How to reference this article
Cignetti, L., Demartini, S. & Puccinelli, D. (2018). Il progetto *Scrivere Come Risorsa Professionale nella Svizzera Italiana*: aspetti linguistici quantitativi e qualitativi delle tesi di laurea nella Scuola Universitaria Professionale della Svizzera Italiana. *Italica Wratislaviensia*, 9(1), 35–50.
DOI: http://dx.doi.org/10.15804/IW.2018.09.02

Luca Cignetti, Silvia Demartini & Daniele Puccinelli
Scuola Universitaria Professionale della Svizzera Italiana

IL PROGETTO *SCRIVERE COME RISORSA PROFESSIONALE NELLA SVIZZERA ITALIANA*: ASPETTI LINGUISTICI QUANTITATIVI E QUALITATIVI DELLE TESI DI LAUREA NELLA SCUOLA UNIVERSITARIA PROFESSIONALE DELLA SVIZZERA ITALIANA

WRITING AS A PROFESSIONAL RESOURCE IN SOUTHERN SWITZERLAND: QUANTITATIVE AND QUALITATIVE ASPECTS OF BACHELOR'S THESES AT THE UNIVERSITY OF APPLIED SCIENCES AND ARTS OF SOUTHERN SWITZERLAND

Abstract: This paper describes the highlights of Project SCRiPSIt (*Writing as a Professional Resource in Southern Switzerland*), led by the Department of Teaching and Learning of the University of Applied Sciences and Arts of Southern Switzerland (SUPSI). Located in the largest Italian-speaking population centre outside of Italy, SUPSI is a higher-learning institution with a strong emphasis on professional development. Project SCRiPSIt investigates a relatively large corpus of SUPSI bachelor's theses by bringing together a heterogeneous research team featuring a combination of qualitative and quantitative research expertise in linguistics as well as in automated text processing. After a description of the key project objectives, we present an overview of the current state of the corpus and of the text-processing pipeline, along with some preliminary results.

Keywords: academic writing, Italian learner corpus, natural language processing, written Italian language, language teaching

Received 26/10/2017; Accepted 2/01/2018; Published 20/06/2018
ISSN 2084-4514 e-ISSN 2450-5943

1. IL PROGETTO DI RICERCA

Il problema della padronanza della lingua italiana degli studenti universitari, in particolare nella sua varietà scritta, è oggi ben noto, e non può essere confinato a specifici corsi ma attraversa trasversalmente tutti i dipartimenti e facoltà. Si tratta di un tema che occupa uno spazio di rilievo in letteratura, secondo varie prospettive: le indagini sull'argomento si configurano sia come parte integrante di una descrizione completa ed esaustiva dell'italiano e delle sue varietà, sia come riflessione più ampia e trasversale su un nodo cruciale del percorso formativo degli studenti, in considerazione del ruolo attribuito alla competenza linguistica nell'elaborazione del pensiero complesso.

In merito alle competenze scritte in lingua italiana in contesto universitario, diverse indagini, già da tempo, mettono in luce da un lato l'insufficienza del livello medio degli studenti e dall'altro l'assenza di proposte didatticamente significative da parte delle istituzioni formative (a questo proposito, si vedano almeno Sobrero, 2009; Morgana & Prada, 2010). A fronte di questo, la maggior parte delle proposte relative all'insegnamento della scrittura funzionale, così come i manuali di scrittura per le tesi di laurea, seguono ancora un'impostazione prevalentemente lineare e progressiva, che spesso prescinde dall'analisi delle reali difficoltà degli studenti. Come già sostenuto in numerosi studi (tra i quali ad esempio Serianni, 2010; Cignetti & Fornara, 2014), per proporre interventi didattici adeguati sarebbe invece necessario partire da un'attenta analisi dei nodi critici della scrittura, e quindi calibrarsi su di essi.

Partendo da simili presupposti, presso la Scuola Universitaria Professionale della Svizzera italiana è stato attivato il progetto SCRiPSIt[1] (acronimo per *Scrivere come Risorsa Professionale nella Svizzera Italiana*), che si propone come obiettivo l'identificazione delle difficoltà e degli errori più ricorrenti nella scrittura delle tesi di laurea in lingua italiana degli studenti di tre suoi dipartimenti e di una scuola affiliata.

---

[1] Il progetto SCRiPSIt è stato finanziato nell'A.A. 2016–2017 dal fondo SUPSI dedicato ai progetti interdipartimentali relativi all'asse di ricerca "Sistemi educativi/formativi".

Partendo dai risultati della ricerca, è prevista l'attivazione di corsi di scrittura accademica e di scrittura funzionale-professionale.

La prima fase della ricerca corrisponde alla raccolta di una selezione delle tesi discusse negli ultimi cinque anni presso il Dipartimento formazione e apprendimento (DFA), il Dipartimento Tecnologie Innovative (DTI), il Dipartimento Economia Aziendale e Scienze Sociali (DEASS) e l'Accademia Teatro Dimitri e il loro trattamento in formato digitale[2].

La seconda fase prevede l'analisi dei dati raccolti e l'identificazione degli errori più ricorrenti attraverso l'interrogazione automatica e manuale dei principali livelli linguistici. Questa analisi prevede un esame quantitativo generale del corpus e specifici rilievi qualitativi, attraverso operazioni condotte in parte in modo automatico ricorrendo a specifici software di *mining testuale*, in parte di tipo manuale, mettendo in rilievo gli aspetti linguistici più significativi, problematici e ricorrenti. Successivamente, saranno selezionati gli errori appartenenti alle categorie più comuni e diffuse, privilegiando l'aspetto diagnostico rispetto a quello più propriamente descrittivo.

La terza fase del lavoro prevede l'elaborazione di proposte didattiche e operative finalizzate a migliorare la qualità della scrittura degli studenti. In questa fase saranno individuate le priorità su cui intervenire e definite di conseguenza le modalità didattiche più adatte ad ogni problema rilevato nel corso delle fasi precedenti e si definiranno gli aspetti pratici di realizzazione dei corsi da attivare entro l'anno successivo alla fine del progetto.

In questo contributo, dopo avere descritto gli obiettivi (par. 1) e lo stato attuale del corpus SCRiPSIt (par. 2), ci si soffermerà sui software impiegati per l'analisi automatica (par. 3), sui primi esiti dell'analisi relativa al lessico (par. 4) e sulle potenzialità dell'impiego della pipeline Tint (par. 5).

---

[2] Per l'analisi testuale sono state scelte le tesi dei dipartimenti SUPSI coinvolti nel progetto di ricerca interdipartimentale, con l'obiettivo di costruire sotto-corpora di dimensioni simili e dunque comparabili (le tesi del DFA sono mediamente più brevi e questo spiega il maggiore numero di testi raccolti presso questo dipartimento); per l'Accademia Teatro Dimitri sono state raccolte tutte le tesi scritte in lingua italiana negli anni di riferimento del progetto.

2. IL CORPUS SCRIPSIT

Il corpus SCRiPSIt include una selezione di tesi di laurea triennale (*bachelor*) di tre dipartimenti della Scuola Universitaria Professionale della Svizzera Italiana (SUPSI) e di una scuola affiliata, per un totale di 543 testi e di 5.540.000 parole grafiche (intendendo con parola grafica "ogni sequenza di caratteri separata dalle altre da uno spazio bianco o da un segno di interpunzione", De Mauro, 2005, p. 14)[3]; tutti i testi sono conservati in formato cartaceo o elettronico presso i rispettivi dipartimenti.

La distribuzione dei testi è la seguente: del Dipartimento Economia Aziendale e Scienze Sociali (DEASS) le tesi degli anni 2014–15 e 2015–16, per un totale di 133 testi; del Dipartimento Tecnologie Innovative (DTI) le tesi degli anni 2012–13, 2013–14, 2014–15 e 2015–16, per un totale di 139 testi; del Dipartimento Formazione e Apprendimento (DFA) le tesi degli anni 2010–11, 2011–12, 2012–13, 2013–14 e 2014–15, per un totale di 258 testi; dell'Accademia Teatro Dimitri, le tesi degli anni 2011–12, 2012–13, 2013–14, 2014–15 e 2015–16, per un totale di 13 testi. La distribuzione, per numero di parole complessive (*tokens*), è equiparabile tra i tre dipartimenti coinvolti, tra i quali è dunque possibile formulare delle considerazioni di carattere comparativo; il sotto-corpus relativo all'Accademia Teatro Dimitri comprende invece un numero inferiore di testi, dovuto al ridotto numero di studenti italofoni presenti in questo corso di studi (dove è anche consuetudine scrivere la tesi in lingue diverse dall'italiano).

3. SOFTWARE ATTUALMENTE IN USO

Per gestire il trattamento e l'analisi di un corpus di dimensioni vaste come quello sopra descritto, l'équipe di ricerca si sta avvalendo dell'uso di alcuni software destinati a diversi tipi di analisi linguistiche. Individuare gli strumenti più efficienti non è sempre semplice: i software proprietari offrono *toolkit* ricchi e performanti, ma poco personalizzabili,

[3] La lemmatizzazione non è ancora completa e, dunque, i dati non possono essere considerati definitivi.

mentre quelli *open source* permettono un maggiore adeguamento a esigenze e richieste specifiche, ma proprio per questo richiedono un lavoro di adattamento non sempre semplice. I programmi più moderni, poi, per lo più sono ancora poco addestrati a lavorare sulla lingua italiana, nelle sue diverse varietà.

Nell'ambito dei vari progetti dedicati alla scrittura in corso presso il Dipartimento Formazione e Apprendimento della SUPSI i software attualmente in uso sono i seguenti:

– Atlas.ti (agevola le analisi qualitative attraverso l'etichettatura dei fenomeni d'interesse).
– TalTac2 (specifico per Text Analysis e Text Mining, offre informazioni lessicali, grammaticali e tematiche).
– TreeTagger (effettua PoS tagging e lemmatizzazione in modo probabilistico, basandosi su *decision trees*).
– T-LAB (permette un'esplorazione multilivello dei testi: temi prevalenti, associazioni, lessico, concordanze, co-occorrenze, contesti elementari ecc.).

Fra questi, per il progetto SCRiPSIt si sono per ora usati in particolare T-LAB (per gli aspetti sintetizzati al par. 3.2) e altri strumenti descritti di seguito nel contributo, che permettono alcuni rilievi d'insieme (ad esempio alcune stime di leggibilità, da considerarsi con cautela) e qualche ricerca più mirata e fine, che scaturisce dalla combinazione di diversi livelli linguistici (ad esempio sintassi e punteggiatura, come si vedrà al par. 5).

3.1. Leggibilità

La leggibilità di un testo e la sua comprensibilità sono parametri estremamente complessi da valutare, come mostrano gli studi al riguardo (cfr. ad esempio la sintesi di Lavinio, 2004, pp. 123–144) e gli indici per stimarli (il più noto dei quali, per l'italiano, è *Gulpease*[4]). Questi si basano in primis su aspetti di superficie (come lunghezza, tipo e numero medio di parole per frase, e lunghezza di frase), abbinando a essi, nelle

[4] Illustrato in Lucisano e Piemontese (1988).

evoluzioni più recenti, informazioni lessicali e sintattiche[5]. È tuttavia evidente come il discorso sulla leggibilità e sulla comprensibilità di un testo specialistico come la tesi di laurea sia qualcosa di ancor più complesso: non si tratta, infatti, di inseguire la semplificazione a tutti i costi come ideale espressivo per questo genere testuale, quanto, piuttosto, di individuare quali elementi entrano legittimamente in gioco (per esempio i tecnicismi o una sintassi non elementare, che riproduce un pensiero articolato) e che cosa, invece, mina la chiarezza e la facilità di lettura, e talvolta la comprensibilità. Questa distinzione non è sempre presente nella sensibilità linguistica degli studenti, che spesso ricorrono a un italiano ingenuamente complesso e artificioso, "da tesi", senza però padroneggiarlo compiutamente.

I primi dati ricavati sulla leggibilità di un sub-corpus di testi del progetto SCRiPSIt (quelli del corso di laurea in Ingegneria Informatica del Dipartimento Tecnologie Innovative) vanno dunque letti alla luce di queste premesse e presi come puramente indicativi di una stima di massima. Lo strumento usato per ricavare il dato è Corrige!it, che effettua sia stime di leggibilità, sia rilievi sul profilo ortografico dei testi; lanciando l'analisi, scopriamo ad esempio che le tesi dell'annata 2014–15 presentano un indice di leggibilità di 47 (nella scala 0–100, un testo è tanto più leggibile quanto più l'indice si avvicina a 100). Questo indice va messo in relazione con il livello d'istruzione dei potenziali lettori: ciò significa che le tesi sarebbero (prevedibilmente) incomprensibili per lettori con la sola licenza elementare, mentre risulterebbero globalmente accessibili a quei lettori con un livello d'istruzione medio superiore. Anche Tint, pipeline di cui parleremo al par. 5, offre la possibilità di effettuare questo tipo di analisi.

Al di là della stima d'insieme, è interessante e utile osservare nel dettaglio i singoli periodi (individuando i più problematici, cosa che i software permettono) ed esplorare la caratterizzazione del lessico. Ciò permetterà di individuare la distribuzione delle parole fra Vocabolario di Base (fondamentale, di alto uso e di alta disponibilità, cfr. De Mauro,

[5] Ad esempio READ-IT, il toolkit sviluppato dall'Istituto di Linguistica Computazionale del CNR di Pisa.

1980 e 2000) e parole a esso estranee, come ad esempio i termini tecnici, ma anche il lessico comune.

3.2. Associazioni lessicali: di che cosa si parla nei testi

Uno strumento come T-LAB permette, fra le varie opzioni d'indagine, di approfondire in un altro senso gli aspetti lessicali: esplorare i testi a livello tematico. Ad esempio, attraverso le associazioni di parole si può individuare l'universo semantico che caratterizza temi salienti, come, per citare un caso, la "ricerca" nelle tesi di un determinato dipartimento. Intorno a questa parola orbita infatti un universo di parole più o meno fortemente associate a essa, che la definiscono: nel caso del Dipartimento Formazione e Apprendimento, ad esempio, emerge come la ricerca sia strettamente legata all'educazione (lo mostra la stretta associazione con parole come *scuola*, *classe*, *allievo*).

Oppure ancora, tornando ad analisi più strettamente linguistiche, T-LAB permette l'esplorazione di tutti i contesti elementari del corpus in cui occorrono specifiche parole. È ad esempio possibile scorrere tutti insieme i frammenti di testo che contengono parole o locuzioni d'interesse perché particolarmente esposte a errore, come, ad esempio, *riguardo a* (insidioso a livello di reggenza).

4. PROBLEMI LESSICALI: ESEMPI DALL'ANALISI MANUALE

Solo un'attenta lettura da parte del ricercatore permette, però, di trovare e analizzare in profondità le peculiarità e le devianze lessicali. Non c'è, infatti, ancora uno strumento che riesca a esaminare le scelte lessicali in contesto (e nel co-testo) e a valutarne la pertinenza; inoltre, quelli che sommariamente possono essere etichettati come "errori lessicali" sono, in realtà, fenomeni di natura diversa e complessa[6].

[6] Un tentativo di tipologizzare i più ricorrenti problemi lessicali riscontrati in un vasto corpus di testi scritti da studenti di scuola dell'obbligo si può trovare in Demartini (2016).

4.1. Imprecisione nella selezione lessicale

Scorriamo qualche esempio eterogeneo, che dà una prima idea delle criticità più ricorrenti e resistenti anche in testi revisionati come le tesi di laurea (fra quadre si trova l'indicazione del dipartimento in cui è stata discussa la tesi). Cominciamo dall'esempio 1., in cui, a una prima lettura, può non essere semplice cogliere le anomalie lessicali:

> 1) Sulla stessa **linea d'onda** si situano anche le affermazioni di Coffey (1968). Secondo lui, un inizio di predilezione per i dipinti figurativi, piuttosto che astratti, si **aggirerebbe** attorno ai 4–5 anni. [DFA, Dipartimento Formazione e Apprendimento]

A parte l'uso di "lui" riferito a uno studioso citato (scelta stilisticamente non idonea), "linea d'onda" e "aggirerebbe" sono le espressioni su cui concentrare l'attenzione. La prima, di fatto, non è corretta, benché non di rado sia circolante anche in questa forma (basta controllare con una semplice ricerca sul web): dovrebbe essere "lunghezza d'onda"; la seconda, invece, viene riferita in modo improprio al sostantivo "inizio": il verbo "aggirarsi" nel senso di "approssimarsi", infatti, può essere legittimamente abbinato solo a un sostantivo di quantità (qui *età*), cosa che qui non accade, generando – come spesso capita − un problema a cavallo fra lessico e sintassi. Sebbene il senso non sia compromesso, una tesi di laurea sarebbe auspicabile non presentasse imperfezioni simili. Qualcosa di simile accade nell'esempio 2, ma con esito ancor più infelice:

> 2) Ancora oggi il tema dell'insuccesso scolastico è un argomento **caustico** […]. [DFA]

Qui la parola target che lo scrivente vorrebbe utilizzare è del tutto mancata, ed è persino difficile ipotizzare quale avrebbe dovuto essere la sua scelta (la somiglianza fonetica farebbe pensare a *ostica*, ma il contesto indurrebbe a inferire piuttosto qualcosa come *centrale* o *urgente*): a essere compromessa è, insomma, per intero, l'efficacia comunicativa. Una porzione più estesa di un altro testo permette di percepire ancor

meglio l'effetto sul lettore di questo tipo di errore di selezione lessicale imprecisa:

> 3) Le molle presenti nell'assieme di costruzione servono a generare una forza durante la fase di azzeramento della macchina in modo tale da garantire un miglior contatto tra i piani di costruzione e proiezione, favorendo il corretto azzeramento. Senza le molle la forza di contatto tra i due elementi corrisponde alla forza peso della parte libera di muoversi dell'assieme di costruzione, lasciando **discutibile** il corretto azzeramento. La forza generata dalla molla deve essere sufficiente da garantire il contatto tra il piano di costruzione e di proiezione […]. [DTI, Ing. Meccanica]

L'aggettivo "discutibile", a proposito del processo descritto, è un errore di adeguatezza: nella spiegazione di questo fenomeno fisico, l'aggettivo "discutibile" (vago, ambiguo) risulta infatti improprio, se si considera che la formulazione corretta (chiara, non passibile di interpretazioni diverse) avrebbe dovuto essere, ad esempio, "impedendo il corretto azzeramento".

Questa debolezza nella gestione del lessico non specialistico adeguato e una certa approssimazione nelle scelte (indice di scarsa consapevolezza) si ritrovano talvolta, in misura minore, anche nell'uso dei termini tecnici (nel caso 3., termini del linguaggio settoriale della linguistica), per i quali si riscontrano imprecisioni o lapsus, cui spesso si potrebbe ovviare con un'attenta rilettura (qui "sorde" e "sonore" andrebbero invertite):

> 4) È risultato che gli errori maggiori sorgono quando devono produrre consonanti **sorde**, poiché le sostituiscono con quelle **sonore** (esempio: "mago", che diventa "mako"). [DFA]

4.2. Regionalismi e forestierismi: un'occasione per riflettere

Nell'esaminare le tesi del corpus, non va dimenticato il fatto che l'italiano parlato e scritto in Canton Ticino presenta alcune peculiarità tipiche di una specifica varietà regionale, che includono anche usi lessicali particolari o non-standard. Vediamone un paio (la locuzione "a dipendenza" e l'uso di "rispettivamente" con valore disgiuntivo):

5) Oltretutto si ha la possibilità di programmare il sistema su 3 assi o più **a dipendenza** delle caratteristiche del robot stesso. [Dipartimento Tecnologie Innovative, Ing. Meccanica]
6) Le giunzioni sono rappresentate da cerchi in cui possono giungere più transizioni, **rispettivamente** partirne di nuove. [DTI, Ing. Elettronica]

Ovviamente, non si tratta qui di errori, quanto, piuttosto, di peculiarità alle quali lo studente (qualsiasi studente, a seconda della varietà regionale che è sua) dovrebbe essere sensibile. Dovrebbe infatti avere gli strumenti per effettuare un'auto-valutazione spontanea di quanto un regionalismo possa essere accettabile a seconda del contesto e dei destinatari, che potrebbero capire o meno un certo uso geo-linguisticamente connotato. E, nel caso si rischino l'oscurità o l'ambiguità, lo scrivente dovrebbe saper cambiare formulazione.

In un certo senso, un'analoga riflessione potrebbe svilupparsi a proposito dell'opportunità dell'inserimento di forestierismi, in particolare, come nell'esempio 7, in lingua inglese:

7) Una volta scelta la compagnia navale e dopo aver ricevuto la conferma del **booking**, viene effettuata la prenotazione, detta **pick up order**, del ritiro presso il **carrier** […]. Nel caso in cui si tratti una spedizione **full container** […] un trasportatore italiano ha l'incarico di portare il vuoto presso l'azienda […] ; quando si tratta di spedizioni **in collect**, Expeditors contatta direttamente un **carrier** sul territorio che avrà il compito di prelevare la merce […] e successivamente portarla […] presso un **co-loader** che avrà invece il compito di effettuare **groupage** […]. [DTI, Ing. Informatica]

Spesso si tratta di termini tecnici settoriali, che dunque sono giustamente insostituibili; tuttavia, una simile pioggia di parole inglesi induce comunque a riconsiderare l'opportunità di qualche caso (perché "booking", ad esempio, e non "prenotazione"?).

4.3. Il lessico e il testo

Da ultimo, esaminiamo ora una porzione testuale più estesa, che compendia in poco spazio una serie di errori molto comuni e ne mostra l'effetto sull'impatto del testo nell'insieme:

> 8) Questo progetto ha **concesso** di affrontare delle **problematiche** relative al mondo del lavoro, confrontarsi con i clienti e riuscire a capire le loro esigenze per poter realizzare il prodotto che desiderano. In questi momenti si comprende **affondo** l'importanza della progettazione, della raccolta e dell'analisi dei requisiti. Inoltre è stato possibile partecipare alla creazione del prodotto non solo a livello di implementazione, ma suggerendo le proprie idee e consigli riguardante alcuni aspetti dello stesso. Un fattore chiave di questo progetto è di aver potuto approfondire diverse **tematiche** come la sicurezza, la localizzazione, la creazione di query complesse e l'analisi dei dati. **Riguardo la** sicurezza è stato necessario implementare un sistema per evitare i robot, inoltre è stato effettuato hashing con salt per evitare attacchi di tipo Rainbow Table, più altre funzionalità descritte nel capitolo "Implementazione". [DTI, Ing. Informatica]

Se, come spiega Silvana Ferreri (2005, pp. 68, 98–94), conoscere una parola significa saperle associare diverse informazioni (dalla forma fonica e grafica, alla corretta morfologia, alla semantica, alle relazioni con altre parole nel testo), allora l'esempio 8 mostra diversi livelli di criticità: si inizia con la scelta di "concesso" (concedere, 'accordare con favore', stride un po' nel contesto e ben esemplifica la tendenza alla ricerca, non sempre consapevole, della variante "alta": "permesso" sarebbe stata l'alternativa migliore); si notano, poi, specifiche e insistenti preferenze morfologiche (l'uso, non raro nell'italiano contemporaneo, di sostantivi come "problematica" e "tematica" usati come sinonimi per "problema" e per "tema"); e poi ancora, l'univerbazione con raddoppiamento fonosintattico di "a fondo" in "affondo" (che, benché accolta come variante lecita nei maggiori dizionari dell'uso, è voce che con buona probabilità riflette la pronuncia dell'italiano centro-meridionale), e, infine, un errore di reggenza ("riguardo la" per "riguardo alla"). Il testo risulta almeno in parte disturbato da una trama lessicale che presenta imperfezioni e anomalie di diverso tipo, forse non eclatanti, ma tanto frequenti da disturbare la lettura e da compromettere la qualità della scrittura.

5. TINT

Sebbene nel nostro corpus non si possa prescindere dall'analisi manuale degli errori, l'utilizzo di strumenti di analisi automatica del testo può rivelarsi molto utile a complemento dell'analisi manuale. Lo strumento scelto dall'équipe di ricerca per l'analisi automatica è Tint, una pipeline di elaborazione del linguaggio naturale (*natural language processing*) applicato a operazioni di *pattern matching*.

Tint è l'acronimo di *The Italian NLP Tool* (Palmero Aprosio & Moretti, 2016). Si tratta di una pipeline per l'elaborazione automatica di testi in lingua italiana basata su Stanford CoreNLP (Manning *et al*., 2014). Il PoS tagger utilizzato da Tint, in particolare, si basa sullo Stanford Log-linear Tagger, che sfrutta il principio di massima entropia. Data una parola e il suo contesto (altre parole nella stessa frase e i loro tag), l'utilizzo del principio di massima entropia assegna una probabilità a ogni tag in un tagset predefinito, rendendo possibile la stima della probabilità di una sequenza di tag data una sequenza di parole (una frase). Tra tutte le possibili distribuzioni che soddisfano un dato insieme di vincoli, viene scelta la distribuzione a massima entropia in quanto essa rappresenta l'assegnazione di probabilità che soddisfa i vincoli e richiede il minor numero di ipotesi aggiuntive.

L'obiettivo della nostra analisi automatica è quello di coadiuvare l'analisi manuale, e quindi di permettere a chi esegue quest'ultima di focalizzare la propria attenzione solo sui passaggi potenzialmente errati. A tale scopo, è importante massimizzare il recupero (*recall*), cioè far tendere all'unità la frazione di passaggi contenenti errori che vengono individuati automaticamente. Per i nostri scopi, la massimizzazione del recupero può avvenire a scapito della precisione, definita come la frazione di passaggi recuperati che effettivamente contengono errori. La precisione fornisce quindi un'indicazione di quanto lavoro di analisi manuale è stato risparmiato dall'analisi automatica.

L'idea fondamentale è quella di utilizzare l'analisi delle dipendenze di Tint per riconoscere il ruolo sintattico delle parti del discorso e di conseguenza isolare i casi di interesse. Come *case study*, l'équipe ha scelto un fenomeno interpuntivo, cioè l'individuazione automatica di casi in

cui il gruppo del soggetto e il gruppo del verbo sono separati da una virgola: casi potenzialmente problematici, in cui la virgola va a separare impropriamente unità semanticamente e sintatticamente coese[7].

Partendo dai testi delle tesi di laurea in formato PDF, prima le si converte in formato testuale tramite un apposito software standard, e poi si effettua una seconda conversione per eliminare le informazioni ridondanti come le intestazioni, le parti in altre lingue, e i contenuti di natura non propriamente testuale, come ad esempio i contenuti delle tabelle o gli algoritmi. Per questa seconda conversione, si utilizza già Tint: contenuti di questo tipo vengono scartati meramente sulla base dell'analisi delle dipendenze. Ad esempio, i passaggi di natura non testuale presentano evidenti anomalie a livello di analisi delle dipendenze, e possono quindi essere filtrati grazie a essa. È a questo punto che si procede alla fase di *pattern matching* ancora con l'analisi delle dipendenze di Tint, al fine di individuare tutti i passaggi sospetti che presentano il pattern soggetto-virgola-verbo.

Seguono un paio di esempi, entrambi tratti da tesi del corso di laurea in Ingegneria Gestionale, estratti automaticamente con il nostro metodo, che sul campione di interesse ha raggiunto un richiamo unitario.

> 9) La quale assieme al impianto [*sic*] di verniciatura, è stata messa a disposizione delle industrie ticinesi, per svolgere lavorazione della lamiera conto terzi. [DTI, Ing. Gestionale]

In questo primo esempio, l'analisi delle dipendenze permette il riconoscimento del soggetto ("*la quale"*) e della relativa radice, vale a dire l'elemento da cui dipende il soggetto nell'economia delle dipendenze, che in questo caso è il predicato verbale ("*è stata messa a disposizione"*). Una volta riconosciuti questi due elementi, non rimane che verificare se sono separati da una virgola o meno. In questo caso lo sono, e quindi la frase viene correttamente segnalata a chi deve effettuare l'analisi manuale.

---

[7] Inevitabilmente, ulteriori aspetti interpuntivi critici sono affrontati dai ricercatori attraverso l'esame manuale del corpus (cfr. per esempio Demartini & Ferrari, in press).

10) Un altro grosso cambiamento rispetto al passato, è l'imprevedibilità del quantitativo di ordini in entrata. [DTI, Ing. Gestionale]

Anche in questo secondo esempio, l'analisi delle dipendenze permette il riconoscimento del soggetto ("cambiamento") e della relativa radice, che in questo caso è il predicato nominale ("è l'imprevedibilità"). Anche in questo caso troviamo una virgola interposta tra gruppo del soggetto e gruppo del verbo, e quindi anche questa frase viene segnalata.

L'analisi è stata estesa ai casi con un numero di virgole diverso da zero e diverso da due, anche se quest'ultima condizione può portare a un richiamo inferiore all'unità in presenza di casi in cui una o entrambe le virgole siano state usate impropriamente dal tesista.

La nostra metodologia ha portato a un richiamo unitario a scapito di una precisione ridotta, principalmente dovuta a casi di errori nell'analisi delle dipendenze dovuti ad errori di sintassi presenti nelle tesi. La massimizzazione della precisione condizionato al raggiungimento unitario è tuttora oggetto di studio per l'équipe di ricerca. Un approccio promettente è quello, già parzialmente intrapreso, di utilizzare molteplici strumenti per l'analisi delle dipendenze.

BIBLIOGRAFIA

Berruto, G. (1993). *Le varietà del repertorio*. In Sobrero (1993), 3–36.

Berruto, G. (1995). *Fondamenti di sociolinguistica*. Roma–Bari: Laterza.

Berruto, G. (2006). *Sociolinguistica dell'italiano contemporaneo*. Roma: Carocci.

Brusco, S., Lucisano, P. & Sposetti, P. (2014), *Le scritture degli studenti laureati: una analisi delle prove di accesso alla Laurea Magistrale in Pedagogia e Scienze dell'Educazione e della Formazione della "Sapienza"*. Roma: Aracne.

Cignetti, L. & Fornara, S. (2014). *Il piacere di scrivere. Guida all'italiano del terzo millennio*. Roma: Carocci.

Demartini, S. (2016). Un repertorio delle difficoltà lessicali ricorrenti. In L. Cignetti, S. Demartini & S. Fornara (ed.), *Come TIscrivo? La scrittura a scuola tra teoria e didattica* (pp. 161–201). Roma: Aracne.

Demartini, S. & Ferrari P.L. (in press). La *virgola splice* nei testi di studenti universitari: un problema solo in apparenza superficiale. In A. Ferrari, L. Letizia, F. Pecorari & R. Stojmenova Weber (ed.), *Punteggiatura, sintassi, testualità nella varietà dei testi contemporanei*. Atti del Convegno internazionale (Basilea, 17–19 gennaio 2018), Firenze: Cesati.

De Mauro, T. (1980). *Guida all'uso delle parole*. Roma: Editori Riuniti.

De Mauro, T. (2000). *Grande Dizionario Italiano dell'Uso* (ed. in CD-ROM). Torino: UTET.

De Mauro, T. (2005) *La fabbrica delle parole. Il lessico e problemi di lessicologia*. Bologna: il Mulino.

Ferreri, S. (2005). *L'alfabetizzazione lessicale. Studi di linguistica educativa*. Roma: Aracne.

Lavinio, C. (2004). *Comunicazione e linguaggi disciplinari. Per un'educazione linguistica trasversale*. Roma: Carocci.

Lucisano, P. & Piemontese, M.E. (1988). GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e città*, 3(31), 110–124.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J. Berthard, S. & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In K. Bontcheva & Z. Jingbo (eds.), *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). Baltimore, Maryland: Association for Computational Linguistics.

Palmero Aprosio, A. & Moretti, G. (2016). Italy goes to Stanford: a collection of CoreNLP modules for Italian. *ArXiv e-prints* [arXiv:1609.06204v2].

Serianni, L. (2010). *L'ora di italiano. Scuola e materie umanistiche*. Roma–Bari: Laterza.

Sobrero, A.A. (A cura di) (1993). *Introduzione all'italiano contemporaneo*, vol. 2. Roma–Bari: Laterza.

Sobrero A.A. (2009). L'incremento della competenza lessicale, con particolare riferimento ai linguaggi scientifici, *Italiano LinguaDue*, vol. 1, 211–225.

Sposetti, P. (2017a). *Le scritture professionali in educazione. Teorie, modelli, pratiche*. Roma: Edizioni Nuova Cultura.

Sposetti, P. (2017b). Quante e quali scritture professionali in educazione. *Italiano LinguaDue* 1/2, 261–271.

Telmon, T. (1993). *Varietà regionali*. In Sobrero (1993), 93–149.

Riassunto: In questo contributo viene introdotto il progetto *Scrivere Come Risorsa Professionale nella Svizzera italiana* (SCRiPSIt), promosso dal Dipartimento formazione e apprendimento (DFA) della Scuola Universitaria Professionale della Svizzera Italiana (SUPSI). Il progetto si propone di identificare le difficoltà e gli errori più ricorrenti nella scrittura delle tesi di laurea in lingua italiana degli studenti di tre dipartimenti della SUPSI (Dipartimento Formazione e Apprendimento, Dipartimento Tecnologie Innovative, Dipartimento Economia Aziendale, Sanità e Sociale) e di una scuola affiliata (Accademia Teatro Dimitri), con l'obiettivo di migliorarne la qualità linguistica e formale. Dopo una descrizione degli obiettivi e dello stato attuale di realizzazione del corpus, ci si sofferma sull'illustrazione dei software impiegati per l'analisi automatica, sui primi esiti dell'analisi relativa al lessico e sulle potenzialità dell'impiego della pipeline Tint. I risultati del progetto consentiranno di attivare corsi specifici dedicati alla scrittura delle tesi di laurea e più in generale alla scrittura di tipo funzionale e professionale, con applicazione potenzialmente estesa a tutti i dipartimenti della SUPSI e alle scuole affiliate.

Parole chiave: scrittura accademica, corpus di apprendenti, analisi automatica del linguaggio, italiano scritto, insegnamento della lingua