AItLA

ASSOCIAZIONE ITALIANA DI LINGUISTICA APPLICATA

XXI CONGRESSO INTERNAZIONALE

UNIVERSITÀ DEGLI STUDI DI BERGAMO | Dipartimento di Lingue, Letterature e Culture Straniere

FARE LINGUISTICA APPLICATA CON LE DIGITAL HUMANITIES

RELAZIONI SU INVITO

JOANNE CLELAND | UNIVERSITY OF STRATHCLYDE

FRANCESCO CUTUGNO | UNIVERSITÀ DEGLI STUDI DI NAPOLI, FEDERICO II

CHRISTOPH DRAXLER | LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

ISTRUZIONI PER LA PARTECIPAZIONE

Per partecipare alle diverse sessioni del XXI Congresso dell'Associazione Italiana di Linguistica Applicata è necessario cliccare sui link che seguono e che rimandano alla piattaforma di comunicazione Microsoft Teams.

L'accesso alla piattaforma è possibile sia tramite la app Microsoft Teams scaricabile in locale, che tramite web browser. In questo secondo caso vi preghiamo di accedere annunciandovi con il vostro nome e cognome e non con uno pseudonimo.

In ogni caso, vi preghiamo di accedere tenendo spenti microfono e webcam.

Per la gestione della fase di discussione delle relazioni vi preghiamo di attenervi alle indicazioni che riceverete dai presidenti di sessione.

GIOVEDÌ 11 FEBBRAIO 2021	
Mattina (9:00 – 13:00)	LINK1
Pomeriggio (14:00 - 17:30)	LINK2
VENERDÌ 12 FEBBRAIO 2021	
Mattina (9:00 – 12:00)	LINK3
Pomeriggio (13:30 – 16:45)	LINK4
Assemblea dei Soci (17:00)	LINK5

PROGRAMMA

Giovedì 11 febbraio

9:00 Saluti di apertura

9:30 *Joanne Cleland*
Clinical Applications of Articulatory Phonetics:
Using Ultrasound Tongue Imaging in the Speech Therapy Clinic

10:30 *Marta Maffia, Massimo Pettorino, Rosa De Micco, Gioacchino Tedeschi, Alessandro Tessitore, Anna De Meo*
Analisi della voce e malattia di Parkinson: uno studio sull'italiano di pazienti in stadi iniziali

11:00 Pausa

11:30 *Claudia Marzi, Loukia Taxitari, Marcello Ferro, Andrea Nadalini, Vito Pirrelli*
Valutare la lettura "in tempo reale": un esempio di integrazione tra linguistica computazionale e linguistica applicata

12:00 *Valentina De Iacovo, Marco Palena*
L'utilizzo di un chatbot Telegram per la didattica assistita per apprendenti di italiano L2 e nella valutazione linguistica delle conoscenze disciplinari

12:30 *Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi*
Profiling-UD: strumento per la ricostruzione automatica del profilo linguistico di corpora multilingue

13:00 Pausa

14:00 *Francesco Cutugno*
Beni culturali, linguistica e nuove tecnologie:
i molti volti della linguistica applicata alle esperienze di cultura digitale

15:00 *Chiara Branchini, Lara Mantovan*
Grammatica della lingua dei segni italiana e digitalizzazione: una nuova frontiera per una più ampia accessibilità

15:30 Pausa

16:00 *Valentina De Iacovo, Elisa Di Nuovo, John Hajek*
L'analisi degli errori in un treebank orale di parlanti italo-australiani attraverso l'annotazione IOB integrata in CoNLL-U

16:30 *Flavia Sciolette, Emiliano Giovannetti*
Un modello per domarli tutti: verso una rappresentazione del testo come esplicitazione di documento, lingua e contenuto

17:00 *Antonio Mastropaolo, Daniele P. Radicioni, Luisa Revelli*
La lingua del legislatore in Valle D'Aosta: studio di caso e prospettive applicative

PROGRAMMA

Venerdì 12 febbraio

9:00 *Christoph Draxler*
Automatic Transcription of Spoken Language Using Publicly Available Web Services

10:00 *Giovina Angela del Rosso, Silvia Brambilla*
L'accuratezza delle trascrizioni ASR sul parlato non-standard. L'italiano nell'OH Portal

10:30 Pausa

11:00 *Nicola Brocca*
LADDER: La costruzione e analisi di un corpus di scritture digitali per l'insegnamento della pragmatica in L2 attraverso le digital humanities

11:30 *Valeria Caruso, Roberta Presta*
Approcci digitali alla lessicografia: sviluppo e valutazione di un dizionario per smartphone

12:00 Pausa

13:30 *Elisa Gugliotta, Angelapia Massaro, Giuliano Mion, Marco Dinarelli*
Un'analisi statistica della definitezza in un corpus di arabish tunisino

14:00 *Elisa Di Nuovo, Cristina Bosco, Elisa Corino*
Analisi sintattica e dell'errore in un corpus di apprendenti di italiano: il sintagma nominale

14:30 *Francesca Pagliara*
Validare schemi di codifica in ambito linguistico: l'Intercoder Agreement applicato alla tassonomia del Cross-Cultural Speech Act Realization Patterns Project (CCSARP)

15:00 Pausa

15:30 *Alessandro Puglisi*
Dentro cascate di testo: una proposta per l'analisi del pivoting in un canale Twitch

16:00 *Simone Ciccolone, Giulia Isabella Grosso*
Macchine parlanti o macchine apprendenti? Prime riflessioni sulle interlingue simulate dei chatbot

16:30 Saluti di chiusura

17:00 Assemblea dei soci AItLA

ABSTRACT DELLE COMUNICAZIONI SELEZIONATE

(IN ORDINE DI PRESENTAZIONE IN PROGRAMMA)

GIOVEDÌ MATTINA

Marta Maffia, Massimo Pettorino, Rosa De Micco, Gioacchino Tedeschi, Alessandro Tessitore, Anna De Meo

Analisi della voce e malattia di Parkinson: uno studio sull'italiano di pazienti in stadi iniziali

Studi sperimentali condotti su diverse lingue hanno evidenziato come i cambiamenti anatomici e fisiologici dovuti all'insorgere della malattia di Parkinson possano determinare variazioni nei parametri acustici della voce, sia a livello segmentale sia soprasegmentale [1; 2]. L'osservazione di tali fenomeni, correlati alla perdita di dopamina nel sistema nervoso centrale, potrebbe rappresentare uno strumento sostenibile e non invasivo di supporto alla diagnosi clinica della patologia, a partire dagli stadi iniziali della malattia [3].
In precedenti studi, la durata media del VtoV, ossia l'intervallo tra due consecutivi punti di inizio vocalici, e la percentuale di porzioni vocaliche sull'enunciato (%V), sono stati identificati come parametri acustici in grado di rendere conto efficacemente delle variazioni ritmiche del parlato parkinsoniano in diverse lingue, quando confrontato con parlato "sano" [4; 5].
Nel presente studio si intende verificare la validità di tali parametri anche nella descrizione di produzioni vocali di soggetti agli stadi iniziali della malattia. A tale scopo, sono stati finora coinvolti 20 pazienti (di cui 6 donne, con un'età media di 65 anni) inseriti in uno specifico percorso di monitoraggio clinico della malattia di Parkinson, diagnosticata negli ultimi due anni. Sono stati inoltre reclutati 20 soggetti sani, comparabili ai pazienti per età e sesso, con funzione di gruppo di controllo. Tutti i partecipanti alla ricerca, residenti in Campania, hanno dichiarato l'italiano come lingua materna.
Il protocollo di raccolta dati ha previsto la registrazione di parlato letto, controllato e spontaneo per ciascun soggetto. L'analisi della voce, per ora effettuata solo sul parlato letto, è consistita nella segmentazione manuale del segnale acustico in porzioni vocaliche e consonantiche, con il software Praat. Dopo aver estratto le durate dei segmenti, sono stati calcolati per ciascun parlante la %V e il valore medio del VtoV.
I primi risultati confermano che la %V è il parametro che cambia in maniera più significativa nei due gruppi di soggetti coinvolti, con valori più alti nel parlato parkinsoniano rispetto a quello "sano": se nel gruppo di controllo i valori vanno dal 42% al 47%, nei soggetti affetti da Parkinson la %V si attesta tra il 48% e il 54%. Rispetto al valore medio del VtoV, è stata riscontrata una tendenza da parte dei soggetti parkisoniani a mostrare valori più alti anche di tale parametro, e quindi a parlare a una velocità più bassa rispetto ai parlanti sani.
Tali dati, andranno ulteriormente validati con l'ampliamento del corpus e con l'estensione dell'analisi acustica anche al parlato controllato e spontaneo.

[1] SKODDA S., GRONHEIT W., SCHLEGEL U. (2012), Impairment of vowel articulation as a possible marker of disease progression in Parkinson's Disease, *PLoS ONE*, 7/2, 1–8.
[2] LISS J. M., WHITE L., MATTYS S. L., LANSFORD K., LOTTO A. J., SPITZER S. M., CAVINESS J. N. (2009), Quantifying speech rhythm abnormalities in the dysarthrias, *Journal of Speech, Language, and Hearing Research*, 52, 1334–1352.
[3] HAREL B. T., CANNIZZARO M. S., COHEN H., REILLY N., SNYDER P. J. (2004), Acoustic characteristics of Parkinsonian speech: a potential biomarker of early disease progression and treatment, *Journal of Neurolinguistics* 17, 439–453.
[4] PETTORINO M., BUSÀ M.G., PELLEGRINO E. (2016), Speech Rhythm in Parkinson's Disease: A Study on Italian, *Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, 1958-1961.

[5] PETTORINO, M., GU W., PÓŁROLA P., FAN P. (2017), Rhythmic Characteristics of Parkinsonian Speech: A Study on Mandarin and Polish, *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*, 3172-3176.

Claudia Marzi, Loukia Taxitari, Marcello Ferro, Andrea Nadalini, Vito Pirrelli

Valutare la lettura "in tempo reale": un esempio di integrazione tra linguistica computazionale e linguistica applicata

In anni recenti, linguistica computazionale e linguistica applicata hanno ampliato i loro rispettivi ambiti d'indagine, utilizzando l'ontologia formale della linguistica teorica e i modelli cognitivi della psicolinguistica per studiare le dinamiche attraverso cui i parlanti si avvalgono delle proprie competenze linguistiche per svolgere "compiti" specifici (ad es. comprendere o tradurre un testo).
Nell'ambito della lettura, le tecnologie per il Trattamento Automatico del Linguaggio (TAL) si sono dimostrate capaci di classificare il livello di leggibilità di un testo, basandosi sulla distribuzione di alcuni parametri linguistici in testi pre-classificati per età dei lettori destinatari, o per grado di scolarità, o per livello di sviluppo cognitivo [1]. Ad esempio, parole o frasi più lunghe, o parole più rare tendono a distribuirsi in testi di più difficile comprensione, o destinati a lettori più maturi. E' possibile così assegnare a un testo, o a ogni singola frase, un punteggio di leggibilità in funzione (inversa) della complessità lessicale, morfologica, sintattica o pragmatica dell'unità testuale analizzata.
In Linguistica Applicata (LA) la valutazione della difficoltà di lettura ha seguito un approccio funzionale [2]. Nel *modello semplice di lettura* [3], ad esempio, la capacità di leggere un testo è analizzata come il prodotto dell'interazione tra *decodifica* e *comprensione*. Attraverso l'osservazione di un campione di bambini impegnati nella lettura, è possibile valutare la loro fluenza in decodifica, gli errori di decodifica e comprensione, e l'efficacia di percorsi educativi personalizzati.
La piattaforma *ReadLet* è stata sviluppata con l'obiettivo di integrare l'approccio classificatorio del TAL con quello funzionale della LA [4]. Il bambino legge un breve testo visualizzato sullo schermo di un tablet, ad alta voce o in modalità silente. In entrambi i casi, al bambino viene chiesto di "tenere il segno" con il dito sullo schermo nel corso della lettura. La traccia tattile è registrata e allineata con il testo visualizzato sullo schermo mediante un algoritmo di convoluzione. Al contempo, il testo è annotato automaticamente per tratti linguistici. Alla fine della sessione di lettura silente, il bambino risponde ad alcune semplici domande sul contenuto del testo. I dati raccolti consentono di valutare le difficoltà (rallentamenti o errori) che il bambino incontra nella lettura, e di mettere in relazione "in tempo reale" queste difficoltà con aspetti linguistici specifici del testo. Un'analisi preliminare dei dati raccolti da *ReadLet* su oltre 400 allievi di alcune scuole elementari toscane e della Svizzera italiana, ha evidenziato il differente "passo" di lettura tra lettori con sviluppo tipico e atipico, e il peso che variabili come lunghezza, frequenza e lessicalità hanno su profili di lettura individuali e aggregati. La possibilità di "controllare" automaticamente la distribuzione di queste variabili nel testo e di correlarle con le difficoltà del singolo bambino consente, infine, di somministrare testi con livelli di difficoltà gradualmente crescenti, rendendo possibili percorsi personalizzati di potenziamento.

[1] Dell'Orletta, F., S. Montemagni, & G. Venturi. (2011). READ–IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, 73–83.
[2] Grabe, W. & F. L. Stoller. (2019) *Teaching and researching reading*, 3rd ed. Routledge.
[3] Hoover, W. A. & P. B. Gough. (1990) The simple view of reading. *Reading and Writing: An Interdisciplinary Journal*, 2, 127–160.
[4] Ferro, M., C. Cappa, S. Giulivi, C. Marzi, O. Nahli, F. A. Cardillo & V. Pirrelli. (2018). Readlet: Reading for understanding. In *IEEE 5th International Congress on Information Science and Technology* (CiSt), 1–6.

Valentina De Iacovo, Marco Palena

L'utilizzo di un chatbot Telegram per la didattica assistita per apprendenti di italiano L2 e nella valutazione linguistica delle conoscenze disciplinari

Confrontare la pronuncia di apprendenti di una lingua straniera con enunciati di parlanti nativi (Delmonte 2009) sta ricevendo sempre più attenzione grazie anche alle numerose applicazioni che nascono nell'ambito della didattica assistita. Parallelamente gli studi glottodidattici sulla variazione prosodica tra più parlanti nativi fanno emergere una variabilità ritmico-intonativa che non può essere ridotta a pochi modelli eligibili ma, al contrario, dovrebbe essere parte del bagaglio linguistico dell'apprendente. Seguendo questa direzione, in questo studio presentiamo un *chatbot* pensato come supporto di apprendimento proattivo per il miglioramento delle competenze orali in italiano L2. Realizzato all'interno dell'applicazione di messaggistica istantanea Telegram, il *chatbot* prevede l'interazione con l'utente attraverso domande e risposte basate sulla valutazione di conoscenze disciplinari. In particolare, propone all'apprendente una serie di domande a risposta chiusa (quiz) che possono avere carattere generale di comprensione linguistica oppure essere legate a un particolare ambito disciplinare. All'individuazione della risposta corretta, l'apprendente ha la possibilità di ascoltare la stessa prodotta da parlanti madrelingua. A questo punto, l'apprendente invia al *chatbot* la propria risposta sotto forma di nota vocale. Questo è in grado di confrontare, in maniera automatica, la risposta dell'utente con un archivio di risposte date da parlanti madrelingua e trovare quindi quella che più si avvicina in termini prosodici a quella dell'apprendente. A partire da questa vicinanza prosodica l'utente riceve infine un riscontro sul proprio livello di competenza orale. Inoltre, l'impostazione a quiz permette la valutazione di eventuali criticità linguistiche come ad esempio la pronuncia di date o formule matematiche.

BOUREUX M. & BATINTI A. (2004), La prosodia. Aspetti teorici e metodologici nell'apprendimento-insegnamento delle lingue straniere, in *Atti delle XIV Giornate di Studio del Gruppo di Fonetica Sperimentale*, Esagrafica, Roma: 233-238.

CAZADE A. (1999), De l'usage des courbes sonores et autres supports graphiques pour aider l'apprenant en langues, in *ALSIC (Apprentissage des Langues et Systèmes d'Information et de Communication, online)*, 2(2): 3-32.

CHUN D. M. (2002), *Discourse Intonation in L2: From theory and research to practice*, Benjamins, Amsterdam.

DELMONTE R. (2009), Prosodic tools for language learning, in *International Journal of Speech Technology* 12(4): 161-184.

FERNOAGĂ V., STELEA GA., GAVRILĂ C. & SANDU F. (2018), Intelligent education assistant powered by chatbots, in *The International Scientific Conference eLearning and Software for Education* 2: 376-383.

LACHERET-DUJOUR A. (2001), Modéliser l'intonation d'une langue. Où commence et où s'arrête l'autonomie du modèle? L'exemple du français parlé, in *Actes du colloque international Journées Prosodie2001*, 57-60.

PEREIRA J. (2016), Leveraging chatbots to improve self-guided learning through conversational quizzes, in *Proceedings of the fourth international conference on technological ecosystems for enhancing multiculturality*, TEEM '16, ACM Press, New York: 911-918.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi

Profiling-UD: strumento per la ricostruzione automatica del profilo linguistico di corpora multilingue

La disponibilità crescente di corpora di grandi dimensioni rappresentativi di diverse varietà d'uso reale della lingua e di strumenti di Trattamento Automatico della Lingua (TAL) sempre più affidabili nell'analisi della struttura linguistica di un testo ha reso possibile l'impiego di approcci computazionali per lo studio di temi storicamente di interesse per la comunità di ricerca in linguistica applicata e delle scienze umane (Digital Humanities). È il caso ad esempio dell'analisi computazionale delle variazioni tra generi e registri anche in una prospettiva diacronica [1], della Sociolinguistica Computazionale [7] impegnata nello studio della variazioni diastratiche della lingua, della Stilometria Computazionale [5] il cui obiettivo è quello di attribuire un testo al proprio autore sulla base delle caratteristiche linguistiche dello stile di scrittura (o di verificarne l'attribuzione), o ancora dell'analisi della complessità linguistica di un testo [4]. Tradizionalmente questi studi si ispirano al framework dell'Analisi multidimensionale/fattoriale [2] basata sul conteggio di indicatori linguistici semplici, quali la frequenza di parole funzionali ricavate da liste predefinite o di n-grammi di caratteri, considerati buone approssimazioni di differenze stilistiche tra varietà funzionali della lingua. Oggi, la sempre maggior affidabilità delle analisi generate da strumenti di TAL, rende possibile il monitoraggio ad ampio raggio di una grande quantità di caratteristiche del testo che ne modellano fenomeni grammaticali, lessicali e semantici distintivi del profilo linguistico [9].
Ponendoci in questo contesto di ricerca, nel presente contributo presentiamo *Profiling-UD* [3], strumento che permette di ricostruire il profilo linguistico di testi a partire dall'estrazione automatica di una vasta gamma di caratteristiche (circa 130) acquisite dal testo linguisticamente annotato. A differenza di altri sistemi esistenti, è espressamente progettato per essere multilingue dal momento che si basa sulla rappresentazione morfo-sintattica e sintattica a dipendenze definita nell'ambito del progetto Universal Dependencies (UD) [8]. Facendo riferimento allo schema di annotazione UD comune per la rappresentazione di più di 90 lingue, lo strumento permette di realizzare confronti intra- e inter-linguistici. Un'altra delle peculiarità è costituita dalla vasta gamma di caratteristiche che è possibile estrarre dai testi analizzati. Tali caratteristiche spaziano tra caratteristiche di base del testo (lunghezza del testo, frasi e parole), informazioni relative alla ricchezza lessicale (rapporto tipi/unità), informazioni morfo-sintattiche (es. distribuzione delle categorie dello schema UD, densità lessicale), anche relative alla struttura di predicati verbali (es. valenza), alla struttura locale (es. lunghezza media delle relazioni di dipendenza) e globale (es. profondità massima) dell'albero sintattico delle frasi di un testo o ancora relative all'uso della subordinazione.
Lo strumento è liberamente accessibile per scopi di ricerca. Al momento è stato, ad esempio, impiegato con successo per studiare l'evoluzione delle abilità di scrittura di studenti di università statunitensi apprendenti lo spagnolo come lingua seconda [6]. Sono in corso specializzazioni, quali l'individuazione di nuovi parametri linguistici oggetto di monitoraggio, a supporto di ulteriori applicazioni.

[1] Argamon, S. (2019). Computational register analysis and synthesis. In: Register Studies, 1(1):100-135.

[2] Biber, D. (1993). Using register-diversified corpora for general language studies. Computational Linguistics, 19(2):219–242.

[3] Brunato D., Cimino A., Dell'Orletta F., Montemagni S., Venturi G. (2020). Profiling-UD: a Tool for Linguistic Profiling of Text. In: Proceedings of 12th Edition of LREC, 11-16 May, 2020, Marseille, France.

[4] Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. In: ITL - International Journal of Applied Linguistics, 165(1):97-135.

[5] Daelemans, W. (2013). Explanation in computational stylometry. In: Computational Linguistics and Intelligent Text Processing, pages 451-462, Berlin, Heidelberg, Springer Berlin Heidelberg.

[6] Miaschi A., Davidson S., Brunato D, Dell'Orletta F., Sagae K., Sanchez-Gutierrez C.H. and Venturi G. (2020). Tracking the Evolution of Written Language Competence in L2 Spanish Learners. In: Proceedings of 15th BEA Workshop (BEA 2020), 10 July 2020.

[7] Nguyen, D., Doğruöz, A. S., Rosé, C. P., and de Jong, F. (2016). Computational sociolinguistics: a survey. In: Computational Linguistics, 42:537-593.

[8] Nivre, J. (2015). Towards a universal grammar for natural language processing. In: International Conference on Intelligent Text Processing and Computational Linguistics, pages 3-16.

[9] van Halteren, H. (2004). Linguistic profiling for author recognition and verification. In: Proceedings of the ACL, pages 200-207.

Chiara Branchini, Lara Mantovan

Grammatica della lingua dei segni italiana e digitalizzazione: una nuova frontiera per una più ampia accessibilità

Il nostro lavoro intende presentare l'applicazione di strumenti digitali alla realizzazione della più ampia grammatica descrittiva della lingua dei segni italiana (LIS), la lingua utilizzata dalla comunità sorda segnante italiana. Per la loro natura visiva, le lingue dei segni rappresentano un settore emblematico di applicazione delle *digital humanities* la cui versatilità e interattività consentono una fruizione innovativa e completa delle loro grammatiche.
La realizzazione della grammatica della LIS nasce all'interno del progetto europeo SIGN-HUB (2016-2020, www.sign-hub.it) a cui hanno collaborato sette Paesi, tra cui l'Italia, con l'obiettivo di creare risorse digitali che consentano di preservare e diffondere il patrimonio linguistico, storico e culturale delle comunità europee di Sordi segnanti. Accanto alla grammatica digitale della LIS, sono state realizzate le grammatiche digitali di altre sei lingue dei segni: catalana, tedesca, spagnola, turca, olandese e francese. Tutte sono state sviluppate seguendo le linee guida raccolte nel manuale "SignGram Blueprint" [1]. Le *digital humanities* hanno consentito di compiere un processo non ancora reso possibile nei diversi Paesi coinvolti nel progetto, quello di divulgare in formato accessibile e interattivo le proprietà linguistiche di lingue che non possiedono una forma scritta, rispettando dunque la loro natura visiva. La mancanza di una forma scritta, accanto alla prolungata discriminazione e isolamento sociale dovuto in molti casi ad un mancato riconoscimento giuridico, ha reso le lingue dei segni da sempre vulnerabili e a rischio di estinzione.
Questo intervento intende: i) consentire alla comunità Sorda di appropriarsi della propria lingua restituendole la dignità e il riconoscimento a lungo sottratto, ii) dare la possibilità alla comunità udente non segnante di scoprire un patrimonio linguistico trasmesso in una diversa modalità e, infine, iii) consegnare ai ricercatori strumenti d'analisi per approfondire lo studio delle caratteristiche linguistiche comuni alle lingue vocali e specifiche della modalità segnica.
In particolare, la grammatica della LIS comprende l'equivalente di circa 800 pagine e copre i seguenti domini: contesto storico e sociale, fonologia, lessico, morfologia, sintassi e pragmatica. Oltre al testo scritto, la piattaforma digitale ospita 2367 esempi glossati, 712 immagini, 1541 video esemplificativi di produzioni linguistiche in LIS, collegamenti ipertestuali che permettono al lettore di muoversi tra parti diverse della grammatica e un glossario dei principali termini linguistici. Il linguaggio utilizzato è diretto e accessibile anche a lettori non provvisti di una preparazione linguistica specifica. Da un punto di vista metodologico, ci siamo avvalsi della consulenza linguistica di sette collaboratori Sordi segnanti nativi provenienti da diverse aree geografiche. Le produzioni elicitate e spontanee dei segnanti nativi sono state videoregistrate e analizzate attraverso il software di annotazione ELAN [2]. La grammatica gratuita e presto accessibile sulla piattaforma digitale del progetto SIGN-HUB (www.sign-hub.eu) rappresenta uno strumento innovativo e utile per molteplici ambiti e finalità, tra cui la promozione dell'identità culturale e linguistica della comunità Sorda, l'educazione linguistica di bambini e ragazzi Sordi, la formazione di interpreti, l'insegnamento della LIS a persone udenti e lo sviluppo di nuovi studi linguistici comparativi.

[1] QUER J., CECCHETTO C., DONATI C., GERACI C., KELEPIR M., PFAU R. & STEINBACH M.(eds.). (2017). *SignGram Blueprint: A guide to sign language grammar writing*. Berlin: De Gruyter.

[2] SLOETJES H. & WITTENBURG P. (2008). Annotation by category: ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC*.

Valentina De Iacovo, Elisa Di Nuovo, John Hajek

L'analisi degli errori in un treebank orale di parlanti italo-australiani attraverso l'annotazione IOB integrata in CoNLL-U

Lo studio dei fenomeni linguistici nei contesti migratori rappresenta da sempre un campo di ricerca fecondo per i suoi potenziali punti di sviluppo indagabili (Vedovelli, 2011). Seguendo questa linea di studio, ci siamo chiesti quale fosse lo stato della lingua delle comunità italo-australiane stabilitesi negli anni '50 in Australia. Si parla ancora dialetto? In che rapporto si trovano la lingua inglese e quella italiana? Per provare a rispondere a queste domande abbiamo predisposto un corpus orale composto da interviste semi-strutturate con l'obiettivo di documentare i principali tratti fonetici, morfologici, lessicali e sintattici di una comunità italo-australiana a Melbourne. Il corpus, accessibile online, è stato trascritto automaticamente con MAUS (Kisler *et al*. 2017) e rivisto manualmente con Praat (Boersma & Weenink, 2001) e adattato alle linee guida di trascrizione di C-ORAL-ROM (Cresti & Moneglia, 2005). Successivamente è stato creato un corpus parallelo normalizzato su due livelli: nel primo livello sono state eliminate le disfluenze tipiche dell'oralità (false partenze, autocorrezioni, parole interrotte); nel secondo invece si è ricreato un corpus equivalente ma grammaticalmente corretto. È quindi stato annotato automaticamente sia il corpus originale che i due livelli paralleli usando UDPipe (Straka & Straková, 2017). Nel file CoNLL-u del testo originale è stata aggiunta manualmente l'annotazione di *code-mixing* (tra dialetto, italiano e inglese) e delle disfluenze orali; questi due fenomeni sono stati trattati in formato IOB (Ramshaw & Marcus, 1995). La stessa annotazione è stata quindi effettuata per gli errori grammaticali, ma inserita nel file CoNLL-u del testo di primo livello. Dal confronto tra questi CoNLL-u siamo giunti a un'analisi quantitativa preliminare delle strutture sintattiche devianti (cfr. Kinder, 1994) considerando anche le possibili interazioni tra dialetto, italiano e inglese. In questa prima fase abbiamo riscontrato sia devianze strettamente grammaticali, sia fenomeni caratteristici dell'italiano colloquiale (Berruto, 2012) o regionale.

BERRUTO G. (2012), *Sociolinguistica dell'italiano contemporaneo*, Carocci, Roma.

BOERSMA P. & WEENINK D. (2001), Praat, a system for doing phonetics by computer, in *Glot International* 5(9/10): 341-345.

CRESTI E. & MONEGLIA M. (2005), *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*, Benjamins Publishing, Amsterdam-Philadelphia.

KINDER J. (1994), Il ricupero della sintassi nell'italiano della seconda generazione in Australia, in GIACALONE RAMAT A. & VEDOVELLI M. (a cura di), *Italiano lingua seconda / lingua straniera*, Bulzoni, Roma: 343-361.

KISLER T, REICHEL U. D. & SCHIEL F. (2017), Multilingual processing of speech via web services, in *Computer Speech & Language*, 45: 326-347.

RAMSHAW L. A. & MARCUS M. P. (1995), Text chunking using transformation-based learning, in *Proceedings of the Third ACL Workshop on Very Large Corpora*, Dublin, Ireland: 82-94.

RUBINO A. (2003), Prospettive di mantenimento linguistico: fase di vita e di comunità come fattori di variabilità tra gli italiani in Australia", in BERNINI B., MOLINELLI P. & VALENTINI A. (a cura di), *Ecologia linguistica*, Bulzoni, Roma: 309-329.

Flavia Sciolette, Emiliano Giovannetti

Un modello per domarli tutti: verso una rappresentazione del testo come esplicitazione di documento, lingua e contenuto

Il presente lavoro nasce sulla scia di progetti di *Digital Humanities* relativi alla traduzione e all'analisi del testo su lingue di scarsa attestazione e domini della conoscenza specializzati (tra cui: Progetto Traduzione Talmud Babilonese e *Dictionary of Old Occitan medico-botanical terminology*). Un problema tipico in questi ambiti è la difficoltà a reperire e utilizzare risorse adeguate, spesso frammentarie, scarsamente integrate tra loro e difficilmente sfruttabili per compiti di NLP.
L'obiettivo della presente ricerca è proporre un modello olistico di rappresentazione dell'informazione testuale. Il modello è concepito per integrare le diverse dimensioni dell'oggetto testo, qui considerato nella sua accezione di *diasistema*. Quest'ultimo si definisce come insieme di elementi ordinati in sistemi distinti e fortemente interconnessi, nei quali ogni elemento del sistema ha un effetto sul comportamento del diasistema stesso nella sua interezza. Nel modello si considerano come parti del diasistema i seguenti sistemi: *segnico*, *linguistico*, *documentale*, *discorsivo*, *concettuale*. Ognuno dei sistemi si divide in *dimensioni*; in questa sede si intende la dimensione come un determinato spazio del sistema entro il quale si situano elementi omogenei che divengono unità di analisi. La nozione di dimensione non presuppone un ordinamento gerarchico. I legami interdimensionali si rappresentano secondo la metafora delle interfacce (es. sintassi-semantica).
L'orientamento teorico è affine alle attuali tendenze di ricerca, tanto in ambito testologico (Orlandi 2010) che computazionale: con l'avvento del *semantic web* e dei *linked data,* appare infatti ormai consolidato un paradigma fondato sull'associazione dell'informazione di varia natura, anche legata ai testi e alle lingue (es. ELEXIS, per la creazione di dizionari, che integra il modello OntoLex-Lemon con la codifica dei testi secondo lo standard TEI). Il modello di cui presentiamo le prime specifiche vuole rappresentare un 'ritorno al testo', inteso come un oggetto i cui dati devono essere esplicitati, resi interoperabili e comprensibili tanto da parte dell'interprete umano che da strumenti di analisi del testo e della lingua.
I primi studi condotti su testi tecnici (es. referti medici, fig. 1) hanno mostrato le potenzialità del modello per l'esplicitazione dell'informazione ivi contenuta, in maniera tale da consentire a un utente umano o a un agente artificiale ricerche complesse che combinino sistemi e dimensioni diverse, ad esempio per correlazione su due o più pattern formulari, eventi o dati quantitativi (elementi che incrociano informazioni di varia natura, dal sistema linguistico, discorsivo e concettuale). Saranno presentati casi analoghi su testi diversi, su tre possibili casi di uso: i) la traduzione del Talmud, ii) la creazione di un lessico di termini medico-botanici dell'antico occitano, iii) la creazione di un'ontologia da uno studio tematico su testi in antico francese.

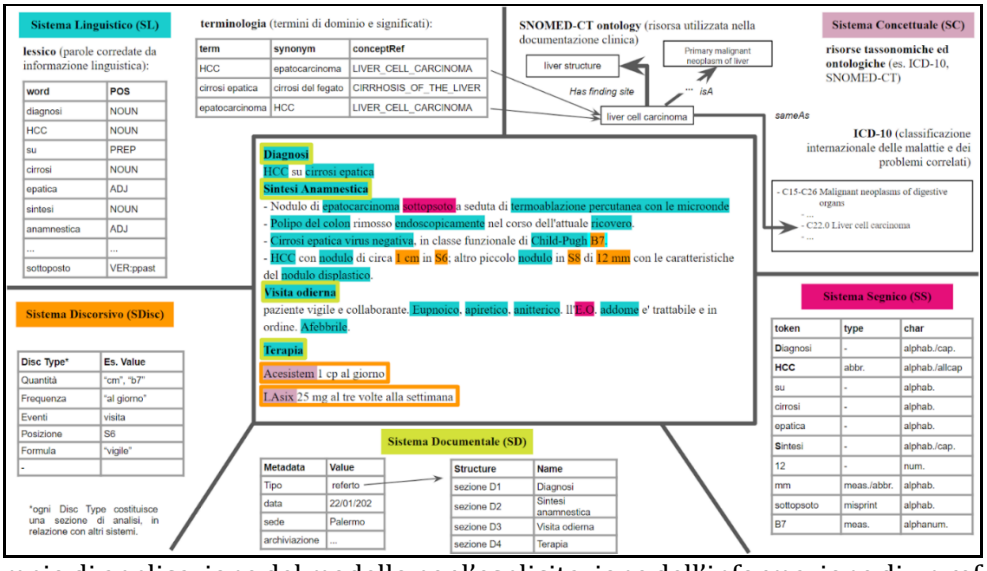

Fig. 1 Un esempio di applicazione del modello per l'esplicitazione dell'informazione di un referto medico.

Bellandi, A., Giovannetti, E., & Weingart, A. (2018), Multilingual and Multiword Phenomena in a lemon Old Occitan Medico-Botanical Lexicon In *Information*, 9(3), 52; doi:10.3390/info9030052.

Giovannetti, E., Albanesi, A., Bellandi, A., & Benotto, G. (2016), Traduco: A collaborative web-based CAT environment for the interpretation and translation of texts, in *Digital Scholarship in the Humanities*, doi:10.1093/llc/fqw054

McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P., Cimiano, P. (2017) *The OntoLex-Lemon Model: development and applications*, http://john.mccr.ae/papers/mccrae2017ontolex.pdf

McCrae, J.P., Tiberius, C., Khan, A. F., Kernerman, I., Declerck, T., Krek, S., Monachini, M. & Ahmadi, S. (2019), *The ELEXIS Interface for Interoperable Lexical Resources, Sixth Biennial Conference on Electronic Lexicography*, 10.5281/zenodo.3518959

Orlandi, T. (2010), *Informatica testuale. teoria e prassi*, Laterza, Roma.

Antonio Mastropaolo, Daniele P. Radicioni, Luisa Revelli

La lingua del legislatore in Valle D'Aosta: studio di caso e prospettive applicative

L'intervento intende presentare il paradigma teorico-metodologico predisposto per un progetto collocato in un'area interdisciplinare all'intersezione fra diritto, linguistica e informatica e centrato sulla produzione legislativa e regolamentare della Regione Valle d'Aosta dell'ultimo ventennio con l'obiettivo di definire un modello descrittivo della varietà di categorie testuali e delle loro relazioni con le soluzioni linguistiche adottate per veicolare contenuti semantico–informativi di matrice giuridica.
A fondamento del progetto si pone il concetto di *comprensibilità dei testi normativi*, che viene correlato alle criticità della sottoderminazione del linguaggio giuridico (Longo et al. 2012) e considerato come principio imprescindibile per la piena intellegibilità della *legge* da parte dei suoi destinatari finali: i cittadini. Nel caso di studio in esame il problema della *comprensibilità* deve trovare soluzioni rispetto a due specificità: lo *statuto di autonomia regionale* e il bilinguismo ufficiale da esso istituito.
A partire dalla costituzione di un corpus digitale di atti del dominio giuridico-amministrativo regionale tipologizzato per testi e generi (Damele et al. 2011) e attraverso il ricorso ad applicazioni per il *Trattamento Automatico del Linguaggio Naturale* disponibili e predisposte ad hoc, il progetto prevede che stili e canoni legislativi locali siano esaminati sui documenti in lingua italiana in riferimento all'efficacia comunicativa, testata attraverso l'applicazione di *indici di leggibilità*, la creazione di un *indice di salienza* dei documenti, basato sulla nozione di *densità semantica* (Rezaii et al. 2019) e sull'adozione di un set di *sense embedding* multilinguistici (Colla et al., 2020) e dalla parallela conduzione di interviste sul campo. L'aderenza dei testi del corpus ai principi generali della *scrittura controllata* e alle *linee-guida* concepite per i testi d'ambito giuridico-amministrativo a livello nazionale è verificata attraverso l'estrazione di un *lessico di frequenza* confrontabile con il *vocabolario di base della lingua italiana* oltre che attraverso annotazioni semantiche volte a tipizzare organizzazione strutturale, articolazione interna e categorie argomentative dei provvedimenti. Un ulteriore livello d'analisi prevede la comparazione delle corrispondenze fra i testi del corpus e la corrispettiva normativa di dimensione nazionale ed europea: in quest'ottica, una volta individuate caratteristiche e percentuali di similarità, sono investigati gli effetti di stratificazione dell'oscurità trasmessi – tanto a livello intralinguistico ed interlinguistico-traduttivo quanto da un punto di vista giuridico-strutturale – attraverso processi di mera riproduzione e vengono messe a fuoco, contemporaneamente e all'opposto, le soluzioni di distanziamento adottate per un miglioramento della qualità e comprensibilità dei testi legislativi sovraordinati.

COLLA D., MENSA E., & RADICIONI D. P. (2020), LESSLEX: Linking multilingual Embeddings to SenSe representations of Lexical items, in *Comput Linguist*, 46(2): 289-333.

DAMELE G., DOGLIANI M., MASTROPAOLO A., PALLANTE F. & RADICIONI D. P. (2011), On legal argumentation techniques: Towards a systematic approach, in BIASIOTTI M.A. & FARO S. (eds), *From Information to Knowledge–Online Access to Legal Information: Methodologies, Trends and Perspectives, Frontiers in Artificial Intelligence and Applications*, IOS Press, Amsterdam: 119-127 .

LONGO F., MASTROPAOLO A. & PALLANTE F. (2012), Incertezze derivanti dalla ineliminabile, ma non adeguatamente contenuta, oscurità linguistica delle disposizioni normative, in: DOGLIANI M., *Il libro delle leggi strapazzato e la sua manutenzione*, Giappichelli, Torino: 43 – 49.

REZAII N., WALKER E., & WOLFF P. (2019). A machine learning approach to predicting psychosis using semantic density and latent content analysis, in *NPJ schizophrenia*, 5(1): 1-12.

Giovina Angela del Rosso, Silvia Brambilla

L'accuratezza delle trascrizioni ASR sul parlato non-standard. L'italiano nell'OH Portal

1. Introduzione

Per interviste registrate in condizioni acustiche ottimali, di parlato standard e monologico, studi precedenti (Ashwel & Elam, 2017; Cosi, 2016) riportano prestazioni accurate nelle trascrizioni con ASR. L'OH Portal offre un nuovo supporto alle trascrizioni, benché rimanga da indagare la sua precisione con «non-standard speech, speech of elderly people or dialects, under-resourced languages» (Draxler *et al.*, 2020: 3358; cfr. Besacier *et al.*, 2014). Nel presente lavoro analizziamo le prestazioni dell'ASR di Google nel portale, su registrazioni di italiano non-standard, elicitate in condizioni variabili, al fine di quantificare e classificare gli errori, secondo un approccio *user-based*.

2. Materiali e metodo

L'indagine si basa su tre corpora:
- L2: 9 dialoghi di apprendenti L2 e nativi;
- BA: 11 dialoghi di parlanti di Bari;
- controllo: subcorpus ortofonico LP e LB del CLIPS[1].

Per valutare l'accuratezza dell'ASR, abbiamo confrontato le trascrizioni manuali (*reference* - REF) con quelle elaborate nel portale (*hypothesis* - HYP). Abbiamo categorizzato le coppie di parole REF-HYP, allineate automaticamente e corrette manualmente, nei tipi di (*mis*)*match* consolidati in letteratura (Palmerini & Savy, 2014): corrispondenza (OK), cancellazione (D), sostituzione (S) e inserzione (I). La correzione, al momento limitata a L2 e LP, ha prodotto un aumento delle coppie da 21.576 a 22.132 (+2,51%), di cui sono state annotate le prime 1100 per parlante (10.614 coppie totali) con parametri linguistici (es. parte del discorso, complessità fonologica) ed extralinguistici (es. ambiente, eventi acustici). Abbiamo infine calcolato il *word error rate* (WER) e la frequenza di occorrenza dei (*mis*)*match* in relazione a tali parametri.

3. Risultati preliminari

L'analisi parziale dei dati mette in luce tendenze nei *pattern* di errore riscontrabili dagli utenti, con variazioni sensibili nell'accuratezza dell'ASR. A livello quantitativo, l'accuratezza è confermata sul controllo (WER_{LP}: 2,5-4,5%; cfr. Draxler *et al.*, 2020); di contro, gli errori aumentano nei testi non-standard (WER_{L2}: 50,7-92,6%), benché sia identificabile una distribuzione regolare (D > OK > S > I). A livello qualitativo, in L2 il quadro dei (*mis*)*match* risulta complesso. Ad esempio, l'interazione tra parametri non rende conto di disallineamenti e *mismatch* multiparola nei dati, legati piuttosto a sostituzioni fonetiche e sintattiche, nonché all'effetto del cotesto.

ASHWELL, T. & ELAM, J. R. (2017), How accurately can the Google Web Speech API recognize and transcribe Japanese L2 English learners' oral production?, in *JALTCALLJournal*, 13(1): 59-76.

BESACIER, L., BARNARD, E., KARPOV, A. & SCHULTZ, T. (2014), Automatic speech recognition for under-resourced languages: A survey, in *Speech Communication*, 56: 85–100.

COSI, P. (2016), Phone Recognition Experiments on ArtiPhon with KALDI, in BASILE, P., CUTUGNO, F., MONTEMAGNI, S., NISSIM, M., PATTI, V. & SPRUGNOLI, R. (eds), *EVALITA*, Accademia University Press, Torino: 26-31.

DRAXLER, C., VAN DEN HEUVEL, H., VAN HESSEN, A., CALAMAI, S., CORTI, L. & SCAGLIOLA, S. (2020), A CLARIN Transcription Portal for Interview Data, in Calzolari, N., Béchet, F., Blache, P., Chouckri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. & Piperidis, S. (eds), *LREC 2020*, ELRA, Paris: 3353-3359.

[1] http://www.clips.unina.it/it/.

PALMERINI, M. & SAVY, R. (2014), Gli errori di un sistema di riconoscimento automatico del parlato. Analisi linguistica e primi risultati di una ricerca interdisciplinare, in BASILI, R., LENCI, A. & MAGNINI, B. (eds), *The First Italian Conference on Computational Linguistics*, Pisa University Press, Pisa: 281-285.

Nicola Brocca

LADDER: La costruzione e analisi di un corpus di scritture digitali per l'insegnamento della pragmatica in L2 attraverso le digital humanities

Molte recenti ricerche (Cortés Velásquez & Nuzzo 2017; Nuzzo & Cortés Velásquez, 2020; Artoni *et al.* 2020) hanno sottolineato come sia utile la creazione ed analisi di corpora per la didattica della pragmatica che, a differenza di altri livelli linguistici, come ad esempio la sintassi, non possono essere spiegate con regole ma con il riferimento a valori tendenziali o scelte più o meno adeguate in un certo contesto. Questo vale a maggior ragione per le interazioni attraverso media digitali, come mail e servizi di messaggistica, per il cui apprendimento i parlanti non nativi hanno pochi modelli di riferimento in quanto scritture private che trovano scarso spazio in manuali o corsi di L2 (Brocca, in stampa; Garofolin & Trubnikova, in preparazione). Studenti di un corso di didattica di It. L2 sono stati coinvolti nella creazione del Corpus LADDER (Brocca, 2020). Nella prima fase del corso gli studenti sono impegnati nella raccolta di un corpus di scritture digitali attraverso questionari online rivolti a apprendenti di italiano tedescofoni dei livelli A2-B2 e a parlanti nativi. I questionari vertono a elicitare attraverso *rolle play* autentici e *Discourse Completion Task* (Taguchi & Roever 2017: 85, 231) atti linguistici della richiesta (Bettoni 2006, Pagliara 2019) e della disdetta (Santoro *et al.* 2019) in livelli crescenti di distanza sociale e diversi media. Nella seconda fase il corpus viene analizzato dagli studenti secondo un'ottica pragmatica, annotato e i dati raccolti vengono elaborati secondo modelli statistici. Il corpus taggato sarà reso disponibile in formato Excel e XML *open access*. Il contributo mostrerà alcuni risultati delle analisi condotte con il corpus LADDER in ottica contrastiva L1/L2 durante i seminari di didattica. Oltre ad illustrare un modo di produzione di strumenti per la conoscenza dell'italiano L1 e L2 attraverso tecnologie digitali il contributo si profila come modello di didattica basata sulle *digital humanities*.

Artoni, D., Benigni, V. & Nuzzo, E. (2020), Pragmatic instruction in L2 Russian: A study on requests and advice, in *Instructed Second Language Acquisition*, 4.

Austin, J. L. (1976), *How to do things with words* (2. ed.): Oxford Univ. Press. Oxford.

Bettoni, C. (2006), *Usare un'altra lingua: guida alla pragmatica interculturale*: Laterza, Roma.

Brocca, N. (2020), LADDER. LeArners' Digital communication: a DatabasE for pRagmatical competences in L2, https://ladder.hypotheses.org/1.

Brocca, N. (in stampa), Insegnare italiano secondo l'approccio per task attraverso media digitali, in Hinzinger-Unterreiner E., (Hr.) *Aufgabenorientierung im Italienischunterricht. Ein theoretischer Einblick mit praktischen Beispielen,* Narr Francke Attempto Verlag, Tübingen.

Brown, P., & Levinson, S. C. (1987), *Politeness: some universals in language usage*. Cambridge University Press, Cambridge.

Cortéz Velásquez D. & Nuzzo E. (2017), Disdire un appuntamento: spunti per la didattica dell'Italiano L2 a partire da un corpus di parlanti nativi, in: *Italiano LinguaDue*, n1. 2017: 17-36.

Garofolin, B. & Trubnikova, V. (in preparazione). Pragmatica linguistica in corsi di lingua italiano L2.

Nuzzo, E. & Cortés Velásquez, D. (2020), Canceling Last Minute in Italian and Colombian Spanish: A Cross-Cultural Account of Pragmalinguistic Strategies, in *Corpus Pragmatics*: 1-26.

Santoro, E., Cortés Velásquez, D. & Nuzzo, E. (2019), Tra consapevolezza pragmatica e intercomprensione: uno studio esplorativo sull'atto linguistico della disdetta con italiani e brasiliani, in NUZZO E. & VEDDER *I.* (a cura di) *Lingua in contesto. La prospettiva pragmatica,* AItLA, Milano, 2019: 169-184.

Taguchi, N. & Röver, C. (2017), *Second language pragmatics*, Oxford University Press, Oxford.

Pagliara, F. (2019). La codifica pragmalinguistica dell' atto di richiesta nelle e-mail degli studenti universitari italiani; in NUZZO E. & VEDDER *I.* (a cura di) *Lingua in contesto. La prospettiva pragmatica,* AItLA, Milano, 2019. 149-168.

Valeria Caruso, Roberta Presta

Approcci digitali alla lessicografia: sviluppo e valutazione di un dizionario per smartphone

Secondo Jones (2014) tra il 2004 e il 2008 le app per smartphone hanno contribuito all'intrusione massiccia dei dati digitali nella vita quotidiana, inaugurando una rivoluzione che ha investito anche la ricerca scientifica con la nascita delle *Digital Humanities*. A dispetto di ciò, i software per dispositivi mobili rappresentano un ambito d'interesse ancora piuttosto marginale per diversi ambiti scientifici. Se si pensa ad esempio ai dizionari si osserva come questi generalmente migrino dalle versioni desktop agli smartphone senza adeguamenti significativi.

Per fornire supporto agli utenti in diverse situazioni comunicative, come la classe di lingue o l'interazione quotidiana, è stata sviluppata *Idiomatica*, prototipo di un dizionario per smartphone delle espressioni idiomatiche italiane indirizzato a parlanti non nativi. La complessità strutturale dei fraseologismi in genere, così come la loro centralità nel parlato spontaneo è stata ampiamente sottolineata in letteratura assieme alle difficoltà di apprendimento che essi pongono agli stranieri (Siyanova-Chanturia & Spina, 2019 per una sintesi su tutti questi aspetti). In particolare, per imparare gli idiom, Nation (2006) raccomanda esplicitamente l'uso di dizionari.

Idiomatica (fig. 1) ha un layout progettato secondo principi lessicografici e linee guida ergonomiche dello *human-centred design* (Giacomin, 2014; ISO 9241:210). La sua microstruttura composita sostituisce gli 'scrollabili' (fig. 2) comunemente in uso per gli smartphone fornendo le informazioni principali nella parte alta dello schermo (significato, un esempio d'uso e eventuali varianti formali, fig. 1), mentre le altre vengono visualizzate su schermate collegate cliccando sulle etichette metalinguistiche di una 'table view' (Human Interface Guidelines). Ulteriori indicatori suggeriscono quali informazioni usare per i compiti di comprensione ("Informazioni per la comprensione") e di produzione ("Informazioni per la produzione").

Dopo aver illustrato l'approccio partecipativo usato per progettare il dizionario (Caruso et al. 2019), si discuteranno le caratteristiche del prototipo rispetto alla sua usabilità, efficacia e 'acceptance' (Hornbæk & Hertzum, 2017) da parte degli utenti. Vengono in tal senso presentati i risultati di uno studio di valutazione formativa (Lazar *et al.*, 2017) con 10 apprendenti cinesi (di livello B1) impegnati da remoto nello svolgimento di cinque task su due espressioni idiomatiche diverse.

Analizzando i risultati degli esercizi e le videoregistrazioni delle attività svolte sono emersi i punti di forza e le debolezze del prototipo realizzato. Gli utenti apprezzano il dizionario per la sua ricchezza informativa, l'accessibilità microstrutturale e l'efficacia descrittiva di alcuni campi informativi, come l'illustrazione della corrispondenza tra la morfo-sintassi e la semantica delle locuzioni basata sui *Frame* di Fillmore. Ugualmente apprezzate sono le indicazioni sui tipi di atti linguistici che le singole locuzioni tipicamente realizzano; mentre le informazioni sul registro non sono state comprese dai partecipanti e le tavole flessive non sembrano sufficienti per usare correttamente le locuzioni più complesse, come quelle provviste di doppio clitico (ad es. 'fasciarsi la testa prima di rompersela').

Elisa Gugliotta, Angelapia Massaro, Giuliano Mion, Marco Dinarelli

Un'analisi statistica della definitezza in un corpus di arabish tunisino

L'arabo tunisino, come di norma avviene nelle varietà neoarabe, realizza il tratto di indefinitezza con modalità non marcata, ossia nomi nudi (Mion, 2009). Allo stesso tempo, è noto che un nome nudo può anche essere interpretato come definito (per esempio teste nello stato costrutto in semitico, o nomi propri non articolati in italiano). Ne consegue che l'assenza della marca di definitezza (DET) non necessariamente implica l'assenza del tratto di definitezza (Def) e che quest'ultimo e la categoria DET non sono esattamente sovrapponibili. Inoltre in un numero di lingue il nome articolato è requisito della struttura argomentale del VP anche in contesti non definiti (Longobardi, 2008), (Raposo, 1998), ([IP *mi[VP piace[DP mangiare [DP[D ø[NP torta]]]]]]). Questo lavoro si propone di verificare, analizzando i nomi nudi presenti nel *Tunisian Arabish Corpus* (TArC) (Gugliotta & Dinarelli, 2020), in che misura essi possano essere considerati definiti. Il corpus utilizzato contiene testi tunisini codificati in *arabish*, sistema tipico della *Computer Mediated Communication* araba, basato su caratteri latini e cifre. Il TArC si compone di vari livelli di annotazione del testo, tra cui *tokenizzazione* e *Part-of-Speech tagging* (PoS), secondo le convenzioni del *Penn Arabic Treebank* (PATB) (Maamouri, 2009). Trattandosi di convenzioni ideate per l'arabo standard, queste includono il *core-tag CASE* e le relative *features* per i casi e le marche di definitezza o indefinitezza: *NOM*, *ACC*, *GEN*, *DEF* e *INDEF*. Considerando l'assenza di marche per la funzione nominale in tunisino (*ʾiʿrab*), per l'annotazione in PoS del TArC, il core-tag CASE e le sue features sono stati rimossi, riducendo l'informazione Def alla presenza o assenza del core-tag DET, 'determinante' per l'articolo **al-*. Questa operazione ha imposto alcune riflessioni riguardo alle modalità di *PoS-tagging* del tratto Def. Per analizzare il tratto Def nel TArC sono quindi state selezionate le sue prime 15.000 parole. Da queste sono state estratte automaticamente tutte le parole contenenti il tag: *NOUN*. Una volta separati i nomi propri, i nomi testa di uno stato costrutto o i nomi che presentavano DET (o *POSS_PRON,* pronomi possessivi) da quelli che ne erano privi, questi ultimi sono stati analizzati e annotati manualmente una seconda volta al fine di verificare l'assenza di Def. Le analisi condotte finora sui dati mostrano che meno dell'80% di questi appare effettivamente indefinito, mentre per il restante 20% circa vanno effettuati ulteriori test al fine di stabilire se l'assenza di DET deriva da: (1) assimilazione fonetica del determinante, (2) appartenenza del nome a una categoria di 'elementi non contabili', come 'liquidi', 'polveri' o 'gas'. Inoltre, la nostra ricerca si occupa di indagare la presenza di quantificatori per i quali potrebbe essere in corso un processo di grammaticalizzazione come marca di indefinitezza (Mion, 2009).

Gugliotta, E., Dinarelli, M., (2020), Tarc: Incrementally and semi-automatically collecting a Tunisian Arabish Corpus. In Proceedings of the *12th Language Resources and Evaluation Conference (LREC 2020),* Marseille, France.

Maamouri, M., et al. (2009), *Penn Arabic treebank guidelines*. Linguistic Data Consortium.

Mion, G., (2009), L'indétermination nominale dans les dialectes arabes. Une vue d'ensemble*. Miscellanea arabica 2009*, 14, 215.

Longobardi, G. (2008), 'Reference to individuals, person, and the variety of mapping parameters'. In Henrik Høeg Müller e Alex Klinge (ed), *Essays on nominal determination: From morphology to discourse management*, 189, 211.

Raposo, E. (1998, July), Definite/zero alternations in Portuguese. In *Romance Linguistics: Theoretical Perspectives. Selected papers from the 27th Linguistic Symposium on Romance Languages (LSRL XXVII),* Irvine, 20 22 February, 1997 (Vol. 160, p. 197). John Benjamins Publishing.

Elisa Di Nuovo, Cristina Bosco, Elisa Corino

Analisi sintattica e dell'errore in un corpus di apprendenti di italiano: il sintagma nominale

Individuare specifici fenomeni linguistici in un Corpus di Apprendenti (CA) è un'operazione resa più ardua dalle caratteristiche intrinseche alle varietà di apprendimento. Considerando che tali fenomeni possono essere pertinenti a diversi livelli d'analisi, la sola annotazione per parti del discorso (*PoS tagging*) sovente applicata ai CA (CAIL2, LOCCLI, COLI, VALICO[2]) non garantisce la possibilità di coglierli e studiarli in modo affidabile. Un significativo miglioramento in termini di precisione di estrazione e analisi si potrebbe ottenere da CA con la struttura di *treebank* che annotino esplicitamente anche le relazioni sintattiche (Aarts & Granger, 1998).
Questo contributo presenta l'annotazione in formato Universal Dependencies[3] (lo standard *de facto* per l'annotazione sintattica) di un sottocorpus di VALICO (Corino & Marello, 2017), attualmente composto di trentasei testi, nove per ognuna delle quattro L1 degli apprendenti selezionate (francese, inglese, spagnolo e tedesco) e suddivisi in base all'anno di studio dell'italiano (tre per prima, seconda e terza annualità), in cui a ogni frase (*Learner Sentence*, LS) è stata abbinata una sua versione corretta (*Target Hypothesis*, TH) e una codifica degli errori, già introdotta in un primo studio e ispirata ai *tag set* descritti in Dagneaux *et al.* (1996) e Nicholls (2003).
Come caso di studio abbiamo esaminato i fenomeni devianti nei sintagmi nominali tramite un approccio computazionale, sul modello di Felice *et al.* (2016), che, grazie al confronto delle LS con le TH, consente di ricavare automaticamente sia errori di accordo (Salvi, 1991), compresi i casi di sovraestensione (1), sia errori interni al sintagma nominale non collegati all'accordo (2) e (3).

(1) LS: Un uomo molto *forto*, [...] aveva preso *sulla sua* spalle [...]
TH: Un uomo molto *forte*, [...] aveva preso *sulle sue* spalle [...]
(2) LS: […] una donna è andata al parco con *il suo* marito […]
TH: […] una donna è andata al parco con *suo* marito […]
(3) LS: […] come se volesse liberarsi da *quel* uomo *in palestra*.
TH: […] come se volesse liberarsi da *quell'*uomo *palestrato*.

Oltre a facilitare e velocizzare l'individuazione delle aree problematiche, l'approccio computazionale descritto basato sull'uso dei testi paralleli è orientato alla preparazione del terreno necessario per lo sviluppo di sistemi automatici di correzione dell'errore (Davidson *et al.*, 2020).

AARTS J. & GRANGER S. (1998), Tag sequences in learner corpora: a key to interlanguage grammar and discourse, in Granger S. (a cura di), *Learner English on Computer*, Routledge, London: 132-141.

CORINO E. & MARELLO C. (2017), *Italiano di stranieri. I corpora VALICO e VINCA*, Guerra, Perugia.

DAVIDSON S., YAMADA A., FERNÁNDEZ-MIRA P., CARANDO A., SÁNCHEZ GUTIÉRREZ C. & SAGAE K. (2020), Developing NLP Tools with a New Corpus of Learner Spanish, in *Proceedings of the 12th Conference on Language Resources and Evaluation*: 7238-7243.

DAGNEAUX E., DENNESS S., GRANGER S. & MEUNIER F. (1996), *Error Tagging Manual Version 1.1*, Université Catholique de Louvain: Centre for English Corpus Linguistics, Louvain-la-Neuve.

FELICE M., BRYANT C. & BRISCOE T. (2016), Automatic Extraction of Learner Errors in ESL Sentences Using Linguistically Enhanced Alignments, in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*: 825-835.

NICHOLLS D. (2003), The Cambridge Learner Corpus: error coding and analysis for lexicography and ELT, in *Proceedings of the Corpus Linguistics 2003 conference*: 572-581.

SALVI G. (1991), L'accordo, in *Grande grammatica italiana di consultazione*, a cura di RENZI L., SALVI G. & CARDINALETTI A., 3 voll., vol. 2°, il Mulino, Bologna: 227-244.

[2] CAIL2, LOCCLI e COLI sono disponibili qui: https://www.unistrapg.it/cqpwebnew/; VALICO qui: http://www.valico.org.

[3] Sito del progetto: https://universaldependencies.org.

Francesca Pagliara

Validare schemi di codifica in ambito linguistico: l'Intercoder Agreement applicato alla tassonomia del Cross-Cultural Speech Act Realization Patterns Project (CCSARP)

La riproducibilità dei risultati di uno studio è un criterio di qualità fondamentale negli studi qualitativi di matrice umanistica che si servono di sistemi di annotazione dei dati. In linguistica, l'annotazione di dati presuppone l'impiego di categorie d'analisi talvolta interpretative e implica la formulazione di giudizi soggettivi. La necessità di stabilire fino a che punto tali giudizi siano affidabili e riproducibili ha assunto crescente importanza nelle discipline sociali e umanistiche, ivi compresa la linguistica, fino a rendere le procedure di validazione una prassi consolidata (Gagliardi, 2018). Nel settore degli studi linguistici questo è avvenuto soprattutto nel ramo della linguistica computazionale, poiché, per addestrare un sistema automatico a svolgere compiti di annotazione, è indispensabile che le tassonomie e le categorie di analisi siano risorse validate. È meno comune, invece, che test di misurazione dell'accordo tra i codificatori siano eseguiti in altri filoni della ricerca linguistica, sebbene siano numerosi gli studi che si avvalgono dell'uso di schemi di codifica e di categorizzazioni per il trattamento dei dati (Kukartz & Radiker, 2019). In letteratura è tuttavia appurato che un alto livello di accordo tra diversi codificatori è indice dell'affidabilità e della riproducibilità di un paradigma di annotazione (Di Eugenio, 2000; Warrens, 2010). Diversi studi (Kukartz & Rädiker, 2019; Gagliardi, 2018) quindi sostengono che, quando si fa una codificazione di dati qualitativi, è opportuno condurre una procedura di convalidazione dello schema di codifica. Un indice statistico che permette di calcolare l'accordo tra codificatori è l'*Intecoder Agreement* (Cohen, 1960).
Il presente contributo ha lo scopo di illustrare la procedura seguita per validare la codifica dell'atto di richiesta in 378 e-mail di studenti universitari PN e PNN di italiano, utilizzando la nota tassonomia del *Cross-Cultural Speech Act Realization Patterns Project* (CCSARP - Blum-Kulka *et. al.*, 1989), ampiamente utilizzato negli studi cross-culturali e interlinguistici, ma ancora scarsamente sottoposto a validazione (Miura, 2019).

Blum-Kulka S., House J. & Kasper G. (1989), *Cross-cultural pragmatics: Requests and apologies*, Norwood, New Jersey: Ablex.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Di Eugenio, B. (2000). On the usage of Kappa to evaluate agreement on coding tasks. In: Calzolari N. *et al.* (eds): P*roceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, ELRA - European Language Resources Association, Paris, pp. 441–444.

Gagliardi G. 2018, Inter-Annotator Agreement in linguistica: una rassegna critica, in: E. Cabrio, A. Mazzei, F. Tamburini (a cura di), *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it,* December 2018, Torino, AILC edizioni, Trento.

Kukartz S. & Radiker S. (2019), Analyzing Qualitative Data with MAXQDA, Springer, Switzerland AG.
Miura, A. (2019). *Corpus Pragmatics: Exploring criterial pragmalinguisticfeatures of requestivespeech acts produced by Japanese learners of English at different proficiency levels*. (Unpublished doctoral dissertation). Tokyo University of Foreign Studies, Tokyo.

Warrens, M. J. (2010), "Inequalities between multirater kappas", *Advances in Data Analysis and Classification*, 4(4):271-286.

Alessandro Puglisi

Dentro cascate di testo: una proposta per l'analisi del pivoting in un canale Twitch

Gli ambiti e i metodi ricompresi nelle *digital humanities* sono aumentati esponenzialmente negli ultimi anni, e già da tempo il dibattito in merito è vivace (Gold, 2012).
Nel contesto della comunicazione multimediale e transmediale, con specifico riferimento alla pratica del *live-streaming*, *Twitch* rappresenta ormai una realtà consolidata (Hamilton *et al.*, 2014) e numerosi studi hanno indagato differenti aspetti della piattaforma, dai modi per incoraggiare o prevenire determinati comportamenti all'interno di essa (Seering *et al.*, 2017) all'*information overload* (Nematzadeh *et al.*, 2019) fino ad una forma di interazione su larga scala che è stata definita "crowdspeak" (Ford *et al.*, 2017).
Il contributo che proponiamo, relativo a una ricerca in corso, prende le mosse dall'idea di "participatory spectatorship" (Georgen, 2014) a più riprese proposta per definire il ruolo assunto dagli utenti-spettatori nell'*online streaming*, oltre che dalle acquisizioni sul rapporto fra *streaming* e *frames* goffmaniani (Karhulahti, 2016). La nostra ricerca guarda, sulla base della metodologia di trascrizione e analisi proposta da Recktenwald (2017), all'attività di *online streaming* su *Twitch* dello *streamer* italiano Xiuder. Si registreranno e trascriveranno, per sette giorni consecutivi, estratti di *stream* che vedono impegnato lo *streamer* in Fortnite, noto videogioco. In questo modo si potrà guardare ai complessi rapporti che si instaurano fra tre elementi cardine: la comunicazione (verbale orale e non verbale) dello *streamer*, la comunicazione scritta condotta dagli utenti in chat e gli eventi di gioco. Particolare attenzione sarà riservata al fenomeno del *pivoting*, come produzione frasale o comunicazione "embodied" prodotta da eventi di gioco e che ad essi, circolarmente, (ri)attribuisce senso.
La nostra proposta si nutre del crescente interesse scientifico per le forme di comunicazione mediata dal computer (CMC) in diretta e, in questo senso, la ricerca intende indagare un fenomeno complesso attraverso strumenti propri tanto della linguistica quanto dell'analisi dei media.

FORD C., GARDNER D., HORGAN L. E., LIU C., TSAASAN A. M., NARDI B. & RICKMAN J. (2017), Chat Speed OP PogChamp: Practices of Coherence in Massive Twitch Chat, in *CHI EA '17: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*: 858-871. DOI: https://dl.acm.org/doi/10.1145/3027063.3052765
GEORGEN C. (2014), Well played and well watched: dota 2, spectatorship and esports, in *Well Played 4(1)*: 179-191.
GOLD M. K. (ed.) (2012), *Debates in the Digital Humanities*, University of Minnesota Press, Minneapolis.
HAMILTON W. A., GARRETSON O. & KERNE A. (2014), Streaming on Twitch: Fostering Participatory Communities of Play within Live Mixed Media, in *CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*: 1315-1324. DOI: https://dl.acm.org/doi/10.1145/2556288.2557048
KARHULAHTI V.-M. (2016), Prank, troll, gross and gore: performance issues in esport live-streaming, in *Proceedings of 1st International Joint Conference of DiGRA and FDG*.
NEMATZADEH A., CIAMPAGLIA G. L., AHN Y-Y. & FLAMMINI A. (2019), Information overload in group communication: from conversation to cacophony in the Twitch chat, in *Royal Society Open Science* 6: 191412. DOI: http://dx.doi.org/10.1098/rsos.191412
RECKTENWALD D. (2017), Towards a transcription and analysis of live streaming on *Twitch*, in *Journal of Pragmatics* 115: 68-81. DOI: http://dx.doi.org/10.1016/j.pragma.2017.01.013
SEERING J., KRAUT R. & DABBISH L. (2017), Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting, in *CSCW '17: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*: 111-125. DOI: https://dl.acm.org/doi/10.1145/2998181.2998277

Simone Ciccolone, Giulia Isabella Grosso

Macchine parlanti o macchine apprendenti? Prime riflessioni sulle interlingue simulate dei chatbot

Un campo d'indagine ancora da esplorare, legato a nuovi contesti di comunicazione digitale, riguarda l'interazione con interfacce conversazionali come assistenti vocali o *chatbot*. La presenza di questi nuovi strumenti permette all'utente umano di interagire con la macchina tramite gli stessi canali usati per la comunicazione tra esseri umani.
Lo sviluppo di sistemi di intelligenza artificiale applicati alla conversazione (le cosiddette *conversational AIs*) si sta gradualmente avvicinando al suo obiettivo finale di simulare un dialogo naturale. A un esame più attento, tuttavia, si può notare come tali sistemi siano spesso pilotati o condizionati dalle competenze linguistiche dei loro autori, che introducono risposte precostruite più complesse di quelle abitualmente generate dal software per mascherare eventuali carenze del sistema; in alcuni casi viene usata una connotazione linguistica della "personalità" del *chatbot* per giustificarne l'incapacità di rispondere in modo pertinente all'utente. Un esempio lampante è il *chatbot* Eugene Goostman, rappresentato come un adolescente ucraino apprendente di inglese, che anche in virtù di tale caratterizzazione ha potuto in più occasioni superare il test di Turing (cfr. Shah *et al.*, 2016).
In modo più o meno esplicito a seconda dei casi, alcune interfacce conversazionali simulano una sorta di "interlingua" (in senso lato) con esiti tuttavia ben diversi rispetto a quella di un apprendente umano. Si pensi ad es. al *Winograd Schema* (cfr. Levesque, 2011) per testare la disambiguazione della coreferenza, problematica anche per le più avanzate AI conversazionali, o alla difficoltà di individuare il *topic* nell'enunciato, di contro alla maggiore competenza pragmatica degli apprendenti umani anche nelle prime fasi dell'acquisizione.
Molte delle competenze qui citate dipendono dalla capacità degli interlocutori di decodificare (o, nella visione di alcuni studiosi interazionisti, di co-costruire) il contesto nell'ambito dell'interazione. Sia che questo si intenda (nell'accezione linguistica) come l'insieme degli elementi linguistici di un testo "e i rapporti che li legano l'uno con l'altro così da essere pienamente significativi solo se presi nel loro complesso" (Pallotti, 2014: 121), sia che si intenda come una co-costruzione di senso che avviene all'interno dell'interazione, è centrale nel concetto stesso di contesto la possibilità, per gli interlocutori, di accedere a un quadro di conoscenze più ampio che permetta una reale comprensione dell'evento comunicativo. In quest'ottica risulta cruciale poter indagare il modo in cui i tratti contestuali possano influire sui comportamenti linguistici osservati.
Scopo di questo contributo è riflettere sulle anomalie e patologie coesive e di progressione tematica degli interventi di un'interfaccia conversazionale nell'interazione con utenti umani, e confrontarne le (presunte) strategie di negoziazione dei significati con quelle prodotte da apprendenti di una L2. L'analisi si concentrerà su alcuni casi specifici emersi in interazioni con *chatbot*, proposti a titolo esemplificativo e come primo campione di fenomeni, comparandoli con quelli provenienti da corpora di italiano parlato da stranieri.

Levesque H. (2011), The Winograd Schema Challenge, *AAAI Spring Symposium* (technical report).

Pallotti G. (2014), Studiare i contesti di apprendimento linguistico: modelli teorici e principi metodologici, in De Meo, A. *et al.* (a cura di), *Varietà dei contesti di apprendimento linguistico*, AItLA, Milano: 121-132.

Shah H., Warwick K., Vallverdú J. & Wu D. (2016), Can machines talk? Comparison of Eliza with modern dialogue systems, in *Computers in Human Behavior*, 58: 278-295.