DAL DATO ALLO STRUMENTO

Quando l'errore è un valore: questioni teoriche e pratiche nell'allestimento del LISSICS (Lessico dell'Italiano Scritto della Svizzera Italiana in Contesto Scolastico)

Luca CIGNETTI*, Silvia* DEMARTINI

ABSTRACT • In this paper we introduce the project TIscrivo (and TIscrivo2.0), providing an overview of its main features and focusing particularly on the preparation of LISSICS (*Dictionary of Italian Language Written in Italian Switzerland at School*): a resource designed to be useful both to scholars and to teachers. The ongoing process of lemmatisation and PoS tagging is facing many challenges, especially in managing orthographic errors. Also, we will illustrate some of the most relevant orthographic mistakes in the corpus, composed of texts written by primary and lower secondary school students.

KEYWORDS • Italian Learner Corpus, Natural Language Processing, Written Italian Language, Orthography, Language Teaching, Language Acquisition, Children's Writing Competences

1. I progetti di ricerca TIscrivo e TIscrivo2.0

L'analisi di corpora dedicati a specifici tipi di testo in cui la lingua presenta caratteristiche spesso non conformi alla lingua *standard* (su cui Berruto 2010) è una possibilità ricca e complessa. È ricca in quanto permette di addentrarsi nell'esame e nello studio di fenomeni che non sempre emergono da raccolte di testi dalle caratteristiche complessivamente vicine alla norma; è complessa perché testi che presentano, a più livelli, molti tratti devianti rispetto alla norma di riferimento pongono non pochi problemi a livello di trattamento automatico.

I testi raccolti nell'ambito delle ricerche TIscrivo (DoRe 13DPD3_136603 La scrittura oggi, tra parlato e lingua mediata dalla rete. Aspetti teorico-descrittivi, diagnosi e interventi didattici) e TIscrivo2.0 (FNS 100012_156247 Scrivere a scuola nel terzo millennio. Descrizione della varietà e del vocabolario dell'italiano scritto in contesto scolastico ticinese e implicazioni didattiche)[1] sono un caso esemplare in questo senso. Infatti, il corpus in esame − d'ora in poi corpus "DFA-TIscrivo" −, raccolto nel 2012, cioè nella fase iniziale del progetto, è composto da 1.735 testi scritti a scuola da bambini e ragazzi tra gli 8 e i 14 anni, in Canton Ticino (cfr. par. 1.1) ed è perciò rappresentativo di una particolare varietà di scrittura: quella dei giovani scriventi in contesto scolastico.

L'obiettivo di fondo dei lavori è quello di rinnovare la riflessione intorno alla didattica della scrittura sulla base di dati di prima mano che contribuiscano a delineare il quadro linguistico

[1] Si tratta di due fasi triennali di uno stesso progetto in continuità (2011-2014 e 2014-2017), finanziato dal Fondo Nazionale Svizzero per la Ricerca Scientifica e condotto dal *Centro di Didattica dell'italiano e delle lingue nella scuola* del DFA (Dipartimento Formazione e Apprendimento) della SUPSI (Scuola Universitaria Professionale della Svizzera Italiana), a Locarno, sotto la responsabilità di Simone Fornara.

attuale, permettendo di elaborare proposte didattiche innovative per la scuola primaria e secondaria di primo grado. Il primo triennio della ricerca, terminato nel 2014, ha portato alla pubblicazione di un volume di saggi dedicati a fenomeni e ad aspetti specifici della lingua usata dai bambini e dai ragazzi (Cignetti *et al*. 2016a), oltre che all'applicazione dei risultati delle analisi dei testi nell'ambito della formazione di base e continua degli insegnati. Invece, la seconda fase dei lavori ha tra gli obiettivi principali quello di realizzare il LISSICS, cioè il *Lessico dell'Italiano Scritto della Svizzera Italiana in Contesto Scolastico*: uno strumento che vorrebbe essere utile tanto agli studiosi di linguistica quanto agli insegnanti, che in esso vedrebbero delineate e sistematizzate alcune delle tendenze di scrittura che emergono quotidianamente nelle produzioni degli allievi[2].

1.1. Il corpus: i dati in analisi

Come si è accennato, il corpus è costituito da testi di tipo narrativo-riflessivo redatti a mano in contesto scolastico in risposta alla consegna di scrittura descritta in Fornara *et al.* (2015) e in Cignetti *et al.* (2016b). Analoga per i diversi ordini di scolarità, la consegna sollecitava lo stesso tipo di azione cognitiva e di produzione testuale, ma era differenziata dalla proposta di due diversi testi-stimolo di partenza: la favola *La tartaruga e la lepre* di Esopo per i bambini di scuola elementare e il racconto *Il giardino segreto* di Italo Calvino per i ragazzi di scuola media. Eccone la formulazione:

> Dopo aver letto e analizzato in classe il racconto di Calvino/la favola di Esopo, ti è stato chiesto di pensare a un episodio che hai vissuto o a cui hai assistito dal quale hai ricavato un insegnamento. Raccontalo ora in forma scritta (minimo una pagina, massimo due pagine) e spiega che cosa ti ha insegnato.

Il campione di testi raccolti in modo omogeneo sul territorio del Canton Ticino risulta così composto (SE sta per scuola elementare, SM sta per scuola media[3]):

Ordine scolastico	SE	SM
Numero istituti	35	21
Numero classi	48 (24 di III, 24 di IV)	51 (25 di II, 26 di IV)
Numero testi	742	993
Totale testi	1.735	

Tab. 1 – Numeri di istituti, classi e testi raccolti del corpus "DFA-TIscrivo".

[2] Date le dimensioni del corpus "DFA-TIscrivo", si tratterà di uno strumento comparabile, per estensione, al LIP di De Mauro *et al.* (1993), al LIPSI di Pandolfi (2009) e al *Lessico elementare* di Marconi *et al.* (1994). Quest'ultimo è lo studio nell'insieme più affine al LISSICS, in quanto offre l'analisi di due corpora da 500.000 parole ciascuno (uno è costituito da componimenti di un campione bambini di scuola elementare, l'altro da una selezione delle più diffuse letture per l'infanzia). Per il LISSICS si prevede tuttavia un formato non solo cartaceo, ma anche consultabile in formato digitale.

[3] Secondo le denominazioni in uso in Canton Ticino, dove la scuola media ha durata quadriennale.

I testi, trascritti in formato elettronico (e successivamente controllati), codificati[4] e predisposti per il trattamento con software specifici, offrono circa 391.250 parole grafiche[5]. Questo numero, che potrà considerarsi definivo solo al termine del lavoro di pulizia e trattamento del corpus, include «ogni sequenza di caratteri separata dalle altre da uno spazio bianco o da un segno di interpunzione» (De Mauro 2005: 14), e tiene anche conto della tokenizzazione separata delle forme elise (*c'era* = due token). La punteggiatura è conteggiata a parte. Le peculiarità dei dati ai diversi livelli linguistici (ortografia, morfologia e sintassi, ma anche semantica e organizzazione testuale) rendono necessaria una progettazione mirata del lessico di frequenza (LISSICS). In particolare, il *Lessico*, oltre a raccogliere le parole dei testi con relativa annotazione per parti del discorso, vorrebbe non perdere la dimensione dell'errore (in particolare grafico e morfologico) e proporre un lemmario quanto più possibile aderente alla realtà linguistica (cioè, per esempio, che lemmatizzi le polirematiche). Il raggiungimento di questi obiettivi va considerato in relazione alle possibilità offerte dal trattamento automatico del linguaggio e, in questa prospettiva, si stanno cercando gli strumenti più efficienti a disposizione per la lingua italiana, con l'obiettivo di focalizzare l'intervento manuale solo alle fasi e ai casi in cui resta necessario; tenendo conto che, limitando l'osservazione ai tagger, spesso la qualità delle performance cala quando operano su testi lontani dall'italiano standard, ci si propone di migliorarne, se possibile, le funzionalità[6].

2. Dai dati alle informazioni: gli obiettivi delle ricerche e i software in uso

L'esplorazione di dati linguistici pone sempre la sfida dell'estrazione di informazioni a partire da dati grezzi, in modo da renderle accessibili e consultabili in modo esaustivo e agevole, perdendo meno informazione possibile. In estrema sintesi, le prospettive scientifiche e applicative delle ricerche *TIscrivo* e *TIscrivo2.0* (alcune delle quali già in parte sviluppate[7]) si possono riassumere nei punti seguenti:

- delineare i tratti tipici della scrittura delle giovani generazioni in contesto scolastico su diversi livelli di analisi (dall'ortografia alla testualità), sulla base del più grande corpus di testi scolastici finora raccolto nella Svizzera Italiana;

[4] Sono a disposizione i seguenti metadati: sesso, ordine scolastico e classe, ubicazione della scuola, informazioni circa la provenienza linguistica dell'allievo (fornite dai docenti).

[5] La scuola elementare contribuisce con il 28% delle parole (circa 109.000) e la scuola media con il 72% (circa 281.000). Analisi sulla lunghezza media di testi e frasi, sulla leggibilità dei testi e sulla caratterizzazione lessicale sono in corso con Corrige!it (http://www.corrige.it/) e con READ-IT (http://www.italianlp.it/demo/read-it/). Ad esempio, una prima analisi con READ-IT del subcorpus di 3SE (per la quale si ringraziano molto Felice Dell'Orletta e Giulia Venturi) rileva, tra le altre cose, che i testi dei bambini più piccoli sono composti mediamente da 6,5 periodi, che ogni periodo è composto mediamente da 25 token e che ogni testo conta in media 136 token. Si tratta di dati interessanti da confrontare in prospettiva evolutiva, mettendoli in relazione con altri aspetti della competenza scrittoria al crescere dell'ordine e del grado di scolarità.

[6] Dal 2017 l'équipe di ricerca si avvale di una collaborazione col Dipartimento Tecnologie Innovative della SUPSI (Manno), cosa che permetterà un più approfondito e specializzato supporto informatico per cercare e sperimentare soluzioni ai fini del progetto.

[7] Oltre al già citato Cignetti, Demartini, Fornara (2016a), che raccoglie un'ampia serie di contributi dedicati ai vari oggetti di studio approfonditi nella prima fase della ricerca, si vedano Fornara *et al.* (2015), Cignetti *et al.* (2016b) e Demartini (2016).

- rilevare e studiare gli eventuali influssi della Comunicazione Mediata dal Computer sulla scrittura manuale a scuola, cioè su un tipo di scrittura legata a un mezzo e a un contesto diversi (in cui possono però verificarsi frequenti mescolamenti impropri di codici e di registri, soprattutto nei testi dei ragazzi più grandi);
- costruire il LISSICS (*Lessico dell'Italiano Scritto della Svizzera Italiana in Contesto Scolastico*), nella duplice prospettiva di affinare l'analisi automatica del tipo di testi in esame e di offrire, poi, uno strumento di consultazione agile, utile sia per gli specialisti, sia per un pubblico più vasto (come può essere quello degli insegnanti);
- fornire indicazioni didattiche mirate a partire dal quadro emerso dalle analisi, in particolare dai principali errori individuati.

Per lavorare a questi obiettivi, al momento il corpus è trattato e analizzato principalmente con l'ausilio di tre software, destinati a operazioni diverse: Atlas.ti, T-LAB e TreeTagger. Il primo strumento è usato per etichettare e codificare fenomeni di interesse presenti nei testi (vale a dire aspetti specifici – a diversi livelli – studiati dai ricercatori, come, per citare qualche esempio, l'uso del *che* relativo, l'uso del *ma*, l'impiego di similitudini e metafore, i titoli dei testi ecc.): il programma permette di consultarli ed esaminarli tutti insieme, mantenendo, al contempo, il riferimento alle produzioni del corpus in cui essi si trovano; il secondo (su cui Lancia 2004) è in uso prevalentemente per effettuare analisi tematiche e alcune analisi lessicali (ad esempio concordanze e contesti elementari); il terzo è quello individuato per la costruzione del lessico: effettua infatti PoS tagging e lemmatizzazione[8].

2.1 L'allestimento del LISSICS: difficoltà e possibilità nel trattamento del corpus

Come si è accennato, l'italiano dei testi del corpus "DFA-TIscrivo" presenta una varietà di tratti significativi per chi si occupa di studiare la lingua scritta da un particolare tipo di apprendenti (cioè da bambini e ragazzi prevalentemente italofoni L1). È fondamentale non perderli, ma, anzi, riuscire a renderne conto sia a fini descrittivi (per tracciare un quadro documentato della scrittura sui banchi di scuola), sia in termini di accesso agevole alle informazioni. Nella pratica, come mostrano anche altre recenti analisi che presentano alcune affinità con quella qui presentata (in particolare Spina 2014 e Barbagli *et al.* 2015), si tratta di un'operazione tutt'altro che semplice.

Per fornire un quadro d'insieme del corpus, il gruppo di ricerca ha preso in considerazione tutti i livelli linguistici, dapprima attraverso studi qualitativi mirati (cioè dedicati a specifici fenomeni ricorrenti nei testi). In questa sede, limitiamo le considerazioni ad alcune questioni pratiche legate all'allestimento del LISSICS, sintetizzando così le principali difficoltà operative: non perdere niente dei dati dal punto di vista grafico e morfologico, e renderli conformi a uno standard completo e funzionale, agevole, poi, per la consultazione (cartacea e digitale). Nei prossimi paragrafi, limitandoci a ciò che è di stretta pertinenza del LISSICS (cioè al PoS tagging e alla lemmatizzazione), ci soffermeremo su alcuni esempi di casi critici e poi in modo specifico sugli errori ortografici, particolarmente caratteristici e significativi nel corpus (pertanto si è scelto di non normalizzarli, bensì di tenerne traccia).

[8] Al momento si stanno testando anche altri tagger per valutarne le prestazioni.

2.1.1. Alcuni esempi di casi critici

Il software attualmente in uso per annotare le forme per parti del discorso e ricondurle a lemma è TreeTagger[9], con i parametri per l'italiano di Marco Baroni. Il primo output del programma, cioè la prima restituzione del corpus (il cui testo si può leggere nella colonna di sinistra) con etichette grammaticali e riconduzione a lemma, si presenta come nelle tabelle seguenti (Tabb. 2 e 3):

Forma	PoS Tag	Lemma
rimasimo	VER:fin	<unknown>
solo	ADV	solo
in	PRE	in
pantaloncini	NOUN	pantaloncino
corti	ADJ	corto
.	SENT	.
Ci	CLI	ci
siamo	AUX:fin	essere
rimasti	VER:ppast	rimanere
un'ora	VER:fin	<unknown>
in	PRE	in
piscina	NOUN	piscina

Tab. 2 - La prima restituzione di TreeTagger: esempi di "unknown".

Io	PRO:pers	io
avevo	VER:fin	avere
paura	NOUN	paura
ma	CON	ma
non	NEG	non

[9] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/, Schmid 1994.

osavo	VER2:fin	osare
dirlo	VER:infi:cli	dire
perché	WH	perché
mi	CLI	mi
avrebbero	AUX:fin	avere
presa	VER:ppast	prendere
in	PRE	in
giro	NOUN	giro

Tab. 3 – La prima restituzione di TreeTagger: l'etichetta *WH* e le espressioni polirematiche.

Questi stralci mostrano alcuni casi rispetto ai quali è stato necessario operare delle scelte e intervenire, quando possibile, in fase di input del corpus, affinando il processo di tokenizzazione (ad esempio delle parole apostrofate). La tabella 2 mostra infatti due casi di "unknown", cioè di forme di fronte alle quali il programma non riesce a fornire una risposta e commette errori: in un caso, si tratta di un passato remoto errato, *rimasimo* (che il software etichetta correttamente ma non lemmatizza), nell'altro, dell'elisione corretta *un'ora*. La tabella 3 mostra invece un caso di etichettatura "WH" (*Wh words*) applicata in modo non pertinente (qui *perché* è una congiunzione causale, e non introduce un'interrogativa) e una polirematica (*avrebbero presa in giro* > *prendere_in_giro*) che, nell'idea di lessico che si vorrebbe realizzare, andrebbe considerata come un'unica unità polilessicale, e che invece il software tratta parola per parola, seguendo un altro criterio.

L'individuazione e la tipologizzazione di casi critici ricorrenti come quelli citati, rispetto ai quali non vi è una restituzione corretta di TreeTagger (o non pienamente idonea al risultato finale desiderato), rende necessario progettare misure correttive. Al momento, si stanno valutando e testando le seguenti possibilità: orientare da subito la lemmatizzazione agli obiettivi finali (ad esempio a livello di identificazione delle polirematiche, che possono essere almeno in parte individuate modificando i parametri del software); procedere con il controllo manuale dell'annotazione di una parte di corpus, in modo da avere un gold standard con cui confrontare le prestazioni di tagger diversi; provare ad automatizzare quanto più possibile l'azione di verifica della distanza tra forme grafiche errate e forme corrette, associando al programma degli algoritmi di controllo [10]. Quest'ultimo intervento sarebbe particolarmente efficace per migliorare il trattamento degli errori ortografici.

2.1.2 Un caso specifico: gli errori ortografici nel processo di lemmatizzazione

Gli errori di ortografia sono estremamente diffusi nel corpus "DFA-TIscrivo", com'è lecito aspettarsi dal tipo di testi che lo compongono, e rappresentano uno scoglio non semplice da superare nell'allestimento del LISSICS. Infatti, essi sono un caso esemplare di contrasto,

[10] Che considerino per esempio la *distanza di Levenshtein* o di *editazione* (l'unità di misura che considera il numero minimo di modifiche elementari che consentono di trasformare una stringa in un'altra, cfr. ad esempio Naldi 2014: 46-49).

significativo e complesso, fra le possibilità di trattamento automatico del linguaggio e la realtà linguistica, cioè fra lo strumento e il dato. Anzi, sarebbe più esatto dire tra i dati e *gli* strumenti, intendendo con "strumenti" sia quelli usati in fase di analisi, sia il prodotto finale (lo strumento di consultazione che, si auspica, verrà ricavato dal trattamento dei dati).

Errori ortografici come *stó*, *fù*, *pò*, *finche*, *inizziato*, *soppratutto*, *pichiarci*, *gnente* ed errate segmentazioni del tipo *linsegnamento*, *laveva*, *perterra*, *glie lo dato*, *emmezo*, *or'mai* sono solo pochissimi esempi delle molte varianti che caratterizzano il materiale in analisi. Si tratta di forme di indubbio interesse e valore per uno studio che si occupa dei tratti dell'italiano scritto dei bambini e dei ragazzi (prevalentemente italofoni), forme che sovente rivelano con evidenza alcune radicate difficoltà nell'acquisizione e nell'affinamento del codice scritto: una grafia come *ce lo* per *ce l'ho* adombra ad esempio un profondo problema di riflessione grammaticale[11] (su questo tipo di errori, detti *ortografico-morfologici*, cfr. Cignetti e Demartini 2016: 75-76), cui conseguono, nella costruzione di uno strumento come il LISSICS, problemi di attribuzione dell'esatta parte del discorso e di lemmatizzazione. Come mantenere la forma grafica (*lo*) e, allo stesso tempo, ricondurre correttamente a lemma ("lo" = *l'ho* = *lo* + *avere*)? E lo stesso vale per un caso di concrezione dell'articolo come quello nella tabella 4 (*laveva*):

il	ART	il
mio	DET:poss	mio
allenatore	NOUN	allenatore
laveva	VER:fin	<unknown>
sgridato	VER:ppast	sgridare

Tab. 4 – L'univerbazione errata *laveva* e la risposta del software.

Com'è lecito aspettarsi, TreeTagger non può gestire automaticamente un'univerbazione simile, che andrebbe correttamente rappresentata secondo questa intenzione[12], conservando la forma realmente scritta ma segmentandola anche nelle parole (clitico + verbo) che la compongono:

```
1-2   laveva
      1     l       > lo
      2     aveva   > avere
```

Due parole (1 e 2) rappresentate in un unico token grafico improprio: si tratta di una delle sfide sottese alla realizzazione di un lessico come quello qui presentato, interessato al mantenimento del repertorio di forme errate; una sfida che mette in luce come le parole vadano intese sia nella loro componente grafica (parola come sequenza di caratteri, per cui *laveva* è una sola parola), sia in quella analitica profonda (per cui *laveva* sono due parole). Nessuna di queste

[11] Non è sempre semplice distinguere e rubricare un errore come meramente grafico oppure, invece, grammaticale in senso esteso. Pur non ambendo, per ora, a una tipologizzazione estremamente dettagliata degli errori, si tiene come riferimento (per l'affinità del materiale in analisi) l'elenco di errori di Barbagli *et al*. (2016: 93).

[12] Il modello di rappresentazione di riferimento è il formato CoNLL-U di Universal Dependencies (http://universaldependencies.org/format.html).

due componenti può essere lasciata da parte in un lavoro il cui fine principale è orientato all'individuazione delle difficoltà degli scriventi.

Analogamente, anche errori ortografici come quelli prima citati (uso delle doppie lettere, accenti sbagliati, *h* mancanti o in eccesso ecc.) compromettono il trattamento automatico dei dati, inducendo il software a restituire un "unknown" accanto alla parola contenente un errore (Tab. 5).

e	CON	e
anchè	VER:fin	<unknown>
mio	DET:poss	mio
papà	NOUN	papà

Tab. 5 – L'accento improprio su *anchè* induce in errore il tagger e causa la mancata lemmatizzazione.

È proprio nel caso dei moltissimi errori grafici come questo, per i quali la revisione manuale rischia di essere non solo molto lunga, ma anche passibile di imprecisioni, che si mostrerebbe particolarmente utile e fruttuoso automatizzare quanto più possibile la riconduzione al lemma di riferimento di forme relativamente poco distanti da quella corretta (*anchè-anche*, *senò-se no*, *pichiarci-picchiarci* ecc.).

3. Le principali tipologie di errore ortografico nel corpus DFA-TIscrivo

Al fine di illustrare le possibilità di estrazione di specifiche tipologie di errore offerte dal corpus "DFA-TIscrivo", mostreremo di seguito alcuni esempi relativi al livello dell'ortografia[13].

In primo luogo, in tutti e quattro i sottocorpora (3a e 5a SE, 2a e 4a SM) sono stati riscontati numerosi errori d'uso dell'accento diacritico sui monosillabi; entro questa categoria, vale da esempio l'errore di accento sulla *e* con valore copulativo, diffuso nei sottocorpora di SE ma attestato anche nel sottocorpus di 4a SM:

(1) il primo tiro l'ho fatto io però non *e* stato un bel tiro. [3a SE]
(2) *e* stato un addio piuttosto doloroso. [4a SM]

Ben distribuiti sono anche gli errori relativi ad altri monosillabi con accento diacritico, come *là ~ la*, *sé ~ se*, *sì ~ si*, *dà ~ da* o *lì ~ li*, quest'ultimo presente nell'esempio (3):

(3) Un giorno sono andato a un pranzo, e *li* c'era un fiume. [3a SE]

Ampiamente attestato è anche l'impiego dell'accento sui monosillabi che non ne richiedono mai la presenza, e anche in questi casi la distribuzione si concentra soprattutto nei sottocorpora di SE, benché non manchino occorrenze in quelli di SM:

(4) 2 settimane *fà* io e la mia clase siamo andati allo stadio di Cornaredo. [3a SE]
(5) Un paio di giorni *fà* ho fatto una gara di ginnastica. [5a SE]

[13] Per approfondimenti circa i dati qui sinteticamente presentati, cfr. Cignetti (2016).

(6) l'anno prossimo giocherò con i due difensori di due anni *fà* che hanno lasciato il Lugano. [4a SM]
(7) terra ovunque e qualche filetto verde *quà* e là. [2a SM]
(8) Quando sono arrivata *quà* non sapevo niente. [4a SM]
(9) A *mè* non è mai capitato di avere veri problemi con gli amici. [2a SM]
(10) non *stò* qui ha raccontare tutto perché sono veramente tante cose. [3a SE]

Per quanto riguarda gli errori d'impiego dell'apostrofo, risultano invece attestate in tutti i 4 sottocorpora le forme *po* e *pò* e le forme *qual'è* e *qual'era*:

(11) la strada era un *po* in pendenza. [4a SM]
(12) ogni minuto si fermava per chiedere *qual'era* la nota. [3a SE]

Numerosi sono anche i casi di apostrofo dopo l'articolo indeterminativo maschile *un*:

(13) Il custode era *un'uomo* sulla cinquantina di statura media, *un'uomo* molto severo e specialmente cattivo con i bambini e con i ragazzi. [2a SM].

Inoltre, limitatamente alla SE, si contano anche alcuni casi di apostrofo dopo *ad*:

(14) *Ad' un certo punto* mia sorella disse. [3a SE]

In merito all'uso della *h*, è emersa la tendenza all'impiego di più grafie alternative per il verbo *avere* (sono per esempio attestate in 3a SE le forme *ò*, *o*, *ànno* e *anno*):

(15) Quindi *o preso* l'ancora e sono risalito su in cima. [3a SE]
(16) Adesso *ò imparato* che se passo nelle scorciatoie che non so bene non ci devo andare. [3a SE]

E anche di questi errori non mancano i casi nei sottocorpora di SM:

(17) non bisogna mentire perché le bugie *anno* le gambe corte. [2a SM]

Né mancano gli impieghi di *h* in eccesso (es. 18) o gli errori dovuti alla collocazione del digramma *ch* prima di *a* e di *o* (ess. 19-21):

(18) Grazie *ha* questo episodio ho imparato che non bisogna mai sottovalutare. [5a SE]
(19) E loro mi hanno detto di *giochare* a calci di rigore. [5a SE]
(20) ho giocato a calcio con dei miei amici, nei *parcho* giochi dei pallazzi. [5a SE]
(21) le ossa mi dolevano e la testa sembrava che dovesse *schoppiarmi.* [4a SM]

Una categoria di errore diffusa quasi esclusivamente nel sottocorpus di 3 SE è invece l'errata segmentazione delle parole, il cui tipo più frequente riguarda l'errata discrezione dell'articolo, tale da dare luogo a forme come *linsegnamento*, *lostesso* e *laltro*:

(22) E se non rispetto *linsegnamento* rischio di perdermi ancora. [3a SE]
(23) con la porta ben chiusa ma sentivo *lostesso* mia mamma che mi diceva: smettila! [5a SE]
(24) se te lo vedi sbalzare fuori da un angolo lo spavento te lo prendi *lostesso*. [4a SM]
(25) Passammo ai rigori un rigore dopo *laltro* lui inizio tiro palo gol. [3a SE]

Sono inoltre attestati esempi di errata fusione dei costituenti di sintagma, come *perquello* e *dasola*:

(26) lei era intenta a giocare *perquello* non mi ascoltò. [3a SE]
(27) C'era un giardino medio, per andare con il bus *dasola*. [3a SE]

Altra categoria di errore molto comune riguarda l'impiego delle *doppie*, che può realizzarsi come errato scempiamento (tra gli esempi le forme *abiamo*, *arabbiata*, *tapeto*, *arivata*: cfr. ess. 28-31) o come errato raddoppiamento (ess. 32-34):

(28) siamo andati al Penz a fare un pic-nic *abiamo* grigliato la carne, le salciccie, i sevelà arancioni e bianchi (3a SE)
(29) Mi ero *arabbiata* tanto e buttavo in giro le cose perché ero furiosa. [3a SE]
(30) Pero sto sempre sdraiata sul *tapeto* perché mi fa sempre calmare. [3a SE]
(31) Quando sono *arivata* a scuola, la maestra mi ha *avertito* di non ritornare a casa con il motorino. [5a SE]

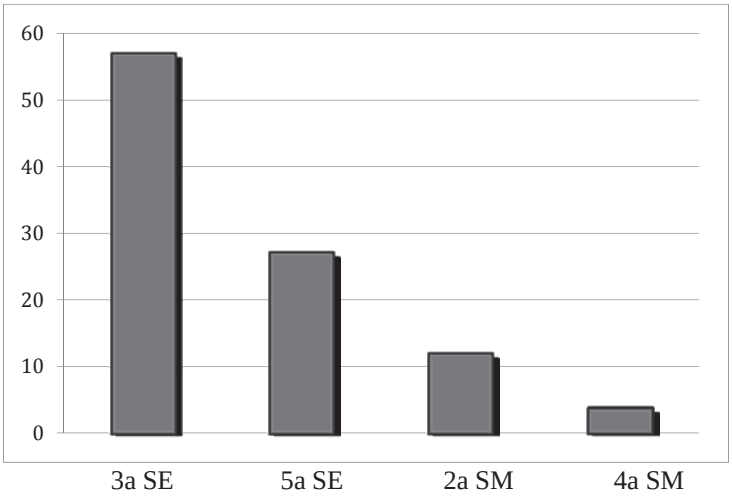
(32) Io mi sono *comportatto* bene solo che lui. [3a SE]
(33) A scuola feci *lezzione* bagnando tuti i fogli che passavano sotto il nio naso. [5a SE]
(34) L'amica di mia madre mi ha raccontato di una vicenda molto *corraggiosa.* [2a SM]

Questa tipologia di errore ortografico è evidentemente dovuta all'influenza della pronuncia regionale ticinese, assimilabile, per questo tratto, a quella dell'Italia settentrionale[14].

4. Distribuzione degli errori ortografici nei diversi livelli scolastici

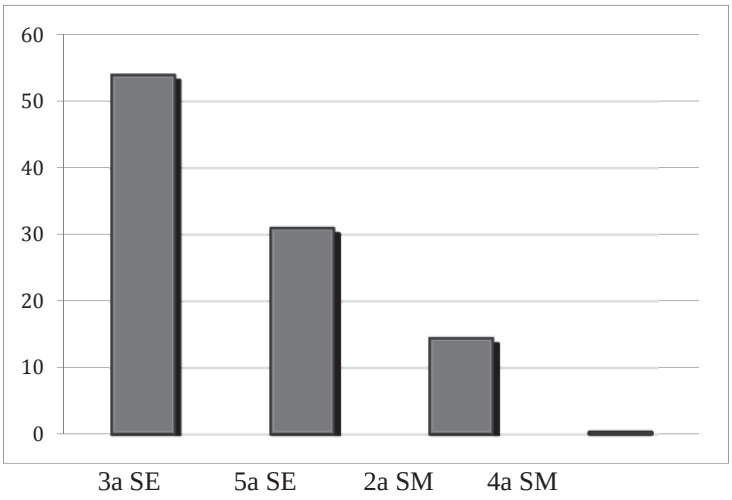
Proponiamo di seguito i risultati relativi ad alcune voci-campione in merito alla distribuzione delle diverse categorie di errore nei quattro sottocorpora. Un primo esempio riguarda l'errore di accento nelle parole *perché*, *però* e *può*. In questo caso, i dati, normalizzati, mostrano che oltre la metà degli errori si concentra nel corpus di 3a SE, con una riduzione piuttosto graduale nelle classi successive (Tab. 5).

[14] Cfr. De Blasi (2014: 65), dove, a proposito dei "fenomeni fonetici presenti in generale in tutta l'Italia settentrionale", si cita "lo scempiamento delle consonanti intense (per es. *afetare* invece di *affettare*)". Sul problema della norma ortografica, cfr. Cignetti & Demartini (2016).

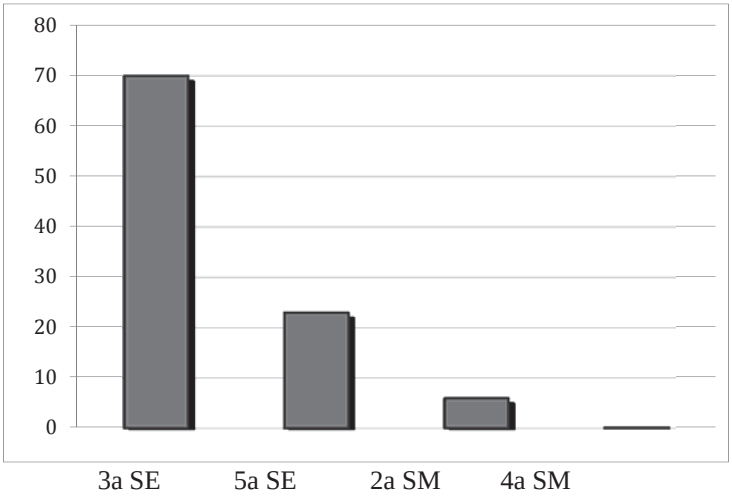


Tab. 5 – Occorrenze delle forme **perche*, **pero*, **puo*.

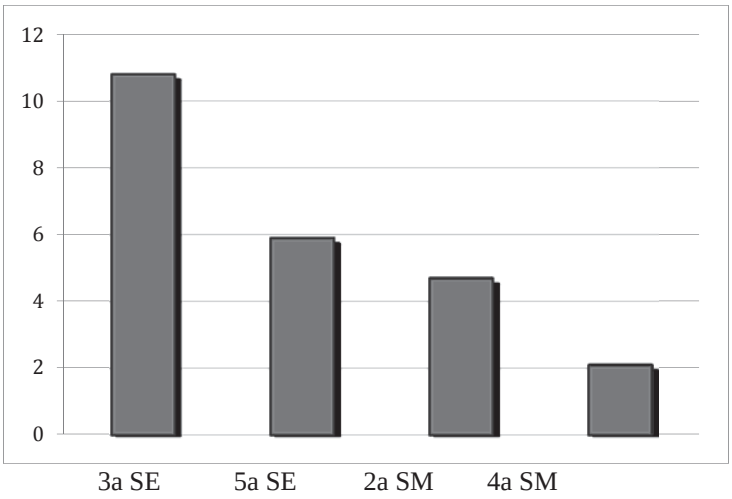
I dati disaggregati permettono di illustrare tale progressione con maggiore precisione, mostrando come la riduzione sia graduale con le voci *perché* e *può*, mentre nel caso di *può* il miglioramento nella fase di passaggio tra SE e SM non appare completamente soddisfacente, un fatto che può essere interpretato come segnale della necessità di un migliore coordinamento tra i diversi cicli scolastici (Tabb. 6-8):



Tab. 6 – Occorrenze della forma **perche*

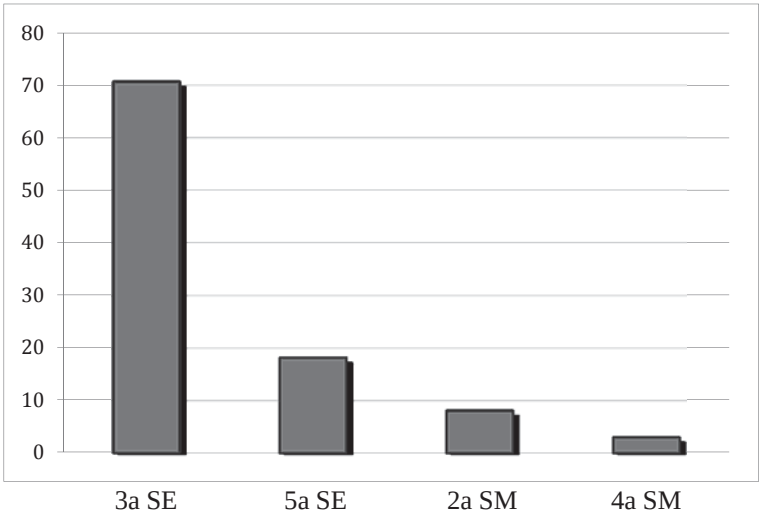


Tab. 7 – Occorrenze della forma **pero*.



Tab. 8 – Occorrenze della forma **puo*.

Ulteriori dati esemplificativi di questa fase della ricerca, relativi agli accenti in eccesso su alcuni monosillabi, sono riprodotti nelle Tab. 9 e 10: nella prima, i dati aggregati di dieci forme campione consentono di osservare una riduzione significativa degli errori tra la 3a SE e la 5a SE, mentre nella seconda l'osservazione della distribuzione degli errori nei diversi cicli di studio permette di comprendere meglio il fenomeno.



Tab. 9 – Occorrenze delle forme *pò*, *sù*, *stò*, *chè*, *dò*, *ò*, *lò*, *và*, *hà*, *nò*.

	3 SE	5 SE	2 SM	4 SM
pò	V	V	V	V
sù	V	V	V	-
stò	V	V	V	-
chè	V	V	-	-
dò	V	V	-	-
ò	V	-	-	-
lò	V	-	-	-
và	V	-	-	-
hà	V	-	-	-
nò	V	-	-	-

Tab. 10 – Distribuzione delle forme *pò*, *sù*, *stò*, *chè*, *dò*, *ò*, *lò*, *và*, *hà*, *nò*.

Se in 4a SM i problemi (tra quelli selezionati in queste dieci forme) si limitano dunque, sostanzialmente, al solo *po'* accentato, in 2a SM risultano invece attestate, oltre a numerosi casi di *pò*, anche diverse occorrenze di *sù* e di *stò*; in 5a SE compaiono molti casi di *dò* e di *chè* (pronome e complementatore) e in 3a SE risultano molto comuni altri problemi d'uso dell'accento, come mostra la presenza delle forme *ò*, *lò*, *và*, *hà*, *nò*.

5. Conclusioni

Le osservazioni sin qui raccolte permettono di cogliere le potenzialità euristiche offerte dalla realizzazione del LISSICS nel campo del trattamento automatico del linguaggio; contestualmente, fanno emergere le potenzialità diagnostiche dello strumento, in quanto anche questi soli primi dati offrono molti elementi utili alla comprensione delle reali competenze ortografiche degli studenti della scuola dell'obbligo del Canton Ticino, che spesso si discostano non poco dagli obiettivi dichiarati, forse un po' ottimisticamente, nei piani di studio ufficiali. La consapevolezza della

reale distribuzione degli errori tra i diversi livelli scolastici resa possibile da questo studio fornirà informazioni preziose per l'elaborazione di proposte didattiche coerenti con lo sviluppo delle singole competenze degli apprendenti.

BIBLIOGRAFIA

Barbagli, A., Lucisano, P., Dell'Orletta, F., Montemagni, S., Venturi, G. (2015), *CItA*: *un Corpus di Produzioni Scritte di Apprendenti l'Italiano L1. Annotato con Errori*, in Bosco, C., Tonelli, S., Zanzotto, F.S. (a c. di), *Proceedings of the Second Italian Conference on Computational Linguistics*, CLiC-it 2015. Torino, Accademia University Press, pp. 31-35.

Barbagli, A., Lucisano, P., Dell'Orletta, F., Montemagni, S., Venturi, G. (2016*), CItA: an L1 Italian Learner Corpus to Study the Development of Writing Competence*, in Calzolari, N. et al. (a c. di), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, ELRA, pp. 88-95.

Berruto, G. (2010), "Italiano standard", in *Enciclopedia dell'Italiano Treccani*, diretta da Raffaele Simone, Roma, Istituto dell'Enciclopedia, pp. 729-731.

Cignetti, L. (2016), *Tipologie e frequenza degli errori di ortografia nella scrittura degli apprendenti*, in Cignetti, L., Demartini, S. e Fornara, S. (a c. di) (2016a), pp. 19-36.

Cignetti, L., Demartini, S. (2016), *L'ortografia*, Roma, Carocci.

Cignetti, L., Demartini, S., Fornara, S. (2016a) (a c. di), *Come TIscrivo? La scrittura a scuola tra teoria e didattica*, Roma, Aracne.

Cignetti, L., Demartini S., Fornara, S. (2016b), *Il lessico di TIscrivo. Caratterizzazione del vocabolario e osservazioni in prospettiva didattica*, in *Atti del Workshop SLI-Giscel* svoltosi durante il XLVII Congresso Internazionale SLI 2013, "Sviluppo della competenza lessicale. Acquisizione, apprendimento, insegnamento", Salerno, 27 settembre 2013, pp. 43-60.

Demartini, S. (2016), *La grammatica nei testi scritti a scuola. Rilievi dall'analisi del* corpus *TIscrivo*, in Benedetti M., Bruno C., Dardano P., Tronci L., (a c. di) *Grammatiche e grammatici: teorie, testi e contesti*, Atti del XXXIX Convegno della Società Italiana di Glottologia, Roma, pp. 197-202.

De Blasi, N. (2014), *Geografia e storia dell'italiano regionale*, Bologna, Il Mulino.

De Mauro, T. (2005), *La fabbrica delle parole. Il lessico e problemi di lessicologia*, Torino, UTET.

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

Fornara, S., Cignetti, L., Demartini, S., Guaita M., Moretti A. (2015), *Costruzione del testo e punteggiatura tra norma, uso e didattica negli elaborati del corpus Tiscrivo*, in "Bulletin Suisse de Linguistique Appliquée", Actes du colloque VALS-ASLA 2014 (Lugano, 12-14 février 2014), No spécial 2015, t. 1, pp. 71-94.

Lancia, F. (2004), *Strumenti per l'analisi dei testi. Introduzione all'uso di T-LAB*, Milano, FrancoAngeli.

Naldi, M. (2014), *Traduzione automatica e traduzione assistita*, Bologna, Esculapio.

Pandolfi, E.M. (2009), *LIPSI. Lessico di frequenza dell'italiano parlato nella Svizzera italiana*, Bellinzona, Osservatorio Linguistico della Svizzera Italiana.

Marconi, L., Ott M., Pesenti E. (1994), *Lessico elementare. Dati statistici sull'italiano scritto e letto dai bambini delle elementar*i, Bologna, Zanichelli.

Schmid, H. (1994), *Probabilistic Part-of-Speech Tagging Using Decision Trees*, in *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf).

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 c. di), *Proceedings of the First Italian Conference on Computational Linguistics, CLiC-it 2014*, Pisa, Pisa University Press, pp. 354-359.

LUCA CIGNETTI • Teacher-researcher in Didactics of Italian language at the University of Applied Sciences and Arts of Southern Switzerland (SUPSI). His main fields of research are Didactics of Italian language, Didactics of writing and Textual linguistics. His recent publications include *L'ortografia* (Roma, Carocci, 2016; with S. Demartini), *Il piacere di scrivere. Guida all'italiano del terzo millennio* (Roma, Carocci, 2014; with S. Fornara) and *L'Inciso. Natura linguistica e funzioni testuali* (Alessandria, Edizioni dell'Orso, 2011).

E-MAIL • luca.cignetti@supsi.ch

SILVIA DEMARTINI • Researcher in Didactics of Italian language at the University of Applied Sciences and Arts of Southern Switzerland (SUPSI). Her main fields of research are Didactics of Italian language, History of Italian grammar and Children's writing competences. Her recent publications include *L'ortografia* (Roma, Carocci, 2016; with L. Cignetti), *Grammatica e grammatiche in Italia nella prima metà del Novecento. Il dibattito linguistico e la produzione testuale* (Firenze, Cesati, 2014) and *La punteggiatura dei bambini. Uso, apprendimento e didattica* (Roma, Carocci, 2013; with S. Fornara).

E-MAIL • silvia.demartini@supsi.ch