GIOVINA ANGELA DEL ROSSO[1] - SILVIA BRAMBILLA[1]

L'accuratezza della trascrizione ASR sul parlato non-standard. L'italiano nell'OH Portal

Abstract

This paper aims at evaluating the transcription accuracy of the Google-IT's ASR service available in the OH Portal. Data on its performance are limited to good quality recordings. Thus, we focus on suboptimal authentic materials, encompassing non-standard conversational speech recorded in noisy environments. We carry out a quanti-qualitative analysis of the linguistics and extra-linguistic parameters affecting accuracy and error distribution. The preliminary results show higher Word Error Rates for non-standard speech and low-quality recordings. Moreover, we seek the error patterns that could ease the transcription correction process for the users.

1. *Introduzione*

In questo contributo[2] presentiamo i risultati preliminari di un lavoro in corso incentrato sulle trascrizioni automatiche di materiali di parlato, uno degli strumenti chiave delle *digital humanities*. Il processo di trascrizione manuale dei materiali di parlato costituisce un passaggio laborioso ma imprescindibile nell'elaborazione delle registrazioni, sebbene spesso precluda l'analisi di grandi quantità di dati. Una nuova risorsa che risponde all'esigenza di ottimizzare la catena di trascrizione è rappresentata dall'OH Portal. Si tratta di un'interfaccia *web* integrata nell'infrastruttura europea CLARIN (clarin.eu), che raccoglie una serie di strumenti per la gestione semi-automatica dei materiali di parlato (Draxler et al. 2020; Scagliola et al. 2020; van den Heuvel 2020). Al suo interno sono disponibili vari servizi, che includono strumenti di terze parti per la trascrizione tramite il riconoscimento automatico del parlato (*Automatic Speech Recognition*, ASR). Tra questi, per il nostro studio, abbiamo usato l'ASR di Google, già riconosciuto come uno degli ASR commerciali più accurati (Ashwell & Elam 2017; Biadsy et al. 2012; Filippidou & Moussiades 2020; Tavosanis 2018). Le valutazioni dell'accuratezza degli ASR implementati nell'OH Portal, incluso quello di Google, sono prevalentemente basate su materiali monologici, di varietà linguistiche tendenti al polo dello standard,

[1] Università degli Studi Roma Tre-Sapienza Università di Roma.

[2] Sebbene il presente contributo sia frutto di un costante lavoro comune, la stesura dei paragrafi 1, 2, 4.1, 5.2.1, 5.2.4 e 6 è da attribuirsi a Giovina Angela del Rosso, mentre i paragrafi 3, 4.2, 5.1, 5.2.2, 5.2.3 sono da attribuirsi a Silvia Brambilla.

registrati in buone condizioni acustiche. Spesso, però, i materiali di parlato spontaneo comprendono interviste e dialoghi, attestano varietà non-standard e non sono registrati in ambienti ideali. Notoriamente, le esitazioni, false partenze, parole interrotte e sovrapposizioni di turno, fenomeni largamente presenti nel parlato spontaneo e dialogico, rendono il riconoscimento automatico complesso e possono contribuire a far diminuire l'accuratezza della trascrizione (Badino 2016; Çetin & Shriberg 2006; Kitaoka et al. 2014; Tavosanis 2018). Pertanto, basandoci su materiali precedentemente raccolti per altre finalità e che rappresentano tali caratteristiche (§ 2), il nostro obiettivo è indagare a livello quanti-qualitativo l'accuratezza della trascrizione automatica e quali fattori la influenzino, adottando una prospettiva basata sull'utente e orientata all'utente.

Un limite intrinseco di questa analisi, tuttavia, deriva dall'oggetto stesso della ricerca. L'o*utput* che osserviamo è il prodotto di un sistema di ASR commerciale, il che comporta che non sono disponibili informazioni dettagliate sul funzionamento e sui materiali di addestramento dell'ASR di Google (Draxler et al. 2020). Ciononostante, tale limite può essere arginato in virtù di due considerazioni. In primo luogo, possiamo avanzare ipotesi inferenziali, fondate sul funzionamento dei riconoscitori automatici di ultima generazione (Abulimiti & Schultz 2020; Këpuska & Bohouta 2017; Kitaoka et al. 2014). In secondo luogo, in un'ottica incentrata sugli utenti, l'esame linguistico dell'*output* è di maggiore utilità rispetto all'analisi delle motivazioni che lo hanno prodotto. L'analisi linguistica, infatti, costituisce una fase euristica necessaria per lo sviluppo di strategie di correzione manuale *ad hoc*, funzionali all'ottimizzazione della catena di trascrizione (semi-)automatica.

L'articolo è organizzato come segue: in § 2 descriviamo il dataset; in § 3 ripercorriamo la metodologia adottata; in § 4 discutiamo i principali risultati finora ottenuti, che riguardano l'accuratezza della trascrizione (§ 4.1) e la distribuzione degli errori (§ 4.2); in § 5 illustriamo e discutiamo tramite esempi casi di errore rilevanti per gli utenti; infine, in § 6 presentiamo le conclusioni preliminari del lavoro e i possibili sviluppi futuri.

2. *Materiali*

Il dataset analizzato è costituito dai corpora descritti nella tab. 1. I corpora BA e L2, precedentemente raccolti per altri scopi, rappresentano varietà di italiano non-standard, rispettivamente di italiano regionale di Bari nativo e L2. Tali corpora sono stati confrontati tra di loro e con i subcorpora LP e LB del CLIPS (Albano Leoni et al. 2007). Questi ultimi fungono da controllo, in quanto rappresentativi dell'italiano standard.

Tabella 1 - *Corpora analizzati*

Corpus L2	*Corpus BA*	*Corpus CLIPS*	
9 registrazioni	10 registrazioni	10 registrazioni	10 registrazioni
6 parlanti avanzati L2 (2 F e 4 M) 3 parlanti nativi (2 F e 1 M)	10 parlanti nativi (5 F e 5 M)	10 parlanti professionisti (5 F e 5 M)	
Dialogo	Dialogo	Monologo	
Parlato spontaneo	Parlato semi-spontaneo (dialoghi semi-guidati)	Parlato letto ortofonico (subcorpus LP)	Parlato letto ortofonico (subcorpus LB)
Italiano regionale di Bari	Italiano regionale di Bari	Italiano standard	
Ambienti chiusi ma rumorosi	Ambienti chiusi e silenziosi, ma non isolati	Camera anecoica (Falcone et al. 2007b)	
Computer portatile con microfono esterno, verso gli intervistati	Registratore Tascam DR-40 con microfoni incorporati, verso gli intervistati	Apparecchiatura professionale (Falcone et al. 2007a)	
44.1 kHz, 16-bit, .wav	44.1 kHz, 24-bit, .wav	22.1 kHz, .wav	

3. *Metodo*

Abbiamo trascritto manualmente le registrazioni dei corpora BA e L2 che, insieme alle trascrizioni del corpus CLIPS, costituiscono i nostri testi di riferimento (*reference texts,* REF). Tutte le registrazioni sono poi state elaborate nell'OH Portal con il sistema ASR di Google-IT, i cui risultati costituiscono i nostri testi di ipotesi (*hypothesis texts*, HYP). I testi REF e HYP sono stati allineati usando un algoritmo[3] di programmazione dinamica (Sakoe & Chiba 1978) che consente di calcolare la distanza di Levensthein tra stringhe (Levenshtein 1966). Le coppie di parole sono state così classificate come corrispondenze (OK) o errori, distinti nei tre tipi consolidati in letteratura (Bechét & Favre 2013; Levis & Suvorov 2012; Palmerini & Savy 2014), cioè cancellazioni (DEL), sostituzioni (SUB) o inserzioni (INS). Questa classificazione è funzionale al calcolo del tasso di errore, per il quale la metrica adottata è stata il WER (*Word Error Rate*):

$$WER = ERR/N$$

in cui ERR è la somma del numero degli errori (INS + SUB + DEL), che viene rapporta a N, cioè il numero totale delle coppie di parole considerate (OK + INS + SUB + DEL). Data la sua diffusione in letteratura, abbiamo scelto il WER come in-

[3] Codice Python *case-insensitive* adattato da https://holianh.github.io/portfolio/Cach-tinh-WER/.

dice di accuratezza per una maggiore comparabilità dei dati, nonostante non manchino proposte di metriche alternative (Filippidou & Moussiades 2020). Le uscite di tale allineamento saranno d'ora in poi chiamate REF-0/HYP-0. In totale, tale processo ha individuato 57248 coppie di parole, automaticamente annotate anche per tipo di errore.

Successivamente, abbiamo riallineato manualmente tali coppie. Ciò ha prodotto un secondo allineamento, d'ora in avanti chiamato REF-1/HYP-1. Abbiamo conseguentemente corretto l'annotazione dei tipi di errore. Nella tab. 2 riportiamo un esempio dei diversi allineamenti, con le relative classificazioni degli errori. Si può notare che in REF-0/HYP-0 mancano le corrispondenze tra *barese-varese* e *guidato-di lato*; quest'ultima è anche un caso di *sostituzione multiparola* (§ 5.2.2). A seguito di questo tipo di correzioni è stato riconteggiato un totale di 57389 coppie.

Tabella 2 - *Confronto tra allineamento automatico e manuale*

automatico			*manuale*		
ERR-0	REF-0	HYP-0	ERR-1	REF-1	HYP-1
DEL	sono	***	DEL	sono	***
OK	una	una	OK	una	una
DEL	barese	***	SUB	barese	varese
DEL	acquisita	***	SUB	acquisita	visita
DEL	tuttavia	***	DEL	tuttavia	***
SUB	non	varese	DEL	non	***
SUB	ho	visita	DEL	ho	***
SUB	mai	di	DEL	mai	***
SUB	guidato	lato	SUB	guidato	di
SUB	qui	si	SUB	SUB	lato
			SUB	qui	si

Per ragioni di chiarezza, aggiungiamo un asterisco (*) all'etichetta del dataset, quando ci riferiamo alle prime 1100 coppie per registrazione di REF-0/HYP-0 o di REF-1/HYP-1 (o alla totalità delle coppie disponibili, se meno di 1100). Le 35040 coppie totali di REF-1*/HYP-1* sono state annotate secondo una serie parametri linguistici ed extralinguistici, il cui effetto sull'accuratezza dei sistemi ASR è stato precedentemente evidenziato in letteratura (Badino 2016; Çetin & Shriberg 2006; Palmerini & Savy 2014; Tavosanis 2018)[4]:

a. parametri linguistici: parte del discorso (PoS, *Part of Speech*) di REF-1* e HYP-1*, normalizzazione o errore ortografico, complessità fonologica, lingua di REF-1* e HYP-1*, ripetizioni di *chunk* in REF-1*, segnale discorsivo;

[4] Per l'annotazione delle parti del discorso abbiamo adottato una versione semplificata del *tagset* ISST-TANL (http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf.); per gli eventi acustici abbiamo seguito le linee guida del CLIPS (Savy 2007); per la distinzione tra segnali discorsivi e disfluenze, cfr. Crible (2016).

b. parametri extralinguistici: turno, numero di partecipanti, distanza del parlante dal microfono, ambiente, evento acustico.

In un'ottica orientata agli utenti, in questa prima fase della ricerca, tuttora in corso, abbiamo lavorato principalmente sull'interpretazione di PoS e turno. Ad esempio, come mostreremo in § 5, sembra che alcune classi di parole siano maggiormente soggette a errore rispetto ad altre.

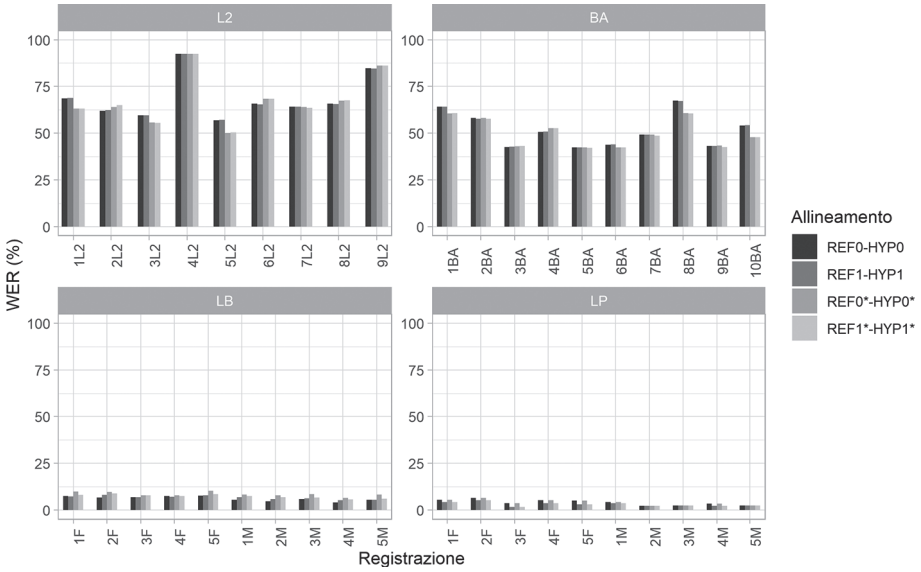
Infine, abbiamo calcolato il WER per registrazione di REF-0/HYP-0, REF-1/HYP-1, REF-0*/HYP-0*, REF-1*/HYP-1*. Per REF-1*/HYP-1* abbiamo anche calcolato il WER rispetto al turno di parola. Nei prossimi paragrafi, mostriamo i risultati dell'analisi quanti-qualitativa condotta col supporto dell'ambiente statistico R (R Core Team 2021).

4. *Risultati*

In questa sezione mostriamo una panoramica complessiva dei risultati preliminari di questo lavoro, che riguardano l'accuratezza delle trascrizioni automatiche (§ 4.1) e l'analisi della tipologia degli errori (§ 4.2). Considerata la natura del dataset, abbiamo utilizzato le singole registrazioni come unità di comparazione, senza aggregare i dati per corpus, anche se impieghiamo le etichette dei corpora per richiamarne tutte le registrazioni.

4.1 Il tasso di errore

Figura 1- *Comparazione dei WER per registrazione, per corpus e per metodo di calcolo*



Dall'analisi finora condotta, come mostrato nella fig. 1, a livello quantitativo si riscontra uniformità tra i tassi di errore risultanti dalla reiterazione del calcolo del

WER di REF-0/HYP-0, REF-1/HYP-1, REF-0*/HYP-0* e REF-1*/HYP-1*. Si veda ad esempio la registrazione 1L2, in cui i WER relativi al dataset annotato sono coincidenti (63.2%), discostandosi di pochi punti percentuali (Δ=5.7%) dai WER relativi alle trascrizioni integrali, a loro volta molto simili (Δ=0.2%).

Dal grafico si rileva, inoltre, sostanziale variazione dei WER sia inter-corpora che intra-corpus. Globalmente, il tasso d'errore delle trascrizioni automatiche è maggiore nei corpora di italiano non-standard rispetto ai corpora di italiano standard: L2 (50.0-92.5%) > BA (42.2-67.4%) > LB (4.0-10.3%) > LP (1.7-6.5%).

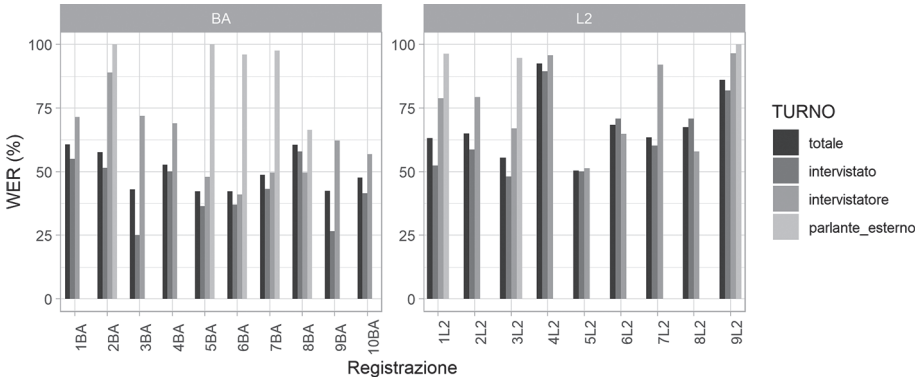
Alla limitata variazione dei WER, però, si contrappongono le differenti distribuzioni degli errori rilevati automaticamente e manualmente (§ 4.2). Considerando ciò, l'analisi delle interazioni tra errori e parametri linguistici ed extralinguistici sarà basata su REF-1*/HYP-1*.

Tra i parametri considerati, il tipo di corpus e il turno di parola influiscono significativamente[5] sul tasso di errore (fig. 2)[6]. Nei nostri dati, tali parametri sono strettamente collegati rispettivamente alla qualità del segnale e alla distanza del parlante dal microfono (cfr. tab. 1), corroborando i risultati già noti delle ricerche sull'argomento (cfr. Li et al. 2016). Inoltre, ciascun partecipante contribuisce alla definizione del tasso di errore globale in misura disuguale, proporzionale alla lunghezza del proprio turno. Da ciò segue che il WER di REF-1*/HYP-1* coincide con la media ponderata dei WER dell'intervistato, dell'intervistatore e degli eventuali parlanti esterni. Nel nostro dataset, il WER dei parlanti esterni influisce limitatamente sul WER totale, sebbene possa raggiungere anche il 100% (come in 2BA, 5BA e 9L2), perché le coppie attribuite ai loro turni sono poche (ad esempio, in 2BA sono solo 6 su 1100). Tuttavia, tassi di errore così elevati sono motivati dalle condizioni di registrazione. Nello specifico, durante le interviste dei corpora BA e L2, gli eventuali parlanti esterni si trovavano in posizioni periferiche rispetto a intervistato e intervistatore e conseguentemente erano molto lontani dal microfono, direzionato sempre verso l'intervistato. A ciò si aggiunge la tendenza dei parlanti esterni a intervenire a bassa voce e a generare non di rado sovrapposizioni di turno o eventi acustici di varia natura.

[5] Dall'analisi preliminare dell'interazione e della gerarchia tra parametri, l'albero di inferenza condizionale (*conditional inference tree*) conferma che anche i nodi corrispondenti al tipo di corpus e al turno di parola esercitano un effetto statisticamente significativo sull'accuratezza della trascrizione. Tuttavia, sarà necessario approfondire tali complesse relazioni in uno studio dedicato, che esula parzialmente dagli scopi di questo lavoro, la cui natura è principalmente linguistica.

[6] Questa considerazione è stata recentemente integrata come *disclaimer* nell'OH Portal: "Known Issues for Google ASR: if the recording is longer than a few minutes we observed that whole stre[t]ches of words are simply omitted in the result; this happens frequently at speaker turns, i.e. when two or more speakers take turns, then often the begin of the turn of the new speaker is compromised." (https://clarin.phonetik.uni-muenchen.de/apps/TranscriptionPortal/, accesso effettuato in data 28/09/2021).

Figura 2 - *Comparazione dei WER per registrazione, per corpus e per turno di parola*



4.2 Distribuzione degli errori

Anche se l'algoritmo ha permesso la classificazione automatica degli errori, la successiva verifica dei risultati ha evidenziato la necessità di correggerla manualmente. Le classificazioni automatiche e manuali divergono qualitativamente sia nell'allineamento sia per la conseguente rietichettatura degli errori (cfr. tab. 2, § 3). Nella tab. 3 sono riportati a titolo esemplificativo i WER minimi e massimi divisi per corpus e per allineamento. Nel caso del corpus LP, l'annotazione è stata esaustiva, quindi REF-0/HYP-0 e REF-1/HYP-1 coincidono rispettivamente con REF-0*/HYP-0* e REF-1*/HYP-1*.

Tabella 3 - *WER minimi e massimi per corpus e allineamento*

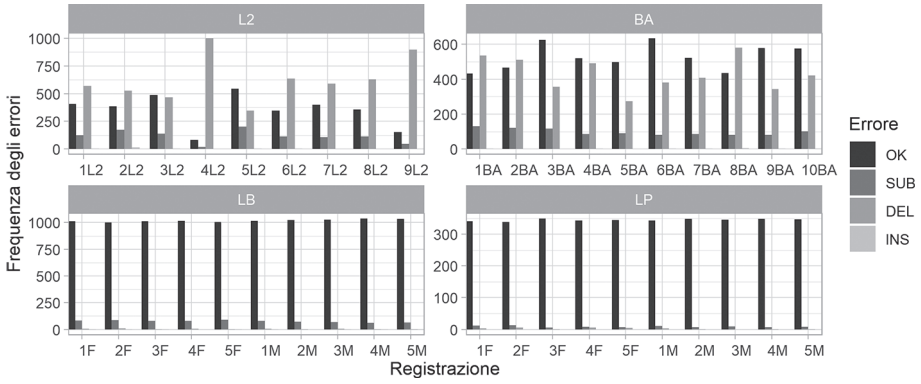
Corpus	*LP*		*LB*		*BA*		*L2*	
Registrazione	2F	3F	4M	5F	5BA	8BA	5L2	4L2
REF-0/HYP-0	3.7	6.5	4.0	7.6	42.3	67.4	56.8	92.4
REF-1/HYP-1	1.7	5.3	5.2	7.9	42.3	67.2	57.0	92.4
REF-0*/HYP-0*	–	–	6.4	10.3	42.3	60.7	50.0	92.5
REF-1*/HYP-1*	–	–	5.7	8.5	42.2	60.5	50.4	92.5

I risultati dell'analisi sono rappresentati nella fig. 3[7]: la distribuzione degli errori nei corpora non è uniforme né da un punto di vista quantitativo né qualitativo. Dai diagrammi a barre, risulta evidente che nei subcorpora del CLIPS prevalgono le corrispondenze (OK) rispetto alle sostituzioni (SUB) e alle cancellazioni (DEL). Diversamente, nei corpora BA e L2 (di italiano non-standard) le cancellazioni incidono maggiormente sul tasso di errore, come rilevabile dal grafico, a differenza delle inserzioni (INS), quasi del tutto assenti. Tuttavia, BA e L2 differiscono per la distribuzione interna di cancellazioni e sostituzioni: al peggiorare delle condizioni

[7] Per ragioni grafiche, la scala dell'asse y non è uniforme.

acustiche, aumenta il numero di cancellazioni, mentre diminuisce il numero di sostituzioni.

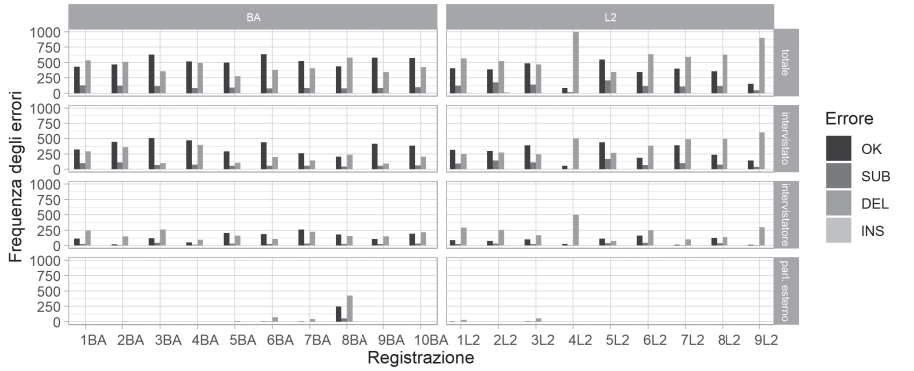
Figura 3 - *Distribuzione degli errori per registrazione e per corpus*



Una tendenza analoga è riscontrabile anche all'aumentare della distanza del parlante dal microfono, fattore rappresentato nel nostro dataset dal parametro "turno di parola". Dato che tale effetto è osservabile solo su materiali non monologici, l'analisi è pertinente solo per BA e L2 (fig. 4). Nei diagrammi a barre riferiti a "parlante esterno" e a "intervistatore", il numero di cancellazioni è maggiore che per l'"intervistato", per la maggiore vicinanza di quest'ultimo al microfono. Quindi, il deteriorarsi delle condizioni acustiche comporta una ridotta variazione dei tipi di errore, come osservabile anche nella fig. 4.

Di conseguenza, gli utenti dell'OH Portal che optino per l'ASR di Google-IT dovranno aspettarsi di ripristinare manualmente nella trascrizione soprattutto gli enunciati prodotti dai parlanti più distanti dal microfono, perché tendenzialmente l'ASR non ne trascrive correttamente o addirittura ne cancella del tutto i turni.

Figura 4 - *Distribuzione degli errori per registrazione e per turno di parola in BA e L2*



5. *Gli errori per gli utenti*

In questa sezione consideriamo l'interazione tra gli errori e i parametri linguistici, anche in funzione del dominio. In particolare, ci soffermiamo sulle inserzioni (§ 5.1) e sulle sostituzioni (§ 5.2) di maggior rilievo per gli utenti per l'uso delle trascrizioni automatiche. Infatti, se nei casi di cancellazione è sufficiente che gli utenti aggiungano le parti omesse dall'ASR, l'individuazione e la successiva correzione delle inserzioni e delle sostituzioni risulta più complessa.

5.1 Inserzioni

In letteratura le inserzioni sono solitamente considerate come errori dovuti alla reinterpretazione linguistica di rumore spurio (Baber & Hone 1993). Nel nostro dataset, le inserzioni sono gli errori più rari. A livello quantitativo le inserzioni individuate tramite l'algoritmo di allineamento sono 153 su 35081 coppie totali di REF-0*/HYP-0* (0.4%). Dopo il riallineamento manuale, il numero delle inserzioni è stato ridotto a 65, con una riduzione delle inserzioni allo 0.2% del totale. Infatti, abbiamo ricondotto la maggior parte delle inserzioni rilevate automaticamente a sostituzioni multiparola (§ 5.2.2).

Le parole inserite sono frequenti nel lessico italiano, perlopiù foneticamente deboli e brevi, quindi spesso appartenenti a classi funzionali più che lessicali (Bybee 2001; Maturi 2014). Questa indicazione di massima può guidare l'utente durante la correzione e coadiuvare l'individuazione di elementi estranei alla traccia audio ma presenti nella trascrizione automatica.

È da notare che, diversamente da quanto atteso, le inserzioni sembrano essere attribuite più all'influsso della probabilità linguistica rispetto a quello della probabilità acustica. Infatti, le inserzioni ricorrono anche in assenza di rumore ambientale e, in generale, di eventi acustici cui potrebbero essere attribuite: sono presenti anche nel subcorpus di controllo LB, benché questo sia stato registrato in camera anecoica.

Dal punto di vista linguistico, possiamo ricondurre gli *output* testuali interessati da inserzioni a tre macro-fenomeni, distinti in base al grado di standardizzazione dei nostri materiali e di grammaticalità dei testi HYP. In particolare, i fenomeni individuati sono:

1. standardizzazione di forme di varietà regionali e di apprendimento: in (1)[8] nella HYP viene inserito il determinante *i* assente nella REF perché omesso dal parlante non nativo. Per l'utente, l'inserzione di *i* nella sequenza comporta una standardizzazione non fedele alla varietà testimoniata dal materiale originale;

(1)	deve	lavorare	per	darle	INS	soldi
	devi	lavorare	per	darle	i	soldi

[8] Negli esempi, dove non specificato diversamente, la prima riga è il testo REF, la seconda è il testo HYP.

2. presenza nell'HYP di strutture più frequenti di quelle della REF, benché entrambe le varianti siano grammaticali: in (2) il determinante *la* precede il sintagma *settimana scorsa*, perché il costrutto *la settimana scorsa* è più probabile rispetto alla variante priva di articolo;

(2)	m'	ha	detto	una	volta	INS	settimana	scorsa	poi
	ma	che	tu	una	volta	la	settimana	scorsa	poi

3. creazione di *chunk* localmente grammaticali ma contestualmente agrammaticali: in (3) la presenza del verbo *andare* innesca l'inserzione della preposizione *a*, che segue frequentemente questo lemma. Tuttavia, se si considera un contesto più ampio del bigramma, la sequenza **andare a prima al mercato* non è grammaticale.

(3)	per	fare	la	spesa	andiamo	INS	prima	al	mercato
	DEL	DEL	DEL	DEL	andiamo	a	prima	al	mercato

5.2 Sostituzioni

5.2.1 Sostituzioni per fattori linguistici e di dominio

L'accuratezza degli ASR è sensibile non solo alle condizioni acustico-ambientali ma anche a fattori linguistici e di dominio (Besacier et al. 2014; Draxler et al. 2020). Per gli utenti, l'(inter-)azione di tali fattori è più evidente nelle sostituzioni.

Abbiamo riscontrato che, nei nostri dati, alcune parole possono essere trascritte con grafie errate. Queste sostituzioni sembrano essere dovute alle procedure di post-elaborazione ortografica applicate all'uscita dell'ASR. Tali errori influiscono largamente sull'accuratezza, perché si riscontrano in parole ad alta e altissima frequenza, come articoli e preposizioni.

In questa casistica rientrano i numerali, riconosciuti correttamente nel 61.5% delle occorrenze e trascritti con alterazioni ortografiche nel 15.6% dei casi. Tra gli errori ortografici che coinvolgono i numerali, sono sistematici quelli di numeri contenenti *dieci* (4a), mentre sono occasionalmente errati altri cardinali (4b) e i numeri pronunciati in sequenza (4c)[9]:

(4a)	dieci, diecimila	>	dici, dici-comma-zero-zero-zero
(4b)	tre (*)	>	iii
(4c)	il sette il settantuno (*)	>	il settecentosettantuno

Un esempio analogo riguarda l'apostrofo, che viene sempre trascritto nel verbo *c'è*, ma sistematicamente omesso in *l'*, *un'* e *po'*. Più complesso è il caso delle preposizioni articolate, in cui alla corretta trascrizione si alternano casi di omissione. Sebbene le grafie errate siano frequenti, gli errori di questo tipo sono facilmente emendabili.

Altrettanto articolato è il caso dell'onomastica. Da un lato, abbiamo individuato sostituzioni *di* nomi propri (5a-b), presumibilmente candidati poco probabili per

[9] Il simbolo (*) affianca i *token* che non sono assenti nel vocabolario (*out of vocabulary*), perché trascritti correttamente altrove.

l'*output*: ad es., l'antroponimo straniero *Ahmad* viene trascritto come *Amazon* (5a). Dall'altro, segnaliamo le sostituzioni *con* nomi propri (5c-d), anche poco diffusi nelle conoscenze enciclopediche degli italofoni ma ben indicizzati su Google Search: ad es., *inverno* viene sostituito dal cognome del giornalista Luciano *Onder* (5c).

(5a) Ahmad > Amazon
(5b) Jamil (*) > già mi
(5c) inverno (*) > Onder
(5d) m'ha detto (*) > Maletto

L'ipotesi del collegamento tra l'ASR e gli altri prodotti Google spiegherebbe anche la maggiore accuratezza di trascrizione dei toponimi rispetto a quella degli antroponimi (tab. 4). Tale differenza si apprezza maggiormente in BA, in cui i toponimi sono trascritti correttamente in circa 1 caso su 2, mentre gli antroponimi solo in 1 su 5.

Tabella 4 - *Distribuzione degli errori nell'onomastica in BA e L2*

	BA				L2			
	OK	SUB	DEL	TOT.	OK	SUB	DEL	TOT.
toponimi	50.2%	8.2%	41.6%	100%	31.1%	8.6%	60.3%	100%
antroponimi	19.0%	14.3%	66.7%	100%	26.7%	13.3%	60.0%	100%

Anche le forme non-standard nel dataset vengono sostituite e ricondotte a forme standard, cui sembrano essere attratte per similarità fonetica (6a-c); ciò comporta, talvolta, un casuale mantenimento del significato (6b).

(6a) er a mett > ero metà
(6b) pad > papà
(6c) cè > cioè

La similarità fonetica spiegherebbe anche le saltuarie sostituzioni di *token*, in presenza di realizzazioni ipoarticolate (Lindblom 1990; per l'italiano, cfr. Albano Leoni & Maturi 1995). La confusione tra suoni può avvenire all'interno di parola (7a) o limitarsi alle vocali atone finali (7b).

(7a) parco (*) > pacco
(7b) famosa (*) > famoso

Infine, anche le disfluenze sono oggetto di errori linguistici (80.5% DEL, 8.1% SUB). Le sostituzioni riguardano false partenze e parole troncate, sempre trascritte con parole intere e foneticamente plausibili, che possono ripristinare (8a) o alterare (8b) il senso originale.

(8a) quartie+[quartiere] > quartiere
(8b) picco+ [piccoli] > picconi

Anche l'individuazione di questi tipi di *pattern* di errore potrebbe supportare gli utenti nello sviluppo delle strategie di correzione.

5.2.2 Le sostituzioni multiparola

L'algoritmo di allineamento considera possibili solo corrispondenze uno-a-uno, non uno-a-molti, in cui l'unità è la parola ortografica delimitata da spazi bianchi. Tuttavia, durante la correzione manuale, considerando il cotesto, abbiamo individuato sostituzioni che coinvolgono più parole, dette *sostituzioni multiparola*. Complessivamente, costituiscono il 9.8% delle sostituzioni.

Le sostituzioni multiparola ricorrono in due situazioni: nella prima, abbiamo allineato una o più parole della REF a due o più parole dell'HYP, come in (9).

(9) Quartierino > quartiere Reno

Nella seconda, in modo speculare, abbiamo allineato due o più parole della REF a una o più parole dell'HYP, come in (10).

(10) ti dovresti > tirolese

In questi casi, alcuni errori, che nell'allineamento automatico erano rispettivamente inserzioni e cancellazioni, sono stati riannotati come sostituzioni. Nel calcolo del WER il numero di errori legato a questa rianalisi è rimasto invariato, dato che un tipo di errore è stato corretto in un tipo diverso: in (9), abbiamo contato due sostituzioni invece che una sostituzione e un'inserzione, mentre in (10) abbiamo contato due sostituzioni invece che una sostituzione e una cancellazione. La soluzione da noi adottata rispetta la tassonomia classica degli errori degli ASR e il suo vantaggio risiede in una migliore comparabilità dei risultati. Tuttavia, restano da indagare gli eventuali benefici di metriche alternative (Besacier et al. 2014; Filippidou & Moussiades 2020), che attribuiscano un peso diverso alle sostituzioni multiparola, il cui *status* sembrerebbe non essere ancora stato considerato esplicitamente in letteratura.

Infine, l'incidenza delle sostituzioni multiparola sugli errori totali si apprezza più nei subcorpora di controllo (12.6% in LP, 8.0% in LB) che nei corpora di italiano non-standard (1.7% in BA, 1.6% in L2). Questo è presumibilmente legato alla maggiore variabilità dei tipi di errore al migliorare delle condizioni acustiche (§ 4.2). L'analisi di questi aspetti e della loro interazione con la similarità fonetica necessita di ulteriori approfondimenti futuri.

5.2.3 Le parti del discorso

Nonostante il ruolo dei modelli acustici nei moderni ASR sia preponderante rispetto a quello dei modelli linguistici, per gli utenti può essere un utile strumento di correzione individuare quali parti del discorso siano maggiormente soggette a sostituzione.

Tabella 5 - *Matrice di confusione delle sostituzioni rispetto alla parte del discorso (PoS)*

SUB		PoS_HYP																
		V	com_N	art_det	agg	prep_art	art_ind	prep	cong	avv	aus	pro_cli	mod	N_pro	num	pro	V_cli	int
PoS_REF	V	385	40	2	18	2	1	16	38	24	2	4	3	18	2	12	3	5
	com_N	47	246	1	26	1	2	15	7	15	2	7	3	38	6	12	1	0
	art_det	6	0	175	7	23	2	7	8	5	1	4	0	0	0	5	0	0
	agg	12	32	3	106	1	1	5	4	10	3	0	0	10	0	13	1	0
	prep_art	0	6	25	0	68	0	11	2	0	1	0	0	0	0	1	0	0
	art_ind	0	0	8	0	0	54	3	1	2	0	0	0	0	1	1	0	0
	prep	9	1	11	3	16	0	49	5	7	2	3	0	1	1	4	0	0
	cong	31	7	0	2	6	0	10	48	15	8	10	5	2	0	3	0	1
	avv	18	13	3	7	4	1	12	23	41	2	14	2	5	0	10	0	0
	aus	4	0	1	0	0	0	0	4	6	41	4	1	0	0	1	0	1
	pro_cli	4	1	2	1	0	0	9	5	4	1	34	0	2	0	2	0	0
	mod	6	6	0	1	0	0	1	3	4	0	0	26	0	0	0	0	0
	N_pro	7	20	1	2	2	0	1	5	9	0	0	0	22	0	3	1	0
	num	6	1	0	3	0	0	0	1	1	0	0	0	1	20	1	0	0
	pro	7	1	0	6	0	0	4	9	6	1	8	1	2	0	19	0	0
	V_cli	10	1	0	0	0	0	0	0	1	0	1	0	0	0	0	4	0
	int	3	3	0	0	0	0	3	36	13	0	3	1	0	0	2	0	1

La matrice di confusione dei casi di sostituzione divisi per parte del discorso (tab. 5) permette di verificarne la consistenza numerica. In particolare, le previsioni (la trascrizione manuale, PoS_REF) sono riportate nelle righe, mentre lo stato effettivo (la trascrizione automatica, PoS_HYP) è riportato nelle colonne. Nella diagonale sono evidenziati i casi in cui, malgrado la sostituzione, è mantenuta la classe di parola originaria. Ad esempio, in 246 casi i nomi comuni della REF sono sostituiti da altri nomi comuni nell'HYP (veri positivi), come esemplificato in (11):

(11a) soccorritori > seguitori
(11b) farmacia (*) > suoneria

Diversamente, 26 nomi comuni della REF sono sostituiti da aggettivi (falsi negativi), come in (12):

(12a) entrata (*) > dentale
(12b) pancia (*) > panica

Viceversa, nell'HYP 32 aggettivi sono sostituiti da nomi comuni (falsi positivi), come in (13).

(13a) equo > eco
(13b) pulito > prurito

Alcune di queste sostituzioni sono intuibili, come quelle frequenti tra la congiunzione *e* e la forma verbale *è*, mentre altri errori di sostituzione esibiscono *pattern* meno regolari. In termini assoluti, la distribuzione più eterogenea riguarda verbi e nomi, che sono anche le parti del discorso più frequenti nel nostro dataset. Un'analisi divisa per classe, tuttavia, restituisce un quadro parzialmente diverso circa l'accuratezza di trascrizione rispetto alla parte del discorso. Una sintesi è fornita nella tab. 6, in

cui la confusione per classe cresce da sinistra verso destra. Per ciascuna PoS, sono riportati i valori di richiamo (*recall*), precisione (*precision*) e F1-score. Nello specifico, si nota un maggiore mantenimento di classe per gli articoli ($F1_{artind}$=82%, $F1_{artdet}$=73%), nonostante siano foneticamente deboli, rispetto a classi di parole più lunghe e piene, come aggettivi e avverbi ($F1_{agg}$=55%, $F1_{avv}$=25%). A livello complessivo, la classe di parola è mantenuta in circa metà delle sostituzioni, come indicato sia dal valore di macro-F1(47%), in cui ciascuna classe ha uguale peso, sia dal valore di micro-F1 (54%), più influenzato dal numero di esemplari per classe.

Tabella 6 - *Accuratezza percentuale di trascrizione per parte del discorso*

metrica		PoS																
		art_ ind	art_ det	V	aus	num	N_ com	mod	prep_ art	agg	pro_ cli	prep	V_ cli	cong	avv	N_ pro	pro	int
per-classe	richiamo	88	75	69	64	68	65	62	55	58	36	33	40	24	25	22	22	12
	precisione	77	72	67	65	60	57	55	60	53	52	44	23	32	26	30	30	1
	F1-score	82	73	68	64	64	61	58	57	55	43	38	29	27	25	25	25	2
media	micro-F1	54																
	macro-F1	47																

5.2.4 I verbi in LB

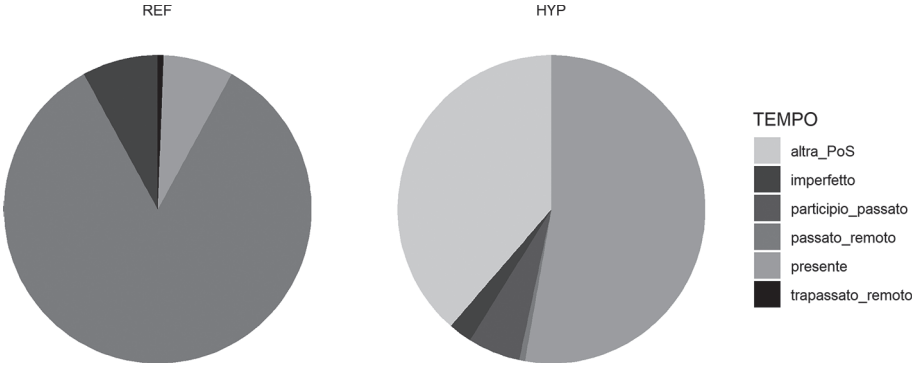
Il subcorpus LB è stato scelto, insieme a LP, come controllo, perché rappresentativo dell'italiano standard e di condizioni acustiche ottimali. Come atteso, il tasso di errore per questi materiali è basso (WER_{LP}=1.5-5.3%; WER_{LB}=5.7-8.8%). Eppure, l'accuratezza di trascrizione per LB è significativamente inferiore rispetto a quella di LP ($\chi^2 = 111.85$, $p < 2.2^{-16}$), sebbene i due subcorpora siano stati registrati nelle stesse condizioni: in camera anecoica, con la medesima strumentazione, con la stessa varietà (italiano standard), con gli stessi parlanti, con lo stesso *task*. Ciò mette in luce il fatto che il riconoscimento può fallire anche in condizioni acustiche ottimali, per forme linguistiche standard, prodotte con un eloquio iperarticolato da parlatori professionisti (cfr. tab. 1). Si può quindi supporre che le cause degli errori per questi subcorpora siano da ricercare nel dominio linguistico e nel modello del linguaggio implementato nell'ASR: alcune forme sono meno frequenti di altre. Approfondire la natura degli errori presenti in LB è di particolare interesse non tanto per la differenza quantitativa, spiegata dal numero maggiore di DEL, ma per le caratteristiche di questo subcorpus. Infatti, LB comprende 120 stimoli, mentre LP ne comprende 20 (cfr. Falcone et al. 2007a, allegato C); il maggior numero di stimoli dà più spazio a strutture linguistiche rare e complesse, come esemplificato nella tab. 7. Quindi ci siamo concentrate sull'analisi degli errori di sostituzione dei verbi in LB, che costituiscono il 23.6% del totale (179 su 758 sostituzioni) e permettono di esemplificare una maggiore varietà di fenomeni.

Tabella 7 - *Confronto tra le frasi di LP e LB*

LP	*LB*
Un mese di vacanza passa in fretta.	Stimammo che il banco del Perù aveva un debito pari a mezzo milione di dollari.
Nel grande parco un bambino giocava con suo padre.	Il frate bevve un blando orzo molto caldo e si scottò il labbro superiore.
Luisa fingeva di guardare da un'altra parte, ma cercava il modo per farsi notare.	Inghiottì la capsula e dopo un paio di secondi cadde sul posto stecchito.

Perciò, abbiamo annotato i verbi della REF di LB per tempo, modo, persona e lemma. Specularmente, abbiamo annotato i corrispondenti *token* dell'HYP: (a) secondo gli stessi parametri, se verbi; (b) con l'etichetta *altra_PoS*, se sostituiti da parole di classi diverse. I risultati sono mostrati nella fig. 5. In particolare, il 67.3% dei verbi della REF è al passato remoto indicativo e tende a essere sostituito nell'HYP da verbi al presente indicativo; in metà dei casi (50.7%) si tratta di forme dello stesso lemma. Ciò sembra riconducibile alla minore frequenza del passato remoto rispetto al presente indicativo nell'italiano contemporaneo (D'Achille 2019).

Figura 5 - *Confronto REF-HYP dei verbi sostituiti in LB*



Dalla fig. 5, inoltre, si può notare che, nelle sostituzioni di verbi in LB, è frequente anche il cambio di parte del discorso (37.9%), talvolta concomitante anche con sostituzioni multiparola (11.2%). Due sostituzioni tipiche di verbi in LB sono esemplificate in (14), dove *bevve* e *scottò* sono ricondotti rispettivamente al presente *beve* e alla preposizione *sotto*.

(14) il frate bevve un blando orzo molto caldo e si scottò il labbro superiore
il frate beve un blando orzo molto caldo e si sotto il labbro superior

Se questo tipo di distribuzioni non viene controbilanciato in modo mirato nei materiali di addestramento o nel modello dei sistemi di ASR, come *output* viene preferito il candidato più frequente rispetto a quello meno frequente, soprattutto in concomitanza di una forte similarità fonetica (Song et al. 2021). Come esemplificato nella tab. 8, tale preferenza può essere sistematica: infatti, se in alcuni casi l'esito

delle sostituzioni è variabile, in altri casi le sostituzioni sono fisse e riguardano tutte le occorrenze di un dato verbo nel dataset (10 su 10).

Tabella 8 - *Verbi in LB sostituiti in tutte le occorrenze*

SUB fisse		*SUB variabili*	
REF	HYP	REF	HYP
bevve	beve	cadde	cade (9) che (1)
chiamai	chiama	convertì	convertiti (7) convertire (2) convertini (1)
colpì	colpi	scottò	sotto (6) scotta (3) Scott (1)
dormii	dormi	stimammo	sti mammo (5) stimiamo (2) sì mamma (1) ti mammo (1) sì ma mo (1)

6. *Conclusioni preliminari*

L'analisi preliminare conferma il divario tra materiali ottimali e non ottimali per quanto riguarda l'accuratezza della trascrizione automatica. I risultati sono in linea con quanto ipotizzato in letteratura (Draxler et al. 2020; Scagliola et al. 2020; Tavosanis 2018). Rimane però da indagare una serie di questioni. Innanzitutto, la definizione di standard e non-standard sul piano della pronuncia e degli ASR potrebbe non essere del tutto sovrapponibile. Tale conflitto emerge se si considera la natura dei corpora BA e L2, che differiscono sia per la qualità delle registrazioni sia per il tipo di italiano non-standard; di contro, la ricerca sulle *under-resourced languages* sembra indicare che la definizione di (non-)standard dipenda primariamente dalla distanza dell'*input* dai materiali di addestramento (Besacier et al. 2014; Biadsy et al. 2012; Tavosanis 2018). Pertanto, lo specifico tipo di *input* sembrerebbe essere trascurabile, ma ulteriori approfondimenti potrebbero confutare tale ipotesi. Inoltre, è necessario individuare la gerarchia dei parametri che influenzano l'accuratezza, sebbene ciò comporti un allontanamento dall'approccio basato sugli utenti. Infine, si conferma la necessità di continuare a sviluppare strumenti di automazione delle trascrizioni, nonché complementari strategie di correzione che ne migliorino la fruizione da parte degli utenti impegnati nello studio di materiali linguistici, in particolare conversazionali e di italiano non-standard.

Bibliografia

Abulimiti, Ayimunishagu & Schultz, Tanja. 2020. Automatic Speech Recognition for Uyghur through Multilingual Acoustic Modeling. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 6444-6449. Marseille.

Albano Leoni, Federico & Maturi, Pietro. 1995. *Manuale di fonetica*. Roma: Carocci.

Albano Leoni, Federico & Sobrero, Alberto A. & Paoloni, Andrea. 2007. Corpora e lessici di italiano parlato e scritto (CLIPS). *Bollettino di italianistica* 2. 122-130. doi:10.7367/71826.

Ashwell, Tim & Elam, Jesse R. 2017. How accurately can the Google Web Speech API recognize and transcribe Japanese L2 English learners' oral production? *JALT CALL Journal* 13(1). 59-76.

Badino, Leonardo. 2016. The ArtiPhon Task at Evalita 2016. In Basile, Pierpaolo & Cutugno, Franco & Nissim, Malvina & Patti, Viviana & Sprugnoli, Rachele (eds.), *EVALITA. Evaluation of NLP and Speech Tools for Italian*, 20-25. Pisa: Accademia University Press. doi:10.4000/books.aaccademia.1930.

Bechet, Frederic & Favre, Benoit. 2013. ASR error segment localization for spoken recovery strategy. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6837-6841. Vancouver, BC, Canada: IEEE. doi:10.1109/ICASSP.2013.6638986.

Besacier, Laurent & Barnard, Etienne & Karpov, Alexey & Schultz, Tanja. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication* 56. 85-100. doi:10.1016/j.specom.2013.07.008.

Biadsy, Fadi & Moreno, Pedro J & Jansche, Martin. 2012. Google's Cross-Dialect Arabic Voice Search. *2012 IEEE Acoustics, Speech, and Signal Processing*, 4441-4444. Kyoto, Japan. doi:10.1109/ICASSP.2012.6288905.

Bybee, Joan L. 2001. *Phonology and language use*. Cambridge, UK: Cambridge University Press.

Çetin, Özgür & Shriberg, Elizabeth. 2006. Speaker Overlaps and ASR Errors in Meetings: Effects Before, During, and After the Overlap. *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, vol. 1, I-357-I-360. Toulouse, France: IEEE. doi:10.1109/ICASSP.2006.1660031.

Crible, Ludivine. 2016. Discourse Markers and Disfluencies. Integrating Functional and Formal Annotations. *Proceedings of the LREC 2016 Workshop ISA-12*, 38-45. Portoroz, Slovenia.

D'Achille, Paolo. 2019. *L'italiano contemporaneo*. Terza edizione. Bologna: il Mulino.

Draxler, Christoph & van den Heuvel, Henk & van Hessen, Arjan & Calamai, Silvia & Corti, Louise & Scagliola, Stefania. 2020. A CLARIN Transcription Portal for Interview Data. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 3353-3359. Marseille.

Falcone, Mauro & Barone, Antonio & Bonomi, Alessandro. 2007a. Definizione e caratterizzazione di un database vocale ortofonico realizzato da parlatori professionisti in camera anecoica. *Progetto CLIPS - W1-a3* 1-14.

Falcone, Mauro & Barone, Antonio & Bonomi, Alessandro. 2007b. Realizzazione del database ortofonico in camera anecoica. *Progetto CLIPS - W1-a3* 17.

Filippidou, Foteini & Moussiades, Lefteris. 2020. A Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems. In Maglogiannis, Ilias & Iliadis, Lazaros & Pimenidis, Elias (eds.), *Artificial Intelligence Applications and Innovations*, vol. 583, 73-82. Cham: Springer. doi:10.1007/978-3-030-49161-1_7.

Kitaoka, Norihide & Enami, Daisuke & Nakagawa, Seiichi. 2014. Effect of acoustic and linguistic contexts on human and machine speech recognition. *Computer Speech & Language* 28(3). 769-787. doi:10.1016/j.csl.2013.09.009.

Levenshtein, Vladimir Iosifovich. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady* 10. 707-710.

Levis, John & Suvorov, Ruslan. 2012. Automatic Speech Recognition. In Chapelle, Carol A. (ed.), *The Encyclopedia of Applied Linguistics*. Oxford, UK: Blackwell. doi:10.1002/9781405198431.wbeal0066.

Li, Jinyu & Deng, Li & Haeb-Umbach, Reinhold & Gong, Yifan. 2016. *Robust automatic speech recognition: a bridge to practical applications*. Waltham, MA: Academic Press.

Maturi, Pietro. 2014. *I suoni delle lingue, i suoni dell'italiano: nuova introduzione alla fonetica*. Bologna: Il Mulino.

Lindblom, B. 1990. Explaining Phonetic Variation: A Sketch of the H&H Theory. In Hardcastle, William J. & Marchal, Alain (eds.), *Speech Production and Speech Modelling*, 403-439. Dordrecht: Springer Netherlands. doi:10.1007/978-94-009-2037-8_16.

Palmerini, Maria & Savy, Renata. 2014. Gli errori di un sistema di riconoscimento automatico del parlato. Analisi linguistica e primi risultati di una ricerca interdisciplinare. In Basili, Roberto & Lenci, Alessandro & Magnini, Bernardo (eds.), *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, vol. I, 281-285. Pisa: Pisa University Press.

Sakoe, Hiroaki & Chiba, Seibi. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26(1). 43-49. doi:10.1109/TASSP.1978.1163055.

Savy, Renata. 2007. Specifiche per la trascrizione ortografica annotata dei testi raccolti. *Progetto CLIPS - W1-a4* 1-28.

Scagliola, Stefania & Corti, Louise & Calamai, Silvia & Karrouche, Norah & Beeken, Jeannine & van Hessen, Arjan & Draxler, Cristoph & van den Heuvel, Henk & Broekhuizen, Max & Truong, Khiet. 2020. Cross disciplinary overtures with interview data: Integrating digital practices and tools in the scholarly workflow. 126-136. doi:10.3384/ecp2020172015.

Song, Minguang & Zhao, Yunxin & Wang, Shaojun & Han, Mei. 2021. Word Similarity Based Label Smoothing in Rnnlm Training for ASR. *2021 IEEE Spoken Language Technology Workshop (SLT)*, 280-285. Shenzhen, China: IEEE. doi:10.1109/SLT48900.2021.9383598

Tavosanis, Mirko. 2018. *Lingue e intelligenza artificiale*. Roma: Carocci.

van den Heuvel, Henk. 2020. Crossing the SSH Bridge with Interview Data. *Proceedings of LR4SSHOC: Workshop about Language Resources for the SSH Cloud*, 42-44.