10.12871/CLICIT2014171

Corpus ICoN: una raccolta di elaborati di italiano L2 prodotti in ambito universitario

Mirko Tavosanis
Università di Pisa
Dipartimento di Filologia,
Letteratura e Linguistica
Via Santa Maria 36, 56126 Pisa
PI

tavosanis@ital.unipi.it

Abstract

Italiano. Il contributo presenta le caratteristiche essenziali del Corpus ICoN. Il corpus raccoglie elaborati realizzati nell'arco di 13 anni da studenti universitari; gli elaborati sono ripartiti in due sottocorpora equivalenti dedicati rispettivamente agli studenti che conoscono l'italiano come L1 e a quelli che lo conoscono come L2/LS.

English. *The paper describes the essential features of the Corpus ICoN. The corpus includes essays created over 13 years by university students; the essays are divided into two comparable subcorpora dedicated respectively to students who speak Italian as L1 and those who know the language as L2/FL.*

1 Introduzione

I corpora di testi realizzati come L2 sono da tempo uno strumento essenziale per lo studio dell'apprendimento delle lingue. Nel caso dell'italiano, tuttavia, anche se esistono prodotti importanti e ottimamente realizzati, il numero di corpora è ancora ritenuto insufficiente per molti tipi di ricerche (per una panoramica: Andorno e Rastelli 2009).

Il corpus in allestimento descritto qui di seguito mira a fornire in contributo in questo senso. Il lavoro si colloca all'interno delle attività del progetto PRIN "Scritture brevi" ed è previsto che il prodotto finale venga usato in primo luogo dall'Istituto di Linguistica Computazionale del CNR di Pisa per la messa a punto di strumenti di valutazione automatica dell'elaborato di apprendenti.

Il lavoro è attualmente ancora in corso. La conclusione delle attività è prevista per la fine del 2015, ma le caratteristiche complessive del corpus sono già ben definite e rendono quindi possibile una presentazione articolata.

2 Composizione del corpus

Il corpus è composto da circa 8000 elaborati complessivi. L'elaborazione e l'eliminazione delle irregolarità sono ancora in corso, ma le dimensioni del corpus finale sono al momento stimate in circa due milioni di token.

Il corpus si divide in due sottocorpora equivalenti tra loro come dimensione (circa un milione di token l'uno). Il primo è composto da elaborati realizzati da studenti che hanno l'italiano come L1; il secondo è composto da elaborati di studenti che conoscono l'italiano come LS/L2 e a un livello almeno pari al B2.

I due sottocorpora sono formati da testi realizzati dai relativi gruppi di studenti in circostanze identiche tra di loro. Ciò rende evidente la possibilità di un confronto tra i due corpora sul modello consolidato VALICO / VINCA.

3 Il Corso di Laurea ICoN

ICoN - Italian Culture on the Net – è un consorzio di università italiane (19, al momento della stesura di questo testo) che opera in collaborazione con il Ministero per gli Affari Esteri. Il Consorzio è stato fondato nel 1999 con il patro-

nato della Camera dei Deputati e con il supporto della Presidenza del Consiglio e del Ministero per l'Università e la Ricerca. Nella pratica, ICoN opera attraverso il proprio sito web, all'indirizzo: www.italicon.it (Tavosanis 2004).

Scopo del Consorzio è "promuovere e diffondere la lingua e la cultura italiana nel mondo" attraverso Internet e iniziative educative specifiche. Le attività mirate a questo scopo sono diverse, e includono per esempio la realizzazione di corsi di lingua e l'erogazione di Master universitari e corsi di aggiornamento. Il servizio più antico del Consorzio è però l'erogazione di un Corso di Laurea triennale in Lingua e cultura italiana per stranieri. Attivo dal 2001, il Corso di Laurea è erogato completamente via Internet ed è rivolto a due precise fasce di studenti: cittadini stranieri e cittadini italiani residenti all'estero. Di fatto, in oltre dieci anni di attività il Corso di Laurea ha avuto tra i propri iscritti un numero grosso modo equivalente di stranieri e di italiani (v. sezione 5). Dalle produzioni didattiche realizzate per il Corso è quindi possibile ricavare due corpora approssimativamente simili come estensione e del tutto comparabili come origine.

3.1 Criteri d'ammissione al Corso

I criteri di ammissione al Corso sono gli stessi di tutti i Corsi di Laurea delle università italiane. Per l'iscrizione è necessario possedere due requisiti: un titolo di studio che consenta l'accesso all'Università in Italia o nel paese di provenienza e una conoscenza della lingua italiana pari o superiore al livello B2.

3.2 Prove d'esame

Le prove scritte d'esame sono state svolte con modalità immutate fin dal primo anno accademico di operatività del Corso. Ogni corso all'interno del piano di studi si è quindi concluso con una prova scritta, che ogni studente ha dovuto realizzare al computer. Le prove si svolgono all'estero (e, in rari casi, in Italia, presso la sede del Consorzio) e sono composte da due parti. Lo studente deve infatti fornire le risposte a una batteria di trenta domande e scrivere un breve elaborato (descritto qui in dettaglio al punto 4). Per svolgere entrambi i compiti sono disponibili complessivamente 90 minuti, che ogni studente può dedicare all'una o all'altra parte nella proporzione che preferisce. Al termine del tempo stabilito il programma impedisce ulteriori modifiche; le prove vengono poi trasmesse in forma criptata alla sede centrale ICoN per la valutazione.

Durante le prove di esame gli studenti si trovano in ambienti controllati, in modo che non possano consultare libri o appunti, e il computer su cui operano è scollegato dalla rete fino al termine delle prove, in modo che sia impossibile fare riferimento a testi disponibili su Internet.

4 Gli elaborati

Gli elaborati del corso di laurea costituiscono il punto di partenza per la costituzione del corpus ICoN.

4.1 Caratteristiche dell'elaborato

La prova scritta è, in pratica, un piccolo tema. Il candidato può scegliere una traccia tra le tre alternative che gli sono proposte.

Esempi tipici di consegna sono:

Il restauro barocco di Maratta e il restauro ottocentesco di Cavalcaselle: metti a confronto due atteggiamenti diversi nei confronti della conservazione dell'opera d'arte.

Analizza il rapporto tra Petrarca e l'Umanesimo.

Illustra il concetto di equivalenza e il suo ruolo nella metodologia del confronto interlinguistico.

Il testo che segue è invece un esempio tipico di inizio di elaborato:

Raffaello Sanzio e' uno dei maggiori rappresentanti internazionali del Rinascimento italiano. Lui ha lavorato, come molti altri artisti famosi a quei tempi, in Roma - sede della Chiesa Cattolica e centro di grandi imprese artistiche com temi teologici. Uno dei posti din concentrazione dell'attivita' era il Vaticano, edificio accanto alla Bassilica di San Pietro. Giulio II era in ufficio nel momento in quale chiese a Raffaello di decorare le stanze private del papa, nel 1508. La prima stanza affrescata e' stata quella della Segantura. Qui le quattro pareti sono state divise usando il modello della divisione del sapere diritto, folosofia, poesia (al posto della medicina) e teologia. Cosi' si fa il percorso del bene, del vero e del bello.

Le caratteristiche testuali attese sono quelle degli elaborati prodotti in ambiente universitario. Nel sistema italiano, esami scritti di questo tipo

sembrano relativamente rari ma esistono anche nei corsi di laurea tradizionali.

4.2 Interfaccia

L'interfaccia di scrittura è formata da una finestra molto semplice. La finestra include un indicatore che mostra il tempo ancora disponibile per completare la prova e un contatore di caratteri che mostra la lunghezza dell'elaborato.

L'interfaccia non include invece strumenti avanzati di gestione del testo (cerca e sostituisci) e strumenti di formattazione (corsivi, grassetti e simili).

Inoltre, l'interfaccia non possiede funzioni di controllo ortografico. Questa caratteristica ha ovvie motivazioni didattiche ma si combina anche con un fattore esterno per produrre risultati linguisticamente interessanti. Poiché gli studenti sono residenti all'estero, le tastiere usate per scrivere le prove sono infatti solo raramente tastiere italiane. Ciò fa sé che spesso per gli studenti sia difficile inserire le lettere accentate. Esempi tipici di scrittura sono quindi:

La famiglia *e'* l'obiettivo principale, la vita professionale ancora non svolse per loro un ruolo di grande rilievo. Dagli anni settanta fino agli anni 90 *e'* diminuito il numero dei nuclei famigliari. Con il tempo si *noto'* sempre di *piu'* il processo della femminizzazione (iniziato *gia'* dopo I Guerra Mondiale), il quale ebbe un forte impatto sul nucleo famigliare.

Le commissioni d'esame sono a conoscenza della situazione e sono quindi invitate a ignorare deviazioni di questo genere dall'ortografia standard in tutti i casi in cui si può ragionevolmente ritenere che le ragioni a monte siano esclusivamente di questo tipo.

4.3 Archiviazione

Al termine delle prove gli elaborati vengono registrati all'interno di un file XML (la DTD di riferimento non è usata per validazioni in corso d'opera). Le prime righe dei file hanno di regola questo aspetto:

```
  <ESAMI idstudente="93969" nomestu-
dente="(eliminato)">
  <ESAME idnucleo="498" nome="Lingua
e letteratura latina/Letteratura la-
tina, medievale e umanistica 1 LET
A" status="F" datainizio="01/02/2011
08.09.28" datafine="01/02/2011
09.39.29" duratamax="90" ultimari-
sposta="00">
  <ESERCIZIO MODULEID="" unita=""
poolid="" testid="9440">
```

I file XML vengono poi criptati e inviati per posta elettronica alla sede operativa ICoN. A consegna avvenuta, i file vengono decriptati, controllati, caricati sul server di archiviazione e resi disponibili alle commissioni.

Le commissioni controllano eventuali anomalie nei test e valutano gli elaborati secondo griglie predefinite. La valutazione viene conservata su file separati.

5 Studenti

I dati anagrafici degli studenti rappresentano la prima fonte di informazione sociolinguistica per il corpus. La segreteria ICoN registra infatti le informazioni principali, incluse dichiarazioni sulla L1 e sulle L2 conosciute. Questi dati, opportunamente anonimizzati ai fini dell'analisi, sono poi raccordati ai singoli elaborati in modo da permettere la selezione dei testi attraverso diversi criteri.

Per la sostanza dei dati, va notato che la provenienza degli studenti è molto varia. I circa 300 laureati che hanno conseguito il titolo di studio entro l'estate del 2014 provengono infatti da 56 paesi diversi. Questa varietà segue una distribuzione spiccatamente da "coda lunga": i primi quattro paesi di provenienza dei laureati (Argentina, Germania, Brasile e Turchia) non solo corrispondono di regola a quattro diverse lingue madri ma forniscono complessivamente poco più di un quarto del totale dei laureati. Le L1 di origine sono quindi quasi altrettanto variate e tra gli studenti sono abbondantemente rappresentate lingue che vanno dal polacco al farsi.

Tuttavia, come accennato al punto 2, è possibile fare una distinzione molto forte tra due categorie di studenti: quelli che hanno l'italiano come L1 e quelli che invece lo conoscono come L2/LS.

5.1 Italiano come L1

In generale, si può dare per scontata la conoscenza dell'italiano a livello madrelingua da parte dei cittadini italiani che abbiano compiuto buona parte del proprio percorso formativo in Italia. Nelle scritture degli studenti residenti da molto tempo all'estero, però, sono presenti occasionalmente esempi di erosione dell'italiano o di interferenza da parte delle L2.

5.2 Italiano come L2/LS

Il livello degli studenti stranieri è molto variabile. Sebbene tutti siano accomunati da una conoscenza della lingua di livello almeno B2, la differenza tra studenti con conoscenza appena sufficiente e studenti con competenze assimilabili a L1 è molto vistosa.

Gestire questa diversità è sicuramente una delle sfide principali nell'elaborazione del corpus. In una prima fase, l'assegnazione del livello di competenze dovrà essere fatta interamente da valutatori umani. In una seconda fase, è possibile che l'operazione possa essere condotta in parte in modo automatico.

Nel corpus definitivo gli studenti che hanno avuto contatti con l'italiano come L2 saranno distinti da quelli per cui l'italiano è stato solo LS. In entrambi i casi inoltre, compatibilmente con la documentazione disponibile, saranno etichettati gli studenti che per vari motivi (origine familiare, ambiente, trasferimenti in Italia) hanno avuto un contatto con l'italiano diverso da quanto normalmente prevedibile per una L2/LS. La granularità di questa etichettatura non è ancora stata definita; soprattutto per gli studenti degli ultimi anni, che come parte della procedura di iscrizione forniscono spesso lettere di motivazione e descrizioni dei propri contatti con la lingua e la cultura italiana, è possibile che possa essere realizzata una descrizione molto dettagliata, probabilmente presentata sotto forma di testo articolato.

6 Inserimento degli elaborati all'interno del corpus

Il corpus prevede che gli elaborati vengano importati come testo semplice con codifica UTF-8. Un punto delicato è la gestione degli errori ortografici collegati a tastiere non italiane e descritti a 4.2. Tuttavia i primi esempi di analisi, condotti con il sistema READ-IT dell'Istituto di Linguistica Computazionale del CNR di Pisa (Dell'Orletta, Montemagni e Venturi 2011), mostrano che gli strumenti oggi esistenti possono ricondurre senza problemi gli errori di questo genere alle forme target, senza che sia necessaria neanche una fase di addestramento.

7 Distribuzione del corpus

Il prodotto finito sarà reso disponibile in forma mediata. Per ragioni collegate alla natura del corpus non sarà quindi possibile il libero scaricamento degli elaborati o il loro collegamento a consegne. Si prevede che la ricerca avvenga attraverso un'interfaccia web e che la dimensione dei contesti venga limitata, in alternativa, ai confini di frase o a un massimo di 300 caratteri.

8 Conclusioni

L'elaborazione del corpus è ancora in corso. Tuttavia, gli assaggi eseguiti fino a questo momento sono molto promettenti e rassicurano sull'utilità del progetto. Di particolare valore sembra la possibilità di confrontare i testi prodotti da studenti che hanno o meno l'italiano come L1 in circostanze in cui i fini comunicativi rispondono a una precisa realtà didattica.

Bibliografia

Cecilia Andorno - Stefano Rastelli, *Corpora di italiano L2. Tecnologie, metodi, spunti teorici*, Perugia: Guerra, 2009.

Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi, *Read-it: Assessing Readability of Italian texts with a View to Text Simplification*. In: *Proceedings of the 2nd Workshop on Speech and Language processing for Assistive Technologies*, Edinburgh, 2011, pp. 73-83.

Mirko Tavosanis, *L'italiano del web*, Roma: Carocci, 2011.

Mirko Tavosanis, *Insegnamento di lingua e cultura italiana a stranieri: l'esperienza di ICoN*. In: *Italiano e italiani nel mondo. Italiani all'estero e stranieri in Italia: identità linguistiche e culturali*. Vol. 1, Roma: Bulzoni, pp. 1-13.