# CItA: un Corpus di Produzioni Scritte di Apprendenti l'Italiano L1 Annotato con Errori

**Alessia Barbagli•, Pietro Lucisano•,**
**Felice Dell'Orletta◇, Simonetta Montemagni◇, Giulia Venturi◇**
•Dipartimento di Psicologia dei processi di Sviluppo e socializzazione, Università di Roma "La Sapienza"
alessia.barbagli@gmail.com,pietro.lucisano@uniroma1.it
◇Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{nome.cognome}@ilc.cnr.it

## Abstract

**English.** In this paper we present CItA the first corpus of written essays by Italian L1 learners in the first and second year of lower secondary school. CItA was annotated with grammatical, orthographic and lexical errors. The corpus peculiarities and its diachronic nature make it particularly suitable for computational linguistics applications and socio–pedagogical studies.

**Italiano.** *In questo articolo presentiamo CItA il primo corpus di produzioni scritte di apprendenti l'italiano L1 del primo e del secondo anno della scuola secondaria di primo grado annotato con errori grammaticali, ortografici e lessicali. Le specificità del corpus e la sua natura diacronica lo rendono particolarmente utile sia per applicazioni linguistico-computazionali sia per studi socio-pedagogici.*

## 1 Introduzione

La costruzione di corpora di produzioni di apprendenti è da sempre al centro delle attività di ricerca della comunità di linguistica computazionale. Un'attenzione particolare è dedicata all'annotazione e classificazione degli errori commessi dagli apprendenti. Corpora annotati con questo tipo di informazione vengono usati tipicamente per lo studio e la creazione di modelli di sviluppo delle abilità di scrittura (Deane and Quinlan, 2010) e per lo sviluppo di sistemi a supporto dell'insegnamento (i cosiddetti *Intelligent Computer–Assisted Language Learning systems*) (Granger, 2003). In questo scenario, un interesse particolare è dedicato alla raccolta e annotazione di corpora di produzioni di apprendenti L2 impiegati come punto di partenza per studi sullo sviluppo dell'interlingua, per attività di riflessione sull'eventuale modifica e/o personalizzazione dell'azione didattica dell'insegnante e per lo sviluppo di sistemi di correzione automatica degli errori. La maggior parte delle attività ha riguardato la costruzione di corpora di apprendenti l'inglese L2, tra cui il più recente e il più ampio è il *NUS Corpus of Learner English (NUCLE)* (Dahlmeier et al., 2013), utilizzato come risorsa di riferimento nel 2013 e 2014 dello "Shared Task on Grammatical Error Correction" (Ng et al., 2013; Ng et al., 2014). Tuttavia, in questi ultimi anni, l'attenzione è stata anche rivolta a L2 diverse dall'inglese, quali ad esempio l'arabo (Zaghouani et al., 2015), il tedesco (Ludeling et al., 2005), l'ungherese (Dickinson and Ledbetter, 2012), il basco (Aldabe et al., 2005) e il ceco e l'italiano (Andorno and Rastelli, 2009; Boyd et al., 2014). Minore attenzione è stata invece dedicata alla costruzione di risorse costituite da produzioni di apprendenti L1. Un'eccezione è rappresentata dal *KoKo* corpus (Abel et al., 2014), collezione di produzioni di apprendenti tedesco L1 dell'ultimo anno della scuola secondaria di secondo grado arricchite con informazioni di sfondo degli apprendenti (es. età, genere, situazione socio–economica), annotazione manuale di errori ortografici e grammaticali, e informazione linguistica annotata in maniera automatica.

Collocandoci in quest'ultimo scenario, in questo articolo presentiamo CItA (*Corpus Italiano di Apprendenti L1*) il primo corpus di produzioni scritte di apprendenti l'italiano L1 annotato manualmente con diverse tipologie di errori e con la relativa correzione. Il corpus, composto da produzioni dei primi due anni della scuola secondaria di primo grado, è a nostra conoscenza non solo il primo corpus italiano di questo tipo ma contiene delle caratteristiche di novità che lo rendono unico anche all'interno del panorama internazionale di ricerca.

## 2 Corpus

Il punto di partenza per la creazione di CItA è rappresentato dalle trascrizioni delle produzioni scritte di studenti di sette diverse scuole secondarie di primo grado di Roma descritte da Barbagli et al. (2014). Le scuole considerate sono rappresentative di un ambiente socio–culturale che può essere definito medio–alto (il centro) e di uno medio–basso (la periferia). Per ogni scuola è stata individuata una classe, per un totale di 77 studenti in centro e 79 in periferia e per ogni studente sono state raccolte due tipologie di produzioni scritte: le tracce assegnate indipendentemente da ogni docente durante l'anno e due prove comuni a tutte le scuole. Il corpus, composto da 1.352 testi (366.335 tokens), comprende i testi prodotti da ogni studente durante il suo primo e secondo anno scolastico. Il corpus è accompagnato da un questionario che raccoglie alcune variabili di sfondo di ogni studente come ad esempio il background familiare (es. il lavoro e il titolo di studio dei genitori), territoriale (zona della scuola), personale (es. numero di libri letti).

Le principale novità di CItA riguardano il tipo di produzioni considerate (quelle di apprendenti di italiano L1), l'annotazione degli errori e l'ordinamento temporale delle produzioni all'interno di due anni scolastici consecutivi. Queste caratteristiche permettono di condurre uno studio sulle variazioni delle frequenze e delle tipologie di errori commessi al mutamento delle competenze linguistiche di ogni studente sia all'interno di uno stesso anno scolastico sia al passaggio dal primo e al secondo anno della scuola secondaria di primo grado. CItA rende inoltre possibile studiare le relazioni tra le variazioni di errori e le variabili di sfondo contenute nel questionario. L'attenzione posta sulla scuola secondaria di primo grado rappresenta un'ulteriore aspetto innovativo. Il primo biennio della scuola media è stato sino ad oggi poco indagato dalle ricerche empiriche nonostante sia un momento cardine nello sviluppo delle abilità linguistiche di uno studente.

## 3 Schema di Annotazione

La definizione dello schema di annotazione degli errori qui presentato si inserisce nel più ampio contesto degli studi condotti in ambito italiano sulla valutazione delle abilità linguistiche nella lingua materna (Corda Costa and Visalberghi, 1995; De Mauro, 1983; GISCEL, 2010; Colombo, 2011). Siccome l'attribuzione di errore ad una forma linguistica è un'operazione delicata poiché si presuppone il riferimento ad un sistema normativo, che di per sé non è oggettivo ma arbitrario, poiché basato su convenzioni sociali, per individuare gli errori abbiamo fatto riferimento al concetto di italiano standard neostandard individuato da Berruto (1987). L'analisi empirica della distribuzione delle varie tipologie di errori in CItA è stato un altro dei criteri adottati nel definire lo schema di annotazione. Sulla base di queste considerazioni, abbiamo scelto di annotare le tipologie di errori a cui si fa tradizionalmente riferimento in letteratura laddove la frequenza di occorrenza nel corpus fosse significativa. Come mostra la Tabella 1, che riporta lo schema di annotazione e le distribuzioni delle diverse categorie di errore considerate, abbiamo scelto di annotare errori riconducibili a tre macro–aree: grammatica, ortografia e lessico. Come indicato anche nel recente Rapporto sulla "Rilevazione degli errori più diffusi nella padronanza della lingua italiana nella prima prova di italiano"[1] redatto nel 2012 dall'INVALSI e dall'Accademia della Crusca, sono queste tre gli ambiti di competenza linguistica rispetto ai quali è possibile valutare la padronanza linguistica di uno studente. Seguendo la ripartizione suggerita dal Rapporto in descrittori specifici, per ciascuna competenza è stata prevista una categoria di errore corrispondente alla categoria morfosintattica coinvolta (colonna *Categoria* della Tabella 1). Inoltre, adottando la strategia suggerita da Granger (2003), per ogni categoria è stato individuato il tipo di modifica proposta per l'errore (colonna *Tipo di modifica*). Il formato di annotazione scelto è ispirato a quello messo a punto in occasione dello "Shared Task on Grammatical Error Correction" 2013. La frase seguente mostra un esempio estratto dal corpus dove sono stati annotati due errori:

> [...] io <M t="3.1" c="dovevo">avevo a</M> salire fin lassù ma mi sono <M t="2.1" c="fatta">fata</M> coraggio [...]

Il tag <*M*> (*Mistake*) e la sua relativa chiusura <*/M*> marcano l'area dell'errore annotato. <*M*> ha due attributi: *t* (*type*) il cui valore corrisponde al codice dell'errore e *c* (*correction*) il cui valore è la correzione dell'errore. In questo caso sono stati annotati due errori: un errore d'uso lessica-

[1] http://www.invalsi.it/download/rapporti/es2_0312/RAPPORTO_ITALIANO_prove_2010.pdf

| | | I anno | | | II anno | | | Totale % |
|---|---|---|---|---|---|---|---|---|
| **Categoria** | **Tipo di modifica** | Freq.% | Media | Dev. | Freq.% | Media | Dev. | |
| **Grammatica** | | | | | | | | |
| Verbi | **Uso dei tempi** | 7,78 (150) | 0,99 | 2,29 | 15,67 (239) | 1,47 | 4,05 | 11,26 (389) |
| | **Uso dei modi** | 4,25 (82) | 0,54 | 1,39 | 4,92 (75) | 0,49 | 0,99 | 4,55 (157) |
| | **Concordanza con il soggetto** | 2,85 (55) | 0,37 | 1,38 | 4 (61) | 0,41 | 1,27 | 3,36 (116) |
| Preposizioni | **Uso errato** | 6,48 (125) | 0,83 | 2,58 | 6,75 (103) | 0,66 | 1,21 | 6,6 (228) |
| | **Omissione o eccesso** | 1,03 (20) | 0,13 | 0,40 | 0,72 (11) | 0,07 | 0,25 | 0,90 (31) |
| Pronomi | Uso errato | 5,09 (98) | 0,65 | 1,13 | 3,54 (54) | 0,36 | 0,97 | 4,4 (152) |
| | **Omissione** | 0,41 (8) | 0,05 | 0,36 | 0,59 (9) | 0,06 | 0,39 | 0,49 (17) |
| | Eccesso | 2,70 (52) | 0,35 | 0,61 | 1,57 (24) | 0,16 | 0,46 | 2,2 (76) |
| | **Uso errato del pronome relativo** | 2,13 (41) | 0,27 | 0,70 | 1,70 (26) | 0,17 | 0,44 | 1,94 (67) |
| Articoli | **Uso errato** | 5,81 (112) | 0,75 | 3,72 | 3,54 (54) | 0,35 | 1,09 | 4,81 (166) |
| Congiunzioni e/o connettivi | Uso errato | 0,57 (11) | 0,07 | 0,33 | 0,52 (8) | 0,05 | 0,23 | 0,55 (19) |
| **Altro** | | 7,31 (141) | 0,94 | 3,66 | 5,18 (79) | 0,49 | 1,79 | 6,37 (220) |
| **Ortografia** | | | | | | | | |
| Doppie | **Uso per difetto** | 6,74 (130) | 0,83 | 2,49 | 5,05 (77) | 0,48 | 1,56 | 5,99 (207) |
| | Eccesso | 3,27 (63) | 0,42 | 0,89 | 3,67 (56) | 0,37 | 1,13 | 3,45 (119) |
| Uso dell'*h* | **Per difetto** | 3,21 (62) | 0,39 | 1,03 | 1,64 (25) | 0,17 | 0,62 | 2,52 (87) |
| | Per eccesso | 1,66 (32) | 0,21 | | 1,11 (17) | 0,10 | | 1,42 (49) |
| Monosillabi | **Uso errato dei monosillabi accentati** | 4,87 (94) | 0,63 | 1,07 | 4,07 (62) | 0,40 | 0,83 | 4,52 (156) |
| | Uso di *po* o *pò* anziché *po'* | 1,66 (32) | 0,21 | 0,72 | 1,64 (25) | 0,17 | 0,52 | 1,65 (57) |
| Apostrofo | **Uso errato** | 4,82 (93) | 0,61 | 1,01 | 4,52 (69) | 0,46 | 0,89 | 4,69 (162) |
| **Altro** | | 21,77 (420) | 2,76 | 4,58 | 23,02 (351) | 2,27 | 4,60 | 22,32 (771) |
| **Lessico** | | | | | | | | |
| Vocabolario | **Uso errato** | 5,60 (108) | 0,70 | 1,64 | 6,56 (100) | 0,66 | 1,09 | 6,02 (208) |
| **Numero totale di errori** | | 1929 | | | 1525 | | | |

Tabella 1: Schema di annotazione degli errori. Per ogni anno scolastico sono riportati: distribuzione percentuale degli errore e numero di occorrenze (*Freq.%*), occorrenza media degli errori per anno (*Media*), deviazione standard delle medie (*Dev.*). La colonna *Totale %* riporta la percentuale e il numero di occorrenze degli errori nei due anni. Gli errori che variano tra i due anni in modo statisticamente significativo all'analisi della varianza ($p<0.05$) sono stati marcati in grassetto.

le (*t="3.1"*) e un errore ortografico nell'uso per difetto delle doppie (*t="2.1"*).

In quanto segue riportiamo alcuni esempi di annotazione estratti da CItA che esemplificano alcune categorie di errori e le relative correzioni.

**Verbi: uso dei tempi**. [...] dopo aver fatto le squadre <M t="11" c="abbiamo">avevamo</M> subito iniziato a giocare [...]

**Verbi: uso dei modi**. [...] il pensiero che mi tormentava di più era che tra poco si <M t="12" c="sarebbe fatto">faceva</M> il campo scuola.

**Verbi: concordanza con il soggetto**. [...] la mia famiglia ed io <M t="13" c="stavamo">stavo</M> al mare a Torvajanica

**Preposizioni: uso errato**. <M t="14" c="in">a</M> Romania sono andata <M t="14" c="in">a</M> agosto [...]

**Pronomi: uso errato**. Proteggere i più deboli è molto coraggioso da parte di chi <M t="16" c="li">lo</M> protegge [...]

**Pronomi: eccesso**. Alla nostra maestra <M t="18" c="canc">gli</M> piaceva tanto la storia

**Pronomi: uso errato del pronome relativo**. La scienza non so perché mi fa pensare a un fenomeno costruito su un'altura <M t="19" c="per cui">che</M> ci vuole molto ingegno.

**Articolo: uso errato**. <M t="111" c="gli">i</M> dei, sapendo che qualcuno aveva preso senza merito il sacro vaso della Giustizia, si rattristarono molto, [...]

**Grammatica: altro**. Quando vedo <M t="10" c="quel">quelle</M> genere di <M t="10" c="persone">persona</M> mi sento strano.

**Vocabolario: uso errato**. C'era molta ombra nel giardino e io mi ci <M t="31" c="addormentavo">addormivo</M> sempre.

## 4 CItA per

Il corpus così annotato può avere diversi tipi di utilizzi. Dal punto di vista applicativo, CItA può essere usato come corpus di riferimento per sviluppare sistemi di identificazione e correzione automatica degli errori per la lingua italiana e/o per costruire modelli predittivi della competenza linguistica di un apprendente L1 (Richter et al., 2015). In quest'ultima direzione vanno gli studi che possono essere condotti confrontando le variazioni degli errori nel passaggio dal primo al secondo anno con i risultati del processo di monitoraggio delle caratteristiche linguistiche estratte dai testi linguisticamente annotati in modo automatico (Barbagli et al., 2014). Un esempio è quello della correlazione statisticamente significativa tra la distribuzione dei pronomi e il loro uso errato che diminuisce tra il primo e il secondo anno. Diminuisce ad esempio l'uso di pronomi personali e clitici, che sono usati in eccesso al primo anno, mentre rimane invariato l'uso di pronomi relativi ma cala la percentuale di errori che li coinvolgono. Il rapporto tra uso dei pronomi e relativi errori risulta pertanto predittivo dell'evoluzione nella competenza d'uso di questa categoria morfosintattica.

CItA può essere utilizzato per monitorare l'evoluzione degli errori nel tempo. La Tabella 1 riporta la distribuzione degli errori sia globalmente sia in ognuno dei due anni scolastici. Analizzando gli errori al passaggio tra primo e secondo anno, si può notare come la distribuzione di quelli ortografici e grammaticali (colonna *Totale* %) sia molto simile (rispettivamente 46,55% e 47,33%) mentre quelli lessicali sono nettamente meno (circa il 6%). Andando a valutare i singoli errori, quelli più frequenti sono quelli ortografici non classificati (22,32%) seguiti dall'uso errato dei tempi verbali (circa la metà dei precedenti), gli errori grammaticali non sottocategorizzati e l'uso errato delle preposizioni. Quando valutiamo la significatività delle variazioni degli errori tra i due anni, vediamo che quasi tutti (quelli marcati in grassetto) variano in maniera significativa, mostrando che esistono delle forti tendenze comuni nel passaggio dal primo al secondo anno. Studiando le distribuzioni di frequenza in modo separato per i due anni (colonna *Freq.*%) e la distribuzione media di ogni errore per anno (colonna *Media*), gli errori più diffusi sono quelli ortografici e grammaticali non sottocategorizzati, l'uso errato dei verbi, delle preposizioni, degli articoli, dei pronomi e l'uso per difetto delle doppie. Sebbene in generale il numero totale degli errori diminuisca nel passaggio dal primo al secondo anno, indagando come variano le distribuzioni di ogni categoria, scopriamo che non tutti i tipi di errore diminuiscono. Il caso più evidente è quello dell'aumento nel secondo anno dell'uso errato dei verbi in generale e dei tempi verbali in particolare. Ciò potrebbe essere riconducibile sia all'evoluzione dello studente sia al diverso tipo di tracce distribuite in classe dai docenti. Mentre nel primo anno le tracce assegnate sono per lo più di tipo narrativo, tipologia testuale che comporta l'uso di una sequenza temporale che potrebbe essere considerata più semplice da riconoscere e da costruire per gli studenti, al secondo anno aumentano le tracce di tipo argomentativo la cui struttura risulta più complessa. Questo ci porta a ipotizzare che gli studenti del secondo anno tentino di utilizzare forme verbali più complesse commettendo più errori. Questo è avvalorato dai risultati del monitoraggio linguistico automatico che rivelano come gli studenti al secondo anno usino strutture verbali più complesse (es. uso di ausiliari in tempi composti).

CItA può inoltre essere utilizzato all'interno di studi socio–pedagogici permettendo di mettere in relazione la distribuzione degli errori con le variabili di sfondo. È possibile così verificare in che misura i cambiamenti che avvengono nella scrittura sono attribuibili a condizioni socio–economiche di sfondo. È ad esempio interessante osservare come le esplorazioni statistiche condotte hanno rivelato che la diminuzione dell'uso errato del lessico dal primo al secondo anno è significativamente correlata con l'abitudine alla lettura. Oppure si può studiare come gli errori grammaticali variano in modo statisticamente significativo rispetto alla collocazione della scuola in centro o periferia di Roma: mentre nelle scuole del centro gli errori diminuiscono nel passaggio dal primo al secondo anno, in due delle quattro scuole in periferia aumentano. Diverso è il caso degli errori ortografici che non variano in modo statisticamente significativo rispetto alle variabili di sfondo considerate. Ciò confermerebbe alcuni studi (Colombo, 2011; Ferreri, 1971; Lavino, 1975; De Mauro, 1977) dove si afferma che la correttezza ortografica è un'abilità che si acquisisce con il tempo poiché richiede la sedimentazione di norme, spesso arbitrarie, che stabiliscono legami non causali tra suono e grafia.

## References


A. Abel, A. Glaznieks, L. Nicolas, and E. Stemle. 2014. KoKo: an L1 Learner Corpus for German. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 26–31.

I. Aldabe, L. Amoros, B. Arrieta, A. Díaz de Ilarraza, M. Maritxalar, M. Oronoz, L. Uria. 2005. Learner and Error Corpora Based Computational Systems. *Proceedings of the PALC 2005 Conference*, Poland.

C. Andorno and S. Rastelli. 2009. *Corpora di Italiano L2: tecnologie, metodi, spunti teorici*. Guerra Edizioni.

A. Barbagli, P. Lucisano, F. Dell'Orletta, S. Montemagni, and G. Venturi. 2014. Tecnologie del linguaggio e monitoraggio dell'evoluzione delle abilità di scrittura nella scuola secondaria di primo grado. *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it)*, 9–10 December, Pisa, Italy.

G. Berruto 1987. Sociolinguistica dell'italiano contemporaneo. Carocci, Roma.

A. Boyd, J. Hana, L. Nicolas, D. Meurers, K. Wisniewski, A. Abel, K. Schöne, B. Štindlová, and C. Vettori. 2014. The MERLIN corpus: Learner Language and the CEFR. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

A. Colombo. 2011. *"A me mi" Dubbi, errori, correzioni nell'italiano scritto*. Franco Angeli editore.

M. Corda Costa and A. Visalberghi. 1995. (a cura di) *Misurare e valutare le competenze linguistiche*. La Nuova Italia, Firenze.

D. Dahlmeier, H.T. Ng, and S.M. Wu. 2013. Building a large annotated corpus of learner English: The NUS Corpus of Learner English. *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 22–31.

P. Deane and T. Quinlan. 2010. What automated analyses of corpora can tell us about student's writing skills. *Journal of Writing Research*, 2(2):151–177.

T. De Mauro. 1983. Per una nuova alfabetizzazione. Gensini S., Vedovelli M.(a cura di) *Teoria e pratica del glotto-kit. Una carta d'identità per l'educazione linguistica*. Franco Angeli Milano.

T. De Mauro. 1977. *Scuola e linguaggio*. Editori Riuniti, Roma.

M. Dickinson and S. Ledbetter. 2012. Annotating Errors in a Hungarian Learner Corpus. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

S. Ferreri. 1971. Italiano standard, italiano regionale e dialetto in una scuola media di Palermo. Medici M.-Simone R. (a cura di) *L'insegnamento dell'italiano in Italia e all'estero*, I, Roma, Bulzoni, 1971, pp. 205-224.

GISCEL Emilia-Romagna. 2010. La correzione dei testi scritti. Lugarini E. (a cura di) *Valutare le competenze linguistiche*. Franco Angeli Milano, pp. 188-203

S. Granger. 2003. Error-tagged Learner Corpora and CALL: A Promising Synergy. *CALICO Journal*, 20:465–480.

C. Lavino. 1975 *L'insegnamento dell'italiano. Un'inchiesta campione in una scuola media sarda*. Edes, Cagliari.

A. Lüdeling, M. Walter, E. Kroymann, and P. Adolphs. 2005. Multi–level error annotation in learner corpora. *Proceedings of Corpus Linguistics 2005*.

H.T. Ng, S.M. Wu, Y. Wu, C. Hadiwinoto, and J. Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–12.

H.T. Ng, S.M. Wu, T. Briscoe, C. Hadiwinoto, R.H. Susanto, and C. Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–14.

S. Richter, A. Cimino, F. Dell'Orletta, and G. Venturi. 2015. Tracking the Evolution of Language Competence: an NLP–based Approach. *Proceedings of the 2nd Italian Conference on Computational Linguistics (CLiC-it)*, 2–3 December, Trento, Italy.

W. Zaghouani, N. Habash, H. Bouamor, A. Rozovskaya, B. Mohit, A. Heider, K. Oflazer. 2015. Correction Annotation for Non-Native Arabic Texts: Guidelines and Corpus. *Proceedings of The 9th Linguistic Annotation Workshop*, pp. 129–139.