A cura di

Nicola Grandi

L'ITALIANO SCRITTO DEGLI STUDENTI UNIVERSITARI

Quadro sociolinguistico, tendenze tipologiche, implicazioni didattiche

Materiali Linguistici
Università di Pavia

FrancoAngeli

3. La raccolta dei dati e la costruzione dei corpora

di *Silvia Ballarè, Nicola Grandi* e *Matteo Pascoli*

1. Introduzione

Uno degli obiettivi del progetto Univers-ITA è stato la creazione di strumenti che possano restituire una fotografia attendibile ed empiricamente solida della lingua effettivamente usata da studenti e studentesse nella redazione di testi formali all'università. Nello specifico, nell'ambito del progetto sono stati realizzati tre corpora, Univers-ITA, Univers-ITA-ProGior e Univers-ITA-ProUniv, che riproducono tre tipologie di testi redatti da studenti e studentesse universitari con diversi corredi di metadati[1]. I tre corpora, che verranno descritti nel dettaglio nelle prossime sezioni di questo contributo, sono consultabili in accesso aperto sulla piattaforma NoSketch Engine a questo indirizzo: https://corpora.ficlit.unibo.it/CUSP.

Per le modalità di consultazione dei corpora si rimanda al vademecum riportato sul sito del progetto: https://site.unibo.it/univers-ita/it/corpora.

2. Il corpus Univers-ITA[2]

Il corpus Univers-ITA raccoglie i testi redatti *ad hoc* per il progetto. L'obiettivo, come già detto nei capitoli precedenti, era selezionare un campione rappresentativo di studenti e studentesse della coorte 2019/20 a cui far redigere un breve testo formale (tra le 250 e le 500 parole) su una traccia

1. Come si è detto nel cap. 1, i corpora Univers-ITAProGior e Univers-ITA-ProUniv raccolgono testi prodotti in una fase generalmente avanzata del percorso universitario (contributi a giornali universitari, relazioni di stage, tesi); il corpus Univers-ITA comprende testi nel segmento iniziale della formazione universitaria.

2. Il corpus Univers-ITA deve essere citato come segue: Grandi, Nicola, Ballarè, Silvia, Chiusaroli, Francesca, Gallina, Francesca, Pascoli, Matteo, Pistolesi, Elena; *Corpus Univers-ITA*. 2023, https://doi.org/10.60760/unibo/univers-ita

comune, accompagnato da un ricco questionario sociobiografico, necessario per tracciare il retroterra di ogni scrivente. La raccolta dati ha avuto luogo nell'a.a. 2020/2021, coinvolgendo, dunque, studenti e studentesse iscritti al secondo anno di corsi triennali o magistrali a ciclo unico. La scelta del secondo anno è stata motivata con la necessità di distanziare temporalmente la rilevazione dall'inizio dell'esperienza universitaria, per attenuare l'effetto della formazione scolastica secondaria di secondo grado sui testi prodotti.

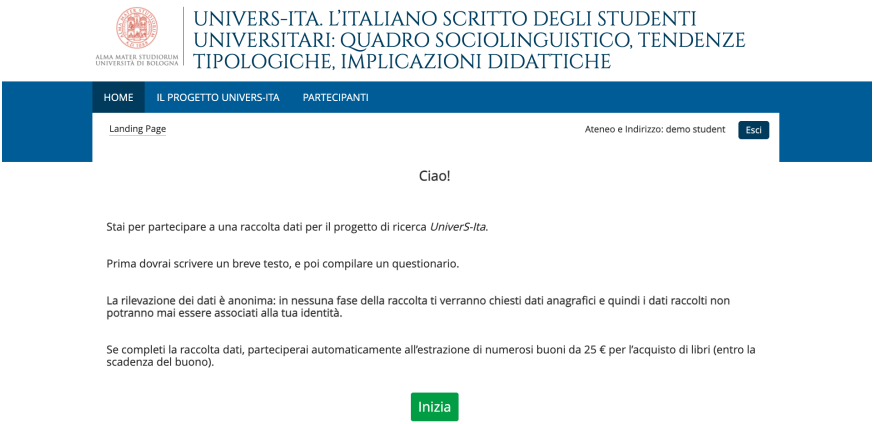
2.1. *La raccolta dei dati*

La rilevazione era originariamente programmata in presenza in atenei rappresentativi delle tre aree geografiche: Milano e Bologna per il Nord, Macerata, Perugia e Pisa per il Centro, e Napoli per il Sud. Come si è visto nel cap. 1, il progetto originario prevedeva di coinvolgere classi dei corsi di studio individuati (L-SNT/1 e LM-13 per l'area sanitaria; L-9, nello specifico ingegneria gestionale, per l'area scientifica; L-11 e LM-85bis per l'area umanistica e L-18, nello specifico economia aziendale, e LMG/01 per l'area sociale) durante una normale lezione curricolare. Tuttavia, a causa delle restrizioni legate alla situazione pandemica esplosa proprio in coincidenza con la campagna di raccolta dati, si è resa necessaria una revisione radicale del piano di lavoro per riprogettare la raccolta dei dati a distanza. Grazie alla collaborazione e alla disponibilità di colleghe e colleghi, il progetto è stato presentato in numerose aule virtuali all'inizio di lezioni curricolari. Nella maggior parte dei casi, studenti e studentesse hanno poi partecipato alla raccolta dati in un secondo momento; solo occasionalmente i dati sono stati raccolti durante la lezione stessa. Questo riadattamento metodologico forzato da contingenze esterne ha avuto conseguenze sostanziali rispetto alla composizione del campione, come già ricordato nel cap. 1. Innanzitutto, si è verificata una certa difficoltà a controllare l'effettiva attività degli studenti e delle studentesse durante la rilevazione: la partecipazione, inizialmente pensata come monitorabile data la presenza in aula dei e delle partecipanti e del ricercatore o della ricercatrice, è divenuta di fatto volontaria e non sorvegliata (e questo può certamente determinare un effetto leggermente distorsivo sui risultati dell'indagine, dal momento che si può supporre che abbiano partecipato, in media, studenti e studentesse più motivati). Inoltre, la modalità a distanza ha ovviamente comportato un decremento del tasso di partecipazione media prevista e ci ha indotto ad ampliare il numero degli atenei coinvolti (oltre 40 alla fine del progetto). Inoltre, per le aree scientifica e sociale si è resa necessaria una revisione delle lauree da campionare. Per l'area scientifica, abbiamo quindi rilevato anche i corsi della classe L-9 con denominazione diversa da ingegneria gestionale e i corsi delle classi L-7 e L-8; per l'area sociale, abbiamo rilevato tutti corsi della classe L-18.

Come si è accennato in precedenza, a studenti e studentesse coinvolti nel progetto è stata sottoposta una traccia comune, che conteneva indicazioni esplicite relative al registro (sorvegliato) e alla tipologia (argomentativa) del testo: è stato infatti chiesto di produrre un testo con uno stile formale ("come se scrivessi per un tuo professore"), immaginando di rispondere a un sondaggio "rivolto a tutti gli studenti sulla didattica a distanza nei mesi di emergenza sanitaria"; con la precisazione di scrivere "in modo non schematico, i vantaggi e gli svantaggi della didattica a distanza" secondo il proprio punto di vista (Grandi *et al.*, 2024)[3]. Per evitare di influenzare il campione, l'accesso al sito web per la raccolta dei dati è stato protetto da username e password forniti di volta in volta agli studenti e alle studentesse.

L'interfaccia era presentata come segue:

Fig. 1 - Prima schermata dell'interfaccia per la raccolta dei dati



Come già accennato, dopo la stesura del testo, i/le partecipanti hanno avuto accesso alla sezione dedicata al questionario sociobiografico (cfr. cap. 2). L'impegno complessivo previsto era di circa 90-120 minuti. Per incentivare la partecipazione è stata creata una sorta di lotteria che metteva in palio buoni di 25 euro per l'acquisto di libri.

3. Nel dettaglio, agli e alle partecipanti è stato richiesto quanto segue:

Devi scrivere un testo di media lunghezza: tra le 250 e le 500 parole. Dovrai cercare di usare uno ***stile formale****: quindi, scrivi nel modo più corretto possibile, come se scrivessi per un tuo professore. Proprio perché la rilevazione è totalmente anonima, sarà impossibile associare il testo alla tua persona e quindi, partecipando alla rilevazione, rinunci alla proprietà intellettuale su di esso. Il testo non sarà mai pubblicato integralmente e sarà utilizzato solo per scopi di ricerca.* (Il grassetto è nell'originale)

Fig. 2 - Seconda schermata dell'interfaccia per la raccolta dei dati

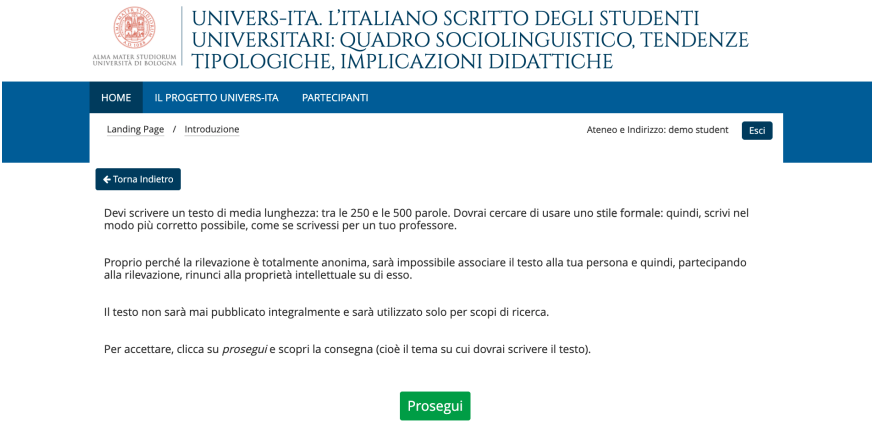


Fig. 3 - L'interfaccia e la traccia per la stesura del testo[4]

4. Traccia: *Immagina che il tuo corso di laurea abbia aperto un sondaggio rivolto a tutti gli studenti, con l'obiettivo di raccogliere opinioni sul funzionamento della didattica a distanza nei mesi di emergenza sanitaria. Scrivi un testo in cui esponi, in modo non schematico, i vantaggi e gli svantaggi della didattica a distanza, secondo il tuo punto di vista*. La definizione di questa traccia e della procedura di raccolta dei dati è stata preceduta da una lunga fase di *testing* che si è svolta nelle sedi del progetto. In questa fase, soprattutto, sono state sperimentate tracce diverse. L'obiettivo era l'individuazione di una traccia che fosse del tutto indipendente dalle aree disciplinari dei corsi di studio da campionare e accessibile a ogni partecipante. Anche l'intervallo di parole (250-500) è stato fissato dopo questa fase di *training*: a gruppi diversi di studenti e studentesse è stato chiesto di redigere, in un'ora, testi di lunghezza diversa. L'obiettivo, in questo caso, era quello di individuare un *range* di parole che potesse essere gestito in un'ora, prevedendo però anche il tempo necessario per una preliminare pianificazione del testo, indispensabile dato il carattere formale del medesimo. In altri termini, tempo e lunghezza dovevano essere tali da evitare che i testi venissero scritti "di getto".

Alla fine della campagna di raccolta dei testi e del necessario processo di "pulitura" dei medesimi[5], sono stati selezionati 2.137 testi validi, per un totale di 810.715 parole. Il corpus Univers-ITA è, dunque, di dimensioni ridotte, ma ha il vantaggio di essere piuttosto bilanciato e di essere corredato da una serie ricchissima di metadati, che verranno presentati nei prossimi paragrafi.

2.2. *L'annotazione automatica e manuale dei testi*

I 2.137 testi raccolti sono stati poi trattati in maniera automatica e manuale. In prima istanza, i testi sono stati lemmatizzati e etichettati per parte del discorso tramite gli strumenti software di Dylan Lab (CNR, Pisa); inoltre, sono stati esaminati automaticamente utilizzando le funzionalità del software READ-IT[6], che valuta la leggibilità di un testo secondo un set di parametri finalizzati a misurare il suo livello di complessità: la suddivisione in frasi, il numero di parole, la lunghezza delle frasi, la lunghezza delle parole, le parti del discorso, alcuni tratti morfo-sintattici, ecc. Di questi parametri, ne sono stati selezionati alcuni, giudicati più affidabili per l'analisi condotta nel progetto e per la natura dei testi: il numero di frasi del testo, il numero di parole diverse, il numero di token verbali, ecc. Lo scopo di questa prima indagine era valutare, ancorché con grana piuttosto grossolana, la complessità dei testi e la loro omogeneità. Il valore che il software restituisce è relativo alla probabilità percentuale che il testo analizzato venga classificato come difficile. I termini di confronto sono identificati con un articolo del quotidiano *La Repubblica* come prototipo di un testo "complesso" e con un testo del periodico *Due Parole*, destinato a lettori con un livello di scolarizzazione basso, come esempio ideale di un testo "semplice". I testi raccolti restituiscono il quadro che segue:

Tab. 1 - Media e mediana indice di leggibilità dei testi

	Mediana indice di leggibilità	*Media indice di leggibilità*
Totale	72,06	72,17
Nord	70,87	71,20
Centro	73,63	73,13
Sud	72,44	72,63

5. Per esempio, alcuni testi sono stati scritti in inglese; altri testi comprendevano una "coda" di caratteri senza senso digitati solo per arrivare alla lunghezza minima prevista, ecc. I testi eliminati sono stati 23.

6. Cfr. Dell'Orletta *et al.* (2011), www.italianlp.it/demo/read-it/

I valori nella tabella indicano, come si è appena detto, la probabilità che un testo appartenga alla classe dei testi di difficile leggibilità. Questo significa, in sintesi, che i testi redatti dai e dalle partecipanti al progetto hanno poco più del 70% di probabilità di essere scarsamente accessibili a un lettore poco colto e istruito.

Pur con i limiti connessi ad una ricognizione puramente quantitativa ed automatica, il quadro che emerge è quello di una sostanziale omogeneità dei testi prodotti, rispetto all'indice di leggibilità: i dati rivelano, cioè, una complessità medio-alta. È sostanzialmente quello che ci si aspetterebbe da una rilevazione di questo tipo, data la natura dei testi oggetto dell'indagine. Questo consente, a prima vista, di identificare la lingua impiegata da studenti e studentesse universitari per redigere un testo formale come un sottoinsieme abbastanza omogeneo del diasistema italiano.

Per avere però un quadro più attendibile, constatati i limiti di una ricognizione puramente quantitativa e automatica, tutti i testi sono stati successivamente riletti e annotati manualmente e qualitativamente da due lettrici con una solida competenza linguistica e metalinguistica acquisita durante un percorso di studi magistrale incentrato sulle scienze linguistiche e con una specifica formazione sulla lingua italiana. L'annotazione manuale è stata preceduta da una lunga fase di *training* e di allineamento delle due annotatrici ai criteri definiti in precedenza dal gruppo di ricerca[7].

Si è scelto programmaticamente di annotare tutti i tratti che configurassero una qualche forma di "devianza" o allontanamento rispetto al risultato atteso, cioè un testo formale, sorvegliato, pienamente conforme a quanto prescritto dalla grammatica normativa o scolastica. Questi tratti si caratterizzano per gradi diversi di "devianza" o allontanamento dal risultato atteso. Per questa ragione, i fenomeni annotati risultano essere eterogenei dal punto di vista della loro caratterizzazione sociolinguistica[8]. Si hanno, cioè, tratti fortemente stigmatizzati che, tipicamente, occorrono in testi decisamente caratterizzati verso il basso in diafasia e diastratia (ad esempio realizzazioni ortografiche

7. Le due lettrici sono state reclutate appositamente per il progetto con contratti di collaborazione e con un livello medio-alto di retribuzione. Per il lavoro di lettura e annotazione dei testi è stato concesso loro un tempo piuttosto ampio. Le due lettrici, periodicamente, verificavano l'una il lavoro dell'altra, a campione, anche allo scopo di mantenere un costante allineamento sulla selezione dei tratti da annotare. A questa fase di verifica, seguiva poi un'ulteriore fase di controllo, sempre a campione, da parte dei membri del progetto. Si è ritenuto che una retribuzione relativamente più alta della media, la possibilità di poter organizzare il lavoro con ritmi non contingentati e un sistema di controlli incrociati ricorrenti e costanti potessero ridurre il rischio che si innescasse una sorta di "automatismo" nell'annotazione dei testi.

8. E per questa ragione, come anticipato nell'introduzione, si è scelto di indicarli con l'etichetta "neutra" di annotazioni e di evitare l'uso dell'etichetta "errore", che avrebbe potuto innescare equivoci nell'interpretazione dei dati.

substandard come *o comprato* senza *h*[9]), ma anche altri che si trovano con una certa frequenza in produzioni mediamente controllate di colti (come, ad esempio, le costruzioni marcate come le dislocazioni o il pronome *gli* generalizzato), benché non di rado condannati dalle grammatiche normative poiché estranei al vecchio standard più rigidamente codificato[10]. Questa ultima categoria di tratti dal punto di vista sociolinguistico è riconducibile all'italiano neo-standard (Berruto 2012; Ballarè 2020; Grandi 2024) e dunque si colloca nella porzione centrale (e non periferica) dell'italiano contemporaneo. In altre parole, questi tratti rivelano che in alcuni ambiti, anche molto controllati, la norma scritta, come avviene sempre in tutte le lingue, sta mutando in maniera significativa.

Nella tabella (2) si riportano i fenomeni annotati nei testi.

Tab. 2 - Criteri di annotazione manuale

Parametri	*Tag*	*Fenomeni*
Organizzazione del testo	PAR1, PAR2, PAR3	Mancata/scorretta suddivisione del testo in paragrafi Esposizione schematica dei contenuti
Ortografia	ORT	Assenza/impiego scorretto dell'apostrofo Uso dell'accento con forme verbali monosillabiche (*fa*, *sa*, *so*) e con la forma apocopata dell'avverbio *poco* Da non considerare: errori di battitura; uso dell'accento grave al posto di quello acuto e viceversa
Registro	REG	Lessico non adeguato al contesto scritto sorvegliato mediamente formale Uso di *gli* sovraesteso per *loro* e *le* Uso del *tu* impersonale
Frasi marcate	MRC	Frasi dislocate a destra/sinistra Frasi scisse e pseudoscisse Frasi a tema sospeso
Lessico	LES	Povertà/eccessiva genericità lessicale Lessico improprio Ripetizioni Plastismi Violazione di collocazioni Malapropismi

9. Queste "deviazioni" sono, di fatto, considerabili "errori" in ogni tipo di testo e in effetti vengono corrette da ogni insegnante, senza eccezioni.

10. Su questi tratti gli interventi degli e delle insegnanti sono meno omogenei: in alcuni casi, cioè, vengono corretti; in altri vengono tollerati (cfr. Grandi e Zucchini, 2022).

Tab. 2 - Segue

Parametri	*Tag*	*Fenomeni*
Punteggiatura	PUN	Omissione dei segni interpuntivi Sostituzione di un segno interpuntivo con un altro Inserimento di segni interpuntivi in contesti incongrui
Morfosintassi	MFS	Mancato accordo per genere e numero Mancato rispetto della consecutio temporum Inadeguata gestione del riferimento (ad esempio, pronomi distanti dai loro antecedenti, pronomi che rimandano a referenti dotati di realtà concettuale anziché testuale) Reggenze preposizionali errate *Che* polivalente
Coerenza	COE	Uso illogico dei connettivi Mancata esplicitazione delle relazioni logiche che intercorrono fra i contenuti espressi (giustapposizione) Contraddittorietà Frammentazione delle informazioni "Cortocircuiti semantici", come in espressioni del tipo "possibile/impossibile" + potere; permettere + potere; ecc.
Sintassi e coesione	SIN	Omissione della preposizione nella coordinazione di sintagmi Mancati o scorretti parallelismi Gerundi assoluti Omissioni argomentali, ad esempio: "ricominciare a recarsi in presenza" Interruzione della continuità sintagmatica, ad esempio: "Basterebbe pensare alle famiglie che vivono, magari anche numerose, in un monolocale"; "Sono, infine, felice delle scelte fatte dai miei professori"

I testi sono stati poi "taggati" utilizzando le annotazioni apposte, come in (1), in cui viene riportato un esempio di annotazione per ogni classe citata in tab. 2 ad eccezione della prima (PAR):

(1) ORT: …molto in soggezione gli studenti e condiziona l'esito dell'esame, **{ORT sopratutto}** per chi riesce meno a gestire l'ansia…

REG: …Spesso durante le lezioni capita che la connessione **{REG va via}** e i ragazzi non riescono più a seguire…

LES: …La didattica a distanza ci ha inseriti in un **{LES cerchio}** che porta l'ansia a dominare tutti i nostri momenti quotidiani…

PUN: …**Personalmente{PUN}** poi, ho trovato la didattica a distanza molto più difficile da seguire{PUN ,} rimanere a casa nella mia esperienza è stato anche…

MRC: …a settembre **{MRC il rettore ha preso lui la decisione}**…

MFS: …paragonati ai disagi che la didattica a distanza **{MFS possa}** recare ad alunni delle scuole inferiori o primarie, **{MFS dove loro}**, nel loro sviluppo e nella loro crescita dal punto di vista sociale…

COE: …portando magari molti studenti a rimandare la lezione del giorno a data da definirsi, **{COE 'tanto è registrata'}**, per poi ritrovarsi a un passo dall'esame con quarantina di registrazioni da dover ascoltare…

SIN: …le aule possono avere un numero non limitato di partecipanti, **{SIN cosa che potrebbe accadere con le aule fisiche}** per problemi di…

Tanto i parametri quantitativi ricavati in automatico quanto le annotazioni apposte manualmente costituiscono uno strumento per la consultazione del corpus Univers-ITA e per l'estrazione dei dati[11].

2.3. *Metadati*

Grazie alle informazioni raccolte con il questionario sociobiografico (cfr. cap. 2), il corpus è consultabile anche attraverso numerosi filtri di ricerca ricavati dalle risposte (eventualmente aggregate) di studenti e studentesse. Di seguito, si riportano i diversi filtri e i loro valori che è possibile selezionare, con eventuali ulteriori specifiche.

i. *Sede degli studi*
   - Nord
   - Centro
   - Sud

   È possibile selezionare anche la città in cui è stato raccolto il testo.

ii. *Corso*
   - Area sanitaria
   - Area scientifica
   - Area sociale
   - Area umanistica

11. Ogni annotazione manuale ha ovviamente un margine di soggettività e necessita, dunque, di revisioni periodiche. A fine 2024 è stata avviata la prima revisione delle annotazioni finalizzata sia ad eliminare alcune incongruenze, sia ad integrare parametri non considerati nella prima annotazione. Le revisioni saranno tracciate sul sito del progetto.

iii. *Genere*
- Maschio
- Femmina
- Altro
- Non dichiarato

iv. *Età*
È possibile selezionare valori singoli (18, 19, ecc.), fino all'ultimo valore (>30).

v. *Luogo di nascita*
- Estero
- Italia
  - Nord
  - Centro
  - Sud

Per ciascuna delle macroaree è possibile selezionare una singola regione.
- Non dichiarato

vi. *Disturbi di lettura*
- Sì
  - Ipovedente
  - DSA
  - Altro
- No
- Non dichiarato

vii. *Scolarizzazione*
- All'estero
- In Italia
- In Italia dalle medie
- In Italia dalle superiori
- Altro

viii. *Scuole superiori*
- Istituto professionale
- Istituto tecnico
- Liceo

Per ciascuna tipologia di scuola superiore, è poi possibile selezionare la fascia del voto di uscita, classificata come segue: sufficiente (60-70), discreto (71-80), buono (81-90), ottimo (91-100).

ix. *Origine della famiglia*
- Estero
- Italia
- Mista

x. *Scolarizzazione dei genitori*
- Entrambi con al massimo la licenza elementare
- Entrambi con la licenza media
- Entrambi diplomati
- Entrambi laureati
- Un genitore diplomato (e l'altro con un titolo di studio inferiore)
- Un genitore laureato (e l'altro con un titolo di studio inferiore)
- Un solo genitore con al massimo la licenza media
- Non dichiarato

xi. *Lingua dei genitori verso i figli*
- Italiano
- Italiano e altra lingua
- Italiano e dialetto o lingua minoritaria
- Italiano, dialetto o lingua minoritaria e altra lingua
- Dialetto o lingua minoritaria
- Altra lingua
- Non dichiarato

xii. *Plurilinguismo*
- No
- Sì

In associazione al valore *sì*, è possibile poi raffinare ulteriormente la ricerca selezionando il numero di lingue coinvolte e il contesto (formale o informale) in cui il rispondente dichiara di utilizzarle.

xiii. *Letture*

In questo filtro, sono raccolti tre parametri:
- Interessi (ad es. fumetti, poesie, ecc.)
- Numero di libri letti all'anno (da meno di cinque a più di 10)
- Propensione familiare alla lettura (questo parametro è stato ricavato chiedendo ai rispondenti di dichiarare se i genitori (uno o entrambi) leggessero abitualmente)

xiv. *Scrittura*

È possibile selezionare testi di studenti e studentesse che hanno frequentano corsi di scrittura (sì/no) e/o dichiarano di scrivere abitualmente (sì/no).

xv. *Appunti*
È possibile selezionare testi di studenti e studentesse che dichiarano di prendere appunti (sì/no) a lezione oppure in preparazione ad un esame e la modalità con cui prendono appunti (ad es. a mano o a computer).

xvi. *Frequenza scrittura universitaria*
- Molto spesso
- Spesso
- Qualche volta
- Mai

xvii. *Redazione*
In questo filtro sono raccolte informazioni circa:
- L'impiego (abituale o meno) del correttore automatico
- Il tipo di correzioni ricevute da docenti su produzioni scritte (come, ad es., chiarezza, coesione, lessico, ecc.)
- L'abitudine di redigere una scaletta per la stesura di testi scritti
- Gli strumenti utilizzati per scrivere (come, ad es., computer, smartphone, ecc.)

xviii. *Annotazione*
Attraverso questo filtro è possibile selezionare i testi in cui sono state inserite annotazioni sui diversi livelli di analisi:
- Coerenza
- Lessico
- Marcatezza
- Morfosintassi
- Ortografia
- Punteggiatura
- Registro
- Sintassi

Questi filtri possono ovviamente combinarsi ai precedenti e, dunque, permettono di effettuare ricerche complesse e piuttosto articolate e di estrarre i dati dal corpus con una granularità realmente molto fine (ad esempio: estrarre tutte le annotazioni relative alle frasi marcate in testi redatti in atenei dell'Italia settentrionale da studenti di genere maschile con poca propensione alla lettura iscritti a corsi di area sociale, ecc.).

Per l'elaborazione statistica dei dati (cfr. cap. 4) è stato costruito un dataset[12] nel quale ogni testo (identificato da un codice numerico) è abbinato ai dati

12. Grandi Nicola, Pascoli Matteo (2025), *Dataset del progetto Univers-ITA*, Università di Bologna, https://doi.org/10.6092/unibo/amsacta/8229

dell'ateneo e del corso di studio di chi lo ha redatto, ai dati ricavati dal questionario sociobiografico appena elencati, alle misurazioni quantitative ottenute da READ-IT e al numero di annotazioni apposte per ognuna delle classi di annotazione indicate in tab. 2.

3. I testi non-*ad hoc*

In questa sezione, si descrivono brevemente gli altri 2 corpora creati all'interno del progetto Univers-ITA. A differenza di quanto visto sino ad ora, in queste due risorse sono contenuti testi prodotti per altre finalità, quindi indipendentemente dal progetto, e che sono stati raccolti ex post. Per questa ragione, in entrambi i casi l'inventario dei metadati è ridotto rispetto al corpus descritto in §2 e dunque anche le possibilità di interrogazione sono più limitate.

3.1. *Univers-ITA ProGior*[13]

Il corpus Univers-ITA ProGior contiene articoli di carattere giornalistico scritti da studenti e studentesse ed estratti da blog, giornali universitari online e siti di informazione/opinione gestiti da studenti e studentesse universitari di diversi atenei italiani. I testi sono stati raccolti grazie alla collaborazione di studenti e studentesse dell'Università di Bologna che hanno preso parte al tirocinio Univers-ITA, attivo per gli anni accademici 2020/2021 e 2021/2022. I e le tirocinanti si sono occupati in prima istanza di selezionare siti web che ospitassero giornali e blog e, successivamente, di volta in volta, hanno contattato i e le responsabili delle diverse testate per illustrare il progetto e fare loro firmare una liberatoria per l'utilizzo dei testi. Infine, hanno classificato i testi raccolti per argomento trattato (ad es. arte-cultura, economia-società, ecc.), per collocazione geografica dell'ateneo associato al sito (Nord, Centro, Sud e Isole) e per anno di redazione del testo (dal 2012 al 2021). Queste tre informazioni sono attualmente utilizzabili come filtri di ricerca per l'interrogazione del corpus. Tutti i testi sono stati poi lemmatizzati ed etichettati per parte del discorso. Il corpus è composto da 1.630 testi per un totale di 1.692.846 parole.

13. Grandi Nicola, Ballarè Silvia, Chiusaroli Francesca, Gallina Francesca, Pascoli Matteo, Pistolesi Elena; *Corpus Univers-ITA-ProGior.* 2023, https://doi.org/10.60760/unibo/univers-ita-progior

3.2. *Univers-ITA ProUniv*[14]

Il corpus Univers-ITA ProUniv è costituito da tesi, tesine, relazioni, ecc. nella versione *non* corretta dai e dalle docenti. Complessivamente, contiene 773 testi per 6.267.765 tokens.

Anche in questo caso, la risorsa è stata costruita grazie al prezioso contributo di studenti e studentesse dell'Università di Bologna che hanno partecipato al tirocinio Univers-ITA (a.a. 2020/2021 e 2021/2022). I e le tirocinanti si sono occupati innanzitutto di cercare i testi facendo affidamento, prevalentemente ma non solo, sulla propria rete di contatti e relazioni, e cercando, per quanto possibile, di diversificare studenti e studentesse coinvolti in base sia alla collocazione geografica degli atenei, sia al corso di laurea frequentato. Nonostante sia sempre stato acquisito il consenso esplicito di autori e autrici all'utilizzo dei testi, per ragioni legate al diritto d'autore e alla normativa di riferimento, i testi non sono accessibili globalmente, ma viene restituito un contesto sufficientemente ampio per analizzare le forme estratte. Inoltre, è bene specificare che alcuni testi sono stati inseriti nel corpus integralmente; altri, invece, solo parzialmente su richiesta degli e delle scriventi (ad esempio per mantenere riservatezza su sezioni di tesi contenenti dati inediti o ipotesi ancora in via di verifica). I testi sono stati ripuliti eliminando le sezioni dedicate ai riferimenti bibliografici e le citazioni molto lunghe (esterne al corpo del testo) e poi classificati secondo i seguenti metadati, utilizzabili come filtri di ricerca[15]:

i. *Tipo*
   - Esame
   - Lettera
   - Progetto
   - Recensione
   - Relazione
   - Tesi
   - Tesina
   - Altro

14. Grandi Nicola, Ballarè Silvia, Chiusaroli Francesca Gallina, Francesca Pascoli Matteo, Pistolesi Elena, *Corpus Univers-ITA-ProUniv.* 2023, https://doi.org/10.60760/unibo/univers-ita-prouniv

15. Il contatto diretto con gli autori e le autrici dei testi raccolti in questa fase ha permesso infatti di raccogliere un buon numero di informazioni che sono attualmente utilizzabili come filtri di ricerca.

ii. *Corso:*
  – Area sanitaria
  – Area scientifica
  – Area sociale
  – Area umanistica

iii. *Sede ateneo:*
  – Nord
  – Centro
  – Sud

iv. *Tipologia del corso:*
  – A ciclo unico
  – Magistrale
  – Triennale

v. *Anno accademico di redazione del testo*: dal 2013-2014 al 2021-2022

vi. *Area di nascita:*
  – Nord
  – Centro
  – Sud
  – Estero

vii. *Genere:*
  – Femmina
  – Maschio
  – Altro/Non risponde

Il corpus può essere consultato in modalità bilanciata o non bilanciata (ovvero nella sua interezza). Infatti, impiegando gli stessi parametri di campionamento adottati per il corpus dei testi *ad hoc* (per cui cfr. §2), è stato creato a posteriori un sottocorpus rappresentativo per area geografica dell'ateneo e per area disciplinare dei corsi di studio. Questo sottocorpus ha dimensioni piuttosto ridotte, è infatti costituito da 254 testi e 2.578.072 parole.

Riferimenti bibliografici

Ballarè S. (2020), "L'italiano neo-standard oggi: stato dell'arte", *Italiano Linguadue*, 12: 469-492, testo disponibile al sito: https://riviste.unimi.it/index.php/promoitals/article/view/15013

Berruto G. (2012 [1987]), *Sociolinguistica dell'italiano contemporaneo*, Carocci, Roma.

Dell'Orletta F., Montemagni S. and Venturi G. (2011), *READ-IT: assessing readability of Italian texts with a view to text simplification*, in Alm N., ed., *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, Association for Computational Linguistics, Edinburgh, 73-83, testo disponibile al sito: https://aclanthology.org/W11-2308/

Grandi N. (2024), *L'italiano neo-standard*, in Ballarè S., Fiorentini I. e Miola E., a cura di, *Le varietà dell'italiano contemporaneo*, Carocci, Roma, 33-48.

Grandi N., Ballarè S., Martari Y. e Miola E. (2024), "Univers-ITA. Descrizione e primi risultati di uno studio dell'italiano scritto di studenti universitari", *Italiano a stranieri*, 35: 19-26, testo disponibile al sito: https://flip.edilingua.it/RIV_35/

Grandi N. e Zucchini E. (2022), "Tratti neostandard nella scrittura formale giovanile. Un'indagine sulle scuole secondarie di Bologna", *Rassegna Italiana di Linguistica Applicata – RILA* 2021, 3: 121-138.

Materiali Linguistici

L'italiano è davvero una lingua in declino? È vero che gli studenti e le studentesse oggi hanno difficoltà a scrivere e comprendere testi complessi? La tecnologia ha davvero impoverito la nostra lingua? Queste domande corrispondono a una serie di luoghi comuni molto diffusi nell'opinione pubblica, ma quasi mai confermati dai dati. Questo volume raccoglie i primi risultati di un progetto di ricerca pluriennale finalizzato a produrre la prima mappatura sistematica mai realizzata delle capacità di scrittura formale della popolazione studentesca universitaria italiana. Un gruppo di ricerca interdisciplinare (composto da linguisti, statistici, informatici e sociologi) ha analizzato sia i principali tratti linguistici e grammaticali, sia i più importanti correlati sociobiografici della produzione scritta formale di un campione di oltre 2.000 studenti e studentesse di 44 atenei italiani, rappresentativo per aree disciplinari e geografiche. Il quadro che emerge è molto complesso e articolato: un confronto con dati del passato pare mostrare, in alcune aree, un seppur lieve miglioramento nelle competenze di scrittura; tuttavia, esistono aree di oggettiva debolezza, come, ad esempio, nelle attività di pianificazione del testo.

Contributi di: S. Ballarè, F. Chiusaroli, F. Da Milano, E. Di Domenico, M. Farnè, F. Gallina, C. Gianollo, N. Grandi, A. Iannella, Y. Martari, E. Miola, S. Scaglione, M. Pascoli, F. Pecorari, M.L. Pierucci, E. Pistolesi, R. Pugliese, M.L. Restivo, M. Tavosanis, L. Tramutoli, S. Tusini, S. Valente.

Nicola Grandi è professore ordinario di Glottologia e Linguistica presso il Dipartimento di Filologia classica e Italianistica dell'Alma Mater Studiorum – Università di Bologna. Si occupa di tipologia linguistica, di sociolinguistica, con particolare riferimento alle tendenze dell'italiano contemporaneo, e di educazione linguistica. Ha fondato e dirige il sito divulgativo *www.linguisticamente.org*.

FrancoAngeli
La passione per le conoscenze