Pubblicato in La Piazza delle lingue *L'italiano degli altri. Firenze, 27-31 maggio 2010.* Atti, a cura di Nicoletta Maraschio e Domenico De Martino, Firenze, Accademia della Crusca, 2011 ("La Piazza delle lingue", 2).pp.49-61

I corpora VALICO e VINCA: stranieri e italiani alle prese con le stesse attività scritte.

Adriano Allora, Simona Colombo, Carla Marello[1]

(Università di Torino)

1. Presentazione dei corpora VALICO e VINCA

Il gruppo di ricerca torinese che ha creato VALICO e VINCA[2] era ed è dell'opinione che l'italiano come quinta lingua straniera più studiata nel mondo debba avere un corpus di scritti di apprendenti stranieri - come l'hanno l'inglese, il francese, il tedesco e altre lingue – e che questo corpus debba essere in rete e di pubblico dominio[3].

I testi attualmente interrogabili in linea nel corpus VALICO sono 3804. È un corpus in espansione, grazie al fatto che gli insegnanti capiscono sempre di più l'importanza di testimoniare la varietà di apprendimento della propria classe e apprezzano la possibilità di averla disponibile in un corpus che la rende comparabile con quella di altri gruppi di apprendenti affini per età o lingua madre e che tramite VINCA la rende comparabile con la produzione di parlanti, anzi scriventi, nativi italofoni.

VINCA, che ha raggiunto ora i 678 testi in rete, è pure in espansione, nell'intento di aumentare la rappresentatività delle varietà regionali di italiano scritto e dei profili di italofoni scriventi (minore età, terza età, punti di raccolta diversi).

Sia fra gli stranieri che fra gli italofoni gli autori minorenni sono pochi rispetto a quanto sarebbe auspicabile, perché raccogliere la richiesta dell'autorizzazione alla messa in rete quando non è l'autore stesso del testo a poterla firmare, ma deve essere il genitore o un legale rappresentante, è più difficile. Non potendo correre il rischio di mettere in rete un corpus illegale, ci siamo di fatto dovuti accontentare, per ora, dei diversi stadi di apprendimento di una popolazione composta in gran parte da studenti universitari. Inoltre per avere la sicurezza che chi scrive per VALICO sia in grado di produrre un testo scritto accettabile di almeno 100 parole, seguendo la traccia iconica data, abbiamo contattato preferibilmente apprendenti di livello A2, meglio ancora B1 e oltre.

I punti di forza di VALICO e VINCA sono da ricercare nel fatto che

a- sono pensati per più utenti (linguisti, glottodidatti, insegnanti, studenti) ed essendo in rete hanno una finestra di ricerca che si adatta a diversi tipi di utenti e a diversi tipi di ricerca;
b- non sono statici, ma hanno anzi un'architettura pensata per l'incremento e le loro versioni interrogabili in rete valorizzano le potenzialità dell'unione fra interrogazione di corpora annotati e basi di dati sociolinguistici;

---

[1] I §§ 1, 2, 2.1, 3 sono di Carla Marello, il § 2.4 è di Adriano Allora e i §§ 2.2, 2.3, 2.5 sono di Simona Colombo. Si ringrazia Elisa Corino, che coordina la raccolta e l'implementazione dei testi, per tutti i dati non accessibili on line.

[2] Rispettivamente **V**arietà di **A**pprendimento della **L**ingua **I**taliana **C**orpus **O**nline e **V**arietà di **I**taliano di **N**ativi **C**orpus **A**ppaiato, consultabili in www.valico.org. La prima presentazione a stampa del corpus è nel numero 4 del 2004 della rivista ITALS, ma la raccolta dei testi e la contemporanea stesura delle linee guida per la trascrizione sono iniziate nel giugno 2003. VALICO e VINCA sono stati creati con CQP *Corpus Query Processor*, un software per l'allestimento di corpora messoci a disposizione dall'IMC, l'Istituto di Linguistica Computazionale dell'Università di Stoccarda. L'ultima versione delle Linee Guida si può scaricare da http://www.bmanuel.org/projects/br-guidelines_74.pdf.

[3] ADIL2 è su DVD allegato a Palermo (2009). Il corpus di Pavia (cf. Andorno / Rastelli 2009) si rivolge prevalentemente ai linguisti.

c- il ricercatore o l'insegnante possono ritagliare al loro interno sottocorpora omogenei facilmente esportabili;
d- seguono una trascrizione dei testi manoscritti molto accurata, attenta alla fisicità del testo; si sono registrate, fra l'altro, le autocorrezioni, le inserzioni, le variazioni degli scriventi, permettendo di cogliere meglio il processo di formazione del testo (si vedano Tosco 2010 e Vučo 2009 per studi sulle autocorrezioni di studenti rispettivamente sinofoni e serbofoni);
e- sono allestiti partendo dagli stessi stimoli iconici e quindi si prestano per istituire confronti fra produzioni di nativi italofoni e di non nativi a livello di lessico, sintassi, organizzazione testuale, eventuali strategie di evitamento;
f- sono annotati e interrogabili liberamente in rete nello stesso modo in cui si possono interrogare i numerosi altri corpora a cui ha lavorato il gruppo torinese (si veda http://www.corpora.unito.it), situano quindi le varietà di apprendimento in un più ampio ventaglio di varietà standard e non standard di italiano;
g- sono all'interno di una piattaforma, www.valico.org, che mira a concretizzare i risvolti applicativi della linguistica dei corpora di apprendenti, a sviluppare l'autoformazione dei docenti di italiano LS/L2 e l'autoapprendimento guidato da computer[4].

2. I metadati: raccolta e interrogazione

I profili sociolinguistici degli scriventi si ricavano on line tramite una base di dati costruita selezionando alcune delle caratteristiche raccolte tramite il questionario che ciascun scrivente compila mentre firma (lui stesso o il suo legale rappresentante) l'autorizzazione alla messa in rete.
In rete è visibile e interrogabile un'ampia selezione dei dati raccolti, quelli che sono parsi più utili al linguista e all'insegnante. Eccoli in un ordine di importanza che ci è stato anche suggerito dai molti contributori a VALICO poi divenuti utenti: lingua madre dell'apprendente, altre lingue conosciute, consegna, età dell'apprendente, sua scolarizzazione in L1, eventuale permanenza in Italia, sesso. Molti altri dati raccolti e disponibili in rete sono utili all'insegnante e al linguista, ma lo sono maggiormente al linguista dei corpora e in particolare a chi fa la manutenzione e l'aggiornamento di VALICO e VINCA on line. Altri dati interrogabili offline (ad es. il dialetto dei genitori e la loro scolarizzazione per VINCA, le condizioni in cui è stato svolto il testo in classe, i suggerimenti dati dall'insegnante, la possibilità di usare un dizionario per VALICO), poi, possono offrire ulteriori chiavi di interpretazione delle varietà di lingua degli apprendenti.
Per quanto riguarda VINCA, il corpus appaiato di italofoni, il questionario è simile. Ovviamente al posto della lingua madre, di cui tuttavia si intende specificare la varietà, ad esempio i testi di italiano raccolto in Svizzera, si chiede quali altre lingue e dialetti lo scrivente conosca e si dà rilievo al luogo di raccolta.

2.1 **La base di dati sociolinguistici**

Il lavoro che il trascrittore fa per alimentare la base di dati da innestare sul corpus di testi è molto importante, ma ripetitivo e soggetto a errore. Consiste nel leggere le risposte date a mano nel questionario e trascriverle in un file .txt, aggiungendo altri particolari come luogo e data di raccolta, nome del raccoglitore, del trascrittore. L'attribuzione all'insieme di dati dell'intestazione di un

[4] Per quanto concerne l'autoformazione dei docenti di italiano L2, e anche, per certi versi, dei docenti di italiano L1, si vedano alcuni dei contributi in Corino / Marello (a cura di) (2009).

numero di riferimento per agganciarlo al testo vien fatta in un secondo momento dall'informatico che si occupa di mettere in rete la versione aumentata e aggiornata del corpus.
Nella Fig. 1 si trova un esempio di come appare un'intestazione. Chiariamo che per *specifiche* si intende il sesso dell'autore del testo e che per *status* si intende status sociale (in base al reddito da indicare con modesto (1), medio (2), alto (3)). La domanda sullo status è spesso non compilata (e il trascrittore come in questo caso mette un punto interrogativo) ed è perciò una delle informazioni non messe in rete, ma l'utilità di sapere tale parametro ci è stata segnalata da Tanya Roy che, operando in India, voleva avere prove concrete del fatto che l'italiano nel suo paese venisse scelto da studenti di fascia economica modesta, non in grado di sborsare le cifre necessarie per studiare prima di entrare all'Università lingue con accesso universitario a numero chiuso. Per *contatto_lingua esposizione= sc* si intende quando e con chi si pratica la lingua italiana. *sc* significa che lo ha, fino al momento della raccolta, usato solo in situazione scolastica, però potrebbero esserci abbreviazioni come *am*, che sta per amici, *med* che sta per mezzi di comunicazione come televisione o internet. Il soggetto a cui si riferiscono i dati in Fig. 1 ha lo spagnolo come lingua madre, conosce anche la lingua catalana e l'inglese, è un'universitaria fra i 19 e i 25 anni, studia l'italiano solo da un anno, ha 5 mesi di permanenza a Torino; il testo è stato prodotto e raccolto come test di ingresso del suo secondo corso di lingua per studenti Erasmus. Chi ha trascritto ha lavorato su una fotocopia del testo scritto a mano.

```
10.<HEAD> <doc-id>
    <charset>ansi</charset>
    <lingua>italiano</lingua>
    <autore>axxx,xxxxa</autore>
    <fornitore>lisa,beltramo</fornitore>
   <trascrittore>sara,frattin</trascrittore>
    <data>?,?,?</data>
    <luogo>torino,IT</luogo>
    <ist>scuola</ist>
    <ist_nome>università di
       torino</ist_nome>
 </doc-id>

 <set-id>
    <corpus>valico</corpus>
    <gruppo_num>1,gn</gruppo_num>
    <gruppo_nome>ierialparco</gruppo_nome>
 </set-id>

 <autore>
    <specifiche>f</specifiche>
    <età>19-25</età>
    <status>?</status>
    <annualità>1</annualità>
    <lingua1>spagnolo</lingua1>
    <lingua2>catalano,inglese</lingua2>
    <scolarizzazione>un</scolarizzazione>
    <permanenza>5,torino</permanenza>
    <esposizione>sc</esposizione>
 </autore>

 <testo>
<tipo_forma>c-lib_descr</tipo_forma>
    <topics>0</topics>
    <keyw>0</keyw>
    <test>ingresso</test>
    <qualità>origFC</qualità>
    <esecuzione>ms</esecuzione>
  </testo>

  <ref>
<stel>lisabeltramo_F.txt,sarafrattin_T.txt,ier
ialparco_G.txt,ingresso_P.txt</stel>
  <cons>love_C.txt</cons>
  <txtext>0</txtext>
  <imgext>0</imgext>
  <txtint>0</txtint>
  <imgint>love</imgint>
  </ref>
</HEAD>
```

Fig. 1. Esempio di intestazione di testo in VALICO

Poiché i trascrittori sono stati e sono studenti della Facoltà di lingue e letterature straniere dell'Università di Torino (che hanno svolto il lavoro di trascrizione come forma di iniziazione alla linguistica dei corpora, essendo praticamente tutti digiuni di nozioni sull'elaborazione informatica delle lingue naturali), dopo un primo periodo in cui si è contato unicamente sulla loro diligente attenzione, pagando l'eccesso di fiducia con un lavoro di controllo posteriore troppo oneroso, si è pensato di aiutarli a sbagliare meno. Mauro Costantino, dopo aver trascritto lui stesso un bel po' di testi, mentre seguiva il dottorato in ingegneria linguistica ha disegnato *Transcript-o'-matic*, un

modulo in HTML che fornisce al trascrittore campi da compilare nel caso delle scelte aperte come nome e cognome dell'autore, località, eccetera, e menu a tendina per i campi con valori prestabiliti, come le date o i dati metatestuali (vedi Costantino 2009).

2.2 Trascrizione dei testi

Le norme di trascrizione dei testi manoscritti filologicamente concepite da Manuel Barbera sono chiare, particolareggiate, complesse da mettere in pratica. Costantino ha creato una parte del modulo *Transcript-o'-matic* per facilitare il rispetto di tali norme. Rimandiamo a Costantino (2009: 179) per l'illustrazione che riproduce la videata su cui il trascrittore di testi lavora e arriva a una trascrizione come quella sotto riprodotta. L'esempio mostra la trascrizione del corpo (body) del testo di cui prima si è vista l'intestazione (head). Non è troppo complesso da trascrivere perché ha due inserzioni (quelle comprese fra <ins>….</ins>) e un numero di autocorrezioni non troppo alto (quelle inglobate in <corr>…… </corr>). L'autrice è andata a capo, una sola volta, alla fine del testo come segnalato dal simbolo #, mentre non ha iniziato a capo quando ha ricominciato il testo con l'incipit della consegna "Ieri al parco"; non ha usato date e quindi non è stato necessario riprodurre i segnali relativi. Ha usato dei nomi propri (quelli inglobati in <anth>……. </anth>) e un toponimo (*Valentino*, inglobato in <topn>…. </topn>). Il trascrittore, anzi la trascrittrice, volendo, avrebbe potuto essere più precisa e inglobare in <lng_spagnolo>cuando</lng_spagnolo> il secondo *cuando*; sarebbe stato utile per future ricerche sull'uso di parole non italiane, interamente straniere, nei testi.

```
<BODY> $001$Ieri al parco un uomo <corr>uommo</corr> che si chiama <anth>Gianni</anth> era andato a piedi a vedere i fiori . Dopo <ins>Dopo</ins> Lui stava in un banco , <corr>allora</corr> leiendo un giornale e subito vedevo <ins>vedevo</ins> un ragazzo <corr> stava </corr> giovani con una ragazza domandando socorso . <anth>Gianni</anth> si levanto e Ieri al parco <topn>Valentino</topn> un uomo che si chiama <anth>Gianni</anth> era andato a piedi a vedere i fiori perche lui stava stanco per el lavoro della settimana . Dopo , lui stava in un banco <corr>cuando</corr> leiendo il giornale cuando ha visto un ragazzo con una ragazza sulla la testa . La ragazza non voleva andare con lui allora <anth>Gianni</anth> ce era molto forte mentre la ragazza gridava Lui la liberò del ragazzo bruto .La donna li ha ditto grazie a <anth>Gianni</anth> mentre il ragazzo stava in terra . Poi ,la donna semblava moleste ,ma io credo che lei stava enamorata di <corr>del</corr> <anth>Gianni</anth> . # Forse <anth>Gianni</anth> ha accompagnato alla ragazza a la sua casa e anno fatto una cena in sieme . <BODY>
```

Fig. 2. Esempio di testo trascritto in VALICO

I raccoglitori di testi sono incoraggiati a fornirci testi manoscritti, e soprattutto non corretti dall'insegnante; ultimamente però ci sono stati proposti anche testi in formato elettronico. Questi vanno comunque trattati, non si possono immettere tali e quali nel corpus: ad esempio se vogliamo rendere la punteggiatura un elemento interrogabile, bisogna isolare il segno di interpunzione dalla parola precedente. Si perdono tutte le autocorrezioni e alcuni apprendenti sottopongono il testo al controllo del correttore grammaticale del sistema di scrittura, per cui si hanno testi impeccabili nella grafia, corretti nell'accordo all'interno del contesto breve del sintagma o poco più, ma con errori di sintassi incongrui rispetto al livello dimostrato dalla correttezza ortografica e di accordo morfologico. Peraltro arrivano da paesi in cui non si usa l'alfabeto latino testi scritti a mano così ordinati da supporre che siano la bella copia di un testo precedente più tormentato. Per ora il sottoinsieme di testi con sospetto di correzione da computer non è tale da giustificare un campo di metadato nel sito in rete, ma siamo comunque in grado di individuare i testi arrivati in formato

elettronico attraverso le intestazioni archiviate, ossia controllando le informazioni <qualità>… </qualità> ed <esecuzione>…..</esecuzione>.

2.3 Finestra di interrogazione dei metadati

L'architettura di interrogazione dei due corpora VALICO e VINCA è stata disegnata da Simona Colombo, responsabile informatico del gruppo di ricerca e del sito attraverso cui si consultano i due corpora. In un primo momento si era pensato di adattare CWB *Corpus Workbench* (cf. Evert 2005, URL: http://cwb.sourceforge.net/), lo strumento di ricerca in corpora sviluppato per corpora molto grandi implementati con CQP e operante su una annotazione morfo-sintattica poco profonda, come illustrato da Heid (2009: 160-163). Colombo ha poi sviluppato, invece, un'interfaccia di interrogazione che combina CWB con una base di dati (si veda Colombo in corso di stampa).
Per quanto concerne ciò che l'utente vede, seguendo le indicazioni di Carla Marello e Elisa Corino, Simona Colombo ha disegnato la finestra di interrogazione dei metadati avendo cura di mettere i campi più spesso richiesti in una posizione "ergonomica". Una volta selezionate le variabili sociolinguistiche che si desiderano attraverso i vari menu a tendina che si srotolano a richiesta, cliccando su "Visualizza testi" si ottiene l'insieme dei testi che soddisfano i requisiti richiesti.

Per esempio, selezionando lingua madre spagnolo e consegna "amore" si ottengono 69 testi; posso chiedere solo i testi scritti da persone di una certa età e noterò che nel menu che si srotola quando seleziono "età" mi compaiono tre possibilità 19-25, 26-30 o "?" (ovvero dato non noto). Supponiamo di scegliere autori fra i 19-25 anni: ottengo 51 testi. Voglio isolare quelli raccolti in Italia, ad esempio a Torino: vado sul campo provenienza e trovo 4 luoghi di raccolta di cui due, Torino e Pollenzo, in Piemonte, uno in Spagna e uno in Francia (i soggiorni Erasmus!).

A fianco di ciascun testo dell'insieme trovato con "Visualizza testi" si trova un'icona cliccando sulla quale si vedono i dati dell'intestazione e quindi è possibile leggere altri dati circa l'autore di quel testo oltre a quelli prescelti. È possibile inoltre risalire ai testi dello stesso gruppo, cioè ai testi analoghi a quello estratto prodotti da tutti gli studenti di una stessa classe, nelle medesime condizioni, sotto la supervisione dello stesso insegnante.
Delle varie indicazioni date nella trascrizione del corpo dei testi la visualizzazione in rete riporta solo le inserzioni (le parole sottolineate) e le correzioni (le parole sottoscritte), perché sono quelle che vale la pena saltino all'occhio. Le versioni non accolte nella versione finale sono sottoscritte.

Fornitore	Lisa,Beltramo	Istituto	scuola
Trascrittore	Sara,Frattin	Nome Istituto	Università di Torino
Data	?	Gruppo	ierialparco
Luogo	Torino,IT	Specifiche	f
Lingua Madre	Spagnolo	Età	?
Lingue	?	Annualità	1
Scolarizzazione	un	Esposizione	sc
Permanenza	(5,Torino)	Consegna	amore_C.txt

001 Ieri al parco un uomo uommo che si chiama Gianni era andato a piedi a vedere i fiori . Dopo Dopo Lui stava in un banco , allora leiendo un giornale e subito vedevo un ragazzo stava giovani con una ragazza domandando socorso . Gianni si levanto e Ieri al parco Valentino un uomo che si chiama Gianni era andato a piedi a vedere i fiori perche lui stava stanco per el lavoro della settimana . Dopo , lui stava in un banco cuando leiendo il giornale cuando ha visto un ragazzo con una ragazza sulla la testa . La ragazza non voleva andare con lui allora Gianni ce era molto forte mentre la ragazza gridava Lui la liberò del ragazzo bruto . La donna li ha ditto grazie a Gianni mentre il ragazzo stava in terra . Poi , la donna semblava moleste , ma io credo che lei stava enamorata di del Gianni . # Forse Gianni ha accompagnato alla ragazza a la sua casa e anno fatto una cena in sieme .

Legenda: **Titolo** Inserzione *Variazione*

Testo Normale Testo Cancellato

Fig. 3. Intestazione e corpo del testo come appaiono in linea

Il testo dell'apprendente spagnola riportato nella fig. 1, grazie alla pazienza della trascrittrice, sarà quindi visualizzato come nella fig. 3. I nomi propri saranno recuperati invece in sede di interrogazione per ricerca linguistica.

2.4 Ricerca linguistica

La ricerca linguistica si può fare o direttamente su tutto il corpus nella sua interezza o sul sottoinsieme individuato tramite la selezione dei metadati e l'ordine "Visualizza testi". La finestra di interrogazione della ricerca linguistica è stata la prima ad essere pianificata, poiché di primario interesse per il gruppo di ricerca torinese e sviluppata non solo per VALICO e VINCA, ma per tutto l'insieme di corpora allestiti. Si prevedeva che i tipi di utenti di questi due corpora fossero più differenziati di quelli di altri corpora; che avrebbero potuto esser consultati anche da studenti e docenti non interessati a questioni di linguistica dei corpora. Tarando la finestra di ricerca linguistica sugli utenti di VALICO, si è reso un servizio anche agli utenti più esperti.
Adriano Allora ha disegnato una prima versione della finestra di interrogazione linguistica nel 2006 e tale versione è stata utilizzata fino all'inverno 2010, quando ne ha elaborato una versione migliorata, quella che si trova attualmente in rete e che è stata adottata anche per VINCA.
La spinta alla maggior trasparenza è venuta dalle richieste di studiosi e insegnanti di italiano che provavano ad interrogare VALICO, ma avevano difficoltà nello scrivere la sintassi di ricerca anche con l'interfaccia facilitata del 2006. Era poi necessario poter aggiungere delle migliorie di rappresentazione dei risultati trovati e creare una sinergia con VINCA che nel frattempo era stato messo in rete.

Il valore aggiunto della finestra di interrogazione attuale sono gli esempi/modelli di interrogazione; più vicini alle possibili ricerche di un insegnante, senza per questo cessare di essere interessanti per il linguista. Ciascun esempio si può copiare e incollare nella finestra di ricerca e cliccando su "Scopri come modificare questa ricerca" ci sono istruzioni per modificare secondo le proprie esigenze. L'elenco delle POS (*part of speech*), che l'etichettatore di CQP/CWB usa, è opportunamente messo a fianco.

I risultati appaiono della lunghezza che si è scelta e nella quantità scelta con o senza annotazioni POS, con o senza lemma a cui far risalire una determinata parola. Queste modalità di visualizzazione sono utili soprattutto per il linguista che si occupa di corpora e per quanti curano la loro manutenzione, ma sono anche una risorsa per il docente che si chieda se la parola *socorso* del testo precedente è ricondotta al lemma *soccorso*, se *bruto* è ricondotto al lemma *brutto* oppure no, o che si chieda come è classificato il *si levanto*[5]. Il fatto che *socorso* sia ricondotto al lemma *soccorso*, è utile perché significa che chi vuol sapere se gli apprendenti usano la parola *soccorso* troverà anche questo contesto, digitando la ricerca [lemma='soccorso']. Che *bruto* sia considerato un aggettivo, e non sia ricondotto a *brutto* dal lemmatizzatore di CWB, non è necessariamente la giusta interpretazione, ma il lemmatizzatore e l'annotatore automatici, benché allenati, hanno sempre molti problemi con varietà non standard di lingua.

[5] L'autrice voleva dire *si alzò* ma ha subìto l'interferenza dello spagnolo *levantarse*; l'annotatore automatico classifica *si* come pronome riflessivo e *levanto* come nome, rifacendosi forse al toponimo ligure.

Fra i miglioramenti apportati vi è anche la visualizzazione in formato KWIC dei risultati della ricerca in VALICO o in VINCA e la possibilità di esportarli (anche in tale formato). Questa possibilità è apprezzabile soprattutto quando si voglia evidenziare la preposizione che segue un verbo o gli oggetti diretti con cui si accompagna un verbo. Ecco i primi sei contesti dei 45 selezionati dalla ricerca [lemma='domandare'] modalità KWIC in VALICO con contesto limitato a una sola riga (si può chiedere di visualizzare anche un contesto di 2000 parole, che ovviamente è pensato per gli altri corpora del gruppo di ricerca torinese).

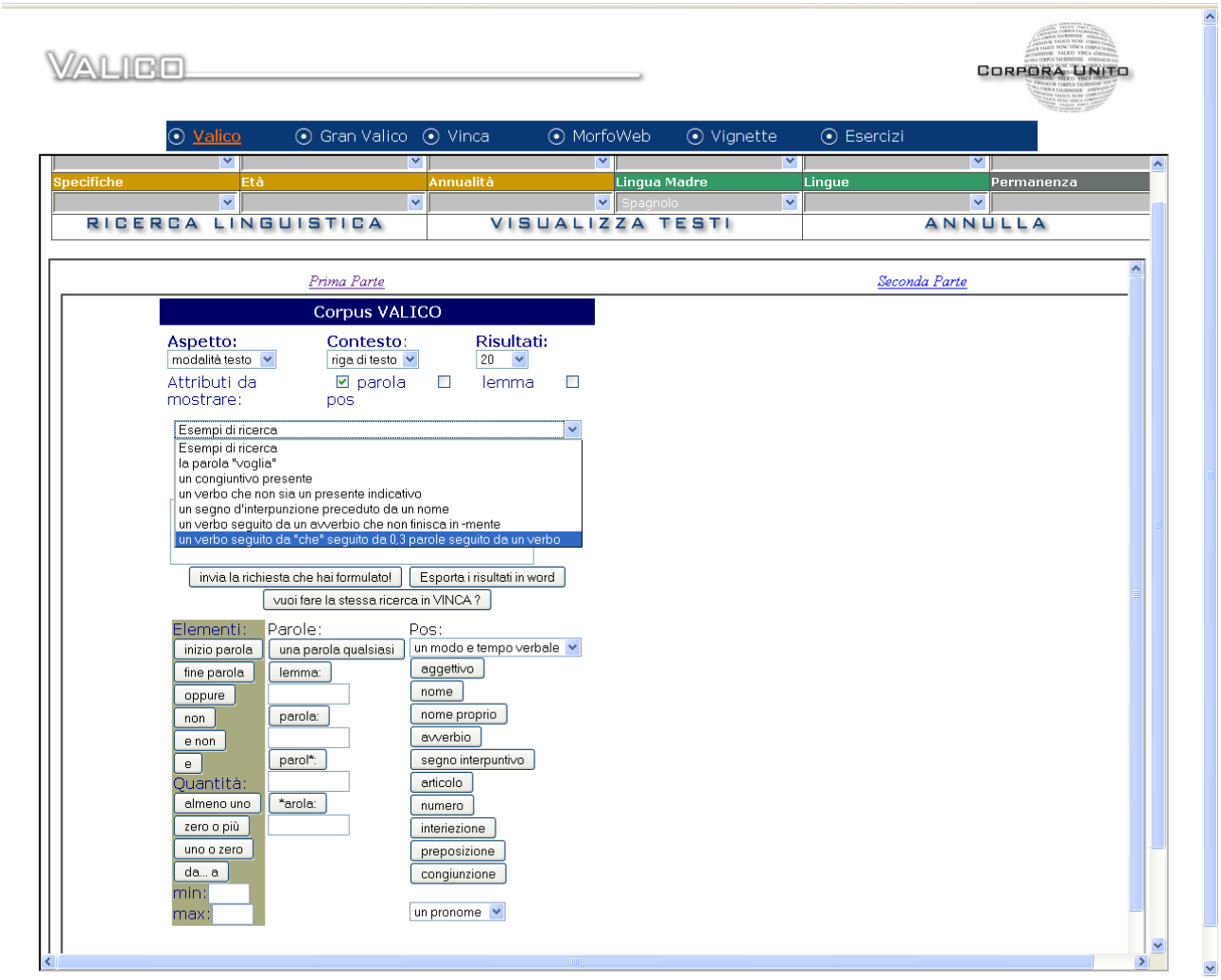

Fig. 4. Finestra di interrogazione linguistica in VALICO

1	59	quadro . E mi vienne una	domanda	perche , forse ce qualco
2	220	pigro . Leo non capì e si	domandò	che cosa era successo .
3	221	una birra . Il cameriere	domandò	il signore perchè era fu
4	485	prenduto una ragazza che	domandava	soccorso , oppure al men
5	490	n ha dito niente a la mia	domanda	e dopo io le ho dato un

Fig. 5. Cinque contesti dei 45 selezionati dalla ricerca [lemma='domandare'] in VALICO

Osserviamo ora gli stessi cinque contesti con la POS esplicitata e capiremo perché anche *la mia domanda* dell'esempio 5 venga pescata dalla ricerca. Gli omonimi sono sempre un problema in assenza di annotatori che agiscano anche con alberi sintattici.

1	59	O vienne/VER:remo una/DET	domanda/VER:pres	perche/ADV ,/PON forse/A
2	220	ER:remo e/CON si/PRO:refl	domandò/VER:remo	che/CON cosa/NOM era/VER
3	221	Il/DET:def cameriere/NOM	domandò/VER:remo	il/DET signore/NOM perch
4	485	a/DET ragazza/NOM che/CON	domandava/VER:impf	soccorso/NOM ,/PON oppur
5	490	f la/DET:def mia/PRO:poss	domanda/VER:pres	e/CON dopo/PRE io/PRO:pe

Fig. 6. I contesti della Fig. 5 con POS esplicitata

Particolarmente utile la possibilità di fare in VINCA la stessa ricerca che si è fatta in VALICO senza doverla riscrivere; un'apposita opzione "trasporta" l'utente nell'altro corpus e l'espressione regolare che stava usando in VALICO nella finestra di ricerca di VINCA, permettendo confronti rapidi fra una certa caratteristica negli scritti degli apprendenti stranieri e la stessa caratteristica negli scritti di italofoni.

2.5 Visualizzazione iconica e "geografica" dei risultati

In fase sperimentale avanzata è la visualizzazione iconica dei risultati resa possibile da software come Wordle per la creazione delle *word clouds*. Quasi tutti i portali istituzionali offrono ormai la possibilità di vedere quali sono le parole più cercate al loro interno non con numeri e percentuali o attraverso statistiche o grafici (almeno non solo), ma con un'immagine che rappresenta la parola più

frequentemente cercata, attribuendole un carattere con corpo proporzionalmente più grande di quello di una parola meno ricercata.

Poiché VALICO e VINCA partono dagli stessi stimoli iconici senza parole, si pensa sia interessante far visualizzare, ad esempio, quali sono le parole più usate da 50 italofoni nei loro testi e quali quelle usate da 50 non italofoni nel descrivere la stessa storia senza parole o quali quelle usate da stranieri nati in Europa rispetto a quelle di stranieri che studiano italiano in paesi dell'Asia.

Con la consultazione di Google Maps ci si abitua a vedere ciò che si cerca su una mappa: intendiamo far vedere le consegne "geograficamente" distribuite su una mappa in relazione ai luoghi di raccolta dei testi e stiamo valutando la possibilità di introdurre un espediente grafico che faccia cogliere immediatamente la quantità di testi raccolti in un luogo per ciascun tipo di consegna. Allo scopo è stato approntato un algoritmo proprietario di creazione dinamica di pagine kml[6] a partire dalle opzioni di interrogazione dei metadati scelte dall'utente fra quelli disponibili nella base di dati associata ai corpora.

3. Sviluppi ulteriori

Dal punto di vista del bilanciamento per consegna, per annualità di studio, per numero di testi di apprendenti di una certa lingua madre, per l'equilibrio fra testi raccolti in Italia e testi raccolti all'estero, c'è ancora lavoro da fare. D'altra parte la grande facilità con cui il ricercatore può ritagliarsi subcorpora omogenei all'interno di VALICO e VINCA compensa quest'aspetto. Non si esclude la possibilità di avere un VALICO orale, anche se la liberatoria per mettere in rete registrazioni di studenti è ancora più complessa da ottenere.

Il sito www.valico.org intende diventare non solo il sito attraverso cui consultare VALICO e VINCA, ma anche il luogo in cui trovare strumenti per mettere a frutto didatticamente le ricerche.

Già vi si trova il software Morfoweb (cf. Colombo 2009, Corino 2006); in futuro si vogliono mettere esempi di esercizi a scelta multipla creati a partire da errori presenti in VALICO (cf. Marello 2009) ed esempi di esercizi creati a partire da ricerche fatte sia in VALICO che in VINCA.

L'intento è non solo formare docenti di Italiano L2 o L1 facendo capire i vantaggi dell'uso di corpora d'apprendenti, ma anche servirsi dei due corpora per la didattica universitaria della linguistica applicata.

Bibliografia

Andorno C. / Rastelli S. (2009) (a cura di), *Corpora di italiano L2: tecnologie, metodi, spunti teorici*, Perugia, Guerra.

Colombo S. (2009),

[6] *Keyhole Markup Language*, un linguaggio basato su XML creato per gestire dati geospaziali in tre dimensioni nei programmi quali Google Earth.

Colombo S. (in corso di stampa), *Storia dell'architettura di VALICO*, in E. Corino, C. Marello (in corso di stampa)

Corino E. (2006), *MorFo. Morfemi fondamentali per capire l'italiano*. In F. Bosc, C.Marello, Mosca S. (a cura di) Saperi per insegnare, Torino, Loescher, pp. 285-297.

Corino E., Marello C. (in corso di stampa), *Italiano di apprendenti. I corpora VALICO e VINCA*, Guerra, Perugia.

Costantino M. (2009), *Transcript-o'-matic: la trascrizione dei testi per VALICO*. In: Valico: studi di linguistica e didattica a cura di Elisa Corino e Carla Marello, Perugia, Guerra Edizioni, 2009, pp. 167-176.

Evert S. (2005),

Heid U. (2009), *Metadata for learner corpora: a case study on VALICO*. In Valico: studi di linguistica e didattica a cura di Elisa Corino e Carla Marello, Perugia, Guerra Edizioni, 2009, pp. 151-165.

Marello C. (2009),

Palermo M. (2009),

Tosco A. (2010), *Autocorrezioni di apprendenti cinesi nel corpus gran VALICO*, in S. Rastelli, a cura di, Italiano di studenti cinesi. Percorsi di didattica acquisizionale, Guerra Edizioni, Perugia, 2010, pp. 123-132.

Vučo J. (2009), *Autocorrezioni di parlanti serbi che scrivono in italiano. Esempi nel corpus VALICO*, in: Valico: studi di linguistica e didattica a cura di Elisa Corino e Carla Marello, Perugia, Guerra Edizioni, 2009, pp. 137-150.