IL *CORPUS* CELI: UNA NUOVA RISORSA PER STUDIARE L'ACQUISIZIONE DELL'ITALIANO L2

Stefania Spina, *Irene Fioravanti*, *Luciana Forti*, *Valentino Santucci*, *Angela Scerra*, *Fabio Zanda*[1]

1. INTRODUZIONE: LA LEARNER *CORPUS* RESEARCH

Quali sono i tratti distintivi dell'italiano L2 nei diversi livelli del *Quadro Comune Europeo di Riferimento per le Lingue* (QCER)? In che modo si sviluppa la competenza linguistica nell'apprendimento dell'italiano L2, rispetto alla progressione tra un livello e un altro?

Secondo la teoria dei sistemi complessi, l'apprendimento linguistico è caratterizzato da dinamicità, adattabilità e non-linearità (Larsen-Freeman, Cameron, 2009; Verspoor, 2017). Tra le varie tipologie di dati che possiamo elicitare per indagare tali proprietà (Mackey, Gass, 2012) troviamo i dati estratti da *corpora*, raccolte sistematiche di dati linguistici autentici, adatti al trattamento informatico e rappresentativi di una lingua o varietà (McEnery *et al.*, 2006: 6). In particolare, alla costruzione di *corpora* contenenti testi prodotti da apprendenti (i cosiddetti *learner corpora*) è stato dato un impulso notevolissimo negli ultimi decenni, al punto tale da definire una nuova disciplina: la *Learner Corpus Research* (LCR). Dell'autonomia di tale disciplina abbiamo contezza grazie a elementi diversi, tutti susseguitisi negli ultimi anni ma che rappresentano il culmine di un percorso di ricerca lungo almeno tre decenni.

Il primo volume sulle potenzialità dei *corpora* nell'ambito della linguistica acquisizionale e della glottodidattica appare nel 1998. Si tratta di *Learner English on Computer* ed è curato da Sylviane Granger (Granger, 1998). Raccoglie contributi su come creare un *learner corpus* e su come analizzarne il contenuto. Il volume riporta anche alcuni dei primi risultati ottenuti nelle analisi sulla prima versione dell'ICLE, *International Corpus of Learner English*, progettato per raccogliere le produzioni scritte di apprendenti di inglese L2, di livello intermedio/avanzato, con diverse L1 (Granger *et al.*, 2020). Il progetto nasce inizialmente per colmare una lacuna dell'ICE – *International Corpus of English*, avviato alla fine degli anni '80 (Greenbaum, Nelson, 1996), che non aveva previsto l'inclusione di testi prodotti da non nativi. In breve tempo, diventa il punto di riferimento per la ricerca acquisizionale basata su *corpora*, e determina la nascita prima e il consolidamento dopo della LCR. Pur accogliendo studiosi dedicati all'apprendimento dell'inglese come L2, il CECL – *Centre for English Corpus Linguistics*, fondato presso l'Université catholique de Louvain (Louvain-la-Neuve, Belgio) da Sylviane Granger, diventa in breve tempo il cuore propulsore della LCR e, negli anni, estende il proprio campo di interesse anche a temi relativi all'acquisizione di lingue diverse dall'inglese. Viene istituito, così, il convegno biennale internazionale di *Learner Corpus Research*, svoltosi nella sua prima edizione nel 2011 a Louvain-la-Neuve, in

[1] Università per Stranieri di Perugia.

Tutti gli autori hanno contribuito alla realizzazione del *corpus* che qui viene presentato, e hanno revisionato il testo dell'articolo nella sua globalità. Più in particolare, Stefania Spina è autrice dei §§ 4, 5.1 e 6; Irene Fioravanti e Fabio Zanda dei §§ 2 e 3; e Luciana Forti dei §§ 1 e 5.2.

celebrazione del ventennale dalla fondazione del CECL. Da allora, lo svolgimento del convegno avviene regolarmente ogni due anni, con successiva pubblicazione degli Atti.

È del 2015, inoltre, la pubblicazione di *The Cambridge Handbook of Learner Corpus Research*, curato da Sylviane Granger, Gaëtanelle Gilquin e Fanny Meunier, contenente capitoli fondanti su come costruire un *corpus* di apprendenti, come analizzarne il contenuto, e come porre il frutto di tali analisi in relazione con la linguistica acquisizionale e la glottodidattica (Granger *et al.*, 2015). Sempre nel 2015, appare il primo numero della rivista *International Journal of Learner Corpus Research*[2], edita dalla casa editrice Benjamins, con due uscite annue. Nel 2018, viene poi fondata la *Learner Corpus Association*[3] (LCA), che fissa, nel proprio documento costitutivo, i seguenti obiettivi:

a) to promote the field of learner *corpus* research by supporting:
   - the compilation of learner *corpora* in a wide range of languages, where possible making this material available to researchers and institutions;
   - the design of innovative methods and tools to analyse the data;
   - the application of the research to relevant domains;
   - attempts to link up learner *corpus* research to neighbouring fields, such as Second Language Acquisition theory, Language Testing and Assessment, Natural Language Processing, Foreign/Second Language Teaching and linguistic theory in general;
   - the wider dissemination of findings to the broader scientific community.

b) to provide a forum for research and discussion on learner *corpus* research, in particular by maintaining a dedicated website known as the LCA website and initiating the biennial Learner *Corpus* Research (LCR) conferences[4].

L'Associazione cura la *Learner Corpus Bibliography*, disponibile anche come collezione Zotero liberamente accessibile[5], e l'archivio *Learner Corpora around the world*, che fornisce un inventario completo e sistematico dei *learner corpora* realizzati in tutto il mondo, in riferimento a qualsiasi lingua. Raccoglie e diffonde, inoltre, eventi legati alla LCR e alle discipline affini. Infine, il 2021 ha visto lo svolgimento della prima *Graduate Student Conference in Learner Corpus Research*[6], organizzata dalla Faculty of Education dell'Inland Norway University of Applied Sciences. L'intento dell'iniziativa è stato quello di creare uno spazio di confronto e di crescita per dottorandi e studenti di magistrale con progetti di ricerca inerenti alla LCR e, più ampiamente, alle intersezioni tra linguistica dei *corpora*, linguistica acquisizionale e glottodidattica.

Sul piano internazionale, dunque, la LCR si identifica sempre di più come ambito di studio ben definito, con un centro di ricerca iniziatore, un convegno, una rivista e un'associazione dedicati, e, da poco, anche un convegno specificamente pensato per dottorandi e studenti di magistrale. Com'è la situazione in Italia? Quali e quante sono state, finora, le esperienze di costruzione di *corpora* per studiare l'italiano L2?

Una ricostruzione del panorama italiano, relativo agli studi sull'italiano L2, viene presentata da Stefania Spina in occasione della sua comunicazione plenaria nell'ambito della LCR conference del 2017. La ricostruzione si focalizza sui *corpora* di italiano L2 realizzati sino a quel momento, partendo dalla prima, importante esperienza di raccolta di dati empirici: la banca dati del Progetto di Pavia (Chini, 2016; Giacalone Ramat, 2003). Il

[2] Pagina web della rivista: https://benjamins.com/catalog/ijlcr.
[3] Sito web dell'Associazione: https://www.learner*corpus*association.org/.
[4] LCA constitution: https://www.learner*corpus*association.org/about/constitution/.
[5] La bibliografia è accessibile alla seguente pagina:
https://uclouvain.be/en/research-institutes/ilc/cecl/learner-*corpus*-bibliography.html.
[6]https://eng.inn.no/conferences/the-graduate-student-conference-in-learner-*corpus*-research-2021.

progetto viene definito come pionieristico dal momento che, per la prima volta in contesto italiano, nella seconda metà degli anni '80 del secolo scorso, raccoglie dati empirici con la finalità di descrivere l'interlingua di apprendenti di italiano L2 sviluppatasi in contesto spontaneo. Grazie a questa banca dati, è stato possibile descrivere aspetti dell'interlingua dell'italiano L2 relativi alla morfologia nominale e verbale, alla sintassi e alla testualità. A tale analisi empirica viene poi attribuito un ruolo centrale rispetto alla caratterizzazione dei percorsi glottodidattici pensati per apprendenti di italiano L2.

Tuttavia, è soltanto a partire dal 2009 che assistiamo alla pubblicazione dei primi *corpora* di italiano L2, di dimensioni più ampie e liberamente accessibili in rete. I criteri di costruzione che tali *corpora* adottano assumono una prospettiva trasversale, nel caso di testi raccolti in un singolo momento da un campione di apprendenti omogeneo, oppure una prospettiva longitudinale, nel caso di testi raccolti in più momenti nel tempo da un singolo campione di apprendenti, oppure ancora una prospettiva pseudo-longitudinale, nel caso di testi raccolti da campioni di apprendenti di diversi livelli di competenza. Consultando il database *Learner Corpora around the world*, infatti, troviamo sei *corpora* contenenti testi prodotti esclusivamente da apprendenti di italiano L2, e tre contenenti testi prodotti anche da apprendenti di lingue target diverse. Tra i primi, troviamo quelli che seguono:

- il LIPS (*Lessico dell'italiano parlato da stranieri*), un *corpus* con impianto longitudinale composto da 1426 testi orali, prodotti da 700 apprendenti, di livelli compresi tra l'A1 e il C2, con lingue materne varie, per un totale di circa 670.000 *token*, realizzato dall'Università per Stranieri di Siena (Gallina, 2015)[7];
- il VALICO (*Varietà Apprendimento Lingua Italiana Corpus Online*), un *corpus* trasversale di 2502 testi, prodotti da altrettanti apprendenti, di livelli vari definiti in base alle annualità di studio della lingua italiana, con lingue materne varie, per un totale di circa 382.000 *token*, realizzato nel 2003 e messo in rete nel 2009 dall'Università degli Studi di Torino (Corino *et al.*, 2017)[8];
- il COLI (*Corpus of Chinese Learners of Italian*), un *corpus* pseudo-longitudinale di testi scritti e orali, realizzati da 30 apprendenti, di lingua materna cinese, distribuiti equamente nei tre livelli B1, B2 e C1, per un totale di 82.300 *token*, realizzato e reso interrogabile in rete nel 2009 dall'Università per Stranieri di Perugia[9];
- il CAIL2 (*Corpus di Apprendenti di Italiano L2*), un *corpus* trasversale di 400 testi scritti, composti da altrettanti apprendenti di livello intermedio-avanzato, con lingue materne varie, per un totale di circa 237.000 *token*, pubblicato nel 2015 dall'Università per Stranieri di Perugia (Bratankova, 2015);
- il LOCCLI (*The Longitudinal Corpus of Chinese Learners of Italian*), un *corpus* longitudinale di 350 testi scritti, composti da 175 apprendenti di lingua materna cinese, in due momenti diversi distanziati di sei mesi, per un totale di 97.000 *token*, pubblicato nel 2015 dall'Università per Stranieri di Perugia (Spina, Siyanova-Chanturia, 2018);
- il CORITE (*Corpus del Italiano de los Españoles*), un *corpus* in parte longitudinale e in parte pseudo-longitudinale di 385 testi scritti, realizzati in vari punti temporali da 90 apprendenti, con livelli di competenza compresi tra l'A1 e il B2, per un totale di

[7] http://www.parlaritaliano.it/index.php/it/*corpora*-di-parlato/653-*corpus*-lips
[8] http://www.valico.org/valico.html
[9] https://www.unistrapg.it/cqpwebnew/ (su questo portale si trovano tutti i *corpora* realizzati presso l'Università per Stranieri di Perugia).

103.147 *token*, realizzato presso l'Università Cattolica del S. Cuore di Milano nel 2018 (Ballini, Frigerio, 2018)[10].

Nel secondo gruppo di *corpora*, contenenti testi di più lingue target, troviamo quelli che seguono:

- il MERLIN (*Multilingual Platform for the European Reference Levels: interlanguage exploration in context*), un *corpus* pseudo-longitudinale di 813 testi scritti, prodotti da altrettanti apprendenti, di lingue materne varie, distribuiti in modo bilanciato nei livelli compresi tra A1 e B1, per un totale di circa 107.000 *token*, pubblicato dal centro di ricerca Eurac di Bolzano nel 2014 nell'ambito di un progetto UE (Wisniewski et al. 2013)[11];
- il LEONIDE (*The Longitudinal LEarner Corpus iN Italiano, Deutsch, English*), un *corpus* longitudinale di circa 2.500 testi scritti prodotti da 163 studenti bilingui (italiano e tedesco) delle scuole superiori, nell'arco di tre anni scolastici (2015-2018), per un totale di 237.000 *token*, realizzato dal centro di ricerca Eurac di Bolzano (Glaznieks *et al.*, 2022)[12];
- il KOLIPSI, un *corpus* di testi scritti raccolto nell'arco di due anni scolastici: il 2007/8 e il 2014/2015. Comprende circa 4.000 testi, prodotti da circa 2.000 studenti, per un totale di circa 800.000 *token*. Il *corpus* è stato realizzato dal centro di ricerca Eurac di Bolzano (Glaznieks *et al.*, in preparazione)[13].

Nel quadro di tale contesto, questo contributo presenta un nuovo *learner corpus* per lo studio dell'acquisizione dell'italiano L2: il *corpus* CELI. Realizzato nell'ambito del progetto PRIN 2017 *PHRAME – Misure di complessità fraseologica in italiano L2. Integrazione di eye tracking, corpora e metodologie computazionali per la creazione di risorse finalizzate all'apprendimento di una seconda lingua,* il *corpus* CELI si compone di testi scritti elicitati in ambito certificatorio, configurandosi dunque come risorsa preziosa per la caratterizzazione sistematica dell'interlingua di apprendenti di italiano L2 a diversi livelli di competenza. Nella prossima sezione, descriviamo il metodo con cui il *corpus* è stato costruito.

2. METODO

2.1. *Criteri di costruzione del corpus*

Il *corpus* si compone di produzioni scritte tratte dagli esami di certificazione linguistica per l'italiano generale CELI (Certificati di Lingua Italiana) destinati ad un pubblico generico di adulti scolarizzati, elaborati e somministrati dal Centro per la Valutazione e le Certificazioni Linguistiche (CVCL[14]) dell'Università per Stranieri di Perugia. I testi sono stati estratti dalle prove d'esame CELI 2, CELI 3, CELI 4 e CELI 5, che attestano la conoscenza della lingua italiana rispettivamente nei livelli B1, B2, C1 e C2 del QCER. La scelta di selezionare questi quattro livelli di certificazione è dipesa dall'obiettivo di raccogliere un *corpus* rappresentativo dell'italiano scritto di apprendenti di italiano di livello intermedio e avanzato.

[10] https://corespiycorite.altervista.org/presentazione/
[11] https://merlin-platform.eu/index.php.
[12] https://www.porta.eurac.edu/lci/leonide/.
[13] https://www.porta.eurac.edu/lci/kolipsi-family/.
[14] https://www.cvcl.it/home-cvcl.

Il *corpus* è stato realizzato adottando i seguenti criteri:

1. la tipologia dei compiti della Prova di Produzione di testi scritti degli esami CELI;
2. la nazionalità del centro d'esame CELI;
3. il punteggio totale della prova d'esame CELI;
4. il punteggio totale assegnato alla produzione scritta.

Nei sottoparagrafi seguenti verrà approfondito ciascuno dei quattro criteri adottati nella realizzazione del *corpus* CELI.

2.1.1. *Criterio 1: la tipologia dei compiti della Prova di Produzione di testi scritti degli esami CELI*

La struttura di tutti gli esami CELI prevede una *Prova Scritta* e una *Prova Orale*. La Prova Scritta è costituita da diverse componenti volte a verificare differenti abilità: nel caso dell'esame CELI 2 si articola in tre "sotto-prove" (Parte A, Parte B e Parte C), e nel caso del CELI 3, CELI 4 e CELI 5 in quattro "sotto-prove" (Parte A, Parte B, Parte C e Parte D) (Tabella 1). Nello specifico, la Prova Scritta del CELI 2[15] è composta da una Prova di Comprensione della lettura (Parte A), una Prova di Produzione di testi scritti (Parte B) e una Prova di Comprensione dell'ascolto (Parte C), mentre l'omonima Prova Scritta del CELI 3[16], CELI 4[17] e CELI 5[18] è invece composta da una Prova di Comprensione della lettura (Parte A), una Prova di Produzione di testi scritti (Parte B), una Prova di Competenza linguistica (Parte C, assente nel CELI 2) e una Prova di Comprensione dell'ascolto (Parte D). Nella scelta dei testi da inserire nel *corpus* CELI sono stati considerati i compiti relativi esclusivamente alla Prova di Produzione di testi scritti (Parte B in tutti i livelli prescelti).

Tabella 1. *Struttura degli esami CELI 2, CELI 3, CELI 4 e CELI 5*

CELI 2	**Sotto-prove**
PROVA SCRITTA	Prova di Comprensione della lettura (Parte A)
	Prova di Produzione di testi scritti (Parte B)
	Prova di Comprensione dell'ascolto (Parte C)
PROVA ORALE	Prova di Produzione orale

CELI 3, 4 e 5	**Sotto-prove**
PROVA SCRITTA	Prova di Comprensione della lettura (Parte A)
	Prova di Produzione di testi scritti (Parte B)
	Prova di Competenza linguistica (Parte C)
	Prova di Comprensione dell'ascolto (Parte D)
PROVA ORALE	Prova di Produzione orale

[15] https://www.unistrapg.it/sites/default/files/docs/certificazioni/celi-2-descrizione-prova.pdf.
[16] https://www.unistrapg.it/sites/default/files/docs/certificazioni/celi-3-descrizione-prova.pdf.
[17] https://www.unistrapg.it/sites/default/files/docs/certificazioni/celi-4-descrizione-prova.pdf.
[18] https://www.unistrapg.it/sites/default/files/docs/certificazioni/celi-5-descrizione-prova.pdf.

La Prova di Produzione di testi scritti (Tabella 2) si articola in: tre compiti di produzione scritta per l'esame di certificazione CELI 2 (*task* B.1: un modulo/questionario su argomenti di interesse generale; *task* B.2: un breve annuncio da scrivere o a cui rispondere su un argomento dato di vita quotidiana; *task* B.3: una breve lettera o e-mail da sviluppare da una traccia predefinita); due compiti per l'esame di certificazione CELI 3 (*task* B.1: una composizione su esperienze personali, situazioni, temi e argomenti di interesse generale; *task* B.2: una lettera/e-mail fortemente contestualizzata con specifici obiettivi comunicativi); due compiti per l'esame di certificazione CELI 4 (*task* B.1: un riassunto di un testo; *task* B.2: una composizione con traccia a scelta che può di volta in volta essere una relazione su problemi e fenomeni della società odierna, un racconto su avvenimenti ed esperienze personali o una lettera formale); due compiti anche per l'esame di certificazione CELI 5 (*task* B.1: una composizione da scegliere tra relazione di un saggio, racconto di fantasia o descrizione di un'esperienza personale; *task* B.2: due lettere/e-mail formali con destinatari e obiettivi definiti) (Grego Bolli e Pelliccia, 2005).

Tabella 2. *Struttura della Prova di Produzione di testi scritti degli esami CELI 2, 3, 4 e 5*

CELI 2	***TASK***	**TIPOLOGIA**
PARTE B: Prova di Produzione di testi scritti	B.1	un modulo/questionario a cui rispondere su temi e argomenti di interesse generale
	B.2	un breve annuncio da scrivere o a cui rispondere su un argomento dato di vita quotidiana (circa 50 parole)
	B.3	**una breve lettera o e-mail da scrivere, seguendo una traccia data** (dalle 90 alle 100 parole)

CELI 3	***TASK***	**TIPOLOGIA**
PARTE B: Prova di Produzione di testi scritti	**B.1**	**una breve composizione su esperienze personali, situazioni, temi e argomenti di interesse generale da scegliere tra due diversi input** (dalle 120 alle 180 parole)
	B.2	una lettera o e-mail fortemente contestualizzata con obiettivi comunicativi diversi: chiedere o dare informazioni, consigli, esprimere opinioni ecc., da scegliere tra tre diversi input (dalle 80 alle 100 parole)

CELI 4	***TASK***	**TIPOLOGIA**
PARTE B: Prova di Produzione di testi scritti	B.1	un testo da riassumere tenendo conto delle informazioni fornite in una traccia (dalle 150 alle 200 parole)
	B.2	**una composizione da scegliere tra due diversi input che possono riguardare una relazione su problemi e fenomeni della società odierna, un racconto su avvenimenti ed esperienze personali o una lettera formale** (dalle 220 alle 250 parole)

CELI 5	*TASK*	TIPOLOGIA
PARTE B: Prova di Produzione di testi scritti	**B.1**	**una composizione libera da scegliere tra tre diversi input che possono riguardare una relazione o un saggio, un racconto di fantasia o una descrizione di esperienze personali anche in relazione ad aspetti della civiltà italiana** (dalle 330 alle 360 parole)
	B.2	una composizione da scegliere tra due diversi input che possono riguardare una relazione su problemi e fenomeni della società odierna, un racconto su avvenimenti ed esperienze personali o una lettera formale (dalle 220 alle 250 parole)

Nello specifico, sono stati selezionati e inclusi nel *corpus* i testi prodotti dai candidati dei quattro livelli di certificazione per rispondere ai seguenti compiti: scrivere una breve lettera/e-mail per il CELI 2 (*task* B.3); una breve composizione su esperienze personali/interessi generali per il CELI 3 (*task* B.1); una composizione su aspetti della società/racconto di esperienze personali per il CELI 4 (*task* B.2); e una relazione di un saggio/racconto di fantasia/descrizione di un'esperienza personale per il CELI 5 (*task* B.1). Ciò ha garantito l'inclusione nel *corpus* di testi non troppo difformi tra loro per ogni livello (numero minimo e massimo di parole da utilizzare specificato), maggiormente articolati e di lunghezza (in parole) più ampia rispetto a quelli prodotti per rispondere alle altre tipologie di compiti proposti nella Prova di produzione scritta (Grego Bolli, Spiti, 2004).

2.1.2. *Criterio 2: la nazionalità del centro d'esame CELI*

I centri d'esame CELI sono distribuiti non solo su suolo italiano, ma anche europeo ed extra-europeo[19]. Tenendo conto di questo dato, sono state incluse nel *corpus* prove d'esame sostenute in centri d'esame CELI sia italiani sia esteri, in modo da poter ottenere una distribuzione il più possibile omogenea della nazionalità dei candidati. Infatti, nella maggioranza dei casi, la nazionalità delle sedi d'esame all'estero combaciava con la nazionalità dei candidati di cui sono stati trascritti i testi e ciò ha contribuito a mantenere, per quanto possibile, un certo equilibrio in termini di bilanciamento per nazionalità degli autori dei testi inclusi nel *corpus*.

2.1.3. *Criteri 3 e 4: il punteggio totale della prova d'esame CELI e il punteggio totale assegnato alla Prova Scritta*

I testi sono stati inclusi solo se il candidato aveva superato l'intera prova d'esame (i.e., aveva ottenuto almeno il punteggio minimo sufficiente sia nella Prova Scritta sia nella Prova Orale) nella stessa sessione d'esame, e solo se il punteggio assegnato alla specifica produzione scritta raggiungeva almeno la sufficienza. Sono stati adottati questi criteri per ragioni di omogeneità tra i candidati che avevano ottenuto la certificazione linguistica per un dato livello.

[19] L'elenco completo dei centri d'esame CELI è disponibile alla seguente pagina: https://www.unistrapg.it/it/certificati-di-conoscenza-della-lingua-italiana/centri-d-esame-celi.

Riassumendo, il *corpus* è stato costruito con l'obiettivo di rappresentare l'italiano scritto di apprendenti di italiano di livello intermedio e avanzato. Sono state quindi selezionate le produzioni scritte dalle quattro prove d'esame CELI (da CELI 2 a CELI 5), che certificano la competenza in italiano nei livelli B1, B2, C1 e C2. All'interno di ciascuna Prova di Produzione di testi scritti, sono state selezionate le produzioni più articolate e più ampie in termini di lunghezza in parole. Sono stati inclusi testi tratti da prove d'esame sostenute nei centri CELI sia italiani sia esteri per ottenere una distribuzione il più possibile omogenea della nazionalità dei candidati. Infine, i testi sono stati inclusi nel *corpus* solo se il candidato aveva ottenuto, nella medesima sessione d'esame, la sufficienza nella specifica produzione scritta, nella cosiddetta Prova Scritta e nell'intera prova d'esame.

2.2. *Criteri di trascrizione dei testi*

Le prove d'esame CELI sono conservate in formato cartaceo presso l'archivio del CVCL dell'Università per Stranieri di Perugia. La selezione dei testi è avvenuta tramite consultazione dell'archivio e le produzioni scritte sono state raccolte seguendo i criteri esposti nel paragrafo precedente. Ogni manoscritto è stato quindi trascritto manualmente nel database digitale di raccolta sulla piattaforma LOL (Learning OnLine) dell'Università per Stranieri di Perugia. Oltre alle trascrizioni dei manoscritti, nel database sono stati inseriti anche i metadati relativi a ogni testo trascritto.

Il principale criterio di trascrizione adottato è stato quello di trascrivere la produzione scritta del candidato il più fedelmente possibile al testo originale. Tuttavia, in un numero limitato di casi sono state normalizzate le forme prodotte dall'apprendente per limitare gli errori nella successiva fase di annotazione per categoria grammaticale (*pos-tagging*; vedi paragrafo 4; Haan, 2000).

È stata eseguita una normalizzazione delle forme prodotte dagli apprendenti nei seguenti casi:

1) Normalizzazione degli errori ortografici: la forma prodotta dall'apprendente conteneva un errore di raddoppiamento o di mancato raddoppiamento o mancanza di accento. Esempi:
   a. Forma originale: **racontano* > Forma normalizzata: *raccontano*
   b. Forma originale: **amicizzia* > Forma normalizzata: *amicizia*
   c. Forma originale: **perche* > Forma normalizzata: *perché*

2) Normalizzazione degli omografi: l'apprendente ha prodotto una forma scorretta con determinata categoria grammaticale derivabile dal contesto, ma omografa a un'altra forma di diversa categoria grammaticale e che poteva essere confusa nella fase di *pos-tagging*. Esempi:
   a. Forma originale: **qui* > Forma normalizzata: *cui* (e viceversa)
   b. Forma originale: **quanto* > Forma normalizzata: *quando* (e viceversa)
   c. Forma originale: **o visitato* > Forma normalizzata: *ho visitato*

3) Normalizzazione degli errori fonografici: la forma prodotta dall'apprendente conteneva un errore fonografico a livello di suono vocalico o consonantico. Esempi:
   a. Forma originale: **dicisamente* > Forma normalizzata: *decisamente*
   b. Forma originale: **ceremonia* > Forma normalizzata: *cerimonia*
   c. Forma originale: **encoraggia* > Forma normalizzata: *incoraggia*

Al contrario, non sono state normalizzate le forme che mostravano una plausibile influenza della L1 del candidato (es.: «Una *lenda antiqua* ma molto interessante»*;* «Com'è possibile che i *caschi polari* siano quasi spariti?»; «Qui è difficile avere dei veri veri amici, ma è facile fare dei *amici per diverzione*»; «Credo che questo sia naturale e forse anche *saludabile*») o le forme non corrette in italiano dal punto di vista lessicale o morfosintattico (es.: **attrattività* turistiche, 'attrazioni turistiche'; **il più meglio studente*, 'il miglior studente'; **mi ho laureato*, 'mi sono laureato'). Sono stati inoltre inseriti dei *tag* interni al testo per indicare casi in cui la trascrizione di una determinata forma non è stata possibile per grafia non leggibile (tag: $UNCLEAR$); per rimuovere il nome dell'apprendente se presente e anonimizzare quindi il testo (tag: $NAME$); per rimuovere indirizzi e-mail dei candidati dal testo se presenti (tag: $EMAIL$); e per omettere possibili link a siti web (URL).

Infine, sono state trascritte nel database di raccolta anche le tracce delle produzioni scritte in modo da poter successivamente ricollegare il testo alla tipologia di traccia scelta e alla sessione d'esame. Le 64 tracce selezionate sono state trascritte dai fascicoli cartacei delle prove d'esame a cui appartenevano le produzioni scritte incluse nel *corpus*.

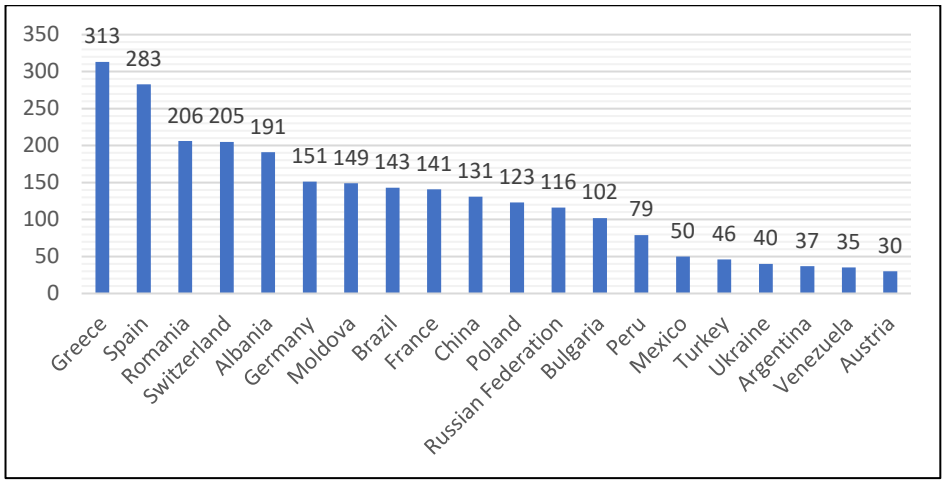
3. COMPOSIZIONE DEL *CORPUS*

Il *corpus* CELI si compone di 3041 testi suddivisi in quattro sezioni corrispondenti ai quattro livelli di competenza dell'italiano B1, B2, C1 e C2. Per ogni testo sono riportati i metadati relativi ai candidati, ai testi e alle tracce. Nei sottoparagrafi seguenti saranno descritte le informazioni contenute nei metadati relativi ai candidati, al testo, alla traccia e ai dati del *corpus*.

3.1. *Metadati relativi ai/alle candidati/e, ai testi e alle tracce*

Nel *corpus* CELI ogni testo è associato al candidato che lo ha prodotto. Per ogni candidato è riportata l'informazione relativa al genere (F=2147; M=894), alla data di nascita, al numero di matricola e alla nazionalità (che non indica necessariamente la lingua nativa dei candidati). Le nazionalità dei partecipanti sono in tutto 104, con una prevalenza delle nazionalità greca, spagnola, romena, svizzera e albanese. Il Grafico 1 mostra la distribuzione delle venti nazionalità più frequenti nel *corpus* CELI.

Grafico 1. *La distribuzione delle prime venti nazionalità dei candidati nel corpus CELI*



Ogni testo è poi registrato nel *corpus* con un numero identificativo, con il codice del centro linguistico in cui si è svolta la prova d'esame, e con il numero del faldone dell'archivio del CVCL in cui è conservata la prova cartacea.

Per ciascun testo sono inoltre riportati i dati relativi alla prova d'esame: il livello linguistico (secondo il QCER per il quale il candidato ha sostenuto l'esame di certificazione linguistica); il punteggio totale dell'intera prova d'esame; la fascia di punteggio (A (ottimo), B (buono) o C (sufficiente)[20] a cui corrisponde il punteggio finale della prova d'esame; il punteggio totale della prova scritta. La Tabella 3 mostra i *range* e le fasce dei punteggi delle prove d'esame per ciascun livello.

Nel *corpus* ogni testo è associato a tre ulteriori metadati: il numero identificativo della traccia della produzione scritta; il punteggio totale assegnato alla produzione scritta; e i punteggi assegnati alla competenza lessicale, alla competenza grammaticale, alla competenza sociolinguistica, e alla coerenza e coesione del testo[21]. La Tabella 4 mostra i *range* del punteggio totale e dei punteggi delle quattro competenze valutate (lessicale, grammaticale, sociolinguistica, coerenza e coesione del testo) delle produzioni scritte.

Infine, nella Tabella 5 sono riassunti i metadati disponibili in relazione ai partecipanti e ai testi.

Tabella 3. *Range dei punteggi e fasce di punteggio delle prove d'esame per livello linguistico*

Livello	***Range* del punteggio dell'intera prova d'esame (Prova Scritta + Prova Orale)**	**Fascia di punteggio dell'intera prova d'esame**	***Range* del punteggio della Prova Scritta**
CELI 2 B1	138 - 160 115 - 137 94 - 114	A B C	72 - 120
CELI 3 B2	173 - 200 144 – 172 117 - 143	A B C	84 - 140
CELI 4 C1	173 - 200 144 – 172 117 - 143	A B C	84 - 140
CELI 5 C2	173 - 200 144 – 172 117 - 143	A B C	89 - 150

[20] https://www.unistrapg.it/sites/default/files/docs/certificazioni/celi-2-valutazione.pdf; https://www.unistrapg.it/sites/default/files/docs/certificazioni/celi-3-valutazione.pdf; https://www.unistrapg.it/sites/default/files/docs/certificazioni/celi-4-valutazione.pdf; https://www.unistrapg.it/sites/default/files/docs/certificazioni/celi-5-valutazione.pdf.

[21] https://www.unistrapg.it/sites/default/files/docs/certificazioni/competenze-punteggi-CELI-2-B1-scritto.pdf; https://www.unistrapg.it/sites/default/files/docs/certificazioni/competenze-punteggi-CELI-3-B2-scritto.pdf; https://www.unistrapg.it/sites/default/files/docs/certificazioni/competenze-punteggi-CELI-4-C1-scritto.pdf; https://www.unistrapg.it/sites/default/files/docs/certificazioni/competenze-punteggi-CELI-5-C2-scritto.pdf.

Tabella 4. *Punteggi della produzione scritta per livello linguistico*

Livello	*Range* punteggio produzione scritta	Competenza lessicale (max.)	Competenza grammaticale (max.)	Competenza sociolinguistica (max.)	Coerenza e coesione (max.)
B1	12 - 20	5	5	5	5
B2	12 - 20	5	5	5	5
C1	18 - 30	8	8	6	8
C2	21 - 35	9	8	9	9

Tabella 5. *Metadati relativi ai partecipanti e ai testi*

Metadato	Valore
ID	Numero identificativo del testo
AUTHOR	Numero di matricola del candidato
AGE	Età del candidato alla data dell'esame
SEX	Genere del candidato
NATIONALITY	Nazionalità del candidato
EXAM_CENTRE	Codice identificativo del centro d'esame dove è stato sostenuto l'esame
FALDONE	Numero del faldone dell'archivio CVCL in cui è conservata la prova d'esame
ITA/EE	Informazione relativa a dove la prova si è svolta (in Italia o all'estero)
CEFR	Livello linguistico per il quale il candidato ha sostenuto l'esame
TOT_SCORE	Risultato finale della prova d'esame (somma tra punteggio della Prova Scritta e Prova Orale)
SCORE_BAND	Fascia di punteggio a cui corrisponde il risultato finale dell'esame
W_SCORE	Punteggio totale della Prova Scritta
ID_TRACCIA	Numero identificativo della traccia
TASK_SCORE	Punteggio totale assegnato alla produzione scritta
LEX_SCORE	Punteggio assegnato alla competenza lessicale
GRAM_SCORE	Punteggio assegnato alla competenza grammaticale
SOCIO_SCORE	Punteggio assegnato alla competenza sociolinguistica
CC_SCORE	Punteggio assegnato alla coerenza e coesione del testo

Per quanto riguarda i metadati delle tracce delle produzioni scritte, nel *corpus* ogni traccia è assegnata a un codice identificativo, che contiene l'informazione relativa alla sessione d'esame (a quando quindi il candidato ha sostenuto la prova di certificazione linguistica) e alla tipologia della traccia stessa. Inoltre, per ogni traccia sono riportate le seguenti informazioni: il livello QCER dell'esame a cui appartiene la traccia; il punteggio massimo che può essere ottenuto nella Prova Scritta; il punteggio massimo che può essere ottenuto nella produzione scritta; i punteggi massimi che possono essere assegnati alla competenza lessicale, grammaticale, sociolinguistica e alla coerenza e coesione del testo; il genere della traccia; la tipologia della traccia (Tabella 6).

I generi di produzione scritta sono stati classificati in lettera, e-mail, *blog*, racconto, articolo e relazione. Invece, le tipologie delle tracce sono state suddivise in argomentativa,

descrittiva e narrativa, o mista (descrittiva-narrativa; argomentativa-narrativa; argomentativa-descrittiva; argomentativa-narrativa-descrittiva).

Tabella 6. *Metadati relativi alle tracce*

Metadato	**Valore**
ID_TRACCIA	Numero identificativo della traccia
SESSIONE	Data della sessione d'esame
CEFR	Livello QCER dell'esame a cui appartiene la traccia
TOT_SCORE_MAX	Punteggio massimo ottenibile nell'intera prova d'esame
TASK_SCORE_MAX	Punteggio massimo ottenibile nella produzione scritta
LEX_SCORE_MAX	Punteggio massimo ottenibile nella competenza lessicale
GRAM_SCORE_MAX	Punteggio massimo ottenibile nella competenza grammaticale
SOCIO_SCORE_MAX	Punteggio massimo ottenibile nella competenza sociolinguistica
CC_SCORE_MAX	Punteggio massimo ottenibile nella coerenza e coesione del testo
GENRE	Genere della produzione scritta
TYPE	Tipologia della produzione scritta

3.2. *Dati*

I 3041 testi del *corpus* CELI comprendono complessivamente 608.614 *token* e 24.698 *type* (forme distinte), che corrispondono mediamente a 200 *token* per testo. Il *corpus* è suddiviso in quattro sezioni, corrispondenti ai quattro livelli di competenza linguistica in italiano B1, B2, C1 e C2 (secondo il QCER). Il numero totale dei *token* è distribuito in modo bilanciato tra le quattro sezioni; ciò le rende comparabili in termini di grandezza (ciascun livello contribuisce circa per il 25% al totale dei *token* del *corpus*). La Tabella 5 presenta un quadro riassuntivo del *corpus*, con i dati relativi a ciascuna delle quattro sezioni.

Tabella 7. *La composizione del corpus CELI e delle quattro sezioni*

Sezione	**n. testi**	***token***	**media *token***	***type***	**TTR**	**frasi**	***token* x frase**
B1	1212	156.612	129,21	7.397	18,69	13.514	11,58
B2	840	152.251	181,25	9.519	24,39	8.438	18,04
C1	585	149.859	256,16	12.546	32,4	7.508	19,95
C2	404	149.892	371,01	14.153	36,55	7.196	20,82
TOTALE	**3041**	**608.614**	**-**	**-**	**-**	**36.656**	**-**

Nota. *La type-token ratio (TTR) è stata calcolata con l'indice di Guiraud (Guiraud, 1954) per ovviare alla non-omogeneità del numero di token.*

La composizione interna di ogni sezione è descritta nei sottoparagrafi seguenti. Per ogni sezione saranno illustrati i dati relativi alle nazionalità dei candidati; a dove è stato svolto l'esame CELI; alle medie dei punteggi delle prove d'esame, delle prove scritte e delle produzioni scritte; ai generi e alle tipologie delle tracce. Un primo riassunto dei dati viene mostrato nella Tabella 8.

Tabella 8. *Totale delle nazionalità dei candidati e medie dei punteggi delle prove d'esame per livello*

Sezione	Tot. Nazionalità Candidati	Media punteggio prova d'esame	Media punteggio prova scritta	Media punteggio produzione scritta
B1	85	124	91	16
B2	65	157	107	16
C1	52	154	104	24
C2	49	153	109	28

3.2.1. *B1*

La sezione del livello B1 comprende testi prodotti da 1212 apprendenti (F=808; M=404) per un totale di 156.612 *token*. La maggior parte degli apprendenti ha svolto l'esame di certificazione presso un centro d'esame estero (73%). L'insieme delle nazionalità degli apprendenti è eterogeneo, per un totale di 85 nazionalità diverse; le dieci nazionalità più frequenti sono: greca (10%); cinese (9%); svizzera (8%); moldava (7%); romena (7%); spagnola (6%); tedesca (5%); albanese (5%); russa (4%); francese (4%).

Per la prova B1, il *range* del punteggio totale dell'intera prova d'esame è compreso fra 94 e 160; la media del punteggio per il livello B1 è di circa 124, collocando la maggior parte degli apprendenti nella fascia di punteggio intermedia B (17%), a seguire il 57% nella fascia di punteggio C e il 26% nella fascia di punteggio A. Invece, la media del punteggio totale assegnato alla prova scritta è di 91 (in un *range* di punteggio compreso fra 72 e 100). Inoltre, la media di punteggio della specifica produzione scritta (compreso fra 12 e 20) è di 16. Infine, le medie dei punteggi nelle quattro competenze testuali valutate nella specifica produzione sono: 3,8 per la competenza lessicale; 3,6 per la competenza grammaticale; 4,3 per la competenza sociolinguistica; 4,5 per la coesione e la coerenza testuale.

Il genere più frequente delle tracce B1 è l'e-mail (85%), seguito dalla lettera (15%). Invece, le tipologie delle tracce delle prove B1 si distribuiscono in: descrittiva-narrativa (54%); argomentativa-descrittiva (17%); argomentativa-descrittiva-narrativa (15%); narrativa (14%).

3.2.2. *B2*

La sezione del livello B2 è costituita da 840 testi, per un totale di 152.251 *token*. La maggior parte degli apprendenti (F=611; M=229) ha sostenuto la prova d'esame presso un centro d'esame estero (79%). Le nazionalità degli apprendenti B2 sono in totale 65; le dieci più frequenti sono: spagnola (9%); romena (9%); brasiliana (8%); greca (7%); albanese (6%); tedesca (6%); francese (6%); polacca (5%); bulgara (5%); e peruviana (4%).

La media del punteggio assegnato all'intera prova d'esame, il cui *range* è compreso fra 117 e 200, è di 157, posizionando la maggior parte degli apprendenti nella fascia di punteggio B (57%), mentre il 22% degli apprendenti ha superato l'esame di livello B2 con la fascia di punteggio A e il 21% con la fascia di punteggio C. La media del punteggio dell'intera prova scritta e di 107 (il *range* è compreso fra 84 e 140) e della specifica produzione scritta, in cui il candidato poteva prendere da un minimo di 12 a un massimo di 20 per superarla, è di 16. Le medie assegnate alle quattro competenze testuali valutate

nella produzione scritta sono di: 3,8 per la competenza lessicale; 3,7 per la competenza grammaticale; 4,3 per la competenza sociolinguistica; 4,3 per la coesione e la coerenza del testo.

Nella sezione del livello B2 i generi delle tracce si suddividono in tema (90%), lettera (8%), articolo (1%), e relazione (1%). Invece le tre tipologie testuali sono così distribuite: il 53% delle tracce è descrittivo-narrativo; il 31% argomentativo-descrittivo; il 7% argomentativo-narrativo; il 5% narrativo; e il 4% argomentativo.

3.2.3. *C1*

La sezione del livello C1 si compone di 585 testi, per un totale di 149.859 *token*. La prova d'esame è stata sostenuta dalla maggior parte degli apprendenti (F=423; M=162) presso un centro d'esame estero (77%). Il *background* delle nazionalità degli apprendenti C1 è eterogeneo, per un totale di 52 nazionalità, tra cui le più frequenti sono: spagnola (16%); greca (13%); albanese (10%); romena (6%); svizzera (6%); tedesca (5%); francese (5%); bulgara (5%); russa (5%); moldava (4%).

Per superare la prova d'esame, gli apprendenti dovevano ottenere un punteggio compreso fra 117 e 200; la media è di 154 e colloca la maggior parte dei candidati nella fascia di punteggio B (56%), il 27% nella fascia C, e il 17% nella fascia A. La media del punteggio dell'intera prova scritta è di 104 (in un *range* compreso fra 84 e 140), e della specifica produzione scritta, in cui l'apprendente poteva ottenere un punteggio compreso fra 18 e 30 per superarla, è di 24. Le medie assegnate alle quattro competenze testuali valutate nella produzione scritta sono di: 6 per la competenza lessicale; 6 per la competenza grammaticale; 5,1 per la competenza sociolinguistica; 6,9 per la coesione e la coerenza del testo.

Nelle prove d'esame C1 i generi delle tracce si suddividono in: *blog* (42%); e-mail (35%); lettera (12%); tema (6%); articolo (5%). Le tipologie delle tracce C1 sono argomentativa (78%) e argomentativa-descrittiva (22%).

3.2.4. *C2*

La sezione del livello C2 è formata da 404 testi, per un totale di 149.892 *token*. La maggior parte degli apprendenti (F=305; M=99) ha sostenuto l'esame di certificazione linguistica presso un centro d'esame estero (64%). Le nazionalità dei candidati sono eterogenee, per un totale di 49 nazionalità diverse, con una prevalenza di nazionalità: greca (14%); spagnola (12%); polacca (9%); bulgara (6%); albanese (5%); romena (5%); francese (4%); tedesca (4%); brasiliana (3%); moldava (3%).

L'intera prova d'esame era considerata superata se gli apprendenti ottenevano un punteggio compreso fra 117 e 200; la media dei punteggi finali è di 153 con la maggior parte degli apprendenti collocati nella fascia di punteggio intermedia B (53%), e a seguire il 31% dei candidati ha passato l'esame ottenendo la fascia di punteggio C e il 16% la fascia di punteggio A. Invece, la media dei punteggi della prova scritta è di 109 (in un *range* compreso fra 89 e 150), e della specifica produzione scritta è di 28 (gli apprendenti superavano la prova se prendevano un punteggio compreso fra 21 e 35). Le medie delle competenze testuali valutate nella produzione scritta sono di: 6,8 per la competenza lessicale; 6,2 per la competenza grammaticale; 7,4 per la competenza sociolinguistica; 7,6 per la coerenza e la coesione del testo.

Per quanto riguarda i generi delle tracce C1, questi si distinguono in: articolo (44%); racconto (24%); tema (15%); lettera (9%); *blog* (8%). La tipologia delle tracce C1 scelta più frequentemente dagli apprendenti è quella argomentativa (64%), a seguire narrativa (24%), argomentativa-descrittiva (9%), e argomentativa-narrativa (3%).

4. ANNOTAZIONE

Dopo la fase di trascrizione delle 3041 prove d'esame, i testi sono stati convertiti in xml, in modo da integrare tutti i metadati riassunti nella Tabella 5 in un linguaggio interpretabile da *software* adatti ad interrogazioni di tipo linguistico. Successivamente, i dati sono stati annotati per categoria grammaticale e lemmatizzati per mezzo di *TreeTagger* (Schmid, 1994), uno dei programmi più diffusi per le operazioni di *pos-tagging*. Le procedure di annotazione per categoria grammaticale, che costituiscono un compito tutt'altro che banale nel caso di dati prodotti da apprendenti (Van Rooy, 2015), sono state svolte usando il *tagset* creato ad hoc per l'annotazione del *Perugia corpus* (Spina, 2014)[22]. Il *tagset* è formato da 54 etichette, ed è quindi piuttosto articolato rispetto a quelli utilizzati per altri *corpora* dell'italiano, come ad esempio il *Lessico di frequenza dell'italiano parlato* (De Mauro *et al.*, 1993). La categoria grammaticale che prevede la differenziazione maggiore è quella dei verbi (vedi Tabella 9), che, a partire da una tripartizione inziale tra verbi, ausiliari e verbi servili, è successivamente suddiviso in verbi di modo finito, infiniti, gerundi e participi, ciascuno dei quali è ulteriormente distinto in forme con o senza pronomi clitici in*corpora*ti. Gli aggettivi, inoltre, sono a loro volta differenziati in qualificativi, possessivi, indefiniti e dimostrativi (queste ultime tre categorie prevedono anche la forma pronominale). L'uso di un *tagset* con ampia differenziazione interna, se da un lato comporta il rischio di una minore accuratezza da parte del *tagger*, consente tuttavia un raffinamento molto maggiore in fase di interrogazione dei dati.

Tabella 9. *Le etichette per le forme verbali nel tagset del CELI*

Etichetta	Descrizione	Esempio
VER:fin	Verbo di modo finito	vado, partirò, terrei, parlassi
VER:fin:cli	Verbo di modo finito con clitico	vacci, dammi, diccelo
AUX:fin	Verbo ausiliare di modo finito	ho (fatto), ero (andato)
VER2:fin	Verbo servile di modo finito	posso, voglio, devo
VER:geru	Verbo al gerundio	andando, temendo, partendo
VER:geru:gli	Verbo al gerundio con clitico	andandoci, temendolo, finendole
AUX:geru	Verbo ausiliare al gerundio	avendo (fatto), essendo (arrivato)
AUX:geru:cli	Verbo ausiliare al gerundio con clitico	avendolo (fatto), essendoci (arrivato)
VER2:geru	Verbo servile al gerundio	volendo, potendo, dovendo
VER2:geru:cli	Verbo servile al gerundio con clitico	volendolo, potendola, dovendoli
VER:infi	Verbo all'infinito	amare, temere, finire
VER:infi:cli	Verbo all'infinito con clitico	amarlo, temerle, finirla

[22] Il *tagset* completo può essere consultato qui:
https://www.unistrapg.it/cqpwebnew/doc_*corpora*/tagset.pdf, ed è basato su un precedente *tagset* creato da Marco Baroni.

AUX:infi	Verbo ausiliare all'infinito	avere (fatto), essere (andato)
AUX:infi:cli	Verbo ausiliare all'infinito con clitico	averlo (fatto), esserci (andato)
VER2:infi	Verbo servile all'infinito	potere, dovere, volere
VER2:infi:cli	Verbo servile all'infinito con clitico	poterci, doverlo, volerla
VER:ppast	Verbo al participio passato	andato, temuto, finire
VER:ppast:cli	Verbo al participio passato con clitico	andatoci, temutolo, finitolo
AUX:ppast	Verbo ausiliare al participio passato	(è) stata (prevista)
VER2:ppast	Verbo servile al participio passato	potuto, dovuto, voluto
VER2:ppast:cli	Verbo servile al part. pass. Con clitico	potutolo, dovutolo
VER:ppre	Verbo al participio presente	riguardante, derivante, stante
VER:ppre:cli	Verbo al participio presente con clitico	autopropagantesi

Un'altra caratteristica dell'annotazione per categoria grammaticale e lemmatizzazione del *corpus* CELI è che, conformemente a quanto già fatto per il *Perugia corpus*[23], un insieme di locuzioni avverbiali, o più in generale di espressioni con significato prevalentemente grammaticale composte da più di una parola, sono state annotate come un lemma unico. È questo, ad esempio, il caso di *un po'* (pronome indefinito), *a galla*, *a fondo*, *al di fuori*, *al di là*, *di sfuggita*, *di gran lunga*, *di recente* ecc. (avverbi). Oltre che ad un principio metodologico, derivante dalla considerazione che queste espressioni sono blocchi di parole separate solo convenzionalmente, ma di fatto riconducibili ad un'unica funzione, tanto è vero che di alcune sono possibili le grafie univerbate accanto a quelle separate (*al di là* / *aldilà*), questa scelta risponde anche a un fine pratico: permette infatti di evitare una serie di errori che il *tagger* potrebbe commettere se queste forme fossero analizzate parola per parola: ad esempio, *fondo* in *a fondo* è una forma molto ambigua, che prevede tre possibili annotazioni, come aggettivo, sostantivo o prima persona singolare dei due verbi *fondare* e *fondere*. Ci sono inoltre locuzioni i cui componenti non sono utilizzati in italiano al di fuori della locuzione stessa (*bizzeffe*, *galla*).

Una volta portate a termine le operazioni di annotazione per categoria grammaticale e lemmatizzazione, i dati annotati sono stati sottoposti a un'ulteriore elaborazione semiautomatica, mirata ad eliminare errori ricorrenti del *tagger* (Spina, 2014). Questa fase di *post-tagging*, che viene comunemente applicata a dati di parlanti nativi e non nativi, ha consentito di correggere quasi 5.000 errori commessi dal *tagger*. Gli errori sono relativi in particolare a forme caratterizzate da forte ambiguità grammaticale (*come*, *che*, *dove*, *perché*, che possono essere congiunzioni subordinanti, avverbi comparativi o interrogativi), o a forme verbali con clitico non presenti nel lessico e quindi ignote al *software*, che prova ad assegnarle a una categoria grammaticale ma non riesce a individuarne il lemma (*spronarsi*, *raccontartene*).

Infine, un'operazione ulteriore di correzione manuale dei dati annotati ha riguardato le rimanenti forme non riconosciute dal *tagger* e classificate come lemmi "sconosciuti". Si tratta prevalentemente di forme scritte dagli apprendenti in modo non standard, e quindi tipiche di dati prodotti da parlanti non nativi. In questi casi, anche quando la forma corrispondeva a un lemma morfologicamente possibile in italiano, la correzione ha tendenzialmente inserito il lemma target (*estraniero* → *straniero*; *espettative* → *aspettativa*; *evolvimenti* → *evoluzione*), anche se in alcuni casi questo non è stato possibile per la mancanza di lemmi singoli a cui ricondurre le forme prodotte. È il caso, ad esempio, degli infiniti

[23] https://www.unistrapg.it/cqpweb/doc_*corpora*/lista_MWE.pdf.

soluzionare, *volontariare*, *disponibilizzare*, il cui lemma è stato lasciato identico alla forma prodotta, benché non esistente in italiano. Alla fine delle due fasi semiautomatiche e manuali di *post-tagging*, un numero di errori di lemma e/o categoria grammaticale pari a circa l'1% del totale delle forme che compongono il *corpus* ha potuto essere corretto, incrementando in tal modo l'accuratezza dell'annotazione[24].

Il risultato di questa complessa e impegnativa fase di annotazione sono dei dati arricchiti da un insieme vasto e articolato di informazioni, sia linguistiche (le categorie grammaticali e i lemmi di ciascuna forma prodotta) che extralinguistiche (i dati anagrafici degli apprendenti, il loro livello di competenza, i punteggi delle singole componenti e della valutazione globale). Il *corpus* costruito su questi dati[25] consente dunque l'esecuzione di interrogazioni multilivello, in grado di integrare criteri linguistici ed extralinguistici. Le varie fasi di costituzione del *corpus*, dalla sua progettazione alla raccolta e al bilanciamento dei dati, dalla loro annotazione fino alla sua pubblicazione per renderlo disponibile alla comunità scientifica, si sono conformate a ciò che la letteratura più recente dell'area della linguistica dei *corpora* (Paquot, Gries, 2020) individua come standard consolidato della disciplina. L'attenzione particolare dedicata all'accuratezza dell'annotazione, infine, deriva da una concezione non esclusivamente quantitativa del dato linguistico: soprattutto nel caso di *corpora* di apprendenti, la cui elaborazione richiede un lavoro aggiuntivo per la natura non omogenea delle loro interlingue, il valore di un *corpus* di dati non risiede solo nella sua estensione, ma anche nella qualità delle informazioni che è in grado di veicolare sugli usi linguistici di specifiche varietà di parlanti.

5. Usi potenziali per la ricerca e per la didattica

In questa sezione, illustreremo gli usi potenziali del *corpus* CELI considerando due prospettive differenti:

- la prospettiva della ricerca linguistico-acquisizionale;
- la prospettiva della glottodidattica.

L'intento della sezione è quello di mostrare la molteplicità e versatilità degli usi del *corpus*, considerando non solo le differenti prospettive individuate, ma anche i loro diversi gradi di integrazione.

5.1. *La prospettiva della ricerca linguistico-acquisizionale*

La ricerca svolta nell'ambito della linguistica acquisizionale mira a verificare teorie e adottare approcci analitici attraverso il riscontro con i dati empirici. Fino ad ora, la maggior parte degli studi empirici relativi all'apprendimento dell'italiano L2/LS si è concentrata su livelli di competenza specifici e aree linguistiche definite. La disponibilità e l'accesso aperto a dati pseudo-longitudinali (Callies, 2015) provvisti di un'annotazione multilivello come quelli del *corpus* CELI consente in primo luogo di estrarre e analizzare

[24] Una valutazione formale dell'accuratezza dell'annotazione grammaticale e della lemmatizzazione esula dagli scopi di questo articolo. Le stesse operazioni (*tagging* automatico e *post-tagging*), svolte usando lo stesso *tagger*, addestrato allo stesso modo, hanno portato, su dati di italiani nativi, a un'accuratezza complessiva superiore al 98% (Autore).

[25] Il *corpus* CELI è stato creato usando Open *Corpus* Workbench (CWB), un insieme di risorse open source per la gestione e l'interrogazione di *corpora* linguistici annotati (https://cwb.sourceforge.io/), e sarà interrogabile dal sito https://www.unistrapg.it/cqpwebnew/.

dati sull'uso dell'italiano di apprendenti relativi a diversi livelli di competenza, raccolti in modo bilanciato. Ciò permette di tracciare lo sviluppo della competenza, e ad esempio di verificare, per tutti i livelli di analisi consentiti da un *corpus* scritto, e con una solida base empirica, la non-linearità nell'apprendimento linguistico, che è uno degli assunti della teoria della complessità (Larsen-Freeman, Cameron, 2009).

Un *learner corpus* come il CELI può dunque servire a verificare in modo deduttivo teorie preesistenti sulle caratteristiche e lo sviluppo dell'interlingua degli apprendenti, ma può anche essere usato, in modo induttivo, *corpus-driven* (Francis, 1993), come mezzo per scoprire regolarità o anomalie nell'uso degli apprendenti, senza teorie a priori da verificare. In entrambi i casi, l'uso di un *corpus* bilanciato e rappresentativo consente di generalizzare i risultati della ricerca, e di rendere l'analisi replicabile (Plonsky, 2012).

Il *corpus* CELI è inoltre uno strumento adatto all'adozione di due degli approcci che si sono sviluppati negli anni all'interno della LCR (Spina, 2020). Il primo è la *Contrastive Interlanguage Analysis* (Granger, 1996, 2015), che prevede due possibili tipi di analisi comparative: il confronto tra l'interlingua degli apprendenti, documentata dal *corpus*, e la lingua di parlanti italiani nativi, relativamente a fenomeni di qualsiasi livello di analisi; e quello tra tipi diversi di interlingue, ad esempio di apprendenti con diverse L1. Il secondo approccio è quello della *Computer-aided Error Analysis* (Dagneaux *et al.* 1998). Questa metodologia è mirata ad un'analisi sistematica degli errori prodotti dagli apprendenti, e prevede l'adozione di uno schema standardizzato di annotazione di tali errori, che possono dunque essere individuati, misurati, confrontati e analizzati all'interno del loro contesto di occorrenza, fornendo in tal modo possibili evidenze sullo sviluppo della competenza linguistica degli apprendenti.

Un'altra importante potenzialità del *corpus* CELI in campo acquisizionale deriva infine dal suo possibile legame con le metodologie della ricerca sperimentale. L'integrazione di approcci diversi nello studio dell'interlingua è infatti da tempo considerata una priorità negli studi acquisizionali (Gilquin, Gries, 2009). La triangolazione di approcci diversi, gli uni basati sui dati tendenzialmente autentici dei *learner corpora*, gli altri sui dati controllati raccolti tramite specifici esperimenti, consente di verificare le ipotesi di ricerca in esame alla luce di prospettive diverse, e di ottenere in tal modo un quadro più articolato dei fenomeni considerati, in grado di rafforzare l'affidabilità dei risultati ottenuti. Oltre a ciò, in modo anche solo strumentale, i dati del *corpus* CELI possono essere utilizzati per costruire gli stimoli di esperimenti di tipo psicolinguistico, e costituiscono, in virtù della loro autenticità, un valore aggiunto per l'efficacia di tali esperimenti.

5.2. *La prospettiva della glottodidattica*

Sul versante della glottodidattica, il *corpus* CELI si presta a numerosi utilizzi inerenti alla pianificazione didattica, allo sviluppo di attività pedagogiche e al *language testing*. La scarsa disponibilità di dati empirici bilanciati per livelli di competenza ha, di fatto, costituito un limite nel tentativo di operare delle generalizzazioni pedagogicamente rilevanti (cfr. Giacalone Ramat, 2003). Grazie al *corpus* CELI, è possibile disporre di una base di confronto bilanciata tra i diversi livelli, consentendo quindi sia allo sviluppatore di materiali didattici, sia all'insegnante di lingua, di osservare regolarità sul piano degli usi linguistici e degli errori più ricorrenti relativi a un livello di competenza specifico. Tramite il confronto con il livello precedente e quello successivo rispetto al livello a cui appartengono i propri studenti, un docente avrà modo di disporre di uno "specchio empirico" con cui confrontarsi, quantitativamente consistente e di sicura attribuzione per

quanto riguarda il livello di competenza, proprio perché fondato su testi prodotti in ambito certificatorio.

Un *learner corpus* come il *corpus* CELI si presta, inoltre, allo sviluppo di attività pedagogiche di vario tipo. Tali attività possono assumere le forme di tutte quelle attività che tradizionalmente troviamo nei manuali in uso nei corsi di lingua, oppure le forme di attività basate sul contatto diretto tra i dati tratti dal *corpus* e gli apprendenti stessi. Nel primo caso, lo sviluppatore di materiali didattici o l'insegnante di lingua può attingere al *corpus* CELI per estrapolare frasi adatte ad apprendenti di un determinato livello di competenza, per poi manipolarle al fine di costruire attività didattiche specifiche. Per esempio, si può estrarre una frase contenente un errore, così da costruire un esercizio in cui all'apprendente viene chiesto di correggere l'errore. Oppure, si possono estrarre frasi corrette, utilizzandole per esercizi di completamento o di abbinamento. I vantaggi di questo tipo di operazione sono molteplici. Innanzitutto, le frasi che verranno presentate all'apprendente sono frasi autentiche, in quanto prodotte da apprendenti di pari livello rispetto agli apprendenti che si hanno in classe. Oltre a ciò, si disporrà anche di errori autentici, relativi a livelli di competenza specifici. Di conseguenza, l'insegnante non si troverebbe più nella condizione di dover ricorrere a frasi inventate per venire incontro ai bisogni degli apprendenti: la frase adeguata al livello di competenza sarebbe già disponibile all'interno del *corpus* e all'insegnante non resterebbe altro da fare che estrarla attraverso una ricerca per forma specifica, per parte del discorso o per stringa.

Una seconda modalità di utilizzo didattico del *corpus* CELI consiste nell'esplorazione diretta del *corpus* da parte degli/delle apprendenti. In questo caso, facciamo riferimento al *Data-driven learning* (Johns, 1991; Boulton, 2017; Forti, 2019, 2021). Nella sua forma più tipica, gli apprendenti esplorano insiemi di occorrenze tratte da un *corpus*, al fine di individuare specifiche regolarità d'uso. Un *learner corpus* bilanciato per livelli di competenza come il *corpus* CELI consente di svolgere attività di scoperta guidata, in cui gruppi di apprendenti si confrontano con i propri pari che hanno prodotto i testi contenuti nel *corpus*. L'insegnante potrà guidare gli apprendenti in attività fondate sulla riflessione metalinguistica, orientata, per esempio, al confronto tra testi di apprendenti e testi di nativi (Ackerley, 2013). Tali attività si potranno svolgere sia su computer, sia su carta. Nel primo caso, è importante che il *corpus* sia stato costruito in modo tale da rendersi adatto all'esplorazione da parte di apprendenti e insegnanti (Forti, Spina, 2019). Nel secondo caso, l'insegnante dovrà predisporre le attività in anticipo, estraendo e selezionando gli insiemi di occorrenze pertinenti a un determinato obiettivo didattico. Il vantaggio principale di questa seconda modalità consiste nella possibilità di controllare e manipolare gli insiemi di occorrenze considerati.

Infine, il *corpus* CELI è in grado di fornire un aiuto prezioso nell'ambito del *language testing*. Come già mostrato in diversi lavori relativi alla lingua inglese (Taylor, Barker, 2008; Barker *et al*., 2015; Callies, Götz, 2015; Gablasova, 2021), e nell'ambito di un progetto di ricerca dell'Università per Stranieri di Siena (Peri *et al*., 2021), i *corpora* costituiscono uno strumento di rilevante interesse per la validazione degli *item* di un test. Nel caso del *corpus* CELI, si ha l'opportunità di attingere a usi linguistici autentici di apprendenti a diversi livelli di competenza, disponendo, in tal modo, dell'opportunità di calibrare gli *item* di un test in modo realistico, sulla base di quanto prodotto nell'ambito di un livello di competenza specifico. Un *learner corpus* come il *corpus* CELI, inoltre, consente l'utilizzo di dati relativi agli errori prodotti dagli apprendenti nell'ambito della costruzione di *item* a risposta multipla: gli errori estratti possono, infatti, costituire distrattori realistici in quanto autentici (Marello, 2009).

6. CONCLUSIONI

In questo articolo abbiamo descritto il *corpus* CELI, contenente 3041 testi di prove di esame per i livelli B1, B2, C1 e C2 di apprendenti di italiano. Il *corpus* è bilanciato per livello di competenza e include testi di apprendenti di 104 nazionalità diverse, con una prevalenza delle nazionalità greca, spagnola, romena, svizzera e albanese. Abbiamo descritto le varie fasi della realizzazione del *corpus*, da quella di progettazione, con l'individuazione dei criteri di composizione, a quella di raccolta, trascrizione e annotazione, che ha arricchito i testi con informazioni linguistiche ed extralinguistiche. Il lavoro di creazione e pubblicazione del *corpus* è stato presentato nel quadro metodologico della LCR, un approccio alla ricerca acquisizionale basato sull'uso sistematico di *corpora* elettronici di apprendenti. Sono state infine delineate alcune potenzialità che l'uso del *corpus* offre nel campo della ricerca acquisizionale e della glottodidattica.

La disponibilità di risorse come il CELI, accessibili alla comunità scientifica, è coerente con una visione della ricerca come attività necessariamente basata su dati aperti e riutilizzabili. In questo senso, il CELI fornisce un importante contributo, in termini di dati quantitativamente rilevanti e costruiti secondo standard qualitativi condivisi, alla ricerca acquisizionale e glottodidattica. In questi campi, l'adozione di approcci che integrino metodologie diversificate è da tempo raccomandata come una delle strade principali da percorrere in futuro. L'impiego sistematico di raccolte vaste e bilanciate di dati linguistici prodotti da apprendenti costituisce inoltre una sorta di ponte, potenzialmente in grado di consentire una migliore integrazione tra LCR e ricerca sull'acquisizione delle L2, con l'obiettivo di renderle in misura maggiore complementari, e di circoscrivere in tal modo alcuni dei loro limiti rispettivi (Myles, 2005): consolidando quindi da un lato le basi teoriche della LCR e fornendo dall'altro dati maggiormente estesi e di qualità alla ricerca acquisizionale.

RIFERIMENTI BIBLIOGRAFICI

Ackerley K. (2013), "A comparison of learner and native speaker writing in online self-presentations: Pedagogical applications", in Granger S., Gilquin G., Meunier F. (eds.), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*, Presses Universitaires de Louvain, Louvain, pp. 1-10.

Bailini S., Frigerio A. (2018), "CORESPI e CORITE, due nuovi strumenti per l'analisi dell'interlingua di lingue affini", in *CHIMERA: Romance Corpora and Linguistic Studies*, 5, 2, pp. 313-319.

Barker F., Salamoura A., Saville N. (2015), "Learner *corpora* and language testing", in Granger, S., Gilquin, G., Meunier, F. (eds.), *The Cambridge Handbook of Learner Corpus Research*, Cambridge University Press, Cambridge, pp. 511-534.

Boulton A. (2017), "Research Timeline. *Corpora* in language teaching and learning", in *Language Teaching*, 50, 4, pp. 483-506.

Bratankova L. (2015), *Le collocazioni Verbo + Nome in apprendenti di italiano L2*, Tesi di dottorato, Università per Stranieri di Perugia.

Callies M. (2015), "Learner *corpus* methodology", in Granger S., Gilquin G., Meunier F. (eds.), *The Cambridge Handbook of Learner Corpus Research*, Cambridge University Press, Cambridge, pp. 35-56.

Callies M., Götz S. (2015), *Learner corpora in language testing and assessment*, John Benjamins, Amsterdam-Philadelphia.

Chini M. (2016), "Elementi utili per una didattica dell'italiano L2 alla luce della ricerca acquisizionale", in *Italiano LinguaDue*, 8, 2, pp. 1-18: https://riviste.unimi.it/index.php/promoitals/article/view/8172.

Corino E., Colombo S., Marello C. (2017), *Italiano di stranieri: i corpora VALICO e VINCA*, Guerra, Perugia.

Dagneaux E., Denness S., Granger S. (1998), "Computer-aided error analysis", in *System*, 26, 2, pp. 163-174.

De Mauro T., Mancini F., Vedovelli M., Voghera M. (1993), *Lessico di frequenza dell'italiano parlato*, EtasLibri, Milano.

Forti L. (2019), *Developing phraseological competence in Italian L2: A study on the effects of Data-driven learning*, Tesi di dottorato, Università per Stranieri di Perugia.

Forti L. (2021), "Apprendere la grammatica attraverso il Data-driven learning", in Giunchi P., Roccaforte M., *La grammatica tra acquisizione e apprendimento. Un percorso verso la consapevolezza linguistica*, Carocci, Roma, pp. 100-113.

Forti L., Spina S. (2019), "*Corpora* for Linguists vs. *Corpora* for Learners: Bridging the Gap in Italian L2 Learning and Teaching", in *EL.LE*, 8, 2, pp. 349-362.

Francis G. (1993), "A *Corpus*-Driven Approach to Grammar. Principles, Methods and Examples", in Baker M., Francis G., Tognini-Bonelli E. (eds.), *Text and Technology: In Honour of John Sinclair*, Benjamins, Amsterdam, pp. 137-156.

Gablasova D. (2021), "*Corpora* for second language assessments", in Tracy-Ventura N., Paquot M. (eds.), *The Routledge Handbook of Second Language Acquisition and Language Testing*, Routledge, Londra-New York, pp. 45-53.

Gallina F. (2015), *Le parole degli stranieri. Il Lessico Italiano Parlato da Stranieri*, Guerra, Perugia.

Giacalone Ramat, A. (a cura di) (2003), *Verso l'italiano. Percorsi e strategie di acquisizione*, Carocci, Roma.

Gilquin G., Gries S. Th. (2009), "*Corpora* and experimental meth- ods: A state-of-the-art review", in Gilquin G. (ed.), *Corpora and Experimental Methods. Special issue of Corpus Linguistics and Linguistic Theory*, 5, 1, pp. 1-26.

Glaznieks A., Frey J.-C., Stopfner M., Zanasi L., Nicolas L. (2022), "LEONIDE: A longitudinal trilingual *corpus* of young learners of Italian, German and English", in *International Journal of Learner Corpus Research*, 8, 1, pp. 97-120.

Glaznieks A., Frey J.-C., Nicolas L., Abel A., Vettori C. (in preparazione), "The Kolipsi *Corpus* Family. A collection of Italian and German L2 learner texts from secondary school pupils".

Granger S. (1996), "From CA to CIA and back: An integrated approach to computerized bilingual and learner *corpora*", in Aijmer K., Altenberg B., Johansson M. (eds.), *Languages in Contrast. Text-based Cross-linguistic Studies*, Lund University Press, Lund, pp. 37-51.

Granger S. (2015), "Contrastive interlanguage analysis: A reappraisal", in *International Journal of Learner Corpus Research*, 1, 1, pp. 7-24.

Granger S. (ed.) (1998), *Learner English on Computer*, Longman, Londra.

Granger S., Dupont M., Meunier F., Naets H., Paquot M. (2020), *The International Corpus of Learner English. Version 3*, Presses Universitaires de Louvain, Louvain-la-Neuve.

Granger S., Gilquin G., Meunier F. (eds.) (2015), *The Cambridge Handbook of Learner Corpus Research*, Cambridge University Press, Cambridge.

Greenbaum S., Nelson G. (1996), "The International *Corpus* of English (ICE) Project", in *World Englishes*, 15, 1, pp. 3-15.

Grego Bolli G., Pelliccia F. (2005), *CELI. Breve guida ai certificati di italiano L2*, Guerra, Perugia.

Grego Bolli G., Spiti M. G. (2004), *La verifica delle competenze linguistiche. Misurare e valutare nella certificazione CELI*, Guerra, Perugia.

Guiraud P. (1954), *Les caractères statistiques du vocabulaire*, Presses Universitaires de France, Paris.

Haan P. (2000), "Tagging non-native English with the TOSCA-ICLE tagger", in Mair Ch., Hundt M. (eds.), *Corpus linguistics and linguistic theory*, Rodopi, Amsterdam, pp. 69-79.

Johns T. (1991), "Should you be persuaded - Two examples of data driven learning materials", in *Classroom Concordancing, English Language Research Journal*, 4, pp. 1-13.

Larsen-Freeman D., Cameron L. (2009), *Complex systems and applied linguistics* (Nachdr.), Oxford University Press, Oxford.

Mackey A., Gass S. (eds.) (2012), *Research methods in second language acquisition: A practical guide*, Blackwell, Oxford.

Marello C. (2009), "Distrattori tratti da *corpora* di apprendenti di italiano LS/L2", in Corino E., Marello C. (a cura di), *VALICO. Studi di linguistica e didattica*, Guerra, Perugia, pp. 177-193.

McEnery A., Xiao R., Tono Y. (2006), *Corpus-based language studies: An advanced resource book*, Routledge, London-New York.

Myles F. (2005), "Interlanguage *corpora* and second language acquisition research", in *Second Language Research*, 21, 4, pp. 373-391.

Paquot M., Gries S. (eds.) (2020), *A Practical Handbook of Corpus Linguistics*, Springer, Berlino.

Peri G., Machetti S., Masillo P. (2021), "*Corpora* di apprendenti e valutazione linguistica. Strumenti per la valutazione dell'italiano di stranieri", presentazione poster, LIV Congresso internazionale della Società di Linguistica Italiana, 8-10 settembre 2021, Università degli Studi di Firenze, Firenze, Italia.

Plonsky L. (2012), "Replication, meta-analysis, and generalizability", in Porte G. K. (ed.), *Replication Research in Applied Linguistics*, Cambridge University Press, Cambridge, pp. 116-132.

Schmid H. (1994), "Probabilistic part-of-speech tagging using decision trees", in *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, U.K.

Spina S. (2014), "Il Perugia *Corpus*: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione", in Basili R., Lenci A., Magnini B. (a cura di), *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it* (Vol. 1), Pisa University Press, Pisa, pp. 354-359.

Spina S. (2020), "The role of Learner *Corpus* Research in the study of L2 phraseology: main contributions and future directions", in *Rivista di psicolingusitica applicata - Journal of Applied Psycholinguistics*, XX, 2, pp. 35-52.

Spina S., Siyanova-Chanturia A. (2018), "The Longitudinal *Corpus* of Chinese Learners of Italian (LOCCLI)", Poster presentato al 13th Teaching and Language *Corpora* conference, Università di Cambridge, Cambridge, U.K.

Taylor L., Barker, F. (2008), "Using *corpora* for language assessment", in Hornberger N. H. (ed.), *Encyclopedia of Language and Education*, Springer, Boston, MA, pp. 2377-2390.

Van Rooy B. (2015), "Annotating learner *corpora*", in Granger S., Gilquin G., Meunier F. (eds.), *The Cambridge Handbook of Learner Corpus Research*, Cambridge University Press, Cambridge, pp. 79-106.

Verspoor M. (2017), "Complex Dynamic Systems Theory and L2 pedagogy. Lessons to be learned", in Ortega L., ZhaoHang H. (eds.), *Complexity Theory and Language Development: In celebration of Diane Larsen-Freeman*, John Benjamins Publishing, Amsterdam, pp. 143-162.

Wisniewski K., Schöne K., Nicolas L., Vettori C., Boyd A., Meurers D., Abel A., Hana J. (2013), "MERLIN: An online trilingual learner *corpus* empirically grounding the European Reference Levels in authentic learner data", in *ICT for Language Learning 2013, Conference Proceedings*, Libreriauniversitaria.it. Edizioni, Firenze, pp. 14-15.