerca sociolinguistici;

b. attraverso la selezione di uno o più campi porta al popolamento degli altri campi con i relativi dati presenti in base dati e la successiva selezione del tasto "Visualizza testi" conduce l'utente ai testi che rispondono ai parametri impostati;

c. ottenuto un sottoinsieme di testi che rispondono ai requisiti impostati come in *b.*, si può procedere a una "Ricerca linguistica" solo su quel sottoinsieme di testi, utilizzando la stessa interfaccia dell'opzione *a*.

Anche con le ricerche *a.* e *c.* è possibile visualizzare i testi interi: basta cliccare sul numero che appare a sinistra prima del contesto e si accede sia all'intestazione (*header*) che al testo. Le informazioni della intestazione sono a loro volta selezionabili: ad esempio, avendo appurato che il testo appartiene a un certo insieme di testi raccolti in una data occasione, è possibile cliccando vedere tutti i testi raccolti in quell'occasione: un'opzione che permette di studiare eventuali condizionamenti del docente o caratteristiche del gruppo[11].

4. Interfaccia georeferenziata

Tutti i precedenti modi di accesso implicano comunque delle selezioni – sia pure facilitate – da parte dell'utente: il vantaggio di passare ad un'interfaccia geografica risiede innanzitutto nell'immediato colpo d'occhio rispetto alla presenza di testi prodotti da parlanti nativi di un paese specifico. Non è più necessario impostare una ricerca, poiché è possibile consultare i dati a partire da una cartina geografica**.**

Questa interfaccia è stata sviluppata, all'interno del gruppo di ricerca[12],

[11] Ad esempio ci si avvede che il gruppo di scritti, elicitati a partire da *Sogno*, raccolti a Budapest da Chiara Fanton nel 2011, hanno tutti la caratteristica di presentare i nomi propri dei personaggi duplicati: Paolo Paolo, Lucia Lucia, Stefano Stefano, Valeria Valeria.

[12] Il lavoro è stato svolto da Simona Colombo nell'ambito del progetto PRADIGEO, sviluppato dalla società *Annoluce* di Torino con la collaborazione del Dipartimento di Scienze Letterarie e Filologiche dell'Università di Torino. La proposta progettuale ha partecipato nel

integrando le API (*Application Programming Interface*) freeware di Google Maps, che costituiscono la cartografia di base, con un linguaggio XML geografico, il KML (*Keyhole Markup Language*) e con una pagina ASP.NET (*Application Service Provider*) avente le medesime caratteristiche di tutte le altre pagine di interrogazione del sito www.valico.org.

Ciò consente di estrarre le informazioni dalla base dati associata ai testi di VALICO e di tradurle in punti geografici corredati di proprietà che possono anch'esse essere visualizzate, tramite icone su una mappa e, a partire da questa, l'utente può consultare i testi con semplici clic. Le icone scelte sono un cuore per la storia *Amore*, un treno per la storia *Stazione*, un punto interrogativo per *Equivoco*, una saetta per *Scontro* e una faccina gialla con occhi chiusi per *Sogno*.

Nella Fig. 4 ad esempio vediamo punti di raccolta in Europa contrassegnati dall'icona della storia per cui ci sono più testi. Si noti che cliccando sulla Gran Bretagna si scopre che, oltre ai testi elicitati da Amore, sono stati raccolti anche testi per *Stazione* e *Scontro* (designato dalla saetta dell'ira). Cliccando su una singola icona della rosa, si approda al testo intero con intestazione.

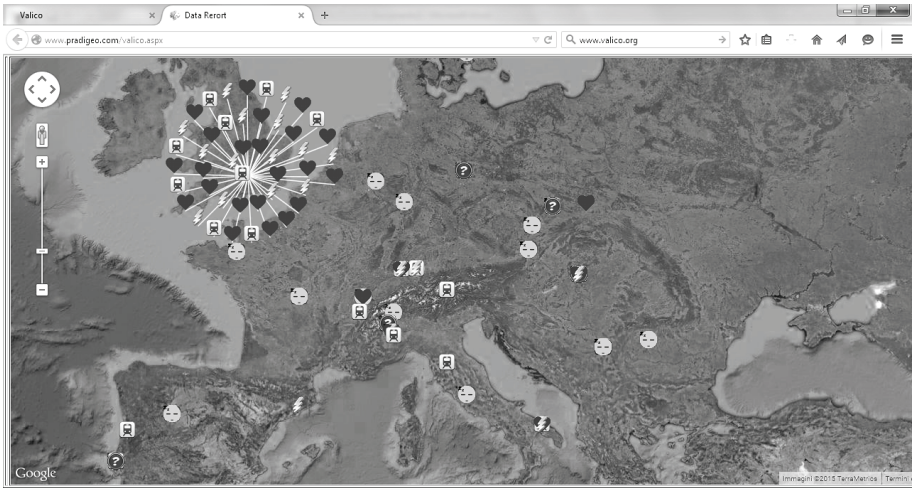

Fig. 4. Visualizzazione a partire dal tipo di consegna

L'approccio georeferenziato è stato adottato anche per accedere agli esercizi di sintassi predisposti per parlanti anglofoni, ispanofoni, ecc. A

2009 ai Bandi Regionali ICT Piemonte (POR FESR 2007-2013) classificandosi 5° assoluta su 240 partecipanti e prima nella sua linea di intervento ("Green web e condotte eco-sostenibili").

fianco di questi esercizi appare l'icona di altoparlante quando abbiamo potuto registrare i ragionamenti degli apprendenti durante lo svolgimento dell'esercizio.

5. conclusioni

Abbiamo mostrato come dall'originaria, unica ammessa, ricerca linguistica per forme e per espressioni regolari, il gruppo di ricerca abbia costantemente cercato di ampliare, senza troppo complicare, le possibilità di interrogazione del corpus VALICO.

Tramite l'interfaccia web georeferenziata è possibile avere una visione della raccolta di testi nel mondo, oppure ingrandire un'area geografica, per entrare nel dettaglio della raccolta. Si tratta di un colpo d'occhio sulla composizione del corpus che, grazie all'uso di carte geografiche e icone vivaci, si propone quale chiave di accesso alternativa per la navigazione fra i dati[13].

[13] Per ulteriori approfondimenti si rimanda a Simona Colombo - Carla Marello, *Un accesso "geografico" alle risorse per l'insegnamento e l'apprendimento di una lingua*, in E. Garavelli - E. Suomela-Härmä (a cura di), *Dal manoscritto al web: canali e modalità di trasmissione dell'italiano. Tecniche, materiali e usi nella storia della lingua*, Atti del XII congresso SILFI Società internazionale di linguistica e filologia italiana, Helsinki, 18-20 giugno 2012, Firenze, Cesati, 2014, pp. 607-17; S. Colombo - C. Marello, *Paesaggi dinamici. Dagli atlanti linguistici alla georeferenziazione di dati linguistici in rete*, in F. Cugno *et al.* (a cura di), *Studi linguistici in onore di Lorenzo Massobrio*, Torino, Istituto dell'Atlante linguistico Italiano, 2014, pp. 231-47.