ISSN: 2612-226X

Studi AISV 9

LA POSIZIONE DEL PARLANTE NELL'INTERAZIONE: ATTEGGIAMENTI, INTENZIONI ED EMOZIONI NELLA COMUNICAZIONE VERBALE

THE POSITION OF THE SPEAKER IN INTERACTION: ATTITUDES, INTENTIONS, AND EMOTIONS IN VERBAL COMMUNICATION

a cura di

Riccardo Orrico e Loredana Schettino

AISV
Associazione Italiana Scienze della Voce

LA POSIZIONE DEL PARLANTE NELL'INTERAZIONE: ATTEGGIAMENTI, INTENZIONI ED EMOZIONI NELLA COMUNICAZIONE VERBALE

THE POSITION OF THE SPEAKER IN INTERACTION: ATTITUDES, INTENTIONS, AND EMOTIONS IN VERBAL COMMUNICATION

a cura di
Riccardo Orrico e Loredana Schettino

Milano 2022

Studi AISV è una collana di volumi collettanei e monografie dedicati alla dimensione sonora del linguaggio e alle diverse interfacce con le altre componenti della grammatica e col discorso. La collana, programmaticamente interdisciplinare, è aperta a molteplici punti di vista e argomenti sul linguaggio: dall'attenzione per la struttura sonora alla variazione sociofonetica e al mutamento storico, dai disturbi della parola alle basi cognitive e neurobiologiche delle rappresentazioni fonologiche, fino alle applicazioni tecnologiche.

I testi sono sottoposti a processi di revisione anonima fra pari che ne assicurano la conformità ai più alti livelli qualitativi del settore.

I volumi sono pubblicati nel sito dell'Associazione Italiana di Scienze della Voce con accesso libero a tutti gli interessati.

c/o LUSI Lab - Dip. di Scienze Fisiche
Complesso Universitario di Monte S. Angelo
via Cynthia snc
80135 Napoli
email: presidente@aisv.it
sito: www.aisv.it

Edizione realizzata da
Officinaventuno
Via F.lli Bazzaro, 18
20128 Milano - Italy
email: info@officinaventuno.com
sito: www.officinaventuno.com

ISBN edizione digitale: 978-88-97657-63-7
ISSN: 2612-226X

Sommario

DOI: 10.17469/O2109AISV000000

IOLANDA ALFANO, FRANCESCO CUTUGNO

Prefazione

Il volume raccoglie alcuni dei contributi discussi durante il XVIII Convegno Nazionale dell'AISV (Associazione Italiana di Scienze della Voce), che si è svolto presso l'Università degli Studi di Napoli "Federico II" nel maggio 2022, dal titolo *La posizione del parlante nell'interazione: atteggiamenti, intenzioni ed emozioni nella comunicazione verbale*. Il convegno invitava alla riflessione su tutti gli aspetti legati in vario modo alla posizione assunta dal parlante negli scambi verbali, vale a dire sulle risorse linguistiche e "paralinguistiche" espresse per affermare la propria (pre)(dis)posizione e il proprio atteggiamento verso l'altro. I lavori si sono svolti in parallelo con il XXII Congresso dell'AItLA (Associazione Italiana di Linguistica Applicata), dal titolo *Vecchie e nuove forme di comunicazione diseguale: canali, strutture e modelli* ed hanno coinvolto un pubblico eterogeneo e particolarmente interessato ai vari temi proposti. Una delle sessioni in comune tra i due convegni ha visto la partecipazione ad una vivace Tavola Rotonda sul tema del linguaggio d'odio, *hate speech*, moderata da Massimo Pettorino, con gli interventi di Franca Orletti (Università di Roma Tre), Chiara Valerio (Marsilio Editore), Paola Villa, Natascia Festa (Corriere del Mezzogiorno/Corriere della Sera). Ricche e stimolanti poi le molteplici relazioni plenarie di ospiti italiani e stranieri. Ha aperto il convegno Oliver Niebuhr (Università della Danimarca Meridionale) con una relazione dal titolo *Influencing influence – Acoustic charisma and its effects in verbal communication*. In ordine di presentazione, Mariapaola D'Imperio (Rutgers University) e Riccardo Orrico (Università degli Studi di Napoli "Federico II") hanno poi presentato alcuni risultati delle loro ricerche nell'ambito di *Intonational Meaning and Individual Variability*. Inoltre, Emanuela Campisi (Università degli Studi di Catania) ha proposto un contributo dal titolo *From the individual to the group: reframing the notion of gesture space for the study of spoken interaction*. Infine, Julia Hirschberg (Columbia University) ha chiuso le relazioni su invito con una comunicazione intitolata *Trusted and Mistrusted Speech: Acoustic-Prosodic and Lexical Cues to the Speech Humans Trust – and the Speech They Do Not?*

Il volume ospita, quindi, alcuni dei lavori presentati in quella sede e sottoposti a *blind peer review*, nei quali trovano posto saggi sul tema del convegno, nonché contributi a tema libero. Nel piano complessivo della raccolta, emergono vari e interessanti spunti di riflessione sulla variazione nel parlato (nativo e non nativo, patologico e normofasico) esaminata considerando svariati indici sul piano fonico, in funzione della situazione comunicativa, della scelta del codice, ma soprattutto degli specifici interlocutori e del loro ruolo nella conversazione.

Apre il libro il contributo di Avesani, Bonifacio, D'Aloia, Olivucci, Calignano, Vayra e Zmarich, che presentano il primo studio longitudinale sullo sviluppo della prominenza lessicale e postlessicale in italiano, analizzando un bambino dell'Italia nord-orientale, registrato ogni 3 mesi dai 18 ai 36 mesi di età. L'analisi si inserisce nel quadro della Fonologia Prosodica e della Teoria Autosegmentale Metrica dell'Intonazione. I risultati mostrano che solo la realizzazione fonetica del livello più alto di prominenza procede in modo lineare: la durata delle vocali nucleari di IP aumenta costantemente nel tempo e quella delle atone diminuisce. A livello lessicale, in linea con le aspettative e coerentemente con la gerarchia prosodica dell'adulto, la traiettoria di sviluppo non segue invece un andamento lineare.

Nel saggio successivo, Cangemi, Dipino, Garassino e Schmid conducono uno studio sui correlati fonetici della complessità narrativa. Gli autori esaminano un corpus di narrazioni in tedesco svizzero elicitate attraverso un paradigma sperimentale mediante vignette caratterizzate da un diverso grado di complessità, sia rispetto al susseguirsi delle vicende sia rispetto al numero di personaggi coinvolti. I risultati forniscono nuove prove del legame tra la complessità della vicenda e quella della narrazione: le storie complesse hanno una maggiore durata complessiva e un maggior numero di unità interpausali, le quali sono di breve durata e contengono un minor numero di sillabe. A dispetto delle variazioni tra i parlanti, emerge per tutti che a maggiore complessità narrativa corrisponde una maggiore lunghezza e una maggiore frammentarietà.

Segue il contributo di Crocco, Gili Fivela e Magistro, che offrono un confronto prosodico tra l'italiano regionale parlato a Venezia e il dialetto veneziano. Utilizzando un compito di lettura in parlanti bilingui, gli autori esaminano enunciati comparabili da un punto di vista lessicale, sintattico e informativo. Confrontano le proprietà ritmiche dei due sistemi ed esaminano diverse metriche, giungendo a dimostrare una diversa organizzazione metrica nei due sistemi in contatto, sensibile ai processi segmentali che li differenziano. Al di là, inoltre, del confronto specifico, il lavoro presenta e discute una metodologia replicabile, applicabile al confronto tra altre varietà dialettali e italiane, che consente di identificare e quantificare mediante l'analisi di diversi parametri differenze prosodiche rilevanti.

Di Salvo esamina, nel contributo successivo, la selezione del codice e la variazione nell'uso del dialetto e dell'inglese in conversazioni raccolte in una comunità italiana del Regno Unito. I risultati evidenziano che la scelta tra italiano e dialetto è condizionata dal tipo di rapporto con l'interlocutore: i parlanti bilingui si rivolgono in italiano a un interlocutore esterno alla comunità, ma preferiscono il dialetto nelle interazioni con compaesani residenti nel contesto di immigrazione. La relazione con l'interlocutore (interno/esterno) condiziona, inoltre, anche l'alternanza con l'inglese, sia sul piano strutturale sia su quello funzionale. Dal punto di vista metodologico, l'autrice evidenzia che il caso in esame suggerisce la necessità di costruire corpora che possano essere rappresentativi della complessità rispetto alla relazione tra i partecipanti, "per comprendere quanto possa essere ampio lo spettro di variazione da un lato e quali siano i valori, funzionali, emotivi, sociali, legati alla

selezione di codice, di alcune varianti specifiche e della lingua dominante della società di accoglienza." (p. 94)

Gili Fivela, D'Apolito e Pagliaro esaminano il grado di accuratezza del parlato disartrico nel caso della produzione di segmenti fonologicamente diversi (occlusive *vs*. fricative) o socio-foneticamente marcati (occlusive aspirate), effettuando due esperimenti, il primo su parlato controllato e il secondo su parlato semispontaneo. I risultati indicano che i soggetti disartrici differenziano fricative e occlusive, ma non differiscono in modo statisticamente significativo dai controlli nella realizzazione dell'aspirazione. Le autrici rilevano, quindi, che l'aspirazione in quanto tratto sociolinguistico possa non essere preservata grazie a strategie di compensazione tanto quanto la differenza tra segmenti fonologicamente rilevanti (occlusive *vs*. fricative), che viene mantenuta nonostante la riduzione dei correlati, preservando le proporzioni attese (durata delle occlusive minore di quella delle fricative) e, soprattutto nel parlato semispontaneo del secondo esperimento, alcuni tratti caratteristici (come il *burst* nelle occlusive).

Il contributo di Maffia e Pettorino si inserisce nel dibattito sul parlato in ambito didattico, offrendo vari spunti di riflessione critica. Il lavoro verte sulle caratteristiche ritmiche e intonative del parlato di due docenti native di lingua italiana che si trovano a interagire con un gruppo di nativi, studenti italofoni, e un gruppo di non nativi, apprendenti stranieri adulti, con l'obiettivo di stabilire se e in che misura il parlato mostri caratteristiche prosodiche diverse in funzione dei due diversi gruppi di interlocutori. Gli autori, diversamente da quanto in alcuni casi riportato in letteratura, non riscontrano differenze nella velocità di eloquio e di articolazione, né nel *range* tonale, ma unicamente nella frequenza e nella durata delle pause silenti. Rispetto alla presenza di fenomeni di disfluenza, le docenti esaminate adattano il proprio eloquio al contesto L2, impiegando però strategie differenti.

A seguire, anche il lavoro di Rossi considera variazioni nel parlato rispetto al tipo di interlocutore, ma in relazione ad aspetti ben diversi. L'autrice esamina, infatti, l'influenza del genere sulla variazione degli indici fonetici coinvolti nella presa di turno in tedesco, esaminando durata, frequenza fondamentale e intensità in un corpus di conversazioni spontanee formate da coppie di parlanti dello stesso sesso oppure miste. I risultati suggeriscono che sia il genere del parlante sia quello dell'interlocutore possono influenzare il modo in cui i potenziali luoghi di transizione sono caratterizzati foneticamente nelle conversazioni.

Nel successivo lavoro, Sbranna, Wehrle e Grice studiano i fenomeni di *backchannels* nel tedesco come lingua straniera, esaminando parlato semispontaneo di apprendenti italofoni di tedesco con diversi livelli di competenza a confronto con parlanti nativi di tedesco. Gli autori analizzano frequenza, lunghezza, tipo lessicale e funzione svolta dal *backchannel*. Mentre la scelta dei tipi lessicali appare diversa in funzione del livello di competenza nella lingua straniera, nel senso che si avvicina alla lingua target unicamente nei parlanti con competenza avanzata, la frequenza e la lunghezza dei *backchannels* sembrano dipendere in misura maggiore dalla diade comunicativa specifica più che dal livello di competenza in sé. Inoltre, la variabilità

riscontrata tra i parlanti di tedesco come L1 induce gli autori a mettere in discussione l'idea stessa di caratteristiche specifiche da acquisire, almeno per quanto riguarda frequenza e lunghezza, diversamente da quanto riscontrato per i tipi lessicali e le loro funzioni.

Chiudono il volume due lavori incentrati sui fenomeni di disfluenza. Il primo, di Schettino, Alfano, Cataldo e Leo è di taglio contrastivo e considera parlato dialogico semispontaneo in italiano e spagnolo, indagando le caratteristiche formali e funzionali di pause piene e allungamenti segmentali. I risultati di questo studio pilota indicano che i parlanti italiani esaminati realizzano un maggior numero di fenomeni di disfluenza osservati rispetto a quelli spagnoli. In entrambe le lingue, al di là della variabilità individuale, gli allungamenti sono globalmente più frequenti delle pause piene. Quanto alla loro realizzazione, non emergono differenze interlinguistiche né nella durata, né rispetto alla posizione preferita, vale a dire vocalica in finale di parola. Le pause piene, invece, mostrano una diversa composizione segmentale tra le due lingue, legata all'inventario fonetico-fonologico specifico e sembrano essere impiegate per svolgere funzioni diverse.

Il secondo lavoro sullo stesso tema, nonché ultimo articolo di questa raccolta, di Schettino, Origlia e Matrone, affronta il tema dell'influenza di tali fenomeni sulla percezione del parlato. Il lavoro verte, infatti, sulla valutazione percettiva di un sistema di sintesi vocale che include allungamenti segmentali, pause silenti e riempitivi. La preparazione degli stimoli sperimentali è stata supportata da un modello computazionale del parlato. In particolare, un sistema di sintesi neurale è stato addestrato a generare enunciati che includono i fenomeni osservati in modo contestualmente plausibile. Successivamente, attraverso un test di discriminazione è stato valutato se e in che misura i fenomeni inseriti come da modello possano influenzare la percezione degli ascoltatori degli enunciati sintetizzati. I risultati mostrano che gli enunciati sintetici che includono disfluenze sono giudicati più naturali rispetto agli enunciati che non ne includono.

Il volume testimonia alcuni interessi di ricerca frutto dell'intensa e proficua collaborazione fra studiosi nel settore della fonetica sperimentale in Italia, sempre pronta ad accettare derive attuali e importanti. Dagli studi di carattere sociolinguistico analizzati sotto la luce specialistica dello studio delle strutture del parlato, passando per gli studi sul parlato patologico, si giunge a lavori con ricadute tecnologiche specificamente nel campo della valutazione della qualità della sintesi vocale. La nostra comunità scientifica di riferimento si arricchisce ogni anno di nuovi giovani studiosi che vanno ad affiancare i soci storici della associazione: la presenza continua di pochi ma volenterosi studiosi che scelgono questi temi è il segno dell'interesse che la fonetica sperimentale ancora solleva, sebbene nei corsi istituzionali di linguistica generale e glottologia (ma anche nei corrispettivi settori dei corsi di laurea ad indirizzo tecnologico-applicativo) allo studio della voce umana e della comunicazione parlata sia riconosciuto un ruolo meno significativo di quanto meriterebbero. Alla folta presenza di studiosi ai convegni AISV, si aggiungono le presenze di ulteriori contributi nei convegni di varie associazioni linguistiche nazionali, come GSCP,

SLI e AItLA, così come numerosi sono i contributi di autori italiani nei principali convegni internazionali di settore come ICPhS, *InterSpeech* e *Speech Prosody*.

Detto ciò, il più anziano degli autori di questa prefazione, avendo qualche decennio di esperienza nella frequentazione della comunità scientifica dei fonetisti italiani, "ruba" un po' di spazio per proporre alcune semplici riflessioni: come si presenta la situazione della ricerca scientifica nel settore in Italia? Da una parte, nel settore tecnologico, si registra purtroppo una diminuzione degli studi sul riconoscimento del parlato, che oramai viene studiato solo in pochi laboratori che hanno raccolto l'eredità di Piero Cosi che da un paio di anni si è ritirato a vita privata. Sempre vivi, invece, sono gli studi di fonetica dei dialetti, settore che l'associazione ha sempre sostenuto e che, con diversa numerosità nel corso del tempo, hanno sempre trovato spazio nei nostri convegni. Per quanto riguarda gli studi sul parlato in produzione, resta attuale il dibattito metodologico relativo alla scelta del tipo di dati su cui basare l'osservazione dei fenomeni oggetto di studio, ovvero fra coloro che decidono di affrontare lo studio del parlato (semi-)spontaneo, o comunque non raccolto *ad hoc* per scopi specifici, e coloro che preferiscono lavorare con parlato laboratoriale controllato. Ad oggi è ancora possibile individuare una distinzione abbastanza netta nell'adozione dell'uno o dell'altro approccio sperimentale, sebbene la consapevolezza dell'influenza del tipo di dato osservato sui risultati analitici ottenuti induca alla necessità di riflessione sulla confrontabilità dei risultati che questi due approcci producono, laddove non è sempre automaticamente possibile estendere i modelli teorici e i sistemi di riferimento generati in uno dei processi nell'altro, e stia spingendo verso maggiore consapevolezza metodologica e la considerazione di più tipi di parlato, per poter osservare quali risultati, e in che misura, siano generalizzabili. Anche gli studi prosodici restano centrali nell'ambito della nostra comunità. Nel panorama della ricerca nazionale l'approccio autosegmentale metrico continua ad essere adottato dalla maggior parte degli autori che studiano prosodia, in particolar modo l'intonazione, sebbene sia possibile registrare l'emergere dell'esigenza di integrare considerazioni fonetiche e la tendenza alla ricerca di nuovi approcci più squisitamente fonetici, precedenti alla determinazione di una complessa tassonomia fonologica.

Per concludere, la raccolta dei lavori qui brevemente delineati rende, a nostro parere, la pubblicazione di questo volume un prezioso contributo alla riflessione scientifica della comunità dei fonetisti e dei linguisti italiani, offrendo validi strumenti metodologici e aprendo nuove strade di ricerca. Il merito di questo valore è da attribuire a tutti coloro che hanno contribuito, che desideriamo ringraziare singolarmente.

I curatori desiderano, infine, ringraziare i membri del Comitato Scientifico del XVIII Convegno AISV del 2022, per la loro generosa collaborazione alla revisione dei lavori.

DOI: 10.17469/O2109AISV000001

CINZIA AVESANI, SERENA BONIFACIO, GIULIA CALIGNANO, VALERIA D'ALOIA, FRANCESCO OLIVUCCI, MARIO VAYRA, CLAUDIO ZMARICH

The emergence of lexical and post-lexical prominence in Italian. A case study

Our study identifies the developmental trajectory of prominence at lexical and post-lexical levels. From very early in life infants are sensitive to lexical stress contrasts, but, due to very limited vocal capabilities, the production of stress contrasts only starts in the second year of age. We address the question of whether, when and how a child learns to differentiate lexical (stress) from post-lexical prominence (accent) by acoustically examining the spontaneous productions of one child from North-East Italy recorded every 3 months from 18 to 36 months of age. Our analysis is cast in the framework of the Autosegmental Metrical Theory of Intonation. Results show that during the child's prosodic development the duration of IP nuclear vowels increases linearly, the duration of unstressed vowels decreases linearly and the duration of stressed, prenuclear and ip nuclear vowels is progressively but non-linearly adjusted, consistent with the adult prosodic hierarchy.

Keywords: development of prosodic prominence, stress, accent, Italian, Autosegmental Metrical Theory of Intonation.

1. *Introduction*

Although studies on the acquisition of prosody are fewer in number than studies on the acquisition of other aspects of language, a large number of experimental studies have shown that children from a very early age are sensitive to prosody. Such sensitivity is rooted in prenatal experience with speech, which consists mainly of prosodic information, and it has been shown to already impact how newborns perceive speech and produce communicative sounds (Gervain, 2015).

Not only do newborns recognize their mother's voice and prefer it to other female voices (De Casper, Fifer, 1980), but they also recognize vowels heard prenatally (Moon, Lagercrantz & Kuhl, 2013) and their native language (Moon, Cooper & Fifer, 1993). Soon after birth, at only two days of age, newborns are sensitive to the alternation of stressed and unstressed syllables: Italian two-day newborns can discriminate between disyllabic and trisyllabic words differing in stress pattern regardless of consonant variations (màma vs. mamà, tàcala vs. tacàla), and seem to be able to categorize words based on their stress pattern (Sansavini, Bertoncini & Giovannelli, 1997).

The preference for the prosody of the maternal language becomes more specific at about 6 months. At this age, infants discriminate and prefer to listen to words of the maternal language rather than to those of a foreign language when the two

languages differ in their global prosody, whereas this preference appears only at 9 months when the two languages mostly differ in their phonetic and phonotactic properties (Jusczyk, Friederici, Wessels, Svenkerud & Jusczyk, 1993).

At 4-9 months of age, infants discriminate between patterns of lexical stress, lexical pitch contours, and lexical tones if these patterns are contrastive in their ambient language, and they seem to consider all acoustic parameters (duration, pitch and intensity) to build their preferences (Bahatara, Boll-Avetisyan, Hohle & Nazzi, 2018). Infants of this age are sensitive to lexical stress contrasts and show a preference for the predominant stress pattern of the ambient language, which, in turn, partly determines the capacity of extracting word forms from fluent speech (Bahatara et al., 2018; Bion, Benavides-Varela & Nespor, M., 2011).

Moreover, 7- to 10-month-old infants prefer to listen to clauses of the maternal language in which artificial pauses are inserted between rather than within clauses Hirsh-Pasek, Kemler Nelson, Jusczyk, Cassidy, Druss & Kennedy, 1987) and 9-month-old infants prefer to listen to phrases of the maternal language in which artificial pauses are inserted between rather than within phrases (Jusczyk, Hirsch-Pasek, Kemler Nelson, Kennedy, Woodward & Piwoz, 1992).

It thus appears that, in their first months of life, infants pay attention to and discriminate the global prosody of speech and become attuned to the prosodic patterns of the maternal language.

To summarize, infants are sensitive to a variety of prosodic cues if they are linguistically relevant in the language environment. Several studies have proposed that they can use them to access lexical and morpho-syntactic information of their native language (e.g. the familiarity with the typical stress patterns of the ambient language enables infants to segment the continuous speech stream into words; by relying on the correlation between the position and the acoustic realization of phrase-level prominences and word order, infants can distinguish Head-Complement from Complement-Head languages). In this view, prosody has been interpreted as a flywheel for language acquisition, because the ability to perceive the native language prosodic patterns is considered the trigger to the acquisition of other language domains ("prosodic bootstrapping": for a review, see Gervain, Christophe & Mazuka 2021).

1.1 Stress

In production, young children seem to show more accurate control of intonation earlier than duration (Prieto, Estrella, Thorson & Vanrell 2012). Toward the end of the first year of life, children of intonational languages begin to produce linguo-specific intonational patterns, which develop rapidly during the second year of life as they learn to associate pragmatic meanings and prosodic features (Esteve-Gibèrt, Prieto, 2018).

The production of contrastive stress patterns occurs later. Several studies have shown that English children begin to produce contrasts of stressed and unstressed syllables at or after 18 months of age: at 18 months according to Kehoe, Stoel-

Gammon & Buder (1995), at 22 months according to Schwartz, Petinou, Goffman, Lazawski & Cartusciello (1996), at 24 months according to Pollock, Brammer & Hageman (1993). In these early stages, duration is the parameter that is used by children to produce accentual oppositions, while F0 and intensity will contribute to accent production only later (Pollock et al., 1993).

In those studies, however, no distinction is made as to whether the prominent syllables under consideration are stressed at a word or phrase level. That is, whether within the utterance they are not only lexically stressed but also associated with melodic configurations that give them a higher degree of prominence.

1.2 Italian

Studies on the acquisition of prosody in Italian children are mainly due to the work of D'Odorico and colleagues (D'Odorico, Carubbi, 2003; D'Odorico, Fasolo & Marchione, 2009; D'Odorico, Fasolo & Zanchi 2010), who studied the development of intonation at 24-36 months of age, adopting a global approach and relating it to children's syntactic and narrative ability.

An analysis of intonation cast within the AM theoretical framework is due to Zanchi, D'Imperio, Zampini & Fasolo (2016), who studied 3- to 4-year-old children. Their results indicate that 3-year-olds master nuclear pitch accents as adults, but they do not produce rising boundary tones in the same measure as adults.

Specifically on stress development, Arciuli and Colombo (2016) analyze the productions of 3- to 5-year-old children to delineate developmental trajectories in the ability to produce stressed and unstressed syllables in trisyllabic words with a trochaic (SW) or iambic (WS) beginning. The approach is a traditional phonetic one: stressed and unstressed syllables are analyzed according to their acoustic characteristics of duration, peak intensity, and F0. The analysis is centered on a relatively late age group (3-5 years), in a stage of development (well past that of the early vocabulary) in which possible discrepancies with adult stress targets may have been resolved. The study has the merit to be one of the first to address the production of stress in Italian children but, by running a group analysis, it fails to delineate individual developmental trajectories.

In a different vein, the study by Olivucci, Pasqualetto, Vayra & Zmarich (2016) delineates the developmental trajectories of lexical stress in the production of 5 young children from 21 to 27 months of age. The authors analyzed duration, intensity, spectral emphasis, F1 and F2 formant trajectories of stressed and unstressed vowels (/a, i, o, u/), and compared them to those of an adult control group. Their results indicate that toddlers differentiate unstressed and stressed vowels starting from 21 months of age, and that the acoustic parameters that most significantly cue such difference are duration and spectral emphasis. In two later studies (Olivucci, Vayra, Avesani & Zmarich, 2018, 2019), three different children have been analyzed, extending the age window to comprise the children's production from 18 to 42 months. Stressed and unstressed vowels were distinguished based on their position in one and multi-word utterances. Results showed that lexical stress distinctions appear

starting from 18 months of age, and that stressed vowels were significantly more prominent than unstressed ones in single-word utterances and in the final words of multi-word utterances, while they were not distinguished in words occurring within the utterance. Duration confirms to be the most reliable correlate of lexical stress, and the developmental trajectory shows that the difference between unstressed and stressed vowels increases with age. Interestingly, the increasing difference is not only due to an increase in the duration of stressed vowels but also to a decrease in the duration of unstressed ones.

2. *Aims and predictions*

The present study builds on the works by Olivucci and colleagues, and extends its focus on the acquisition of post-lexical prominence in Italian. We have two aims: by enlarging the sample of analysis, we would like to confirm the developmental trajectory of lexical stress. Our second aim is to go further and address the acquisition of post-lexical prominence. That is, we would like to understand when, in individual development, stressed syllables start acquiring relatively different degrees of prominence once the words to which they belong form part of a structured sentence. Triggered by the observations in Olivucci et al. (2018, 2019) that stress differences emerge in the strongest prosodic position in multi-word utterances (i.e. only in the utterance-final word), we aim at uncovering the emergence of the prosodic structuring of the sentence in the child's speech by looking at the acoustic properties of the lexically stressed syllables in one, two and multi-word utterances.

We adopt the framework of Prosodic Phonology (Selkirk, 1984; Nespor, Vogel, 2007[2]) and the Autosegmental Metrical Theory of Intonation (Beckman, Pierrehumbert, 1986; Ladd, 2008[2]). Prosody constitutes the organizational structure of spoken language, and Prosodic Phonology, integrated into AM theory, considers it to consist of several hierarchically ordered metrical constituents that, above the Syllable and the Foot, scholars agree to be the Prosodic Word (W), the Phonological or Intermediate Phrase (respectively, ϕ or ip, depending on the version of the theory) which consist in one or more Prosodic Words, and the Intonational Phrase (IP), which consists of one or more Phonological Phrases.

Each constituent is endowed with a head or nucleus (the strongest element in the constituent). In Italian, the head of the Word is the syllable designated in the lexicon to carry the lexical stress (**S**tressed), while at the higher levels of hierarchically ordered constituents the head is (the stressed syllable of) the last word in each constituent. These postlexically metrically strong syllables are designated to be the exponents of phrasal prominence and are associated with linguo-specific tonal configurations that give rise to pitch accents. We code the head syllable of an IP, or *nuclear* in the IP, as **N** and the head syllable of ip, or nuclear in ip, as **Nphp**. Metrically weaker syllables in IPs and ips can be associated with a pitch accent even if they are not the head of their constituent, and by virtue of this association, they acquire greater prominence than a lexically stressed syllable that does not have a pitch accent (e.g. Beckman,

1996). These syllables, stronger than the lexically stressed ones and weaker than the nuclearly accented head ones, are named prenuclear (**P**).

In Fig 1, the head syllables of each constituent (W, ip and IP) are indicated by a star.

Figure1 - *Schematic representation of the higher levels of prosodic structure: Prosodic Word (W), Phonological/Intermediate Phrase (ip) and Intonational Phrase (IP). Stars represent the head elements in each constituent*

Based on their position in the hierarchy of prosodic constituents, metrically strong syllables carry a progressively higher degree of structural prominence, according to the following progression:

IP Nuclear > ip Nuclear > Prenuclear > Stressed (> Unstressed)

In adult speech, the hierarchy of word- and phrase-level prominence is substantiated by significant differences in a set of acoustic and articulatory parameters. In Italian, at the lexical level, stressed vowels show longer acoustic durations, more intensity, and more spectral emphasis than unstressed vowels. Moreover, they show less centralization in F1-F2 formant space: low stressed vowels show a higher degree of opening, with a higher F1, and high vowels a higher degree of fronting (with a higher F2) than unstressed vowels, and less C-V coarticulation (Farnetani, Kori, 1982, Vayra, Fowler, 1987; Vayra, 1991; Savy, Cutugno, 1996; Vayra, Avesani & Fowler, 1999; Tamburini, 2009).

Articulatorily, palatographic data show that stressed vowels are more open than unstressed ones and that tongue contacts decrease for unstressed high vowels and increases for unstressed low vowels (Farnetani, Faber, 1992). Kinematic data show that the jaw is lower in stressed low vowels (Vayra, Fowler, 1992; Magno Caldognetto, Vagges & Zmarich, 1995) and their opening gesture has a longer duration, higher peak velocity and more displacement as compared to unstressed vowels (Avesani, Vayra & Zmarich, 2009; Zmarich, Avesani, 2015).

At the postlexical level, sentence-level prominence manifests acoustically as enhanced acoustic parameters. Accented low vowels which are IP nuclear are longer, with a more extreme F1 trajectory, and with more energy in the high-frequency bands (lower values of spectral balance, leading to more prominence) than lexically stressed vowels (Avesani, Vayra, 2013). IP nuclear vowels are also significantly longer and with lower values of spectral tilt (more energy in the high frequencies of the spectrum, leading to more prominence) than prenuclearly accented vowels. However, the latter ones are not significantly longer than stressed vowels while there is evidence of their having more spectral emphasis in the high energy bands

(less spectral balance) than stressed ones (Bocci, Avesani, 2011). Articulatorily, IP nuclear vowels show opening gestures that are longer, with higher peak velocity and more displacement than stressed vowels.

Summarizing, the literature on Italian lexical and postlexical prominence shows that postlexical strong vowels are phonetically realized with acoustic and articulatory parameter values that directly correlate with their position in the prosodic hierarchy: accented IP nuclear vowels are more prominent than prenuclearly accented vowels than lexically stressed ones.

In the present study, we build on the results offered by the literature on the infants' perception of phrasal prosody (Gervain et al., 2021; Chen, Esteve-Gibert, Prieto & Redford, 2021) to investigate when a child begins to produce utterances which are prosodically phrased, and when the internal structure of the prosodic constituents in terms of relatively strong and weak elements will emerge. As reported by Gervain et al. (2021), infants perceive intonational phrase boundaries from 5 months of age (Hirsh-Pasek et al., 1987) and intermediate phrase boundaries from 9 months of age (Gerken, Jusczyk & Mandel 1994; Shukla, Wite & Aslin, 2011). At 20 months of age, French and English toddlers are able to get the correct interpretation of ambiguous sentences by relying on their prosody (de Carvalho, Lidz, Tieu, Bleam & Christophe, 2016; de Carvalho, Dautriche, Lin & Christophe, 2017), and at 18 months of age are able to exploit the syntactic structure accessed through phrasal prosody to guess the meaning of a novel word (de Carvalho, He, Lidz & Christophe, 2015).

We expect that, in production, prosodic phrasing will consistently appear when the child will produce not only utterances formed by the simple juxtaposition of two words but utterances effectively endowed with argument structure, progressively longer and syntactically more complex. At this stage, we expect utterances will be phrased into prosodic constituents in a hierarchical relationship to each other. Along with the emergence of prosodic phrasing, we also expect that the head syllables of each hierarchically ordered prosodic constituent will be marked by a degree of prominence coherent with their position in the prosodic hierarchy and that the preceding words can be optionally endowed with relatively weaker prenuclear accents.

We predict that, in the path of linguistic and prosodic development, the child will attain to produce lexically stressed, prenuclear and nuclear vowels which will be differentiated for duration. Moreover, we want to discover when this progressive differentiation would occur and how. One hypothesis is that the postlexically strong vowels will lengthen linearly as the linguistic and prosodic development procede. But speech development is hardly a linear process, and therefore an alternative hypothesis is that in the acquisition of prosodic prominence the adult target could be attained nonlinearly.

3. *Method*

3.1 Corpus and recordings

Data analyzed in this study are part of a corpus collected by Serena Bonifacio at Trieste, from 2007 to 2009. The corpus includes 10 Italian children, 4 males and 6 females, recorded every three months from 18 to 48 months of age. Parents compiled the MacArthur CDI surveys, in which they reported the most frequent words produced by their child at each developmental stage (Caselli, Casadio, 1995) and filled out a questionnaire aimed at verifying their normal psycho-physical and linguistic development. When they were 18-months-old, the children underwent audiologic screening to exclude the presence of hearing impairments (Ling, 1976).

In the semi-structured recording sessions (Schmitt, Meline, 1990), the child interacts with the clinician in front of a set of toys. These objects were chosen based on the list of words compiled by the parents on the MacArthur CDI and were presented on the basis of decreasing frequency of the semantic categories that resulted from CDI. Besides the most common words and for all developmental stages, children were invited to repeat (five times each at least) 12 minimal pseudo-word, initially stressed, contrasting labial, dental and velar voiced and voiceless stops: 'papa', 'baba', 'pipi', 'bibi', 'tata', 'dada', 'titi', 'didi', 'kaka', 'gaga', 'kiki', 'gigi'. In the last sessions, children were engaged in more structured conversations about their interests (holidays, movies, cartoons, family and so on). Each recording session lasted on average about an hour. The speech samples collected at 18 and 21 months were considered valid and representative of the child's linguistic abilities of those developmental stages only if the number of lexical forms produced represented at least 50% of the words in the lexical list compiled by the parent. Speech samples were recorded with an Edirol R-09 digital recorder, at 16-bit sample size and 44.1 kHz sampling frequency.

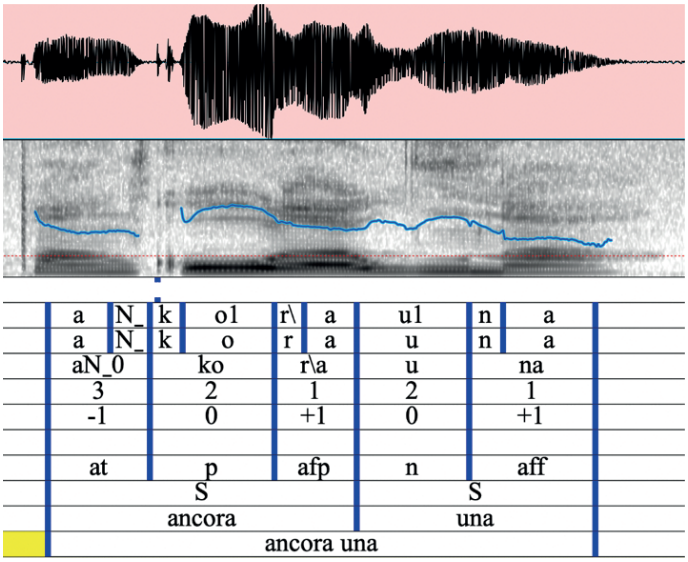
3.2 Data selection, coding, analysis

For the present study, we selected one of the most talkative children in the corpus (BS) and analyzed his productions at the stages of 18, 21, 24, 27, 30, 33 and 36 months. We considered bi-, tri- and quadrisyllabic words, with stress on the penultimate and ante-penultimate syllable, and uttered spontaneously (we did not consider words produced by repetition). The target words occurred in one-, two-, three- and multi-word utterances. Of the total number of syllabic occurrences (3454), the analysis focuses on open syllables, with C = voiced and voiceless stops, fricatives, nasals, liquids, affricates, and V = all Italian vowels (1157 syllables). Due to the typical distortion processes operated by children at the ages considered here, as well as the connected speech style of several utterances, the quality of some vowels does not completely match the quality of vowels in the adult phonological system.

We phonetically and prosodically annotated in *Praat* (Boersma, Weenink, 2021) the entire utterance in which the target words occur. An example is given in Fig. 1. We created 11 *tiers* in which we indicated, in order: the burst of the stop

consonant; the transcription (SAMPA) of the actual phone produced by the child; the adult target; the actual syllable; the distance of the syllable from the word end (1 = end); the distance of the syllable from the lexical stress (0 = stress); possible observations; the accentual status of the syllable: unstressed (**U**), stressed (**S**), prenuclearly accented (**P**), nuclearly accented in intermediate/phonological phrases (**Nphp**), and nuclearly accented in intonational phrases (**N**). Besides, we coded as **postF** the prominent syllable(s) which eventually follow the sentence focus. In Fig. 1, the prosodic labels also indicate the position of the syllable in the word and in the utterance (afp = unstressed and word-final; aff = unstressed and utterance-final). In the final *tiers* we coded, respectively, the modality of production (S = spontaneous), the adult target word and the adult target sentence (orthographic transcription). In order to assess the prominence status of the target syllables, 2 of the authors have annotated the whole corpus. In case of disagreement, the final decision on the prominence status of a syllable was reached by discussing each singular case.

Figure 2 - *An example of phonetic and prosodic annotation with Praat. Waveform, spectrogram and F0 of the utterance "ancora una", transcribed and coded in 11 tiers (see main text for details)*



For each of the 1149 vowels eligible for analysis we automatically measured: 1) duration, 2) F1 and F2 at vowel midpoint; 3a) spectral emphasis, which was calculated as spectral tilt (H1* – A3*) according to Jessen, Marasek (1997), with some adjustments to make calculations suitable for children; and 3b) as spectral balance (difference in dB between four contiguous frequency bands; Sluijter, van Heuven, 1996). Bands values are set based on the analysis of F0 and formant values for each developmental stage); 4) F1 and F2 trajectories (ten equidistant points over the vowel duration).

In the present paper, we analyze only the results of vowel duration, as the previous study by Olivucci et al. (2016) indicated that duration is the most robust cue to prominence among other acoustic cues (formant trajectories, spectral tilt and spectral balance). The data span across 18 months of child production, starting from the 18-month-recording in which the child's speech includes almost exclusively one-word utterances, to the 36-month-recording in which the child produces complex sentences. To factor out differences in speech rate, for each developmental stage we normalized 1) the duration of the vowels on the duration of the syllables in which they occur, and 2) the duration of each vowel on the duration of the nuclearly accented one. The correlation between normalized and row durations across the different developmental stages is very high in both types of normalizations ($R = 0.89$, $p < 0.0001$). Therefore, results will be reported in terms of raw duration values (ms).

4. *Results*

In what follows we offer a brief description of the child's prosodic development and present the results of vowel duration, separately for each recording session. Vowel duration for unstressed (**U**), lexically stressed (**S**), prenuclearly accented (**P**), nuclearly accented in intermediate/phonological phrase (**Nphp**), and nuclearly accented in intonational phrases (**N**), is presented in Fig. 3 in separate box plots per stage.

18 months. At this stage, the child produces all one-word utterances in response to the clinician's solicitations (e.g.: *Come si chiama questo? Cosa si fa con questo?*). Only one two-word utterance with reduplicated words (*baba baba*) is attested in the recording session in which the child interacts with the clinician for about an hour. Adult target words were disyllables with penultimate stress and trisyllables with stress on any syllable (eg: *fòrbice, poltróna, biberòn*).

Trisyllables as uttered by the child undergo a deletion process by which unstressed syllables (Weak) preceding the stressed ones (Strong) are omitted. Accordingly, SWW trisyllables retain the target word structure (*fòr.bi.ce* > *bò.ci.ci*), WSW trisyllables loose the pre-stress weak syllable (*pol.trò.na* > *tòn.na*), WWS oxitons are reduced to the final strong syllable (*bi.be.ròn* > *òn*).

Adult target bisyllables are trochees (SW, eg. *dàdo, càpra, tòpo*), but not all of them are uttered by the child as such: in 22% of the cases (18 out of 82) the stress pattern is reversed, from SW to WS: *tàta* > *tatà*. In 33% of cases (31 out of 82), both syllables are perceived by the transcribers of equal prominence (*dà.do* > *dà.dò)*. In the remaining cases (33 out of 82) the child produces the word according to the stress structure of the adult target, but these cases amount to only 40% of the total.

Due to the process of weak syllable deletion and our choice to exclude word- and sentence-final vowels from the computation, we were left with very few cases of unstressed vowels corresponding to the adult target. In order to arrive at a more

balanced set, we included the word-initial weak syllables of words uttered with stress inversion (ta.tà). As for stressed vowels, we considered them both as the metrical heads of the lexical word (i.e. stressed) and the metrical heads of the intonational phrase that wraps the one-word utterance (i.e. accented). This operational choice will need to be more thoroughly evaluated in future research.

The analysis of a total of 39 syllables shows that the average duration of unstressed vowels (n = 8, M = 118.25 ms, SD = 36.79 ms) is shorter than nuclearly accented ones (n = 30, M = 193.37 ms, SD = 47.82 ms). The median difference is 71.5 ms.

21 months. At the developental stage represented in the 21-month recording, two- and multi-word utterances appear. The child produces cases of Det + N, N + Adj, N + N, such as *la spazzola, la tata, la mucca, gallina bella, bibi questo, babbo natale,* as well as Prepositional Phrases, Adverbial Phrases and complex NPs such as the following: *sotto il fungo, come la tata, ghighi e la cova, gaga con la voce, scarpe de mamma*.

There are still some cases in which the stress structure of a word is different from the adult target: disyllabic paroxytone words that become oxytones (1 case), bisyllabic words in which both syllables have the same degree of perceived prominence (4 cases).

Along with nominal structures, the child begins to produce utterances with argument structure: two-word utterances in which the argument is in canonical position (object in a postverbal position such as *chiama la zebra, fa pipì, chiudi gli occhi*), and multi-word utterances with non-canonical order i.e. utterances in which at least one element appears in non-canonical order such as Subject in post-verbal position, and/or Object in pre-verbal position, (D'Odorico, Fasolo, Marchione, 2009). Examples are: *mamma palla damme; biberon fa la mamma; limone ti do questo; cucchiaio mano a Stefano questo*.

With the appearance of the phrasal structure, the structuring of the utterance into prosodic constituents also appears, as well as the modulation of postlexical strong syllables in different degrees of prominence. In (1) the utterance is phrased in two intonational phrases separated by a pause in which each IP-final word is nuclearly accented; in (2) the multi-word utterance is not internally phrased, but each lexical word is endowed with a different level of prominence: stress on the utterance initial word *cucchiaio*, prenuclear accent on *mano* and *Stefano*, nuclear accent on *questo*.

(1) [[mamma **pal**la]$_{IP}$ [**dam**me]$_{IP}$

(2) [cu**cchia**io$_{S}$ **ma**no$_{P}$ a **Ste**fano$_{P}$ **que**sto$_{N}$]$_{IP}$

Acoustically, vowels in metrically strong positions (here nuclearly accented in intonational phrase) are longer than the other ones (untressed, stressed and prenuclearly accented). Average values of the 102 eligible tokens are: U (n = 44, M = 140.68, SD = 49.16), S (n = 10, M = 221.30, SD = 54.03), P (n = 10, M = 184.30, SD = 68.64), N (n = 37, M = 234.27, SD = 47.00), With respect to the

18 month stage, the difference between U (M = 140.68, SD = 49.16) and N (M = 234.27, SD = 47.00) increases (median difference = 99 ms).

24 months. Along with uttering two-word utterances (Det+N), at 24 months the child produces simple and coordinated Prepositional Phrases (*con il cappello, della scatola e della bici*), short questions and exclamatives (*Che si chiama Chicchi? Che bon!*), VO (*racconta una storia*) and SVO sentences (*la tata fa pipì*), and sentences with direct and indirect arguments (*mette il vino nella bottiglia*).

Prosodically, at lexical level words with stress inversion are no longer present, but there remain 3 cases of word-final unstressed vowels that are perceptually as prominent as the word-initial stressed ones (e.g p**à**pp**a**).

At postlexical level, two clear cases of phrase accents appear. In the example (3), the stressed syllabe [gi] is the head of an intermediate/phonological phrase and is associated with an H*(+L) pitch accent:

(3) [[chiama **gh**ighi$_{Nphp}$]$_{ip}$ [con le **scar**pe$_{N}$] $_{ip}$]$_{IP}$

For the first time in his development the child produces a sentence with information focus on a left-dislocated object *(pipì)* (4b) in response to the question posed by the clinician in (4a)

(4a) Cosa fa? Ti ricordi? Come si dice?
(4b) [[la pi**pì**]$_{ip}$ [fa] $_{ip}$]$_{IP}$

It could be argued that in (4b) the non-canonical order results from a not yet fully developed competence of argument structure, rather than from an option selected by the child to mark the object pragmatically. In a study on Italian by D'Odorico et al (2009), for example, utterances with non-canonical order are attested, but "the distinction between non-canonical and canonical orders is hardly marked at all from a prosodic point of view" (ibid: 326). On the contrary, (4b) appears to be a true case of focus-background partition: first, the pitch countour conforms to the adult norm, as the nuclear syllable "pì" is marked by an L+H* pitch accent, and the post-focal pitch contour is lowered as in the adult language; second, in the same recording session, the child answers the question in (4) with a canonical VO sentence *fa pipì.* Therefore he shows to be able to pragmatically and prosodically master utterances with focus *in-situ* and *out-of-situ*.

Post-focal vowels have not been included in the acoustical analysis that counts 134 eligible tokens. Vowels in strong metrical positions N (n = 51, M = 205.49, SD = 42.95) and Nphp (n = 2, M = 215.0, SD = 9.90) are longer than vowels in metrically weaker positions U (n = 63, M = 105.59, SD = 36.69), S (n = 11, M = 132.73, SD = 53.60), P (n = 7, M = 142.14, SD = 41.16), while no clear differentiation in the duration of the vowels within each group is observable. The median difference between N and U is increasing with respect to the previous stages: 101 ms.

27 months. At 27 months the child interacts more with the clinician and is very talkative. The total number of eligible CV syllables in this recording session is 249. There are no longer words that do not prosodically conform to the adult

target: all words are produced with the expected stress structure. At syntactic level the child produces utterances with progressive verb forms (*sta mettendo in testa l'ombrello; sta stirando la camicia*) and sentences with subordinate clauses (*sta cucinando per andare a casa; bambino per andare in bicicletta*). Utterances with a sentence-initial information focus in response to a question posed by the clinician are more frequent:

(5a) C: questa, cosè questa?
(5b) BS: [**tor**ta]$_{FOC}$ [sè]$_{Backgr}$

(6a) C: e dove vola via?
(6b) BS: [sotto la **mac**china]$_{FOC}$ [va]$_{Background}$

(7) [tutte (le) **bam**bole]$_{FOC}$ [voglio]$_{Background}$

And a case of in-situ contrastive focus occurs, which is marked by a L+H* pitch accent with a large pitch span:

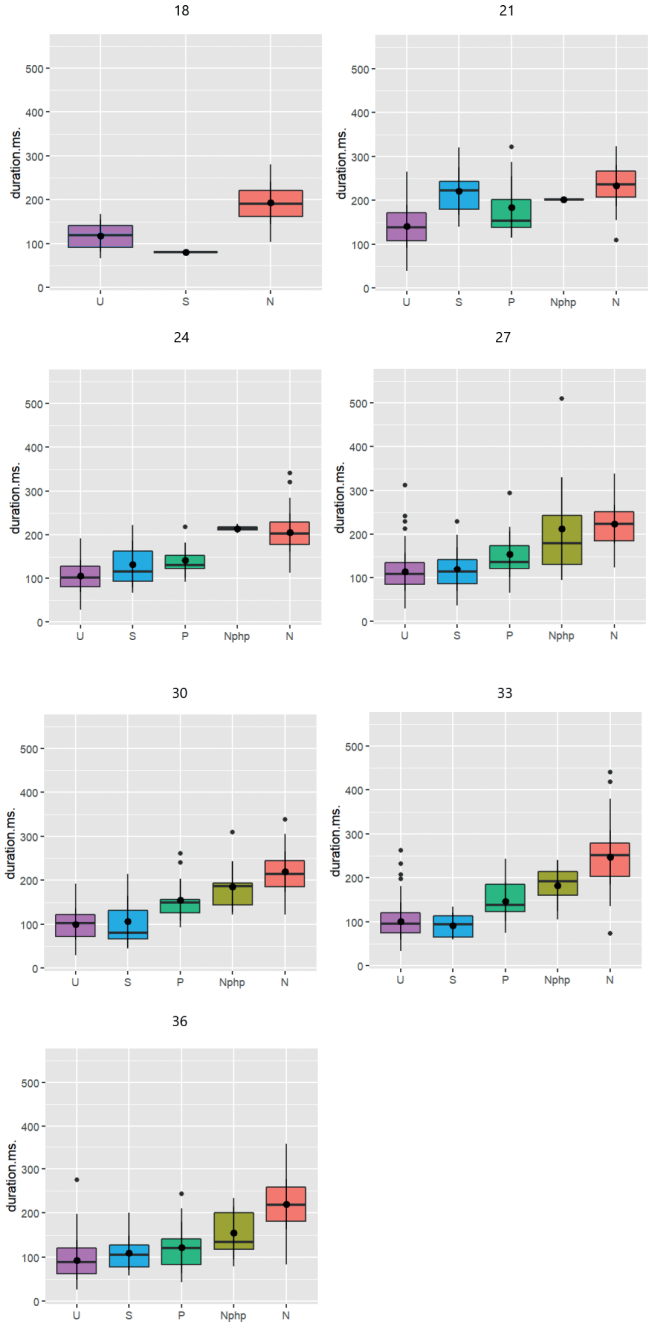
(10) tu sai suonare il pianoforte?
(10b) Son [PIU' GRANDE]$_{ContrFoc}$ [sa]?

A higher number of nuclear accents in intermediate phrases is observed (10 cases), related to a more frequent use of prosodic phrasing in the utterance. Utterances with coordinated NPs foster the placement of a prosodic boundary after the first NP. Moreover, the child's new syntactic achievement, i.e. the production of sentences composed of a main and a subordinate clause, fosters the presence of a prosodic boundary at the major syntactic boundary. In (11), for example, a boundary tone is placed between the two clauses "sta cucinando" and "per andare a casa". Each clause is consequently wrapped in an intermediate phrase: "nan" is the head of the first ip and constitues a case of ip nuclear accent (Nphp); "ca" is the head of the second ip but at the same time it is the head of the IP that wraps the whole utterance and therefore constitues a case of IP nuclear accent (N).

(11) [[sta cuci**nan**do$_{Nphp}$]$_{ip}$ [per andare a **ca**sa$_{N}$]$_{ip}$]$_{IP}$

Vowels' duration begins to be differentiated postlexically: while untressed (M = 113.54, SD = 43.70) and stressed vowels (M = 120.25, SD = 50.44) are of equal duration, a progression in duration is observable from prenuclearly accented (M = 154.65, SD = 53.47) to IP-nuclearly accented (M = 212.60, SD = 118.35). The median distance from U to N vowels rises to 116 ms.

Figure 3 - *Box-plot of the duration of the syllable in milliseconds (split for the five vowels of interest U, S, P, Nphp and N) according to Age (18, 21, 24, 27, 30, 33 and 36 months). Lower and upper box boundaries represent the 25th and 75th percentile, respectively; the line inside box represents the median; lower and upper error lines represent the 10th and 90th percentile, respectively; filled circles represent data falling outside the 10th and 90th percentile*



30 months. The child continues with his production of complex NPs, VO and SVO declarative sentences, sentences with a main and a subordinate clause, questions and exclamatives. Examples of sentence-initial focus attested in this recording session are the following:

(13) [[una **mac**china]$_{Foc}$ [voglio]]

(14a) C: Senti, conosci il Libro della Giungla?
(14b) BS: [[Ba**lù**]$_{Foc}$ [sè]]$_{IP}$ [[Ba**lù**]]$_{IP}$

The eligible number of vowel tokens in this recording session is less than in the previous one: 197. As for their duration, unstressed (n = 91, M = 100.63, SD = 35.32) and stressed vowels (n = 11, M = 106.45, SD = 58.64) are still undifferentiated, while the progressively higher average duration of P (n = 13, M = 155.46, SD = 47.44), Nphp (n = 8, M = 185.62, SD = 57.42) and N (n = 74, M = 219.31, SD = 47.70) does not show any appreciable difference with respect to the durations of the 27-month stage. Also the median difference between U and N vowels shows no appreciable variation (111.5 ms).

Generally, in the period ranging from 27 to 30 months of age, the child does not show any significant variation in his linguistic and prosodic development.

33 months. In the recording session at 33 months of age, the child talks more than in the preceding session, and the number of eligible vowels is 249. The type of utterances he produces is in line with those of the 30 and 27 months: noun phrases, sentences with direct and indirect arguments, exclamatives and sentences with sentence-initial information focus.

The average duration of unstressed (n = 120, M = 101.77, SD = 41.71) and stressed vowels (n = 9, M = 92.22, SD = 28.81) is still undifferentiated, but the duration of vowels in IP nuclear position (n = 95, M = 247.32, SD = 61.26) is longer than in the 30 months recording. The median difference between the duration of U and N vowels increases at 156 ms.

36 months. At 36 months the exchange with the clinician is richer and the number of sentences increases. In (15) and (16) we report two examples, with the indication of the focal structure, the prosodic phrasing and the tonal structure of each child's utterance. Pitch accents (H*, L+H*) refer to the nuclear accent (in bold) and boundary tones (H-, L-, LL%) refer to the following prosodic boundary.

(15) C: cosa hai fatto ieri?
BS: [mi hanno fatto la pun**tu**ra]$_{Broad\ Focus}$ H* H- [tutta tutta]
C: la puntura? Che puntura te ga fatto?
BS: quella [della **me**dica]$_{Narrow\ Information\ Focus}$ L+H* LL%
C: quella...??
BS: quella [della **me**dica]$_{Narrow\ Information\ Focus}$ H*+L LL%
C: medica?? E la puntura per cosa servi?
BS: [la **me**dica]$_{Narrow\ Information\ Focus}$ L+H* L- [quella quella che ho avuto dopo]
C: ahh! La vaccinazione te ga fatto!

(16) C: cosa è questo qua?
BS: [un **to**po]$_{\text{Broad Focus}}$ H* LL%
C: e chi va a mangiare i topi?
BS: [i **gat**ti]$_{\text{Information Focus}}$ L+H* L- [va a mangià i **to**pi]$_{\text{Background}}$ L* LL%

The number of eligible vowels is 187. At this age, it appears that the child has acquired the control of vowels' duration according to their role in the hierarchy of prosodic domains. As in adult's speech, BS shows a steady progression in vowels' duration from weak positions (Unstressed, n = 72, M = 93.69, SD = 44.30) to strong positions in Words (n = 22, M = 109.73, SD = 38.20), Intermediate Phrases (n = 7, M = 122.58, SD = 58.1) and Intonational Phrases (n = 74, M = 154.71, SD = 50.09).

4.1 Statistical modelling

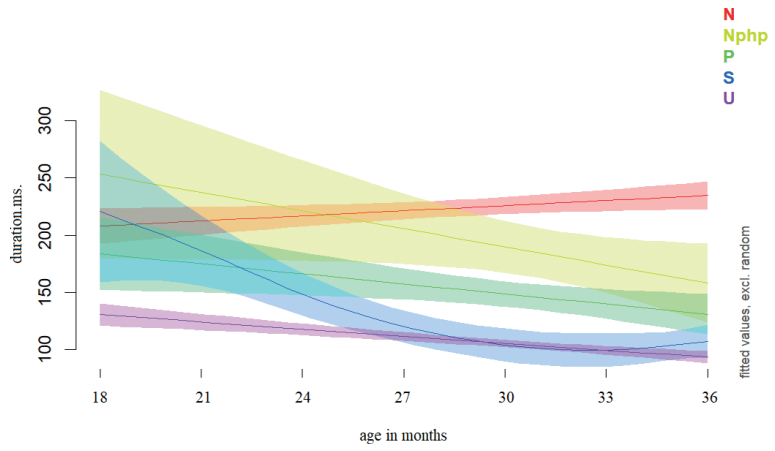
Data were analyzed with the R software (R Core Team, 2020) using generalized additive mixed-effects models (GAMMs) with the *mgcv package* (version 1.8-38, Wood, Wood, 2015). GAMMs approaches allow to model data for random and fixed effects as a function of time. Of note, those methods fit with multiple factors and unbalanced data sets, which are the norm rather than the exception in early child research, and stabilize the estimation of the parameters under investigation (Bates, 2010). In addition, those approaches are indicated to analyze longitudinal n-of-1 or single-subject study (idiographic) in which the target population consists of some larger set of time periods within a person's life (i.e., population-of-one studies (Daza, 2019), as in the case of the present study. Moreover, the mixed effect approach to statistical modelling allows to select a theoretical family distribution that better fits the empirical distribution of residuals. This in turn reduces the violation of the statistical assumptions, usually encountered with traditional approaches e.g. ANOVA, and the likelihood of obtaining false positive results (Boisgontier, Cheval, 2016). Accordingly, in the Statistical Appendix Fig. a and b, respectively, show the good fit of the Gamma (vs Lognormal) theoretical distribution family with the empirical distribution.

To find the best approximation to the true model, we followed a model comparison approach with AIC (Akaike Information Criterion) and AIC weight as indexes of goodness of fit. The AIC and AIC weight compare all the models at once and give information on a model's relative evidence (i.e., likelihood and parsimony), so that the model with the lowest AIC and the highest AIC weight is to be preferred (Wagenmakers, Farrell, 2004). We started from the simplest model with only random factors and proceeded by adding predictors. Specifically, we tested two models: i) the null model that uniquely includes smoothing components throughout Age in months and a random slope for Vowel type (U, S, P, Nphp and N); ii) the model with smoothing components throughout Age in months and a random slope for vowel type and with the interaction between Age in months and the type of Vowels i.e. U, S, P, Nphp and N.

The model comparison indicates that the last model with the interaction Age x Vowels better approximates the observed duration in milliseconds (dAIC = 0, AIC weight = 1, R-sq.(adj) = -0.123, Deviance explained = 45.2%) compared with the null model (dAIC = 169.6, AIC weight = 0, R-sq.(adj) = 0.562, Deviance explained = 51.7%). Following the best practice in order to correctly interprete GAMMs results (Van Rij et al., 2019), we visually inspected the model estimates between vowels. As a sanity check, Fig. c in the Statistical Appendix shows the autocorrelation plot of the selected model meeting the statistical assumption of a close to zero autocorrelation between the regressed variables. Fig. 4 shows how Age in months substantially interacts with the selected vowel tokens, i.e. U, S, P, Nphp and N by predicting statistically significant differences in duration.

On the one hand, the estimated smoothed effect indicates that U vowels predicted shorter duration compared to all the others since 18 months of age. On the other hand, N vowels predicted a significant increase in duration from 18 to 36 months of age compared to all the other vowels. The Nphp, P, and S vowels respectively, show a significant difference from 24 months and stay statistically different till 36 months of age. In addition, the Nphp, P and S vowels predicted an overall decrease in duration across time. Finally, the substantial effect indicating shorter duration for U vs S syllables disappeared around 27 months of age.

Figure 4 - *Partial effects (fixed effects only) of the initial GAMM showing the nonlinear regression lines for each of the five syllables (U, S, P Nphp, and N) with pointwise 95% confidence intervals*



5. *Discussion and conluding remarks*

Our aims at the start of this study were rather exploratory, being this the first study on the production of lexical and postlexical prominence in Italian children. Based on the results of our previous studies, we expected that the child we analyzed would distinguish between unstressed and stressed vowels since the earliest recording

stage, i.e. 18 months, and that along his linguistic and prosodic development vowels with different prominence degrees would emerge according to the acquisition of the utterance prosodic and syntactic organization.

At the 18-month stage, the child has not yet mastered the production of lexical stress: cases of stress inversion and equal prominence in disyllables, and weak syllable deletion in trisyllables occur. Weak syllable deletion in multisyllabic words is a well-known phonological process in acquisition, attested in many languages. Within a cognitive view of language acquisition, it has been formalized by proposing the existence of a shape constraint such that children's word productions conform to a consistent size and rhythmic pattern. Gerken (1994) formalized it as the output of the application of a metrical constraint known as "trochaic bias": children would align a trochaic (SW) metrical constraint at the beginning of an intended word and weak syllables that do not fit the template are omitted. In a Natural Phonology approach, weak syllable deletion is one of the natural processes which are systematically applied in speech production until children learn to suppress them. Within an alternative usage-based approach, cast in an emergentist framework which does not assume universal constraints on production (and perception), Vihman proposes that the acquisition of prosodic structure "appeals to learning based on both initial perceptual biases of possible evolutionary origin (...) and infants' experience, along with neurophysiological maturation, of vocal production practice (...)" (Vihman, 2018:185). Weak syllable deletion appears to be suppressed by 21 months of age.

At 18 months the child is still in the stage of one-word utterances. In a phrasal perspective, the only possible rythmic difference within one-word utterances is between unstressed and nuclearly accented vowels. Such opposition is clearly characterized by a durational difference: nuclear vowels are significantly longer than unstressed ones.

With the emergence of argument structure and multi-word utterances at 21 months of age, prosodic phrasing appears, and syllables endowed with stress, prenuclear accents, and IP-nuclear accents occur. However, the duration of the prominent vowels, although in the expected direction, is not significantly different between S, P and N. The only significant difference is still that between unstressed and nuclearly accented vowels.

A pivotal stage for this child prosodic development is represented by the 24-month recording: the model predicts that U, S, P, Nphp and N vowels are significantly different, but nuclear vowels in intonational phrases are shorter than in intermediate phrases.

At 27-months of age the prominence hierarchy appears to be in place, as in the adult speech, with increasing duration from U to N vowels.

But the system is not stable yet: at 30- and 33-month recordings, while the higher levels of prominence remain statistically different and IP-nuclear vowels increase in duration, unstressed and stressed vowels loose their distinction and become durationally indifferentiated. This in in line with our expectations that the process of stress/accent acquistion is nonlinear.

Finally, in the 36-month recording the child's production is adult-like: the distribution of prominences within the prosodic constituents is what is expected in adult speech, with a linear increase in the duration from unstressed, to stressed, prenuclear, ip-nuclear to IP-nuclear vowels.

As for the developmental trajectory of postlexical prominence, we had no predictions as to whether it would proceed linearly or not. That is, whether the duration of the strong vowels would lengthen in a linear fashion, progressively differentiating from unstressed vowels and from each other. Our results show that only the phonetic realization of the highest level of prominence proceeds linearly: the duration of IP-nuclear vowels steadily increases across months. At lexical level, also the phonetic realization of unstressed vowels evolves linearly, but in the opposite direction: their duration steadily decreases across months. Along with the progressive divergence in the duration of the lowest and highest degrees of prominence, in his prosodic development the child progressively adjusts the phonetic content (i.e. vowel duration) of the intervening levels of prominence. Two stages appear to be important: the one represented in the 24-month recording, where for the first time all the prominence levels are significantly different, but in which IP-nuclear vowels are shorter than ip-nuclear ones; and the 27-month recording, where the phonetic content is consistent with the place of the vowels in the prosodic hierarchy. At 3 years of age the process of prosodic prominence acquisition appears to be completed, after a temporary "regression" at 30- and 33-months, which is in line with the nature of speech acquisition.

In our future research we will examine the other acoustic correlates that cue prominence which have not been anayzed in the present paper, and will expand the present study to include the data up to 48 months of age. At the same time, we aim to analyse the other Italian children who have been recorded in the Bonifacio's database.

References

Arciuli, J., Colombo, L. (2016). An acoustic investigation of the developmental trajectory of lexical stress contrastivity in Italian. In *Speech Communication*, 80, 22-33.

Avesani C., Vayra M. (2013). Prosodic prominence and its articulatory bases. Poster presented at the International Conference "pS-prominenceS. Prominences in Linguistics", University of Tuscia (Viterbo, Italy), December 12-13, 2013.

Avesani C., Vayra M. & Zmarich C. (2009). Coordinazione vocale-consonante e prominenza accentuale in italiano. La sfida della Articulatory Phonology, in G. Ferrari, G., Benatti, R. & Mosca. M. (Eds.) (2006), *Linguistica e modelli tecnologici di ricerca.* Atti del XL Congresso Internazionale della SLI (Società di Linguistica Italiana) (Vercelli, 21-23 settembre 2006), Roma, Bulzoni, 365-399.

Bahatara, A., Boll-Avetisyan, B., Höhle, B. & Nazzi, T. (2018). Early sensitivity and acquisition of prosodic patterns at the lexical level. In Prieto, P., Esteve-Gibert, N. (Eds.), *The development of Prosody in First Language Acquisition,* Amsterdam/Philadelphia: John Benjamins, 38-57.

Bates, D.M. (2010). lme4: Mixed-effects modeling with R.

Beckman. M.E. (1996). The parsing of prosody. In *Language and Cognitive Processes*, 11, 17-68.

Beckman, M., Pierrehumbert, J. (1986). Intonational structure in English and Japanese. In *Phonology Yearbook*, 3, 255–310.

Bion, R.A.H., Benavides-Varela, S. & Nespor, M. (2011). Acoustic markers of prominence influence infants' and adults' segmentation of speech sequences. In *Language and Speech*, *54*(1), 123-140.

Bocci, G, Avesani, C. (2011). Phrasal prominences do not need pitch movements post-focal phrasal heads in Italian. In Cosi, P., De Mori, R., Di Fabbrizio, G. & Pieraccini, R. (Eds.), *Proceedings of Interspeech 2011*, Firenze, 27-31 August 2011, International Speech Communication Association, 1357-1360.

Boersma, P., weenink, D. (2021). PRAAT: doing phonetics by computer. [Computer program] Version 6.1.34. https://www.praat.org

Boisgontier, M.P. & Cheval, B. (2016). The anova to mixed model transition. In *Neuroscience & Biobehavioral Reviews*, 68, 1004-1005.

Caselli, M.C., Casadio, P. (1995). *Il Primo Vocabolario del Bambino*. Milano: Franco Angeli.

Chen, A., Esteve-Gibert, N., Prieto, P & Redford, M. (2021). Development of phrase-level prosody from infancy to late childhood. In Gussenhoven, C, Chen, A. (Eds.). *The Oxford handbook of language prosody*. Oxford: Oxford University Press, 553-562.

D'Odorico, L., Carubbi, S. (2003). Prosodic characteristics of early multi-word utterances in Italian children. In *First Language*, 23(1), 97-116.

D'Odorico, L., Fasolo, M. & Marchione, D. (2009). The prosody of early multi-word speech: word order and its intonational realization in the speech of Italian children. In *Enfance*, 3, 319-327.

D'Odorico, L., Fasolo, M. & Zanchi, P. (2010). Prosodic characteristics of multi-argument utterances in Italian children. In *Child Language Seminar*, London, UK.

Daza, E.J. (2019). Person as Population: A Longitudinal View of Single-Subject Causal Inference for Analyzing Self-Tracked Health Data. *arXiv preprint arXiv:1901.03423*.

DeCasper, A.J., Fifer, W.P. (1980). Of human bonding: Newborns prefer their mothers' voices. In *Science,* 208(4448), 1174-1176.

de Carvalho, A., He, A.X., Lidz, J. & Christophe, A. (2015). 18-month-olds use phrasal prosody as a cue to constrain the acquisition of novel word meanings. Paper presented at the *Boston University Conference on Language Development*, Boston.

de Carvalho, A., Lidz, J., Tieu, L., Bleam,T. & Christophe, A. (2016). English-speaking preschoolers can use phrasal prosody for syntactic parsing. In *Journal of the Acoustical Society of America,* 139(6), EL 216-EL 222.

de Carvalho, A., Dautriche, I., Lin, J. & Christophe, A. (2017). Phrasal prosody constrains syntactic analysis in toddlers. In *Cognition* 163, 63-79

Esteve-Gibert, N. Prieto, P. (2018), Early development of the prosody-meaning Interface. In Prieto, P. and Esteve-Gibert, N. (a cura di), *The development of Prosody in First Language Acquisition,* Amsterdam/Philadelphia: John Benjamins, 227-246.

Farnetani, E., Kori, S. (1982). Lexical stress in spoken sentences: a study on duration and vowel formant pattern. In *Quaderni del Centro di Studio per le Ricerche di Fonetica del CNR*, 1,106-133.

Farnetani, E., Faber, A. (1992). Tongue-jaw coordination in vowel production: isolated words versus connected speech. In *Speech Communication*, 11, 401-410

Frota, S., M. Cruz, N. Matos & M. Vigário. (2016). Early Prosodic Development: Emerging intonation and phrasing in European Portuguese. In Armstrong, M.E., N. Henriksen, N. & Vanrell, M.M. (Eds.), *Intonational grammar in Ibero-Romance: Approaches across linguistic subfields*. Amsterdam: Benjamins, 295-324

Gerken, L. (1994). A metrical template account of children's weak syllable omissions from multisyllabic words. In *Journal of Child Language*, 21, 565–584

Gerken, L., P., Jusczyk, W. & Mandel, D.R. (1994). When prosody fails to cue syntactic structure: 9-month olds' sensitivity to phonological vs syntactic phrases. In *Cognition*, 51, 237-265.

Gervain, J. (2015) Plasticity in early language acquisition: the effects of prenatal and early childhood experience. In *Current Opinion in Neurobiology*, 35, 13-20.

Gervain, J., Cristophe, A & Mazuka R. (2021). Prosodic bootstrapping. In Gussenhoven, C, Chen, A. (Eds.), *The Oxford handbook of language prosody*. Oxford: Oxford University Press, 563-573.

Hirsh-Pasek, K., Kemler Nelson, D.G., Jusczyk, P.W., Cassidy, K.W., Druss, B. & Kennedy, L. (1987). Clauses are perceptual units for young infants. In *Cognition*, 26, 269–286.

Jessen, M., Marasek,K. (1997). Voice quality correlates of word stress and tense versulìs lax vowels in German. In *Larynx*, Marseille, France, June 16-18, 127-130.

Jusczyk, P.W., Friederici, A.D., Wessels, J.M., Svenkerud, V.Y. & Jusczyk, A.M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language, 32*(3), 402–420.

Jusczyk, P., Hirsch-Pasek, K., Kemler Nelson, D., Kennedy, L., Woodward, A. & Piwoz, J. (1992). Perception of acoustic correlates of major phrasal units by young infants, *Cognitive Psychology*, 24, 252-293.

Kehoe, M., Stoel-Gammon, C. & Buder, E.H. (1995). Acoustic correlates of stress in young children's speech. In *Journal of Speech, Language and Hearing Research*, 38(2), 338-350.

Ladd, R. (2008[2]). *Intonational Phonology*. Cambridge: Cambridge University Press.

Ling, D. (1976). *Speech and the hearing-impaired child: Theory and practice. Washington,* DC: Alexander Graham Bell Association for the Deaf.

Magno Caldognetto, E., Vagges, K. & Zmarich, C. (1995). Visible articulatory characteristics of the Italian stressed and unstressed vowels. *In Proceedings of the XIIIth International Congress of Phonetic Science*s, 1, 366-369.

Moon, C., Cooper, R.P. & Fifer, W. (1993). Two-day-olds prefer their native language. In *Infant Behavior and Development*, 16, 495-500.

Moon, C., Lagercrantz H. & Kuhl P. (2013). Language experience *in utero* affects vowel perception after birth: a two-country study. In *Acta Pediatrica*, 101(2), 156-160.

Nespor, M., Vogel, I. (1986). *Prosodic phonology*. Dordrecht: Foris. Berlin: Mouton de Gruyter.

Olivucci, F., Pasqualetto, F., Vayra, M. & Zmarich C. (2016). Lo sviluppo dell'accento lessicale nel bambino in età prescolare: una prospettiva fonetico-acustica. In Savy, R. and Alfano, I. (Eds.). *La fonetica nell'appendimento delle lingue / Phonetics and Language Learning*, *Studi AISV 6,* Milano: AISV, 219-228.

Olivucci, F., Vayra, M., Avesani, C. & Zmarich (2018). Acoustic correlates of word stress in young Italian children's productions. Presented at the *40th Annual Conference of the German Linguistics Society*, Stuttgart, March 9 2018.

Olivucci, F, Vayra, M., Avesani, C. & Zmarich (2019). The development of lexical stress in young Italian children. Poster presented at the *2019 Conference on Phonetics and Phonology in Europe (PaPE 2019)*, Lecce, June 17-19 2019.

Pollock, K.E., Brammer, D.M. & Hageman, C.F. (1993). An acoustic analysis of young children's productions of word stress. In *Journal of Phonetics*, 21,183-203.

Prieto, P., Estrella, A., Thorson, J. & Vanrell, M.M. (2012). Is prosodic development correlated with grammatical development? Evidence from emerging intonation in Catalan and Spanish. In *Journal of Child Language*, *39*(2), 221-257.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Sansavini, A., Bertoncini, J. & Giovannelli, G. (1997). Newborns discriminate the rhythm of multisyllabic stressed words. In *Developmental Psychology*, 33 (1), 3-11.

Savy, R., Cutugno F. (1997). Ipoarticolazione, riduzione vocalica, centralizzazione: come interagiscono nella variazione diafasica. In Cutugno, F. (Ed.) *Fonetica e fonologia degli stili dell'italiano parlato.* Atti VII Giornate di Studio del GFS (Napoli, 14-15 novembre 1996), Roma: Esagrafica, 177-194.

Schmitt, J.F., Meline, T.J. (1990). Subject descriptions, control groups, and research designs in published studies of language-impaired children. In *Journal of Communication* Disorders, 23(6), 365-382.

Schwartz, R.G., Petinou, K., Goffman, L., Lazawski, G. & Cartusciello, C. (1996). Young children's production of syllable stress: An acoustic analysis. In *The Journal of the Acoustical Society of America*, 99(5), 3192-3200.

Selkirk, E. (1984). *Prosody and syntax: The relation between sound and structure*. Cambridge, MA: MIT Press.

Shukla. M., Wite, K.S. & Aslin, R.N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. In *Proceedings of the National Academy of Sciences of the United States of America,* 108 (15), 6038-6043.

Slujiter, A.M., Van Heuven, V.J. (1996). Spectral balance as an acoustic correlate of linguistic stress. In *The Journal of the Acoustical society of America*, 100(4), 2471-2485.

Tamburini, F. (2009). Prominenza frasale e tipologia prosodica: un approccio acustico. In Ferrari G. (Ed.) *Linguistica e modelli tecnologici di ricerca.* Atti del XL Congresso internazionale di studi della Società di linguistica italiana (SLI) (Vercelli, 21-23 settembre 2006), Roma: Bulzoni, 437-455

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R.H. & Wood, S.N. (2019). Analyzing the time course of pupillometric data. In *Trends in Hearing*, 23, 2331216519832483. https://doi.org/10.1177/2331216519832483

Vayra, M. (1991). Appunti su un effetto di 'centralizzazione' nel vocalismo dell'italiano standard. In L. Giannelli, N. Maraschio, T. Poggi Salani & M. Vedovelli (Eds), *Tra Rinascimento e strutture attuali della lingua. Atti del I Convegno Internazionale della S.I.L.F.I.* (Siena, 28-31 March 1989), Torino, Rosenberg & Sellier, 195-212.

Vayra, M., Fowler, C. (1987). The word-level interplay of stress, coarticulation, vowel height and vowel position in Italian. In *Proceedings of the XIth International Congress of Phonetic Sciences*, (Tallinn, 1-7 August 1987), 4, 24-27.

Vayra, M., Fowler, C. (1992). Declination of supralaryngeal gestures in spoken Italian. In *Phonetica*, 49(1), 48-60.

Vayra, M., Avesani, C. & Fowler, C.., C. (1999). On the phonetic bases of vowel-consonant cooordination in Italian: a study of stress and compensatory shortening. In *Proceedings of 14th ICPhS* (San Francisco, USA, 1-7 August 1999), 495-498.

Vihman, M. (2018). The development of prosodic structure. A usage.-based approach. In Prieto, P. and Esteve-Gibert, N. (a cura di), *The development of Prosody in First Language Acquisition,* Amsterdam/Philadelphia: John Benjamins, 185-206.

Wagenmakers, E.-J., Farrell, S. (2004). AIC model selection using Akaike weights. In *Psychonomic Bulletin & Review*, *11*(1), 192–196. https://doi.org/10.3758/BF03206482

Wood, S., Wood, M.S. (2015). Package 'mgcv'. R package version, 1, 29.

Zanchi, P., D'Imperio, M.P., Zampini, L. & Fasolo, M. (2016). L'intonazione delle narrazioni di bambini e adulti italiani. In Savy, R. and Alfano, I. (a cura di). *La fonetica nell'appendimento delle lingue / Phonetics and Language Learning*, *Studi AISV 6,* Milano: AISV, 179-189.

Zmarich C., Avesani C. (2015), L'influenza della durata consonantica sulla coarticolazione della sillaba CV con gradi diversi di prominenza prosodica. In A. Romano, M. Rivoira, I. Meandri (a cura di), *Aspetti prosodici e testuali del raccontare: dalla letteratura orale al parlato dei media*, Alessandria, Edizioni dell'Orso, 305-318.

Appendix

Figure a - *Density distribution plot of the overall duration residual included in the statistical modelling. The vertical line indicates the median value (144) of the whole distribution of vowel duration in milliseconds*

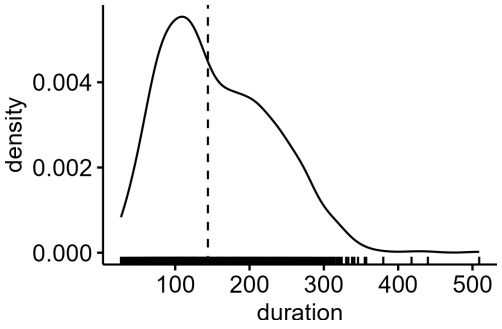


Figure b - *Selection of the best fitting family (Gamma vs Lognormal) distribution for the positive shewed distribution of the duration residuals. The Gamma family emerged to adequately fit the data*

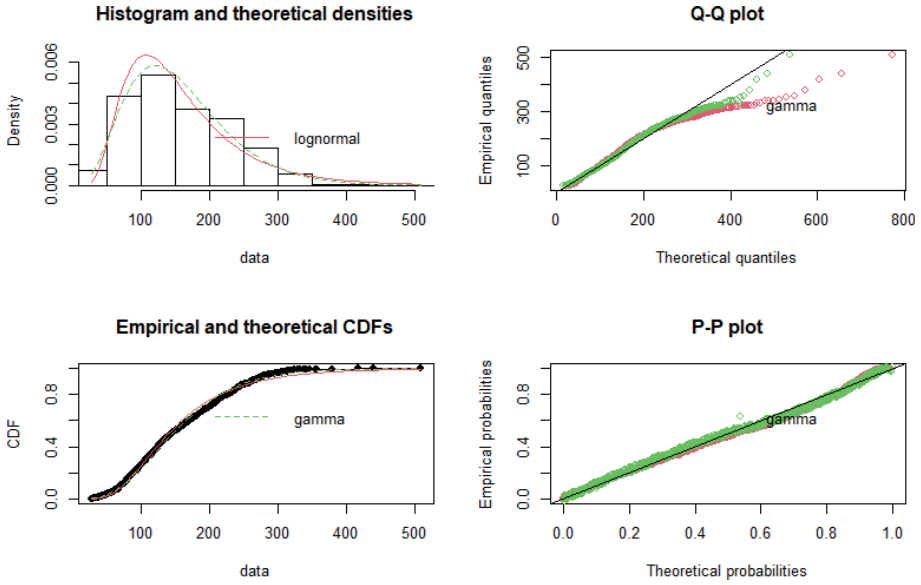
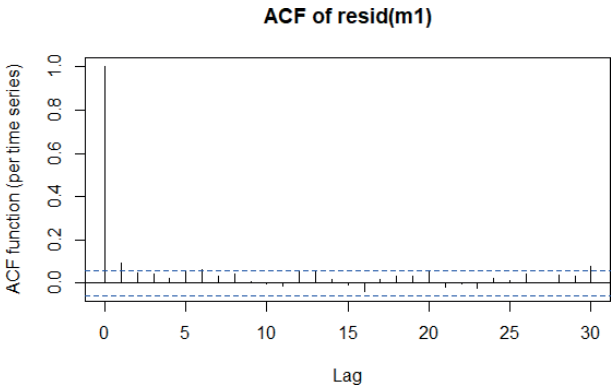


Figure c - *Autocorrelation plot showing an adequate low level of autocorrelation between Age (in months) and vowel duration (in milliseconds) as estimated by the generalized additive model (GAMM)*



DOI: 10.17469/O2109AISV000002

FRANCESCO CANGEMI, DALILA DIPINO, DAVIDE GARASSINO,
STEPHAN SCHMID

Raccontare la complessità. Correlati fonetici della complessità narrativa in un corpus di narrazioni orali in tedesco standard svizzero[1]

Narrating complexity. Phonetic correlates of narrative complexity in a corpus of Swiss Standard German storytelling

We carried out an exploratory study of narrative complexity, focussing on coarse phonetic measures such as the duration and number of interpausal units. Our corpus features picture-based narratives produced by twenty German speakers from Switzerland. The task was designed to elicit simple and complex narratives, depending on the order of appearance of the events and the number of characters in the pictures. Our results show that complex stories have longer overall duration and higher number of interpausal units. A closer look reveals that most of the additional interpausal units in complex stories are short in duration and contain few syllables. Despite inter-speaker variation, this trend is also confirmed at the individual level. In our interpretation, even coarse quantitative phonetic metrics suggest that narrative complexity results not only into more material (i.e. duration), but also into less cohesion (i.e. fragmentation).

Keywords: storytelling, narrative complexity, fragmentation, interpausal units, data visualization.

1. *Introduzione*

Trasformare eventi in parole, come avviene nella narrazione, è sempre un'operazione delicata. E quando gli eventi sono particolarmente complessi, la loro narrazione tende a farsi essa stessa complessa. Nella prefazione a *La cognizione del dolore*, ad esempio, C.E. Gadda presenta la natura barocca della sua scrittura come un riflesso della vicenda da narrare:

> Ma il barocco e il grottesco albergano già nelle cose, nelle singole trovate di una fenomenologia a noi esterna: nelle stesse espressioni del costume, nella nozione accettata «comunemente» dai pochi o dai molti: e nelle lettere, umane o disumane che siano: grottesco e barocco non ascrivibili a una premeditata volontà o tendenza

[1] L'impianto teorico e metodologico di questo studio è opera del primo autore. Il quarto autore ha gestito la raccolta dei dati. La segmentazione e l'annotazione dei file sonori sono opera della seconda autrice e del terzo autore. La prima stesura, l'analisi e la visualizzazione dei dati sono state effettuate dal primo autore. Le osservazioni finali sono frutto della discussione tra tutti gli autori, che hanno rivisto insieme il testo e approvato l'ultima versione.

> espressiva dell'autore, ma legati alla natura e alla storia [...] talché il grido-parola d'ordine "barocco è il G.!" potrebbe commutarsi nel più ragionevole e pacato asserto "barocco è il mondo, e il G. ne ha percepito e ritratto la baroccaggine" (Gadda, 1963: 32).

Il legame tra complessità del mondo e complessità della narrazione ritorna nella prefazione a *Il manoscritto di Brodie* di J.L. Borges:

> Ho cercato, non so con quanto successo, di redigere racconti lineari. Non mi azzarderò a dire che sono semplici; sulla terra non c'è una sola pagina, una sola parola che lo sia, giacché tutte postulano l'universo, il cui attributo più noto è la complessità (Borges, 1970: 11).

La complessità della narrazione non è un tema di sola pertinenza della letteratura, ma può essere esaminata da molteplici punti di vista, da quello sociale a quello cognitivo (Walsh, Stepney, 2018; Grishakova, Poulaki, 2019; Abbott, 2021). Gli studi in merito si concentrano su materiali di genere assai diverso, dalle serie televisive alle campagne d'informazione nel campo della sanità pubblica (Mittell, 2006; Gubrium, Gubrium, 2021). In ambito linguistico, la complessità narrativa viene studiata soprattutto in chiave acquisizionale (Pallotti, 2015), servendo spesso come indice per l'apprendimento di competenze linguistiche in lingue seconde. Tradizionalmente, particolare attenzione è stata dedicata alla dimensione sintattica (Ortega, 2003), a quella morfologica (Bulté, Housen, 2012) e a quella lessicale (McCarthy, Jarvis 2010), mentre lavori recenti cominciano ad occuparsi anche della dimensione semantico-discorsiva (Ryshina-Pankova, 2015). Secondo gli approcci più praticati, la complessità può essere catturata da formule riconducibili a tre tipi di metriche: lunghezza (e.g. delle unità discorsive), subordinazione (e.g. nelle unità sintattiche) e frequenza (e.g. di forme morfologicamente o lessicalmente rare). Simili metriche vengono impiegate nella valutazione automatica della leggibilità dei testi (Dell'Orletta, Montemagni & Venturi, 2011) ed in applicazioni commerciali per la semplificazione di testi scritti, e riscontrano particolare interesse in ambito legale (Kimble, 1992). Tuttavia, questo approccio è stato criticato come eccessivamente riduzionista da Ortega (2012), che invita a modulare il concetto di complessità in base alle competenze linguistiche del soggetto (e.g. apprendenti principianti, avanzati, bilingui) e al tipo di attività linguistica osservata (e.g. scritta, parlata, spontanea, elicitata).

In questo studio offriamo un'esplorazione della complessità narrativa dal punto di vista fonetico, comparativamente poco rappresentato nella letteratura sul tema. Nel tentativo di ridurre le dimensioni di variazione appena citate, ci limiteremo ad analizzare le produzioni di soggetti madrelingua, in modalità parlata (nello specifico: nella varietà standard del tedesco parlato in Svizzera), ed elicitate attraverso il compito sperimentale dettagliato al § 2. Nello stesso paragrafo esporremo i criteri impiegati per l'annotazione dei dati e le metriche estratte dalle annotazioni. Nel paragrafo successivo (§ 3) presenteremo i risultati basati su lunghezza delle narrazioni, numero di unità interpausali e numero di sillabe per unità, con particolare attenzione alla variabilità dei risultati in chiave individuale. Al § 4, infine, offriremo

delle brevi riflessioni conclusive sull'equilibrio tra dettaglio e sintesi nello studio della complessità narrativa.

Tutti i materiali alla base di questo contributo (le illustrazioni usate per l'elicitazione delle narrative, i metadati relativi ai parlanti, i file audio, le annotazioni, gli script per l'estrazione e la visualizzazione dei dati) sono raccolti in un Archivio accessibile all'indirizzo *osf.io/ufdr9*.

2. *Metodo*

2.1 Preparazione

Le narrazioni usate in questo studio sono state elicitate attraverso un paradigma sperimentale mutuato da Fossard, Achim, Rousier-Vercruyssen, Gonzalez, Bureau & Champagne-Lavau (2018). Il soggetto e un assistente di ricerca siedono alle estremità opposte di un tavolo, separati da uno schermo che impedisce la vista reciproca. Ai partecipanti viene consegnata una tavola composta da sei vignette numerate. Anche l'assistente riceve le stesse immagini, ritagliate però in vignette individuali, prive di numerazione e disposte in maniera aleatoria. Il compito del soggetto è di narrare la storia rappresentata dalle sei vignette, in modo da permettere all'assistente di ricostruirne l'ordine. Al termine della narrazione, l'assistente solleva lo schermo e chiede al soggetto di confermare l'accuratezza della ricostruzione. L'esperimento originale prevede la narrazione di 18 storie in totale.

Crucialmente per i nostri fini, le storie contengono diversi gradi di complessità. Alcune tavole presentano vicende in cui le vignette seguono un ragionevole ordine cronologico. Ad esempio, nella Fig. 1 (in alto) è mostrata la storia di un uomo che fa compere al supermercato; la storia inizia con il recupero di un carrello all'entrata (vignetta 1) e termina con l'attesa in coda per il pagamento (vignetta 6).

Altre tavole sono composte da vignette che non seguono l'ordine cronologico. Ad esempio, la Fig. 1 (in basso) mostra una donna e una bambina che preparano un dolce ma, in maniera inattesa, la torta viene estratta dalla teglia (vignetta 3) prima di essere infornata (vignetta 4). Questi materiali presentano un'ulteriore importante differenza: alcune storie includono un unico personaggio (Fig. 1, in alto), mentre altre presentano due personaggi dello stesso sesso (Fig. 1, in basso). In quest'ultimo caso la gestione delle catene anaforiche è più difficile, dal momento che è impossibile disambiguare i due personaggi utilizzando solo un pronome.

Figura 1 - *Esempio di storia Semplice (in alto) e Complessa (in basso), da Fossard et al. (2018)*

Lo studio originale di Fossard et al. (2018) prevede anche la narrazione di storie con un livello intermedio di difficoltà anaforica, ovvero contenenti due personaggi di sesso diverso. Per ognuna delle 3 condizioni di personaggi (un solo referente, due di sesso diverso, due di sesso uguale) e per ognuna delle 2 condizioni logiche (ordine cronologico, ordine non cronologico), lo studio di Fossard et al. (2018) offre 3 diverse storie, per un totale di 3 * 2 * 3 = 18 storie. Le nove storie in ordine cronologico vengono introdotte in un primo blocco, mentre le nove storie in ordine non cronologico vengono presentate nella seconda metà dell'esperimento; inoltre, il primo blocco si apre sempre con una storia con personaggio singolo. Nell'intenzione dello studio originale, queste misure servono a familiarizzare il parlante con il compito sperimentale. Pur avendo raccolto dati per tutte le storie, in questo nostro lavoro ci limitiamo ad analizzare le 6 storie massimamente divergenti, ovvero le 3 storie

più Semplici, con un unico personaggio e vignette disposte secondo un ordine cronologico (e.g. Fig. 1, in alto) e le 3 storie più Complesse, con due personaggi dello stesso sesso e una disposizione delle vignette in ordine non cronologico (e.g. Fig. 1, in basso). Le tavole relative alle 6 storie utilizzate in questo lavoro sono disponibili nella cartella *Sources* dell'Archivio.

2.2 Raccolta

In questo studio analizziamo 6 storie narrate in tedesco standard da 20 parlanti svizzeri, equamente divisi per genere, di età compresa tra 21 e 64 anni (età media: 28.3, deviazione standard: 9.9), e cresciuti prevalentemente nel Canton Zurigo (le 5 eccezioni provengono dai cantoni Argovia, Lucerna, Turgovia, San Gallo e Vallese). I metadati dettagliati, opportunamente anonimizzati, sono disponibili nella cartella *Audio* dell'Archivio. Le registrazioni hanno avuto luogo al Laboratorio di Fonetica dell'Università di Zurigo nella primavera del 2015; sono stati utilizzati un registratore digitale Fostex FR-2LE e due microfoni a cravatta Sennheiser MKE-2-P-C (gamma di frequenza 20–20.000 Hz ±3dB, coefficiente di trasmissione a vuoto 10 mV/Pa ±2.5 dB). Le narrazioni sono state registrate in formato stereo, con una frequenza di campionamento di 44.100 Hz, e salvate come file .wav. Come indicato sopra, queste 120 narrazioni rappresentano un sottoinsieme del corpus registrato, composto nella sua interezza da tutte le 18 storie dello studio di Fossard et al. (2018), narrate da 30 parlanti di svizzero tedesco, per un totale di 540 storie.

A ogni file sonoro è stato assegnato un codice composto di due cifre identificative del parlante (da 01 a 30), una lettera per il tipo di storia (L per quelle in ordine cronologico, N per quelle in ordine non cronologico) nonché una cifra per il grado di complessità anaforica (1 per personaggio singolo, 2 per due personaggi di sesso diverso, 3 per due personaggi dello stesso sesso). Per fare un esempio: le tre storie Semplici corrispondono al codice L1 e le tre storie Complesse al codice N3. Abbiamo inoltre aggiunto un carattere per identificare la storia all'interno di questi due gruppi (A, B o C) e un numero per l'ordine di presentazione all'interno dell'esperimento (da 1 a 9, separatamente per i due blocchi).

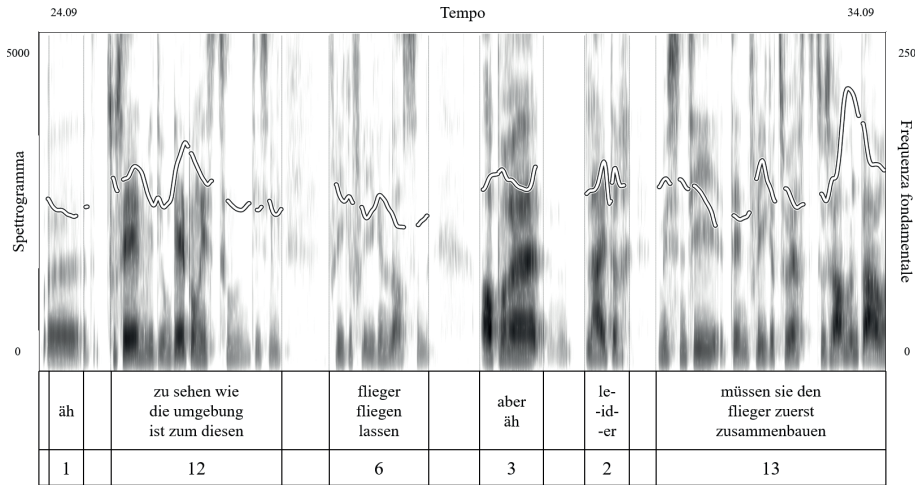
Per ogni registrazione, la porzione relativa ad ognuna delle 6 storie d'interesse è stata esportata in un file audio individuale. Ad esclusione di sporadici segnali non-verbali di assenso, comprensione od incoraggiamento, le tracce audio relative all'assistente di ricerca sono silenti durante le narrazioni. I file audio sono disponibili su richiesta all'indirizzo indicato nella cartella *Audio* dell'Archivio. Il corpus finale contiene 120 file audio (20 parlanti * 2 livelli di complessità * 3 storie) per un totale di circa 2 ore di narrazione.

2.3 Annotazione

Una volta estratte, le narrazioni prodotte dai soggetti sono state segmentate in unità interpausali e annotate su appositi TextGrid nel software *Praat* (Boersma, Weenink, 2022). Utilizzando lo script *A* disponibile nell'Archivio, le unità interpausali sono state combinate in modo da essere precedute e seguite da silenzi di almeno 200 ms

(Duez, 1982, Campione, Véronis, 2002). Pause piene, allungamenti vocalici e altri segnali di esitazione sono stati considerati come appartenenti alla porzione parlata. Ulteriori segnali non verbali come schiarimenti di gola e colpi di tosse sono stati invece considerati parte della porzione silente, data la loro minore rilevanza linguistica in questo contesto essenzialmente monologico. Per le unità interpausali si è fornita una trascrizione ortografica semplificata, utilizzando <äh> o <ähm> per le pause piene e <...> per gli allungamenti. Questa segmentazione è stata quindi duplicata in un secondo livello di annotazione, in cui le trascrizioni ortografiche sono state sostituite dal numero di sillabe percepite dall'annotatrice (di madrelingua tedesca), dunque tenendo conto di possibili fenomeni di riduzione. La Fig. 2 mostra un esempio con 9 secondi di parlato tratto dalla narrazione di una storia Complessa prodotta da un parlante maschio giovane.

Figura 2 - *Esempio di annotazione con segmentazione in unità interpausali, trascrizione ortografica e conteggio delle sillabe (storia N3C, parlante M14)*



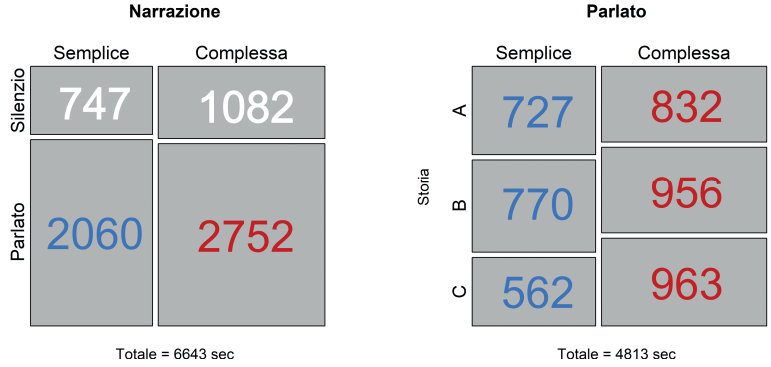
Attraverso lo script *B* disponibile nell'Archivio sono state estratte le informazioni relative ad ogni unità interpausale: posizione nella narrazione, durata in millisecondi e numero di sillabe, nonché la trascrizione ortografica. Ad esclusione dell'ultima unità di ogni narrazione abbiamo estratto anche la durata del silenzio immediatamente successivo. Utilizzando lo script in *R* (R Core Team, 2022) disponibile nell'Archivio, la tabella è stata riorganizzata aggregando i dati per storia e per parlante, e quindi impiegata per la rappresentazione dei risultati in forma grafica.

3. *Risultati*

L'intero corpus ha una durata di 6.634 secondi. La Fig. 3 (pannello di sinistra) mostra che, compatibilmente con l'intuizione menzionata nel § 1, le storie Complesse hanno tempi di narrazione più lunghi, ricoprendo circa il 58% della durata totale.

La maggiore durata delle narrazioni Complesse riguarda in maniera lineare sia la porzione parlata che quella silente. Infatti, indipendentemente dal tipo di storia, le narrazioni presentano una percentuale simile di Silenzio, intorno al 27%. Nel seguito analizziamo in dettaglio la porzione parlata, valutando l'effetto della complessità al livello delle singole storie (§ 3.1), delle singole unità interpausali (§ 3.2) e dei singoli parlanti (§ 3.3).

Figura 3 - *Durate arrotondate al secondo e aggregate su tutti i parlanti. Pannello di sinistra: durata totale delle narrazioni (comprensiva di silenzi), separatamente per le storie Semplici (colonna a sinistra) e Complesse (colonna a destra), e separatamente per Silenzio (riga in alto, numeri in bianco) e Parlato (riga in basso, numeri in blu per storie Semplici ed in rosso per storie Complesse). Pannello di destra: durata totale del Parlato (esclusi i silenzi), separatamente per le storie Semplici (colonna a sinistra, numeri in blu) e Complesse (colonna a destra, numeri in rosso), e separatamente per le tre storie (righe)*

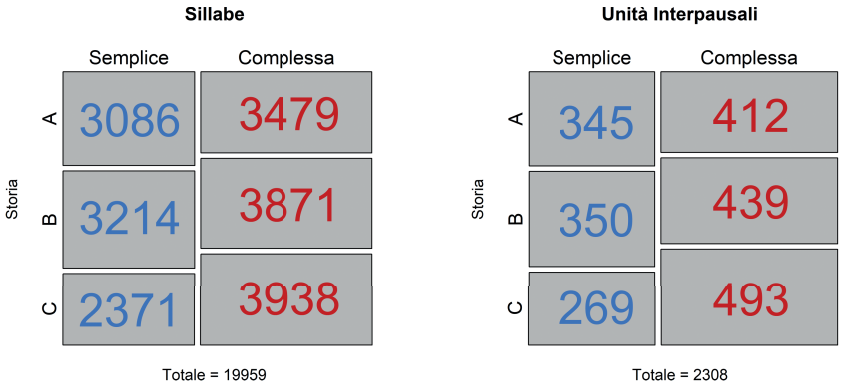


3.1 Analisi per storia

Separando i tempi di Parlato per le tre diverse storie (Fig. 3, pannello di destra), si osserva che la storia Semplice C ha una durata minore delle storie Semplici A e B. Questo risultato non sembra essere rilevante ai fini dell'effetto della complessità. Infatti, ogni storia Complessa ha durata di narrazione maggiore di ogni storia Semplice, confermando la robustezza dell'effetto. Simili risultati si ottengono contando per le diverse storie il numero di sillabe (Fig. 4, pannello di sinistra) o il numero di unità interpausali (Fig. 4, pannello di destra).

In altri termini, tutte le storie Complesse mostrano tempi di Parlato più lunghi, che si riflettono in maniera lineare sul numero di sillabe e di unità interpausali.

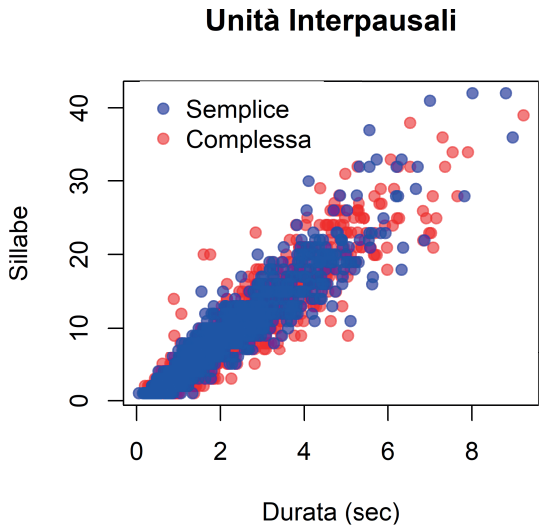
Figura 4 - *Il pannello di sinistra mostra il numero di sillabe, quello di destra il numero di unità interpausali. I dati sono rappresentati separatamente per le storie Semplici (colonna a sinistra, numeri in blu) e Complesse (colonna a destra, numeri in rosso), e separatamente per le tre storie (righe)*



3.2 Analisi per unità interpausale

La natura lineare dell'effetto di complessità sembra confermata dalla Fig. 5, che suggerisce un rapporto ugualmente lineare tra la durata delle unità interpausali e il numero di sillabe in esse contenute. Ad un aumento di durata delle singole unità si accompagna infatti un maggior numero di sillabe, indipendentemente dal tipo di storia. Le differenze di durata totale sembrerebbero quindi ascrivibili a un mero incremento del numero di unità interpausali nelle narrazioni Complesse.

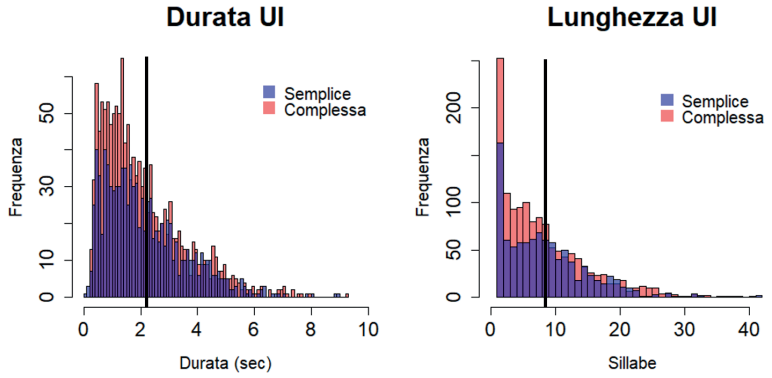
Figura 5 - *Durata (ascissa) e numero di sillabe (ordinata) per ogni unità interpausale, separatamente per le storie Semplici (blu) e Complesse (rosso)*



Tuttavia, calcolando i valori medi di durata, si nota che le unità interpausali hanno una durata media leggermente maggiore nelle narrazioni Semplici (2,13 sec) rispetto a quelle Complesse (2,04 sec). Questa differenza, inferiore al 5%, potrebbe risultare a prima vista di poco conto, ma si rivela interessante perché procede in direzione opposta all'effetto della complessità sui tempi totali di narrazione. Risultati simili si ottengono se si calcola il numero medio di sillabe per unità interpausale, ottenendo un valore di 9 per le narrazioni Semplici e di 8,4 per le narrazioni Complesse, quindi con una differenza vicina al 7%[2].

Per esplorare questo risultato nel dettaglio, abbiamo rappresentato la distribuzione delle durate nelle unità interpausali in Fig. 6 (sinistra). L'istogramma per le narrazioni Semplici (blu) è sovrapposto a quello per le narrazioni Complesse (rosso). In questo modo, i valori condivisi da entrambi i tipi di narrazione appaiono in viola. La presenza di un maggior numero di barre rosse (rispetto alle barre blu) indica che le narrazioni Complesse fanno uso di un maggior numero di unità interpausali. Tuttavia, la maggior parte di queste unità interpausali aggiuntive si concentra nella parte sinistra del grafico, per valori di durata compresi tra 0,5 e 1,5 sec, ovvero per valori al di sotto della durata media delle unità interpausali (indicata dalla linea verticale nera).

Figura 6 - *Distribuzione delle proprietà (ascissa) delle unità interpausali (sinistra: durata in secondi; destra: lunghezza in sillabe), separatamente per narrazioni Semplici (in blu) e Complesse (in rosso). In viola sono rappresentati i valori condivisi da entrambi i tipi. Le linee nere indicano i vaolri medi per le unità interpausali nell'intero corpus*



La Fig. 6 (destra) mostra la distribuzione della lunghezza (espressa in numero di sillabe) per unità interpausale. Parallelamente a quanto osservato per i valori medi, in questo caso l'effetto è ancora più robusto che nel caso delle durate. Come indica

[2] I dati nelle Figure 3-4 permettono di calcolare la velocità d'eloquio media per l'intero corpus (narrazioni Semplici: 3.1 sillabe al secondo; Complesse: 2.9). Risultati comparabili si ottengono per la velocità di articolazione (Semplici: 4.2; Complesse: 4.1). Questi dati non tengono però conto delle interazioni con la durata (in secondi) o la lunghezza (in sillabe) delle unità interpausali. Un'analisi approfondita di questi aspetti, impraticabile in questa sede, resta possibile attraverso l'uso dei materiali messi a disposizione nell'Archivio.

la massa rossa visibile a sinistra della linea nera, le narrazioni Complesse contengono non soltanto un maggior numero di unità interpausali, ma soprattutto un maggior numero di unità interpausali con poche sillabe (cioè tra 2 e 6).

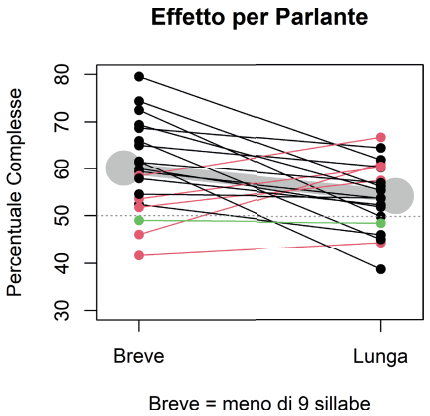
Questi risultati suggeriscono che, al crescere della complessità, la narrazione si faccia non solo complessivamente più lunga, ma anche internamente più frammentata.

3.3 Analisi per parlante

L'ispezione delle Figg. 3-6 suggerisce che al livello dell'intero corpus esista un'interazione tra lunghezza delle unità interpausali, durata della narrazione e complessità della storia. In altre parole, l'analisi superficiale dei dati suggerisce che le narrazioni Complesse abbiano durata maggiore, ma un'osservazione più attenta mostra che tale effetto è il risultato di un maggior numero di unità interpausali brevi. Così come un sottoinsieme di unità interpausali (Brevi) contribuisce in maniera cruciale all'effetto osservato sull'intero corpus, è anche possibile che solo un sottoinsieme di parlanti determini questi risultati. Nel seguito esploriamo quindi la robustezza di questa interazione per i 20 parlanti del nostro corpus.

La Fig. 7 rappresenta ogni parlante con una linea, mentre la linea grigia mostra i risultati relativi all'intero corpus, indipendentemente dal parlante. In ordinata rappresentiamo la percentuale di unità interpausali che il parlante ha prodotto nelle storie Complesse. Valori al di sopra del 50% indicano che il parlante ha prodotto più unità interpausali nella condizione Complessa che nella condizione Semplice. Il conteggio delle unità interpausali è fatto separatamente per le unità Brevi (a sinistra) e Lunghe (a destra), utilizzando come discrimine la stessa soglia impiegata al § 3.2, ovvero, in questo caso, massimo 8 vs. almeno 9 sillabe per unità interpausale. I risultati ottenuti per l'intero corpus (linea grigia) mostrano che, indipendentemente dalla lunghezza, la narrazione di storie Complesse richiede più unità interpausali, dal momento che entrambi i punti, sia quello a sinistra (unità Brevi) che quello a destra (unità Lunghe), si collocano al di sopra della linea del 50%. Inoltre, le unità Brevi sono ancora più frequenti nella condizione Complessa, per cui la linea risultante ha pendenza negativa.

Figura 7 - *Interazione tra tipo di storia (Complessa) e lunghezza delle unità interpausali (Brevi) per parlante. La linea grigia indica i risultati per l'intero corpus. I parlanti indicati in rosso mostrano la tendenza opposta*



Analizzando i risultati più da vicino, notiamo che 5 parlanti su 20 divergono dallo schema atteso. In questi casi, rappresentati in rosso in Fig. 7, la pendenza delle linee è positiva, visto che le narrazioni Complesse contengono un numero comparativamente maggiore di unità interpausali Lunghe. In un caso (rappresentato in verde) sembra invece non esserci alcun effetto, dal momento che entrambi i valori sono vicini al 50%. Nei restanti 14 casi, i parlanti sembrano conformarsi all'effetto atteso, producendo un maggior numero di unità interpausali Brevi nelle narrazioni di storie Complesse.

L'analisi della durata delle unità interpausali, invece che del numero di sillabe, fornisce risultati comparabili, ma meno robusti, dato che la tendenza generale è riscontrata in 12 parlanti. Il grafico relativo può essere creato usando lo script *C* disponibile nell'Archivio.

4. *Discussione*

In sintesi, la maggioranza dei parlanti ha prodotto per le storie Complesse delle narrazioni più lunghe e frammentate in un maggior numero di unità interpausali Brevi. Questi risultati portano nuove prove a sostegno dell'intuizione offerta dal senso comune, secondo cui la complessità della vicenda si rispecchia nella complessità della narrazione. È interessante notare però che tale complessità non si riduce alla mera lunghezza narrativa. Come nota Ryshina-Pankova (2015), unendosi alle voci di altri ricercatori nell'ambito dell'acquisizione delle lingue seconde, "complexity measures are often used and interpreted in simplistic terms reduced to *the longer the better* and *the more the better* arguments". Anche nel nostro caso, pur essendoci limitati ad analizzare misure semplici come la durata (di Silenzio, Parlato e unità interpausali) o il numero di sillabe, notiamo che la maggiore lunghezza delle narrazioni è solo uno

dei correlati della complessità delle storie e che un'analisi più approfondita, non meramente fonetica, potrebbe rivelare ulteriori dimensioni di variazione.

Ad esempio, la nostra analisi delle unità interpausali potrebbe essere integrata da un'esplorazione della distribuzione e della durata delle singole pause. Sia silenzi che unità interpausali potrebbero essere esaminati al livello delle singole storie, invece che al livello dei singoli parlanti, come in questo studio. Un'analisi più approfondita delle storie, ad esempio, potrebbe mostrare che alcune vignette in particolare hanno un effetto sulla narrazione, come quelle che mostrano una chiara rottura dell'ordine causale. Questo è il caso, ad esempio, della vignetta in cui una donna inforna una torta dopo averla sfornata nelle vignette immediatamente precedenti (Fig. 1, in basso). Infine, potrebbe essere interessante esplorare la tenuta di questi risultati al variare dei criteri di definizione delle unità interpausali. Il lettore interessato è invitato a utilizzare i dati messi a disposizione nell'Archivio per esplorare queste tracce o per verificare con statistiche induttive i risultati presentati in questo studio.

Inoltre, estendendo l'analisi anche alle restanti 12 storie disponibili nell'intero corpus zurighese, si potrebbe valutare il ruolo della complessità anaforica (vale a dire il numero e il sesso dei personaggi nelle storie) e di quella logica (ossia la sequenzialità delle vignette nelle storie). Infine, prendendo in considerazione più soggetti sarebbe possibile ricavare conclusioni più attendibili sul genere o sull'età dei parlanti.

Si noti come tutte queste piste di ricerca si limitano all'analisi della dimensione puramente macrofonetica delle narrazioni, ignorando non solo altri aspetti importanti del segnale acustico, come la prosodia, ma anche tutta la dimensione contenutistica e discorsiva. A titolo di esempio, riportiamo qui di seguito due storie raccontate dallo stesso soggetto (F09) per la condizione Semplice e Complessa, suddivise in unità interpausali

Storia Semplice L1A_1_F09

Der blonde Mann ging einmal einkaufen	'L'uomo biondo è andato a fare la spesa'
und als erstes suchte er sich	'e per prima cosa ha scelto'
den besten Wagen aus	'il miglior carrello'
dann ging er	'poi è andato'
in die Gemüseabteilung	'al reparto verdure'
und suchte sich die saftigste Orange aus	'e ha scelto l'arancia più succosa'
als nächstes kamen die Gurken dran	'in seguito è stata la volta dei i cetrioli'
da	'allora'
nahm er die grünste und dickste Gurke	'ha preso il cetriolo più verde e più grosso'
und dann ging er in die Fischabteilung	'e poi è andato al reparto pesce'
und suchte sich	'e ha scelto'
einen dicken Fisch	'un pesce grosso'
als nächstes war dann die Milchabteilung dran	'poi è arrivato il turno del reparto latticini'
wo er sich ein Joghurt holte	'dove ha preso uno yogurt'
und dann ging es auch schon an die Kasse um zu bezahlen	'e poi è già arrivato il momento di andare alla cassa per pagare'

Storia Complessa N3A_3_F09	
Ein braunhaariger Mann	'Un uomo dai capelli castani'
in einem Streifenhemd	'con una camicia a righe'
Öffnet	'apre'
Ähm	'ehm'
das...	'lo...'
die Gepäckauslage des Autos	'il portabagagli dell'auto'
wo zwei Schaufeln darauf leg...	'dove giacev... due pale...'
Ähm	'ehm'
der braunhaarige Mann	'l'uomo dai capelli castani'
trägt einen Tannenbaum	'porta un abete'
Der	'che'
in einer Schnur eingewickelt ist	'è avvolto da una corda'
und neben ihm	'e accanto a lui'
läuft ein	'cammina un'
hellbraun-haariger Junge	'ragazzo dai capelli castano chiaro'
der	'il'
Junge schaufelt den Baum wieder	'ragazzo pianta di nuovo l'albero'
in die Erde ein	'nella terra'
während der dunkelbraun-haarige Mann	'mentre l'uomo dai capelli scuri'
ihm	'lo'
mit einer Schaufel haltend dabei zusieht	'guarda con in mano una pala'
Der	'l''
ältere Mann	'uomo più anziano'
sitzt	'è seduto'
lächelnd in einem Auto	'sorridente in un'auto'
während der Junge	'mentre il ragazzo'
die Tür öffnet und ins Auto hineinsteigt	'apre la porta e sale in macchina'
der Junge schaut dem	'il ragazzo osserva l''
älteren Mann dabei zu	'uomo più anziano'
wie er die Schaufel aus der Be...	'mentre va a prendere la pala dal va...'
Gepäckauslage des Autos holt	'portabagagli dell'auto'
wo noch eine Schaufel und ein Seil sich befinden	'dove c'è un'altra pala e una corda'
der ältere Mann	'l'uomo più anziano'
bindet den Tannenbaum aufs Dach des Autos	'lega l'abete sul tettuccio dell'auto'
während ihm der Junge dabei zusieht	'mentre il ragazzo lo guarda'

La narrazione L1A rappresenta la prima delle storie Semplici narrate dal soggetto F09 nonché la prima storia in assoluto a lui presentata (per il significato dei codici v. § 2.2), mentre la seconda, N3A, corrisponde alla seconda storia Complessa incontrata nel corso dell'esperimento, con ordine non cronologico delle vignette e due personaggi dello stesso sesso.

A conferma di quanto illustrato in precedenza, la differenza più vistosa è la diversa lunghezza delle due storie. Rispetto alla storia Semplice, quella Complessa contiene più del doppio delle unità interpausali. Oltre ai parametri già utilizzati in

questo studio, tuttavia, il confronto tra i due tipi di storie offre ulteriori possibilità d'analisi.

A livello sintattico, ad esempio, nella storia Semplice ad ogni unità interpausale corrisponde all'incirca una proposizione. Nella storia Complessa, invece, la narrazione procede in maniera più spezzata: le proposizioni risultano frammentate, le unità interpausali sono spesso costituite da una sola parola e, a differenza della storia Semplice in cui predomina la paratassi, vengono impiegate numerose subordinate (relative e temporali), sia esplicite (introdotte dai connettivi *wo, der, während, wie*, ...) sia implicite (*haltend*).

Sul piano delle strategie di comunicazione, inoltre, rispetto alle storie Semplici è interessante notare una maggiore presenza di esitazioni (*ähm*), false partenze (*das... die Gepäckauslage*), correzioni (*aus der Be... Gepäckauslage*) e parole funzionali in isolamento (*das*, *der*, *der*) (De Iacovo, Colonna & Romano, 2020). In questi segnali di esitazione sembra visibile lo sforzo cognitivo del parlante al momento di pianificare narrazioni più elaborate. Nel tentativo di riportare a un senso logico la sequenza di immagini disordinata che ha di fronte, il soggetto prende tempo, facendo ricorso alle pause piene, e cerca di ricostruire le relazioni tra gli eventi, organizzando il discorso in maniera più strutturata.

Al momento abbiamo solo scalfito la molteplicità delle dimensioni di variazione lungo le quali si muove il concetto di 'complessità', che meriterebbe uno studio ben più articolato per essere definito e compreso. Tuttavia, pur essendo in accordo col monito di Ortega (2012) sul rischio di riduzionismo che si nasconde nell'uso acritico di semplici metriche di lunghezza, riteniamo che un'analisi dei dati centrata sulle singole storie, sui singoli parlanti e sulle singole unità interpausali abbia già abbastanza da offrire per un primo studio dei correlati fonetici della complessità narrativa.

Ringraziamenti

Il lavoro del primo autore è stato finanziato dal Centro di Ricerca Collaborativa (SFB) 1252 "Prominence in Language" del Fondo di Ricerca Tedesco (DFG) all'Università di Colonia. Le registrazioni e l'interazione con i parlanti sono state realizzate da Andrea Bizzeti, mentre la trascrizione ortografica delle storie è stata fornita da Harriet Hanekamp. Siamo grati ai 20 parlanti che hanno partecipato alla raccolta dati. Ringraziamo infine Chiara Celata ed un Revisore anonimo per i loro cortesi ed utili consigli.

Riferimenti bibliografici

Abbott, P.H. (2021). *The Cambridge Introduction to Narrative*, terza edizione. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108913928

Boersma, P., Weenink, D. (2022). PRAAT: doing phonetics by computer. [software] Versione 6.2.14. https://www.praat.org

Borges, J.L. (1970). *El informe de Brodie*. Buenos Aires: Emecé. [Trad. it.: Melis, A., Lorenzini, L. (a cura di) (1999). *Il manoscritto di Brodie*. Milano: Adelphi].

Bulté, B., Housen, A. (2012). Defining and operationalising L2 complexity. In Housen, A., Kuiken, F., & Vedder, I. (a cura di), *Dimensions of L2 Performance and Proficiency Investigating Complexity, Accuracy and Fluency in SLA*. Amsterdam/Philadelphia: Benjamins, 2146. https://doi.org/10.1075/lllt.32.02bul

Campione E., Véronis J. (2002). A large-scale multilingual study of pause duration. In *Proceedings of the 1st International Conference on Speech Prosody*, Aix-en Provence, France, 11-13 aprile 2002, 199-202.

De Iacovo, V., Colonna, V. & Romano, A. (2020). La pausazione. *Bollettino LFSAG* 5, 41-48. https://www.lfsag.unito.it/ricerca/phonews/05/5_5.pdf

Dell'Orletta, F., Montemagni, S. & Venturi, G. (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*, Edinburgh, Scotland, UK, 73-83.

Duez, D. (1982). Silent pauses and non-silent pauses in three speech styles. In *Language and Speech*, 25(7), 11-28.

Fossard, M., Achim, A.M., Rousier-Vercruyssen, L., Gonzalez, S., Bureau, A. & Champagne-Lavau, M. (2018). Referential choices in a collaborative storytelling task: Discourse stages and referential complexity matter. In *Frontiers in Psychology,* 9*,* 176. https://doi.org/10.3389/fpsyg.2018.00176

Gadda, C.E. (1963). *La cognizione del dolore*. Torino: Einaudi.

Grishakova, M., Poulaki, M. (a cura di) (2019). *Narrative Complexity: Cognition, Embodiment, Evolution*. Lincoln: University of Nebraska Press. https://doi.org/10.2307/j.ctvhktjh6

Gubrium, A., Gubrium, E. (2021). Narrative complexity in the time of COVID-19. In *The Lancet,* 397(10291)*,* 2244-2245. https://doi.org/10.1016/S0140-6736(21)01287-3

Kimble, J. (1992). Plain English: A Charter for Clear Writing. In *Michigan Bar Journal* (Dec. 1992), 1302-1307.

McCarthy, P.M., Jarvis, S. (2010). MTLD, vocdD, and HDD: A validation study of sophisticated approaches to lexical diversity assessment. In *Behavior Research Methods,* 42(2)*,* 381-392. https://doi.org/10.3758/BRM.42.2.381

Mittell, J. (2008). Narrative Complexity in Contemporary American Television. In *The Velvet Light Trap,* 58*,* 29-40. https://doi.org/10.1353/vlt.2006.0032

Ortega, J. (2003). Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing. In *Applied Linguistics,* 24(4)*,* 492-518. https://doi.org/10.1093/applin/24.4.492

Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In Kortmann, B., Szmrecsanyi, B. (a cura di), *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Berlin/Boston: De Gruyter, 127-155. https://doi.org/10.1515/9783110229226.127

Pallotti, G. (2015). A simple view of linguistic complexity. In *Second Language Research,* 31(1)*,* 117-134. https://doi.org/10.1177/0267658314536435

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [software] Versione 4.2.1. https://www.R-project.org/

Ryshina-Pankova, M. (2015). A meaning-based approach to the study of complexity in L2 writing: The case of grammatical metaphor. In *Journal of Second Language Writing,* 29, 51-63. https://doi.org/10.1016/j.jslw.2015.06.005

Walsh, R., Stepney, S. (a cura di) (2018). *Narrating Complexity*. Berlin: Springer. https://doi.org/10.1007/978-3-319-64714-2

DOI: 10.17469/O2109AISV000003

CLAUDIA CROCCO, BARBARA GILI FIVELA, GIUSEPPE MAGISTRO

Comparing dialectal and Italian prosody: the case of Venetian

The following paper aims at setting out a novel methodology in the prosodic comparison between two varieties in contact, the dialect spoken in Venice and the regional Italian spoken in Venice. By deploying a reading task, we compare the rhythmical properties of the two systems and review different metrics. We show that speakers can switch their metrical organization when switching language, but this is sensitive to those segmental processes which differentiate the two systems in contact.

Keywords: contact, rhythm, vanishing "l", "elle evanescente".

1. *Introduction*

The present study aims at comparing dialect and Italian from the prosodic point of view, with a primary methodological goal: we aim at testing a set of suitable procedures to identify and possibly quantify prosodic differences between dialectal and Italian varieties.

As a case-study, we consider Venetian Italian (VI) and urban Venetian dialect (VD). VD is an Italo-Romance vernacular or primary dialect (Coseriu, 1981), i.e., a sister language of Tuscan, from which Italian stems from (Serianni, Trifone, 1993). The comparison presented in this article, therefore, involves two historically related and yet grammatically and phonologically distinct linguistic systems that are in long-standing contact and co-exist in the city of Venice as varieties widely spoken in everyday conversation (Berruto, 2012, Ferguson, 2007). Compared to other Italo-Romance dialects, VD has enjoyed a certain prestige both in the past and today (Cortelazzo, Paccagnella, 1992, Dal Negro, Vietti, 2011). Accordingly, dialect-standard bilingualism is widespread in the region and in the city of Venice, and dialectal speech does not suffer a social stigma (Dal Negro, Vietti, 2011, ISTAT 2017).

VD is characterized in its pronunciation by the so-called *cadenza* (or *calada* in VD; Ferguson, 2007), i.e., a sing-song rhythmical cadence, a feature hinting to prosodic properties of VD partially diverging from those ascribed to Italian varieties (Gili Fivela, Avesani, Barone, Bocci, Crocco, D'Imperio, Giordano, Marotta, Savino & Sorianello, 2015). Magistro and Crocco (2022) explored the rising movements characterizing the final stretch of statements in Veneto dialects, proposing that they may play a role in

the *calada*. In this paper, we focus on durational differences between VD and VI and examine their possible impact on the rhythmic organization of the languages at stake.

The paper is structured as follows: in § 2, we discuss the link between durational variation and rhythm organization. In § 3, several relevant phonetic and phonotactic features of VD are presented. § 4 is dedicated to the methodology adopted to collect (§ 4.1) and pre-process the dialectal and Italian datasets (§ 4.2), with particular attention to the procedure adopted in specific cases (§ 4.2.1). Subsequently, we present the results of a statistical analysis of the durational measurements (§ 5) and formulate hypotheses about the possible source of the observed durational variation (§ 5.1). After exploring these hypotheses (§ 6), we discuss the implications of the results and draw the conclusions of the study (§ 7).

2. *Levels of rhythm organization*

Along the lines of Clarke (1999), Kohler (2009) and Arvaniti (2009), a distinction can be made between timing and rhythm: the former has to do with the duration of events (i.e., durational variability), while the latter regards the regular pattern extracted by the listener from, a.o., durational features. Although linguistic rhythm cannot be reduced to durational variability (Arvaniti, 2009), durational patterns are likely to be relevant for the organization and the perception of rhythm, as this phenomenon unfolds over time (Turk, Shattuck-Hufnagel, 2013).

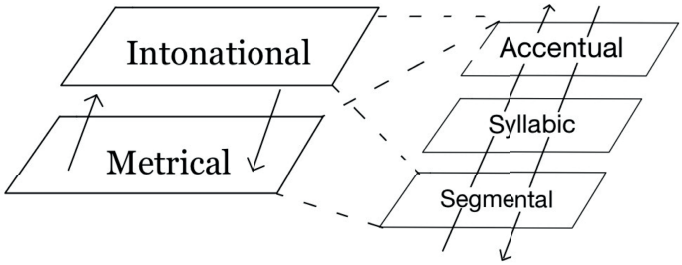
Languages may differ in their prominence-marking strategies (Andreeva, Barry & Koreman, 2014) and, therefore, in their use of duration to mark stressed syllables. In languages such as Italian, for instance, duration is essentially linked to prominence, as it is the most stable correlate of lexical stress both in production and perception (Bertinetto, 1981, Krämer, 2009, D'Imperio, Rosenthall, 1999). On the other hand, lengthening of stressed syllables is far less evident in Spanish than in Italian (Alfano, Savy & Llisterri, 2009; White, Payne & Mattys, 2009; see also Schmid, 2014), and differences are reported in relation to varieties of Italian as well, as prominence-related lengthening seems far more evident in Sicilian than in Venetian Italian (White et al., 2009).

Further, rhythmic differences across languages concern both low-level differences in temporal organization, and high-level differences involving, e.g., the spacing of stresses and accents or the tolerance to arrhythmic configurations (Jun, 2012, 2014, Falk, Rathcke, Dalla Bella, 2014, Frota, Moraes, 2016, Arvaniti, 2007, 2009, Turk, Shattuck-Hufnagel, 2013). Rhythmic analysis should therefore include the interaction between such layers and try to address the question how exactly the low-level temporal features are related to the high-level rhythmic organization, i.e., to the prominence hierarchy and its implementation (Fletcher, 2010, Rathcke, Smith, 2015), and the other way around, that is how the high-level organization, e.g., spacing of accents, is related to the rhythmic differences. In line with this wide perspective on the issue, in this paper we also try to trace back the possible sources of the observed temporal variation at the segmental level, by linking the experimental

results to what is known about the phonetics and phonology of the languages at stake (cf. Turk, Shattuck-Hufnagel, 2013).

Models such as the coupled oscillators proposed by Barbosa (2002, 2007) try to account for the interaction between prominence patterns governed by higher linguistic levels (phrase stress oscillator) and the syllabic sequence organized around vowel onsets (syllabic oscillator). In this model, cross-linguistic differences concerning vowel and consonant reduction are specified in the gestural lexicon. This intrinsic, language-specific level of timing interacts with the prosodic organization to produce actual segmental durations. However, segmental adjustments involving a.o. vowel lengthening, vowel insertion etc. can also be driven by the need to expand the text to provide more site for the tune realization (Grice, Savino, Roettger 2018, Roettger, Grice 2019). Therefore, just as the accentual pattern can be adapted to the words, syllables and segments composing the text, as it happens in well-known cases of tonal truncation or repulsion, adjustments such the insertion or the lengthening of a vowel can also be induced in the text by the tune (see Fig. 1). This suggests that timing effects, while being indeed language-specific, may be not as such independent from the high-level rhythmic organization.

Figure 1 - *Interactions between metrical and tonal tiers and between high-level and low-level components of rhythm*



2.1 Measurements of consonant-vowel ratio

Starting from Dauer's (1983, 1987) observation that vowel reduction may play a role in the perception of rhythm, several scholars have tried to quantify the consonant-vowel ratio by means of specifically developed metrics, in order to assign languages to rhythm classes (a.o. Dellwo, 2004, Dellwo, Wagner, 2003, Grabe, Low, 2002, Ramus, Nespor & Mehler, 1999). Such metrics provide different measures under the common assumption that the consonant-vowel ratio is a direct reflection of the language rhythm organization. The first rhythm metric was proposed by Ramus, Nespor & Mehler in 1999. This metric is based on some of the claims made by Dauer (1984) concerning possible phonetic and phonological correlates of stress-timed languages. The Deltas calculate the standard deviation of vocalic intervals (ΔV), the standard deviation of consonantal or intervocalic intervals (ΔC) and the percentage of vocalic intervals (%V). In the framework of the rhythm classes hypothesis, Ramus and colleagues hypothesize that the duration of vocalic and consonantal intervals

would show a stronger variation in stress-timed than in syllable-timed languages. Accordingly, higher values of ΔV and ΔC are expected in stress-timed languages compared to syllable-timed languages. Additionally, syllable-timed languages, which are characterized by less complex consonant clusters, would present a higher vocalic percentage %V compared to stress-timed languages. Since Deltas are extremely sensitive to speech rate variations, Dellwo and Wagner (2003) and Dellwo (2006) tried to improve the metric by normalizing data for speech rate. The normalized Deltas are called Varcos. Parallel to the Deltas, higher values of VarcoC and VarcoV are expected in stress-timed compared to syllable-timed languages.

A further metric is the so-called Pairwise Variability Index (PVI), proposed by Grabe and Low (2002), originally conceived to grasp the timing differences between closely related dialects of English. The PVI differs from Deltas and Varcos as it also considers the temporal sequence of vocalic and consonantal intervals. The formula of the row PVI (rPVI) computes the difference in duration between one interval and the following in a pairwise fashion, and then calculates the average of all differences. Since vowels are expected to be more sensitive to speech rate variations, Grabe and Low (2002) propose a nPVI (normalized PVI) for the calculation of vocalic intervals. As in the cases of Deltas and Varcos, also for the PVI stress-timed languages are expected to show lower values of rPVI and nPVI than syllable-timed languages.

Finally, a different metric has been proposed by Bertinetto & Bertini (2008). This metric differs substantially from other previously proposed in that it introduces a phonological dimension in the quantification of durational facts and distances itself from the stress-timing/syllable-timing dichotomy. The Control and Compensation Index (CCI) is a modification of the rPVI proposed by Grabe and Low (2008). In the CCI, the duration of each vocalic or consonantal interval is divided by the number of phonological segments included in the interval. Accordingly, geminate consonants and phonologically long vowels count as two segments. The CCI represents the level of "compression" allowed in a language/variety, i.e., the extent to which vocalic and consonantal segments can be lengthened or shortened in the context where they occur. Differences in the level of compression account for differences across languages; according to the authors' hypothesis, controlling languages allow for a low level of compression, whereas compensating languages allow for a high level of compression. Considering a space organized along the two dimensions of vocalic control and compensation (VCCI), and consonantal control and compensation (CCCI), controlling languages are expected to be scattered along the bisector, whereas compensating languages are expected to cluster below the bisector, in the lower right quadrant of the chart.

However, rhythm metrics as an instrument to assign languages to rhythm classes has received strong criticism, along with the rhythm class hypothesis itself (Kohler, 2009, Arvaniti, 2009, 2012, Rathcke, Smith, 2015). While the metrics provide relatively consistent results when applied to prototypical cases of stress-timed and syllable-timed languages, such as English and Spanish, they fail to reliably assign other languages, such as Greek or Thai to a specific rhythm class. Furthermore, the

results obtained through the metrics are heavily influenced by the characteristics of the analyzed data, in terms of speech rate, speech style, segmental composition etc., and consequently have a weak predictive power (Arvaniti, 2009, 2012). These results, besides indicating that rhythm metrics are not reliable as a tool to assign a language to a given rhythm class, also cast doubts on the validity of the rhythm classes hypothesis itself. Additionally, since the same durational measurements can derive from different sources in different languages, an additional shortcoming of such measurements is represented by their inability to identify the source of the observed variation (Arvaniti, 2009, Turk, Shattuck-Hufnagel, 2013).

In our view, the arguments put forward by the above-mentioned scholars are compelling. Rhythm metrics only provide a rough measurement of the durational facts, that are not *per se* typologically or phonologically relevant. Moreover, durational variation can derive from different sources, and is not the sole phonetic exponent of linguistic rhythm (Turk, Shattuck-Hufnagel, 2013). Such measurements, in fact, simply characterize the durational organization of a speech sample in terms of vocalic and consonantal intervals. Keeping this in mind, in this study we approach this set of measurements in a way that decidedly diverges from the preceding studies, using them to explore the hypothesis that one and the same bilingual speaker of VD and VI can adapt his/her durational organization when speaking Italian vs. dialect. Therefore, we use the rhythm metrics described above (Deltas, Varcos, PVI and CCI; implemented in the program *Correlatore*, Mairano, Romano, 2010) exclusively as a tool to quantify temporal differences between VD and VI at the segmental level in terms of vocalic and consonantal intervals; in principle, other measurements could have been used too. Accordingly, we will not frame the results in the rhythm classes hypothesis.

3. *Venetian Dialect: Phonetics and phonotactics*

Although systematic studies on VD phonology are still missing, the available research indicates that VD and Italian differ in several aspects that can be relevant for their rhythmic and prosodic organization. The phonotactics of VD, in particular, is fairly simple when compared to that of other northern dialects such as Romagnolo. Venetian presents 24 syllabic types (Schmid, 2014), most of which are in common with Italian (Schmid, 1998). Previous measurements on VD based on the rhythm metrics (Schmid, 2014), indicate for this dialect a relatively high proportion of vowels (%V) compared to other Italo-Romance dialects, and relatively low variability of the consonantal and vocalic durations (low standard deviation values for consonantal and vocalic duration, ΔV and ΔC). Overall, Schmid's (2014) measurements place VD in the area of syllable-based dialects.

VD has a number of morphological-phonological properties that are absent in Italian, e.g., final vowel/syllable apocope in several contexts and the so-called *'l' evanescente* (vanishing 'l'), i.e., elision/approximant realization of intervocalic /l/ (Zamboni, 1988). Such properties affect the way syllables are realized in VD

with respect to Italian. As for apocope, while Italian allows reduction and re-syllabification within and across words in informal, hypo-speech contexts, in VD re-syllabification phenomena show a (more) systematic character and can lead to fixed, univerbated forms (cf. Ferguson, 2007):

(1) *nol* < *no* + *el* "not the",
pel < *per el* "for the",
naltra < *un'altra* "another" [fem.],
chel < *che el* "that the"

Such cases provide relevant indications of possible differences in the syllable count between Italian and VD in comparable utterances.

A specific issue in VD is represented by the so-called *'l' evanescente* (see e.g., Tomasin, 2010), which is one of the three allophonic variants of the phoneme /l/ in this dialect. According to Tomasin (2010), such variants are:

- [l] in pre- or post-consonantal position, as in *folpo* 'octopus', *cantarla* 'to sing it';
- [ɬ] ([ɻ] in Tomasin's transcription) initial and intervocalic position, excepted when one of the vowels is a palatal. Such *'l' evanescente* is described as an approximant («approssimante dorsopalatale rilassata»), as in [gondoła] 'gondola';
- finally, /l/ is canceled ("dileguo") in intervocalic position when it precedes or follows a palatal vowel, as in *fiàr* 'spin' or *vea* 'sail'.

In the Italo-Romance domain, l-vocalization in intervocalic position is attested in several northern and southern dialects, in which variants including [ɬ], [j] and deletion are possible in different contexts (Rohlfs, 1966: 305ff.). Rohlfs' data suggest that this development is driven by the presence of a palatal vowel, which will then diverge from the development of /l/ into the labio-velar [w]. In Veneto, the approximant realization of /l/ and its deletion are likely to be relatively recent developments, since they are not attested in ancient text and they also lack in Goldoni's language (Rohlfs, 1966: 308), which may be considered as a most prominent example of literary use of the variety. The feature seems to spread from Venice to other varieties spoken in the region (e.g., Paduan; Tomasin, 2010: 731).

To the best of our knowledge, phonetic correlates of l-vocalization in intervocalic position have not yet been experimentally investigated in Romance. Experimental research on the realization of lateral consonants mostly concerns the case of l-velarization in coda position (Recasens, 2012). As for Veneto, the current account of l-vocalization is mostly based on descriptive studies such as Lepschy (1962) and Zamboni (1988), while experimental investigations of the phenomenon are still missing.

The allophonic realization of /l/ as [ɬ] or its deletion represent a potentially relevant issue when it comes to the temporal and, more generally, to the prosodic comparison between regional VI and VD. Although it can be argued that non-syllabic [ɬ] functionally acts as a consonantal incipit, its phonetic realization seems indeed *vocalized*, possibly contributing durational differences in vocalic and consonantal intervals in the two relevel varieties.

4. *Methodology*

In order to verify if bilingual speakers can switch temporal organization when speaking Italian vs. dialect, we measured segmental durations in two sets of VD and VI utterances closely comparable from the lexical, syntactic and informational point of view, and compared durational variability *within each individual* examined for the study. We first identified the metrics that better distinguish between VI and VD using statistical methods. Subsequently, we interpreted the results of these metrics against the background provided by the available knowledge of the phonetics and phonology of VD and Italian. In particular, we focused on the impact of segmental differences on syllable realization and syllable count.

4.1 Data collection

We recorded 5 bilingual speakers reading a set of dialogues corresponding to question-answer pairs. The recordings were taken at 44100Hz 16-bit *wav* format with a Blue Yeti microphone with the monodirectional polar pattern. All speakers declared a high level of proficiency in both Italian and Venetian, as well as claiming to use dialect on a regular basis. Speakers were aged 50-75 and coming from the neighborhood of Castello, in the historical center of Venice. The experiment consisted in two recording sessions (separated at least one day from the other). In the first session, speakers were prompted to read dialogues in Venetian or Italian, and in the second one, they were asked to record the other language. The order of the blocks was presented randomly to avoid potential saturation biases. Although the analysis of natural speech represents the ideal goal to aim for, we chose to rely on read speech for this study to control the experimental layout. This choice is due to the fact that durational measurements hare sensitive to segmental composition and speech style. Given the exploratory nature of the study, it was necessary to start building hypotheses with "clean data", without having to disentangle complicating factors. Furthermore, a rigid segmental layout is an offset to the limited number of recorded speakers. The experimental corpus consisted in a set of 15 target dialogues, eliciting three types of declaratives, 5 Broad Focus declaratives (1), 5 Contrastive Focus (2), and 5 Narrow-Informational focus (3). These dialogues had the same meaning across languages and were designed to keep as much similar segmental layout as possible. For example, the syllable count was kept constant across languages, together with lexical items. As appreciable in the reference examples, each target sentence was made of a bi-syllabic verb followed by the article '*la*' and a trisyllabic noun starting with nasal and having lexical stress on the antepenultimate syllable (Fig. 2). Besides 15 target dialogues, an equal number of fillers was elicited.

Table 1 - *Example of sentences*

Italian	Venetian	Translation
Broad Focus		
A: *Cosa fai stasera?* B: **Cucio la manica.**	A: *Cossa ti fa de sera?* **B: Cuzo la manega.**	'What will you do tonight?' 'I will sew the sleeve'
Contrastive Focus		
A: *Cuci il bottone stasera*? **B: Cucio la manica, stasera.**	A: *Ti te cuzi 'l boton stasera*? **B: Cuzo la manega, stasera.**	'Will you sew the button today?' 'I will sew the sleeve, tonight'
Narrow Informational Focus		
A: Cosa cuci stasera? **B: Cucio la manica, stasera.**	*A: Cossa ti cuzi stasera?* **B: Cuzo la manega, stasera.**	'What will you sew tonight?' 'I will sew the sleeve, tonight'

4.2 Data pre-processing

The target sentences were manually cut in Praat (Boersma, Weenink, 2022). Thereafter they were automatically segmented using the Forced Aligner MAUS (Schiel, 1999) trained for Italian. The intervals were sanity-checked by means of the following procedure: a small sample of 10 TextGrids was visually and auditorily inspected in Praat by each author. By joint comparison of the corrected TextGrids, the major issues of the automatic alignment within the corpus were discussed and debugged. Once the guidelines of manual correction were outlined, each author corrected another subsample of 40 sentences. Mutual annotation agreement was calculated using Intraclass Correlation Agreement (ICC) using the package *irr* in R (Gamer, Lemon, Fellows & Singh, 2019). After checking the main statistical assumptions, the ICC test was performed. The choice of the test was driven by the continuous nature of time-aligned segmentation: while K coefficient is well-suited for categorical variables, ICC provides a valid statistic for continuous dimensions. Specifically, a two-ways random ICC test for absolute agreement was performed and an F value $F(920, 1840) = 10{,}7$ with ICC = .76, $p < 0{,}001$, indicating substantial agreement on boundary placement. Once ascertained the inter-annotator agreement, each author corrected individually an equal part of remaining items (33 items per author = 99 + 1+ 50 jointly corrected = 150). The rationale behind this complex procedure is threefold. First, it alleviates the task of manual segmentation of the phonemes, where only sanity-check must be performed by researchers. Second, by examining the 30% of the automated output in distinct steps, it is possible to ensure that the human intervention on the data was consistent. Finally, a reliable segmentation is essential when computing metrics from duration values as we intended to do.

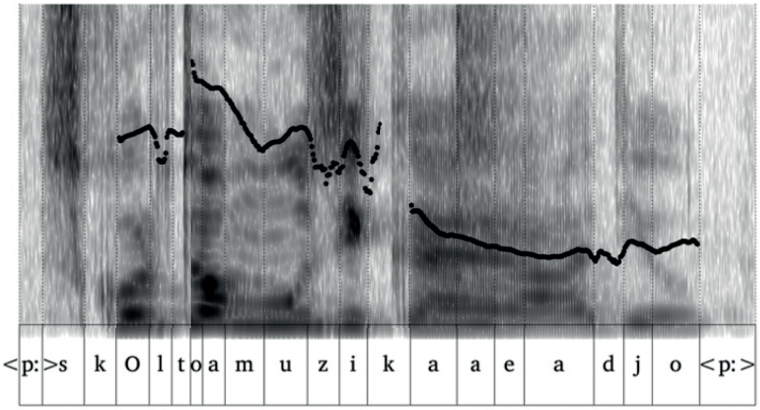
4.2.1 Annotation procedure for *'l' evanescente*

As discussed in § 2, /l/ in VD can be realized as [l], [ɬ], or can be deleted, according to the phonotactic context. In line with the literature (see § 3), we expect /l/ to be realized as [ɬ] in the target sentences, i.e., as an approximant consonant in articles (*la*, "the") and prepositions (*ala*, "to the").

For the annotation of the *'l' evanescente* we adopted the following procedure during the manual check of the Forced Aligner MAUS segmentation (Schiel, 1999): the annotator assessed the realization of the article *ea* "the" by listening to the prosodic phrase ending with the target word. In several cases, listening and instrumental inspection indicate cancellation of /l/ also in contexts in which [ɬ] is expected. In such cases, no allophone of /l/ was reported in the segmentation. An illustration of multiple l-cancelation is the utterance presented in Fig. 2.

Since the metrics implemented in the *Correlatore* compute the V-C proportions, the attribution of a vocalic or consonantal status to /l/ is likely to affect the result of the comparison between Italian and dialect. To avoid introducing biases in the comparison, we adopted a conservative approach by creating two copies of the Venetian dataset to feed the *Correlatore*, the first with /l/ labeled as lateral consonant [l], and the second with /l/ labeled as vocalic [e]. When the syllabic incipit /l/ was absent, only the syllabic nucleus was segmented and annotated, and the same TextGrid (with no interval corresponding to /l/) was used to feed the *Correlatore*.

Figure 2 - */l/ vocalization in the utterance Scolto la musica ala radio "I listen to music on the radio" (female speaker; 3vecfmus)*



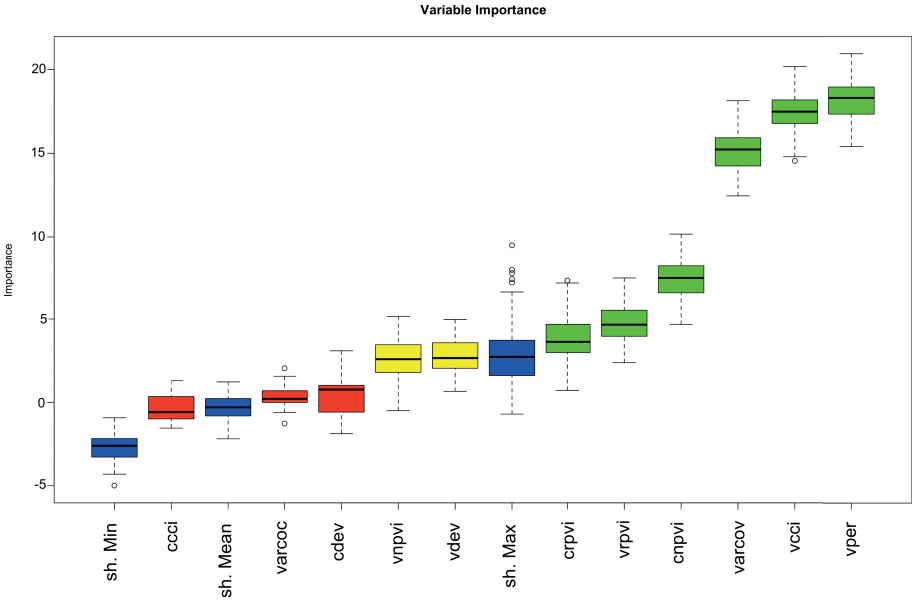
5. *Results: Rhythmic measurements*

5.1 Which dimension matters?

The 150 corrected TextGrids were batch-processed in *Correlatore* (Mairano, Romano, 2010) to extract the relevant rhythmic measures. The metrics were then exported in R, where they were tested to evaluate their capability in distinguishing between VI and VD. To this end, a feature selection algorithm was applied. We implemented a Boruta algorithm using the homonymous package in R. Boruta is a Random Forest-based algorithm, which compares the importance of each feature of the dataset with irrelevant randomly mixed features (the so-called *shadows*) to evaluate which real features are relevant for the classification (Kursa,

Rudnicki, 2010). Boruta, in comparison with other feature selection algorithms, is advantageous for our case: it is a method which does not seek to restrain the dimensionality to the minimum, i.e., it does not take out unnecessary features when others already do the job. This is a desideratum, because the algorithm will not penalize redundant features, which are naturally present in the metrics (the metrics all depend on the length of vowels and consonants, but with different formulas). We report in Fig. 3 the output by Boruta.

Figure 3 - *Results of feature selection with a Boruta algorithm*



The blue boxplots represent the Shadow features, used by Boruta as baseline. The red features are irrelevant for the classification VD-VI, while the yellow ones have slightly more chance to be relevant. The green ones, on the other hand, represent the most relevant features. In our paper, we will focus on the three most important ones, that is %V, VarcoV and VCCI. For the reader's convenience, we recall here that these measurements refer, respectively, to the percent of vocalic intervals in the measured speech (%V), the normalized standard deviation of vocalic interval duration divided by the mean (VarcoV), and the average duration of each vocalic interval divided by the number of phonological segments included in the interval (VCCI).[1] Of course, the algorithm cannot tell us anything about the direction

[1] We also checked whether there is an interaction between the sentence type (broad, contrastive and narrow focus) and the variety in the rhythmic distribution. This was performed with a clustering algorithm, kNN (see next section). The results showed a scattered distribution where the dimension of sentence type cannot represent a grouping factor accounting for the observed variability. We then continue to explore the effects of the continuous rhythmic indexes spotted by Boruta. We will test again whether the pragmatic condition has ultimately an effect in § 5.3.1.

of the relationship between the categorical dependent variables and the relevant features. In other words, while the algorithm helps us to identify the most useful features to distinguish VI from VD, it does not provide information about which of the two varieties has e.g., a higher %V. Next sections will aim at covering this gap.

5.2.1 Clustering methodology: kNN

In order to capture the directionality of the relevant features, we used a k-Nearest Neighbour Machine Learning algorithm (henceforth kNN). kNN is a clustering technique which learns how to distinguish two or more classes on the basis of a discriminating boundary. Starting from the number of near similar occurrences (the parameter k), kNN sets the boundaries of the pattern of distribution. The dataset underwent preliminary feature scaling and was split into 75% for training with k = 40 and 25% for test. We ran the algorithm each time for each relevant feature found by means of Boruta. Given the multi-dimensional nature of the test, we coupled each vocalic feature with its respective consonantal one (e.g. VarcoV coupled with VarcoC). The two-dimensional algorithm (and its plot) allows us to appreciate further the fact that only one dimension has more predictive power than the other. For example, if the boundary line of the kNN plot is orthogonal to only dimension, the other feature will be proven as insignificant in the clustering, confirming the output of Boruta. Furthermore, by drawing boundaries, kNN also expresses the direction of each dimension as a grouping factor, showing for example that values falling within a certain range will probably result in a specific cluster.

5.2.2 VarcoV

The least effective feature among the selected three is VarcoV. Compared to %V and VCCI, the vocalic standard deviations are therefore less efficient in distinguishing between VD and VI.

The results of the kNN training for VarcoV are appreciable in Fig. 4. The areas and dots in red represent VD and the ones in blue represent VI. The lightly colored dots on the background represent the probable class that a point would belong to if it were there. Furthermore, the dot radio represents the likelihood of the classification. For example, the radio is smaller for those dots near the boundary, since the classification is uncertain there. As more extreme values are more likely to belong to a specific group, the class likelihood displays a bigger radio. Fig. 4 shows that the only discriminant line between varieties is orthogonal to VarcoV. VD utterances in the corpus have lower VarcoV in opposition to VI, which occupies the higher area on the x-axis. As expected, VarcoC does not seem to play a discriminant role in the distribution, since the scattered dots occupy the same area on the y-axis. Although two areas can be detected, there is confusion near the boundary, indicating that some VI sentences can appear in the red area and vice versa. Such confusion also leads to a scarce accuracy of 50% in the test set, which is not satisfying for the classification task.

To check whether the language shift within each speaker corresponds to a change in the durational-rhythmic properties of the utterance, we report in Fig. 5 the boxplots for each speaker in a multi-facet grid (speakers are indicated with different numbers). As already suggested by kNN, speakers tend to have higher VarcoV when reading sentences in VI. While this tendency is homogeneous across speakers (see e.g., the median lines), the error bars of VD and VI occupy similar areas, making it difficult to draw clear-cut boundaries. To conclude, while we can see higher VarcoV values for Italian indicating a trend in the data, there is still a fuzzy zone, confirming VarcoV as a non-ideally reliable discriminant dimension.

Figure 4 - *Scatterplot of the sentences along the dimensions of VarcoV with the decision boundary and likelihood area provided by kNN*

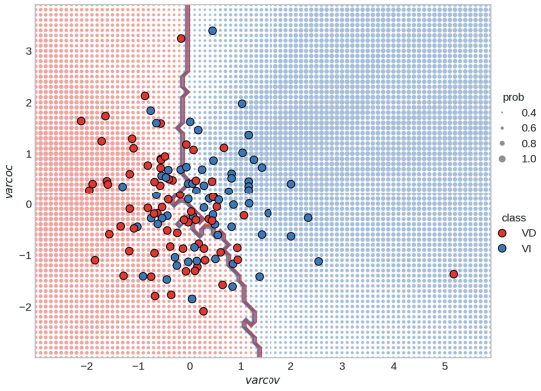
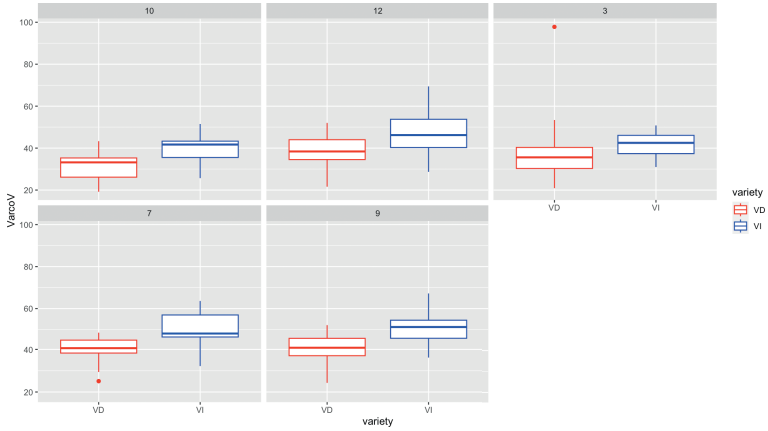


Figure 5 - *Boxplots divided by speaker indicating the range of VarcoV for both VD and VI*

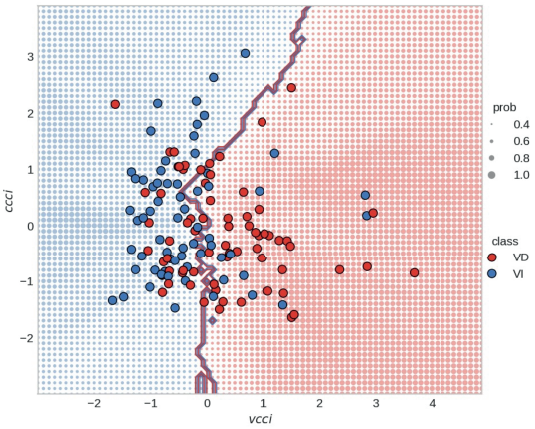


5.2.3 VCCI

We can now turn to kNN for the Compensation and Control Indexes: this time the model scored an accuracy level of 86%, which we consider as valid for discriminating the varieties in our data. Interestingly, kNN shows that VD has higher VCCI levels,

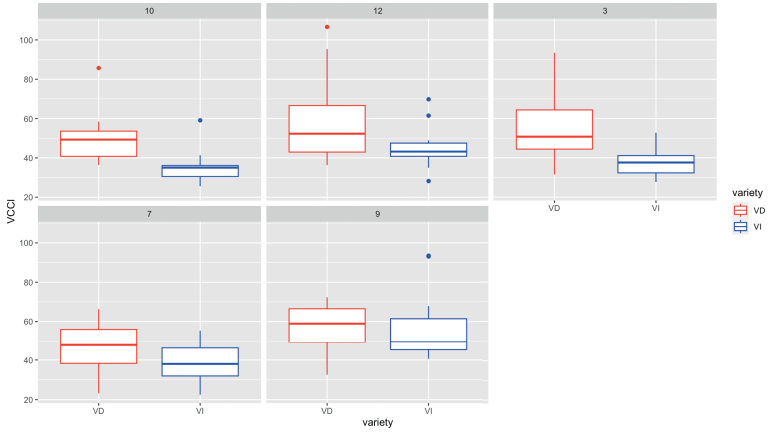
while VI occupies lower areas. Here again, the boundary is orthogonal to VCCI, while the distribution of data does not seem to be sensitive to CCCI, as expected based on the results of the feature selection.

Figure 6 - *Scatterplot of the sentences along the dimensions of VCCI and CCCI with the decision boundary and likelihood area provided by kNN: training set*



Also in this case, however, there is much confusion around the boundary area (Fig. 6). Let us explore the distribution of VCCI per speaker (Fig. 7). The tendency described by kNN corresponds to the homogeneous behavior of each speaker: the median values of VCCI are higher for VD. Note also that while some speakers make a clear-cut distinction (e.g., speakers 3 and 10), others have overlapping areas. This shows that the tendency is fairly constant, but the differentiation rate within speakers is subject to interindividual variation.

Figure 7 - *Boxplots divided by speaker indicating the range of VCCI for both VD and VI*



5.2.4 %V

Finally, we can turn to the most indicative predictor in distinguishing between VI and VD. We employed the same methodology as the previous dimensions. The kNN predictor reached an accuracy level of 92% for the test set, confirming that this is the most reliable dimension. Note that *Correlatore* does not have an inborn function to calculate %C, the opposite dimension of %V, so we calculated with the formula %C = 100 - %V. This explains why in the kNN plot the two dimensions are linearly dependent. Naturally, the decision boundary is a diagonal (Fig. 8): %V and %C are strictly dependent, and the decision boundary is drawn along both dimensions. In particular, VD sentences are more likely to appear with higher values of %V, while VI correlates with higher %C. This tendency seems to be well maintained by each speaker (Fig. 9).

Figure 8 - *Scatterplot of the sentences along the dimensions of %V and %C with the decision boundary and likelihood area provided by kNN*

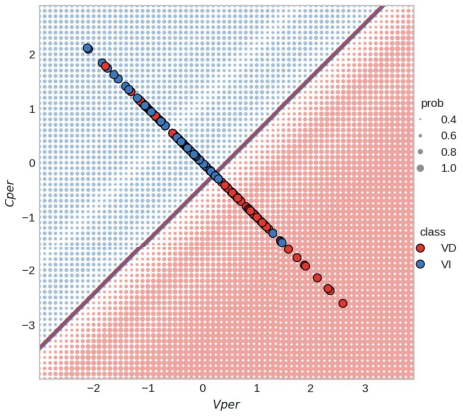
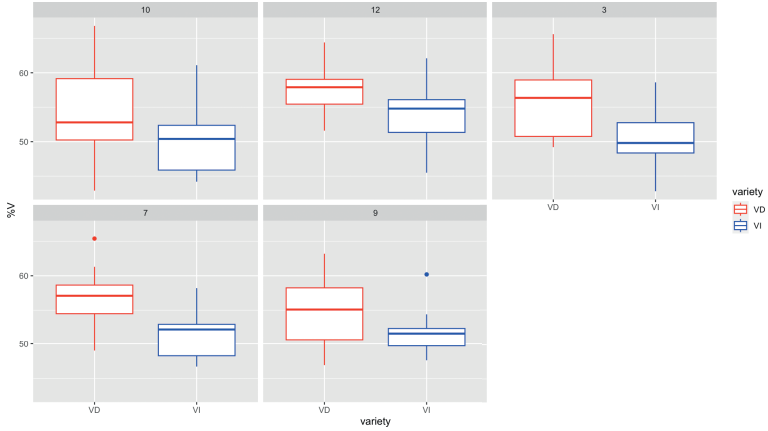


Figure 9 - *Boxplots divided by speaker indicating the range of V% for both VD and VI*



5.3 Intermediate discussion: from timing to rhythm

To the goal of comparing dialect and Italian from the prosodic point of view, we tested a number of rhythm metrics and found measurements capturing possible durational differences between VD and VI within each speaker. The results show that three of these measurements highlight consistent differences in the temporal properties when the speakers switch language. Even if the speakers adapt their production when speaking VD or VI to different extents, the overall tendency seems to be commonly shared.

The analysis also revealed that metrics perform very differently in distinguishing between the VD and VI sample. VD and VI can be distinguished only on the basis of a subset of the metrics tested in this study, namely %V, VCCI and, to a limited extent, VarcoV. It is noteworthy that vocalic and consonantal measurements do not work in tandem: the most important dimensions all rely on vocalic intervals, where the consonantal intervals seem less important in discriminating between VI and VD.

Overall, the analysis of durations of vocalic and consonantal intervals in our datasets indicates differences between VI and VD as far as vowel length is concerned, in line with results in the literature pointing to the relevance of vowels in characterizing VD (Schmid, 2014). As discussed in § 3, however, durational measurements taken as such do not provide relevant information about the rhythmic organization of a language.

Interestingly, the most performant feature in the identification of the variety is %V, which is the most straightforward metric simply displaying the amount of vocalic quantity in the sentences. This linear segmental dimension suggests that VD either has more vocalic intervals than VI, or that vocalic intervals are longer, or both. Similarly, VD displays higher values compared to VI also for VCCI, indicating a higher vocalic ratio in the dialect also when the number of phonological segments composing the interval is considered. Although caution is needed when interpreting the results from these metrics (Arvaniti, 2009, Prieto et al., 2012), we take them at face value because the segmental material was strictly controlled (cf. Fig. 3). Since the lexical material and the consonant-vowel ratio is kept constant across varieties, then this outcome might be attributed to either vocalization of (expected) consonantal segments or lengthening of vocalic intervals, or to a combination of the two. The datasets annotation shows that only a small amount of /l/ in the Venetian is realized in articles and prepositions. In most of the cases (97%), /l/ is not realized and only [a] is present (cf. Fig. 2), also when no palatal vowel is present in the segmental environment, contra our expectations based on the literature (Tomasin, 2010). Although the lack of the consonant /l/ has arguably an impact the consonant-vowel ratio in VD *vis-à-vis* VI, as reflected in the results provided by %V and VCCI, the measurements discussed in the preceding section do not provide information about the actual realization of /l/. The presence of [a] alone hints indeed to two possible scenarios: in the first, /l/ in the article *ła* "the" is simply deleted. In this case, we can expect the duration of the vowel to be unaffected and be therefore comparable to that of the other unstressed vowels of the utterance. In the alternative scenario, /l/ is

realized as a part of [a], which would consequently be lengthened. This hypothesis is coherent with the picture given by the metrics, since both the percentage of vocalic intervals, and the V-C ratio *vis-à-vis* the phonological segments would be higher in VD than in VI. We will verify this hypothesis in the following section (5.3.1).

While ΔC was not a discriminant factor, ΔV performed better but not at a satisfactory level to consider it essential in the distinction between VI and VD (cf. Fig. 4), while the normalized version of ΔV, VarcoV, reached a higher level of importance. VD correlates with lower VarcoV values, indicating that vowels in VI might have a more variable duration than in VD. This parameter, however, had a low predictive power compared to %V and VCCI. Altogether, the metrics suggest that vowels in VD might be more abundant, or longer, or both, and less variable in length compared to VI. It is not straightforward to elaborate about a scenario that can account for all these features. Along the lines of Alfano, Savy and Llisterri (2009), however, it may be hypothesized that the observed durational properties result (at least partially) from different durational strategies in marking stress. In what follows, we will also explore this point.

5.3.1 Vowel length in VD vs. VI

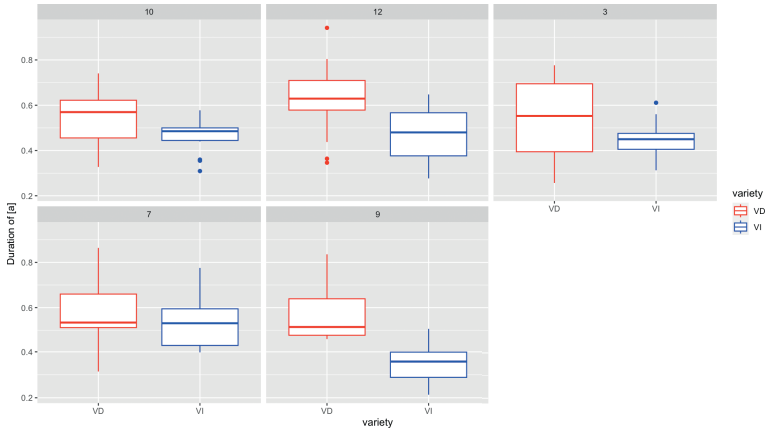
We compared the length of [a] in the article *ła* (VD) and *la* (Italian) "the" to that of the other unstressed vowels of the utterance, to verify if [a] in VD is lengthened. Given the relatively little sample dimension, we cannot test frequentist hypotheses to make inferences on a population level. However, to make sense of the data while controlling speaker variation and the effect of predictors other than the variety, we performed a Bayesian linear mixed model (McNelsh, 2016) to predict the length of the vowel [a] with the variety and sentence type. The model also included the speaker and item in the random structure. Since we expect variation in the segmental length given by the speech rate of each speaker and utterance, we decided to center this dimension. To do so, we followed the procedure displayed in the formula. We first calculated the mean of all unstressed vowels (n) in the sentence. In the calculation of the mean (l), we factored out the actual interval of interest (duration of /a/) by subtracting it. Then, the length of [a] for each utterance was respectively divided by the utterance mean.

(1) $$Centr\Delta t_i = \frac{\Delta t_i}{l} \quad where\ l = \frac{\sum_{j=1}^{n} \Delta t_j - \Delta t_i}{n-1}$$

Back to the statistical modeling, we expect a posterior probability of at least 95% to conclude that these data confirm the hypothesis that [a] is lengthened. The model was built in Stan using the R interface *brms* (Bürkner, 2017), using a Monte Carlo Markov Chain with 4 chains and 2000 observations for each chain. A weakly-informative prior was specified, given the exploratory nature of study (hyperparameters set at a normal distribution with mean = 0 and sd = 1).

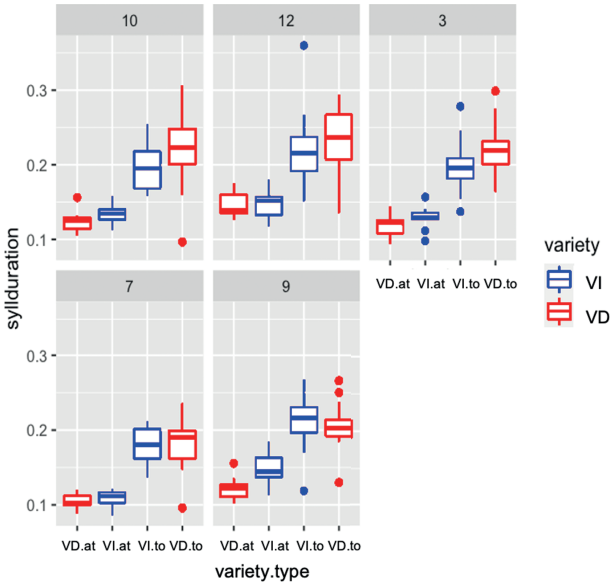
The model, which obtained a R-hat value of 1, specifies an intercept set at the VI with the sentence type BF at 0.46, est. error = 0.03, and a credible interval [0.40, 0.52]. While the effect of the sentence type did not yield any credible difference on the dataset (Category CF CI[-0.06, 0.05], Category NF CI[-0.01, 0.09]). The effect of variety produced a strong effect: the level VD yielded an estimate of 0.12, std. error = 0.02, and CI[0.07, 0.16], probability = 95%). We can conclude that the data and model support the hypothesis that [a] in VD *la* has longer duration than in VI. Moreover, the pragmatic category (e.g. whether the sentence was Broad, Contrastive, or Narrow focus) did not yield any credible effect in the distribution of the length. The differentiation between varieties seems also stable across the speakers (Fig. 10). These results support the hypothesis that /l/ is not simply canceled, but triggers a durational readjustment, i.e., the lengthening of [a].

Figure 10 - *Boxplots divided by speaker indicating the duration of [a] in VD and VI*



To further explore the durational properties of VD and VI and try to trace back the sources of the differences highlighted by the metrics, we compare the duration of stressed and unstressed syllables in VD and VI (Fig. 11). In doing so, we excluded [a] from the dataset as it would obviously have an impact on the results concerning unstressed syllables. Although realizations vary to a certain extent (see speaker 9), unstressed syllables tend to be shorter or similar in length to those of VI, while stressed syllables tend to be longer in VD. This result indicates that speakers can indeed adopt different temporal strategies to mark stress when switching from one language to the other.

Figure 11- *Boxplots divided by speaker indicating the duration of stressed ("to") and unstressed ("at") syllables in VD and VI*



6. *Discussion*

In this section we try to link the phonetic measurements described so far to the higher-level rhythmic organization, i.e., to the prominence hierarchy and its implementation.

The exploration of durational features highlighted several differences between VI and VD. Firstly, the analysis showed that vocalic intervals are globally longer in VD than in VI. Although this result is coherent with the well-known lack of consonantal gemination in VD (Zamboni 1988), it is worth noting that the Italian target sentences used in the present study did not include geminated consonants and, therefore, no role can be attributed to gemination in the results obtained on this dataset. Lack of gemination and longer vocalic intervals, however, fit the picture of VD and VI with divergent temporal organizations at the segmental level. Note that, differently from VD, in VI the distinction between singleton and geminate consonants is consistently made (Mairano, De Iacovo, this volume; in line with what observed for other varieties of Italian spoken in the same area; Zmarich, Gili Fivela, 2005).

Secondly, duration seems a more relevant cue to mark stress in VD than in VI, since stressed syllables tend to be longer in the dialect, and therefore more pronouncedly distinguished from unstressed syllables in this variety. As pointed out by Alfano, Savy and Llisterri (2009), related languages which are structurally similar, such as Italian and Spanish, may differ in their temporal organization and in their use of duration as a cue to stress in production and perception. Our data go in the

same direction, showing that differences concerning the role of duration in marking stress can emerge also from the comparison of an Italo-Romance dialect and the corresponding Italian variety. Additionally, the data indicates that vowel duration in VD might be less variable than in VI. This result can in principle be in line with the scenario of a clearer durational distinction between stressed and unstressed vowels in VD. However, this link is speculative, inasmuch the relationship between the VarcoV values and the stressed/unstressed distinction has not yet been investigated. We leave this topic open for future research.

The results presented in this work also add a piece in the description of the so-called *'l' evanescente*, showing that consonantal deletion can take place also when there is no palatal vowel in the segmental environment, differently from what has been reported in the literature so far (Tomasin, 2010). These results, however, also raise the question about why the vowel [a] in the article is lengthened. A possible explanation is that lengthening is necessary in order maintain the perceptual salience of the article, which would be otherwise too short to be identified in the speech flow. This explanation is in line with the idea that durational adjustments can be driven by linguistic factors above the segmental level.

A further observation for future work concerns the possibility that /l/ is not the only consonant subject to weakening in VD, besides the well-known lack of geminate consonants. The example provided in Fig. 2 suggests that weakening may also involve rhotics. If this is confirmed, consonantal weakening in VD might regard a larger set of sonorants.

Altogether, these results indicate differences in syllable realization (consonantal weakening, vowel lengthening), stress realization (different role of duration in marking stress). The observed differences in timing, therefore, are likely to influence the implementation of the prosodic hierarchy in VD and VI. Consistently, [a] lengthening is likely to be driven by linguistic forces above the level of the segment, as this adjustment could be due to the necessity of preserving the article's perceptual salience at the level of the phonological words.

At present, little work has been done on the intonational phonology of Italo-Romance dialects, including VD. Preliminary research on Veneto dialects (Magistro, Crocco, 2022) suggests that broad focus statements in these varieties might be more dynamic than Italian in the realization of edge tones. On the other hand, the boosting of duration in stressed syllables might also be part of a strategy to enhance prominence, as proposed for Neapolitan dialect by Crocco, Gili Fivela and D'Imperio (2022). Therefore, future research needs to delve into the link between longer duration in stressed syllables and the intonational organization of the dialect, to pinpoint possible influences going from the tune to the text, or the other way round (Grice et al. 2018, Roettger, Grice 2019).

7. *Conclusions*

In this paper we outlined a methodological pipeline based on the use of machine learning techniques to assess the power of durational measurement to distinguish between two languages, with the goal of linking phonetic variability in segmental duration to the higher levels of the prosodic organization.

Assuming that different languages are likely to differ also in their prosodic properties, and hence also in their temporal and rhythmic organization, we tested the hypothesis that bilingual speakers can adapt their temporal organization of when speaking one or the other language. To test this hypothesis, we examined the production of a group of bilingual speakers, asking them to read aloud a set of controlled sentences in both the languages we wanted to investigate (in our case: VI and VD). The material was then carefully segmented to extract duration measurements of consonantal and vocalic intervals. Using a set of machine learning techniques, we subsequently identified the durational parameters that are more likely to distinguish the languages under investigation from the durational point of view. Finally, we linked the results provided by the durational analysis to what is known about the phonology of the examined languages, providing a basis for future investigations aimed at understanding the role of durational facts in the prosodic organization of a language.

We assumed a layered organization of linguistic rhythm, but intentionally no claim was made concerning rhythmic typology and the possibility to assign a language to one or the other rhythm class. Instead, we further convolute rhythmic properties by combining segmental processes and, possibly, salience factors by pointing at the multi-faceted nature of rhythm.

We tested the so-called rhythm metric focusing on the measurements they are based on and exploring the predictive power of such measurements separately through statistical analysis. Machine learning algorithms proved to be a useful tool to assess the possibility that two languages can be distinguished based on a given feature. Our preliminary results indicate that by keeping other linguistic factors controlled and focusing on duration at the segmental level, we can indeed highlight divergences in the temporal organization of languages, which are likely to be, though indirectly, relevant for the rhythmic structure. Although further research is needed to bridge the gap between phonetics and phonology, fine-grained analyses such as those presented in this article can represent a viable methodological option to proceed on this path.

Of course, duration does not exhaust the phonetic features that can be potentially relevant in the rhythmic differentiation between languages or varieties. Other measurements can indeed be performed that could be complementary and perhaps equally useful and need therefore to be integrated. The idea behind this paper, to be further investigated, is that timing effects and rhythm may be dependent from the high-level rhythmic organization. Since languages rely on duration to a different extent to realize stress, other prosodic parameters need to be included in the picture to link realization and function.

References

Alfano, I., Savy, R. & Llisterri, J. (2009). Sulla realtà acustica dell'accento lessicale in italiano ed in spagnolo: La durata vocalica in produzione e percezione. In Romito, L., Galatà, V. & Lio, R. (eds.), *La fonetica sperimentale. Metodo e applicazioni*. *Atti del IV convegno nazionale AISV* – Arcavacata di Rende (CS). [CD-ROM]. Torriana: EDK Editore. 22-39.

Andreeva, B., Barry, W. & Koreman, J. (2014). A Cross-language Corpus for Studying the Phonetics and Phonology of Prominence. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), pages 326–330, Reykjavik, Iceland. European Language Resources Association (ELRA).

Arvaniti, A. (1994). Acoustic features of Greek rhythmic structure. In *Journal of Phonetics,* 22, 239-268. https://doi.org/10.1016/S0095-4470(19)30203-7

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. In *Phonetica* 66, 46-63. doi: 10.1159/000208930

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. In *Journal of Phonetics,* 40, 351-373. https://doi.org/10.1016/j.wocn.2012.02.003

Barbosa, P. (2002). Explaining Brazilian Portuguese resistance to stress shift with a coupled-oscillator model of speech rhythm production. In *Cadernos de estudos linguisticos*, *43*, 71-92. https://doi.org/10.20396/cel.v43i0.8637150.

Barbosa, P. (2007). From syntax to acoustic duration: A dynamical model of speech rhythm production. In *Speech Communication*, 49, 725–742. https://doi.org/10.1016/j.specom.2007.04.013

Berruto, G. (2012). *Sociolinguistica dell'italiano contemporaneo* (2nd ed.) Roma: Carocci.

Bertinetto, P.M. (1981). *Strutture prosodiche dell'italiano. Accento, quantità, sillaba, giuntura, fondamenti metrici.* Firenze: Accademia della Crusca.

Bertinetto, P.M., Bertini, C. (2008). On modeling the rhythm of natural languages. In *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brasil, 427-430.

Boersma, P., Weenink, D. (2022). Praat: doing phonetics by computer [software]. Version 6.2.09, http://www.praat.org/

Bürkner, P.C. (2017). brms: An R package for Bayesian multilevel models using Stan. In *Journal of statistical software*, 80, 1-28.

Clarke, E.F. (1999). Rhythm and timing in music. In Deutsch, D. (Eds.) *The psychology of music*, New York: Academic press, 473–500.

Cortelazzo, M., Paccagnella, I. (1992). Il Veneto. In Bruni, F. (Eds.) *L'italiano nelle regioni*, Torino: UTET, vol. I, 263-310.

Coseriu, E. (1981). Los conceptos de «dialecto», «nivel» y «estilo de lengua» y el sentido propio de la dialectología. In *Lingüística Española Actual*, 3, 1–32.

Crocco, C., Gili Fivela B. & D'Imperio, M. (2022). Comparing prosody of Italian varieties and dialects: data from Neapolitan. In *Proceedings of the 11th International Conference on Speech Prosody*, Lisbon: Portugal, 140-144.

Dal Negro, S. Vietti, A. (2011). Italian and Italo-Romance dialects. In *International Journal of Sociology of Language*, 210, 71- 92. https://doi.org/10.1515/ijsl.2011.031

Dauer, R.M. (1983). Stress-timing and syllable-timing reanalyzed. In *Journal of Phonetics, 11*(1), 51–62.

Dauer, R.M. (1987). Phonetic and phonological components of language rhythm. In *Proceedings of the 11th International Congress of Phonetic Sciences,* Tillinn, Estonia, 447-4450.

Dellwo, V., Wagner, P. (2003). Relations between language rhythm and speech rate. In S, V., Wagner Solé, P., Recasens, D., Romero (eds.), *Proceedings of 15th International Congress of Phonetic Science*s, 471-474.

Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for ΔC. In *Proceedings of the 38th Linguistic Colloquium*, Budapest, 231-241.

D'Imperio, M., Rosenthall, S. (1999). Phonetics and phonology of main stress in Italian. In *Phonology*, 16, 1-28.

Falk, S., Rathcke, T. & Dalla Bella, S. (2014). When speech sounds like music. In *Journal of Experimental Psychology*, 40(4), 1491-1506. http://dx.doi.org/10.1037/a0036858

Ferguson, R. (2017). A Linguistic History of Venice. Firenze: L.S. Olschki. http://digital.casalini.it/9788822256454

Dellwo, V., Wagner, P. (2003). "Relations between language rhythm and speech rate," in *Proceedings of the 15th International Congress of Phonetic Sciences*, 471–474.

Fletcher, J. (2010). The prosody of speech: timing and rhythm. In Laver, J., Gibbon, E.F. (eds.), *The Handbook of Phonetic Sciences* (2nd ed.), West Sussex, UK: Blackwell, 521–602. https://doi.org/10.1002/9781444317251.ch15

Frota S., Moraes, J.A. (2016). Intonation of European and Brazilian Portuguese. In Wetzels, L., Costa, J., Menuzzi, S. (Eds.), *The Handbook of Portuguese Linguistics*. Willey-Blackwell. https://doi.org/10.1002/9781118791844.ch9

Gamer, M., Lemon, J., Fellows, I., Singh, P. (2019). irr: Various Coefficients of Interrater Reliability and Agreement. [R package] version 0.84.1. https://CRAN.R-project.org/package=irr

Gili Fivela, B., Avesani, C., Barone, M., Bocci, G., Crocco, C., D'Imperio, M., Giordano, R., Marotta, G., Savino, M., Sorianello, P. (2015). Intonational phonology of the regional varieties of Italian, in Frota, S., Prieto, P. (Eds.) *Romance intonation*, Oxford: Oxford University Press, pp. 140–197.

Grabe, E., Low, E.L. (2002). Acoustic correlates of rhythm class, in *Laboratory Phonology*, 7, 515–546.

Grice, M., Savino, M. & Roettger T. (2018). Word final schwa is driven by intonation: The case of Bari Italian. In *JASA*, 143(4), 2474-2486. https://doi.org/10.1121/1.5030923

ISTAT. (2017). https://www.istat.it/it/archivio/207961, last access: 30 August 2022.

Jun, S.A. (2012). Prosodic Typology Revisited: Adding Macro-rhythm. In *Proceedings of 6th International Conference on Speech Prosody*.

Jun, S.A. (2014). Prosodic Typology: By Prominence Type, Word Prosody, and Macro-rhythm. In Jun, S.A. (eds.) *Prosodic Typology* II. Oxford: OUP, 520-539. https://doi.org/10.1093/acprof:oso/9780199567300.003.0017

Ladd, R.D. (2008). *Intonational phonology* (2nd ed.) Cambridge: CUP. https://doi.org/10.1017/CBO9780511808814

Kohler, K. (2009), Whither Speech Rhythm Research? In *Phonetica*, 66, 5–14.

Kramer, M. (2009), The phonology of Italian, Oxford: Oxford University Press.

Kursa, M.B, Rudnicki, W.R. (2010). Feature selection with the Boruta package. In *Journal of Statistical Software*, 36, 1-13.

Lepschy, G.C. (1962). Fonematica veneziana, in *L'Italia dialettale,* 25 (1962), pp. 1-22.

Magistro G., Crocco, C. (2022). Rising and rising-falling declaratives in Veneto dialects In *Proceedings of the 11th International Conference on Speech Prosody*, 175-179.

Mairano, P., Romano, A. (2010) Un confronto tra diverse metriche ritmiche usando Correlatore. In: Schmid, S., Schwarzenbach, M. & Studer, D. (Eds.) *La dimensione temporale del parlato*, Proceedings of the V National AISV Congress, University of Zurich, Collegiengebaude, 4-6 February 2009, Torriana (RN): EDK, 79-100.

McNeish, D. (2016). On using Bayesian methods to address small sample problems. In *Structural Equation Modeling: A Multidisciplinary Journal*, 23(5), 750-773. https://doi.org/10.1080/10705511.2016.1186549

Ramus, F., Nespor, M. & Mehler, J. (1999). *Correlates of linguistic rhythm in the speech signal*, In *Cognition*, 73(3), 265–292. https://doi.org/10.1016/S0010-0277(00)00101-3

Rathcke, T., Smith, R.H. (2015). Speech timing and linguistic rhythm: On the acoustic bases of rhythm typologies. In *JASA,* 137: 28-34. http://doi.org/10.1121/1.4919322

Recasens, D. (2012). A cross-language acoustic study of initial and final allophones of /l/. In *Speech Communication*, 54(3), 368-383. https://doi.org/10.1016/j.specom.2011.10.001

Roettger, T.B., Grice, M. (2019). The tune drives the text – Competing information channels of speech shape phonological systems. In Good, J., Greenhill, S. (Eds.) *Language Dynamics and Change*, 9(2), 265-298. doi: https://doi.org/10.1163/22105832-00902006

Rohlfs, G. (1966). *Grammatica storica della lingua italiana e dei suoi dialetti: fonetica*. Torino: Einaudi.

Schiel, F. (1999). Automatic Phonetic Transcription of Non-Prompted Speech. In *Proceedings of the International Congress of Phonetic Sciences*. San Francisco, August, 607-610.

Schmid, S. (1998). Tipi sillabici nei dialetti dell'Italia settentrionale. In: Ruffino, G. (Eds.) *Atti del XXI Congresso Internazionale di Linguistica e Filologia Romanza*. Tübingen: De Gruyter, 613-625.

Schmid, S. (2014). Syllable typology and the rhythm class hypothesis: Evidence from Italo-Romance dialects. In Caro Reina, J., Szczepaniak, R. (Eds.) *Syllable and Word Languages*. Berlin: Mouton de Gruyter, 421-454. https://doi.org/10.1515/9783110346992.421

Serianni, L., Trifone, P. (1993). *Storia della lingua italiana*, Torino: Einaudi.

Tomasin L. (2010). La cosiddetta elle evanescente del veneziano: fra dialettologia e storia linguistica". In Ruffino, G. and D'Agostino, M. (Eds.) *Storia della lingua italiana e dialettologia*. Palermo: Centro di studi filologici e linguistici siciliani. 729-51.

Turk, A., Shattuck-Hufnagel, S. (2013). What is speech rhythm? A commentary on Arvaniti and Rodriquez, Krivokapić, and Goswami and Leong. In *Laboratory Phonology*, 4(1), 93-118. https://doi.org/10.1515/lp-2013-0005

White, L., Payne, E. & Mattys, S. (2009). Rhythmic and prosodic contrast in Venetian and Sicilian Italian. In Vigário, M., Frota, S. & Freitas, M.J. (Eds.) *Phonetics and Phonology: Interactions and interrelations*, Amsterdam: John Benjamins. pp. 137–158. https://doi.org/10.1075/cilt.306.07whi

Zamboni, A. (1988). Veneto, in Holtus, G. Metzelin & M. Schmitt, C. (Eds.) *Lexicon der romanistischen Linguistik*, Tübingen: Niemeyer, 517-38.

Zmarich, C., Gili Fivela, B. (2005). Consonanti scempie e geminate in italiano: studio cinematico e percettivo dell'articolazione bilabiale e labiodentale. In *Misura dei parametri, Atti del Convegno Nazionale AISV*, Padova, 2-4 dicembre 2004, Torriana (RN): EDK, pp. 429-448.

DOI: 10.17469/O2109AISV000004

MARGHERITA DI SALVO

La situazione comunicativa e la scelta del codice: italiano e dialetto in una comunità migrante

Communicative situation and code selection: Italian and dialect in a migrant community

In this paper, I analyzed code selection and variation in use of dialect and English in four different conversations collected in the Italian community in Bedford (UK). The linguistic variables taken into account in the study are: code selection (Italian vs dialect), use of English, features of the local dialect. The first conversation deals with an interview and the others are spontaneous conversations that one single speaker, Giovanni, has with friends from his own village of origin. All the recordings have in common one speaker, Giovanni, whose linguistic behaviour has been deeply investigated. The results gave evidence that Giovanni prefers dialect with his friends born in Montefalcione and Italian with the interviewer who is not perceived as a member of the Italian community of Bedford. Variation was found also in use of English and in the selection of more conservative variants of the local dialect.

Keywords: selezione di codice, posizionamento identitario, repertorio linguistico, variazione diafasica.

1. *Introduzione*

Negli studi sui meccanismi della selezione di codice nelle società contraddistinte da un bilinguismo stabile, grande importanza è assegnata alla situazione comunicativa e all'interlocutore, fattori che spingono i locutori a selezionare, tra le varietà in loro possesso, quella reputata più appropriata al singolo evento comunicativo. Già Gumperz, distinguendo tra commutazione legata al discorso e commutazione legata all'organizzazione conversazionale, ha dimostrato come la selezione di codice nelle società multilingui possa essere influenzata dalle relazioni sociali tra gli attori coinvolti nell'attività interazionale (Gumperz, 1964, 1982). Per la situazione italoromanza, i lavori sull'alternanza di codice, a partire da Alfonzetti (1992), dimostrano come la selezione del codice non sia casuale e che sia dipendente dall'organizzazione della conversazione e della costellazione dei partecipanti all'evento comunicativo.

La selezione di codice può essere inquadrata all'interno della teoria della rappresentazione del sé in pubblico elaborata dal sociologo Ervin Goffman (1997). Essa si basa sul presupposto che ogni attore sociale, in ogni interazione pubblica, mette in scena il suo personaggio. Inoltre, le teorie costruzioniste dell'identità

(Remotti, 1996), così come sono state declinate all'interno di paradigmi di ricerca sociolinguistici (De Fina, 2007, 2015; Rubino, 2014), hanno evidenziato come la selezione del codice spesso venga adoperata dai parlanti per motivazioni correlate all'espressione della propria identità (sociale, etnica). In questi studi, oggetto privilegiato di indagine sono spesso le comunità migranti e l'alternanza presa in esame coincide, nella maggioranza dei casi, con lo studio dei meccanismi di alternanza tra la lingua di origine e la lingua dominante della società. Questo riflette la tendenza, alimentata anche dal recente paradigma che definisce le lingue deterritorializzate come *heritage languages* o lingue ereditarie (Aalberse, Muysken & Backus, 2019; Polinsky, 2018; Rothman, 2009; Polinsky, Scontras, 2020), a considerare le comunità migranti come portatrici di un bilinguismo sbilanciato tra due sole varietà, la lingua ereditaria da un lato e la lingua dominante dall'altro. In questo paradigma, non risulta contemplata la *dilalia* di italiano-dialetto (Berruto, 1995), che è tuttavia un carattere intrinseco della situazione sociolinguistica italiana. Se, infatti, è comunemente accettato che il repertorio linguistico interno ai confini nazionali sia contraddistinto da un plurilinguismo endogeno di italiano, dialetto/minoranze linguistiche (De Blasi, 2009; D'Agostino, 2015; Berruto, 1995, 2012) e che la relazione tra italiano e dialetti italoromanzi è inquadrabile, in accordo con la proposta di Gaetano Berruto (1995), nella definizione di dilalia, tale complessità non viene sufficientemente problematizzata nei modelli di repertorio linguistico degli italiani all'estero. Gli studi sulle comunità italiane nel mondo sembrano infatti polarizzarsi intorno ai due estremi del segmento, con, a un lato, quegli studi che assumono ad oggetto di ricerca il dialetto X parlato oltre i confini nazionali (Goria, 2015), e con, all'altro, quelli che adottano come oggetto di analisi le varietà dell'italiano all'estero (Bonfatti-Sabbioni, 2018; Caruso, 2010). Più raramente la bibliografia sulle comunità italiane nel mondo ha tenuto conto della compresenza, all'interno dei repertori migrati, delle due varietà (il dialetto e l'italiano). Tale compresenza è stata descritta ora nell'ottica del contatto linguistico, soprattutto in quelle forme che coinvolgono la lingua dominante del Paese di arrivo (Di Salvo, 2012, 2018; Rubino, 2014; Cerruti, Goria 2021), ora nell'ottica di descrizione del repertorio linguistico delle comunità migrate (Di Salvo, 2019). Si collocano, ad esempio, in questa seconda prospettiva, le considerazioni di Haller (1987: 396) sulla storia (linguistica) della comunità italiana di New York, che permettono di individuare varietà di italiano, varietà di dialetto e soprattutto varietà frutto di un processo di koineizzazione dei diversi dialetti presenti nello spazio sociale italiano della Grande Mela.

Tale quadro composito è anche confermato dalla più recente indagine condotta nell'ambito del progetto ERC "Microcontact", diretto da Roberta D'Alessandro (Andriani et al., 2022), che documenta anche una diversa distribuzione tra il dialetto adoperato nella comunicazione familiare, e "Italian koine", che funge da "community language". Tuttavia, in questo studio tale formulazione non è supportata da una prolungata osservazione sul campo che, attraverso l'adozione di una prospettiva etnografica, avrebbe potuto consentire di capire i meccanismi che influenzano la scelta dell'una o dell'altra varietà nelle diverse ondate migratorie e nelle diverse

generazioni di parlanti. Lo stesso spazio composito, definito in termini di spazio linguistico globale italiano, è quello documentato da un gruppo di ricercatori coordinati da Turchetta e Vedovelli (2018) che, attraverso molteplici metodi di ricerca (osservazione partecipante e non, questionari percettivi, descrizione dei panorami linguistici urbani, raccolta e analisi di interviste biografiche), hanno evidenziato come a Toronto lo spazio linguistico italiano è formato in realtà da un continuum di varietà, forme di italiano neostandard, dialetti più o meno conservativi che, per effetto del contatto con la lingua dominante e con le altre lingue presenti nel tessuto urbano, si presentano spesso ricchi di innovazioni (cfr. Di Salvo, 2022 e Nagy, 2022 per alcune innovazioni a proposito della marcatura differenziale dell'oggetto).

Inoltre, il modello tripartito della *Storia linguistica dell'emigrazione italiana* (Vedovelli, 2011) ha il merito di sottolineare come, nell'arco di oltre centocinquant'anni di emigrazione, siano partite persone con profili sociolinguistici diversi. A ridosso dell'unità nazionale e fino alla prima guerra mondiale, la dialettofonia era prevalente rispetto dell'italofonia (De Mauro, 1963), con la conseguenza che a partire erano, in questa fase, soprattutto dialettofoni. Con la diffusione dell'italiano entro i confini nazionali, è da rilevare una competenza di forme più o meno regionali di italiano all'interno dei repertori migrati.

Una composizione bipartita, quella dei repertori italiani all'estero, che il paradigma ancorato alla nozione di lingua ereditaria (Rothman, 2009; Polinsky, Scontras, 2020; Aalberse, Muysken & Backus, 2019) non sembra riconoscere in quanto presuppone, di prassi, una lingua non dominante, trasmessa spontaneamente, e diversa dalla lingua dominante della società, come nella definizione di Polinsky (2018: 3):

> Narrowly defined, heritage speakers are individuals who were raised in homes where a language other than the dominant community language was spoken, resulting in some degree of bilingualism in the heritage language and the dominant language (Valdés, 2000). A heritage speaker may also be the child of an immigrant family who abruptly shifted from her first language to the dominant language of her new community. Crucially, the heritage speaker began learning the heritage language before, or concurrently with, the language which would become the stronger language. That bilingualism may be imbalanced, even heavily imbalanced, in favor of the dominant language, but some abilities in the heritage language persist.

Sia nel filone di ricerca più orientato in senso formalista sia nel filone di ricerca sociolinguistico, si presuppone che i migranti all'estero siano monolingui e che quindi la lingua ereditaria sia solo una, solitamente definita *italiano*: nello studio di Bonfatti-Sabbioni (2018), considerato rappresentativo del primo approccio, non si fa riferimento alla possibilità che, per lo meno i membri della I generazione abbiano una competenza di un qualsiasi dialetto italoromanzo; dall'altro, nel lavoro sui clitici nel friulano in Argentina e in Brasile di Frasson, D'Alessandro e van Osch (2021), il tema del contatto tra dialetto e italiano viene parimenti sottaciuto.

Inoltre, soprattutto alla luce del recente dibattito sull'italiano quale lingua ereditaria, sembra utile riconoscere la funzione del dialetto e dell'italiano nella trasmissione familiare: se, infatti, per poter individuare una *lingua ereditaria,* è necessario

che essa venga trasmessa spontaneamente dalla prima generazione alla successiva e che essa sia successivamente sostituita nella quotidianità dalla lingua dominante della società, allora è importante capire se i migranti italiani tendano a trasmettere l'italiano o il dialetto e se tale comportamento sia costante nelle tante comunità italiane o, se, piuttosto, vi siano comportamenti diversi in relazione a precise caratteristiche della società di approdo. Questa seconda posizione è supportata da alcune evidenze relative alla trasmissione di italiano e dialetto nel mondo: studi precedenti hanno dimostrato una maggiore tendenza alla conservazione delle varietà di origine in Europa rispetto al Nord America (Di Salvo, 2020) e al Canada in particolare, ma dall'altro una significativa variazione da contesto a contesto all'interno dell'Europa (Moreno, Di Salvo, 2015) e dello stesso Paese europeo (Di Salvo, 2012). I primi studi, da un lato, hanno dimostrato che in Canada l'italiano è conservato meno che in Europa, anche per la frattura più profonda con il Paese di origine che contraddistingue i contesti extraeuropei: adoperando un questionario percettivo, infatti, è stato evidenziato come non solo i migranti stanziati in Canada si considerano meno italiani di quelli in Europa, ma riportano di usare meno che in altri contesti europei sia l'italiano sia il dialetto; per quanto riguarda il versante della trasmissione intergenerazionale, i migranti italiani in Canada trasmettono meno l'italiano e il dialetto rispetto a coloro che, al contrario, sono stanziati in Europa. Tuttavia, anche il contesto europeo è contraddistinto da una forte variazione in quanto i tassi di conservazione e di trasmissione di italiano e dialetto variano da contesto a contesto: lo dimostra, ad esempio, lo studio comparativo condotto da Moreno e Di Salvo (2015) che, a partire da un comune strumento di rilevazione, hanno evidenziato come a Liegi l'italiano sia meno usato che a Bedford, contesto, quest'ultimo, contraddistinto da una forte tendenza alla conservazione delle lingue di origine.

La maggiore tendenza alla conservazione della comunità italiana di Bedford viene anche confermata in studi condotti a partire dal comportamento linguistico dei migranti: in Di Salvo (2012), ad esempio, viene mostrato che nella comunità di Bedford il dialetto sia una sorta di codice simbolo della coesione della comunità italiana e ciò si traduce in una tendenza alla conservazione di questa varietà piuttosto che dell'italiano; a Cambridge, al contrario, i parlanti dichiarano di usare poco il dialetto e di ricorrere soprattutto all'italiano che è preferito nella trasmissione intergenerazionale (Di Salvo, 2012). Questo studio, basato su dati di tipo percettivo e sulla comparazione del comportamento in sede di intervista, non permette tuttavia di verificare i diversi meccanismi di selezione di codice all'interno della comunità italiana in quanto non fornisce dati sull'uso concreto in diverse situazioni comunicative. Tuttavia, la variazione tra comunità è alla base della recente proposta di sistematizzazione teorica di Aalberse, Backus e Musken (2019) che propongono di analizzare le singole comunità migranti al fine di individuare le tendenze specifiche che, accanto a quelle più generali, caratterizzano i singoli scenari migratori.

2. *Obiettivi*

Il presente contributo indaga i meccanismi di selezione del codice e la preferenza di caratteristiche strutturali (cfr. §4) più o meno conservative del dialetto adoperato da un migrante irpino residente nella città inglese di Bedford. Obiettivo consiste nel dimostrare che, per la descrizione delle comunità italiane all'estero, si debba proiettare, nel contesto di immigrazione, il bilinguismo endogeno di italiano e dialetto. Si ipotizza infatti che italiano e dialetto siano compresenti nel repertorio della prima generazione migrata e che la selezione di una varietà piuttosto che dell'altra possa essere determinata dalla situazione comunicativa, dalla relazione con l'interlocutore e dal tipo di evento (formale/informale) in cui i parlanti sono impegnati. È altrettanto verosimile che tali variabili esterne (relazione con l'interlocutore, situazione comunicativa) possano influenzare anche la selezione di tratti del dialetto più o meno conservativi: nelle pagine che seguono, mi propongo di verificare se esiti conservativi del dialetto siano più ricorrenti nelle conversazioni con i compaesani piuttosto che nelle interazioni con persone percepite come esterne alla comunità di appartenenza. Ciò avrebbe come conseguenza il dover declinare al plurale il concetto di lingua ereditaria, per lo meno per quei casi, ampiamente documentati in letteratura (Di Salvo, in stampa) in cui a spostarsi sono persone in possesso di un repertorio multilingue e che, di prassi, alternano (anche con i propri figli) le proprie lingue di origine.

L'analisi mira a descrivere la selezione di italiano e dialetto e la variazione che si osserva nell'alternanza con l'inglese e nella preferenza per esiti più o meno conservativi del dialetto in situazioni comunicative contraddistinte da un diverso livello di formalità e da una diversa relazione tra parlanti. Si è scelto di comparare una situazione formale, l'intervista, con interazioni dal carattere più spontaneo che vedono impegnati migranti provenienti dallo stesso paese e amici. La mia ipotesi è che la prima situazione comunicativa, anche sulla base del profilo sociolinguistico della comunità riassunto al paragrafo successivo, possa spingere i migranti di I generazione verso l'italiano (regionale), mentre la seconda possa invece favorire la selezione del dialetto. Intendo anche dimostrare come nel primo caso le varianti conservative del dialetto sono meno presenti, mentre nel secondo esse tendono a essere più frequenti. Infine, mi propongo di verificare se anche le forme dovute al contatto con la lingua dominante (l'inglese) siano soggette a variazione nei due tipi di eventi comunicativi osservati. Si potrebbe infatti supporre che l'uso dell'inglese sia limitato a forme necessarie sul piano pragmatico in quanto relative ad un lessico specifico (Backus, 1999, 2001; Del Vecchio, 2023) nell'intervista, mentre la conversazione con compaesani e membri della propria rete possa incoraggiare l'adozione di forme conversazionali e pragmatiche che rinforzano la cooperazione conversazionale e la presenza di una conoscenza condivisa come i segnali discorsivi e le marche pragmatiche.

3. *Il contesto della ricerca*

La storia della comunità italiana di Bedford ha inizio a giugno 1951, quando arrivò nella città inglese il primo contingente di immigrati, reclutati mediante un'agenzia di collocamento aperta a Napoli dalla principale industria di mattoni britannica con l'obiettivo di arruolare manodopera poco qualificata nell'ambito di accordi intergovernativi stipulati tra il Ministero del lavoro britannico e il governo italiano a tale scopo. L'immigrazione da lavoro riguardò sia gli uomini sia le donne, che, nell'ambito di questa tipologia di accordo intergovernativo, furono destinate alla fabbricazione di dolciumi e successivamente al settore delle pulizie di uffici e di strutture pubbliche.

Per l'immigrazione maschile, prevalente nella prima metà degli anni Cinquanta, gli accordi intergovernativi prevedevano che i migranti, per i primi quattro anni, dovessero rimanere legati all'azienda che li aveva reclutati; erano anche costretti a vivere in ostelli che le industrie di mattoni mettevano a loro disposizione. Solo dopo questo periodo di tempo, gli italiani furono liberi di lasciare gli ostelli e si spostarono gradualmente nel quartiere a ridosso della stazione ferroviaria che divenne ben presto il cuore pulsante della comunità.

A partire dalla seconda metà degli anni Cinquanta, il sistema ufficiale di reclutamento fu gradualmente soppiantato dalle catene migratorie che favorirono l'arrivo da aree specifiche del meridione italiano: la provincia di Avellino e, al suo interno, il comune di Montefalcione; la provincia di Campobasso e il comune di Busso; l'agrigentino e il comune di Sant'Angelo Muxaro (cfr. Colpi, 1991).

I migranti arrivati tra gli anni Cinquanta e Sessanta hanno avuto generalmente come lingua materna il dialetto e, al momento dell'arrivo in Inghilterra, non avevano nessuna competenza dell'inglese, lingua che non è diventata dominante, per lo meno per la prima generazione migrata: studi precedenti basati su una combinazione di questionari percettivi e analisi del comportamento della I generazione (Di Salvo 2011, 2012) hanno infatti dimostrato che i membri della I generazione si considerano solo raramente capaci di comprendere e parlare la lingua dominante della società, che è anche scarsamente adoperata in sede di intervista (Di Salvo, 2012). I dati di inchieste precedenti (Di Salvo, 2012, 2019) dimostrano anche come il dialetto sia particolarmente vitale: si riportano, a sostegno di tale ipotesi, i risultati dello studio di Di Salvo (2012) che sono stati ottenuti mediante un questionario auto-valutativo raccolto con 150 migranti di I generazione che riportano una scarsa competenza percepita in inglese da un lato e una maggiore competenza in dialetto dall'altro, tanto nella comprensione quanto nella produzione attiva-(il parlato). La tabella 2, tratta dal medesimo studio, contiene i risultati relativi all'uso (percepito) in alcuni domini (con i paesani, con amici corregionali, con estranei, ...) e dimostra una correlazione percepita tra preferenza del dialetto e interlocutore paesano. Non solo quindi il dialetto è considerato, dai membri della I generazione, il codice di cui si ha una maggiore competenza, ma è anche quello preferito per rinsaldare la coesione su scala paesana.

Tabella 1 - *Percezione relativa alla capacità di comprendere italiano, dialetto e inglese da parte di 150 migranti di I generazione residenti a Bedford (valori percenutali)*

	Varietà	*Bene*	*Così e così*	*Poco*	*Niente*
Comprensione	*Dia*	87	12	0	1
	Ita	89	10	1	0
	Ing	53	43	4	0
Parlato	*Dia*	85	13	1	1
	Ita	84	10	6	0
	Ing	52	43	5	0

Tabella 2 - *Percezione del comportamento linguistico interno/esterno alla famiglia (valori percentuali)*

	DIA	*ITA*	*ING*	*ITA E DIA*	*ITA E ING*	*DIA E ING*	*N.P.*
A casa	45,52	17,93	13,10	8,97	5,52	6,21	2,76
Con i paesani	63,45	15,17	1,38	9,66	1,38	2,07	6,90
Con i corregionali	54,48	18,62	1,38	14,48	2,07	7,59	1,38
Con gli italiani	14,48	37,24	1,38	35,86	4,14	6,90	0,00
Con gli estranei	2,07	0,00	97,93	0,00	0,00	0,00	0,00
Con un estraneo italiano	7,59	73,10	2,07	10,34	0,69	6,21	0,00
Nei negozi italiani	6,90	69,66	2,07	6,90	7,59	5,52	1,38

Sulla base di questi dati, è stato ipotizzato che il dialetto sia, più dell'italiano, da considerare come lingua ereditaria in quanto, come molti membri della seconda generazione hanno confermato, la loro socializzazione primaria interna alla famiglia è spesso avvenuta prevalentemente (anche se non esclusivamente) in dialetto; la competenza dell'italiano è stata successivamente consolidata attraverso percorsi di istruzione formale, generalmente offerti dal Consolato italiano. L'inglese è diventato dominante solo nella seconda generazione, mentre è scarsamente adoperato dai membri della prima, oggi molto anziani.

L'uso del dialetto sembra essere prevalente, sulla base dei dati già ricordati e confermati in successive campagne di osservazione (Di Salvo, 2019), anche nelle interazioni informali con i migranti di origine paesana, mentre l'italiano è riservato alle conversazioni con gli estranei. Sembra sussistere quindi un sistema di selezione di codice tra italiano e dialetto che va indagato mediante appositi strumenti che permettano sia di confermare tale correlazione sia di comprendere la variazione sulla base della situazione comunicativa in relazione a specifiche variabili linguistiche (cfr. §4).

4. *Metodi della ricerca*

In questo contributo, assumiamo come punto di osservazione il comportamento di un unico informatore, Giovanni, nato a Montefalcione nel 1950 ed emigrato nella città inglese di Bedford nel 1969. Il suo comportamento è osservato a partire da quattro diverse interazioni, diverse in base al livello di formalità e ai ruoli dei partecipanti. La prima è da una tradizionale intervista libera condotta dalla sottoscritta, poco dopo aver conosciuto il parlante. Le restanti conversazioni sono costituite da interazioni spontanee in cui Giovanni ha interagito con alcuni compaesani. In queste conversazioni il raccoglitore era presente, ma non ha condotto l'interazione e ha lasciato che fosse il parlante intervistato in prima battuta a raccogliere dati di parlato. Il parlante osservato, in particolare, conversa con suoi compaesani che egli stesso ha coinvolto nella ricerca e ai quali ha presentato il raccoglitore. In queste conversazioni, quindi, il parlante non è più l'osservato, ma diventa in qualche modo il regista della conversazione.

È opportuno sottolineare la distanza delle situazioni osservate: nel primo caso (l'intervista), il parlante ha davanti a sé un ricercatore esterno alla comunità e questo concorre a rendere lo scambio comunicativo più formale; i ruoli son ben definiti in quanto il raccoglitore fa domande mentre il parlante risponde, con una rigida conformazione degli scambi di turno. Nel secondo caso, sono state generalmente osservate occasioni in cui i parlanti interagiscono con i membri più vicini della loro rete sociale, familiari o amici, spesso compaesani.

Di seguito si fornisce uno schema riassuntivo delle principali caratteristiche situazionali delle quattro registrazioni analizzate:

Tabella 3 - *Caratteristiche del campione, tipo di intervista e costellazione dei partecipanti*

Registrazione	*Tipo*	*Presenti*	*Luogo*	*Relazione con interlocutore*
1	Intervista	Giovanni Raccoglitore (per una porzione minima e finale un parlante nato a Bedford da genitori irpini)	Casa di Giovanni	Il raccoglitore (R), da poco arrivato nella comunità, concorda un appuntamento con Giovanni per un'intervista sugli italiani e l'italiano a Bedford

Registrazione	*Tipo*	*Presenti*	*Luogo*	*Relazione con interlocutore*
2	Conversazione spontanea	Giovanni Giuseppe Armando	Casa di Giuseppe	Giovanni conduce il raccoglitore a conoscere altri migranti nati a Montefalcione
3	Conversazione spontanea	I coniugi Tonino Carmela	Casa di Tonino e Carmela	
4	Conversazione spontanea	Raffaella e Maria (sorelle) e, per la prima porzione di conversazione, anche il marito di Raffaella	Casa di Raffaella	

Di seguito un prospetto di tutti i parlanti coinvolti nella ricerca, con particolare riferimento ai parametri biografici e alla relazione con Giovanni:

Tabella 4 - *Prospetto delle caratteristiche sociobiografiche dei parlanti coinvolti nello studio*

Parlante	*Comune di nascita*	*Genere*	*Generazione*	*Età*	*Anno di arrivo in Inghilterra*	*Relazione con Giovanni*
Giovanni	Montefalcione	M	I gen	70	1969	
Giuseppe	Montefalcione	M	I gen	80	1954	Parentela
Armando	Montefalcione	M	I gen	65	1964	Amicizia
Tonino	Montefalcione	M	I gen	50-60	1964	Amicizia
Carmela	Bedford	F	0 gen	50-60	Nata a Bedford	Amicizia
Raffaella	Montefalcione	F	I gen	72	1963	Amicizia
Maria	Montefalcione	F	I gen	70	1953	Amicizia
Marito di Raffaella	Calvi di Sotto	M	I gen	-	-	Amicizia

4. *Le variabili osservate e il corpus*

Scopo dell'analisi è valutare l'eventuale variazione tra la prima intervista e tre registrazioni successive con particolare riferimento alle seguenti variabili linguistiche:

a. Selezione del codice (italiano, dialetto e inglese);
b. Usi e forme dell'inglese;
c. Caratteristiche conservative del dialetto in relazione a: esiti del dimostrativo, enclisi del modificatore possessivo; forme del pronome tonico soggetto.

Il presente contributo intende preliminarmente indagare se la situazione comunicativa e la relazione con l'interlocutore spingano il parlante osservato a optare per l'italiano o per il dialetto, in modo da fornire una descrizione dei valori comunitari associati ad entrambi i codici. In secondo luogo, l'analisi si propone di capire se le

variabili esterne connesse alla situazione comunicativa influenzino anche le forme del contatto con l'inglese, sia per quanto riguarda la quantità di materiale della lingua dominante sia per quanto riguarda le funzioni pragmatiche che essa riveste. Su questo aspetto, in particolare, intendo dimostrare che l'inglese è correlato con particolari campi semantici (il lavoro in primo luogo) solo se il parlante interagisce con un interlocutore esterno alla comunità (il raccoglitore), mentre le forme dell'inglese riguardano anche altri settori della vita quando, invece, si rivolge a suoi compaesani.

Per quanto riguarda l'ultima variabile, lo studio si propone di verificare se, in conversazioni di natura diversa (intervista vs conversazione spontanea con compaesani), il parlante selezioni variabili più o meno vicine al dialetto di Montefalcione in relazione a tre specifici varianti che distinguono questa varietà dal napoletano, da altre varietà linguistiche limitrofe e soprattutto dall'italiano: l'obiettivo di questa sezione del lavoro consiste nel capire se la selezione di varianti più o meno italianizzate sia condizionata dalla situazione comunicativa e, in particolare, se sussista una correlazione tra varianti conservative del dialetto e interlocutori montefalcionesi da un lato, e tra varianti italianizzate e interlocutore esterno alla comunità. Per questa porzione dell'analisi, lo studio è limitato alle tre varietà menzionate in precedenza.

Per quanto riguarda il dimostrativo, studi precedenti (Di Salvo, 2019, 2022) hanno dimostrato come la forma patrimoniale del dimostrativo del dialetto di Montefalcione, *kwiro* per il maschile e per il neutro e *kwira* per il femminile, sia contraddistinta da due diversi fenomeni linguistici che lo distinguono da quello presente in aree limitrofe, il mantenimento della labiovelare dopo l'occlusiva velare e il rotacismo del nesso latino -LL-[1]. Sulla base di studi precedenti condotti sulla comunità montefalcionese a Bedford, è stato dimostrato che, accanto a tali esiti patrimoniali, sono attestati in parlanti di origine montefalcionese residenti stabilmente nella città inglese anche varianti diverse, che si polarizzano:

a. sul modello italiano (*kwello* per il maschile e per il neutro, *kwella* per il femminile);
b. sul dialetto napoletano (*killo* per il maschile e per il neutro, *killa* per il femminile).

Meritano un ulteriore approfondimento le forme *kwillo/kwilla* e *kiro/kira* che presentano, in maniera diversa, solo uno dei processi propri del dialetto montefalcionese (il mantenimento dell'approssimante labiovelare la prima; il rotacismo del nesso laterale geminato la seconda). Secondo la schematizzazione proposta in tabella, queste varianti sono disposte lungo un continuum che va dall'italiano (*kwello/kwella*) al dialetto (*kwiro/kwira),* passando per forme intermedie tra italiano e dialetto (*kwillo/kwilla; kiro/kira*), e napoletano (*killo/killa*).

[1] Una descrizione dell'estensione diatopica di tali esiti è fornita in Di Salvo e Guzzo (2021).

Tabella 5 - *Prospetto delle forme del dimostrativo*

Codice	*Forma*	*Fenomeni presenti*	
		Mantenimento dell'approssimante labiovelare dopo velare	Rotacizzazione del nesso latino geminato -LL-
Italiano	*kwello/kwella*	+	-
Dialetto montefalcionese	*kwiro/kwira*	+	+
Dialetto napoletano	*killo/killa*	-	-
Forme ibride	*kwillo/kwilla*		-
	kiro/kira	-	+

Per quanto riguarda l'enclisi del modificatore possessivo, in area avellinese e nel dialetto di Montefalcione essa è presente sia con i nomi di parentela sia con il lessema *casa;* nelle forme dell'italiano regionale, ancora, l'aggettivo è di prassi collocato in posizione postnominale (Telmon, 1993), mentre in italiano standard il modificatore è posto tra il determinante e il nome.

Tabella 6 - *Prospetto della collocazione del modificatore possessivo in dialetto, italiano regionale e italiano standard*

Dialetto montefalcionese	**Italiano regionale**	**Italiano standard**
Enclisi del possessivo	Det + N + Modificatore possessivo	Det + Modificatore possessivo + N

Tre varianti sono anche quelle osservate nella comunità montefalcionese residente a Bedford (Di Salvo, 2019): il tipo italiano *lui/lei,* la forma patrimoniale del dialetto *illo/illa* e, infine, la variante *isso/issa* che, dal napoletano, si sta gradualmente estendendo anche in area irpina.

Il corpus sottoposto ad analisi è formato da quattro registrazioni che hanno un partecipante comune, Giovanni, nato a Montefalcione e residente nella comunità di Bedford. Le registrazioni hanno una durata compresa tra i 60 e i 110 minuti.

I materiali raccolti sono stati sottoposti ad una duplice analisi: in una fase preliminare, sono stati considerati tutti gli scambi di turno compresi nei primi 30 minuti di ciascuna intervista (per un totale di 2 ore di parlato spontaneo). Per ognuna di esse, è stato individuato il codice del turno di Giovanni, il codice del turno precedente per valutare l'adeguamento del parlante osservato al suo interlocutore. Per procedere con l'analisi quantitativa, i turni (di Giovanni e del suo predecessore) sono stati distinti in base ai parametri riassunti alla figura successiva:

Tabella 7 - *Selezione di codice: varianti individuate e sigle adoperate*

Variabile	*Varianti*	*Sigla adoperata*
Interlocutore	Raccoglitore	R
	Compaesano	P
Codice del turno	Dialetto	DIA
	Italiano	ITA
	Inglese	ING
	Dialetto e italiano	DIA – ITA
	Italiano e inglese	ITA – ING
	Dialetto e inglese	DIA – ING

L'analisi delle variabili al punto c è stata limitata alle due ore individuate (30 minuti per ciascuna intervista) mentre l'analisi dell'alternanza con l'inglese all'intero corpus raccolto (circa 5 ore).

6. *Risultati dell'analisi*

6.1 La selezione di codice

Sono stati sottoposti ad analisi 976 turni (488 realizzati da Giovanni e 488 da altro partecipante). Per quanto riguarda i turni realizzati da Giovanni, essi sono successivi/diretti al raccoglitore nel 37,3% dei casi (182 turni) e a quelli di un compaesano nel 62,7% di casi (306 turni). La selezione di codice, nella nostra prospettiva di analisi, non è meccanica ma frutto di una negoziazione tra i partecipanti all'interazione e, per il suo studio, pertanto, è necessario prendere in esame non solo il comportamento del singolo parlante osservato ma più in generale il sistema di coppie adiacenti, secondo quanto indicato dagli studi di analisi della conversazione.

I dati quantitativi, la cui lettura permette di far emergere le tendenze e la regolarità di alcune scelte, mostrano come, nelle conversazioni informali, i montefalcionesi tendono a selezionare nella quasi totalità dei turni (82,35%) il dialetto, mentre il raccoglitore fa scelte opposte:

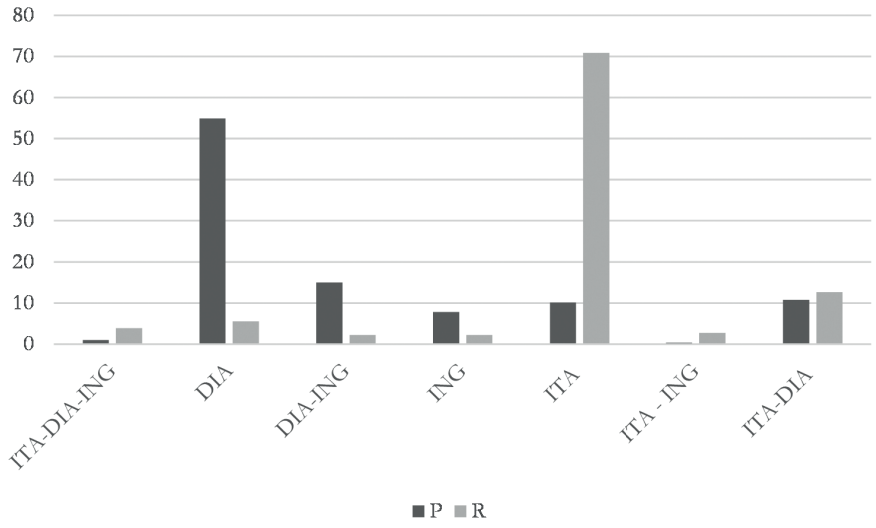
Tabella 8 - *Selezione di codice tra italiano (ITA), dialetto (DIA) e inglese (ING) da parte degli interlocutori di Giovanni*

	Selezione di parlanti nati a Montefalcione	*Selezione da parte del raccoglitore*
DIA	82,35	1,65
ING	2,94	0,00
ITA	14,71	98,35

I turni in dialetto precedenti a quelli di Giovanni sono quasi prevalentemente realizzati da compaesani, mentre l'80% di quelli in italiano è realizzato dal raccoglitore.

I riflessi di questa scelta divergente, non effettuata da sola dagli interlocutori ma negoziata con Giovanni, sono evidenti nel comportamento di quest'ultimo che, con regolarità, preferisce il dialetto con i compaesani e l'italiano con il raccoglitore (v. Fig, 1):

Figura 1 - *Scelta del codice con compaesani (P) e raccoglitore (R) (valori percentuali)*



Il grafico evidenzia la disparità tra italiano e dialetto e la scarsa tendenza ad alternare le varietà: in particolare, i turni in tre lingue (italiano, dialetto e inglese) rappresentano una percentuale irrisoria del totale (10 turni su 488).

In dettaglio, alla tabella successiva, il quadro delle corrispondenze tra scelte di Giovanni in relazione al suo predecessore, conferma la negoziazione che prevede un accomodamento continuo: alla selezione del dialetto corrisponde il dialetto e così parimenti per l'italiano.

Tabella 9 - *Selezione di codice da parte di Giovanni in relazione al codice usato al turno precedente (valori percentuali)*

Turni di Giovanni per codice linguistico	*DIA*	*ING*	*ITA*
ITA-DIA-ING	10,00	0,00	90,00
DIA	84,83	3,93	11,24
DIA-ING	72,00	4,00	24,00
ING	52,63	0,00	47,37
ITA	10,00	0,00	90,00
ITA-ING	16,67	0,00	83,33
ITA-DIA	53,57	0,00	46,43

La preferenza del dialetto tra compaesani sembra dipendere dal valore sociale attribuito dai parlanti al dialetto che, come confermato anche dai dati quantitativi elicitati mediante inchieste percettive, assolve alla funzione di codice che rinsalda i vincoli della solidarietà paesana, fungendo da *we code*.

6.2 Il contatto con l'inglese

L'alternanza di codice è stata indagata a partire dal quadro teorico elaborato da Backus (1999, 2000, 2001) e Muysken (2000) che distinguono tra *alternation, insertion*, e *congruent lexicalization*. Tali forme sono state studiate tenendo insieme sia il piano strutturale sia gli aspetti semantici legati al contatto: questi ultimi, infatti, sembrano essere di estrema importanza nelle comunità migranti in quanto, come suggerito da Backus (1999, 2001), motivazioni semantiche possono incoraggiare l'utilizzo della lingua dominante.

Le forme dovute a contatto sono 170, con una prevalenza (111, pari al 65,2% del totale) nelle interazioni con i compaesani. Tuttavia, se si considera la disparità del numero di conversazioni (3 interazioni spontanee e 1 intervista guidata), il dato va ridimensionato: il numero medio di forme dell'inglese per interazione spontanea è di 37, inferiore a quello rilevato nell'intervista condotta dal raccoglitore (59). L'analisi della distribuzione delle forme dovute a contatto per tipologia strutturale consente alcune riflessioni: non sembra esserci variazione in base all'interlocutore per quanto riguarda il *codeswitching* interfrasale (*alternation*) e l'*insertional codemixing* (*insertion*). La sola variazione quantitativa riguarda la maggiore concentrazione dei segnali discorsivi inglesi con gli interlocutori originari di Montefalcione: la presenza di segnali discorsivi inglesi accomuna tutti i parlanti intervistati, che, come mostrato in alcuni studi condotti a partire dal corpus da me raccolto nella comunità italiana di Bedford (Di Salvo 2013), sono adoperati con le medesime funzioni pragmatiche descritte per l'inglese parlato[2].

L'analisi qualitativa dimostra come questa tipologia contattuale sia utilizzata da Giovanni (e dagli altri parlanti) per consolidare la loro coesione e il loro reciproco riconoscersi in un gruppo: questo comportamento accomuna Giovanni agli altri parlanti intervistati.

Nei turni di Giovanni, non sussiste una correlazione tra ricorso all'inglese da parte del parlante e presenza dell'inglese al turno precedente: questa condizione, infatti, si verifica solo nel primo esempio, ma non negli altri due. Si potrebbe quindi ipotizzare che la presenza di questo tipo di commutazione da parte di Giovanni sia il riflesso di un comportamento oramai abituale, in accordo con la prospettiva incentrata sull'uso e la frequenza proposta da Backus (1996), la cui validità è confermata anche da studi empirici condotti nelle comunità italiane in Inghilterra (Del Vecchio, 2023).

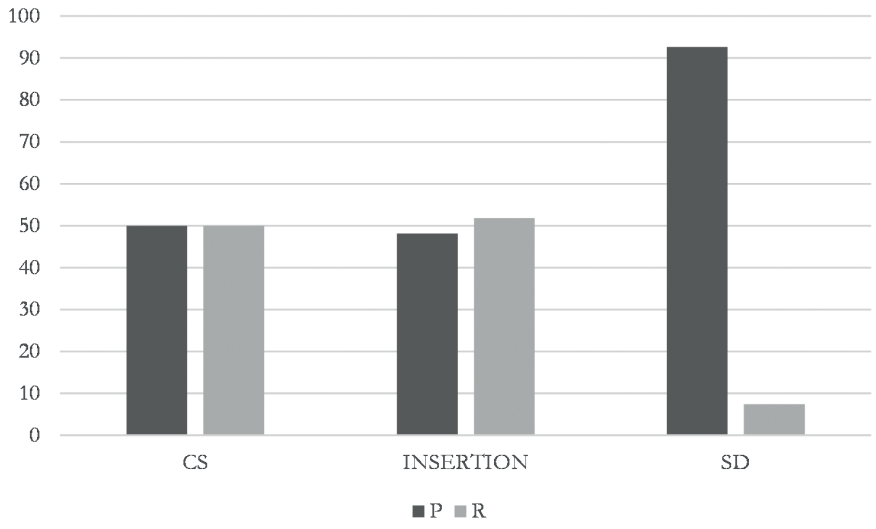
[2] Uno studio dedicato al segnale discorsivo *yeah* è in preparazione: la scelta di dedicare uno studio approfondito a tale segnale discorsivo dipende dall'elevata frequenta con cui compare nei testi e dall'ampio ventaglio di caratteristiche comunicative che esso assume all'interno dei testi.

(1) Giovanni: quelle so...[3]
R: però non pizzicano
Giovanni: no quelle si chiamano zanzare ... dragon
Carmela: dragon fly
Tonino: dragon fly / yeah
Giovanni: yeah / l'elicottero / longa longa

(2) Giovanni: mett e... e fasulə/ e fann a tutte parte
R: ma pəcché so fasulə?
Armando: so fagioli eh
Giovanni: yeah/ quelli grossi però / so quelli gruossi kwisti

(3) Giovanni: specialmente a Maria/ co' Maria egg fatt duj ann re... re scole elementari
Raffaella: a l'Italia?
Giovanni: yeah/ ero piccolo/ po' però/ cocchi vot ievo a Candida/ i dda funtana fu la...a funtana Licia c'ha...mh / sarebbe... a sor / a sorgente / ten a sorgente a piglià l'acqua per papà/ e mamma dicimo / e tənemm a terra a cartularo / proprio sott a funtana / chiù sott a funtan / e po' tenemm a terra... chiù luntan vvicin a stazion ro trenə

Figura 2 - *Codeswitching (CS), inserzioni (insertion) e segnali discorsivi (SD) inglesi per interlocutore(P: compaesano; R: raccoglitore) (valori percentuali)*



Per l'*insertional codemixing*, l'analisi del campo semantico in cui le forme dell'inglese appartengono ha dimostrato che i dati quantitativi nascondono, in realtà, comporta-

[3] Si riportano le convenzioni adoperata per le trascrizioni: con i nomi propri si indicano i parlanti, con R il ricercatore, con *I* la pausa breve, con // la pausa lunga, con ... le esitazioni, con # i mutamenti di progetto.

menti diversi. Se, infatti, con il raccoglitore, Giovanni utilizza l'inglese in corrispondenza quasi esclusiva al campo semantico lavoro, con i compaesani, al contrario, vi è un maggior numero di campi semantici investiti dal contatto con l'inglese:

(4) Giovanni: *fork lift* / c'ha la forma / c'ho pure il modello io / me lo so comprato pe tenello / *fork lift* è una macchina che c'aveva dieci forche / così / no // proprio come se fosse diecə: ... di ferro / e affianco cə so' e camere d'ariə / e sotto / a terra / ci mettevano ... si chiamava... o *steak up* / erano mattoni alzati / una fila di nove qua / nove... / nove file erano / e c'erano venti mattonə pə partə / questo / chesta forca veniva là / poi sopra ci mettevano mille mattonə / quanto ... ci mettevi l'aria comə l'aria de o ... compressore / tenevə o compressore piccolino / quello là / no / la forca dietro / che... co... / mettevi l'aria / e quello si allargavano / e sə prendevə tutto / non lasciava niente a terrə / tu guardavi dentro il forno / o mettevi dentro / levavi l'aria / e... e si pusava a terra / e stavano mille mattuni / che si chiamava il *bottom* / sarebbe quello di sotto
R: *bottom* / giù
Giovanni: *bottom* / giù // e poi cə stevə il *top* / altrə ... ci mettevamo quelli mattoni colorati / *sun faicing* / e chiammavanə / un'altra qualità di mattoni / però // e questa terra veniva: ... veniva bagnato / i mattoni verdi / e poi c'era questa *mascina* cu lo ... / con l'aria / compressore che c'è ... compressore grande che buttava aria a tuttə parti / e quei tubi buttavano questa sabbia / stava dentro un secchio grande sopra / come un imbuto / ca scendeva piano piano piano / e quest'aria la buttava vicin e mattunə / quelli arrivavano bagnati / perché c'era ... e.... c'era... lo *spray* / che bagnava i mattoni e: ... e facevamo e *sun face* / erano i mattoni colorati / erano rossi / c'erano pure marrò / c'erano *gold buff* / era u colore d'oro / u *gold* / è l'oro / *gold* / ci stavano... tanti colori / verde... mamma mia! / facìumə nottə e giornə

(5) Tonino: non ci ho fatto mai attenzione / invece mia moglie ce l'ha / lei ... lei
R: perché è nata qua?
Tonino: è nata qua / esatto
R: in automatico
Tonino: si sì è in automatico
Giovanni: è *Britìsh*
R: no perché se in Italia uno nasce da immigrati non è italiano
Giovanni: è *Britìsh* / è come fosse... ah non lo fanno in Italia

(6) Raffaella: ma che tien o diabet?
Giovanni: *yeah*// mo sto sei... *six*... sei pu... sei punti cinque/ sei e sei sei e cinque
Marito Raffaella: va be è chiù bass e riec... dieci a nov è...
Giovanni: ma n'amm arrivat ancor
Marito Raffaella: quann mangian nu poc e chiù a ser / arriv pur a triric a volte
Giovanni: a serə / a serə issa mangià poc// chell se putesse mangià pure quatt vote al giorno però a sera tea mangià na fett e pane/ ma manc o pan... pcché o pan ten o ... o *glutìn* arintə
Raffaella: o zuccər arind pur
Giovanni: no ma manc o zuccr o *glutìn* sta arind / dà fastidio quello

Accomuna, tuttavia, entrambe le casistiche la forte tendenza a inserire l'inglese prevalentemente (ma non in maniera esclusiva) in corrispondenza di domini semantici appresi durante l'esperienza migratoria: nell'intervista, il lavoro tuttavia è l'unico settore della vita in cui i dati evidenziano una concentrazione di forme imputabili al contatto, mentre nelle restanti conversazioni vi è una maggioranza di domini semantici interessati.

6.3 Le caratteristiche del dialetto

L'esame delle caratteristiche del dialetto necessita una premessa: sulla base dei dati discussi in 6.1, è stato rilevato come Giovanni seleziona il dialetto quasi esclusivamente nelle interazioni con migranti montefalcionesi, mentre l'uso dell'italiano è limitato all'intervista condotta dal raccoglitore. Come conseguenza di ciò, le varianti italiane del pronome tonico da un lato e le forme enclitiche del modificatore possessivo, tipiche del dialetto, sono presenti in maniera esclusiva rispettivamente nell'intervista e nelle conversazioni spontanee:

Tabella 10 - *Distribuzione delle forme enclitiche del modificatore possessivo con i vari interlocutori (valori percentuali)*

	[-enclisi]	*[+enclisi]*
P	86,21	13,79
R	100,00	0,00

Tabella 11 - *Distribuzione delle forme enclitiche del modificatore possessivo con i vari interlocutori (valori percentuali)*

	ISSO	*ILLO*	*LUI*
R	0	0	100
P	18,75	12,5	68,75

Appare particolarmente interessante, però, la compresenza, nelle interazioni con i migranti montefalcionesi, sia della variante del pronome tonico del dialetto di origine sia di quella di tipo napoletano. Tale compresenza è, a mio parere, sintomatica delle innovazioni linguistiche che contraddistinguono anche i parlanti di I generazione, il cui dialetto diverge da quello di origine: le comunità italiane all'estero infatti sono il luogo in cui cogliere i processi di livellamento interdialettale tra i numerosi dialetti che sono compresenti nei singoli contesti migratori (Goria, Di Salvo, 2023). Nel caso specifico del pronome di III persona singolare, si assiste alla compresenza del tipo montefalcionese e del tipo napoletano: tale compresenza potrebbe essere il frutto di un contatto tra i due dialetti che sono parlati nel medesimo contesto migratorio.

Questo viene confermato dai dati relativi alla distribuzione delle varianti dell'aggettivo e del pronome dimostrativo: distinguendo unicamente sulla base dell'interlocutore, la variante italiana è quasi equamente distribuita nelle due tipologie di

conversazione (intervista vs interazione spontanea) e quindi compare sia con interlocutori interni alla comunità sia con il ricercatore esterno. Ciò è indicativo della presenza, anche negli scambi più orientati sul dialetto quali quelli che contraddistinguono la relazione tra interlocutori montefalcionesi, di tratti dovuti a interferenza con l'italiano. Al contrario, la variante conservativa del dialetto montefalcionese è usata in modo esclusivo con interlocutori che provengono dal medesimo comune di origine. Con questi interlocutori, vengono adoperate anche le varianti innovative, *kiro* e *kwillo,* che non sono attestate nel dialetto di origine (Di Salvo, 2022) e che rappresentano, quindi, un'innovazione del montefalcionese usato come lingua ereditaria. Anche la forma napoletana (*killo*) è presente soprattutto, ma non in maniera esclusiva, nelle interazioni con parlanti nati a Montefalcione.

Tabella 12 - *Distribuzione percentuale delle varianti del pronome e dell'aggettivo dimostrativo in relazione all'interlocutore*

	P	*R*
KILLO	85,29	14,71
KIRO	100,00	0,00
KWELLO	47,42	52,58
KWILLO	100,00	0,00
KWIRO	100,00	0,00

Scorporando i dati in base alle diverse conversazioni registrate, essi evidenziano che la variante conservativa del dialetto (*kwiro*) è adoperata soprattutto con i parlanti più giovani e con la sola interlocutrice di seconda generazione; la variante napoletana, al contrario, è adoperata con interlocutori anziani nati a Montefalcione. Ulteriori studi su un numero maggiore di parlanti permetteranno di indagare la concentrazione delle forme patrimoniali del dialetto montefalcionese nelle interazioni tra paesani.

Non sembra quindi sussistere una correlazione tra caratteristiche anagrafiche dell'interlocutore e selezione di una variante più o meno conservativa.

Tabella 13 - *Distribuzione delle forme enclitiche del dimostrativo nelle diverse registrazioni (valori percentuali)*

	Conversazione con...			
	Intervista guidata	*Armando e Giuseppe*	*Tonino e Carmela*	*Raffaella e Maria*
KILLO	14,71	52,94	8,82	23,53
KIRO	0,00	75,00	0,00	25,00
KWELLO	52,58	31,96	5,15	10,31
KWILLO	0,00	25,00	50,00	25,00
KWIRO	0,00	0,00	75,00	25,00

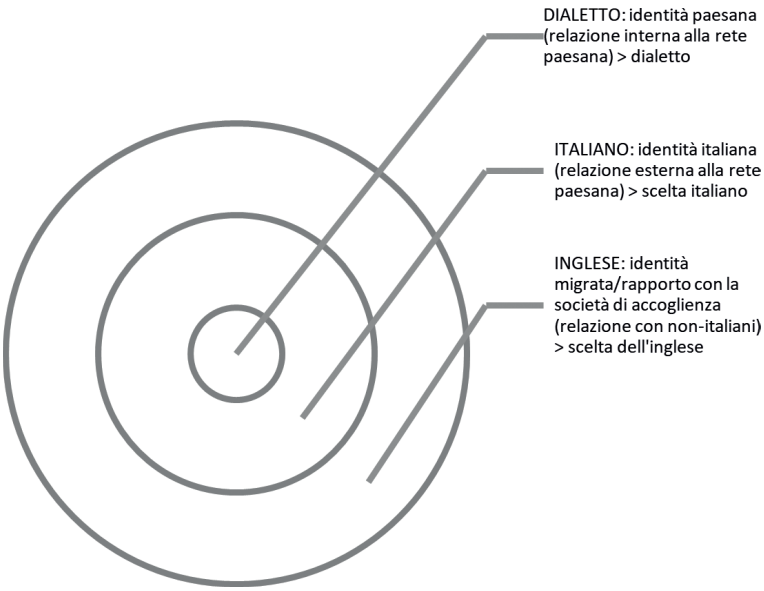
7. *Conclusioni*

I risultati dell'analisi hanno evidenziato in primo luogo che la scelta tra italiano e dialetto è condizionata dal tipo di rapporto con l'interlocutore: i parlanti bilingui italiano-dialetto si rivolgono in italiano a un interlocutore esterno alla comunità, come può essere il ricercatore sul campo, ma preferiscono il dialetto quando si trovano ad interagire con compaesani residenti nel contesto di immigrazione. In relazione ai meccanismi di selezione di codice, l'osservazione condotta sul campo conferma i dati emersi in Di Salvo (2012) in cui tale comportamento è stato individuato e discusso ma solamente a partire da dati di natura percettiva.

Le scelte dei parlanti sono condizionate dalle situazioni comunicative in cui sono impegnati come attori sociali. I migranti, inoltre, adottando una varietà o l'altra, assumono identità mutevoli che negoziano continuamente nell'interazione e che sono condizionate anche dal tipo di relazione sociale che gli interlocutori hanno e dal posizionamento interno/esterno alla comunità dei partecipanti allo scambio comunicativo.

I meccanismi della selezione di codice, inoltre, riflettono le dinamiche sociali interne alle singole comunità: nel caso specifico di Bedford, in particolare, il dialetto funge da *we-code,* mentre l'italiano è riservato alle interazioni con un parlante considerato esterno al gruppo di appartenenza. In questa chiave di lettura, i dati confermano la presenza di un modello a cerchi concentrici, che non si escludono a vicenda, come dimostrato in studi di carattere antropologico sulle identità migranti (Signorelli, 2006).

Figura 3 - *Selezione di codice con interlocutore interno/esterno alla comunità*



La relazione con l'interlocutore (interno/esterno) condiziona non solo la scelta in favore dell'italiano o del dialetto ma anche l'alternanza con l'inglese, sia sul piano strutturale che su quello funzionale: sul piano strutturale, appare particolarmente interessante l'uso delle marche pragmatiche inglesi, utilizzate dai parlanti per sottolineare la solidarietà in-group come rilevato in precedenti studi (Di Salvo 2013). Rispetto allo studio già ricordato, in particolare, la trattazione qui presentata e in particolar modo gli esempi 1-4 sembrano indicare che, accanto agli usi già studiati di *you know,* anche la marca prgmatica *yeah/yes* sembra adoperata da Giovanni per sottolineare una comune prassi comunicativa consolidata dall'uso inter-comunitario: tuttavia, per validare tale ipotesi interpretativa mi sembra necessario ampliare l'analisi ad altri parlanti.

Sul piano funzionale, la concentrazione di casi di inserzione di elementi lessicali appartenenti al campo lavorativo nell'intervista con il ricercatore esterno è sintomatica di una forte specificità che si evince soprattutto nella relazione con un membro esterno della comunità.

Il risultato più generale è che il paradigma ancorato al concetto di *heritage language* sembra essere problematico per descrivere la dilalia italiano-dialetto. Entrambi i codici hanno, nei repertori individuali e nella vita quotidiana dei parlanti, valori e funzioni comunicative diverse. Nel paradigma teorico legato alla nozione di *heritage language* al contrario è implicito un monolinguismo che, di fatto, non è supportato dalle evidenze rintracciate nelle comunità italiane nel mondo che conservano e trasmettono alla generazione successiva due diversi codici, l'italiano e il dialetto. Sicuramente la comunità italiana di Bedford presenta delle specificità dipendenti dalle condizioni storiche e sociali legate soprattutto alla concentrazione di migranti da specifiche aree italiane, dalla forte chiusura interetnica e dal forte legame con la madrepatria (Di Salvo, 2012, 2019) e sono quindi necessari ulteriori studi su altre comunità per verificare la validità generale dei risultati qui presentati. La nostra ipotesi è che, per quanto i valori comunicativi associati all'italiano e al dialetto possano essere diversi da comunità a comunità in accordo con l'approccio fortemente legato al singolo contesto di elocuzione qui sostenuto e supportato da recenti formulazioni teoriche (Aalberse, Backus & Muysken, 2019), appare chiaro che la compresenza di due lingue migrate vada considerata come un elemento costitutivo delle comunità italiane nel mondo.

Sul piano metodologico, il caso in esame dimostra che situazioni diverse per ciò che concerne la relazione tra i partecipanti spingono i parlanti a comportamenti diversi: ciò suggerisce la necessità di costruire corpora che possano essere rappresentativi di tale complessità per comprendere quanto possa essere ampio lo spettro di variazione da un lato e quali siano i valori, funzionali, emotivi, sociali, legati alla selezione di codice, di alcune varianti specifiche e della lingua dominante della società di accoglienza.

Riferimenti bibliografici

Aalberse, S., Backus, A., Muysken, P. (2019). *Heritage Languages: A language contact approach*. Amsterdam: Benjamins.

Andriani, L., Casalicchio, J., Ciconte, F., D'Alessandro, R., Frasson, A., van Osch, B., Sorgini, L., Terenghi, S. (2022). Documenting Italo-Romance heritage languages in the Americas. In Coler, M., Nevins, A. (a cura di), *Contemporary research in minority and disapora languages of Europe.* Language Science Press.

Alfonzetti, G. (1992). *Il discorso bilingue. Italiano e dialetto a Catania*. Milano: Franco Angeli.

Backus, A. (1999). The intergenerational codeswitching continuum in an immigrant community. In Extra, G., Verhoeven, L. (a cura di), *Bilingualism and migration*. Berlin: Mouton de Gruyter, 261-279.

Backus, A. (2000). Insertional code-switching in an immigrant language: 'just' borrowing or lexical reorientation?. In *Bilingualism, Language and Cognition* 3, 103-105.

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

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

Bettoni, C., Rubino, A. (1996). *Emigrazione e comportamento linguistico*. Galatina: Congedo.

Bonfatti Sabbioni, M.T. (2018). *Italian as Heritage Language Spoken in the US*. PhD Dissertation, University of Wisconsin: https://dc.uwm.edu/etd/1757

Cerruti, M., Goria, E. (2021), Varietà italoromanze in contesto migratorio: il piemontese d'Argentina a contatto con lo spagnolo. In Favilla, E., Machetti, S. (a cura di), *Lingue in contatto e linguistica applicata: individui e società.* Milano: Officinaventuno, 125-140.

Colpi, T. (1991). *The Italian Factor*. Londra: Mainstream Publishing.

D'Agostino, M. (2015). *Sociolinguistica dell'Italia contemporanea*. Bologna: Il Mulino.

De Blasi, N. (2009). *Geografia e storia dell'italiano regionale.* Bologna: Il Mulino.

De Fina, A. (2007). La lingua non fa il monaco. Funzioni simboliche dell'alternanza linguistica in comunità di origine italiana all'estero. In *Studi Italiani di Linguistica Teorica e Applicata* XXXVI (3), 401-419.

De Fina, A. (2012). Family interaction and engagement with the heritage language: A case study. In *Multilingua,* 31(4), 349-379.

De Fina, A. (2015). Language ideologies and practices in a transnational community. In Marquez, R., Martin Rojo, L. (a cura di), *A sociolinguistics of diaspora*. New York: Routledge, 48–65.

Del Vecchio, V. (2023). Code-mixing and intergenerational variation within an Italian community in Bletchley (UK). In Goria E., Di Salvo M. (a cura di), *Italian Journal of Linguistics*, *numero monografico*, 35(1): 71-90.

De Mauro, T. (1963). *Storia linguistica dell'Italia Unita*. Bari: Laterza.

Di Salvo, M. (2011). Tra mantenimento e perdita: dinamiche linguistiche e culturali in tre comunità italiane in contesto inglese. In *Bollettino Linguistico Campano* 19/20, 31-53.

Di Salvo, M. (2012). *'Le mani parlavano inglese'. Percorsi linguistici e culturali tra gli italiani d'Inghilterra.* Roma: Il Calamo.

Di Salvo, M. (2013). Segnali discorsivi inglesi tra gli italiani di Bedford e Cambridge. *Rassegna Italiana di Linguistica Applicata* 2-3. 65-84.

Di Salvo, M. (2019). *Repertori linguistici degli italiani all'estero*. Pisa: Pacini.

Di Salvo, M. (2022). Contatto interdialettale e cambiamento linguistico in un dialetto italiano all'estero. In Romito, L. (a cura di), *La variazione linguistica in condizioni di contatto: contesti acquisizionali, lingue, dialetti e minoranze in Italia e nel mondo*. Milano: Officinaventuno, 44-57.

Di Salvo, M. (in stampa). *Il contributo della sociolinguistica per lo studio linguistico dei movimenti migratori*, in Puolato, D. (a cura di), "Napoli città ibrida".

Di Salvo, M., Guzzo, S., (2021). Italian Return Migration: Discourse, Phonology and Recontextualised Identities. In Fiorentino G. Fruttaldo A. (a cura di), *Languaging the Cityscapes: Changing Linguistic Landscapes in Public Discourses*. Firenze: Franco Cesati, 89-110.

Frasson, A., D'Alessandro, R., Van Osch, B., (2021). Subject Clitics in Microcontac: A case Study from Heritage Friulan in Argentina and Brazil. In *Heritage Language Journal* 18, 1-36.

Goffman, E. (1997). La *vita umana come rappresentazione*. Bologna: Il Mulino.

Goria, E., (2015). *Il piemontese di Argentina: considerazioni generali e analisi di un caso.* In *Rivista Italiana di Dialettologia. Lingue dialetti società* 39, 127-158.

Goria, E., Di Salvo M. (2023). An Italo-Romance perspective on heritage languages. in *Italian Journal of linguistics,* 35/1, 45-70.

Gumperz, J.J. (1964). Linguistic and Social Interaction in Two Communities. In *American Anthropologist*, 66, 137-153.

Gumperz, J.J. (1982). *Discourse strategies*. Cambridge: Cambridge University Press.

Haller, H. (1987), Italian Speech Varieties in the United States and the Italian-American Lingua Franca. In *Italica* 64, 393-409.

Moreno, P., Di Salvo, M. (2015). Repertori e comportamento linguistico in due comunità italiane all'estero. In *Rivista italiana di dialettologia. Lingue dialetti società* 39, 105-124.

Muysken, P. (2000). *Bilingual speech: A typology of code-mixing*. Cambridge: Cambridge University Press.

Nagy, N. (2011). Lexical Change and Language Contact: Faetar in Italy and Canada. In *Journal of Sociolinguistics* 15, 366-382.

Polinsky, M. (2018). *Heritage language and their speakers*. Cambridge: Cambridge University Press.

Polinsky, M., Scontras, G., (2020). Understanding heritage languages. In *Bilingualism: Language and Cognition*, 23(1), 4-20.

Remotti, F., (1996). *Contro l'identità.* Milano: Feltrinelli.

Rothman J. (2009). Understanding the nature and outcomes of early bilingualism: Romance languages as heritage languages. In *International Journal of Bilingualism* 13,155–63.

Rubino, A. (2014). *Trilingual Talk in Sicilian-Australian Migrant Families. Playing Out Identities Through Language Alternatio*n. Houndmills: Palgrave Macmillan.

Signorelli, A. (2006). L'identità. L'isola che non c'è. In Giacomarra, M. (a cura di), *Isole*. Palermo: Acta Diurna, 35-42.

Turchetta, B. (2018). Modelli linguistici interpretativi della migrazione italiana. In Turchetta, B., Vedovelli, M. (a cura di), *Lo spazio linguistico dell'italiano globale: il caso dell'Ontario.* Pisa: Pacini, 73-104.

Valdés G. (2000), The teaching of heritage languages: An introduction for Slavic teaching professionals. In Kagan O., Rifkin B. (a cura di), *The Learning and Teaching of Slavic Languages and Cultures,* 375-403.

Vedovelli, M. (2011). *Storia linguistica dell'emigrazione italiana.* Roma: Carocci.

DOI: 10.17469/O2109AISV000005

BARBARA GILI FIVELA, SONIA D'APOLITO, ANNA CHIARA PAGLIARO

Tra economia dello sforzo e accuratezza del parlato nella disartria ipocinetica

Between economy of effort and speech accuracy in hypokinetic dysarthria

This paper aims at observing if dysarthric speakers affected by Parkinson's disease maintain phonological distinctions (/s/ vs. /t/) and sociophonetic features ([t^h]), since they both require a precise control of fine gestures though have a different impact on communication.Acoustic data collected on speech corpora representing different speech styles were analysed with regard to the duration of consonants and to the COG in fricatives as well as in the VOT interval identified in plosives. Results show that pathological speakers distinguish plosives from fricatives even though with an overall shorter duration in comparison with controls. Moreover, data suggest that aspiration as a sociolinguistic marker may be not preserved through compensatory strategies as much as the difference between phonologically relevant segments.

Keywords: dysarthria in Parkinson's disease, plosives, fricatives, aspiration, sociophonetic markers.

Introduzione

L'ipocinesia e la bradicinesia caratterizzano la disartria ipocinetica, che spesso colpisce i soggetti affetti dalla malattia di Parkinson (Ackermann, Ziegler, 1991; Duffy, 2005). Le conseguenze principali relativamente alla produzione di vocali e consonanti consistono nella distorsione delle vocali e nell'imprecisa articolazione delle consonanti, benché le distinzioni fonologiche vengano preservate il più a lungo possibile (Duffy, 2005), grazie a strategie di compensazione. Benché l'accuratezza del parlato sia, quindi, limitata, bisogna considerare che segmenti differenti richiedono sforzi articolatori diversi.

La produzione delle occlusive, ad esempio, è articolatoriamente più semplice di quella delle fricative, come mostrato dalla letteratura sull'argomento, in quanto realizzare una collisione richiede meno precisione che produrre e mantenere una posizione che implichi una stretta diaframmatica e la realizzazione di una costrizione (Fuchs, Perrie, Geng & Mooshammer, 2006). Da questo punto di vista, il controllo dell'ampiezza dei gesti, ad esempio, è di cruciale importanza. Se poi si valutano le caratteristiche delle consonanti occlusive aspirate, si osservano casi nei quali le consonanti sono caratterizzate da una coordinazione temporale dei gesti sopralaringei e glottidali che differisce da quella tipica delle altre occlusive (Best, 1995; Browman, Goldstein, 1986; Kent, Weismer, Kent & Rosebek, 1989; Forrest, Weismer, Turner, 1989). Oltre al controllo fine delle variazioni di ampiezza, per l'aspirazione è chia-

ramente necessario, quindi, un accurato controllo dell'organizzazione temporale dei gesti articolatori.

Al di là degli aspetti prettamente fonetici, per la produzione dei segmenti consonantici appena menzionati può essere rilevante anche lo status fonologico o allofonico del segmento, dato il diverso impatto che l'accuratezza nella produzione di fonemi e allofoni può avere dal punto di vita comunicativo per via del ruolo distintivo dei fonemi e non degli allofoni. In italiano, le occlusive e le fricative sono fonologicamente rilevanti, mentre le occlusive aspirate rappresentano varianti allofoniche solo in alcune varietà e dialetti, come nei dialetti parlati nell'area di Cosenza, nel dialetto e nell'italiano regionale dell'area di Lecce e in alcuni dialetti toscani. In particolare, a Lecce, dove sono stati raccolti i dati analizzati in questo studio, l'aspirazione si osserva maggiormente all'interno di parola e soprattutto in posizione post-tonica (Rohlfs, 1966; Sobrero, Romanello, 1981).

L'accuratezza nella produzione di fricative e occlusive, nonché di occlusive aspirate e non, quindi, è garantita da diversi vincoli di tipo fonetico e veicola informazioni differenti anche dal punto di vista linguistico.

Si consideri, inoltre, che l'accuratezza nella produzione di un qualsiasi segmento, consonantico o vocalico, varia in base al contesto comunicativo. È noto che i parlanti modificano le caratteristiche dell'eloquio nella direzione dell'iper- o dell'ipoarticolazione a seconda delle esigenze comunicative (H&H Theory di Lindblom, 1990). Ad esempio, l'accuratezza può cambiare in contesti nei quali il parlante stimi che i destinatari del messaggio abbiano difficoltà nel sentire e comprendere il parlato prodotto, o in contesti nei quali sia richiesto un maggior sforzo per la pianificazione della produzione. Compiti sperimentali diversi, quindi, che implichino la produzione di parlato letto o semispontaneo (Picheny, Durlach & Braida, 1986), rappresentano condizioni potenzialmente funzionali alla modificazione dell'accuratezza nel parlato.

La costante modulazione delle caratteristiche di produzione tipica del parlato normofasico rappresenta una sfida per il parlante patologico, in particolare per il paziente disartrico, che deve superare difficoltà nella gestione della sincronizzazione e dell'ampiezza dei suoi gesti articolatori. L'analisi del parlato di soggetti disartrici affetti da Parkinson ha mostrato, infatti, che le difficoltà possono riguardare la realizzazione della chiusura diaframmatica nelle occlusive (Antolik, Fougeron, 2013 per il francese), ma anche il mantenimento della stretta nel caso delle fricative (Logemann, Fisher, 1981 per l'inglese). Per quanto riguarda l'aspirazione, inoltre, alcune indagini hanno messo in evidenza la presenza di occlusive e di intervalli di *Voice Onset Time* (VOT) di durata maggiore nei soggetti affetti da disartria rispetto ai soggetti di controllo (cfr. discussione in Kent et al., 1989). Non ci risulta che le indagini abbiano però valutato gli aspetti suddetti in stili di eloquio diversi e nel caso di una differente produttività funzionale dei segmenti analizzati. Questo studio rappresenta un primo passo volto a migliorare la nostra comprensione delle caratteristiche di produzione di parlato da parte di soggetti disartrici, analizzando la realiz-

zazione di segmenti fonologicamente (occlusive vs. fricative) o socio-foneticamente rilevanti (occlusive aspirate) in contesti comunicativi differenti.

1. *Obiettivi e ipotesi*

Il primo obiettivo di questo studio è osservare se i parlanti disartrici siano in grado di distinguere accuratamente le occlusive e le fricative alveolari o se, nonostante il carico funzionale dell'opposizione, siano poco accurati nel produrre occlusive e fricative, mostrando più difficoltà dei soggetti di controllo nel mantenere la stretta necessaria alla produzione delle fricative. Inoltre, nella varietà di italiano qui considerata le occlusive possono essere realizzate con aspirazione, ma, come accennato in §1, dal punto di vista motorio le occlusive aspirate richiedono una precisa coordinazione tra i gesti laringali e sopralaringali, poiché è proprio la differenza in termini di *timing* che permette di distinguere le occlusive non aspirate da quelle aspirate.

Il secondo obiettivo riguarda il confronto diretto tra parlanti affetti da disartria e soggetti di controllo nella realizzazione di segmenti utili a fornire informazioni fonologicamente o socio-foneticamente rilevanti. In particolare, si intende osservare se i soggetti disartrici mostrino maggiori difficoltà dei soggetti di controllo nel mantenere le opposizioni fonologiche e il tratto sociolinguistico ([tʰ]) o se, ad esempio, lo sacrifichino con maggior facilità rispetto ai soggetti di controllo, dato che non è utile a veicolare significati dal punto di vista strettamente linguistico.

Sulla base della letteratura scientifica sull'argomento (cfr. §1), si ipotizza che i parlanti disartrici cerchino di mantenere il più possibile la distinzione tra /s/ e /t/, ma che, nonostante le difficoltà nella realizzazione della chiusura nelle occlusive, il controllo motorio relativo alla distanza tra lingua e parete fissa, utile a creare turbolenza necessaria per la produzione di fricative, possa essere più difficile rispetto alla realizzazione della collisione funzionale alla produzione delle occlusive. Per quanto riguarda le occlusive aspirate, ci si aspetta una maggiore durata della fase di occlusione e del VOT nelle produzioni dei parlanti affetti da disartria rispetto a quelle dei soggetti di controllo, dato che la durata dell'aspirazione o frizione può essere correlata allo stato dell'aspirazione stessa, così come le caratteristiche di concentrazione del rumore nella fase di aspirazione o frizione possono fornire indicazioni circa il luogo di articolazione (§3); in questo senso, oltre alla maggior durata, ci si potrebbe aspettare un luogo leggermente più arretrato nei disartrici, per via di un possibile *target undershoot* dovuto all'attesa ipoarticolazione. Circa l'opposizione fonologica (/s/ vs. /t/) e la realizzazione degli allofoni ([t], [tʰ]), ci si aspetta una maggiore stabilità nella realizzazione dell'opposizione fonologica e, di conseguenza, delle differenze maggiori per l'opposizione sociolinguistica tra parlanti disartrici e di controllo[1]. Relativamente all'aspirazione, in base alla letteratura ci si aspetta inoltre una maggior durata in sillaba post-tonica.

[1] Come osservato da un revisore, l'analisi sincronica effettuata in questo studio potrebbe non garantire che un soggetto disartrico abbia modificato una marca presente nel suo idioletto o che il suo idioletto

Come anticipato, la distinzione tra occlusive, con e senza aspirazione, e fricative, viene analizzata proponendo due diversi contesti comunicativi, che nel nostro caso si riferiscono a due diversi disegni sperimentali, nei quali varia lo stile d'eloquio. In particolare, un compito sperimentale prevede la produzione di parlato letto controllato, mentre l'altro richiede la produzione di parlato semispontaneo, elicitato grazie a Map Task e a compiti di descrizione di immagini. Si tratta di materiali differenti, che prendiamo in esame per acquisire diversi punti di osservazione sugli effetti della disartria, ma che, in questa fase, non analizziamo per effettuare un confronto diretto degli effetti dello stile di eloquio sui fenomeni indagati. Sulla base dei materiali descritti in questo articolo, infatti, non siamo in grado di effettuare un'analisi comparativa diretta tra i dati ricavati nei due compiti sperimentali, parlato controllato e semispontaneo, poiché sono stati raccolti grazie a parlanti diversi (anche in termini di genere). Peraltro, l'analisi di parlato semispontaneo rende impossibile garantire un numero specifico di osservazioni per ogni segmento studiato. I risultati del secondo esperimento forniranno quindi solo osservazioni aggiuntive rispetto a quanto sarà discusso in relazione al primo studio. Tuttavia, riteniamo che l'accuratezza nella produzione dei soggetti disartrici possa diminuire in modo evidente nel parlato semispontaneo, per via della richiesta di maggiori risorse cognitive per lo svolgimento del compito sperimentale. Pur non potendo effettivamente verificare alcuna ipotesi relativa all'impatto dovuto allo stile d'eloquio, quindi, intendiamo comunque descrivere il comportamento dei parlanti disartrici nel parlato semispontaneo, pensando che questo possa chiarire la direzione di un andamento magari solo intuibile nel parlato letto.

2. *Metodo*

In questo studio abbiamo realizzato due diversi compiti sperimentali per la raccolta di parlato controllato e semispontaneo.

Nel primo, la fricativa /s/ e l'occlusiva /t/ sono state elicitate in posizione iniziale e in sillaba tonica e in posizione mediana in sillaba post-tonica, all'interno della stessa parola. In tutti i casi, i segmenti bersaglio sono stati inseriti in pseudo-parole e nel contesto vocalico /a/-/a/ ed elicitate come scempie (*sasa*, *tata*) all'interno di una frase cornice (es. *La sasa blu*). Hanno preso parte all'esperimento 5 parlanti di sesso maschile, affetti da disartria e provenienti da Lecce o zone limitrofe (età media 71.6). I parlanti disartrici sono stati inclusi nello studio in quanto affetti da

abbia mai incluso tale marca. Tuttavia, l'analisi qui presentata è parte di un progetto più ampio nel quale la propensione verso il polo dialettale, e quindi l'incidenza di marche dialettali, è stimata sulla base di un profilo sociofonetico individuato tramite apposito questionario e di analisi relative anche a tratti inerenti al vocalismo e alle caratteristiche intonative. Ne consegue che lo studio qui presentato è volutamente ristretto all'aspirazione delle occlusive, ma gli altri dati in nostro possesso garantiscono la presenza di marche complessive di natura sociofonetica. Non potendo attualmente effettuare un'analisi longitudinale, si assume, quindi, che l'analisi dei dati qui presentati fornisca parte dell'informazione relativa alle suddette marche.

Parkinson e disartria ipocinetica (*Activities of Daily Living*: tra 4 e 6, *Instrumental Activities of Daily Living*: tra 4 e 8; *Robertson Dysarthria profile* punteggio medio per quesito: 1-2). In base alle caratteristiche dei soggetti affetti da disartria, sono stati individuati 4 soggetti di controllo, non affetti da patologie note, di sesso maschile e provenienti da Lecce o zone limitrofe (età media 69.75). Sono stati registrati simultaneamente dati acustici e articolatori grazie ad Articulografia ElettroMagnetica – EMA (AG501) e per i soggetti disartrici la registrazione è avvenuta in fase ON. Il corpus di stimoli è stato letto da un minimo di 5 a un massimo di 7 volte e in questa sede saranno riportati solo dati acustici (cfr. Gili Fivela, d'Apolito & Pagliaro, 2023, per un'analisi di dati articolatori).

Nel secondo esperimento, il parlato semispontaneo è stato elicitato grazie a Map Task e a compiti di descrizioni di immagini realizzati in forma dialogica tra un soggetto sperimentale e uno sperimentatore; all'interno di questo materiale i segmenti (/s/, /t/) sono stati osservati in posizione iniziale di parole reali (*sana*, *tana*), dalla struttura simile a quella delle pseudoparole considerate nell'esperimento precedente, ma estrapolate da frasi semplici o complesse. Le registrazioni acustiche hanno coinvolto due soggetti affetti da disartria (età media 65.6; *Nijmegen Dysartrhia Scale Therapy Outcomes Measures* – DisTOM 3-4) e due di controllo (età media 63.8). Tutti i parlanti sono di sesso femminile e provengono da Lecce o zone limitrofe.

I dati acustici sono stati analizzati in PRAAT (Boersma, Weenink, 2022) individuando i confini di tutti i segmenti della frase cornice per il primo task e della sola parola *target* per il secondo task. Per ciascun segmento è stata calcolata la durata complessiva, utilizzata anche per il calcolo della velocità articolatoria (numero di sillabe/durata totale della frase), e per le occlusive è stata calcolata anche la durata del VOT, individuato come l'intervallo tra il rilascio dell'occlusiva e l'inizio della vocale successiva. Le durate sono state considerate come assolute e normalizzate in base alla durata della parola. Inoltre, è stato calcolato anche il Centro di Gravità (COG) per osservare se le frequenze alle quali si osserva la concentrazione di energia possa differire tra i parlanti affetti da disartria e i soggetti dicontrollo, indicando differenze nel luogo di articolazione tra disartrici e controlli nella realizzazione delle fricative o nel rilascio delle occlusive. Pur non essendoci un solo parametro acustico correlato al luogo di articolazione di fricative e affricate (Jongman, Wayland & Wong, 2000), infatti, il COG è stato individuato come uno dei cinque (di nove) parametri utili a distinguere i tre luoghi di articolazione delle affricate del mandarino (Li, Gu, 2015, insieme ad ampiezza normalizzata, picco, dispersione e skewness). Nel nostro studio, il COG è stato quindi considerato in un primo tentativo di individuare se esistano differenze nel luogo di articolazione tra soggetti disartrici e di controllo (a parità di contesto intervocalico); la durata è stata invece considerata come indicativa dell'entità dell'aspirazione (lo stato di aspirazione in Li, Gu 2015, benché gli autori mettano in relazione la diminuzione di COG anche con la presenza di aspirazione).

I risultati statistici sono stati ottenuti applicando modelli misti grazie al software R (R Core Team 2015) e il pacchetto *lme4* (Bates, Machler, Bolker & Walker, 2015).

Per quanto il numero di dati sia esiguo, abbiamo infatti preferito adottare lo stesso metodo già usato per l'analisi di altre porzioni dello stesso corpus, confidando nei risultati statistici solo al fine di descrivere differenze degne di nota e non per individuare risultati necessariamente generalizzabili su larga scala. I valori di significatività sono stati ottenuti attraverso i test del chi-quadro implementato nella funzione *Anova*. Nelle analisi relative al primo compito sperimentale (§3.1), i fattori fissi considerati sono stati:1) gruppo (soggetti disartrici vs soggetti di controllo); 2) costrizione (fricativa vs occlusiva); 3) posizione (iniziale vs mediana) e 4) numero delle ripetizioni. Al fine di tener conto della variabilità intra-parlante, i parlanti sono stati considerati come *random slope*. Per il secondo *task* (§3.2), si forniscono informazioni quantitative ma non si fa riferimento a modelli misti, dato il numero particolarmente limitato di osservazioni, mentre per il controllo effettuato (§3.3) in merito alla durata del VOT (solo occlusive) e al COG separatamente per fricative e occlusive è stato effettuato un t-test a campioni indipendenti dal momento che l'unico fattore indagato è stato il gruppo (infatti, non è stato considerato il fattore costrizione).

3. *Risultati*

3.1 Esperimento I – parlato letto controllato

Come riportato in letteratura e menzionato in §1, i parlanti con disartria possono realizzare le consonanti in maniera imprecisa, ad esempio con una chiusura incompleta nel caso delle occlusive. La tabella 1 riassume i casi in cui i parlanti, sia disartrici che di controllo, realizzano o meno la chiusura completa nell'occlusiva /t/ in sillaba tonica e in post-tonica. Come si può osservare, l'occlusiva è prodotta in modo accurato e presenta uno scoppio individuabile in tutte le produzioni dei controlli; al contrario, nei parlanti affetti da disartria lo scoppio è stato realizzato nel 57,57% e nel 54,54% dei casi in sillaba tonica e post-tonica rispettivamente. In particolare, i parlanti PD2 e PD5 mostrano le maggiori difficoltà.

I risultati statistici relativi alle misure acustiche suggeriscono che la durata della consonante, sia normalizzata che assoluta, varia in base alla posizione (assoluta: $\chi2(1)=27,44$ p=,000; normalizzata: $\chi2(1)=32,19$ p=,000) ed è maggiore per la consonante in sillaba tonica (assoluta: 10,24ms ±1,89 S.E.; normalizzata: 2,34% ±0,39 S.E.). Inoltre, la durata assoluta differisce in modo significativo in base al gruppo ($\chi2(1)=4,81$ p=,02) e alla costrizione ($\chi2(1)=14,34$ p=,000). La durata, infatti, risulta essere più breve nei parlanti affetti da disartria rispetto ai controlli (-2,89ms ±1,14 S.E.) e nelle occlusive rispetto alle fricative (-7,36ms ±1,91 S.E.) – Fig. 1 e 2. Le interazioni non risultano significative.

Tabella 1 - *Numero di burst realizzati per gruppo*

Parlante	*Posizione*	*Occlusive*		
		N	*Burst realizzato*	*%*
CTR	Tonica	27	27	100
	Post-tonica	27	27	100
PD	Tonica	33	19	57,57
	Post-tonica	33	18	54,54

Figura 1 - *Grafico a scatole per la durata non normalizzata della consonante rispetto alla costrizione (CTR = bianco; PD = grigio; a sinistra le fricative e a destra le occlusive)*

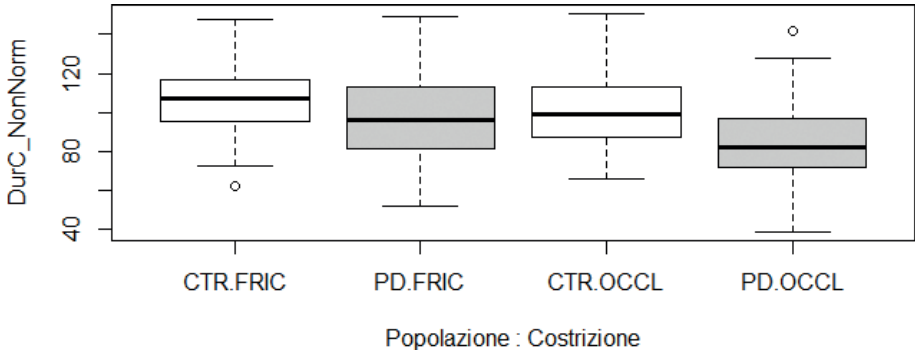
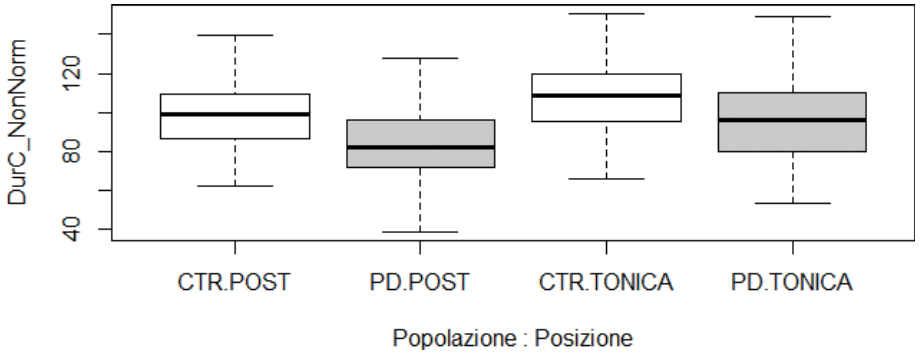
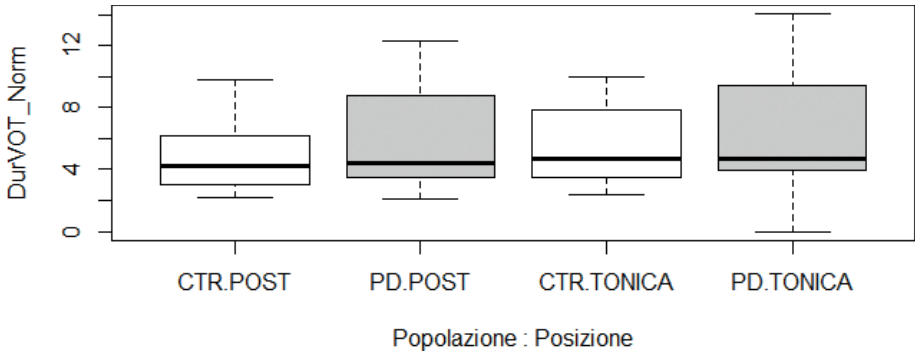


Figura 2 - *Grafico a scatole per la durata non normalizzata della consonante rispetto alla posizione (CTR = bianco; PD = grigio; a sinistra la sillaba post-tonica e a destra la sillaba tonica)*



La durata del VOT differisce in modo significativo solo in base alla posizione (assoluta: $\chi 2(1)=6{,}65$ p=,009; normalizzata: $\chi 2(1)=7{,}62$ p=,005), ed è più lunga per la sillaba tonica iniziale (assoluta: 3,83ms ±1,45 S.E.; normalizzata: 1,09%- ±0,38 S.E.) – Fig. 3. Anche per la durata del VOT non si riscontrano interazioni significative.

Figura 3 - *Grafico a scatole per la durata normalizzata del VOT rispetto alla posizione (CTR = bianco; PD = grigio; a sinistra la sillaba post-tonica e a destra la sillabatonica)*



Al fine di gettare luce sulle misure assolute di durata, è stata considerata anche la velocità di articolazione (numero di sillabe/durata totale della frase) - Tab. 2). Le analisi statistiche non mostrano alcuna differenza significativa né rispetto alla popolazione né rispetto alla costrizione. In generale, la velocità di articolazione è leggermente maggiore nei parlanti disartrici (4,54 d.s. 0,78) rispetto ai controlli (4,49, d.s. 0,65) ed è caratterizzata anche da deviazione standard più elevata. Osservando meglio i valori medi, si riscontra che la velocità di articolazione nei parlanti disartrici risulta leggermente superiore nelle frasi con occlusiva, mentre è leggermente inferiore nelle frasi con fricativa.

Tabella 2 - *Media e deviazione standard per la velocità di articolazione (sill/sec)*

Parlante	*Costrizione*	*Media*	*Dev. st.*
CTR	Occlusive	4,45	0,67
	Fricative	4,53	0,65
PD	Occlusive	4,66	0,77
	Fricative	4,40	0,78

Il centro di gravità (COG) è stato calcolato per le fricative e per la fase di rilascio delle occlusive, poiché è una misurazione relativa alla concentrazione di energia che potrebbe dare indicazioni sul luogo di articolazione del segmento consonantico. Le analisi statistiche sono state effettuate separatamente per le fricative e la fase di rilascio delle occlusive.

Per quanto riguarda le occlusive, il COG risulta significativamente maggiore in posizione di sillaba tonica ($\chi2(1)=42,67$ p=,000; 932,07Hz ± 124,53 S.E.) – Fig. 4 – e varia significativamente anche in base alla ripetizione ($\chi2(6)=13,76$ p=,03). Anche per le fricative il COG varia significativamente in base alla posizione ($\chi2(1)=60,92$ p=,000) e, come per il VOT, risulta essere maggiore quando /s/ compare in posizione iniziale di sillaba tonica (776,12Hz ± 84,34 S.E.) – Fig. 5.

Figura 4 - *Grafico a scatole per il COG per la fase di rilascio dell'occlusiva rispetto alla posizione (CTR = bianco; PD = grigio; a sinistra la sillaba post-tonica e a destra la sillaba tonica)*

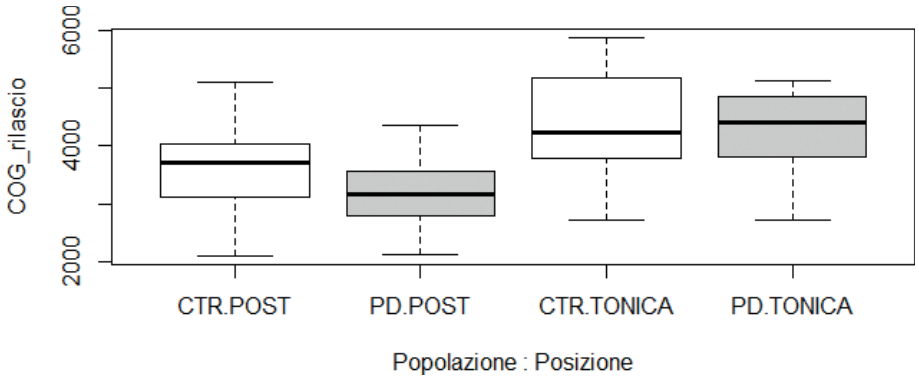
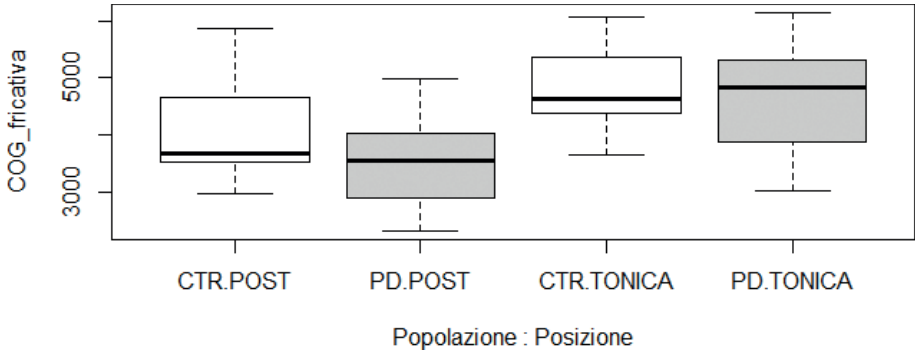


Figura 5 - *Grafico a scatole per il COG delle fricative rispetto alla posizione (CTR = bianco; PD = grigio; a sinistra la sillaba post-tonica e a destra la sillaba tonica)*



Quindi, nel parlato letto il tipico scoppio delle occlusive manca in circa la metà delle produzioni dei soggetti affetti da disartria. I risultati delle misurazioni acustiche mostrano che le durate, delle consonanti e del VOT, e il COG hanno valori maggiori nel caso in cui la sillaba sia tonica. Inoltre, le occlusive hanno una durata più breve rispetto alle fricative e, in generale, le durate consonantiche in valori assoluti hanno una durata minore per i parlanti disartrici rispetto ai controlli.

3.2 Esperimento II – parlato semispontaneo

Ricordiamo che per il parlato semispontaneo sono state analizzate parole reali in cui /s/ o /t/ compaiono solo in sillaba tonica in posizione iniziale di parola, all'interno di dialoghi Map Task o in descrizioni di immagini. Anche in questo caso per le occlusive si è cercato di individuare il rilascio e, come si può osservare dalla tabella 3, i parlanti affetti da disartria non mostrano difficoltà nella sua realizzazione.

Tabella 3 - *Numero di burst realizzati per gruppo*

Parlante	*Posizione*	*Occlusive*		
		N	*Burst realizzato*	*%*
CTR	Tonica	6	6	100
PD	Tonica	8	7	87,5

In linea con i risultati del primo esperimento, le durate sono in generale più brevi per i parlanti affetti da disartria (assoluta: -46,41ms ±8,53; normalizzata: -7,92%) e per le occlusive rispetto alle fricative (assoluta: -44,59ms; normalizzata: -7,42%) – Fig. 6. Per quanto il VOT, la durata assoluta della fase di rilascio delle occlusive è tendenzialmente più breve nei parlanti affetti da disartria (-18,71ms) che nei soggetti di controllo (media 25,6ms) – Fig. 7.

Figura 6 - *Grafico a scatole per la durata normalizzata della consonante rispetto alla costrizione (CTR = bianco; PD = grigio; a sinistra la fricativa e a destra l'occlusiva)*

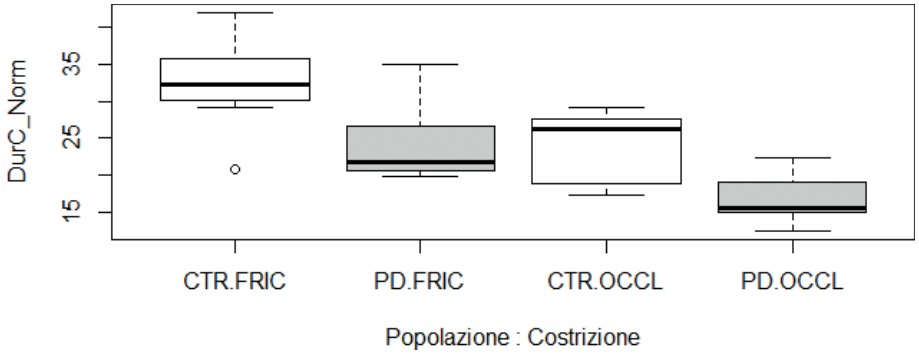
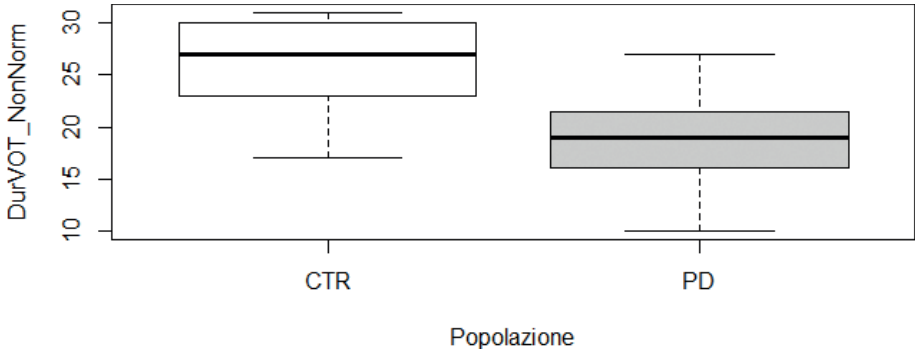


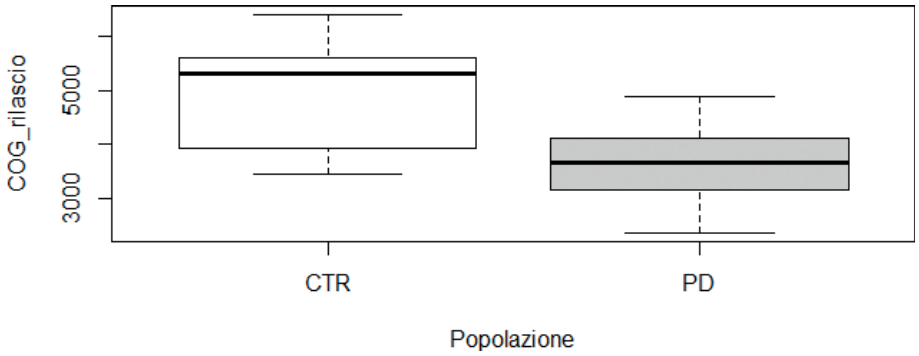
Figura 7 - *Grafico a scatole per la durata non normalizzatadel VOT rispetto al gruppo (CTR = bianco; PD = grigio)*



Infine, il COG nella fase di rilascio delle occlusive presenta valori minori nelle produzioni dei parlanti affetti da disartria (3645,80Hz) rispetto a quelle del gruppo di

controllo (4995,10Hz) – Fig. 8. La stessa tendenza si riscontra per le fricative (media PD: 6508.17Hz; CTR: 7327,80Hz).

Figura 8 - *Grafico a scatole per il COG nel VOT rispetto al gruppo (CTR = bianco; PD = grigio)*



I risultati del secondo task mostrano, quindi, che i parlanti affetti da disartria si comportano come i controlli per quanto riguarda la realizzazione del *burst* delle occlusive, ma producono segmenti consonantici più brevi sia nel caso di occlusive che nel caso di fricative. I soggetti disartrici differiscono anche per quanto riguarda il VOT, che risulta più breve di quello prodotto dai controlli e con COG minore, ad indicare un luogo di articolazione più arretrato che nel caso dei parlanti disartrici (benché Li, Gu 2015, per il Mandarino, mettano in relazione la diminuzione di COG anche con la presenza di aspirazione). La stessa tendenza ad una diminuzione del COG nelle produzioni dei soggetti affetti da disartriasi osserva nell'analisi delle fricative. Questi risultati differiscono da quelli ottenuti per il corpus raccolto nel primo esperimento, nel quale, tuttavia, le consonanti in esame comparivano sia in sillaba tonica iniziale che in atona post-tonica. Per comprendere a pieno se i risultati relativi al COG possano effettivamente differire per via dello stile di eloquio del parlato analizzato, è stata quindi effettuata una verifica sulle misure di COG ricavate nel primo corpus, effettuando l'analisi relativa al parlato letto separatamente in base alla posizione della consonante.

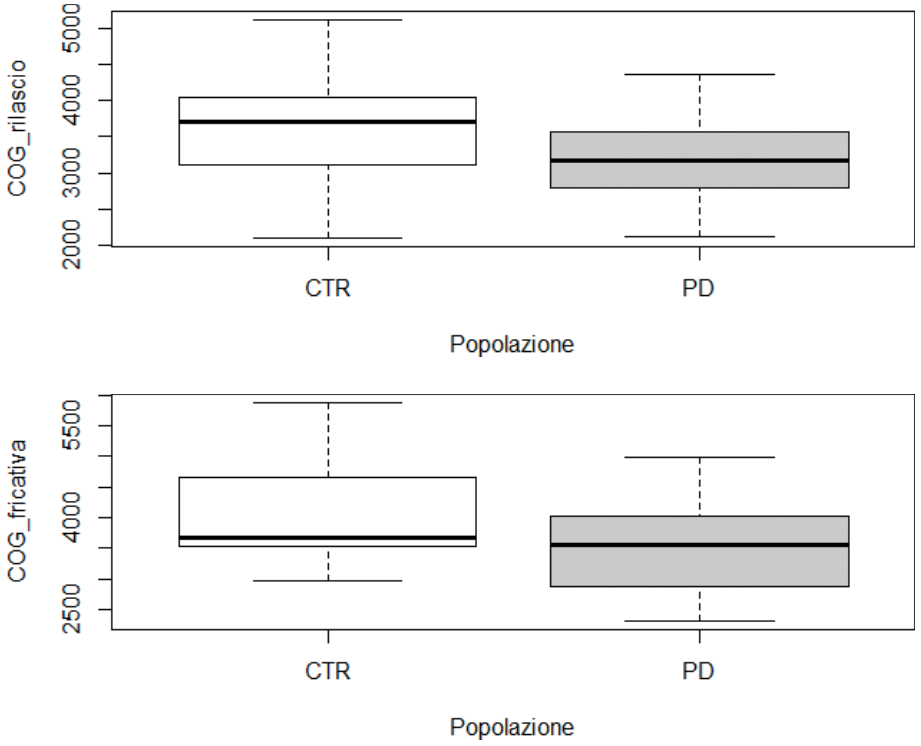
3.3 Verifica sulle occlusive e fricative prodotte nel parlato letto

I risultati del t-test effettuato separando i dati in base alla posizione indicano che per le occlusive in posizione tonica iniziale il gruppo non differisce in modo significativo (p=0.7), con valori di COG medi di 4283,29Hz (d.s. 692,89Hz) nei PD e di 4344,83Hz (d.s 824,25Hz) nei CTR. Al contrario, il t-test mostra una differenza significativa nei dati relativi alla posizione post-tonica (t(1)=2.42 p=,02) con i PD ai quali corrispondono valori più bassi (3133,45Hz,d.s. 576,29Hz) rispetto ai quelli dei controlli (3604,14Hz,d.s. 724,24Hz) – Figura 9, in alto.

Il t-test per /s/ in sillaba tonica non è significativo (p=,30), con il COG nelle produzioni dei PD e dei CTR che corrisponde a valori medi di 4626,80Hz (d.s.

846,77Hz) e 4832,52Hz (d.s. 608,01Hz) rispettivamente. In sillaba post-tonica, il test è invece significativo (t(1)=3,08 p=,003), con valori medi minori nei soggetti disartrici (3448,76Hz, d.s. 684,29Hz) rispetto ai soggetti di controllo (4026,01Hz, d.s. 719,20Hz)– Fig. 9, in basso.

Figura 9 - *Grafico a scatole relativo al COG nel VOT (alto) e nella fricativa (basso) in sillaba post-tonica (CTR = bianco; PD = grigio)*



Il controllo effettuato sui dati di parlato letto conferma, quindi, che i valori di COG in posizione post-tonica sono minori per i soggetti disartrici che per i soggetti di controllo e in modo significativo sia per le occlusive che per le fricative.

4. *Discussione e conclusioni*

I due esperimenti descritti in questo articolo sono stati effettuati per studiare le modificazioni nell'accuratezza del parlato disartrico nel caso della produzione di segmenti fonologicamente diversi (occlusive vs. fricative) o socio-foneticamente marcati (occlusive aspirate), effettuando l'analisi acustica di parlato prodotto in contesti comunicativi differenti, nei quali sono previsti cambiamenti nello stile d'eloquio. Si tratta di materiali presi in esame per osservare gli effetti della disartria in "contesti" diversi, peraltro anche al variare del genere del parlante, e non per effettuare un confronto diretto degli effetti dello stile di eloquio (o del genere) sui fenomeni indagati.

Nel primo esperimento, è stato elicitato parlato letto, nel quale si è osservato che i soggetti disartrici realizzano il tipico scoppio delle occlusive in circa la metà dei casi rispetto ai soggetti di controllo. I parlanti affetti da disartria, tuttavia, sembrano differenziare occlusive e fricative, benché le durate assolute delle consonanti che producono sia inferiore rispetto alla durata delle consonanti realizzate dai parlanti di controllo. Per entrambe le popolazioni, le occlusive hanno durata inferiore a quella delle fricative e, in generale, la presenza della consonante in sillaba tonica implica valori maggiori di durata, VOT e COG. La presenza dello scoppio in un minor numero di occlusive conferma la ridotta pressione e forza articolatoria nelle produzioni dei soggetti disartrici, e la minor durata assoluta delle loro consonanti sembra dovuta alla velocità di articolazione, visto che quest'ultima risulta maggiore soprattutto nelle frasi che includono occlusive per alcuni soggetti affetti da disartria (su velocità di articolazione e di eloquio, si veda anche Gili Fivela, Pagliaro, d'Apolito, Sallustio, Fiorella, in stampa). Inoltre, VOT e COG variano in modo analogo in controlli e disartrici, ma sono influenzati dalla condizione accentuale della sillaba piuttosto che da variazioni sociolinguistiche, per le quali era attesa una differenza in base alla posizione e allo status accentuale della sillaba (es. durata maggiore in sillaba post-tonica). Le misure acustiche sui dati di parlato letto, quindi, suggeriscono che i soggetti disartrici differenzino fricative e occlusive, nonostante variazioni relative alla durata e, ad esempio, al numero di *burst* realizzati nelle occlusive; tuttavia, almeno sulla base delle prime analisi effettuate sul parlato letto, i disartrici non differiscono in modo statisticamente significativo dai controlli nella realizzazione dell'aspirazione.

I risultati del secondo esperimento mostrano che i parlanti affetti da disartria si comportano in modo analogo ai controlli per quanto riguarda la realizzazione del *burst* delle occlusive, diversamente da quanto ci si potrebbe aspettare in un compito cognitivamente complesso (e, per quanto non si effettui un confronto diretto, diversamente da quanto osservato nel primo esperimento). In ogni caso, producono segmenti consonatici più brevi, sia nel caso delle occlusive che nel caso delle fricative, mantenendo la differenza tra le due classi di consonanti. Tuttavia, per quanto riguarda le occlusive, dal secondo esperimento risulta che nelle produzioni dei soggetti affetti da disartria il VOT sia più breve e il COG (misurato nel VOT) corrisponda a valori minori di quello ricavato nel VOT dei controlli (coerentemente con quanto osservato anche nelle fricative), ad indicare un luogo di articolazione più arretrato nelle produzioni dei soggetti disartrici[2].

L'analisi aggiuntiva inerente alle misure di COG ricavate nel primo corpus, effettuata considerando separatamente le misure relative alle diverse posizioni della consonante (§3.3), conferma che i parlanti disartrici possono presentare valori di COG minori rispetto ai controlli, sia per la fase di rilascio delle occlusive che per le fricative. Nel caso di quest'analisi, però, il possibile arretramento del luogo di

[2] Si noti che Li e Gu (2015) mettono in relazione la diminuzione di COG con la variazione di luogo, ma anche, secondariamente, con la presenza di aspirazione; nel nostro caso, la diminuzione riguarda anche le fricative, che non sono aspirate ma possono variare in termini di luogo di articolazione.

articolazione nei soggetti affetti da disartria riguarda la sola posizione post-tonica. Questi risultati confermano che la fase di aspirazione è più breve nei parlanti disartrici che nei parlanti di controllo, ed è realizzata in posizione leggermente arretrata nei soggetti affetti da disartria, benché allo stato attuale dell'analisi non sia chiaro il perché questo accada in sillaba tonica iniziale in un esperimento (il secondo) e nella post-tonica mediana nell'altro (il primo). Tuttavia, è interessante notare che, rispetto alle ipotesi relative al tratto sociolinguistico atteso nella varietà, i risultati del primo e del secondo esperimento suggeriscono tendenze diverse, ma comunque illuminanti. Il primo esperimento suggerisce che i parlanti disartrici si comportino come i controlli, con un incremento del VOT e del COG in sillaba tonica, mentre il secondo esperimento e l'analisi effettuata scorporando i dati relativi alla posizione del primo esperimento suggeriscono che i parlanti affetti da disartria modifichino l'articolazione della fase di aspirazione in sillaba post-tonica (con arretramento del luogo di articolazione, che suggerisce una realizzazione ipoarticolata), andando a modificare la realizzazione dell'aspirazione proprio nella posizione in cui la sua presenza dovrebbe rappresentare il tratto sociofonetico più marcato (cfr. §1).

Per quanto riguarda gli obiettivi della ricerca, possiamo quindi concludere che 1) i parlanti disartrici distinguono le occlusive e le fricative (alveolari), benché la durata dei segmenti consonantici prodotti sia complessivamente ridotta. Relativamente alle 1a) possibili difficoltà nella realizzazione delle fricative rispetto alle occlusive, osserviamo che la durata è ridotta sia nelle prime che nelle seconde e, quindi, le durate non permettono di identificare difficoltà specifiche che dipendano dalla modalità di articolazione. Peraltro, nel produrre le consonanti i disartrici sembrano realizzare l'occlusione o la stretta diaframmatica in un luogo articolatorio leggermente arretrato (cfr. COG), almeno nel secondo esperimento, mentre nel primo realizzano un minor numero di *burst*, e quindi occlusive sorde articolate con un rilascio repentino, caratterizzato da una forte dispersione di energia. Anche considerando correlati acustici più chiaramente attribuibili a caratteristiche articolatorie, quindi, non sembrano emergere chiari indici della presenza di difficoltà specifiche in relazione alle fricative piuttosto che alle occlusive.

Circa la possibilità che 1b) i disartrici siano poco precisi nella coordinazione tra gesti sopra-laringei e laringei in una varietà di italiano in cui le occlusive sono aspirate, i dati offrono spunti di riflessione interessanti. Solo i risultati del secondo esperimento suggeriscono che il VOT possa essere effettivamente più breve nei disartrici, con COG inferiore (luogo di articolazione arretrato) nelle sillabe toniche iniziali, mentre nel primo esperimento l'arretramento risulta riguardare le post-toniche atone.

Come discusso in §2, non effettuiamo un confronto diretto dell'effetto dello stile d'eloquio sui fenomeni indagati, ma possiamo osservare che, indipendentemente dai materiali considerati, l'arretramento suggerisce effettivamente una scarsa accuratezza nella produzione delle consonanti aspirate e che l'aspirazione possa essere modificata in tutte le posizioni e condizioni accentuali, verosimilmente per effettive difficoltà nella coordinazione tra gesti sopra-laringei e laringei. Tuttavia, il fatto che la verifica relativa al primo esperimento (§3.3) mostri che l'arretramento

sia significativamente differente solo nelle post-toniche atone suggerisce che la minor accuratezza riguardi proprio la posizione e condizione accentuale nella quale ci si aspetterebbe il tratto sociolinguistico più marcato. Rispetto all'obiettivo 2), quindi, i dati suggeriscono che l'aspirazione in quanto tratto sociolinguistico possa non essere preservata grazie a strategie di compensazione tanto quanto la differenza tra segmenti fonologicamente rilevanti (fricative vs occlusive). Quest'utima viene mantenuta nonostante la riduzione dei correlati (es. durata), preservando le proporzioni attese (durata delle occlusive minore di quella delle fricative) e, soprattutto nel parlato semispontaneo del secondo esperimento, alcuni tratti caratteristici (come il *burst* nelle occlusive).

I materiali discussi in questo articolo sono parte di un progetto più ampio e solo l'analisi del corpus completo, nel quale saranno gli stessi soggetti ad affrontare diverse situazioni comunicative, permetterà di chiarire tutte le questioni oggetto di riflessione in questa sede. Peraltro, anche un'indagine percettiva accurata inerente alla rilevanza uditiva dei fenomeni indagati è necessaria per fornire un quadro completo della problematica in esame. In ogni caso, l'indagine preliminare presentata in questo articolo ha permesso di sviluppare le prime riflessioni relativamente a ciò che accade alla distinzione fonologicamente rilevante fra occlusive e fricative alveodentali sorde e alla realizzazione di una marca sociofoneticamente rilevante come l'aspirazione delle occlusive in stili di eloquio diversi utilizzati da parlanti disartrici affetti da Parkinson.

Ringraziamenti

Il presente lavoro è finanziato dal Progetto PRIN 2017 – JNKCYZ. Vogliamo ringraziare tutti i parlanti che hanno partecipato a questo studio. Ringraziamo, inoltre, L'Ing. F. Sigona per il suo supporto tecnico, M. Iraci per aver raccolto parte dei materiali audio e per aver effettuato un'analisi preliminare e A. Mazzone per aver segmentato parte dei dati.

Bibliografia

Ackermann, H., Ziegler, W. (1991).Articulatory deficits in Parkinsonian dysarthria: an acoustic analysis. In *Journal of Neurology, Neurosurgery, and Psychiatry*, 54, 1093–8.

Antolík, T., Fougeron, C. (2013). Consonant distortions in dysarthria due to Parkinson's disease, Amyotrophic Lateral Sclerosis and Cerebellar Ataxia. In *Proceeding of Interspeech 2013*, August, Lyon, France, 2152-2156.

Bates, D.S., Maechler, M., Bolker & B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. In *Journal of Statistical Software*, 67, 1, 1-48.

Best, C.T. (1995). A direct realist view of cross-language speech perception. InStrange, W. (Ed.).*Speech perception and linguistic experience: issues in cross-language research*. York Press, 171–204.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. In *Glot International*,5 no. 9/10, 341-345.

Browman, C.P., Goldstein, L. (1986). Articulatory gestures as phonological units. In *Phonology*, 6, 151-206.

Duffy, J.R. (2005). *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. 2°Ed., Elsevier Mosby.

Forrest, K., Weismer, G. & Turner, G. (1989). Kinematic, acoustic, and perceptual analyses of connected speech produced by Parkinsonian and normal geriatric males. In *Journal of the Acoustical Society of America*, 85, 2608-2622.

Fuchs, S., Perrier, P., Geng, C. & Mooshammer, C. (2006). What role does the palate play in speech motorcontrol? Insights from tongue kinematics for German alveolar obstruents. In Harrington, J., Tabain, M., (Eds.). *Speech Production: Models, Phonetic Processes, and Techniques,* Psychology Press, 149-164.

Gili Fivela, B., d'Apolito, S. & Pagliaro, A.C. (2023). Phonological and sociophonetic information in dysarthric speech: A first articulatory investigation in Italian. In *Proceeding of ICPhS 2023*, August, Prague, Czech Republic, 3937-3941.

Gili Fivela, B., Pagliaro A.C., d'Apolito S., Sallustio V. & Fiorella M. (in stampa). Identità e parlato nella disartria ipocinetica. In Dovetto, F.M., (Ed.) *Tra medici e linguisti. Parole dentro, parole fuori.*

Jongman, A., Wayland, R. & Wong, S. (2000). Acoustic characteristics of English fricatives. In *The Journal of the Acoustical Society of America,* 125, 3962-3973.

Kent, R.D., Weismer, G., Kent, J.F. & Rosenbek, J.C. (1989). Toward phonetic intelligibility testing in dysarthria. *JSHD*, 54, 482-499.

Li, S., Gu, W. (2015). Acoustic Analysis of Mandarin Affricates. In *Proceeding of Interspeech 2015*, August, Dresden, Germany, *1680-1684.*

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In Chapter in *Speech Production and Speech Modelling*, Hardcastle W.J., and Marhcal A. (Eds). Netherlands: Springer), 403–439.

Logemann, J.A., Fisher, H.B. (1981). Vocal tract control in Parkinson's Disease: Phonetic feature analysis of misarticulations. In*JSHD*, 46, 348-352.

Picheny M.A., Durlach N.I. & Braida L.D. (1989). Speaking Clearly for the Hard of Hearing III. An Attempt to Determine the Contribution of Speaking Rate to Differences in Intelligibility between Clear and Conversational Speech. In *Journal of Speech, Language, and Hearing Research*, 1.

R Core Team, (2019). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria.

Rohlfs, G. (1966). *Grammatica storica della lingua italiana e dei suoi dialetti*. *Fonetica*. Einaudi, v. 1.

Sobrero, A., Romanello, M.T. (1981). *L'italiano come si parla in Salento*. Milella.

DOI: 10.17469/O2109AISV000006

MARTA MAFFIA, MASSIMO PETTORINO

Il parlato dei docenti di lingua italiana. Un confronto ritmico-prosodico tra contesto L1 e L2

Italian Teacher Talk. A rhythmical-prosodic comparison between L1 and L2 contexts

Teacher talk shares many of the linguistic features of other kinds of simplified registers (foreigner talk, baby talk, elderspeak), in which listener-oriented modifications are used with different addressees who may not be fully competent in the language. The present study intends to investigate rhythmical-prosodic features in the speech of two female teachers of Italian, both from the Campania region, aged 44 and 48. They were recorded in two different educational settings: in secondary schools, teaching to native Italian students; in L2 Italian classes for immigrants hosted by a voluntary association in Naples. Eight monologic samples were selected from the corpus and spectroacustically analysed, allowing the calculation of articulation rate, speech rate, fluency, speech time composition and tonal range. Results show that when teachers speak to non-native learners, they do not modify the rate at which they articulate phones; instead they use longer and more frequent silent pauses and a more reduced tonal range than with native students.

Keywords: teacher talk, Italian, native and non-native listeners, prosody, intonation.

1. *Introduzione*

Il parlato nel contesto della classe è stato ampiamente studiato ed è solitamente descritto come fortemente marcato da variabili diafasiche, legate al ruolo dei partecipanti all'interazione (il docente e i discenti) e alla loro relazione di asimmetria istituzionalizzata (Diadori, 2004: 72). Facendo riferimento alla classificazione proposta dal LIP (*Lessico di frequenza dell'Italiano Parlato* – De Mauro, Mancini, Vedovelli & Voghera, 1993), tale parlato rientra principalmente nelle categorie di "scambio bidirezionale faccia a faccia con presa di parola non libera da parte degli studenti" e di "comunicazione unidirezionale del docente in presenza degli studenti" (Diadori, 2007: 339). È il docente, infatti, in quanto "regista" (Orletti, 2000), "orchestratore" (Pallotti, 1998) della comunicazione didattica, a trovarsi in una posizione caratterizzata da dominanze multiple, di natura sia discorsiva sia sociale (Ciliberti, 1981; 2003; Sinclair, Brazil, 1982; Baker, Freebody, 1989):

– da un punto di vista *quantitativo*, l'insegnante è spesso colui/colei che in classe occupa il maggior numero di turni di parola nonché le sequenze più estese;

– la dominanza *interazionale* ha a che vedere con la responsabilità del docente di gestire gli scambi comunicativi nella classe, ad esempio selezionando i locutori o riportando l'ordine;
– dalla prospettiva *semantica*, è l'insegnante a selezionare gli argomenti oggetto di riflessione e discussione in aula;
– con dominanza *strategica* si intende la pianificazione dell'evento comunicativo operata dal docente, che ha in mente un determinato obiettivo da raggiungere o un percorso da seguire, di cui spesso gli studenti non sono consapevoli;
– la dominanza *conoscitiva* è, inoltre, legata allo sbilanciamento di determinate competenze e conoscenze a favore del docente.

Nel contesto di una classe di lingua seconda/straniera, inoltre, l'insegnante madrelingua detiene anche una dominanza più specifica che si potrebbe definire "nativa", data dall'essere l'unico possibile depositario dell'effettiva competenza linguistico-comunicativa su cui basare la descrizione delle forme e delle funzioni che saranno oggetto della lezione o del corso.

Nell'insegnamento delle lingue seconde/straniere, infatti, il parlato dei docenti rappresenta un caso particolare per il suo trovarsi in qualche modo a metà strada tra il *teacher talk* (Ferguson, 1975) e il *foreigner talk* o xenoletto, la varietà di lingua proposta a interlocutori non-nativi (Ferguson, 1971; Larsen-Freeman, Long, 1991), varietà che condividono alcune caratteristiche di parlato "modificato"[1]. Le modifiche che il docente compie nel proprio modo di parlare sono orientate agli ascoltatori, gli studenti, e rispondono, o meglio dovrebbero rispondere, alla necessità di offrire a questi ultimi un *input* che sia in primo luogo comprensibile (Krashen, 1985). Attraverso processi di semplificazione, regolarizzazione ed elaborazione (o chiarificazione – Voghera, 1992) l'insegnante può/deve tarare il proprio parlato sulle competenze linguistiche della classe cui si rivolge, proponendo enunciati formalmente corretti e funzionalmente adeguati ma anche il più possibile realistici e naturali (Diadori, 2004; Bettoni, 2001).

La presenza di tali modifiche, insieme alla negoziazione interazionale dei significati e delle forme, dovrebbe favorire la trasformazione dell'*input* in *intake* e quindi facilitare il processo di apprendimento linguistico (Gass, 1988; Swain, 1985).

La voce e il parlato sono, quindi, fondamentali strumenti didattici per un docente di lingua: essi costituiscono, infatti, il primo e forse più determinante *input* orale a disposizione degli apprendenti e rappresentano per questi ultimi un vero e proprio modello di riferimento (Vedovelli, 1994; 1999)[2].

[1] Altri esempi di varietà modificate, utilizzate per adeguarsi a una (presunta o reale) limitata competenza linguistica dell'interlocutore sono il *baby talk*, il parlato rivolto a bambini, e l'*elderspeak*, rivolto agli anziani (Cohen, Faulkner, 1986). Ferguson descrive queste varietà come "*a family of simplified registers which are used with various kinds of addressee who, for one reason to another, lack full competence in the language*" (1981: 10).

[2] La centralità dell'aspetto fonico si è, inoltre, ancor più evidenziata nelle recenti esperienze diffuse di didattica a distanza.

Il parlato dei docenti di italiano come lingua seconda o straniera è già stato oggetto di studio, in particolare nell'ambito del progetto CLODIS (*Corpus di Lingua Orale dei Docenti di Italiano a Stranieri* – Diadori, 2004) e in contributi più recenti (Corradi, 2012; Mertelj, 2020; Salvati, Russo 2021)[3]. Le analisi sono state finora principalmente condotte a livello morfosintattico, pragmatico/ conversazionale e a livello lessicale sebbene siano state individuate anche alcune caratteristiche fonetiche e prosodiche di tale varietà, tra cui[4]:

– iperarticolazione e assenza di forme contratte;
– velocità di articolazione e velocità di eloquio moderate;
– frequente uso di pause silenti;
– ampio *range* tonale e uso dell'intonazione funzionale a evidenziare gli elementi di novità, parole chiave e correzioni e favorire, quindi, il fenomeno del *noticing* (Schmidt, 1995; 2001).

2. *Lo studio*

Collocandosi all'interno di questo quadro teorico, il presente contributo riporta i risultati di uno studio che ha come obiettivo l'osservazione delle caratteristiche ritmiche e intonative del parlato di docenti di lingua italiana, parlanti nativi, che si trovano a interagire con due diversi target: studenti italofoni e apprendenti stranieri adulti. La novità rispetto alla letteratura già esistente sta nel tentativo di individuare e descrivere le eventuali modifiche che hanno luogo nella prosodia e nell'intonazione di una insegnante in una precisa situazione comunicativa come conseguenza della presenza di un diverso gruppo di interlocutori.

2.1 Due docenti e due contesti

Al fine di raggiungere tale obiettivo sono state coinvolte nella ricerca due docenti di lingua italiana, entrambe donne di origine campana, di 44 e 48 anni. Le docenti, che saranno identificate dalle iniziali CM e PS, sono state selezionate perché in possesso di un'esperienza almeno quinquennale di insegnamento in due diversi contesti:

– in quello che sarà da ora in poi indicato come contesto L1 (C_L1), cioè in scuole secondarie di secondo grado di Napoli e provincia;
– in contesto L2 (C_L2), ossia nell'ambito di una associazione di volontariato che offre corsi di lingua seconda per apprendenti immigrati adulti a Napoli. Nel caso specifico, le docenti coinvolte, al momento della raccolta dati, erano impegnate in lezioni di italiano L2 rivolte a classi con una competenza linguistica di livello A2 o B1 del Quadro Comune Europeo di Riferimento per le Lingue (Council

[3] Si veda anche il volume di Grassi (2007) per il caso specifico del parlato dei docenti nel contesto scolastico italiano, in presenza di allievi di origine straniera.

[4] Per una rassegna degli studi sulle caratteristiche acustiche del parlato rivolto a stranieri si veda Piazza, Martin & Kalashnikova (2022).

of Europe, 2001) e hanno entrambe dichiarato di utilizzare un approccio glotto-didattico di stampo comunicativo.

2.2 I dati e l'analisi acustica

Per registrare il parlato delle docenti, si è scelto di utilizzare un microfono Lavalier omnidirezionale con clip, che le due partecipanti hanno acconsentito a indossare e collegare a uno smartphone[5]. Questo ha permesso alle insegnanti di muoversi liberamente nello spazio dell'aula nel corso della lezione, non pregiudicando la qualità della registrazione ed evitando strumentazioni più invadenti. Per non compromettere i risultati, non sono stati resi noti alle due docenti gli scopi specifici della ricerca[6]. La volontà di preservare l'usuale *setting* lavorativo delle docenti coinvolte non ha reso possibile un controllo diretto sulle caratteristiche ambientali (ampiezza dell'aula, eventuale presenza di rumori di fondo). Nonostante ciò, il parlato delle docenti è risultato sempre intellegibile e le registrazioni percettivamente di buona qualità.

Sono state registrate otto lezioni di lingua italiana (due per ciascuna docente in ciascun contesto) rivolte a gruppi composti da un numero variabile di studenti (da 10 a 20), per un totale di circa tre ore di parlato[7]. Ai fini del presente contributo sono state selezionate dagli autori le porzioni che rispondessero ai seguenti criteri:

– presenza di un parlato esclusivamente/prevalentemente monologico, ossia della cosiddetta "comunicazione unidirezionale del docente in presenza degli studenti" (Diadori, 2007: 339);
– corrispondenza alla fase didattica della spiegazione di aspetti formali della lingua.

Sono state inoltre escluse le porzioni iniziali delle registrazioni (di circa 10 minuti), per evitare un eccessivo controllo del proprio parlato da parte delle docenti o, al contrario, una reazione emotiva all'avvio della registrazione.

Le otto porzioni, una per ciascuna lezione, corrispondono, in totale, a circa 490 secondi, 2.493 sillabe e 409 catene foniche[8]. In Tab. 1 sono riportate alcune informazioni sulla grandezza dei campioni selezionati, suddivise per docente e per contesto.

[5] È stato utilizzato uno smartphone Android (Xiami Redmi Note 7) e le registrazioni sono state sempre effettuate con l'applicazione "di serie", con frequenza di campionamento di 44,1 kHz e velocità di trasmissione di ~100kbit/s. Alle docenti è stato richiesto di attivare autonomamente la registrazione, prima di cominciare la lezione.

[6] Le registrazioni sono state effettuate in seguito al consenso scritto delle docenti e dopo aver ricevuto l'autorizzazione dei Dirigenti scolastici. Il parlato degli studenti e delle studentesse, peraltro spesso non intellegibile, non è stato specificamente oggetto di registrazione né di analisi.

[7] La varietà di parlato delle docenti può essere identificata come un italiano standard, con inflessioni regionali campane.

[8] Per catena fonica si intende una porzione di enunciato delimitata da due pause silenti successive (Pettorino, Giannini, 2005).

Tabella 1 - *Informazioni quantitative sui dati analizzati, per docente e contesto*

docente	*contesto*	*n. catene foniche*	*n. sillabe*	*n. silenzi*	*n. disfluenze*
CM	C_L1	116	721	122	19
	C_L2	114	595	115	6
PS	C_L1	88	625	90	9
	C_L2	91	552	94	14

L'analisi acustica, condotta manualmente con *Praat* (Boersma, Weenink, 2021), ha permesso di misurare:
- la durata di ciascuna catena fonica;
- la durata sillabica;
- la durata delle pause silenti;
- il valore massimo, minimo e medio di f_0 per ciascuna catena fonica.

Dai dati ottenuti sono stati calcolati, nel parlato delle due docenti e in entrambi i contesti, i seguenti indici, ripresi da Pettorino, Giannini (2005) e scelti per parziale analogia con quelli riportati nella letteratura sul parlato dei docenti (cfr. § 1):
- la Velocità di Articolazione (VdA), ottenuta dividendo il numero di sillabe per la durata delle catene foniche;
- la Velocità di Eloquio (VdE), calcolata dividendo il numero di sillabe per la durata totale dell'enunciato;
- la fluenza, data dal rapporto tra numero di sillabe e numero di catene foniche;
- la composizione del parlato, calcolata in valori percentuali di durata delle diverse componenti dell'enunciato, ossia le sequenze articolate, i silenzi e le disfluenze (prolungamenti, nasalizzazioni, vocalizzazioni, etc.);
- il *range* tonale.

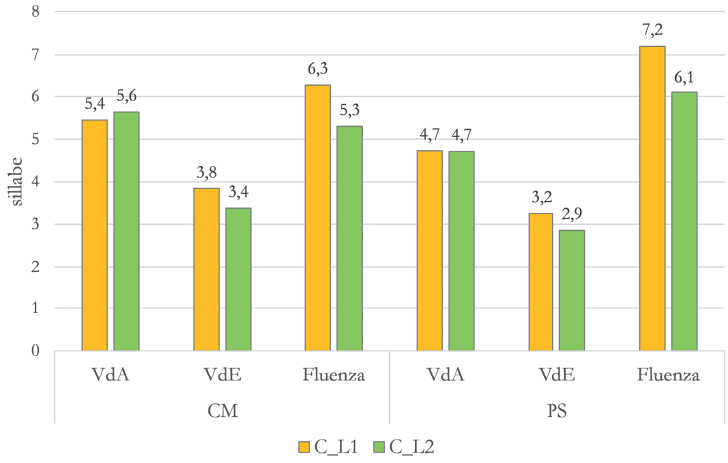
Per verificare la significatività delle differenze osservate tra i due contesti di comunicazione C_L1 e C_L2 sono stati applicati *t*-test per campioni appaiati con il software R, nella versione 4.2.2. La soglia di significatività, come di consueto nelle scienze umane, è stata fissata a 0,05.

2.3 Risultati

In Fig. 1 sono riportati i valori medi della VdA, della VdE e della fluenza nei dati analizzati. È possibile osservare come il diverso contesto e la presenza di interlocutori non nativi, contrariamente a quanto riportato nella letteratura sul tema, non sembrino influenzare la velocità con la quale le due docenti articolano i foni della lingua italiana. I valori di VdA nei due contesti sono infatti molto simili nelle produzioni di CM e in media addirittura identici per PS. Anche la VdE non differisce di molto tra C_L1 e C_L2 negli enunciati delle due partecipanti.

È invece la fluenza l'indice maggiormente condizionato dal contesto. Nel parlato di entrambe le docenti, infatti, la fluenza è maggiore in C_L1 rispetto a C_L2: di esattamente una sillaba per CM, di poco più per PS (1,1). I risultati dei *t*-test effettuati per la VdA, per la VdE e per la fluenza risultano, comunque, non significativi.

Figura 1 - *Velocità di articolazione, velocità di eloquio e fluenza nel parlato delle due docenti (CM e PS) nei due contesti (C_L1 e C_L2)*



Il dato sulla maggiore fluenza con interlocutori italofoni è confermato dalla minore durata media delle catene foniche in C_L2 per entrambe le docenti, come si osserva in Fig. 2. Dal punto di vista statistico, solo nel caso di CM si osserva una minima significatività nel confronto tra la durata delle catene foniche in contesto C_L1 (M=1,14; SE=0,069) e quelle in contesto C_L2 (M=0,92; SE=0,058) (t(228)=2,35; *p*=0,01). Nel parlato di PS la differenza tra i due contesti non risulta significativa.

Figura 2 - *Durata media delle catene foniche e deviazione standard nel parlato delle due docenti (CM e PS) nei due contesti (C_L1 e C_L2)*

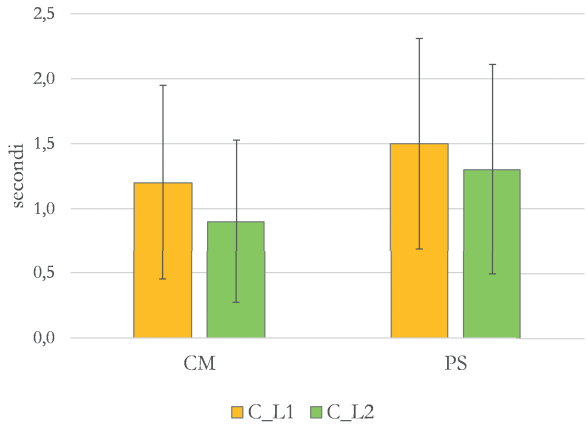
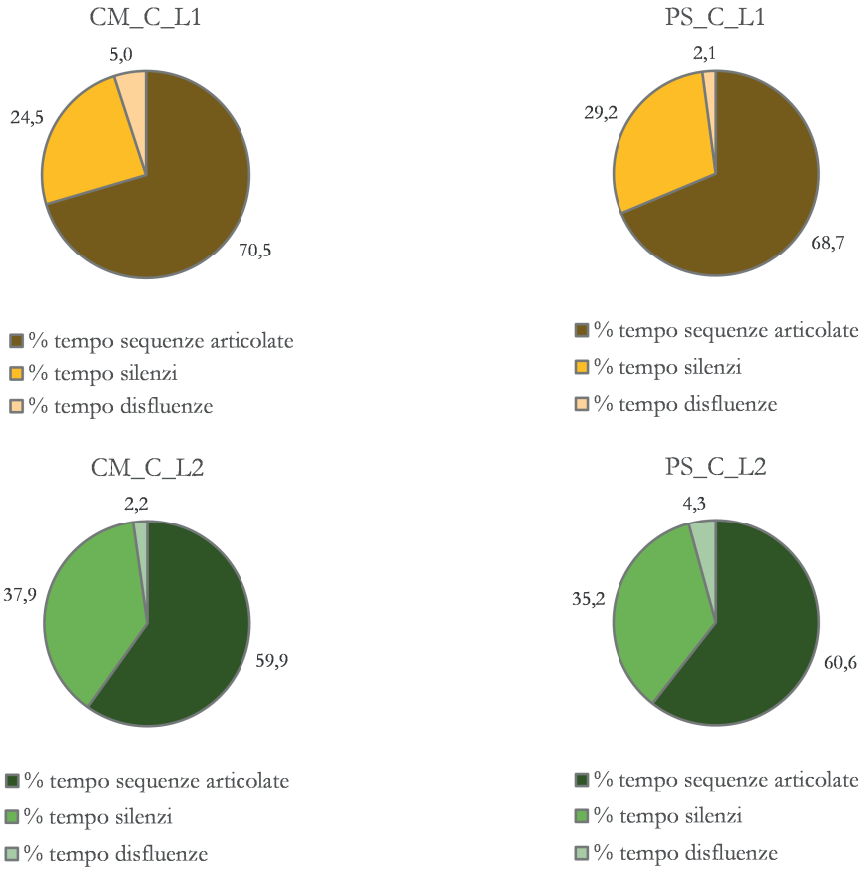


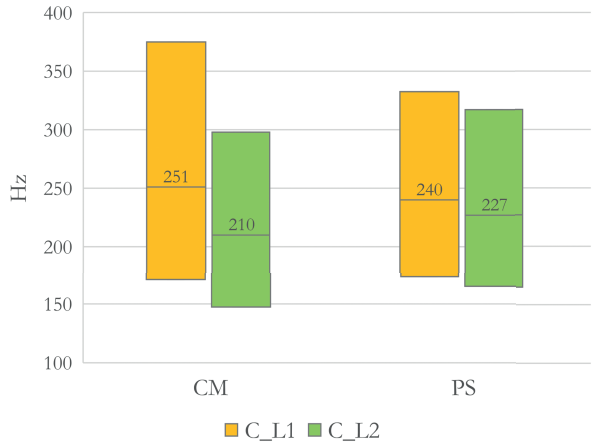
Figura 3 - *Composizione del parlato delle due docenti (CM e PS) nei due contesti (C_L1 e C_L2): valori percentuali di durata*



In Fig. 3 sono riportati i valori relativi alla composizione del parlato nei dati analizzati, dai quali si evince come i silenzi, oltre ad essere più frequenti in C_L2, aumentino anche in percentuale di durata nel caso della comunicazione rivolta ad apprendenti non nativi. Ciò avviene sia nel caso di CM, in cui l'aumento di pause silenti in C_L2 rispetto a C_L1 è di circa il 14%, sia nel parlato di PS, dove si assiste a un aumento del 6 %.

I dati relativi al *range* tonale e i valori medi di f_0 nel corpus considerato sono riportati in Fig. 4.

Figura 4 - Range *tonale e valori medi di f_0 in Hz nel parlato delle due docenti (CM e PS) nei due contesti (C_L1 e C_L2)*



È interessante osservare che entrambe le insegnanti presentano una maggiore escursione tonale e un valore medio di f_0 più alto nel contesto L1 rispetto al contesto L2. Contrariamente a quanto riportato in letteratura in relazione alle caratteristiche acustiche del *foreigner talk* (Piazza, Martin & Kalashnikova, 2022), i risultati di questo studio sembrano indicare che, nella situazione comunicativa della spiegazione grammaticale in una classe di lingua, la presenza di un interlocutore non nativo induce a un parlato meno modulato e con un tono mediamente più basso da parte dell'insegnante madrelingua. Anche in questo caso differenze più accentuate e statisticamente più significative si riscontrano nel parlato di CM (per la f_0 media, $t(228)=6{,}41$; $p<0{,}0001$) rispetto a PS ($t(176)=2{,}47$; $p=0{,}01$).

2.4 Discussione e interpretazione dei risultati

I risultati dell'analisi acustica condotta sul parlato di due docenti di lingua italiana in presenza di due diversi gruppi di interlocutori, nativi e non nativi, ha permesso di evidenziare comportamenti comuni, alcuni dei quali in parziale contraddizione con quanto riportato negli studi precedenti sul tema. Non è stata riscontrata, infatti, la differenza attesa tra gli enunciati prodotti da entrambe le insegnanti nei due diversi contesti in relazione alla velocità di articolazione e alla velocità di eloquio.

Una modifica che, invece, le due docenti apportano al proprio parlato in presenza di ascoltatori/interlocutori non italofoni e con un livello elementare/intermedio di competenza della lingua italiana riguarda la frequenza e la durata delle pause silenti. Come riportato anche da Corradi, l'insegnante "[...] in porzioni di parlato lungo inserisce sovente pause che danno una forma essenzialmente paratattica alla sintassi, la quale procede per giustapposizione di brevi enunciati" (2012: 250). Consenta il lettore qualche ipotesi interpretativa rispetto alla maggiore presenza di silenzi in C_L2: le numerose pause permettono probabilmente alle docenti

di pianificare il discorso, perché risulti comprensibile e comunicativamente efficace a un pubblico non nativo e con una competenza non avanzata nella lingua italiana; i silenzi, inoltre, potrebbero aiutare gli apprendenti a "metabolizzare" le informazioni ricevute nella seconda lingua. La minore frequenza e durata dei silenzi in contesto L1 potrebbero essere invece indice di un minore bisogno di programmazione *online* (da parte delle docenti) e di elaborazione (da parte degli studenti) delle forme linguistiche utilizzate nella spiegazione grammaticale.

Un aspetto interessante, che fa emergere invece stili individuali di parlato, è quello relativo alle disfluenze. Dal confronto tra le due docenti, infatti, emergono due diversi comportamenti: nel caso di CM la percentuale di parlato disfluente si dimezza in contesto L2 (da 5% in C_L1 a 2,2% in C_L2); viceversa, nel caso di PS, il valore percentuale relativo alle disfuenze raddoppia (da 2,2% in C_L1 a 4,3% in C_L2). Naturalmente, in assenza di un più ampio corpus di dati, non si può far altro che prendere atto di tale variabilità, peraltro tutt'altro che insolita nell'occorrenza dei fenomeni di disfuenza (a tal proposito, si vedano sull'italiano, ad esempio, Pettorino, Giannini, 2005; Schettino, Betz, Cutugno & Wagner, 2021).

Tuttavia, considerando i valori complessivi relativi alla composizione dell'enunciato (Fig. 3), si vede come le variazioni relative alla componente "disfluenze" vadano a scapito o a favore principalmente delle pause silenti. In altre parole, questi dati sembrano indicare un diverso modo di gestire le pause o "sospensioni" nel parlato da parte delle due parlanti. Se entrambe adattano il proprio eloquio al contesto L2 e, come già evidenziato, dedicano probabilmente più tempo alla programmazione dell'enunciato, i dati mostrano anche che le strategie adottate per "prendere tempo" differiscono nelle due docenti: CM, in contesto L2, fa ricorso quasi esclusivamente al silenzio; PS, invece, sembra riempire i silenzi con la voce, forse, inconsapevolmente, per non dare l'impressione a chi ascolta di essere incerta su cosa dire. Come riportato anche da Desideri,

> [...] per evitare il silenzio, che comunque palesa l'imbarazzante frammentarietà del discorso, il docente allunga le vocali finali di parola, riempie le pause di segnali generici come *ehm*, *mhm*, indicanti che la lezione continua ed è ancorata a quanto già enunciato (1992: 15).

I risultati relativi all'andamento intonativo e al *range* tonale appaiono in disaccordo con quanto riportato in letteratura, in relazione all'uso enfatico dell'intonazione con interlocutori non nativi, per veicolare le informazioni nuove, per stimolare i processi di attenzione e per esemplificare la realizzazione di diversi atti linguistici. La presenza di una più ridotta escursione tonale in contesto L2 rispetto al contesto L1 potrebbe essere ricondotta anch'essa alla necessità di maggior controllo degli enunciati in presenza di non nativi di cui si è discusso finora: una maggiore attenzione forse su "cosa" dire da parte delle docenti potrebbe essere causa di una meno marcata spontaneità rispetto al "come" tale cosa è comunicata attraverso la voce.

Questo aspetto, anche se comprensibile, suscita comunque alcuni interrogativi e delle perplessità. Ci si potrebbe domandare, infatti, oltrepassando di gran lunga i confini della presente ricerca, se effettivamente l'attenzione che una docente (di

lingua) pone alla scelta delle parole e delle forme giuste per veicolare i contenuti del proprio insegnamento sia affiancata da uno stesso livello di attenzione verso il modo in cui tale messaggio è veicolato nella concretezza della realizzazione vocale. In un contesto di didattica delle lingue seconde, il rischio nel proporre agli apprendenti un parlato prodotto in una fascia tonale piuttosto bassa (rassicurante), non "deformato" da picchi intonativi accentuati, caratterizzato da silenzi frequenti e lunghi, potrebbe essere quello di non fornire un *input* che sia sufficientemente realistico dal punto di vista ritmico-prosodico né rappresentativo degli usi linguistici al di fuori dell'aula.

Se quella degli autori sia una preoccupazione fondata, potrà dirlo solo l'analisi di dati più cospicui, che interessino tutte le fasi della didattica e non solo quella relativa alla spiegazione degli aspetti formali.

3. *Conclusioni*

Lo studio presentato in questo contributo ha avuto l'obiettivo di descrivere il parlato di due docenti di lingua italiana rivolto a diversi interlocutori, studenti nativi e non nativi con una competenza nella L2 di livello elementare/intermedio. I dati raccolti hanno permesso di evidenziare sia tendenze comuni nel comportamento verbale delle due insegnanti, parzialmente in disaccordo con i risultati di studi precedenti, sia stili individuali. Naturalmente, le riflessioni proposte e le interpretazioni personali fornite dagli autori sull'esiguo campione di dati raccolti andrebbero avvalorate o confutate attraverso la raccolta e l'osservazione di un corpus più ampio, con il coinvolgimento di più docenti, nonché estendendo e completando la descrizione del corpus già esistente su altri livelli di analisi.

Se un punto di forza della metodologia di raccolta dati proposta in questo studio è stato sicuramente quello di preservare l'autenticità della situazione comunicativa, non creando dei gruppi *ad hoc* ma effettuando registrazioni in classi già esistenti e già seguite dalle due docenti, questa procedura ha comportato almeno due limiti: la difficoltà nel reclutamento di insegnanti che fossero già impegnate nei due contesti ricercati; come già evidenziato nel §2.2, il mancato controllo diretto delle caratteristiche ambientali delle due aule di insegnamento (dimensioni effettive, interferenze acustiche, etc.).

In conclusione, con questa ricerca si vuole contribuire a rimarcare la mancanza di un vero dibattito sul parlato in ambito didattico, che rimane legato alle capacità comunicative personali del singolo docente, e l'assenza di "un consenso – e dunque di aspettative condivise – sul formato che i vari generi (*didattici*) debbono/dovrebbero assumere" (Ciliberti, Anderson, 1999: 26). Si vuole ribadire la necessità, invece, di una riflessione che sia fondata su dati empirici e che possa costituire una parte importante della formazione dei docenti.

Ringraziamenti

Gli autori intendono ringraziare le due docenti coinvolte in questa ricerca per la disponibilità, la collaborazione e la proficua discussione sul tema.

Riferimenti bibliografici

Baker, C., Freebody, P. (1989). Talk around text: Constructions of textual and teacher authority in classroom discourse. In De Castell, S., Luke, A., Luke, C., (eds.), *Language, authority and criticism*. London: Falmer Press, 263-283.

Bettoni, C. (2001). *Imparare un'altra lingua*. Roma-Bari: Laterza.

Boersma, P., Weenink, D. (2021). PRAAT: doing phonetics by computer. [software] Versione 6.1.40. https://www.praat.org/

Ciliberti, A. (1981). Approcci teorici nella descrizione del "linguaggio scientifico" e loro utilizzazione didattica. In Ciliberti, A. (a cura di), *L'insegnamento linguistico per "scopi speciali"*. Bologna: Zanichelli, 7-36.

Ciliberti, A. (2003). Collaborazione e coinvolgimento nella classe multilingue. In Ciliberti, A., Pugliese, R. & Anderson, L. (a cura di), *Le lingue in classe. Discorso, apprendimento, socializzazione*. Roma: Carocci, 123-142.

Ciliberti, A., Anderson, L. (a cura di) (1999). *Le forme della comunicazione accademica. Ricerche linguistiche sulla didattica universitaria in ambito umanistico*. Milano: Franco Angeli.

Cohen, G., Faulkner, D. (1986). Does "elder- speak" work? The effect of intonation and stress on comprehension and recall of spoken discourse in old age. In *Language and Communication*, 6, 91-98.

Corradi, D. (2012). Il parlato dell'insegnante nella classe di lingua. In *Italiano LinguaDue*, 4(2), 226-257.

Council of Europe (2001). *Common European Framework for Languages: Learning, Teaching, Assessment*. Strasbourg: Council for Cultural Co-operation, Education Committee, Modern Languages Division.

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

Desideri, P. (1992). Lo statuto linguistico della lezione: tecniche e operazioni pragmatiche dell'interazione verbale in classe. In Brasca L., Zambelli M.L. (a cura di), *Grammatica del parlare e dell'ascoltare a scuola*, Quaderni del Giscel. Firenze: La Nuova Italia, 187-199.

Diadori P. (a cura di) (2007). *La DITALS risponde 5*. Perugia: Guerra Edizioni.

Diadori, P. (2004). Teacher-talk/foreigner-talk nell'insegnamento dell'italiano L2: un'ipotesi di ricerca. In Maddii, L. (a cura di), *Apprendimento e insegnamento dell'italiano L2*. Firenze-Atene: IRRE Toscana-Edilingua Edizioni, 71-102.

Ferguson, C.A. (1971). Absence of copula and the notion of simplicity. In Hymes, D. (ed.), *Pidginization and creolization of languages.* Cambridge: Cambridge University Press, 141-150.

Ferguson, C.A. (1975). Towards a characterization of English foreigner talk. In *Anthropological Linguistics*, 17, 1-14.

Ferguson, C.A. (1981). 'Foreigner Talk' as the name of a simplified register. In *International Journal of the Sociology of Language*, 28, 9-18.

Gass, S.M. (1988). Integrating research areas: A framework for second language studies. In *Applied Linguistics*, 9(2), 198-217.

Grassi, R. (2007). *Parlare all'allievo straniero. Strategie di adattamento linguistico nella classe plurilingue*. Perugia: Guerra Edizioni.

Krashen, S. (1985). *The input hypothesis: Issues and implications*. Beverly Hills: Laredo Publishing Company.

Larsen-Freeman, D., Long, M.H. (1991). *An introduction to second language acquistion research*. London: Longman.

Mertelj, D. (2020). L1 e traduzione didattica nel teacher talk degli insegnanti d'italiano e d'inglese come lingue straniere per scopi specialistici. In *Italiano LinguaDue*, 2, 313-324.

Orletti, F. (2000). *La conversazione diseguale*. Roma: Carocci.

Pallotti, G. (1998). *La seconda lingua*. Milano: Bompiani.

Piazza, G., Martin, C.D., Kalashnikova, M. (2022). The acoustic features and didactic function of foreigner-directed speech: A scoping review. In *Journal of speech, language, and hearing research: JSLHR*, 65(8), 2896–2918.

Pettorino, M., Giannini, A. (2005). Analisi delle disfluenze e del ritmo di un dialogo romano. In Albano Leoni, F., Giordano, R. (a cura di), *Italiano parlato. Analisi di un dialogo*. Napoli: Liguori editore, 89-104.

Salvati, L., Russo, I. (2021). Indicatori di complessità nel parlato degli insegnanti di italiano L2: un'analisi quantitativa. In *Italiano LinguaDue*, 2, 122-132.

Schettino, L., Betz, S., Cutugno, F. & Wagner, P. (2021). Hesitations and individual variability in Italian tourist guides' speech. In Bernardasci, C., Dipino, D., Garassino, D., Negrinelli, S., Pellegrino, E. & Schmid, S. (a cura di), *Speaker individuality in phonetics and speech sciences: Speech technology and forensic applications*, Studi AISV 8. Milano: Officinaventuno, 243-262.

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In Schmidt, R. (a cura di), *Attention and awareness in foreign language learning.* Honolulu: University of Hawaii, Second Language Teaching & Curriculum Center, 1-63.

Schmidt, R. (2001). Attention. In Robinson, P. (a cura di), *Cognition and second language Instruction*. Cambridge: Cambridge University Press, 3-32.

Sinclair, J., Brazil, D. (1982). *Teacher talk.* Oxford: Oxford University Press.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In Gass, S., Madden, C. (eds.), *Input in Second Language Acquisition*. New York: Newbury House, 235-256.

Vedovelli, M. (1994). L'italiano parlato dagli italiani e l'italiano appreso dai non italiani. In De Mauro, T. (a cura di), *Come parlano gli italiani*. Firenze: La Nuova Italia, 87-98.

Vedovelli, M. (1999). Il parlato nella didattica della L2: le ragioni della naturalezza e dell'apprendimento. In Vedovelli, M. (a cura di), *Indagini sociolinguistiche nella scuola e nella società italiana in evoluzione*. Milano: Franco Angeli, 225-238.

Voghera, M. (1992). *Sintassi e intonazione dell'italiano parlato*. Bologna: Il Mulino.

DOI: 10.17469/O2109AISV000007

MARTINA ROSSI

Gender influence on phonetic turn-taking cues at potential transition locations in German

Phonetic cues have been found to play a fundamental role in the turn-taking mechanism by helping to signal to the interlocutor(s) the intentions of the current speaker for the upcoming turn. Since previous sociolinguistic research described the existence of gender specific behaviors in interactions, it could be the case that interlocutors' genders might influence the way different speakers use turn-taking cues to signal their conversational intentions. This research aims at investigating the influence of the gender of the speaker and the gender of the interlocutor on phonetic turn-taking cues, i.e., F0 movements, intensity and segmental duration, towards potential transition locations in two-party conversations between German native speakers. The results suggest that both the speakers' and the interlocutors' genders might influence the way potential transition locations are phonetically marked in conversations.

Keywords: turn-taking, German, interaction, phonetic variation, gender influence.

1. *Introduction*

Conversational interactions are a fundamental part of human social behavior. On a daily basis, individuals are involved in face to face or mediated conversations, producing about 1200 talk spurts amounting to around 2 to 3 hours of speech (Mehl, Vazire, Ramírez-Esparza, Slatcher & Pennebaker, 2007; Levinson, Torreira, 2015). It is clear from even the most casual exchanges that people in conversation do not speak at random: in general, it can be expected that there will be one person talking at a time and that the interlocutor(s) will wait for them to finish speaking to launch their turn, usually trying to keep silent gaps and speech overlaps reduced to a minimum (Sacks, Schegloff & Jefferson, 1974). In order to achieve a smooth exchange of turns, speakers seem to be able to predict the approach of a potential transition location (PTL from now on), i.e., a point in the current speakers' turn when speaker change becomes a possibility (Schiffrin, 1987; Transition Relevance Places, Sacks et al. 1974; Potential Turn Boundaries, Zellers, 2016), which allows them to start planning their next conversational move. This can either be to take up the next turn, to remain silent and let the speaker continue, or to produce a non-interrupting backchannel to signal attention. In fact, interlocutors do not wait for the current speaker to finish with their turn before deciding what to do next; they start encoding their following turn while the current one is still ongoing (Levinson, Torreira, 2015), so that they are able to respond appropriately at the appropriate time. Evidence in favor of early planning within the current turn is given by the discrepancy between the language production system's latencies for the encoding

of a single word and of a simple short clause, which are respectively of around 600 ms (Indefrey, Levelt, 2004; Schnur, Costa & Caramazza, 2006; Indefrey, 2011) and 1500 ms (Griffin, Bock, 2000), and the average duration of a silent gap in between turns across several languages, which amounts to around 200 ms (Stivers, Enfield, Brown et al., 2009; Heldner, Edlund, 2010).

Predictions about the approach of a PTL and about what will come after it are made by interlocutors on the basis of turn-taking cues in the current speaker's turn. Previous studies have identified several communicative means that are used by speakers to signal their conversational intentions for the next turn: gestural (Hadar, Steiner, Grant & Rose, 1984; Paggio, Navarretta, 2011; Edlund, Beskow, 2007, 2009; Mondada, 2007; Zellers, Gorisch, House & Peters, 2019), linguistic (Ford, Thompson, 1996; Local, Walker, 2012; Levinson, 2013; Levinson, Torreira, 2015) and phonetic (Yngve 1970; Local, Kelly & Wells, 1986; Hjalmarsson, 2011; Gravano, Hirschberg, 2011). Among the phonetic cues, the variation of F0, intensity and segmental duration have been found to significantly contribute to the turn-taking mechanism in several languages (e.g., Gravano, Hirschberg, 2009; Zellers, 2016; Brusco, Vidal, Beňuš & Gravano, 2020; Ishimoto, Teraoka & Enomoto, 2017), including German (Kohler, 1983; Selting, 1996; Niebuhr, Görs & Graupe, 2013; Peters, 2006; Dombrowski, Niebuhr, 2005). In particular, PTLs followed by a speaker change tend to be marked by either a rising or falling intonation and by a decrease in intensity, while turn holds are preceded by level F0 contours and higher intensity profiles. Backchannels also seem to be preceded by determinate sets of prosodic cues, such as regions of low pitch and some cases of uptalk in English and Japanese spontaneous conversations (Ward, Tsukahara, 2000), and final rising intonation (Skanze, Schlangen, 2009), together with higher intensity values (Gravano, Hirschberg, 2011) in task-oriented interactions in Swedish and in English. Mixed patterns of variation towards PTLs are found for segmental duration. While some studies give evidence for pre-boundary lengthening before turn yields (e.g., Local et al., 1986, for British English; Gravano, Hirschberg, 2011, for American English; Niebuhr et al., 2013 for German), others find increased segmental duration before turn holds and faster speech rate before speaker changes (e.g., Koiso, Horiuchi, Tutiya, Ichikawa & Den, 1998; Zellers, 2016; Brusco et al., 2020).

1.1 Gender variation in conversational behavior

Sociolinguistic research on verbal interactions has revealed the existence of social variation in the way turn-taking takes place: in particular, the interlocutors' genders appear to have a strong role in conversational behavior. Tannen (1994, 1998) proposed that, as a result of their different social and cultural background, men and women tend to use and interpret linguistic strategies such as interruptions, taciturnity and indirectness in contrasting ways. For instance, all-female talk has been observed to be often characterized by a collaborative floor with co-constructed utterances (Edelsky, 1993), cooperative overlaps and frequent minimal responses (Menz, Al-Roubaie, 2008; Stubbe, 2013), while men interacting with each other tend to stick to the one-

speaker-at-a-time model, with either long uninterrupted turns or short rapid ones, where overlaps are rare and perceived as deviant (Coates, 2004). Thus, while in some contexts conversational strategies like interruptions might be used to show support and achieve cooperation, in others they can also be employed to reinforce asymmetry and establish dominance over the other speaker. Asymmetrical interactional patterns have been often observed in mixed-gender conversations, with one speaker saying too much or too little, interrupting the other or virtually withdrawing from the conversation (e.g., Waara, Shaw, 2006; Coates, 2004; Zimmerman, West, 1975). The configuration of same-gender and mixed-gender interactions, however, does not always follow these exact patterns: results from different studies give oftentimes seemingly inconsistent results, since the influence of speakers' genders on their turn-taking behavior seems to be mediated by other social and contextual factors, such as the conversational goal and interactional setting, institutionalized roles and gender identity salience (Plug, Stommel, Lucassen, Dulmen & Das, 2021).

The majority of the scientific literature on gender variation in turn-taking has strongly focused on pragmatics, addressing, for example, talkativeness, the use of tentative language, and interruptions. It is not clear, however, if the similarities and the differences between genders also extend to the use they make of turn-taking cues, such as phonetic ones, to signal to the interlocutor their conversational intentions of offering the next turn or continuing to speak. We are aware of the overall differences in F0, intensity and segmental duration in men's and women's speech (e.g., women tend to have a higher mean F0 than men [Weirich, Simpson, 2019], men have higher conversational intensity levels than women [Gelfer, Young, 1997] and slightly lower durational values as well [Pépiot, 2014]). However, the possible influence of gender on the phonetic variation of these features as turn-taking cues has, to the best of the author's knowledge, never been empirically investigated in any language. Thus, the analysis presented in the current paper aims at proposing a first sociophonetic exploration of turn-taking by testing if, and how, social variables such as the gender of both interlocutors in a dialogue might have an effect on the acoustic cues that speakers use to signal their intentions for the next conversational turn.

2. *Methods*

In order to offer more insight into the variation of phonetic turn-taking cues in spontaneous interactions in German and to investigate the possible influence of gender on such variation, a dataset composed of two-party conversations between German native speakers, balanced for the gender of the interlocutors, was annotated, and values for F0, intensity and segmental duration were extracted at different test locations approaching a PTL.

2.1 Dataset

The two-party conversations analyzed for this study are part of the German sub-corpus of the DUEL Multi-lingual Multimodal Dialogue Corpus (Hough, Tian, De

Ruiter, Betz, Kousidis, Schlangen & Ginzburg, 2016). The DUEL corpus consists of naturalistic, face-to-face conversations based off loosely-directed tasks assigned to the participants. The tasks were specifically designed to provide subjects with a theme to discuss without, at the same time, being constrained in how to develop their interactions (*ibid*). In particular, in the portions of dialogues analyzed for the present research, the speaker pairs dealt with either the "Dream Apartment" or the "Film Script" tasks, in which participants had to imagine and plan out the organization, furniture and decoration of a hypothetical shared apartment or imagine and describe a movie scene centered on an embarrassing situation, respectively. The two interactants in each dyad were recorded on separate channels, which allows the phonetic analysis of speech even when the participants talk in overlap with each other. For the present study, 6 dyads have been annotated and analyzed, for a total of 12 different speakers, all German native speakers and (at the time of the recording) students at the University of Bielefeld, aged 21-28. The dyads were selected with the aim of having both speakers of the same gender in conversation with each other, as well as subjects of different genders interacting with each other. The dyads thus included in the dataset for the study are 2 male-to-male (MM) conversations, 2 female-to-female (FF) conversations and 2 mixed-gender (MF) conversations (see Tab. 1). The first 5 minutes of the tasks have been taken into consideration for this analysis.

2.2 Annotation

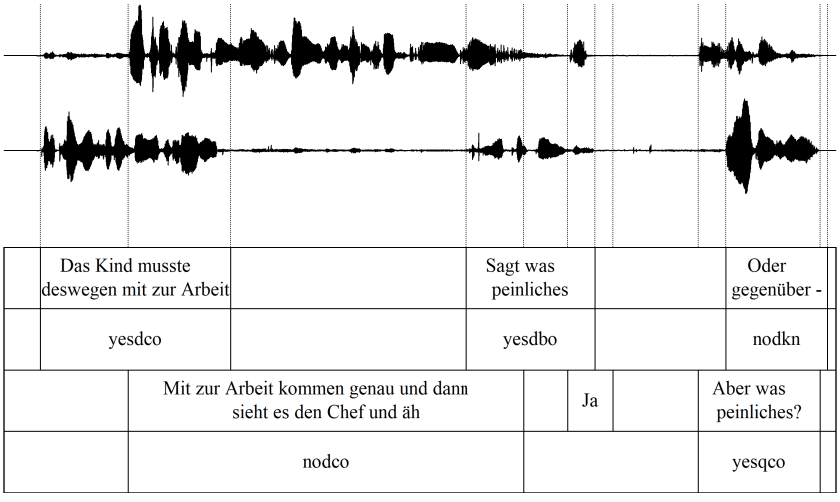
The annotation was carried out in *Praat* (Boersma, Weenink, 2022; Boersma, 2001) and it started with the manual annotation of PTLs (Schiffrin, 1987). As described by Sacks et al. (1974), such locations in conversations, defined by them as Transition Relevance Spaces, arise at the end of Turn Constructional Units, when speaker change becomes possible, though not mandatory. Following Zellers' (2016) classification of Potential Turn Boundaries, these locations were identified not only after syntactically complete utterances, but also after syntactically incomplete utterances which functioned as semantically or pragmatically complete in the local context of the interaction. Such locations were identified in the data using the right boundaries of the "X-utts" tier provided in the DUEL corpus, which contained the segmentation and the orthographic transcription of speakers' utterances within each turn (annotated in the "X-turns" tier). The guidelines for the segmentation of the utterances given to the annotators of the DUEL corpus follow the notion of "slash unit" described by Meteer et al. (1995), in which an utterance should be comprised by maximally a complete sentence or a smaller unit, which may not be syntactically complete but is judged as complete in context by the annotators (Hough et al., 2016). Turns in the "X-turns" tier, instead, included all continuous stretches of speech by one speaker until the other one takes up the floor, or up until a silent gap after which either the current speaker continues talking, or the interlocutor takes up the next turn. By assuming the right boundaries of the "X-utts" tier as PTLs, it was possible to include in the analysis instances considered by the annotators as complete in context, but that were not necessarily followed by a speaker change or by a silent gap.

Once PTLs were identified, they were annotated using a set of labels (Feindt, Rossi & Zellers, 2021; Rossi, Feindt & Zellers, 2022a) describing the utterance by the current speaker and what came after it (see Fig. 1):

- Completeness label: indicated the syntactic/semantic completion of the utterance in context ("yes" for complete utterances and "no" for incomplete utterances);
- Sentence Type label: described the form of the utterance ("d" for declaratives, "q" for questions and "t" for tag questions);
- Sequential Structure label: indicated the conversational action by the current speaker or by the other participant which followed the PTL, i.e., the other participant took up the next full turn (speaker changes, labelled with "c", also referred to as turn yields), the current speaker held the floor (keeps, labelled with "k", also referred to as turn holds), the other participant produced a minimal, non-interrupting response (backchannels, labelled with "b");
- Transition Type label: described the way in which the conversational action following the PTL took place, i.e. with a silent gap ("g"), with a speech overlap ("o"), in a smooth way, without any perceivable silences or speech overlaps (no-gap-no-overlap, "n"); in particular, only those silences longer than 120ms were annotated as gaps, and only those stretches of overlapped speech longer than 120ms were labelled as overlaps, as it has been shown that gaps and overlaps shorter than 120ms are not perceived as such by listeners (Heldner, 2011). Transitions occurring with possible silences or overlaps shorter than 120ms were considered as no-gap-no-overlaps.

For each utterance, words, syllables and segments were also annotated.

Figure 1 - *Example of the annotation of PTLs in Praat, from one of the dialogues in the dataset. The first part of the label(s) refers to the completeness in context ("yes" or "no"), the following letter refers to the sentence type ("d" for declarative, "q" for questions, "t" for tag questions), the third part of the label refers to the sequential structure that followed ("c" for speaker change, "k" for keep and "b" for backchannel), and the final letter describes the transition type ("g" for gaps, "o" for overlaps, "n" for no-gap-no-overlaps)*



2.3 Data extraction

As a part of a wider research project aimed at investigating the location and the extension of the transition space in conversation and testing the hypotheses related to the relevance for turn-taking of time windows against that of phonological categories (issues that will not be discussed in the current paper), the present analysis observes the variation of the phonetic parameters using three different time-points approaching a PTL as reference (Feindt et al., 2021). The phonetic parameters investigated are F0, intensity and segmental duration, and the three test time-points approaching a PTL are located at 500 ms before the end of the utterance, at 200 ms before the end of the utterance, and at the end of the utterance. The phonetic values were extracted automatically using a *Praat* script from utterances with a duration of 1 second and up, so that it was possible to extract data from all three test locations without them being too close to the start of the utterance (Feindt et al., 2021). F0 readings, extracted using *Praat* with the settings for semitones (st) above 1Hz, were then normalized with the individual speaker's baseline in order to exclude the influence of physiological factors. Using a sample of F0 datapoints in semitones extracted from 2 minutes of clear speech for each subject, the individual baseline was calculated as 0.75 times the first quartile of the data, following Zellers and Schweitzer (2017) and Zellers (2021). Intensity was extracted in decibels (dB) and normalized with the speakers' mean (Ludusan, Dupoux, 2015). Values for segmental duration were also extracted at three different test intervals approaching the PTL: the average segmental duration, in milliseconds (ms), was extracted over the last 500 ms and over the last 200 ms (i.e., from the offset of speech to 500 ms before that and from the offset to 200 ms before that) and, finally, the duration of the last segment at the end of the utterance was extracted. This way, it was possible to observe how average segmental duration varied towards PTLs. No normalization was carried out at this stage for segmental duration; however, the randomness of the segments included in the analysis allows us to make general preliminary observations about the variation of segmental duration approaching PTLs, excluding the potential influence of the different segment types.

3. *Analysis and results*

A total of 489 PTLs with the related phonetic values at the three test locations were extracted. The qualitative and quantitative analysis of the data was carried out in *R* (RStudio Team, 2020). Linear mixed effects models with the subject as a random factor were used for the quantitative analysis of the data using the *R* package lmerTest (Kuznetsova, Brockhoff & Christensen, 2017).

From a first qualitative exploration of the data, differences between the behavior of the male and female speakers in conversation appear. For instance, the numbers of speaker changes and keeps in FF conversation were equal: female speakers tend to yield the turn as much as they hold it (78 changes and 70 keeps); on the contrary, in MM conversation there is a bigger disproportion between turn holds and speaker

changes, with male speakers keeping the floor for longer, and more frequently than they are ceding it (44 changes and 92 keeps). Moreover, it was observed that smooth transitions (i.e. no gap no overlap) occurred more frequently in same-gender conversation, thus 26% of the transitions are smooth in MM conversation and 34% in FF conversations, while only 16% of transitions were smooth in mixed-gender conversations (see Tab. 2). Silent gaps occur more frequently in MM conversations than in FF and MF ones, while speech overlaps are observed more frequently in FF dyads (see Tab. 1).

Table 1 - *Transition type (gaps, overlaps, no-gap-no-overlaps) distribution in same-gender ("MM" for male-to-male, "FF" for female-to-female) and mixed-gender ("FM" for female-to-male) dyads*

	gaps	*overlaps*	*no-gap-no-overlaps*
MM	61% (97)	13% (20)	26% (41)
FF	39% (64)	28% (46)	34% (56)
MF	56% (93)	27% (45)	16% (27)

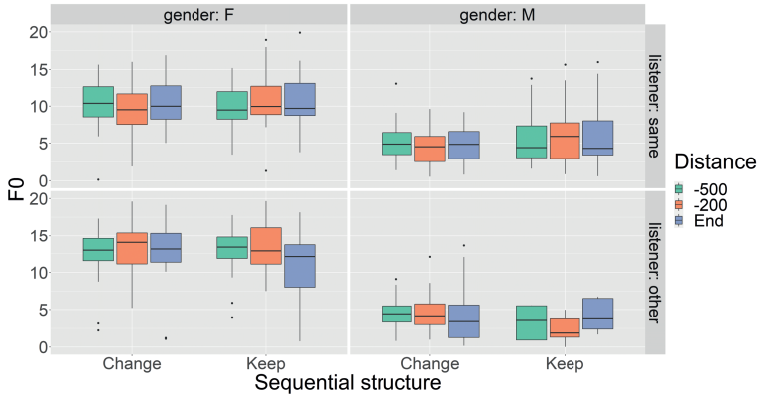
The analysis reported below focusses on the pre-final phonetic marking of syntactically/semantically complete utterances in declarative form preceding speaker change and keep cases, and how the gender of the interlocutors influences that.[1] The PTLs thus considered for the following analysis are 239, including 131 followed by speaker change cases and 108 by turn holds, constituted by syntactically/semantically complete utterances in declarative form.

3.1 F0 values

The variation of the normalized F0 values towards PTLs revealed significant differences both between the speaker changes and keeps, and between male and female interlocutors and the way in which they approach different sequential structures. Male speakers' values at the three test locations analyzed (at the end of the utterance, at 200 ms from the boundary and at 500 ms from the end of the utterance) are closer to their baseline, at around 5 st, while the average values for female speakers are higher, mostly above 10 st.

[1] Due to space constraints, the analysis involving transition types (i.e., gaps, overlap and smooth transitions) will not be reported, and, for the purposes of this study, sentence types such as questions and tag questions are momentarily set aside, in an attempt to limit the sources of prosodic variation. Finally, the exclusion of TRPs preceding backchannels is due to limitations in the current annotation scheme and analysis structure. In fact, several factors have been found to influence backchannel's placement (e.g., the lexical or non-lexical content of the backchannel, or its modality, cfr. Truong, Poppe, De Kok & Heylen, 2011; Ferré, Renaudier, 2017), and would thus have to be considered when targeting their distribution and the cues preceding them. Since such factors are not present in this annotation scheme, nor they would fit into the rest of the analysis, the description that would have resulted would have been biased and ambiguous. Further developments of this research, though, include a more precise annotation of backchannels, as well as their analysis.

Figure 2 - *F0 datapoints distribution (in st, above the speaker's baseline) for speaker change and keep cases at the three test locations ("Distance": -500 ms, -200 ms, End), for female speakers (gender: F) in same (listener: same) and mixed-gender dialogues (listener: other), and for male speakers (gender: M) in same (listener: same) and mixed-gender dialogues (listener: other)*



Considering both the gender of the speaker and the gender of the interlocutor, in speaker change cases the values of male speakers are closer to their baseline, both in MM and MF conversations. In all conditions, the contour created by the three datapoints tends to be flat, though with a certain degree of variability, with the exception of male speakers in MF conversations, where the values of F0 at the End location fall closer to the speakers' baselines, at 2.5 st, while for female speakers they tend to remain higher, closer to the previous ones, at 12.5 st (see Fig. 2). Moreover, in both speaker change cases and keep cases (see Fig. 2), it appears that female speakers are always higher than males, but they are even higher when in conversation with a male speaker; similarly, male speakers in MM conversation show low values, but males in conversation with females are even lower.

Table 2 - *Summary of the linear mixed model for F0 in speaker change cases. A significant three-way interaction is found between the value of the final F0 test location (End), the gender of the speaker and the gender of the interlocutor. The speaker is included as a random factor in the model. Formula: lmer(F0 ~ distance * gender * listener + (1 | speaker))*

	Estimate	*Std. Error*	*DF*	*t-value*	*Pr(>|t|)*
(Intercept)	11.0322	1.6712	9.2173	6.601	8.88e-05 ***
-200: Gender M	2.4714	0.8134	1465.0749	3.038	0.002420 **
End: Gender M	4.4082	0.8134	1465.0749	5.420	6.98e-08 ***
-200 listener other	1.9198	0.8134	1465.0749	2.360	0.018391 *
End: listener other	4.5724	0.8134	1465.0749	5.621	2.26e-08 ***

	Estimate	*Std. Error*	*DF*	*t-value*	*Pr(>|t|)*
Gender M: listener other	-2.2018	4.0219	8.5934	-0.54	0.597998 ns
*-200*gender M: listener other*	-1.5617	1.2932	1465.0749	-1.20	0.227398 ns
*End*gender M: listener other*	-7.1274	1.2932	1465.0749	-5.51	4.20e-08 ***

A linear mixed model shows that the value of F0 at the PTL differ significantly for male speakers in mixed-gender conversations (see Tab. 2). A significant three-way interaction between the distance from the PTL, the gender of the speaker and the gender of the interlocutor is found for keep cases, too: the height of F0 at the –200 datapoint differ significantly from the intercept for male speakers in mixed-gender conversations (see Tab. 3). In these interactions, F0 is rising at the end of the utterances for male speakers, while it is falling for female speakers (see Fig. 2).

Table 3 - *Summary of the linear mixed model for F0 in turn hold cases. A significant three-way interaction is found between the value of the penultimate F0 test location (-200), the gender of the speaker and the gender of the interlocutor. The speaker is included as a random factor in the model. Formula: lmer(F0 ~ distance * gender * listener + (1 | speaker))*

	Estimate	*Std. Error*	*DF*	*t-value*	*Pr(>|t|)*
(Intercept)	9.1696	1.2859	7.3667	7.13	0.000148 ***
-200: Gender M	-0.8469	0.4724	2896.9797	-1.79	0.073108 .
End: Gender M	-1.4471	0.4724	2896.9797	-3.06	0.002209 **
-200 listener other	-0.3076	0.5112	2896.9797	-0.60	0.547383 ns
End: listener other	-2.3349	0.5112	2896.9797	-4.56	5.15e-06 ***
Gender M: listener other	-2.5707	3.6319	7.3266	-0.70	0.500957 ns
*-200*gender M: listener other*	-3.6392	0.9743	2896.9797	-3.73	0.000191 ***
*End*gender M: listener other*	1.3532	0.9743	2896.9797	1.389	0.164965 ns

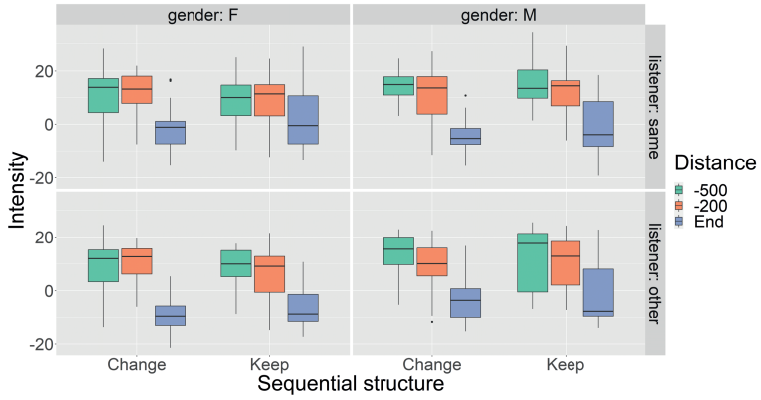
3.2 Intensity

For each of the sequential structures analyzed, intensity decreases approaching the end of the utterance, with the values of the last test location dropping below the speakers' means. The decrease is higher for PTLs preceding speaker change cases, where it appears that speakers lower their voices much more than before keeps.

In speaker change cases, male speakers appear to be louder than females, except at the end of the turn in MM conversations, where their values drop at –5 db below their means, while the final values for females in FF conversations are closer to their

means. On the contrary, female speakers talking to a male conversational partner tend to lower their intensity much more approaching the end of the turn, while males in conversation with a female partner end their turns with a higher intensity, closer to their means (see Fig. 3).

Figure 3 - *Intensity datapoints distribution (in dB, normalized to the speaker's mean) for speaker change and keep cases at the three test locations ("Distance": -500 ms, -200 ms, End), for female speakers (gender: F) in same (listener: same) and mixed-gender dialogues (listener: other), and for male speakers (gender: M) in same (listener: same) and mixed-gender dialogues (listener: other)*



For keep cases, when the speaker held the floor, intensity decreases less towards the PTL. In fact, in these cases the intensity values at the last test location fall below the speakers' means, but they remain closer to it than in speaker change cases. Considering both the speakers' and the interlocutors' gender, however, it is possible to detect an exception in this trend: intensity values at the end of the utterance for male speakers in MF dyads decrease well below their mean, more than it does for female speakers in MF dyads.

For both speaker change cases and keep cases, two separate linear mixed effects models show a three-way interaction between the gender of the speaker, the gender of the interlocutor and the intensity at the final test location (see Tab. 4 and Tab. 5).

Table 4 - *Summary of the linear mixed model for intensity in speaker change cases. A significant three-way interaction is found between the value of the final test location, the gender of the speaker and the gender of the interlocutor. The speaker is included as a random factor in the model. Formula: lmer(intensity ~ distance * gender * listener + (1 | speaker))*

	Estimate	*Std. Error*	*DF*	*t-value*	*Pr(>|t|)*
(Intercept)	10.1865	1.062	9.7709	9.58	2.78e-06 ***
-200: Gender M	-3.2728	0.858	3517.058	-3.81	0.000139 ***

	Estimate	*Std. Error*	*DF*	*t-value*	*Pr(>|t|)*
End: Gender M	-6.6951	0.858	3517.058	-7.80	7.88e-15 ***
-200 listener other	0.1572	0.850	3517.058	0.18	0.853326 ns
End: listener other	-6.7422	0.850	3517.058	-7.92	2.94e-15 ***
Gender M: listener other	1.0235	2.576	9.3741	0.37	0.700063 ns
*-200*gender M: listener other*	-1.8130	1.274	3517.058	-1.42	0.154837 ns
*End*gender M: listener other*	8.8455	1.274	3517.058	6.94	4.57e-12 ***

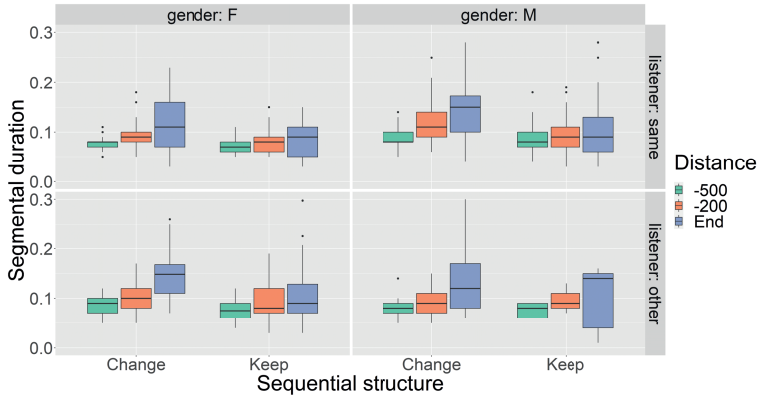
Table 5 - *Summary of the linear mixed model for intensity in turn hold cases. A significant three-way interaction is found between the value of the final test location, the gender of the speaker and the gender of the interlocutor. The speaker is included as a random factor in the model. Formula: lmer(intensity ~ distance * gender * listener + (1 | speaker))*

	Estimate	*Std. Error*	*DF*	*t-value*	*Pr(>|t|)*
(Intercept)	10.3673	2.121	7.5234	4.88	0.00145 **
-200: Gender M	-2.6697	0.943	2896.948	-2.83	0.00468 **
End: Gender M	-7.3551	0.943	2896.948	-7.79	8.69e-15 ***
-200 listener other	-2.7057	1.020	2896.948	-2.65	0.00807 **
End: listener other	-6.9076	1.020	2896.948	-6.76	1.58e-11 ***
Gender M: listener other	-0.4589	5.988	7.4643	-0.07	0.94093 ns
*-200*gender M: listener other*	3.8497	1.945	2896.948	1.97	0.04790 *
*End*gender M: listener other*	8.0387	1.945	2896.948	4.13	3.69e-05 ***

3.3 Segmental duration

Segmental duration is significantly influenced by the distance from the PTL and the different sequential structures that follow the PTL. Duration increases over the three test locations up until the end of the utterance for PTLs preceding speaker change cases. Average final segmental duration is higher also for turn holds, but it remains closer to the average of the previous test location (-200). This indicates that speakers do slow down their speech when approaching a PTL before a keep, but not as much as they do before a speaker change (see Fig. 4).

Figure 4 - *Average segmental duration in ms for speaker change and keep cases over the three test locations ("Distance": -500 ms, -200 ms, End), for female speakers (gender: F) in same (listener: same) and mixed-gender dialogues (listener: other), and for male speakers (gender: M) in same (listener: same) and mixed-gender dialogues (listener: other)*



For speaker change cases, final segmental duration increases to a higher degree for male speakers in MM conversations than for female speakers in FF conversations. In MF dyads, on the contrary, duration at the final test location increases for female speakers, while its average remains closer to the previous locations for male speakers (see Fig. 4).

A linear mixed effects model shows a three-way interaction between the gender of the speaker, the gender of the interlocutor and the duration values at the final test location (Tab. 7).

Table 7 - *Summary of the linear mixed model for segmental duration in speaker change cases. A significant three-way interaction is found between the value of the final test location, the gender of the speaker and the gender of the interlocutor. The speaker is included as a random factor in the model. Formula: lmer(duration~ distance * gender * listener + (1 | speaker))*

	Estimate	*Std. Error*	*DF*	*t-value*	*Pr(>|t|)*
(Intercept)	7.724e-02	6.451e-03	9.551e+00	11.974	4.57e-07 ***
-200: Gender M	1.480e-02	5.137e-03	3.517e+03	2.881	0.00399 **
End: Gender M	4.744e-02	5.137e-03	3.517e+03	9.235	< 2e-16 ***
-200 listener other	4.184e-03	5.091e-03	3.517e+03	0.822	0.41120 ns
End: listener other	3.020e-02	5.091e-03	3.517e+03	5.932	3.29e-09 ***
Gender M: listener other	-1.568e-02	1.564e-02	9.173e+00	-1.003	0.34177 ns
*200*gender M: listener other*	-1.979e-02	7.629e-03	3.517e+03	-2.594	0.00953 **
*End*gender M: listener other*	-5.805e-02	7.629e-03	3.517e+03	-7.609	3.53e-14 ***

In keep cases, as already mentioned, final duration does not increase as much as in speaker change instances. However, this does not seem to be true for male speakers in MF dialogues. In fact, they seem to reach the potential turn end with an increased segmental duration also before turn holds (see Fig. 4). Even if it is possible to detect qualitative differences, in this case, they did not reach statistical significance.

3.4 Summary of the results

The phonetic cues preceding speaker changes and those preceding keep cases show different patterns of variation. Intensity tends to decrease towards the end of the utterance, but to a lesser extent when the current speaker wants to continue talking. Segmental duration increases towards the end, but to a lesser extent, again, when the current speaker has the intention of continuing. F0 values show a high degree of variation, so it is harder to get a clear pattern of variation related to the turn-taking structure. As for the sociophonetic variation of turn-taking cues, there are significant differences in the way male and female speakers in same-gender and mixed-gender conversations mark the test locations for both keep and speaker change cases. In particular, the results obtained using linear mixed models, indicate that, in speaker changes, the interaction between the gender of the speaker, the gender of the interlocutor and the final datapoint, i.e., the end of the utterance, was significant for F0 values (est.: -7.1274; t value: -5.511; $p < .0001$) for intensity (est.: 8.8455; t value: 6.942; $p < .0001$) and segmental duration (est.: -0.05805; t value: -7.609; $p < .0001$). For segmental duration, the interaction of the speaker's and the interlocutor's genders with the "-200" location was significant as well (est.: -0.01979; t value: -2.594; $p < .01$). For keep cases, the variation of F0 values is influenced by the interaction of the gender of the interlocutors and the "-200" location (est.: 1.3532; t value: 1.389; $p < .0001$), and, for intensity, by the gender of the interlocutors and the "-200" (est.: 3.8497; t value: 1.979; $p < .05$) and "End" (est.: 8.0387; t value: 4.133; $p < .0001$) locations.

4. *Discussion*

The results related to the phonetic variation of turn-taking cues in this dataset show patterns that had already been observed in previous research, i.e., speakers tend to maintain their intensity higher (Gravano, Hirschberg, 2009; 2011) and their speech rate faster (Local et al., 1986; Niebuhr et al., 2013) in order to hold the floor. This configuration of the phonetic turn-taking cues approaching PTLs is in line with the idea that speakers might put more effort into marking turn holds than into signaling turn-yielding intentions because speaker change might be "a kind of default option" (Zellers, 2016: 13) in spontaneous conversation. This might be especially relevant when the speaker's utterance is syntactically/semantically complete, as was the case for the utterances considered in this study. In fact, in the presence of syntactic/semantic completion, it might be more likely that the interlocutor will assume that the current speaker has finished with their turn, so

a more prominent phonetic marking is needed. Patterns for F0 variation related to turn-taking are less clear. In fact, no recurrent or systematic patterns emerged from the qualitative analysis of the data. This observation is in line with the results obtained by Feindt et al. (2021) who suggest that, since F0 movements in German are not restricted by any phonological function, as it is the case for other languages, e.g., Swedish, speakers in conversation tend to use F0 more freely towards turn ends and make use of their entire F0 range. Moreover, the fact that only syntactically/semantically complete utterances were taken into consideration for the present study gives a further explanation to the high degree of variation obtained from this data sample: in fact, it has been observed that when the upcoming completion of an utterances is signaled by syntactic, semantic or pragmatic means, F0 patterns are less restricted for turn-taking function, and speakers tend to show a higher degree of variability towards the end of their turn (Selting, 1996; Feindt et al., 2021; Rossi et al., 2022a; Rossi, Feindt & Zellers, 2022b).

As for gender variation, a few differences were observed in the phonetic patterns of the parameters analyzed, which were seemingly influenced not only by the gender of the speaker, but also by the gender of the interlocutor. First of all, the datapoints extracted for F0 (in semitones and normalized with the speakers' baseline to exclude the influence of physiological factors) were always higher for female speakers and closer to the baseline for male speakers, a result consistent with other studies on German and other languages (Weirich, Simpson, 2019; Andreeva et al., 2014; Pépiot, 2014; Pépiot, Arnold, 2021). The difference in F0 height was particularly striking in MF conversations, where F0 was closer to the baseline for male speakers than in MM conversations, while for female subjects it was higher in MF than in FF conversations. Moreover, more rises and falls in F0 were observed for male speakers than for female speakers and, in a few occasions, F0 even showed a more consistent patterning for them, as for example in the keep cases in MF dyads (see Fig. 2).

Variation related to turn-taking for intensity and segmental duration, instead, seemed to be employed more consistently by female speakers in both FF and MF conversations, and by male speakers in MF conversations. Interestingly, in fact, male speakers in MF dyads seemed to use these values to a smaller extent to signal their conversational intentions, e.g., they tend to not lower their final intensity as much as it happens in other conditions (e.g., in MM dyads), as well as do not vary their speech rate in the direction it would be expected, or as much as their female conversational partners. On the contrary, male speakers make use of these cues in line with the expectations in MM dyads and, in these cases, their variation related to turn-taking is much more striking than it is for female speakers in FF dyads. Thus, in general, it appears from these observations that female speakers make more use of duration and intensity as turn-taking cues, while it is possible that male speakers may vary more with F0 towards turn ends to signal their intentions. This observation would be in line with the results obtained by Whiteside (1995; 1996) for British English. She finds that, in a reading task, women mark syntactic boundaries by pausing and using phrase-final lengthening, while male subjects

rely more on F0 shifts and pause less frequently, which, in a conversational setting, would make it harder for the interactant to take up the next turn (Whiteside, 1995). Moreover, their more consistent use of duration and intensity in MF dialogues may suggest that female speakers in this data sample are doing more work in terms of signaling their conversational intentions than male speakers do. However, similar patterns of variation, even more accentuated, are observed for male speakers in MM conversations. So, in general, it appears that all subjects put more effort in terms of signaling their intentions to either cede or hold the floor when the other interlocutor was a male.

As previously mentioned (see §1.1), the interactional setting should be taken into consideration when discussing gender influence on conversational behavior. Previous studies on talkativeness and assertiveness and their relation to gender identity reported that, in mixed-gender dyads, gender identity was salient for the participants, since the effects of controlled factors were significantly bigger than in same-gender ones (Leaper, Ayres, 2007; Hannah, Murachver, 2007). Applying this to our results may suggest that female speakers in mixed conversations, where gender might have been more salient, tended to take up a more cooperative and facilitative role (Holmes, 1995), focused on keeping the flow of the conversations as smooth as possible by making their turn-taking intentions clear. However, the structure of the exchanges of same-gender dyads suggest some influence of the gender of the two interlocutors, too. For example, a greater number of overlaps were found in female-to-female speech, while a lower number of turn yielding cases accompanied by a higher number of holds was observed for male-to-male dyads, which is similar to the descriptions of same-sex conversations provided by the literature (see §1.1). In fact, it has also been indicated how also experimental activities seem to interact differently with gender, with task-oriented dialogues making interlocutors' genders more relevant than non-structured conversations (Leaper, Robnett, 2011; Leaper, Ayres, 2007). Finally, in this context, the result of male speakers doing more signaling work than female speakers in MM conversation is interesting and unexpected. A possible explanation could be that the goal of the conversation, i.e., discuss with each other to complete a cooperative task, was salient and thus the interactants might have focused on that, and tried to cooperate as much as possible with each other by facilitating turn-taking and putting more effort into signaling their conversational intentions.

5. *Conclusions and outlook*

From this first exploration of the influence of gender on the variation of turn-taking cues, it appears that male and female speakers in this data sample share both commonalities and differences. All the speakers display the characteristics of potential turn ends already investigated for German, with certain patterns of variation of F0, intensity and segmental duration before turn holds and speaker changes. A few differences were observed, however, in the degrees of variation of such cues, especially

when the gender of the interlocutor was taken into consideration as well. Even if quantitatively significant, it must be kept in mind that these results concern a relatively small set of data and speaker sample. Moreover, since it seems that no previous study has tested how social and individual variables influence the variation of phonetic turn-taking cues, it is not possible to generalize or make broader claims on the issue. However, these results provide some preliminary evidence for a possible influence of gender on the way male and female speakers mark PTLs in spontaneous interactions in German. Thus, it can be a matter worth of being investigated further, since it might give us some more insight into what contributes to produce the different structures that same-gender and mixed interactions sometimes seem to assume.

Further developments of this research will focus on a larger set of data and will include more phonetic parameters in the analysis, such as voice quality and segmental reduction measures, in order to observe their variation towards PTLs in German and to test the influence of the interlocutors' genders on their use. Then, a better tailored annotation and analysis will focus on backchannels, both vocal and gestural (in the form of head movements) and on the phonetic marking of the speech preceding them. In line with the current one, the investigation of backchannels in German will deal with their form, their distribution in the conversation, and the extent to which the gender of both interlocutors might have an effect on backchannel production and the phonetic cues preceding it.

Acknowledgements

This research was supported by the project "Sound Patterns and Linguistic Structures at the Transition Space in Conversation" funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Council; project no. 444631148). Many thanks to Margaret Zellers and Kathrin Feindt for their valuable suggestions and discussions on the methods and the analysis of the data, to Margareta Ehlers-Karmanova for her work on the segmental annotations, and to the reviewers for their insightful comments and suggestions.

References

Andreeva, B., Demenko, G., Wolska, M., Möbius, B., Zimmerer, F., Jügler, J., Jastrzebska, M. & Trouvain, J. (2017). Comparison of pitch range and pitch variation in Slavic and Germanic languages. *Proceedings of Speech Prosody*, Dublin, Ireland, 20-23 May 2014. https://doi.org/10.21437/speechprosody.2014-143

Boersma, P. (2001). Praat, a system for doing phonetics by computer. In *Glot International* 5:9/10, 341-345.

Boersma, P., Weenink, D. (2022). Praat: doing phonetics by computer [Computer program]. Version 6.2.14. https://www.praat.org/

Brusco, P., Vidal, J., Beňuš, S. & Gravano, A. (2020). A cross-linguistic analysis of the temporal dynamics of turn-taking cues using machine learning as a descriptive tool. In *Speech Communication*, 125, 24-40. https://doi.org/10.1016/j.specom.2020.09.004

Coates, J. (2004). *Women, men and language: a sociolinguistic account of gender differences in language.* Harlow: Pearson Longman.

Dombrowski, E., Niebuhr, O. (2005). Acoustic patterns and communicative functions of phrase-final rises in German: activating and restricting contours. In *Phonetica*, 62, 176-195. https://doi.org/10.1159/000090097

Edelsky, C. (1993). Who's got the floor? In: Tannen, D. (Ed.) *Gender and Conversational Interaction*, 189–227. Oxford: Oxford University Press.

Edlund, J., Beskow, J. (2007). Pushy versus meek-using avatars to influence turn-taking behavior. *Proceedings of Interspeech,* Antwerp, Belgium, 27-31 August 2007. https://doi.org/10.21437/Interspeech.2007-289

Edlund, J., Beskow, J. (2009). Mushypeek: A framework for online investigation of audiovisual dialogue phenomena. In *Language and Speech*, 52(2-3), 351-367.

Feindt, K., Rossi, M. & Zellers, M. (2021) The time course of pitch variation towards possible places of speaker transition in German and Swedish. *Proceedings of Tone and Intonation*, Sonderburg, Denmark, 6-9 December 2021. https://doi.org/10.21437/tai.2021-23

Ferré, G., Renaudier, S. (2017). Unimodal and bimodal backchannels in conversational English. *Proceedings of SEMDial*, Saarbrücken, Germany, 15-17 August 2017. https://doi.org/10.21437/SEMDIAL.2017-3

Ford, C., Thompson, S. (1996). Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In E. Ochs, E. Schegloff, S.A., Thompson (Eds.), *Interaction and grammar,* 134–184. Cambridge: Cambridge University Press.

Gelfer, M.P., Young, S.R. (1997). Comparisons of intensity measures and their stability in male and female sneakers. *Journal of Voice*, 11(2), 178-186.

Gravano, A., Hirschberg, J. (2009). Turn-yielding cues in task-oriented dialogue. *Proceedings of SIGDial,* London, UK, 11-12 September 2009. https://doi.org/10.3115/1708376.1708412

Gravano, A., Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech and Language*, 25, 601–634. https://doi.org/10.1016/j.csl.2010.10.003

Griffin, Z.M., Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, 4, 274–279. https://doi.org/10.1111/1467-9280.00255

Hadar, U., Steiner, T.J., Grant, E.C. & Rose, F.C. (1984). The timing of shifts of head postures during conversations. *Human Movement Science*, 3, 237–245.

Hannah, A., Murachver, T. (2007). Gender preferential responses to speech. *Journal of Language and Social Psychology,* 26(3), 274-290. https://doi.org/10.1177/0261927X06303457

Heldner, M. (2011). Detection thresholds for gaps, overlaps, and no-gap-no-overlaps. *Journal of the Acoustical Society of America*, 130(1), 508–513. https://doi.org/10.1121/1.3598457

Heldner, M., Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics,* 38, 555–568. https://doi.org/10.1016/j.wocn.2010.08.002

Hough, J., Tian, Y., De Ruiter, L., Betz, S., Kousidis, S., Schlangen, D. & Ginzburg, J. (2016). DUEL: A Multilingual Multimodal Dialogue Corpus for Disfluency, Exclamations and Laughter. *Proceedings of Language Resources and Evaluation Conference,* Portoroz, Slovenia, 23-28 May 2016.

Holmes, J. (1995). *Women, Men and Language.* London: Longman.

Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, 53, 23–35. https://doi.org/10.1016/j.specom.2010.08.003

Indefrey, P. (2011). The spatial and temporal signatures of word production components: a critical update. *Frontiers in Psychology,* 2, 255. https://doi.org/10.3389/fpsyg.2011.00255

Indefrey, P., Levelt, W.J. (2004). The spatial and temporal signatures of word production components. *Cognition*, 92(1), 101-144. https://doi.org/10.1016/j.cognition.2002.06.001

Ishimoto, Y., Teraoka, T. & Enomoto, M. (2017). End-of-Utterance Prediction by Prosodic Features and Phrase-Dependency Structure in Spontaneous Japanese Speech. *Proceedings of Interspeech*, Stockholm, Sweden, 20-24 August 2017. https://doi.org/10.21437/Interspeech.2017-837

Kohler, k. J. (1983). Prosodic boundary signals in German. In *Phonetica*, 40, 89–134.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A. & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. In *Language and Speech*, 41, 295–321.

Kuznetsova, A., Brockhoff, P.B. & Christensen, R.H.B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13), 1–26. https://doi.org/10.18637/JSS.V082.I13

Leaper, C., Ayres, M.M. (2007). A meta-analytic review of gender variations in adults' language use: Talkativeness, affiliative speech, and assertive speech. In *Personality and Social Psychology Review*, 11(4), 328-363. https://doi.org/10.1177/1088868307302221

Leaper, C., Robnett, R.D. (2011). Women are more likely than men to use tentative language, aren't they? A meta-analysis testing for gender differences and moderators. In *Psychology of Women Quarterly*, 35(1), 129-142. https://doi.org/10.1177/0361684310392728

Levinson, S.C. (2013). Action formation and ascription. In Sidnell, J. & Stivers, T. (Eds.) *The Handbook of Conversation Analysis*, 103-130. Oxford: Wiley-Blackwell. https://doi.org/10.1002/9781118325001.CH6

Levinson, S.C., Torreira, F. (2015) Timing in turn-taking and its implications for processing models of language. In *Frontiers in Psychology*, 6. https://doi.org/10.3389/fpsyg.2015.00731

Local, J.K., Kelly, J. & Wells, W.H. (1986). Towards a phonology of conversation: turn-taking in Tyneside English. In *Journal of Linguistics*, 22(2), 411-437.

Local, J., Walker, G. (2012). How phonetic features project more talk. In *Journal of the International Phonetic Association*, 42, 255–280. https://doi.org/10.1017/S0025100312000187

Ludusan, B., Dupoux, E. (2015). A multilingual study on intensity as a cue for marking prosodic boundaries. *Proceedings of the International Congress of Phonetic Sciences*, Glasgow, UK, 10-14 August 2015.

Meteer, M.W., Taylor, A.A., MacIntyre, R. & Iyer, R. (1995). *Disfluency annotation stylebook for the switchboard corpus.* University of Pennsylvania.

Mehl, M.R., Vazire, S., Ramírez-Esparza, N., Slatcher, R.B. & Pennebaker,

J.W. (2007). Are women really more talkative than man? In *Science*, 317, 82. https://doi.org/10.1126/science.1139940

Menz, F., Al-Roubaie, A. (2008). Interruptions, status and gender in medical interviews: The harder you brake, the longer it takes. In *Discourse & Society*, 19(5), 645-666. https://doi.org/10.1177/0957926508092247

Mondada, L. (2007). Multimodal resources for turn-taking: pointing and the emergence of possible next speakers. In *Discourse Studies*, 9(2), 194-225.

Niebuhr, O., Görs, K. & Graupe, E. (2013). Speech reduction, intensity, and F0 shape are cues to turn-taking. *Proceedings of SIGDial*, Metz, France, 22-14 August 2013.

Paggio, P., Navarretta, C. (2013). Head movements, facial expressions and feedback in conversations: empirical evidence from Danish multimodal data. In J*ournal of Multimodal User Interfaces*, 7, 29–37. https://doi.org/10.1007/s12193-012-0105-9

Pépiot, E. (2014) Male and female speech: a study of mean f0, f0 range, phonation type and speech rate in Parisian French and American English speakers. In *Proceedings of Speech Prosody*, Dublin, Ireland, 20-23 May, 2014. https://doi.org/10.21437/speechprosody.2014-48

Pépiot, E., Arnold, A. (2021). Cross-Gender Differences in English/French Bilingual Speakers: A Multiparametric Study. In *Perceptual and Motor Skills*, 128(1), 153–177. https://doi.org/10.1177/0031512520973514

Peters, B. (2006). *Form und Funktion prosodischer Grenzen im Gespräch*. Phd Dissertation. Christian-Albrechts-Universität zu Kiel, Germany.

Plug, I., Stommel, W., Lucassen, P.L., Dulmen, S.V. & Das, E. (2021). Do women and men use language differently in spoken face-to-face interaction? A scoping review. In *Review of Communication Research*, 9, 43-79. https://doi.org/10.12840/issn.2255-4165.026

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. http://www.rstudio.com/

Rossi, M., Feindt, K. & Zellers, M. (2022a). Individual variation in F0 marking of turn-taking in natural conversation in German and Swedish. *Proceedings of Speech Prosody* 2022, Lisbon, Portugal, 23-26 May 2022. https://doi.org/10.21437/speechprosody.2022-38

Rossi, M., Feindt, K. & Zellers, M. (2022b). Perception of F0 movements towards potential turn boundaries in German and Swedish conversation: background and methods for an eye-tracking study. *Proceedings of FONETIK*, Stockholm, Sweden, 13-15 June 2022.

Sacks, H., Schegloff, E., Jefferson, G. (1974). A simplest systematics for the organization of turn-taking in conversation. In *Language* 50, 696–735.

Schiffrin, D. (1987). *Discourse markers.* Cambridge: Cambridge University Press.

Schnur, T.T., Costa, A., Caramazza, A. (2006). Planning at the phonological level during sentence production. In *Journal of Psycholinguistic Research,* 35, 189–213. https://doi.org/10.1007/S10936-005-9011-6

Skantze, D., Schlangen, D. (2009). Incremental dialogue processing in a micro-domain. *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics,* Athens, Greece, 30-31 March / 1-3 April 2009. https://doi.org/10.3115/1609067.1609150

Stivers, T., Enfield, N.J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J.P., Yoon, K. & Levinson, S.C. (2009). Universals and cultural variation in turn-taking in conversation. In *Proceedings of the National Academy of Science*s, 106, 10587–10592. https://doi.org/10.1073/pnas.0903616106

Stubbe, M. (2013). Active listening in conversation: gender and the use of verbal feedback. In S. Yamazaki and R. Sigley (Eds). *Linguistic Insights – Studies in Language and Communication: Approaching language variation through corpora: A Festschrift in honor of Toshio Saito,* 365-416. Bern: Peter Lang.

Tannen, D. (1994). *Gender and discourse.* New York: Oxford University Press.

Tannen, D. (1998). The relativity of linguistic strategies: Rethinking power and solidarity in gender and dominance. In D. Cameron (Ed.) *The Feminist critique of language: A reader.* 261–279. London: Routledge.

Truong, K.P., Poppe, R. de Kok, I. & Heylen, D. (2011). A Multimodal Analysis of Vocal and Visual Backchannels in Spontaneous Dialogs. *Proceedings of Interspeech*, Florence, Italy, 27-31 August 2011. https://doi.org/10.21437/Interspeech.2011-744

Weirich, M., Simpson, A. (2019). Effects of gender, parental role, and time on infant-and adult-directed read and spontaneous speech. In *Journal of Speech, Language, and Hearing Research, 62(11)*, 4001-4014. https://doi.org/10.1044/2019_JSLHR-S-19-0047

Yngve, V.H. (1970). On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*. Chicago: Chicago Linguistic Society.

Ward, N.G., Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. In *Journal of Pragmatics*, 32, 1177-1207. https://doi.org/10.1016/S0378-2166%2899%2900109-5

Waara, E., Shaw, P. (2006). Male and Female Witnesses' Speech in Swedish Criminal Trials. In *Journal of Language and Communication in Business*, (36), 129-156. https://doi.org/10.7146/HJLCB.V19I36.25842

Whiteside, S.P. (1995). Temporal-based speaker sex differences in read speech: A sociophonetic approach. *Proceedings of the International Congress of Phonetic Sciences*, Stockholm, Sweden, 14-19 August 1995.

Whiteside, S.P. (1996). Temporal-based acoustic-phonetic patterns in read speech: Some evidence for speaker sex differences. In *Journal of the International Phonetic Association*, 26(1): 23–40.

Zellers, M. (2016). Prosodic variation and segmental reduction and their roles in cuing turn transition in Swedish. In *Language and Speech* 60(3), 454-478. https://doi.org/10.1177/0023830916658680

Zellers, M. (2021). An overview of forms, functions and configurations of backchannels in Ruruuli/Lunyala. In *Journal of Pragmatics* 175(2021), 38-52. https://doi.org/10.1016/j.pragma.2021.01.012

Zellers, M., Gorisch, J., House, D. & Peters, B. (2019). Hand gestures and pitch contours and their distribution at possible speaker change locations: a first investigation. *Proceedings of Gesture and Speech in Interaction*, Paderborn, Germany, 11-13 September 2019.

Zellers, M., Schweitzer, A. (2017). An investigation of pitch matching across adjacent turns in a corpus of spontaneous German. *Proceedings of Interspeech,* Stockholm, Sweden, 20-24 August 2017. https://doi.org/10.21437/Interspeech.2017-811

Zimmerman, D., West, C. (1975). Sex roles, interruptions and silences in conversation. In Thorne, B., Henley, N. (Eds) *Language and Sex: Difference and Dominance*, 105–29. Rowley: Newbury House.

DOI: 10.17469/O2109AISV000008

SIMONA SBRANNA, SIMON WEHRLE, MARTINE GRICE

The use of Backchannels and other Very Short Utterances by Italian Learners of German[1]

Backchannels (BCs) positively contribute to fluency in social interactions. However, their realisation is language-specific, which can cause miscommunication in intercultural contexts. Nevertheless, backchanneling is not formally taught in most classroom settings. To find out whether L2 learners still manage to acquire a target-like BC behaviour, we carried out an exploratory study on Italian learners of L2 German. We recorded Map Task dialogues performed by 6 dyads speaking L1 Italian and L2 German at different proficiency levels and 5 dyads speaking L1 German. We extracted BCs, defined as acknowledgment tokens, and other very short utterances (VSUs) with the same lexical realisation as BCs, but different functions. We analysed their frequency, length and lexical type according to their function. Preliminary results suggest that dyad-specific patterns play a larger role than L2 proficiency when predicting BC frequency and length. As for lexical choice of BC types, L2 learners prefer items shared with their L1 Italian. Specifically German types are only used by advanced learners, indicating a role of proficiency in this aspect of BC production.

Keywords: backchannels, very short utterances, L2 acquisition, individual variability, communicative competence.

1. *Introduction*

One issue in second language acquisition (SLA) research has been the question of how to assess communicative competence in a quantitative and homogeneous way, while taking into account idiosyncratic and contextual factors impacting the L2 learning process as well as L2 oral performance. Fluency has been widely recognised as a central aspect in the assessment of L2 oral performance (De Jong, 2016) and is listed under the abilities for oral interaction in the official European guidelines of language competence, the Common European Framework of Reference for Languages (Council of Europe, 2001; Figueras, North, Takala, Van Avermaet & Verhelst, 2009). However, fluency is a complex phenomenon, as further demonstrated by the lack of a clear-cut definition (e.g. Lennon, 1990, Lennon, 2000, Wood, 2001; Wolf, 2008).

[1] The authors covered different roles in the realisation of this article. Following the CRediT authorship contribution statement, Simona Sbranna contributed with conceptualisation, data curation, visualisation, writing – original draft; Simon Wehrle contributed with conceptualisation, methodology, visualisation, writing – review and editing; Martine Grice contributed with conceptualisation, methodology, supervision.

Most studies on L2 fluency have focussed on individual measures of fluency, such as speech and articulation rate; amount, location and duration of pauses; or repairs and repetitions (for a comprehensive list, see Saito, Ilkan, Magne, Tran & Suzuki, 2018). However, the majority of our real-life oral performances are interactions and much less often monologues. For this reason, 1) training and testing learners in monologic settings does not sufficiently help them to develop L2 fluency and 2) using individual measures in L2 fluency research risks providing only a partial picture of learners' oral competence, ignoring many other contributing factors.

Apart from idiosyncratic and contextual factors, such as a speaker's status, L2 proficiency level, L1 background as well as topic and setting (He, Young, 1998), fluency in dialogue is determined to a great extent by the specific interactional mechanisms that influence a conversation between two interlocutors. Therefore, unique dyad-related factors play a fundamental role, along with strictly individual factors. It is thus clear that the contribution of both interlocutors to the interaction, and the possible accommodation to the conversation partner's speech style, cannot be ignored when studying fluency in a conversation (for an extensive discussion on this topic see Sbranna, Cangemi & Grice, 2020). The specificity of the cooperation between parties shaping a conversation together has been described by Jacoby and Ochs (1995), who view interaction as a form of co-construction and a joint creation of discourse between interlocutors; by Hall (1993, 1995 as reported in He, Young, 1998), who argues that interactional competence emerges in varied interactive practices to which participants contribute with the appropriate linguistic and pragmatic resources; and by McCarthy (2009), who defines interactional flow as a jointly achieved harmonisation of tempo.

The smoothness of a conversation is achieved, among other factors, through the rhythm of turn-taking (Sacks, Schegloff, & Jefferson, 1974). In particular, smooth or disfluent turn transitions can take place at turn-boundaries (also called transition relevance places–TRP–in conversational analysis), and interlocutors have to appropriately foresee the end of the other party's turn and react to it quickly and accordingly (Levinson, 2015; Bögels, Torreira, 2015).

Despite usually going unnoticed in conversation (Shelley, Gonzalez, 2013), one important linguistic means that can facilitate the flow of conversation is the use of so-called backchannels. Backchannels are very short lexical and non-lexical utterances, like 'okay' or 'mm-hm', which have traditionally been described as non-intrusive tokens—that is, as not claiming a floor transfer—used to signal the listeners' active engagement by showing acknowledgement and understanding (Yngve, 1970, Schegloff, 1982). By supporting the ongoing turn of the interlocutor, backchannels positively contribute to fluency in social interactions (Amador-Moreno, McCarthy & O'Keeffe, 2013) as they maintain flow and contribute to creating a shared structure in dyadic conversation (Sacks, Schegloff, & Jefferson, 1974; Schegloff, 1982; Kraut, Lewis & Swezy, 1982).

On the other hand, backchannels can be potentially misleading in cross-cultural contexts, where different culturally-shaped communicative conventions come into

contact (Cutrone, 2005, 2014; Ha, Ebner & Grice, 2016; Li 2006; among others). Research has indeed provided evidence for language- or variety-specific backchannel characteristics concerning length, duration, frequency, location, intonation and function, and these can potentially have negative social implications in a communicational setting where the interlocutors' linguistic backgrounds diverge.

For these reasons, the importance of backchannels in L2 learning becomes clear. The CEFR (Figueras et al., 2009) already lists the use of feedback expressions under passive competence at the A2 level, an early stage in the learning process. However, backchannels are not explicitly taught or thematised in most L2 classroom settings and it cannot be taken for granted that learners will acquire appropriate backchannelling behaviour simply from exposure to the target language. Moreover, teachers are not always native speakers, and input on this particular interactional feature might be completely absent from many classroom settings.

Against this background, two possible outputs in the learners' interlanguage can be expected. On the one hand, it is possible to assume that backchannels go unnoticed in conversation, resulting in a transfer of features from the L1 to the L2. On the other hand, backchannels might be perceived by learners as salient features of foreign speech and receive an appropriate level of attention, which would favour an adaptation to target language patterns. In the latter case, more target-like backchannel behaviour should be observed, especially at an advanced level, i.e. with more experience and exposure to the target language.

With these two scenarios in mind, we will explore the use of backchannels in second-language learning. The paper is structured as follows: in section two, we offer a brief overview of the literature about the phenomenon of backchannels, their differences across cultures and languages and their acquisition in L2 learning; in section three, we present the methods used, including information on participants, data collection, corpus and measures; in section four, we present our results; and in section five we conclude the paper by summarising the findings and discussing their implications.

2. *Background research on backchannels*

In the literature, there is little agreement on the definition of backchannels (as noticed by Lennon, 1990, 2000; Rühlemann, 2007; Wolf, 2008 among others), resulting in a variety of names and categorisations that are often imprecise and overlapping.

In his analysis of telephone conversations, Fries (1952) was probably the first to recognise 'signals of attention' that do not interrupt the speaker's talk. Since then, other terms have been used to define this phenomenon, such as 'accompaniment signals' (Kendon, 1967), 'receipt tokens' (Heritage, 1984), 'minimal responses' (Fellegy, 1995), 'reactive tokens' (Clancy, Thompson, Suzuki & Tao, 1996), 'response tokens' (Gardner, 2001), 'engaged listenership' (Lambertz, 2011) and 'active listening responses' (Simon, 2018).

The term 'backchannel communication' was first coined by Yngve (1970) to define the channel of communication used by the listener and recipient to give

useful information to their interlocutor without claiming a turn, in opposition to the main channel used by the speaker holding the floor.

Early research on backchannels was mostly conducted on American English (Fries, 1952; Yngve, 1970; Duncan, 1974; Duncan, Fiske, 1977; Schegloff, 1982; Jefferson, 1984; Goodwin, 1986). These studies aimed at defining the phenomenon and tried to offer a categorisation of backchannel types, generally based on their pragmatic function or formal realisation.

Schegloff (1982) noted that these short utterances were mainly used by the listener not only to acknowledge the interlocutor's turn, but also to invite the primary speaker to carry on with his turn. For this reason, he defined the minimal utterances used in the specific contexts of an ongoing turn by the interlocutor as "continuers". Jefferson (1984) introduced the term "acknowledgement tokens". Indeed, in its narrow use the term backchannel refers to tokens used to signal acknowledgement and understanding of what the interlocutor is saying, while inviting the main speaker to continue (also used in this sense by Beňuš, Gravano & Hirschberg, 2007, Hasegawa, 2014, and others).

In its broader use, the term backchannel has also been matched to numerous other functions, and some attempts at establishing a function-based categorisation have been made. For example, Drummond and Hopper (1993) further distinguish acknowledgement tokens marking 'passive recipiency', as in the case of continuers, from those marking 'incipient speakership', signalling a listener's intention to start a turn of their own. Maynard (1997) categorises backchannels according to the functions of continuer, understanding, agreement, support, strong emotional answer and minor additions. Kjellmer (2009) recognises five functions of backchannels: regulative, supportive, confirmatory, attention-showing and empathetic. Tolins and Fox Tree (2014) distinguish context-generic backchannels, used as continuers and promoting the production of new information, and context-specific backchannels, also called assessments in previous studies (Goodwin, 1986), such as 'really' or 'wow', eliciting further elaboration of what has just been said.

As far as their formal realisation is concerned, backchannels present a high degree of lexical variability, but they can also be realised through vocal noises (Wong, Peters, 2007), visual modalities such as facial expressions, head movements, gestures (Tolins, Fox Tree, 2014) and responsive laughter (Hasegawa, 2014). Some structurally motivated proposals of classification have been advanced to categorise backchannel utterances. Tottie (1991) classifies them into simple, double and complex types. Simple backchannels are composed of one single utterance, e.g. 'yes', double backchannels are repeated simple types, e.g. 'okay okay', and complex backchannels are a combination of different simple types, such as 'okay yes right'. Wong and Peters (2007) differentiate between minimal, lexical and grammatical types. Minimal types are defined as non-lexical items that are semantically empty and items expressing polarity, e.g. 'mmhm', 'yes', and 'no'. Lexical types are considered to be all single words that are codified in dictionaries and show an increase in semantic weight, such as 'really', 'right', and 'good'. Finally, by grammatical types they mean

predications in the form of short codified phrases, such as 'I see', brief questions, repetitions, sentence completions and commentaries.

Later studies shifted their attention towards languages other than English and revealed that there are cross-cultural and cross-linguistic differences in the use of backchannels (e.g. Tao, Thompson, 1991; Tottie, 1991; Berry, 1994; Clancy et al., 1996; Ward, Tsukahara, 2000; Heinz, 2003; Cutrone, 2005, 2014; Li, 2006; Nurjaleka, 2019; Kraaz, Bernaisch, 2022).

2.1 Backchannel use across languages and cultures

Since then, one main question in the field of backchannel research has been variation across languages and cultures, and differences in backchannel use have been identified regarding frequency, duration, location, lexical type, function and intonation.

Being bound to culture, backchannel behaviour has been found to diverge even across varieties of the same language. Tottie (1991) reports differences with regard to frequency and type across American and British English, showing that in American conversations there was an average of sixteen backchannels per minute, compared with just five backchannels per minute in British conversations. Similarly, differences were observed across Sri Lankan and Indian English in type, frequency and function (Kraaz, Bernaisch, 2022).

Some studies report the impact of different backchanneling behaviour on the turn-taking system. For instance, in a cross-linguistic study on Spanish and North-American English, Berry (1994) found that backchannels were more frequent and longer among Spanish speakers, resulting also in longer stretches of overlapping speech. In turn, American English speakers were shown to use more overlapping backchannels than Germans, as reported in a comparative study by Heinz (2003).

These differences lead to the hypothesis of a potentially negative effect on communication in intercultural conversations. In a study on responsive tokens in English, Mandarin and Japanese, Clancy et al. (1996) observed that Japanese people produced the most frequent reactive tokens, placing them in the middle of the interlocutor's speech. Mandarin speakers, in contrast, produced the fewest backchannels and mostly at TRPs, i.e. at the end of interlocutors' turns. American English lay between the two other languages with regard to frequency, and reactive tokens were placed within interlocutors' turns and at TRPs, but preferably at grammatical competition points. The authors speculate that, in Japanese, backchannels are used as a form of emotional support and cooperation, whereas, on the opposite pole, Mandarin speakers might perceive Japanese backchannels as intrusive in comparison to their tendency to not interrupt the other speaker out of respect. American English speakers, likewise, might find Japanese speakers disruptive while the scarce reactions of Mandarin speakers might leave them wondering what their listeners are thinking (Clancy et al., 1996: 383). Similar hypotheses were tested in a study on backchannel intonation, in which Ha et al. (2016) found differences across Vietnamese and German. While Vietnamese continuers were consistently level or falling, German equivalents were tendentially rising. Based on the results

of a previous perception experiment (Ha, 2012), the authors hypothesise probable misunderstandings in intercultural dialogues. In Vietnamese, rising pitch as used by Germans might be interpreted as impolite. Conversely, for German natives, the level/falling pitch used by Vietnamese speakers might cause irritation (Stocksmeier, Kopp & Gibbon, 2007) and could be interpreted as showing disinterest or as an attempt to end the interlocutor's turn.

Given the observed differences across languages, the immediate next step in research was to put the consequences of this variation in intercultural conversations to the test and find out whether and to what extent differences in backchannel use can lead to miscommunication and/or have negative social implications. Li (2006) conducted a study on Canadian and Chinese speakers in intra- and intercultural conversations and showed that backchannels facilitated communication among speakers of the same language. However, when Canadian speakers were paired with Chinese speakers, the opposite effect was observed, leading to the claim that backchannel responses can be misleading in intercultural conversations and cause miscommunication. It was also found that the Chinese speakers produced the most backchannels and the Canadians the fewest, but when crossed, speakers tended to produce a number of backchannels in between. In a follow-up study providing an analysis of backchannel types (Li, Cui & Wang, 2010), it was found that, in intercultural conversations, both Canadian and Chinese speakers used other backchannels than in their respective native languages, showing some degree of speech convergence for both frequency and lexical type.

However, accommodation in intercultural conversations does not always take place automatically, as knowledge of language- and culture-specific conventions is likely to be essential. For example, White (1989) reports that Japanese speakers did not adapt their active listening style in conversations with Americans, while Americans did, because "they clearly have the linguistic ability to do so" (White, 1989:74), suggesting that language proficiency might be a prerequisite for accommodation. A high level of L2 proficiency can, indeed, provide the speaker with diverse linguistic means to select verbal tokens according to the respective context, and with the flexibility to recognise and switch among language conventions.

2.2 Backchannel productions by L2 speakers

To date, only relatively few studies have investigated backchannels in L2 speech. The relevant findings reinforce the assumptions made on the basis of intercultural studies in showing that L1 backchannel behaviour is generally carried over to the L2, frequently causing miscommunication and misperceptions.

For example, Cutrone (2005) examined the use of backchannels in dyadic interactions between Japanese EFL (English as a Foreign Language) and British speakers. Differences were found in frequency, type and location, and these negatively affected intercultural communication. The frequent backchannels used by the Japanese participants were interpreted as interruptions by the British speakers, and their interlocutors were perceived as impatient. In a follow-up study, Cutrone

(2014) reports that Japanese EFL speakers used a greater number of backchannels because it helped them to feel comfortable as listeners, showing a behaviour similar to the one reported for L1 Japanese (Clancy et al., 1996).

Wehrle and Grice (2019) also report on the negative effect of transfer on intercultural communication. In a pilot experiment they compared the intonation of backchannels in L2 German spoken by Vietnamese and observed that Vietnamese learners produced twice as many non-lexical backchannels (e.g. 'mmhm') with a flat intonation contour as German native speakers, showing a transfer from their L1. As previously mentioned, and corroborated in Wehrle and Grice (2019) through a mouse-tracking experiment, a flat backchannel contour in German might be interpreted as a signal of disinterest and cause irritation (Ha et al. 2016).

Another study that hypothesises a transfer of backchannel features from the L1 to the L2 was conducted by Castello and Gesuato (2019). They investigated the frequency and lexical types of 'expressions of convergence' in Chinese, Indian and Italian learners of English in a language examination setting. They found that Chinese learners used the most backchannels and Indian learners the least, with Italian learners lying between these two groups. They also observed differences in the choice of backchannel types between groups, suggesting an effect of L1 background.

A similar conclusion is reached by Shelley and Gonzales (2013), who analysed backchannel functions in informal interviews in four ESL (English as a Second Language) speakers with different L1 backgrounds as well as one American native speaker of English. They identify four backchannel functions: continuers (the listener is paying attention and does not hold the floor), acknowledgements (the listener agrees or understands), newsmakers (the listener communicates an emotional reaction) and change of activity (the listener signals to move toward a new topic). They report an effect of the L1 as differences in the preferred backchannel functions across the four speakers were found.

Finally, there are studies showing that higher proficiency in the L2 implies a better ability to use backchannels. Galaczi (2014) compared the frequency of backchannels and expressions of confirmation among learners of English with different proficiency levels. Results show that intermediate learners provided less feedback than highly proficient learners, among which the "ability to act as supportive listeners through backchanneling and confirmations of comprehension was found to be more fully developed" (Galaczi, 2014: 570).

To summarise, previous research on various languages provides converging evidence 1) for miscomprehension and misperception of the interlocutor's intentions due to a use of backchannels that diverges from the native conventions, 2) for a transfer of the L1 backchanneling behaviour to the L2, and 3) for proficiency as a positive factor in the improvement of learners' L2 backchannelling ability.

At the same time, these studies have some limitations. Their results are not easily comparable as they differ considerably in design and methodology: how participants in the dialogue were matched, their status, their proficiency level in the language of the conversation, the setting of the dialogue, the method used for dialogue elicitation

and aspects of backchannels analysed. Moreover, most studies have focussed on subjects with different L1 backgrounds, which is useful for detecting cultural-specific differences among groups of learners, but does not permit differentiation between transfer phenomena and cross-linguistic, speaker-specific characteristics.

Still, these findings have significant implications for the relevance of backchannels in language teaching environments. In order to better understand the mechanism behind cross-cultural backchannel behaviour, it is important to shed light on how the backchannelling ability develops in interlanguages, with the goals of raising awareness in multicultural communicative contexts and improving L2 speakers' interactional skills.

Therefore, in the present paper we try to overcome some of the limitations mentioned and relate the results to language pedagogy. In order to assess transfer phenomena and/or the acquisition of target-like backchannel features, we carried out an exploratory study using a within-subjects design. In particular, we investigate backchannel use across Italian learners' L1 and L2 German and compare their realisation to a German native group. We pay particular attention to dyad-specific behaviour in order to differentiate idiosyncratic factors from actual transfer or the acquisition of patterns. Finally, we take into account several aspects of BCs to offer a more comprehensive view of the phenomenon, i.e. backchannel frequency, length, type and function.

3. *Method*

The definition of the term 'backchannel' varies considerably in previous literature. For the purpose of this exploratory study, we will adopt the term 'very short utterances' (VSUs), proposed by Edlund, Heldner and Pelcé (2010) as a loose definition for the wide variety of interactional dialogue phenomena providing feedback to the interlocutors. According to this definition, backchannels are to be considered as a specific sub-category of VSUs with an acknowledging function. Words such as 'yes', which are used both as BCs and as VSUs with a different function – in our corpus as positive answers to yes-no and tag questions – will also be analysed to perform a comparison. The difference between BCs and positive replies is motivated by the fact that BCs are 'unsolicited', whereas in the case of replies the primary speaker gives up their turn by asking a question. For this reason, we will refer to 'backchannels' and 'acknowledgments' interchangeably, while we will refer to tokens that fulfil a function other than acknowledgment as 'other VSUs'.

3.1 Participants

For exploration purposes, we selected 22 speakers from a larger corpus: 12 speakers of L1 Italian and L2 German at different proficiency levels (6 beginner and 6 advanced), as well as 10 speakers of native German as a control group.

All Italian speakers had grown up in the province of Naples with parents of the same origin, ruling out variation in their L2 resulting from the native linguistic substratum. Learners were studying L2 German either at university level at the faculty of foreign languages and literatures or at the Goethe Institute in Naples. Their proficiency levels

ranged from A2 to C1 and were established on the basis of the language courses they were attending at the time of the recordings, corresponding to the levels described by the CEFR (A1-A2: beginner; B1-B2: intermediate; C1-C2: advanced). For the sake of determining potential effects of proficiency by using two balanced groups, we categorised them into beginners (from A1 to B1 levels) and advanced (from B2 to C2 levels). We acknowledge that this not a very precise way of identifying the proficiency of individual learners, as there is likely to be a high degree of variability with regards to different language skills both within and across language courses. Nevertheless, the classification used here can serve as a valid starting point, especially as we have a stated interest in how the definitions and demarcations outlined in the CEFR relate to language production in naturalistic conversational interactions.

L1 German participants had grown up in North Rhine-Westphalia and were students at the University of Cologne.

3.2 Data collection and Corpus

Recordings were performed using headset microphones (AKG C 544 L) connected through an audio interface (Alesis iO2 Express) to a computer running Praat (Boersma, Weenink 2022). All participants were recorded in pairs, with L2 learners being matched by their proficiency level.

To collect data, we used the Map Task (Anderson et al., 1991; Grice, Savino, 2003 for set up, map layout and instructions), which matches the goal-oriented cooperation task described in the CEFR. For the task, participants sit opposite one another but have no eye contact. They are given two maps showing several landmarks, but only one map has a path drawn on it. The objective is to co-operate so that the participant without the path can reproduce it on their own map with the help of instructions given by the partner. The task is made more difficult as some landmarks are intentionally not identical across the two maps. The participants are not informed of these mismatches, as the purpose is to create situations requiring collaborative problem solving. This task is particularly useful when dealing with a mixed group of learners including beginners, as it can be performed at every proficiency level. Indeed, learners should have acquired the grammar and vocabulary necessary for giving directions at the beginner level, according to the CEFR.

Italian learners were recorded at the Goethe Institute in Naples. Learners first read the game instructions and carried out the task in Italian. Afterwards, before performing the task in their L2, they watched a video with a German native speaker (S.W.) explaining the instructions again in German to help them get into the language mode and reduce L1 bias. German native speakers were recorded at the University of Cologne. They watched the same German language video instructions and then played the game.

The resulting corpus for the 22 selected speakers includes 6 dialogues in L1 Italian (30 minutes in total); 6 dialogues in L2 German, with 3 performed by beginners (39 minutes in total) and 3 by advanced learners (22 minutes in total); and 5 dialogues in L1 German (52 minutes in total). We extracted and analysed a total of 924 VSUs, of which 646 were BCs.

3.3 Procedure

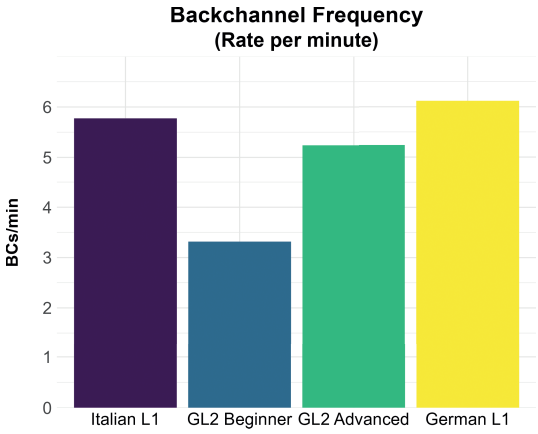
In our analysis, we took into account different aspects of backchannels, i.e. frequency, length, type and function. Specifically, frequency is operationalised as backchannel rate per minute. Length is their duration in milliseconds. Type refers to the lexical and non-lexical realisations. In our corpus, we found that the most frequently used lexical types were 'ja' and 'sì' ('yes' in German and Italian, respectively), 'genau' and 'esatto' ('exactly'), and 'okay'. The most common non-lexical type was 'mmhm'. These token types cover 92% of the whole corpus. We used a category called 'other' for the remaining tokens. Finally, we distinguished three possible functions: acknowledgement, categorised as 'BC', as well as positive replies to tag questions and to yes–no questions, both categorised as 'other VSUs'.

4. *Results*

4.1 Backchannel frequency

Figure 1 shows Backchannel frequency across groups, i.e. the rate of BCs per minute of dialogue. The rate of BCs is very similar across native languages (5.7 BCs per minute for L1 Italian and 6.11 BCs per minute for L1 German). The learners' BC rate is lower than that of both native groups. Beginners produced the fewest BCs, at a rate that was almost half of their output in the target language (3.31 BCs per minute). Advanced learners showed a rate of BCs more similar to the German L1 control group (5.32 BCs per minute). This result might lead to the appealing, but simplistic conclusion that learners acquire a native-like backchanneling behaviour with increasing proficiency. However, this characterisation is incomplete.

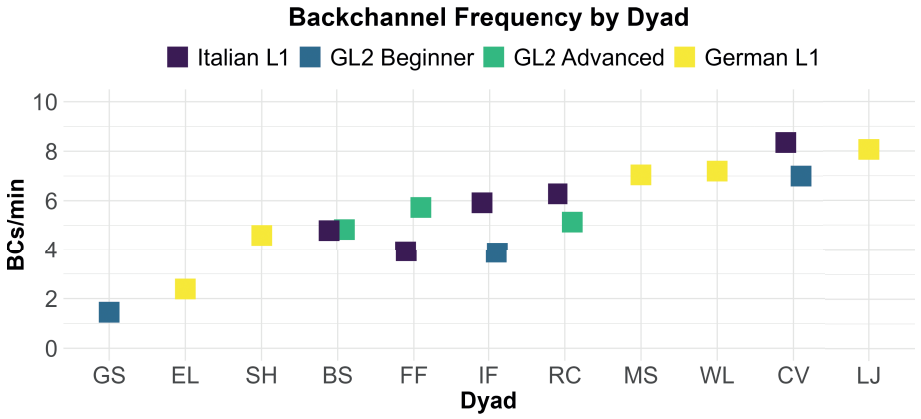
Figure 1 - *Backchannel frequency operationalised as rate per minute of dialogue. The number of BCs per minute is displayed on the y-axis. Language groups are shown on the x-axis and are colour-coded: violet for Italian learners' native speech; blue for beginner learners in L2 German; green for advanced learners in L2 German and yellow for the native German control group*



A closer look at BC rates by dyad [Fig. 2] reveals that dyad-specific behaviour plays a crucial role, in all groups. This becomes obvious when considering the fact that the BC rates across learners' L1 and L2 are strikingly similar. Consider, for example, dyad BS, with almost identical values across the two languages (visible in the overlapping squares). Moreover, the low BC rate in the beginner group as a whole is partly due to the peculiar behaviour of beginner dyad GS, who produced no backchannels whatsoever in their L1 Italian conversation—and only a few other VSUs, which are not displayed in the graph—and also only very few BCs in their L2. This L2 output is most likely not a consequence of low proficiency in German, given that this dyad produced no BCs at all in their L1. Therefore, this peculiarity can be ascribed to dyad-specific behaviour. On the other extreme end, we also observed that another beginner dyad, CV, produced the highest BC rate across all groups, which would not be predicted from group-level results.

A high degree of by-dyad variability can also be seen in the German L1 group. Importantly, one dyad, EL, produced a very similar rate to the beginner dyad GS, showing that a very low BC frequency can also be present among German native speakers, which calls into question the idea of a specific target behaviour to be reached by learners.

Figure 2 - *Backchannel frequency by dyad operationalised as rate per minute of dialogue. The number of BCs per minute is displayed on the y-axis. Dyads are shown on the x-axis and language is colour-coded: violet for Italian learners' native speech, blue for beginner learners in L2 German, green for advanced learners in L2 German, and yellow for the native German control group. Two values are shown for learners, corresponding to speech in their L1 and L2 and distinguished by the colour of the square*



4.2 Backchannel length

Figure 3 shows BC length, i.e. duration in milliseconds (ms). Here, the two native-language groups differ from one another, with L1 Italian speakers producing longer BCs (455 ms) than L1 German speakers (372 ms). A closer inspection of the tokens in the dataset revealed that this difference is mostly due to the fact that Italian

speakers tended to use more complex or repeated BCs, such as 'sì sì okay' or 'okay okay okay'.

Akin to what we already observed for frequency, length values for the two L2 groups seem to suggest an effect of proficiency. There is, indeed, a gradual decrease in BC duration across proficiency levels, from values more similar to the native Italian baseline in the beginner group (416 ms) to values approximating the target in advanced learners (392 ms). This hypothesis would be more appropriate than in the case of BC frequency previously discussed, due to the relatively clear difference between the native and the target languages, which allows an L1 target to be identified. Nevertheless, this claim would again provide an incomplete characterisation in the present case.

Looking at by-dyad values for length, displayed in Figure 4, it appears that dyad-specific behaviour again provides a better explanation than proficiency. Learners' by-dyad values are very similar across their L1 and L2, with one exception, dyad FF, who did show an evident reduction in BC length when speaking in L2 German. However, one side note on this dyad is required. In their case, the long BC duration in L1 Italian is not due to complex or repeated BCs, but to an atypical use of prolonged 'okay' tokens by the instruction follower (only), who did not replicate this behaviour in L2 German.

Figure 3 - *Backchannel length operationalised as duration in ms (on the y-axis). Language groups are shown on the x-axis and are colour coded: violet for Italian learners' native speech, blue for beginner learners in L2 German, green for advanced learners in L2 German, and yellow for the native German control group*

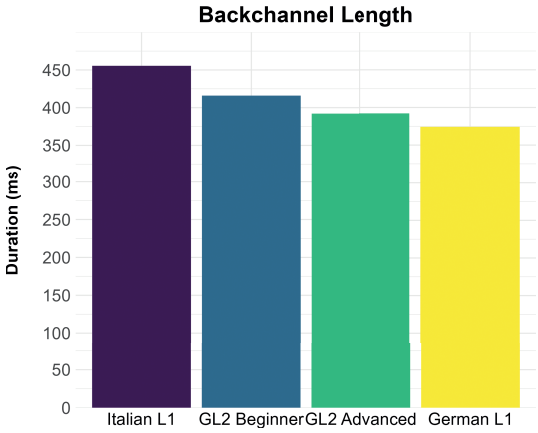
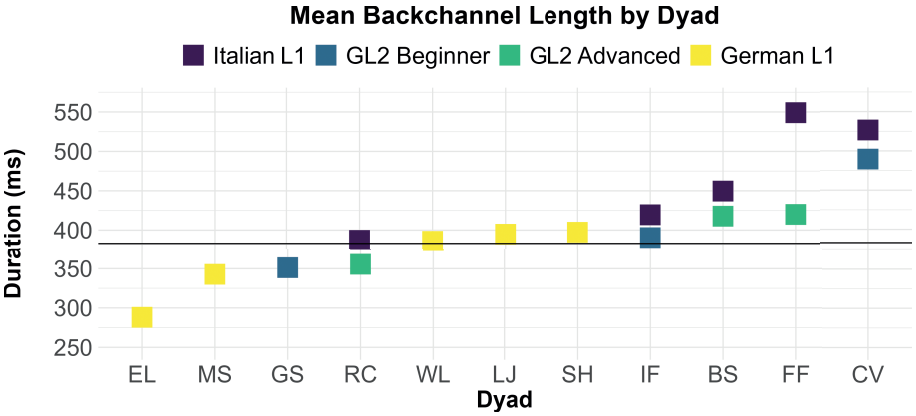


Figure 4 - *Backchannel length by-dyad operationalised as their duration in ms. Mean BC duration in milliseconds is displayed on the y-axis. Dyads are shown on the x-axis and the respective language groups are colour coded: violet for Italian learners' native speech, blue for beginner learners in L2 German, green for advanced learners in L2 German, and yellow for the native German control group. Italian learners of L2 German present two values corresponding to their L1 and L2 speech, distinguished by the colour of the square. The horizontal black line corresponds to the mean BC duration of the L1 German group pooled across all speakers, as a reference for learner productions*

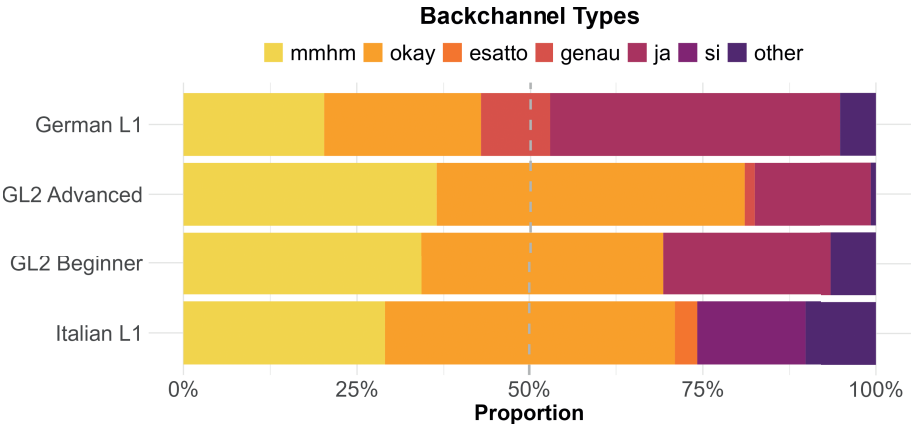


4.3 Backchannel Type

Figure 5 shows the proportions of BC types across groups. Comparing first the two L1s, it can be seen that they diverge regarding the preferred BC type. Both groups show a similar proportion of 'mmhm' (20% in L1 German and 29% in L1 Italian), but in L1 Italian there is a preference of 'okay' (41%) over 'sì' (15%), while the opposite is true in L1 German ('ja' 41%, 'okay' 22%). A type which seems to be typical for L1 German is 'genau' (10%), as the correspondent Italian 'esatto' is not used as much (3%).

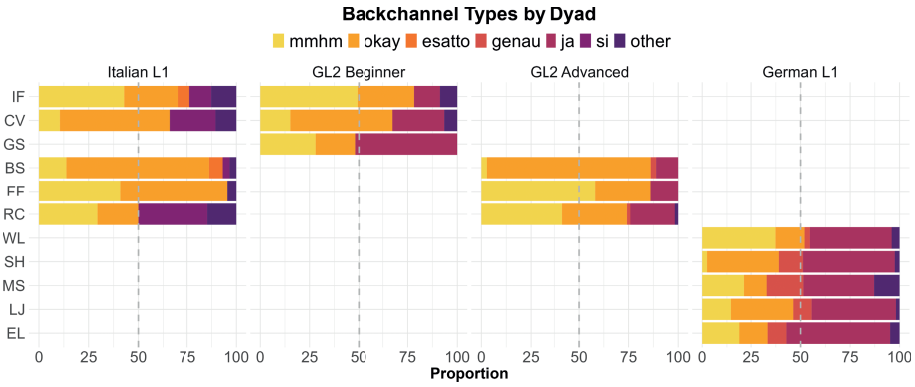
It is evident that L2 learners transfer their choice of BC type from the L1, showing similar proportions of 'mmhm', 'okay' and 'ja' across their L1 and L2 (beginners: 34%, 35% and 24% respectively; advanced: 36%, 44% and 16% respectively). Moreover, the word 'genau' is not used by beginners at all and only represents 1% of BC occurrences in the advanced learner group. This is probably a result of the fact that the Italian equivalent is used very rarely, meaning that learners need more experience and exposure to the L2 to start using this type of BC in the target language.

Figure 5 - *Backchannel Type. Proportions of BC types are shown in percentages on the x-axis. Language groups are shown on the y-axis and are each assigned a bar. The most frequently used BC types are listed in the legend and are colour-coded. The category "other" refers to types that were used only rarely*



The choice of BC type by dyad depicted in Figure 6 reveals highly similar patterns across the L1 and L2 within dyads, especially in the cases of IF, CV and BS, providing support for the transfer hypothesis and showing that dyad-specific patterns in the L1 tend to be reproduced in the L2. One difference between Italian L1 and German L1 concerns the proportions across types in a by-dyad comparison. In detail, it seems that the choice of BC type in L1 German is more consistent across dyads, whereas it appears more variable and dyad-dependent in Italian L1.

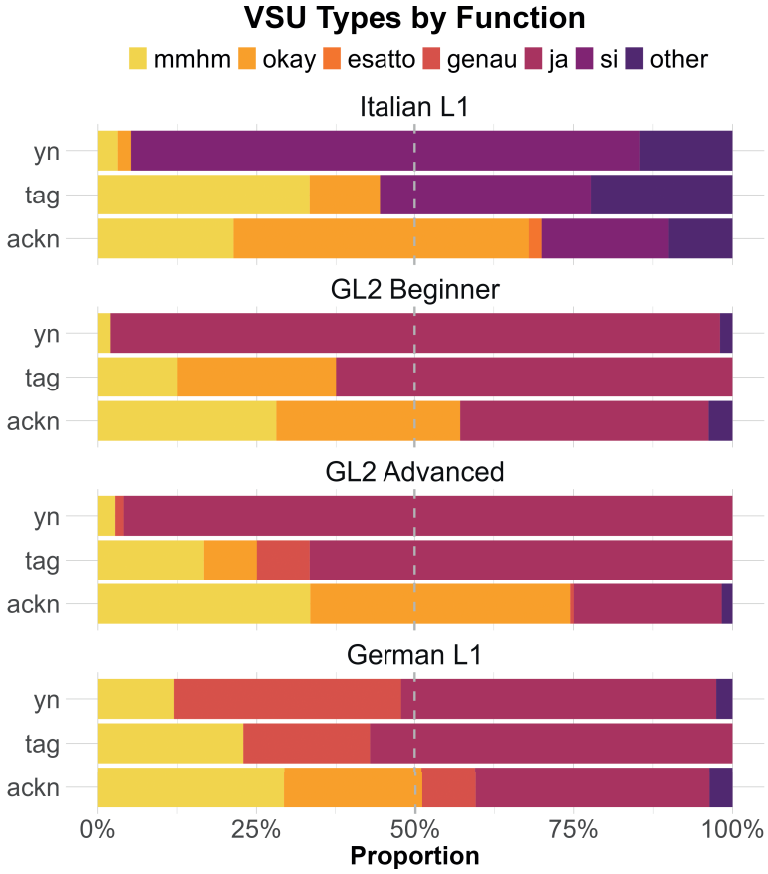
Figure 6 - *Backchannel type by dyad. Proportions of BC types are shown in percentages on the x-axis. Dyads arranged by language group are shown on the y-axis and are assigned a bar each. The most frequently used BC types are listed in the legend and are colour-coded. The category "other" refers to types that were used only rarely*



4.5 BC and VSU by type and function

Figure 5 shows the choice of BCs (acknowledgements) and other VSUs (replies to yes–no and tag questions) by functions and across language groups. The bars representing acknowledgments correspond to those in Fig. 7 and are repeated here to enable a direct comparison.

Figure 7 - *Very Short Utterance types by function for all language groups. Proportions of VSU types are shown in percentages on the x-axis. Functions are shown on the y-axis and are assigned a bar each: replies to yes–no questions, replies to tag questions and acknowledgements (BCs) to compare with. The most frequently used VSU types are listed in the legend and are colour-coded. The category "other" refers to types that were used only rarely.*



The two native languages differ in the proportion of types across the two VSU replies. For yes–no replies, Italians almost exclusively used 'sì' (80%), which is also transferred to the L2 by both beginner and advanced learners (96% respectively), while in L1 German there is more variety with only a few instances of "mmhm" (12%), more predominant 'ja' (50%) and many 'genau' (35%) utterances. For tag replies, Italians seem to prefer 'sì' and 'mmhm' (both 33%), while German

speakers clearly tended to use 'ja' (57%), a pattern which is, somewhat surprisingly, reproduced by L2 learners (beginners: 62%; advanced: 66%). Only advanced learners show some instances of the more typical German 'genau', especially as a response to tag questions (8%); "genau" is mostly used for yes–no and tag replies in the target language (36% yes–no and 20% tag).

Finally, comparing the two VSU categories to the acknowledgement category, it is evident that the choice of lexical type changes across the three functions, suggesting a relationship between type and function.

5. *Conclusion*

In this contribution, we conducted an exploratory analysis of BCs and other VSUs in dyadic interactions in German L2 spoken by Italian learners at two different proficiency levels, and across Italian and German as native languages to compare the learners' output with and assess possible transfer phenomena or the acquisition of target patterns. We took into account frequency, length and lexical type of BCs and of tokens presenting the same lexical types as BCs, but with functions other than acknowledgements, i.e. positive replies to yes-no and tag questions. We paid special attention to dyad-specific variability since individuals' behaviour in a conversation depends not only on idiosyncratic factors and speakers-specific speech style, but, more importantly, on the unique mechanisms that arise from the interaction between the two specific parties in the conversation.

With regards to frequency and length, we observed that dyad-specific behaviour is similar across learners' L1 and L2. Importantly, concentrating on group-level results could have led to the misleading conclusion that target-like patterns of BC frequency and length are achieved in the L2 along with increasing proficiency. A by-dyad analysis suggests, instead, that dyad-specific patterns are more important than proficiency levels when predicting the rate and the length of BCs produced. Moreover, the by-dyad variability found within the group of native German speakers calls into question the idea of target features to be acquired by learners regarding these aspects of BCs.

Differently, for lexical type we observed preferred type-function relations which are language-specific, thus representing a target for learners. In most cases, L2 learners tend to prefer types that are shared with their L1 Italian over specifically German ones, such as 'genau'. The use of specifically German BCs is only present in advanced learners, indicating a positive effect of proficiency for this aspect of BC production.

This first exploration was based on a limited sample of speakers. Therefore, an extension of the analysis to the whole corpus will clarify whether the trends observed are robust, including statistical testing using Bayesian linear regression modelling. However, we set a groundwork from which a few suggestions for further studies can be derived. First, we found that there seem to be preferential co-occurrences among single aspects of BCs, so future studies should address the relation among them. With respect to this point, one further aspect that we did not take into

account and should be addressed in the future is the prosodic realisation of BCs and how it might interact with lexical type and function. Secondly, our preliminary results suggest that dyad-specific patterns are more important than proficiency level when predicting some BC aspects in L2. For this reason, it is important to consider learners' L1 as baseline and investigate dyad-specific behaviour to set apart individual variability from the transfer or acquisition of patterns. Finally, in line with the literature, we showed that there are aspects of BC use which are similar across-languages, but also that some language-specific aspects are not correctly reproduced in L2. Therefore, more comparative studies of varied language pairs can be useful for L2 pedagogy applications.

Acknowledgement

Financial support for carrying out the research is gratefully acknowledged from a.r.t.e.s Graduate School and the German Research Foundation (DFG), grant number 281511265, SFB 1252 Prominence in Language. We thank the Goethe Institute of Naples for making rooms available for the recordings and Prof. Silvia Palermo for recruitment of participants at the University of Naples.

References

Amador-Moreno, C.P., Mccarthy, m., & O'Keeffe, A. (2013). Can English provide a framework for Spanish response tokens?. In *Yearbook of Corpus Linguistics and Pragmatics 2013*. Dordrecht: Springer. 175-201.

Anderson, A.H., Bader, M., Bard, E.G., Boyle, E.H., Doherty, G.M., Garrod, S.C., & Weinert, R. (1991). The HCRC map task corpus. In *Language and speech*, 34(4), 351-366.

Běnuš, Š, Gravano, A., & Hirschberg, J. (2007). The prosody of backchannels in American English. In J. Trouvain & W.J. Barry (Eds.), *16th International Congress of Phonetic Sciences* (pp. 1065–1068). Saarbrücken: Universität des Saarlandes.

Berry, A. (1994). Spanish and American turn-taking styles: A comparative study. In L.F. Bouton, (ed.), *Pragmatics and Language Learning Monograph Series*, Urbana: University of Illinois, 180-190.

Bögels, S., Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. In *Journal of Phonetics*, *52*, 46-57.

Boersma, P., Weenink, D. (2022). PRAAT*:* doing phonetics by computer. [software] Version 6.2. https://www.praat.org/

Castello, E., Gesuato, S. (2019). Holding up one's end of the conversation in spoken English: Lexical backchannels in L2 examination discourse. In *International Journal of Learner Corpus Research,* 5, 231–252.

Clancy, P.M., Thompson, S., Suzuki, R. & Tao, H. (1996) The conversational use of reactive tokens in English, Japanese and Mandarin. In *Journal of Pragmatics,* 26, 355-387. https://doi.org/10.1016/0378-2166(95)00036-4

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge, U.K: Press Syndicate of the University of Cambridge.

Cutrone, P. (2005). A case study examining backchannels in conversations between Japanese–British dyads. In *Multilingua. Journal of Cross-Cultural and Interlanguage Communication*, 24(3), 237-274. https://doi.org/10.1515/mult.2005.24.3.237

Cutrone, P. (2014). A cross-cultural examination of the backchannel behavior of Japanese and Americans: Considerations for Japanese EFL learners. *Intercultural Pragmatics, 11*(1), 83-120. doi: 10.1515/ip-2014-0004

De Jong, N.H. (2016). Fluency in second language assessment. In Tsagari, D., Banerjee, J. (Eds.), *Handbook of second language assessment.* Mouton de Gruyter, 203-218.

Drummond, K., Hopper, R. (1993). Back channels revisited: Acknowledgement tokens and speakership incipiency. In *Research on Language and Social Interaction*, 26(2), 157-177.

Duncan, S. (1974). On the structure of speaker-auditor interaction during speaking turns. In *Language in Society,* 2, 161–180.

Duncan, S., Fiske, D.W. (1977). *Face to face interaction: research, methods and theory*. New Jersey: Lawrence Erlbaum.

Edlund, J., Heldner, M. & Pelcé, A. (2009). Prosodic features of very short utterances in dialogue. In *Nordic Prosody: Proceedings of the Xth Conference*, Helsinki, Finland, August 2008, 57–68.

Figueras, N., North, B., Takala, S., Van Avermaet, P. & Verhelst, N. (2009). *Relating language examinations to the common European framework of reference for languages: learning, teaching, assessment (CEFR): a manual.* Strasbourg, France: Council of Europe, Language policy division.

Fellegy, A.M. (1995). Patterns and functions of minimal response. In *American Speech International Journal of Educational Best Practices*, 2(1), 186-199.

Fries, C.C. (1952). *The structure of English*. London: Longmans, Green and Company.

Galaczi, E.D. (2014). Interactional Competence across Proficiency Levels: How do Learners Manage Interaction in Paired Speaking Tests?, In *Applied Linguistics*, 35(5), 553–574.

Gardner, R. (2001). *When listeners talk: Response tokens and listener stance*. Amsterdam: J. Benjamins Publishing Company.

Goodwin, C. (1986). Audience diversity, participation and interpretation. In *Text & Talk, 6*, 283–316.

Grice, M., Savino, M. (2003). Map tasks in Italian: Asking questions about given, accessible and new information. In *Catalan journal of linguistics*, *2*, 153-180.

Ha, K.P. (2012). *Prosody in Vietnamese – Intonational Form and Function of Short Utterances in Conversation*. Asia-Pacific Linguistics 002 (SEAMLES 001). PhD thesis. Canberra: The Australian National University.

Ha, K.P., Ebner, S. & Grice, M. (2016). Speech prosody and possible misunderstandings in intercultural talk – A study of listener behaviour in Vietnamese and German dialogues. *In Proceedings of Speech Prosody 8*, Boston, USA, 31 May-3 Jun 2016, 801-805.

Hall, J.K. (1993). The Role of Oral Practices in the Accomplishment of Our Everyday Lives: The Sociocultural Dimension of Interaction with Implications for the Learning of Another Language. In *Applied Linguistics*, 14, 145-166.

Hall, J.K. (1995). (Re)creating Our Worlds with Words: A Sociohistorical Perspective of Face-to-face Interaction. In *Applied Linguistics*, 16, 206-232.

Hasegawa, Y. (2014). *Japanese: A linguistic introduction.* Cambridge: Cambridge University Press.

He, A.W., Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A.W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency.* Amsterdam: John Benjamins, 1-24.

Heinz, B. (2003). Backchannel responses as strategic responses in bilingual speakers' conversations. In *Journal of Pragmatics*, 35, 1113–1142.

Heritage, J. (1984). A change-of-state token and aspects of its sequential placement. In Atkinson, J.M. and Heritage, J. (Eds.) *Structures of Social Interaction: Studies in Conversation Analysis*. Cambridge: Cambridge University Press, 299-345.

Jacoby, S., Ochs, E. (1995). Co-construction: An introduction. *Research on language and social interaction*, *28*(3), 171-183.

Jefferson, G. (1984). Notes on a systematic deployment of the acknowledgement tokens "yeah"; and "mm hm". In *Paper in Linguistics*. 17, 197–216. https://doi.org/10.1080/08351818409389201

Kendon, A. (1967) Some functions of gaze-direction in social interaction. In *Acta Psycholoigia* 26, 22-63. https://doi.org/10.1016/0001-69186790005-4

Kjellmer, G. (2009). Where do we backchannel? On the use of mm, mhm, uh huh and such like. In *International Journal of Corpus Linguistics*, 14, 81–112.

Kraaz, M. & Bernaisch, T. (2022). Backchannels and the pragmatics of South Asian Englishes. In *World Englishes*, 41(2), 224-243.

Kraut, R.E., Lewis, S.H. & Swezey, L.W. (1982). Listener responsiveness and the coordination of conversation. *Journal of Personality and Social Psychology*, 43(4), 718–731. https://doi.org/10.1037/0022-3514.43.4.718

Lambertz, K. (2011). Back-channelling: The use of yeah and mm to portray engaged listenership. In *Griffith Working Papers in Pragmatics and Intercultural Communication*, 4, 11-18.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. In *Language Learning*, 40, 387-412.

Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.). *Perspectives on fluency*. Michigan: The University of Michigan Press, 25-42.

Levinson, S. (2015). Turn-taking in Human Communication – Origins and Implications for Language Processing. Trends in Cognitive Sciences. 20. 10.1016/j.tics.2015.10.010.

Li, H.Z., (2006). Backchannel responses as misleading feedback in intercultural discourse. In *Journal of Intercultural Communication Research*, 35(2), 99-116.

Li, H.Z., Cui, Y. & Wang, Z. (2010). Backchannel responses and enjoyment of the conversation: The more does not necessarily mean the better. In *International journal of psychological studies*, *2*(1), 25.

Maynard, S. (1997). Analyzing interactional management in native/non-native English conversation: A case of listener response. In *IRAL- International Review of Applied Linguistics in Language Teaching*, 35(1), 37-60. https://doi.org/10.1515/iral.1997.35.1.37

Mccarthy, M. (2009). Rethinking spoken fluency. In *ELIA, 9, 11-29*.

Nurjaleka, L. (2019). Backchannel Behavior in Interview Discourse: A contrastive study between Japanese and Indonesian. In *Eleventh Conference on Applied Linguistics (CONAPLIN 2018)*, Atlantis Press, 451-457.

Rühlemann, C. (2007). *Conversation in context: A corpus-driven approach*. London: Longman. International Journal of Educational Best Practices, 2, 39–53.

Sacks, H., Schegloff, E.A. & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696-735. https://doi.org/10.2307/412243

Saito, K., Ilkan, M., Magne, V., Tran, M.N. & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency. In *Applied Psycholinguistics*, 39(3), 593-617.

Sbranna, S., Cangemi F. & Grice, M. (2020). Quantifying L2 interactional competence. In Romito, L. (Eds.), *Language change under contact conditions: acquisitional contexts, languages, dialects and minorities in Italy and around the world*, Collana Studi AISV 7, Milano: Officinaventuno, 383-405. DOI: 10.17469/O2107AISV000018

Schegloff, E.A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In D. Tannen (Ed.), *Analysing discourse: Text and talk* (pp. 71-93). Washington, D.C.: Georgetown University Press.

Shelley, L., Gonzalez, F. (2013). Back Channeling: Function of Back Channeling and L1 Effects on Back Channeling in L2. In *Linguistic Portfolios*, 2(1), 9.

Simon, C. (2018). The functions of active listening responses. In *Behavioural Processes,* 157, 47–53. https://doi.org/10.1016/j.beproc.2018.08.013

Stocksmeier, T., Kopp S. & Gibbon, D. (2007). Synthesis of prosodic attitudinal variants in German backchannel ja. In *Interspeech*, 1290-1293.

Tao, H., Thompson, S.A. (1991). English backchannels in Mandarin conversations: A case study of superstratum pragmatic 'interference'. In *Journal of Pragmatics*, 16(3), 209-223. https://doi.org/10.1016/0378-2166(91)90093-D

Tolins, J., Fox Tree, J.E. (2014). Addressee backchannels steer narrative development. In *Journal of Pragmatics*, 70, 152–164.

Tottie, G. (1991). Conversation style in British and American English: The case of backchannels. In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics: Studies in honour of Jan Svartvik*. London: Longman, 254–271.

Ward, N., Tsukuhara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. In *Journal of Pragmatics*, 32, 1177–1207. https://doi.org/10.1016/s0378-2166(99)00109-5

Wehrle, S., Grice, M. (2019). Function and Prosodic Form of Backchannels in L1and L2 German. Poster at *Hanyang International Symposium on Phonetics and Cognitive Sciences of Language*, Seoul, South Korea, 2019.

White, S. (1989). Backchannels across cultures: A study of Americans and Japanese. In *Language in Society*, 18(1), 59-76. https://doi.org/10.2307/4168001

Wood, D.A. (2001). In Search of Fluency: What is it and How Can we Teach It? Canadian Modern Language Review-revue Canadienne Des Langues Vivantes, 57, 573-589.

Wolf, J.P. (2008) The effects of backchannels on fluency in L2 oral task production, *System*, Volume 36, Issue 2, Pages 279-294, ISSN 0346-251X

Wong, D., Peters, P. (2007). A study of backchannels in regional varieties of English, using corpus mark-up as the means of identification. In *International Journal of Corpus Linguistics*, 12, 479–509.

Yngve, V.H. (1970). On getting a word in edgewise. In *Chicago Linguistics Society*, 6, 567–578.

CRediT authorship contribution statement

Simona Sbranna
Conceptualisation
Data curation
Visualisation
Writing – original draft

Simon Wehrle
Conceptualisation
Methodology
Visualisation
Writing – review and editing

Martine Grice
Conceptualisation
Methodology
Supervision

DOI: 10.17469/O2109AISV000009

LOREDANA SCHETTINO, IOLANDA ALFANO, VIOLETTA CATALDO, GIOVANNI LEO

A Crosslinguistic Study on Filled Pauses and Prolongations in Italian and Spanish[1]

Although disfluencies exhibit universal properties, comparative studies have demonstrated cross-linguistic differences as well. This study provides a first comparison between Italian and Spanish, investigating formal and functional features of filled pauses (FPs) and prolongations (PRLs) in dialogical speech. For both phenomena, duration was examined. As regards FPs, we looked at their segmental composition and the surrounding context; for PRLs, we considered the lexical category of the word and the position within the word in which they occur. Beyond individual variability, both Italian and Spanish speakers use more PRLs than FPs, with no interlinguistic duration differences. Furthermore, vocalic final-word PRLs are cross-linguistically preferred. However, Italian speakers present a general higher rate of disfluencies. Finally, FPs show a different segmental composition – related to the language-specific phonetic/phonological inventory –, and seem to be involved in different sub-functions.

Keywords: disfluencies, filled pauses, prolongations, Italian, Spanish.

1. *Introduction*

This study concerns two particular types of speech phenomena in Italian and Spanish task-oriented dialogues, that is Filled Pauses and Prolongations. These elements are usually considered to belong to a heterogeneous class of phenomena that characterize human spontaneous speech and either suspend or edit its production, thus apparently affecting its "fluency"; for this reason, they have been commonly referred to as "disfluencies" (see Lickley, 2015 for an overview).

In the last decades, however, numerous studies have highlighted the significance of such phenomena in the economy of speech. More specifically, human spontaneous speech results from complex online processes including planning, coding, and articulation (Levelt, 1989). Some disfluencies represent flexible and efficient tools which speakers can use to manage their own production (Allwood, Nivre & Ahlsén, 1990; Crocco, Savy, 2003; Voghera, 2017), by either editing something

[1] This article is the result of the collaboration among the authors. However, for academic purposes only, Loredana Schettino is responsible for §§ 1, 2.1, 4.2 and 5.1, Iolanda Alfano for §§ 2.2 and 5.2, Violetta Cataldo for §§ 4.3 and 5.3, Giovanni Leo for §§ 3, 4.1 and 5.4. All the authors are responsible for §§ 6 and 7.

already uttered (also known as "Repairs" or "Backward-Looking Disfluencies") or by monitoring for something about to be uttered (also identified as "Hesitations" or "Forward-Looking Disfluencies", Ginzburg, Fernández & Schlangen, 2014). In particular, filled pauses suspend speech with non-verbal vocalizations and/or nasalizations, i.e., *eeh, ehm, mhh,* whereas prolongations consist of marked lengthening of segmental material (Betz, 2020). Both delay the message delivery, thus reducing the temporal pressure resulting from the simultaneity of planning, production, and reception processes. On the one hand, they gain valuable time for speakers to manage the online process of speech production; on the other hand, they provide extra time for listeners to process information.

The occurrences and the surface realizations of these phenomena are found to vary due to contextual factors and the related cognitive demands (see Bortfeld, Leon, Bloom, Schober & Brennan, 2001), the contextually-determined functions in discourse (see Schettino, Betz, Cutugno & Wagner, 2021), individual factors (see Braun, Rosin, 2015), and language-specific features (see Clark, Fox Tree, 2002).

This study concerns the common and language-specific uses of filled pauses and prolongations in Italian and Spanish dialogues.

2. *Related Work*

2.1 Filled Pauses and Prolongations

Among the different phenomena that may generate suspensions in speech, "non-lexical vocalizations", also referred to as "filled pauses" or "fillers", represent the most salient, most recognizable, and – may be for this reason – most studied expressions of "Forward-Looking" or prospective disfluency. Since the influential study by Clark and Fox Tree (2002), arguing for the speakers' use of different types of filled pauses to signal an upcoming short, "uh", or long, "uhm", delay, it has been intensively debated on the "signal" *versus* "symptom" hypotheses. That is, studies have searched for and provided evidence supporting the use of filled pauses as proper words (Clark, Fox Tree, 2002), or as a by-product of speech production processes (O'Connell, Kowal, 2005; Silber-Varod, Gósy & Lerner, 2021), which nonetheless may provide listeners with valuable information about the ongoing discourse (see Corley, Stewart, 2008; Finlayson, Corley, 2012). More recently, Tottie (2019, 2020) has proposed a different "word-class categorization" for spoken and written uses of filled pauses. According to this view, the former are ascribable to the "fuzzy category of inserts" and the latter to the category of stance adverbs (Tottie, 2019: 128). However, it has been observed that these different uses should be interpreted as preferential and not necessarily linked to the written and spoken modality (Voghera, 2017). Generally, filled pauses have been acknowledged to mark suspensions due to speech planning, corrections, turn management, discourse structure, upcoming new or prominent information, and attention recall (Kjellmer, 2003; Schegloff, 2010; Kosmala, Morgenstern, 2018; Tottie, 2020; Belz, 2021).

From a formal point of view, filled pauses have been described in the literature as non-lexical elements that are realized by a close to a mid-central vocalic and/or nasal phone (see Lickley, 2015; Belz, 2021). However, these vocalizations have been observed to vary across different languages and varieties, as speakers tend to use articulatory models that are strictly linked to their native phonological inventory (Clark, Fox Tree, 2002; Giannini, 2003a; Lickley, 2015). Furthermore, the phonetic-prosodic features of these phenomena have been observed to correlate with functional differences. In particular, a longer duration characterizes occurrences involved in cognitively demanding production processes, i.e., when they occur at the beginning of a phrase (in Dutch monologues, Swerts, 1998), when they introduce new information (in Hungarian monologues, Horváth, 2010), or when they signal problems in the retrieval of a specific word (in Italian monologues, Cataldo, Schettino, Savy, Poggi, Origlia, Ansani, Sessa & Chiera, 2019).

Speakers can temporarily suspend their speech also by prolonging segments beyond their normal duration. Such a lengthening may serve different functions, e.g., marking prosodic phrase boundaries (Albano Leoni, Maturi, 2002), cueing prominence (Bishop, Kuo & Kim, 2020), or signalling hesitation (Eklund, 2004). "Disfluent" lengthening has been described as "a marked prolongation of one or more phones, resulting in above-average syllable and word duration [...] This coincides with a local reduction in speech rate that is not expected by the listener, causing an impression of disfluency and hesitation" (Betz, 2020: 14). Disfluent and non-disfluent lengthening has been observed to be characterized by specific pitch features: prolongations are generally realized with lower pitch range and/or a slowly falling contour, whereas phrase-final lengthenings are usually associated with higher pitch range and boundary (e.g., rising/falling) pitch contour (Savino, Refice, 2000; Shriberg, 2001; Giannini, 2003b; Moniz, 2013; Betz, Eklund & Wagner, 2017). However, given the difficulties and relevance of context in discriminating prolongations, using a "pragmatic" approach based on perceptive criteria has been suggested to provide a "safer" and more suitable solution than relying on temporal thresholds or pitch features (Lickley, 2015).

In different languages, filled pauses have been found to have longer average duration than lengthenings (Swedish, Tok Pisin: Eklund, 2001, 2004; German: Betz et al., 2017; European Portuguese: Moniz, Mata & Viana, 2007; Italian: Giannini, 2003a, 2003b; Cataldo et al., 2019; Di Napoli, 2020). This suggests that these two types of disfluency phenomena may be involved in different ways in the online speech planning. Moreover, in his thesis, Betz (2020) argues that lengthenings represent less salient disfluency elements than silences or filled pauses "which are islands in the speech signal, whereas lengthening stretches the message by ongoing phonation [...] [and is] the softest measure a speaker can apply to solve problems in speech planning" (Betz, 2020: 14).

Besides discourse contextual factors, even individual psycho- and socio-linguistic demands can shape speech planning and production strategies (McDougall, Duckworth, 2017), resulting in speaker-specific uses of disfluency phenomena (see Van Donzel,

Koopmans-van Beinum, 1996; Betz, Lopez Gambino, 2016; Llisterri, Machuca & Ríos, 2019a, b). Hence, various studies suggest that disfluencies may represent a further tool forensic phoneticians may utilize for the identification of speakers (Ishihara, Kinoshita, 2010; Braun, Rosin, 2015; McDougall, Duckworth, 2017).

Moreover, it has been observed that disfluency phenomena may be subjected to the structural and usage constraints imposed by different linguistic systems, as described in the next paragraph.

2.2 Language-specific Features

As noted by McDougall and Duckworth (2017), comparisons across studies are difficult because of differences in the taxonomies used, in the speech style, and in the ways each study compares the occurrence of disfluencies against the whole speech sample (e.g., per 100 syllables, per 100 words, or per minute of speech). Furthermore, some authors rely on perceptive criteria, while others use acoustic duration thresholds.

Even if disfluencies seem to exhibit universal properties, comparative studies have demonstrated cross-linguistic differences, thus giving support to Clark and Fox Tree's (2002) argument that fillers are language-specific. Besides, it has been shown that the use of different hesitation markers and their interplay may respond not only to phonological, syntactic, and semantic constraints, but also to pragmatic culture-specific dynamics (Betz et al., 2021).

McDougall and Duckworth (2017) for British English and Llisterri, Machuca and Ríos (2022) for Spanish analyze disfluency phenomena with the aim of testing the extent to which they are employable for forensic discrimination across speakers. The results of these studies indicate that each speaker presents specific disfluency traits, although not all of these contribute equally to speaker identification. Cross-linguistically, the most frequent disfluencies in Spanish are vocalic prolongations and silent pauses (Llisterri et al., 2022), while the results on British English reveal the presence of silent and filled pauses, but a lower occurrence of segmental prolongations compared to Spanish (McDougall, Duckworth, 2017).

Language-specific patterns may concern not only a different distribution, but also a specific phonetic realization. Candea, Vasilescu and Adda-Decker (2005) study the vocalic peculiarities of "autonomous fillers" (our filled pauses) in several languages, i.e., the realization of a central or non-central timbre of their vocalic support, nasalizations or diphthongized segments. They examine the following eight languages: standard Arabic, Mandarin Chinese, French, German, Italian, European Portuguese, American English, and Latin American Spanish; they base fillers extraction on a minimum duration threshold of 200 ms. They consider three parameters: the duration, the F1/F2 values of the fillers' vocalic segments, and the fundamental frequency (f_0). Whereas f_0 and duration do not show significant differences among the eight languages, the acoustic analysis of F1/F2 reveals language-dependent characteristics. The number of occurrences of disfluencies per language is not considerable for Italian and Spanish corpora (57 and 93 occurrences,

respectively). However, their results strengthen the hypothesis of cross-linguistic timbre differences of the vocalic support of the autonomous fillers, showing that different languages admit various vocalic realizations. Spanish employs the closed-mid vowel [e], while Italian makes use of both central and non-central vocalic supports, i.e., the front open-mid vowel [ɛ].

De Leeuw (2007) reports significant differences among English, German, and Dutch hesitation markers in the proportion of vocalic, vocalic-nasal, and nasal markers, as well as in their positioning.

In the last decades, Eklund (2001) and various colleagues conducted a number of investigations on the characteristics and uses of prolongations in different languages. It has been observed that all types of segments could be subjected to lengthening, but vowels and sonorants are usually more prone to it within the language-specific phonological constraints. For instance, since vowel length is a distinctive feature in German, prolongation tends to be avoided on short vowels (Betz et al., 2017). Lengthened segments generally occur in the word-final syllable, although the distribution within word-initial/medial/final segments seems to be highly dependent on language-specific syllable structure and phonotactic rules. Prolongations tend to occur on functional rather than content words in English (O'Shaughnessy, 1995), Swedish, Tok Pisin (Eklund, 2001), German (Betz et al., 2017), Hungarian (Gósy, Eklund, 2017), whereas no clear-cut distinction was found in Mandarin (Lee, He, Huang, Tseng & Eklund, 2004) and Japanese (Den, 2003).

As far as Italian and Spanish are concerned, the relevant literature on fillers indicates that the most frequent vocalization is a vocalic element representable in Spanish as [eː] (Rebollo, 1997; Machuca, Llisterri & Ríos, 2015) and in Italian as [əː] or [əmː] (Cataldo et al., 2019). Regarding the lexical category of the word affected by lengthening, most cases in Spanish were found to correspond to functional words, in particular, prepositions represent the highest percentage of cases (Machuca et al., 2015); in Italian, instead, a rather balanced distribution among open class (most frequently verbs, nouns, and adverbs) and closed class (mostly prepositions, conjunctions, and determiners) words was found (Di Napoli, 2020). As far as we know, there are no studies comparing Italian and Spanish by considering the same speech style and adopting the same annotation scheme.

3. *Research Aim*

This study sets out to investigate formal and functional features of filled pauses (FPs) and prolongations (PRLs) in dialogical speech of Italian and Spanish from a crosslinguistic perspective. As shown in §1, FPs and PRLs, like other disfluency phenomena, are present in every spoken language, but at the same time possess language-specific traits, which are strongly related to the linguistic (e.g., phoneme inventory, syntactic structures) and extralinguistic constraints (e.g., communicative context) of the language(s) in question. Hence, the investigation of FPs and PRLs in two genealogically related languages, Italian and Spanish, will add to the pre-existing

knowledge on the phenomena by highlighting commonalities and specificities in formal and functional terms by taking count of several factors, among which distributional and acoustic features (§4.3). To summarize, this work will attempt to provide an answer to the following questions:

– What are the common traits exhibited by FPs and PRLs in Italian and Spanish? What are, on the other hand, the language-specific ones?
– What level(s) (e.g., formal or functional) do the aforementioned patterns refer to?

4. *Method*

4.1 Corpus

The corpus for the analysis consists of 4 task-oriented dialogues (2 in Italian and 2 in Spanish), elicited through the "spot the difference" technique (Péan, Williams & Eskénazi, 1993). Table 1 summarizes the dataset by showing the duration of each dialogue per language. The Italian dialogues belong to the Neapolitan variety, whereas the Spanish ones belong to the variety of Spanish spoken in Barcelona. In all, 50 minutes of speech have been examined, about 25 minutes in Italian and about 26 in Spanish.

Table 1 - *Dataset for the analysis*

Italian Dialogue	*Duration (min)*	*Spanish Dialogue*	*Duration (min)*
TDA01N	*14.18*	*TDA01BCN*	*12.14*
TDA02N	*10.16*	*TDA02BCN*	*14.03*
total	*24.5*	*total*	*26.3*

The elicitation and the transcription of each dialogue has been conducted within the framework of the CLIPS project (Savy, Cutugno, 2009). The "spot the difference" elicitation method consists in a game, during which the communicating dyad has to identify the differences on two apparently identical pictures by relying solely on the verbal channel, without seeing each other. The peculiarity of this method is that both participants are placed on a par, which results in a highly interactive task. Although this elicitation technique has limits, it provides a certain naturalness as for the phonetic-prosodic traits, thus resulting in semi-spontaneous speech, with a low degree of discourse planning.

4.2 Disfluency Annotation

Given the cross-linguistic nature of the present study, we identified formal, functional, and structural parameters for filled pauses (FPs) and prolongations (PRLs) that can exhibit common or language-specific uses in Italian and Spanish dialogues.

We carried out the analysis by relying on an annotation scheme designed in previous studies (Cataldo et al., 2019; Schettino et al., 2021; Schettino, 2022). The scheme allows for a multi-layered annotation of disfluency phenomena, based on both

their formal structure and their contextual functions. Accordingly, the ELAN software was used (Sloetjes, Wittenburg, 2008), which permits multilevel linguistic annotation.

For the present study, we have focused on FPs and PRLs (Eklund, 2004; Betz, 2020). Each occurrence has been identified and labelled as follows: FP for non-verbal fillers realized as vocalization and/or nasalization; PRL for marked segmental lengthening. Since the annotation was previously designed for and applied to a different type of speech, namely almost monologic speech of Italian tourist guides (Cataldo et al., 2019; Schettino et al., 2021), we tested the inter-annotator agreement for the identification of the types of disfluency phenomena in dialogic speech conducted. The measured Cohen's ϰ reached 0.92, which stands for "high agreement" according to Landis, Koch (1977).

The two disfluency types serve a broad Forward-Looking function, namely they mark a temporary speech delay (Ginzburg et al., 2014). In the annotation system, such a function is further specified on the basis of the context of occurrence (for more details, see Schettino et al., 2021; Schettino, 2022).

a. Word Searching (WS), when disfluencies are involved in lexical retrieval or lexical selection purposes (Tottie, 2020).
b. Structuring (STR), for disfluencies occurring at the boundaries of syntactic or information structure, e.g., clauses and topic-comment, respectively.
c. Focusing (FOC), associated with disfluencies marking upcoming "semantically heavy concepts or words" (Kjellmer, 2003)[2].
d. Hesitative (HES), for disfluencies not fitting into any of the preceding sub-functions but triggered only by broad speech planning.

The inter-rater's agreement on the assignment of sub-functions to disfluency items reached Cohen's ϰ of 0.78 (substantial agreement, Landis, Koch, 1977).

4.3 Analysis Parameters

FPs and PRLs were then analyzed in Praat (Boersma, Weenink, 2018). Due to the structural differences of the disfluency types under investigation, we selected both general and disfluency-specific parameters. Accordingly, for both FPs and PRLs, we considered:

– Duration (ms).

As regards FP, we looked at:

– Segmental composition, namely the phonetic realization of vocalization and nasalization; specifically, we labelled the sequence of phones constituting the FP, e.g., [em], [ə];
– The surrounding context, i.e., preceding and following the FP, for each occurrence we annotated whether the FP was preceded or followed by speech or silence.

As for PRL, the analysis concerned:

[2] The label FOC was not assigned to phenomena signalling properly focalized elements, but rather to items involved in the planning and production of key information, e.g., new or emphasized elements, independently from syntactic structures.

– The lexical category of the word affected by lengthening, in order to test whether this disfluency type is more likely to affect functional (e.g., articles, conjunctions) or content words (e.g., nouns, verbs).
– The position of PRL within the affected word, namely initial, medial or final position.

The statistical analysis was performed in *R* (R Core Team, 2020), by means of Generalized Linear and Linear Mixed Models in order to control for individual variability ('lme4' package, Bates, Maechler, Bolker & Walker, 2015). Disfluency Type and Duration were set as the dependent variable, Function and Language as interacting independent variables, while Speaker was set as a random effect.

5. *Results*

A total of 406 disfluency items were analyzed, namely 142 FPs and 264 PRLs. First, the overall results on the frequency of occurrence and on the functions of the disfluencies are presented. Then, results per parameter are reported.

The Italian data present a higher incidence of disfluency rate (10,8 per minute) than Spanish data (5,4 per minute). Figure 1 shows the frequency of occurrence of the two types of phenomena in Italian and Spanish dialogues. Generally, PRLs are more frequent than FPs in both languages, although this difference is significantly larger in Spanish than in Italian (estimate: -2.34; SE: 0.58; z-value: -4.03; $p < 0.001$).

Actually, Italian speakers make greater use of both types of disfluencies (FPs = 110, PRLs = 154) with respect to Spanish (FPs = 31, PRLs = 110). The higher incidence per speaker in Italian compared to Spanish dialogues is reported in Table 2.

The observed cross-linguistic difference is significant beyond individual variability. Indeed, all speakers confirm this trend, especially for Spanish speakers presenting a higher incidence of PRLs compared to FPs.

Figure 1 - *Frequency of occurrence of FP and PRL in Italian (left) and Spanish (right) dialogues*

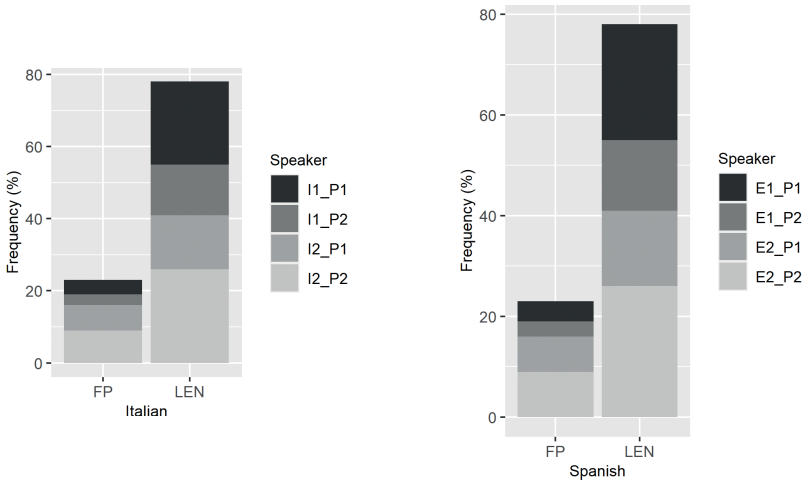
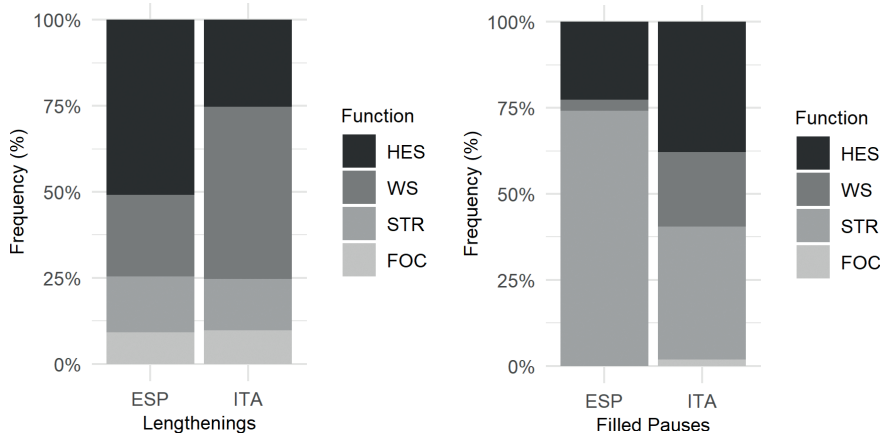


Table 2 - *Incidence (%) of FP and PRL across speakers of Italian (left) and Spanish (right)*

	I1_P1	*I1_P2*	*I2_P1*	*I2_P2*	*E1_P1*	*E1_P2*	*E2_P1*	*E2_P2*
FP (%)	31	29	10	41	5	4	10	12
PRL (%)	32	57	28	37	32	20	21	37

As for sub-functions (Fig. 2), in Spanish, FPs mostly serve a structuring function, whereas in Italian FPs exhibit a more even distribution. Note that FPs are rarely associated with the focusing function, which is, in any case, an infrequent sub-function in this corpus. On the other hand, PRLs are in fact significantly more used for hesitative, generic planning, and word searching, and far less for structuring. Indeed, comparing the use of the two selected phenomena, FPs are associated with the structuring function significantly more than PRLs in both Italian and Spanish (estimate: -2.54; SE: 0.56; z-value: -4.58; $p < 0.001$).

Figure 2 - *Sub-functions of FP (left) and PRL (right) in Italian and Spanish dialogues*



5.1 Duration

Duration was found to vary significantly between disfluency types but, crucially, not between the two languages. Indeed, in both Italian and Spanish, FPs are significantly longer than PRLs (estimate: -255.16; SE: 41.43; t value: -6.159), as shown in Table 3 and Figure 3.

Table 3 - *Duration per disfluency type and language*

Language	*Hes type*	*Dur (ms)*	*St. dev.*	*SE*
ESP	*FP*	530.11	336.65	60.46
ESP	*PRL*	272.07	111.09	10.59
ITA	*FP*	495.86	294.09	27.91
ITA	*PRL*	280.31	137.66	11.09

We also looked at the correlation between the duration of disfluencies and the corresponding sub-functions (Fig. 4). FPs are longer than PRLs regardless of their sub-function, whereas PRLs involved in word searching are on average longer than any other PRLs. This difference is significant for Italian PRLs (estimate: 199.36; SE: 66.05; t value: 3.018), while in Spanish it is a tendency that does not reach statistical significance.

Figure 3 - *Duration (ms) of FP and PRL in Spanish and Italian dialogues*

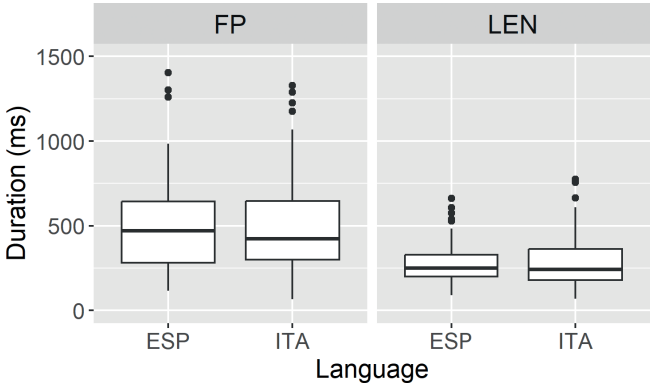
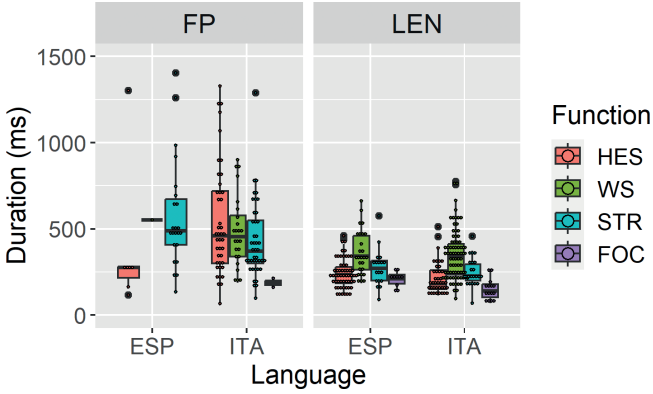


Figure 4 - *Duration (ms) of FP and PRL per function in Spanish and Italian dialogues*



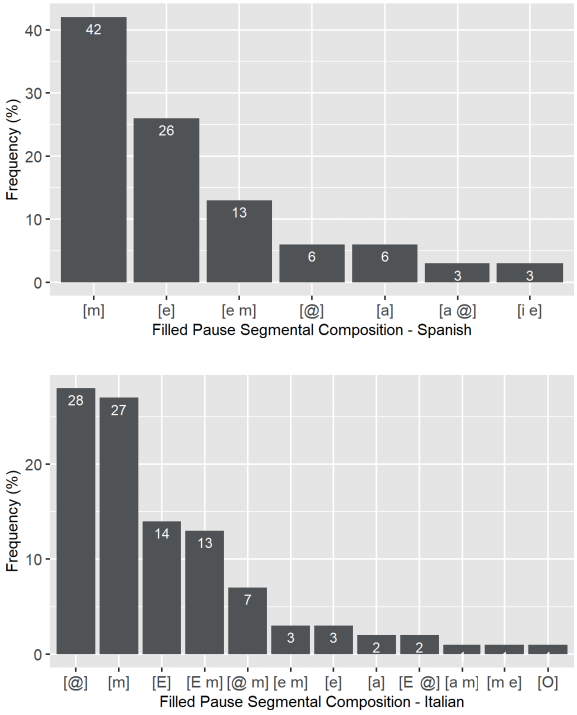
5.2 Segmental Composition

In Spanish, non-verbal vocalizations are slightly more frequently realized by nasalizations, namely [m] (55% of cases), followed by [e] and by their sequence, [em], and less frequently by [ə] (6% of cases).

On the other hand, in Italian, mid central vowel, schwa, turns to be the most frequent realization for FPs (28% of cases), followed by [m], [ɛ], and [ɛm] sequences. Generally, in the distinction between nasalized and non-nasalized items, FPs are evenly realized by nasalizations and vocalic sounds (51% and 49% of cases, respectively). Results on FP segmental composition are provided in Figure 5.

In these datasets, no differences were found in the correlation between specific phonetic realizations and sub-functions.

Figure 5 - *FP segmental composition in Spanish (top panel) and Italian (bottom panel) dialogues*

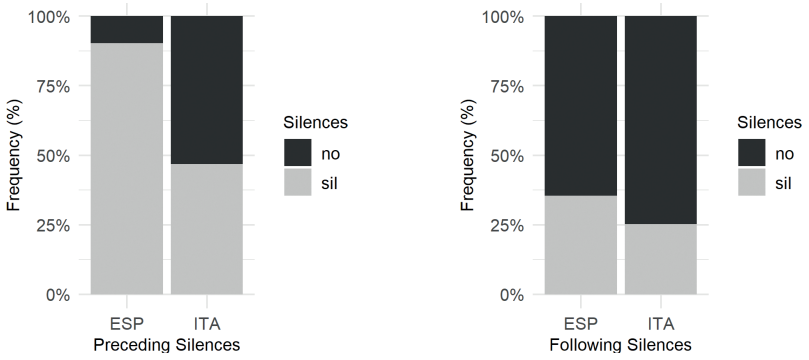


5.3 Surrounding Context

Considering the contextual elements (Fig. 6), Spanish FPs are almost exclusively preceded by a silence, whereas in Italian they are preceded by silences as often as speech. This cross-linguistic difference is statistically significant (estimate: -2.08; SE: 0.73; z value: -2.84; $p < 0.005$). Moreover, regardless of the language, the presence of a silence before a FP is significantly related to the structuring function (estimate: 0.81; SE: 0.43; z value: 2.11; $p < 0.035$).

As far as the following context is concerned, in both languages FPs are mostly followed by speech and not by silences.

Figure 6 - *Preceding (left) and following (right) FP in Spanish and Italian dialogues*

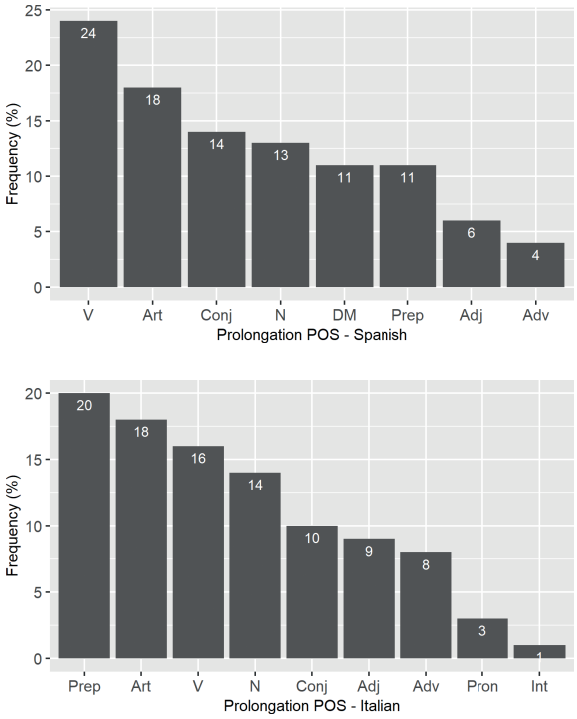


5.4 Prolongation Word Category and Position

As for the lexical categories affected by prolongations (Fig. 7), in Italian and Spanish, PRLs equally occur both on functional words (53% and 54%, respectively) and on content words (47% and 46%, respectively). The Italian distribution confirms previous studies (Di Napoli, 2020), while the distribution in Spanish does not show a marked preference for functional words and prepositions (contra Machuca et al., 2015).

Specifically, in the two languages, PRLs more commonly affect verbs, articles, and nouns.

Figure 7 - *PRL lexical category in Spanish (top panel) and Italian (bottom panel) dialogues*



As for the within-word position of PRLs, similar tendencies were observed in the two languages. Few cases of prolongations in initial (4% in Spanish and 3% in Italian) and medial (7% in Spanish and 1% in Italian) position. The vast majority of PRLs occur in word-final position, namely in 89% of cases in Spanish and 96% of cases in Italian.

6. *Discussion*

In this study, we explored speech management strategies adopted by speakers of different languages in the same communicative situation, that is in task-oriented dialogues in Italian and Spanish. More specifically, we focused on forward-looking disfluencies that generate a suspension in speech due to planning and production demands by means of non-lexical vocalizations (Filled Pauses) and marked lengthenings of word segments (Prolongations); we then analysed the way these phenomena are used by Italian and Spanish speakers based on their distribution, their contextually determined functions, their phonetic aspects, and possible correlations between these features.

From the results emerges, as expected, some individual variability in the use of both Filled pauses and Prolongations (see Betz, Lopez Gambino, 2016; McDougall, Duckworth, 2017). Beyond this source of variability, the analysis highlights, on the one hand, uses of these speech management phenomena that are common to Italian and Spanish speakers, and on the other hand, language-specific structures and dynamics.

Both Italian and Spanish speakers consistently use disfluency phenomena. In particular, segmental prolongation emerges as a more frequent hesitation strategy compared to filled pauses (as also found by Eklund, 2004). Speakers mostly rely on final-word prolongations for delaying speech due to generic planning and word searching, whereas filled pauses are more frequently used to take more time for planning processes involving the introduction of a new proposition.

Furthermore, in both languages, duration values characterize the two disfluency types differently: filled pauses are on average significantly longer than prolongations (as also found in previous literature involving different languages and styles, see §2.1).

These findings support the interpretation of prolongations as a more convenient and subtle means of taking time as opposed to filled pauses that represent, instead, more "salient" phenomena used to provide extra time when the planning and construction processes require it.

Then, prolongations occur quite evenly on content words (mostly verbs and nouns) and functional words (mostly articles, conjunctions, and prepositions). These findings are similar to those on Italian dialogues by Di Napoli (2020), but quite different from previous observations on Spanish by Machuca and colleagues (2015). However, this result may be due to the specific speech style of the analysed data and the resulting different frequency of lexical categories. Indeed, the observed task-oriented dialogues, aimed at the identification of differing details in similar

pictures, may be characterized by a relatively higher rate of content words, e.g., nouns and verbs.

Lastly, final-word position for segmental prolongation is cross-linguistically preferred, which may be due to the fact that Italian and Spanish display a similar distribution of syllabic structures, i.e., prevalently CV structures. However, the CVC type is slightly more frequent in Spanish, which reaches around 20%, than in Italian. The difference increases when considering the final-word position in stressed syllables, a position for which Spanish presents a clear majority of the CVC type (Alfano, 2008). Given the preference to lengthen vowel segments, one would have expected a different finding in this respect. Our data indicate rather that the preference for prolongation in word-final position seems to be not directly dependent on syllabic composition.

Besides these common patterns, a number of language-dependent uses and features have emerged from the analysis.

Firstly, compared to the Spanish dialogues, the Italian ones show a higher rate of disfluencies. Some differences also concern the distribution of the sub-functions, especially for filled pauses. While in Italian this type of forward-looking disfluency may be almost evenly involved in all the considered functions, in Spanish, filled pauses seem to correlate more frequenly with the structuring function. Moreover, in the Spanish dialogues, filled pauses are almost exclusively preceded by a silence, whereas they seem to be equally preceded by speech or a silence in Italian. These observations may suggest a more controlled use of disfluency phenomena by Spanish speakers and, conversely, a higher tolerance for these elements by Italian speakers.

A feature that appears to be peculiar to Italian concerns the correlation between the duration of prolongations and specific sub-functions. As already observed in Cataldo et al. (2019), longer prolongations are more likely to be involved in lexical retrieval processes. Accordingly, this emerges as a robust feature in Italian monologic as well as dialogic speech.

Then, the feature that most of all reveals language-related specificities is the segmental composition of filled pauses. Indeed, it corroborates the assumption that speakers of different languages tend to generate realizations that are strictly linked to their phonological inventory even for the production of non-lexical vocalizations (Clark, Fox Tree 2002; Giannini, 2003a; Giannini 2003b; Ginzburg et al., 2014). In line with the literature, Spanish mostly employs [eː] and [mː] (Rebollo, 1997; Machuca et al., 2015), while in Italian the most frequent sounds are [əː], [ɛː], followed by nasals, e.g., [əːm], [ɛːm] (Giannini, 2003; Cataldo et al., 2019). In fact, the mid-central [əː] phone is not acknowledged as an Italian phoneme. Therefore, in their contrastive corpus study, Candea et al. (2005) claim that Italian is the only language with a vocalic support which is not part of the vocalic system. However, this finding should be interpreted in the light of the regional variety examined in their work, which cannot be deduced in the paper. Indeed, it has been attested that the mid-central vocalic variant characterizes the dialectal substrate of the Neapolitan variety of Italian (Pellegrini, 1977; Loporcaro, 2009, 2016; Ledgeway, 2016). So, as

already observed by Giannini (2003), this realization appears to be connected to underlying dialectal sounds percolating into the local variety of Italian, rather than to other processes, such as the speaker's articulatory economy.

7. *Conclusions*

To conclude, this study suggests cross-linguistically shared uses of filled pauses and prolongations, in terms of phonetic "salience", and common characteristics resulting from similarities between the Italian and Spanish linguistic structures or the specific communicative context and goals. Language-specific uses concern, instead, the relative tolerance for the observed phenomena and phonological differences between Italian and Spanish.

This study has benefitted from the consideration of Italian and Spanish datasets collected by using the same elicitation technique and, therefore, constituting comparable speech data. Given the difficulties in comparing cross-study findings on disfluencies' uses and features, since they rely on different speech types and adopt different approaches, focused cross-linguistic investigations seem to be particularly relevant to the literature and provide more systematic and feasible comparisons in order to shed light on common and divergent disfluency features across languages. Nonetheless, caution is required in the interpretation of the described findings considering the rather small size of the datasets involved. Also, this investigation concerned two types of forward-looking disfluencies, but speakers may use other phenomena to suspend their speech for planning, such as silences or lexical fillers. Hence, the results emerged are only an indication of how and when the phenomena in question may be used in these languages. So, studying the different hesitation tools at speakers' disposal, and their interplay, may contribute to shedding light on more cross-linguistical different uses. For example, Spanish speakers may compensate for the scarce amount of filled pauses by relying on "lexical fillers", such as discourse markers that serve a planning function. Future investigations may follow this lead by employing larger datasets and including other disfluency types in order to deepen our understanding and gain a clearer picture of the strategies speakers may enact to manage their speech.

References

Albano Leoni, F., Maturi, P. (2002). *Manuale di fonetica*. Roma: Carocci.

Alfano, I. (2008). Strutture sillabiche ed accentuali in italiano e in spagnolo. In *Testi e Linguaggi*, 2. Roma: Carocci, 18-36.

Allwood, J., Nivre, J. & Ahlsén, E. (1990). Speech management – on the non-written life of speech. In *Nordic Journal of Linguistics*, 13, 3-48. https://doi.org/10.1017/S0332586500002092

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. In *Journal of Statistical Software*, 67(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Belz, M. (2021). *Die phonetik von äh und ähm: Akustische variation von füllpartikeln im deutschen*. Berlin: Springer Nature.

Betz, S. (2020). Hesitations in Spoken Dialogue Systems. PhD Dissertation, University of Bielefeld.

Betz, S., Bryhadyr, N., Kosmala. L. & Schettino, L. (2021). A Crosslinguistic Study on the Interplay of Fillers and Silences. In Rose, R., Eklund, R. (Eds.), *Proceedings of DiSS 2021, The 10th Workshop on Disfluency in Spontaneous Speech*, Paris, France, 25-27 August 2021, 47-52.

Betz, S., Eklund, R. & Wagner, P. (2017). Prolongation in German. In Rose, R., Eklund, R. (Eds.), *Proceedings of DiSS 2017, The 8th Workshop on Disfluency in Spontaneous Speech*, Stockholm, Sweden, 18-19 August 2017, 13-16.

Betz, S., Lopez Gambino, M.S. (2016). Are we all disfluent in our own special way and should dialogue systems also be?. In *Proceedings of Elektronische Sprachsignalverarbeitung (ESSV)*, Leipzig, Germany, 2-4 March 2016, 81.

Bishop, J., Kuo, G. & Kim, B. (2020). Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: Evidence from rapid prosody transcription. In *Journal of Phonetics*, 82, 100977.

Bortfeld, H., Leon, S.D., Bloom, J.E., Schober, M.F. & Brennan, S.E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. In *Language and speech*, 44(2), 123-147. https://doi.org/10.1177/00238309010440020101

Braun, A. & Rosin, A. (2015). On the speaker specificity of hesitation markers. In Maria Wolters, M., Livingstone J., Beattie, B., Smith, R., MacMahon, M., Stuart-Smith, J., Scobbie, J., M. (Ed.), *Proceedings of the International Congress of Phonetic Sciences (ICPhS 2015)*, Glasgow, UK, 10-14 August 2015.

Candea, M., Vasilescu i. & Adda-Decker, m. (2005). Inter- and intra-language acoustic analysis of autonomous fillers. In *Proceedings of the 4th Workshop on Disfluency in Spontaneous Speech (DiSS 2005),* Aix-en-Provence, France, 10-12 September 2005, 47–51.

Cataldo, V., Schettino, L., Savy, R., Poggi, I., Origlia, A., Ansani, A., Sessa, I. & Chiera, A. (2019). Phonetic and functional features of pauses, and concurrent gestures, in tourist guides' speech. In Piccardi, D., Ardolino, F. & Calamai, S. (Eds.), *XV Convegno Nazionale AISV. Gli archivi sonori al crocevia tra scienze fonetiche, informatica umanistica e patrimonio digitale.* Studi AISV 6. Milano: Officinaventuno, 205-231.

Clark, H.H., Fox Tree, J.E. (2002). Using uh and um in spontaneous speaking. In *Cognition*, 84(1), 73-111. https://doi.org/10.1016/S0010-0277(02)00017-3

Corley, M., Stewart, O.W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. In *Language and Linguistics Compass,* 2(4), 589-602. https://doi.org/10.1111/j.1749-818X.2008.00068.x

Crocco, C., Savy, R. (2003). Fenomeni di esitazione e dintorni: una rassegna bibliografica. In Crocco, C., Savy, R. & Cutugno, F. (Eds.), *API. Archivio di Parlato Italiano,* DVD.

De Leeuw, E. (2007). Hesitation Markers in English, German, and Dutch. In *Journal of Germanic Linguistics*, 19(2), 85-114.

Den, Y. (2003). Some strategies in prolonging speech segments in spontaneous speech. In Eklund, R. (Ed.) *Proceedings of the ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech (DiSS 2003)*, Göteborg, Sweden, 5-8 September 2003, 87-90.

Di Napoli, J. (2020). Filled pauses and prolongations in Roman Italian task-oriented dialogue. In *Proceedings of the Laughter and Other Non-Verbal Vocalisations Workshop*, Bielefeld, 5 October 2020, 24-27.

Eklund, R. (2004). Disfluency in Swedish human-human and human-machine travel booking dialogues. PhD Dissertation, Linköping University: Electronic Press.

Giannini, A. (2003a). Hesitation phenomena in spontaneous Italian. In Solé, M.J., Recasens, D. & Romero, J. (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*, Barcelona, Spain, 3-9 August 2003, 2653-2656.

Giannini, A. (2003b). Vocalizzazioni e prolungamenti vocalici. In Magno Caldognetto E., Cosi P. (Eds.), *Voce, canto, parlato. Studi in onore di Franco Ferrero. Padova: Unipress*, 163-172.

Ginzburg, J., Fernández, R. & Schlangen, D. (2014). Disfluencies as intra-utterance dialogue moves. In *Semantics and Pragmatics*, 7(9), 1-64. https://doi.org/10.3765/sp.7.9

Horváth, V. (2010). Filled pauses in Hungarian: Their phonetic form and function. In *Acta Linguistica Hungarica* (Since 2017 *Acta Linguistica Academica*), 57(2-3), 288-306. https://doi.org/10.1556/ALing.57.2010.2-3.6

Ishihara, S. & Kinoshita, Y. (2010). Filler words as a speaker classification feature. In Tabain, M., Fletcher, J., Grayden, J., Hayek, J. & Butcher, A (Eds.), *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*, Melbourne, Australia, 14-16 December 2010, 34-37.

Kjellmer. G. (2003). Hesitation. in defence of er and erm. In *English Studies*, 84(2), 170-198. https://doi.org/10.1076/enst.84.2.170.14903

Kosmala, L., Morgenstern, A. (2018). Should 'uh' and 'um' be categorized as markers of disfluency? the use of fillers in a challenging conversational context. In Degand, L., Gilquin, G., Meurant, L. & Simon, A., C. (Eds.), *Fluency andDisfluency across Languages and Language Varieties. Corpora and Language in Use.* Louvain-la-Neuve: Presses universitaires de Louvain, 67–89.

Landis, J., Koch, G. (1977). The measurement of observer agreement for categorical data. In *Biometrics*, 33(1), 159-74. https://doi.org/10.2307/2529310

Ledgeway, A. (2016). The dialects of southern Italy. In Ledgeway, A. & Maiden, M. (Eds.), *The Oxford guide to the Romance languages*, Vol. 1, 245-269, Oxford: Oxford University Press.

Lee, T.-L., He, Y.-F., Huang, Y.-J., Tseng, S.-C. & Eklund, R. (2004). Prolongation in spontaneous Mandarin. In *Proceedings of Interspeech 2004*, Jeju Island, Korea, 4-8 October 2004, 2181-2184.

Levelt, W.J. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press. https://doi.org/10.1016/0010-0277(83)90026-4

Lickley, R.J. (2015). Fluency and Disfluency. In Redford, M.A. (Ed.), *The handbook of speech production*. Chichester: John Wiley & Sons, 445-474. https://doi.org/10.1002/9781118584156.ch20

Llisterri, J., Machuca, M.J. & Ríos, A. (2019a). Caracterización del hablante con fines judiciales: fenómenos fónicos propios del habla espontánea. e-AESLA. In *Revista digital de lingüística aplicada*, 5, 265–278.

Llisterri, J., Machuca, M.J. & Ríos, A. (2019b). VILE-P: un corpus para el estudio prosódico de la variación inter e intralocutor. En Lahoz-Bengoechea, J.M. & Pérez Ramón, R.J. (Eds.), *Subsidia: Tools and Resources for Speech Sciences / Subsidia: herramientas y recursos para las ciencias del habla*, Universidad de Málaga, 117-123. https://hdl.handle.net/10630/18177

Llisterri, J., Machuca, M.J. & Ríos, A. (2022). La función de las hesitaciones en la identificación del hablante. In Blecua, B., Cicres, J., Espejel, M. & Machucha, M.J. (Eds.), *Propuestas en fonética experimental: enfoques metodológicos y nuevas tecnologías*, Universitat de Girona, 160-164. http://hdl.handle.net/10256/20770

Loporcaro, M. (2009). *Profilo lingüístico dei dialetti italiani*. Roma-Bari: Laterza.

Loporcaro, M. (2016). L'Italia Dialettale. In Lubello, S. (Ed.), *Manuale di linguistica italiana*. Berlin: De Gruyter, 275-300.

Machuca, M.J., Llisterri, J. & Ríos, A. (2015). Las pausas sonoras y los alargamientos en español: un estudio preliminar. In *Normas. Revista de Estudios Lingüísticos Hispánicos, 5*, 81-96.

McDougall, K., Duckworth, M. (2017). Profiling fluency: an analysis of individual variation in disfluencies in adult males. In *Speech Communication*, 95, 16-27. https://doi.org/10.1016/j.specom.2017.10.001

Moniz, H. (2013). Processing disfluencies in European Portuguese. PhD Dissertation, University of Lisbon.

Moniz, H., Mata, A.I. & Céu Viana, M.C. (2007). On filled-pauses and prolongations in European Portuguese. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association*. Antwerp, Belgium, 27-31 August 2007, 2820-2824 https://doi.org/10.21437/Interspeech.2007-695

O'Connell, D.C., Kowal, S. (2005). Uh and um revisited: Are they interjections forsignaling delay? In *Journal of Psycholinguistic Research*, 34(6), 555-576. https://doi.org/10.1007/s10936-005-9164-3

O'Shaughnessy, D. (1992). Recognition of hesitations in spontaneous speech. In *Proceedings: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP-92), San Francisco, USA, 23-26 March 1992, 521-524. https://doi.org/10.1109/ICASSP.1992.225857

Péan, V., Williams S. & Eskenazi; M. (1993). The design and recording of ICY, a corpus for the study of intraspeaker variability and the characterization of speaking styles. In *Proceedings of Eurospeech 2003*, Geneva, Switzerland, 1-4 September 2003, 627-630.

Pellegrini, G.B. (1977). *Carta dei dialetti d'Italia*. Pisa: Pacini.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rebollo, L. (1997). Pausas y ritmo en la lengua oral. Didáctica de la pronunciación. In Moreno, F., Gil, M. & Alonso, K. (Eds.), *El español como lengua extranjera: del pasado al futuro. Actas del VIII Congreso Internacional de la Asociación para la Enseñanza del Español como Lengua Extranjera*, Alcalá de Henares, Spain, 17-20 September 1997, 667-676.

Savy, R., Cutugno, F. (2009). Diatopic, diamesic and diaphasic variations in spoken Italian. In Mahlberg, M., González-Díaz, V. & Smith, C. (Eds.), *Proceedings of the 5th Corpus Linguistics Conference (CL2009)*, Liverpool, UK, 20-23 July 2009, 20-23.

Schegloff, E.A. (2010). Some other "uh (m)"s. In *Discourse Processes*, 47(2), 130-174. https://doi.org/10.1080/01638530903223380

Schettino, L., Betz, S., Cutugno, F. & Wagner, P. (2021). Hesitations and individual variability in Italian tourist guides' speech. In Bernadasci, C., Dipino, D., Garassino, D., Negrinelli, S. & Schmid, S. (Eds.), *Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications.* Studi AISV 8, Milano: Officinaventuno, 243-262.

Schettino, L. (2022). The role of disfluencies in Italian discourse. Modelling and speech synthesis applications. PhD Dissertation, University of Salerno.

Shriberg, E. (2001). To 'errrr' is human: ecology and acoustics of speech disfluencies. In *Journal of the International Phonetic Association*, 31, 153-169. https://doi.org/10.1017/S0025100301001128

Silber-Varod, V., Gósy, M. & Lerner, A. (2021). Is it a filler or a pause? A quantitative analysis of filled pauses in Hebrew. In Karpov, A., Potapova, R. (Eds.), *Proceedings of the 23rd International Conference on Speech and Computer*, SPECOM 2021, St. Petersburg, Russia, 27-30 September 2021, 638-648.

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 2008)*, Marrakesh, Marocco, 28-30 May 2008, 816-820.

Swerts, M. (1998). Filled pauses as markers of discourse structure. In *Journal of Pragmatics*, 30(4), 485-496. https://doi.org/10.1016/S0378-2166(98)00014-9

Tottie, G. (2019). From pause to word: Uh, um and er in written American English. In *English Language & Linguistics*, 23(1), 105-130.

Tottie, G. (2020). Word-Search As Word-Formation?: The Case Of "Uh" And "Um". In Núñez-Pertejo, P., López-Couso, M.J., Méndez-Naya, B. & Pérez-Guerra, J. (Eds.), *Crossing linguistic boundaries: systemic, synchronic and diachronic variation in english*. London: Bloomsbury Academic, 29–42. https://doi.org/10.5040/9781350053885.ch-002

Van Donzel, M.E., Koopmans-Van Beinum, F.J. (1996). Pausing strategies in discourse in Dutch. In *Proceeding of Fourth International Conference on Spoken Language Processing.* ICSLP'96. 2, Philadelphia, PA, USA, 3-6 October 1996, 1029-1032.

Voghera, M. (2017). *Dal parlato alla grammatica*. Roma: Carocci.

DOI: 10.17469/O2109AISV000010

LOREDANA SCHETTINO, ANTONIO ORIGLIA, GIACOMO MATRONE

Modeling Hesitations. Speech Synthesis Application and Evaluation[1]

Studies have shown that elements like silent pauses, segmental lengthenings, and fillers are naturally involved in the economy of speech and, in specific patterns, may contribute to communication in both human-human and human-machine interactions. Therefore, research on speech synthesis aimed at developing more natural-sounding systems by inserting hesitation phenomena. However, audio issues were found to arise when synthesising filled pauses. Only recently, speech synthesisers based on Deep Neural Networks achieved better performances. In this study, we provide a first perceptual evaluation of a model of occurrence of hesitations (lengthenings, silent pauses as well as fillers) in Italian utterances using a state-of-the-art neural TTS system. A set of experimental stimuli were synthesized and subjected to listeners' evaluations in a discrimination test. Results show that synthetic utterances that include hesitations, according to the linguistic model, are judged as more natural sounding than utterances that do not include any.

Keywords: disfluency, pauses, speech synthesis, Deep Neural Network, perception.

1. *Introduction*

Spontaneous human speech is usually characterised by the occurrence of pauses, fillers, repetitions, corrections, change of planning, various phenomena that seem to alter its fluency and, hence, have been commonly referred to as speech "disfluencies". However, studies on spontaneous speech in different languages have highlighted that the occurrence of disfluency phenomena is not to be considered as exceptional with reference to "normal fluency". Indeed, they report a rate of around 6 to 10 phenomena per 100 words, which suggests that "fluency is the exception, rather than the rule" (Lickley, 2015: 451). Moreover, it has been observed that disfluencies may occur in regular patterns, as they actually serve as tools that the speakers may use to monitor and manage their own speech production by repairing something already uttered, abandoning already started utterance, taking extra-time needed for the planning and construction of the message that is about to be conveyed (Levelt, 1993; Shriberg, 1994).

In particular, speakers can temporarily delay the speech delivery by producing fillers, prolonging speech segments or just being silent. These pauses, prolongations

[1] This article is the result of the collaboration among the authors. However, for academic purposes only, Loredana Schettino is responsible for § 1, 2, 3, 4, 5 and 6 and all the authors are responsible for § 3.1 and 6.

and fillers are also commonly referred to as "hesitation phenomena" (Lickley, 2015). They contribute to communication and can be considered to be beneficial for both speakers and listeners by gaining extra-time for planning as well as for information processing. This claim is corroborated by the fact that these phenomena were also observed to consistently occur in informative speech, e.g., lecturers' or tourist guides' speech (Moniz, Batista, Mata & Trancoso, 2014; Schettino, Betz, Cutugno & Wagner, 2021a). Moreover, studies have shown that hesitation phenomena can bear procedural meaning and convey information about speech planning, structuring, and speakers' disposition (Chafe, 1980; Levelt, 1993; Schegloff, 2010; Tottie, 2016) as listeners learn to exploit the regular occurrence of such phenomena and use it for the interpretation of the ongoing discourse (Corley, Stewart, 2008; Finlayson, Corley, 2012).

These observations on the relevance of hesitation phenomena in communication sparked the interest in developing synthesis systems that were able to insert hesitations in synthesised utterances, in order to obtain more natural-sounding, likeable productions and, eventually, more effective human-machine interactions.

This study integrates the linguistic and computational perspectives while testing the hypothesis that utterances produced using a neural TTS synthesis system that is trained to synthesise disfluencies and where selected phenomena are inserted according to a previously proposed model based on corpus observation are perceived as more natural-sounding and more desirable.

The paper is structured as follows: § 2 provides an overview on previous studies concerning listeners' perception of disfluency phenomena occurring either in spontaneous stimuli or in synthesised ones. Then, in § 3, the approach adopted to evaluate how specific disfluency patterns may affect listeners' perception is described, including the linguistic model, the computational model (the speech synthesiser) and the experimental setting. Finally, the experimental results are presented and discussed in § 4 and § 5.

2. *Disfluency perception and speech synthesis*

Considering that linguistic perception does not necessarily correspond with what has been actually produced but is rather constantly influenced by the communicative context and listeners' selective attention (Levelt, 1993; Voghera, 2017), Lickley (2015) interprets fluency, and disfluency, as a multidimensional concept. The author distinguishes three dimensions that are related to the underlying processes of speech planning, production and perception identified by Levelt (1993) and highlights that speech may be perceived as *fluent* even when containing minor surface disfluencies only detectable on closer inspection of the speech signal, which means that *perception fluency*, does not necessarily imply *planning* and/or *surface fluency*.

In fact, it has been experimentally observed that in disfluency detection tasks, listeners tend to miss out various phenomena (see Collard, 2009 for an overview).

In particular, listeners' perception and awareness of fillers was found to depend on whether they attended to discourse content or style of delivery. Christenfeld (1995) shows that filled and silent pauses negatively affected the perception of the speaker only when listeners' attention was focused on style but not when it was drawn on content. In the latter case, filled pauses tended to be missed. Furthermore, the author observes that filled pauses are perceived as a more «relaxed-sounding» time-buying strategy than silent pauses.

Moreover, it was found that some hesitation phenomena may be perceived as more disruptive than others according to their phonetic features and positioning in sentences and discourse.

Investigating hesitation phenomena in European Portuguese in spontaneous and prepared non-scripted speech, Moniz et al. (2009, 2010) showed that the prosodic phrasing and contour shape exert an influence on participants' ratings of speakers' fluency, defined as *ease of expression*. Lengthenings, filled pauses, and repetitions were most likely rated as *felicitous* when occurring at prosodic breaks and with a flat or ascending pitch contour shape and *infelicitous* when occurring within intonation units or with descending contours.

More recently, Niebuhr and Fischer (2019) investigated the effect of filled pause occurrences on listeners' perception of a speaker's public-speaking and found that shorter and largely nasal filled pause realisations made listeners underestimate their actual number and improved their ratings of the speaker's performance, which was assumed to derive from the lower "saliency" linked to such realisations.

The acknowledgement of the role played by hesitation phenomena in speech challenges rhetoricians' warning against littering speech with them. In fact, it raised the attention of researchers interested in modelling human communicative behaviours to develop speech synthesis systems that would support human-machine interaction systems. So, different state-of-the-art synthesis methods and approaches were implemented to insert hesitations in speech synthesis, most of which focused on the synthesis of filled pauses.

Among the first attempts in this direction is the system developed by Adell, Escudero and Bonafonte (2012) within a rule-based framework. They built a model to generate filled pauses based on the modelling of human fillers prosodic features. A perception rating test conducted to evaluate the system showed that filled pauses introduced with this approach did not increase the degree of listening effort necessary to process the sentences nor decreased their naturalness.

On another account, Dall, Tomalin and Wester (2016) tested the synthesis of filled pauses using HMM-based Speech Synthesis System conducting various evaluation perception experiments. They first found that a voice trained on standard read speech was judged more natural than one trained on spontaneous speech, even when including filled pauses. Hence, the authors tested data-mixing techniques which consisted in combining a synthesis system based on read speech corpora, for the synthesis of general speech, and a system trained on spontaneous speech, for the synthesis of fillers. They observed that this approach together with

obtaining a better phonetic representation of filled pauses improved the overall quality of the synthesis. However, the developed system did not apparently produce satisfying performances.

A HMM-based Speech Synthesis System was also proposed by Betz, Carlmeyer, Wagner and Wrede (2018) who developed and evaluated a model for hesitation insertion in Incremental Spoken Dialogue Systems. The original model included lengthenings, silences and filled pauses, but the perceptual experimental evaluation involved a reduced model without fillers because of the acoustic artefacts produced when synthesising fillers.

Only more recently, Székely, Henter, Beskow and Gustafson (2019) have developed a neural TTS system (Tacotron) trained on a large single-speaker corpus of spontaneous conversational speech. They evaluated the synthesis of filled pauses obtained using models trained on the basis of different types of filled pauses annotation by conducting a pairwise listening test with utterances that both contained filled pauses but were produced using different models: one where the annotation did not account for non-verbal elements so that the system would generate them automatically given a fluent text as input; one where the annotation included two different labels for «uh» and «uhm» instances which allows control on location and type of filled pauses; another one based on an annotation that associated all types of non-verbal vocalisations with one generic label, which only allows to control for their location and was found to provide more natural sounding utterances.

3. *Method*

The evaluation of the way the insertion of disfluency phenomena can affect listeners' perception is no easy task. Common approaches to prepare the experimental stimuli concern repetition tasks and signal manipulations (Fraundorf, Watson, 2011; Mühlack, Elmers, Drenhaus, Trouvain, van Os, Werner, Ryzhova & Möbius, 2021), whereby, however, different issues arise. Repetition tasks consist in asking subjects to listen to a disfluent recording of themselves in a natural setting and to repeat their utterances without disfluencies, which alters the recording setup and lacks in spontaneousness. The second approach consists in intervening on the signal by manually removing/inserting disfluencies, which only involves the disfluent portion of the speech signal thus leaving the immediate prosodic context unchanged and not considering the way it is influenced by the presence of disfluency phenomena. A possible solution to avoid these problems is synthesising experimental stimuli using a system that can be trained to generate utterances, also containing disfluencies, in a highly plausible way.

More specifically, Deep Neural Networks (DNNs) can learn a speaker's intonational patterns both from disfluent and non-disfluent portions of the training data. This allows to generate speech stimuli with and without disfluencies using a probabilistic, trained model generating the most *probable* speech signal that would

correspond to an input text. By providing as input to the synthesiser two versions of the same text with and without annotated disfluencies, the obtained stimuli come from a coherent model of natural speech, unbiased by previous productions and contextually coherent. These can be used to investigate perceptual differences with a solid set of stimuli.

So, to test the effect of disfluency phenomena on listeners' perception, a discrimination test has been conducted whereby subjects were asked to rate stimuli produced using neural synthesis and where disfluency phenomena were inserted according to a linguistic model derived from previous corpus-based observations (Schettino et al. 2021a; Schettino, Betz & Wagner, 2021b). The next sections briefly describe how the synthesis system works (§ 3.1), the previously defined linguistic model (§ 3.2), and the setting of the perception experiment (§ 3.3).

3.1 Computational Model

Neural speech synthesisers represent one of many applications of DNNs and can be trained to replicate the voice of a target speaker or even speaking styles. Previous works have concerned the building of models of the same speaker in different speaking styles (Wang, Stanton, Zhang, Ryan, Battenberg, Shor, Xiao, Jia, Ren & Saurous, 2018). However, it is also possible to target a specific style by training the model using only data representative of that style. In our case, the model has been trained on data that represent the *informative* speaking style.

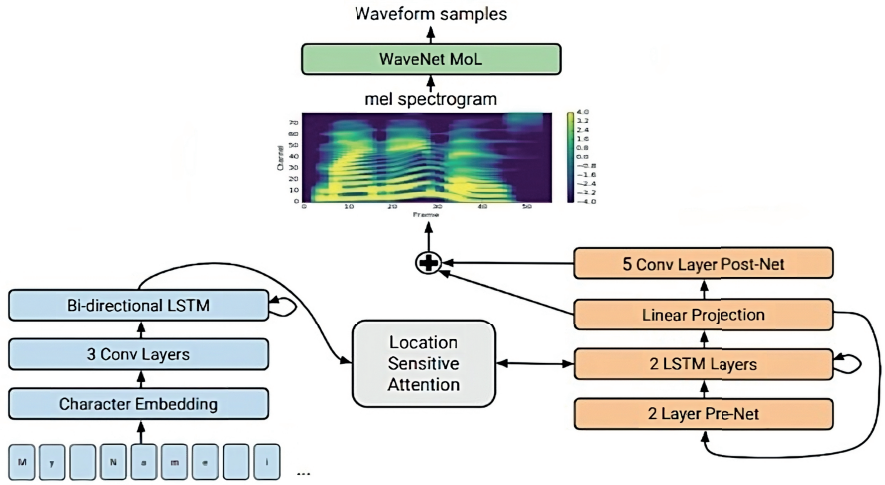
More specifically, the model has been trained on two-and-a-half-hour single-speaker speech extracted from the CHROME corpus (Origlia, Savy, Poggi, Cutugno, Alfano, D'Errico, Vincze & Cataldo, 2018). It consists of Italian semi-spontaneous speech by a female expert guide leading visits at San Martino's Charterhouse and is supplied of orthographic transcriptions (Savy, 2005) and disfluency annotations (Schettino et al., 2021a). In particular, vowel lengthenings, filled pauses, and silent pauses were manually annotated by expert linguists and labelled using grapheme sequences that do not occur anywhere in the corpus other than in correspondence of the considered phenomena:

- Lengthenings (LEN), marked prolongation of segmental material (Betz, Wagner & Eklund, 2017), labelled with repetitions of vocalic sounds for prolongations, i.e., "vv";
- Filled Pauses (FP), non-verbal filler, vocalisations and/or nasalizations, i.e., *eeh, ehm*, annotated using a generic label for both the nasalized and non-nasalized versions of filled pauses, i.e., "ehm";
- Silent Pauses (SP), marked silences perceived as stalling pause in the context of occurrence (Lickley, 2015), labelled using the sequence "hh".

In this study, utterances were synthesised using a state-of-the-art system, namely Tacotron 2 (Shen, Pang, Weiss, Schuster, Jaitly, Yang, Chen, Zhang & Wang, 2018). A network pre-trained on English was fine-tuned on the CHROME transcribed audio material, including disfluency annotations, to generate Italian

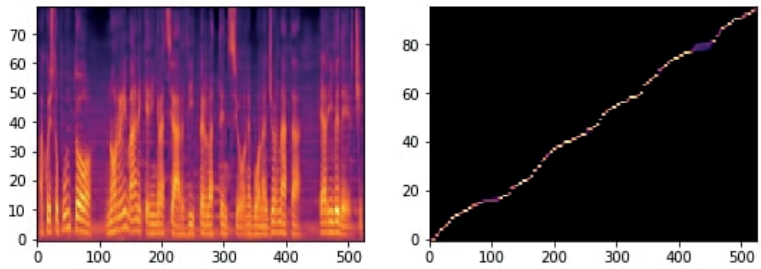
speech spectrograms. These were, then, fed to a Waveglow model (Prenger, Valle & Catanzaro, 2019) to produce the waveform and to apply denoising. At inference time, the model can be used to generate new speech sounds mimicking both voice and speaking style from the example speaker (see Fig. 1).

Figure 1 - *The Tacotron 2 architecture as described in (Shen et al., 2018, p. 4780). In this work, the WaveNet model is replaced by the more recent Waveglow model*



The synthetic spectrogram is accompanied by the corresponding *alignment graph* (Fig. 2) that represents to what extent the network decoder was able to select the correct states, among all the possible ones, making reference to frames in the training audio. The *hotter* the pixel, the higher the attention given to a certain state. Ideally, a diagonal line of *hot* pixels indicates that the decoder focused on the correct states from the encoder to generate the mel spectrum.

Figure 2 - *A synthetic spectrogram (left) together with its alignment graph (right). A good alignment between the encoder vectors (y axis) and the decoder steps (x axis) is represented by a line that tends to a diagonal*



Summarising, DNNs produce the most probable speech output given a character sequence. So, the same text can be submitted, with and without disfluencies, to the same network that is able to generate productions that are not influenced by what has been synthesised before, as would happen with humans. Also, the whole utterance prosodic representation is coherent with the presence or absence of the disfluency phenomena, as opposed to interventions with manual cuts.

3.2 Linguistic Model

The described machine learning solution allows the generation of stimuli where hesitation phenomena are produced after specification of their location in the text and, given their context of occurrence, the synthesis system computes their surface realisation.

This work concerns the perceptual evaluation of the following patterns of hesitations occurrence that emerged from previous corpus analyses:

- Lengthenings are placed: a) before semantically *heavy*, key constituents (*focusing function*, Schettino et al. 2021a); b) toward the end of the clause (Schettino et al. 2021b); c) on content or functional words (following the distribution found in the dataset, i.e., content words: 51%, function word: 49%);
- Silent or Filled Pauses are placed between two clauses (*structuring function*, Schettino et al. 2021a, b). Half of the stimuli also contained a Filled Pause and the other half a Silent Pause.

In the following two utterances of examples, hesitations phenomena are placed according to the just described patterns:

(1) *"Nella prima metà del diciottesimo secolo i lavori passarono a**aa** Nicola Tagliacozzi Canale **ehm** che farà rifare gli**ii** spazi del priore."*

"In the first half of the Eighteenth century the work was handed to[**LEN**] Nicola Tagliacozzi Canale [**FP**] who will redo the[**LEN**] prior's place."

(2) *"La certosa fu inaugurata e consacrata**aa** nel 1368 **hh** seppur i certosini avevano preso possesso del monastero**oo** dal 1337"*

"The Cartherhouse was inaugurated and consecrated[**LEN**] in 1368 [**SP**] although the Carthusians had taken possession of the monastery[**LEN**] since 1337."

3.3 Experimental Evaluation

3.3.1 Experimental Setting

The evaluation of synthesised utterances commonly involves judgements of their perceived *naturalness*. However, Wagner, Beskow, Betz, Edlund, Gustafson, Henter, Le Maguer, Malisz, Székely, Tånnander & Voße (2019) highlight that naturalness is not an inherent property of speech, but is specified with reference to the communicative context of application. Therefore, the evaluation of synthesis

systems should refer to the principle of *contextual appropriateness* given a specific situation or application.

In this study, the context of application for the evaluation of the *disfluent* vs. *non-disfluent* synthetic utterances is to provide voice to a Virtual Avatar, i.e., Embodied Conversational Agent, serving visitors in museums. Given this context, the perception experiment consists in explicitly asking the listeners whether the system meets the estimated needs of *naturalness* and *appropriateness* with reference to the envisioned application.

More specifically, participants were subjected to a pairwise listening test where they were asked to listen to pairs of synthetic utterances and then select the one that sounded more natural to them (*naturalness*), and the one they would choose to give voice to a Virtual Avatar serving in museums like the San Martino Charterhouse in Naples (*appropriateness*).

The set of stimuli is composed by complex phrases, meaning a main clause and a dependent (mostly relative) clause, which describe point of interest of the Charterhouse and comprised 10 target pairs of utterances, one with hesitations (*Disf* condition) and one without any (*no_Disf* condition) and 10 filler couples paired as *Disf-Disf* and *no_Disf-no_Disf*. These stimuli were presented to participants in randomised order.

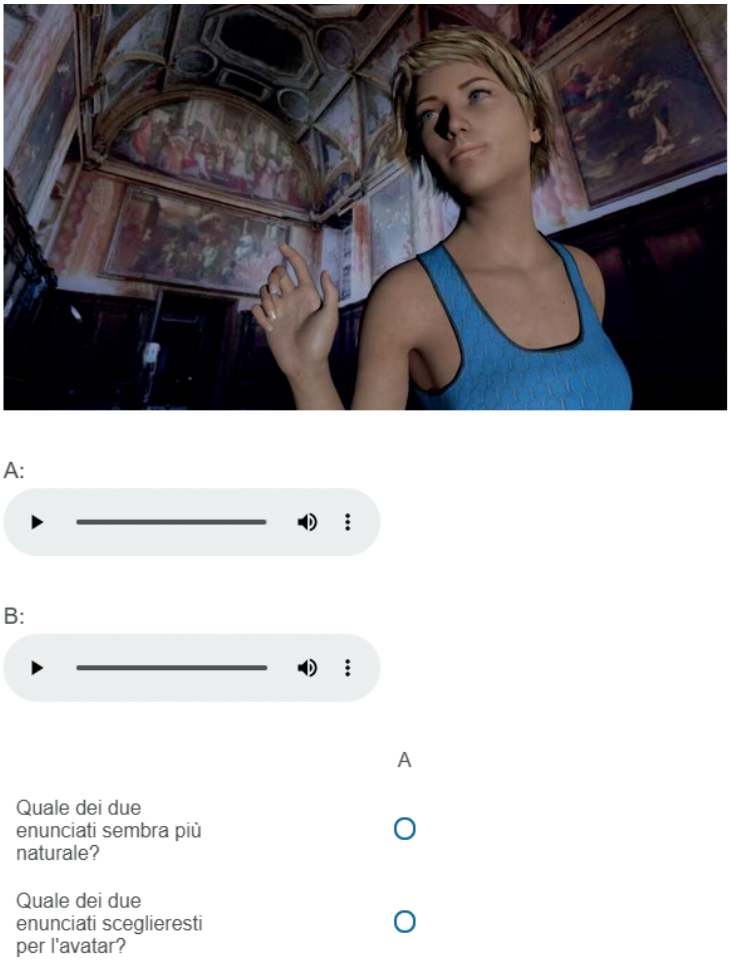
The test was set up and distributed on social media channels for university students using the QUALTRICS software for online surveys.[2]

Participants were asked to fill in a sociolinguistic questionnaire collecting information concerning the age, sex, country and city where they spent most of their life, whether they regularly listen to synthetic voices such as Siri or Cortana. Then, they were asked to make sure they were in a quiet closed area and wearing headphones throughout the experiment duration (approximately 15 minutes, including an initial training phase).

A picture of the CHROME avatar "Maya" (Origlia et al., 2018) was enclosed to each question to provide graphical support to the contextualization (Fig. 3).

[2] Version [2021] of Qualtrics – www.qualtrics.com

Figure 3 - *Example of a question of the Discrimination Task as visible by participants on Qualtrics*



3.3.2 Statistical Analysis

The statistical analysis is conducted using the R software (R Core Team 2021). A Generalised Linear Mixed Model ("lme4" package, Bates, Mächler, Bolker & Walker, 2015) is built including subjects' responses, i.e., the condition chosen between *Disf* and *no_Disf*, as dependent variable; the question, i.e., *naturalness* or *appropriateness*, and type of phenomena occurring in the disfluent stimuli, i.e., *LEN_FP* or *LEN_SP* as interacting independent variables and, to control for individual variability, participants as random effect. Sociolinguistic variables are also controlled considering sex, age, and familiarity with synthetic voices ("yes" or "no") as independent variables.

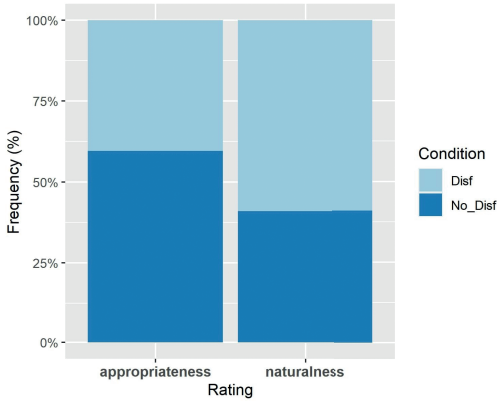
Post-hoc analyses are conducted to inspect the levels within the main effects and the interactions using pairwise comparisons (emmeans package, Lenth, Singmann, Love, Buerkner & Herve, 2018). P-values were calculated using Tukey's HSD adjustment.

4. *Discrimination Test Results*

The experiment was conducted with 22 participants (7 female, 15 male) with age ranging between 22 and 50 years (M = 30, StDev = 7). Among these participants, only 6 (27%) reported to regularly listen to synthetic voices such as Siri or Cortana.

As reported in Table 1, the statistical analysis yields significant results. Synthesised disfluent utterances are significantly more often judged as more natural. Conversely, non-disfluent utterances are significantly more frequently selected to give voice to a virtual agent, than disfluent ones (see Fig. 4).

Figure 4 - *Frequency (%) of appropriateness and naturalness ratings per condition*



Also, this effect is only significant when considering ratings of utterances that contain filled pauses and represents just a trend for judgements related to the utterances with silent pauses.

Table 1 - *Pairwise comparison among the levels of the question variable in interaction with the LEN_FP and LEN_SP stimuli groups*

Stimuli	*Contrast*	*Estimate*	*Std. Error*	*z value*	*p value*
LEN_FP, LEN_SP	*appropriateness – naturalness*	0.809	0.202	4.002	0.0001
LEN_FP	*appropriateness – naturalness*	1.270	0.292	4.353	<.0001
LEN_SP	*appropriateness – naturalness*	0.349	0.279	1.249	0.2115

5. *Discussion*

The participants' responses to the discrimination test attest an inverse direction for naturalness and appropriateness judgements. The insertion of hesitation phenomena according to the linguistic model significantly affects the listeners' perception of the synthesised utterances in that they are perceived as more natural sounding than

non-disfluent utterances, but less luckily to be associated with a virtual avatar. These opposed tendencies may reflect the fact that people would not expect a *machine* to produce spontaneous physiologic (*natural*) elements such as disfluencies so that synthetic utterances containing disfluencies are not (yet) customarily associated with a virtual system, despite being rated as more natural sounding.

In fact, while testing suitable voices for robots, studies have found that a robotic voice is often preferred over human-like voices (Hönemann, Wagner, 2015; Wagner et al., 2019). According to Moore (2017) using human-like voices for artefacts might lead users to overestimate their abilities. Hence, intelligible but robotic voices could be considered more *appropriate* and better systems to manage the users' expectations of *conversational* artefacts, like *Google Now* or *Amazon Echo*. On the other hand, Rodero (2017) showed that human voices perform better in narrative tasks, such as telling an advertising story, as they are rated as more effective and enhance listeners' attention and recall. Based on this picture, synthesis systems that are able to generate human-like voices by reproducing plausible prosodic realisations including hesitation phenomena could provide effective and desirable voices for informative speech.

Looking more closely at the results, the observed effect of the insertion of hesitations on participants achieve significant values only for stimuli pairs where the disfluent utterance contains a filled pause (*LEN_FP*), whereas for those where the disfluent utterance contains a silent pause (*LEN_SP*), a weak trend emerges. This may suggest that filled pauses, being a voiced element that is perceptually independent from the other sounds in the speech chain, are more evident phenomena within the immediate prosodic surroundings and are more likely to be detected as speech planning devices, unlike lengthenings and silent pauses which may be considered as more subtle phenomena. Hence, disfluent *LEN_FP* utterances would stand out more clearly from non-disfluent ones with respect to disfluent *LEN_SP* utterances. The latter, instead, seem to be perceived as less markedly different than non-disfluent utterances, which would hamper the emergence of clear-cut preferences for selecting one type of utterances over the other (disfluent vs. non-disfluent).

6. *Conclusions*

The study described in this article has been conducted to provide a perceptual evaluation of previously observed patterns of occurrence of hesitation phenomena, i.e., silences, fillers, lengthening, in informative speech (Schettino et al., 2021a, b).

The preparation of the sets of experimental stimuli has been supported by a computational model of speech. More specifically, a neural synthesis system has been trained to generate utterances including hesitations in a contextually plausible way. Then, a discrimination test was designed to evaluate how lengthenings, filled pauses and silent pauses inserted according to the corpus-based model can affect the listeners' perception of the synthesized utterances.

The main results of the pairwise listening test highlight that disfluent utterances, especially when containing filled pauses, are perceived as more natural sounding, though less appropriate to the specific application in supporting virtual avatars serving in museums. However, Wagner and colleagues (2019) suggest that an accurate evaluation of synthesis, beside being based on participants' subjective ratings, should also consider behavioural assessment, which consists in the indirect evaluation of the users' comprehension and preferences while fulfilling a task. Therefore, a follow-up test has been designed to integrate the subjective evaluation with a behavioural one involving a task more closely related to the specific application, that is to give a voice for a virtual agent designed to serve visitors in cultural sites by showing relevant points of interest.

More generally, the study provides first evidence that modern technologies, such as neural synthesis systems, being able to produce highly plausible and, to a certain extent, controllable stimuli, may represent valuable tools for testing relevant hypothesis of linguistic, and phonetic, interest, especially when concerning speech phenomena that are difficult to elicit in natural settings such as disfluency phenomena (Malisz, Henter, Valentini-Botinhao, Watts, Beskow & Gustafson, 2019).

Acknowledgements

Funding: This work was supported by the Italian National Project PRIN Cultural Heritage Resources Orienting Multimodal Experiences (CHROME) [grant number B52F15000450001].

References

Adell, J., Escudero, D. & Bonafonte, A. (2012). Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. In *Speech Communication*, 54, 459-476. doi:https://doi.org/10.1016/j.specom.2011.10.010.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. In *Journal of Statistical Software*, 67, 1-48. doi:https://doi.org/10.18637/jss.v067.i01.

Betz, S., Carlmeyer, B., Wagner, P. & Wrede, B. (2018). Interactive hesitation synthesis: modelling and evaluation. In *Multimodal Technologies and Interaction*, 2 (1), 9. doi:https://doi.org/10.3390/mti2010009.

Betz, S., Eklund, R. & Wagner, P. (2017). Prolongation in German. In Rose, R., Eklund, R. (Eds.), *Proceedings of the 8th Workshop on Disfluency in Spontaneous Speech*, Stockholm, Sweden, 18–19 August 2017, 13-16.

Christenfeld, N. (1995). Does it hurt to say um? In *Journal of Nonverbal Behavior*, 19, 171-186. doi:https://doi.org/10.1007/BF02175503.

Collard, P. (2009). Disfluency and listeners' attention: An investigation of the immediate and lasting effects of hesitations in speech. Ph.D dissertation, The University of Edinburgh.

Corley, M., Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. In *Language and Linguistics Compass*, 2, 589-602. doi:https://doi.org/10.1111/j.1749-818X.2008.00068.x.

Dall, R., Tomalin, M. & Wester, M. (2016). Synthesising filled pauses: Representation and datamixing. In Bonafonte, A, Prahallad, K (Eds.), *Proceedings of 9th Speech Synthesis Workshop*, Sunnyvale, California, USA, 13–15 Septembre 2016, 7-13. doi:https://doi.org/10.21437/SSW.2016-2.

Finlayson, I. R., Corley, M. (2012). Disfluency in dialogue: An intentional signal from the speaker? In *Psychonomic bulletin & review*, 19, 921-928. doi:https://doi.org/10.3758/s13423-012-0279-x.

Fraundorf, S. H., Watson, D. G. (2011). The disfluent discourse: Effects of filled pauses on recall. In *Journal of memory and language*, 65, 161-175. doi:https://doi.org/10.1016/j.jml.2011.03.004.

Hönemann, A., Wagner, P. (2015). Adaptive Speech Synthesis in a Cognitive Robotic Service Apartment: An Overview and First Steps Towards Voice Selection. In *Tagungsband Elektronische Sprachsignalverarbeitung ESSV 2015*, 135-142.

Lenth, R., Singmann, H., Love, J., Buerkner, P. & Herve, M. (2018). Emmeans: Estimated marginal means, aka least-squares means. R package version, 1 (1), 1-97.

Levelt, W. J. (1993). *Speaking: From intention to articulation*. Cambridge/London: MIT press. doi:https://doi.org/10.7551/mitpress/6393.001.0001.

Lickley, R. J. (2015). Fluency and disfluency. In Redford M. A. (Ed.), *The handbook of speech production*. Chichester: Wiley Online Library, 445-474. doi:https://doi.org/10.1002/9781118584156.ch20.

Malisz, Z., Henter, G. E., Valentini-Botinhao, C., Watts, O., Beskow, J. & Gustafson, J. (2019). Modern speech synthesis for phonetic sciences: A discussion and an evaluation. In Calhoun, S., Escudero, P., Tabain, M. & Warren, P. (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences*. Melbourne, Australia, 487-491. doi:https://doi.org/10.31234/osf.io/dxvhc.

Moniz, H., Batista, F., Mata, A. I. & Trancoso, I. (2014). Speaking style effects in the production of disfluencies. In *Speech Communication*, 65, 20-35. doi:https://doi.org/10.1016/j.specom.2014.05.004.

Moniz, H., Trancoso, I. & Mata, A. I. (2009). Classification of disfluent phenomena as fluent communicative devices in specific prosodic contexts. In *Proceedings of Interspeech 2009*. Brighton, United Kingdom, 6-10 September 2009. doi:https://doi.org/10.21437/Interspeech.2009-518.

Moniz, H., Trancoso, I. & Mata, A. I. (2010). Disfluencies and the perspective of prosodic fluency. In Esposito, A., Campbell, N., Vogel, C., Hussain, A. & Nijholt, A. (Eds.), *Development of multimodal interfaces: active listening and synchrony*. Berlin, Heidelberg: Springer, 382–396. doi:https://doi.org/10.1007/978-3-642-12397-9_33.

Moore, R. K. (2017). Appropriate voices for artefacts: some key insights. In Dassow, A., Marxer, R. & Moore, R., K. (Eds.), *Proceedings of the 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*. Skövde, Sweden, 25-26 August 2017, 7-11. doi:https://doi.org/10.3389/frobt.2016.00061.

Mühlack, B., Elmers, M., Drenhaus, H., Trouvain, J., van Os, M., Werner, R., Ryzhova, M., & Möbius, B. (2021). Revisiting recall effects of filler particles in German and English. In *Proceedings of Interspeech 2021.* Brno, Czechia, 30 August / 3 September 2021, 2021-1056. doi:https://doi.org/10.21437/Interspeech.

Niebuhr, O., & Fischer, K. (2019). Do not hesitate!-unless you do it shortly or nasally: How the phonetics of filled pauses determine their subjective frequency and perceived speaker performance. In *Proceedings of Interspeech 2019,* 15-19 September 2019 Graz, Austria, 2019-1194. doi:https://doi.org/10.21437/Interspeech.

Origlia, A., Savy, R., Poggi, I., Cutugno, F., Alfano, I., D'Errico, F., Vincze, L., & Cataldo, V. (2018). An audiovisual corpus of guided tours in cultural sites: Data collection protocols in the CHROME Project. In *Proceedings of the AVI-CH Workshop on Advanced Visual Interfaces for Cultural Heritage (vol. 2091)*. Grosseto, Italy, 1-4.

Prenger, R., Valle, R., & Catanzaro, B. (2019). Waveglow: A flowbased generative network for speech synthesis. In *Proceedings of the 44th International Conference on Acoustics, Speech and Signal Processing*. Brighton, United Kingdom, 12-17 May 2019, 3617-3621. doi:https://doi.org/10.1109/ICASSP.2019.8683143.

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. URL: https://www.R-project.org/.

Rodero, E. (2017). Effectiveness, attention, and recall of human and artificial voices in an advertising story. Prosody influence and functions of voices. In *Computers in Human Behavior*, 77, 336–346. doi:https://doi.org/10.1016/j.chb.2017.08.044.

Savy, R. (2005). Specifiche per la trascrizione ortografica annotata dei testi. In Albano Leoni, F., Giordano, R. (Eds.), Italianoparlato. Analisi di un dialogo. Napoli: Liguori, 1-37.

Schettino, L., Betz, S., Cutugno, F., & Wagner, P. (2021a). Hesitations and individual variability in Italian tourist guides' speech. In Bernardasci, C., Dipino, D., Garassino, D., Negrinelli, S., Pellegrino, E., & Schmid, S. (Eds.), *Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications, STUDI AISV 8*. Milano: Officinaventuno, 243-262.

Schettino, L., Betz, S., & Wagner, P. (2021b). Hesitations distribution in Italian discourse. In *Proceedings of the 10th Workshop on Disfluency in Spontaneous Speech*. Paris, France, 25-27 August 2021, 29-34.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, A.R., Agiomyrgiannakis, Y., & Wu, Y. (2018). Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. In *Proceedings of the 43th International Conference on Acoustics, Speech and Signal Processing*. Calgary, Canada, 15-20 April 2018, 4779-4783. doi:https://doi.org/10.1109/ICASSP.2018.8461368.

Shriberg, E. E. (1994). Preliminaries to a theory of speech disfluencies. PhD dissertation. University of California.

Student (1908). The probable error of a mean. In *Biometrika*, 6, 1-25. doi:https://doi.org/10.2307/2331554.

Székely, É., Henter, G. E., Beskow, J., & Gustafson, J. (2019). How to train your fillers: uh and um in spontaneous speech synthesis. In *Proceedings of the 10th Speech*

Synthesis Workshop. Vienna, Austria, 20-22 September 2019, 245–250. doi:https://doi.org/10.21437/SSW.2019-44.

Voghera, M. (2017). *Dal parlato alla grammatica*. Roma: Carocci.

Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., Eje Henter, G., Le Maguer, S., Malisz, Z., Székely, É., Tånnander, A., Vosse, J. (2019). Speech synthesis evaluation—state-of-the-art assessment and suggestion for a novel research program. In *Proceedings of the 10th Speech Synthesis Workshop*. Vienna, Austria, 20-22 September 2019, 105-110. doi:https://doi.org/10.21437/SSW.2019-19.

Wang, Y., Stanton, D., Zhang, Y., Ryan, R.-S., Battenberg, E., Shor, J., Xiao, Y., Jia, Y., Ren, F., & Saurous, R. A. (2018). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In Dy, J., Krause, A. (Eds.), *International Conference on Machine Learning*. Stockholm, Sweden, 10-15 July 2018, 5180-5189.

Appendix 1

Target Stimuli

Disf_No Disf

LEN SP

1. La Certosa di San Martino costituisce in assoluto uno deiii maggiori complessi monumentali religiosi della città hh che è fra i più riusciti esempi diii architettura e arte barocca.
2. La certosa fu inaugurata e consacrataaa nel mille trecento sessantotto hh seppur i certosini avevano preso possesso del monasterooo dal mille trecento trentasette.
3. Il pavimento fu realizzato da Bonaventura Prestiii in preziosi marmi di diversi colori hh che produconooo un'apparente tridimensionalità.
4. La seconda cappella a sinistra della navata è quella diii San Bruno hh le cui decorazioni marmoree sonooo del Fanzago.
5. La seconda cappella di destra è quellaaa di San Giovanni Battista hh che fu decorata dal Fanzagooo nel mille seicento trentuno.

LEN FP

1. All'inizio del Seicento la direzione del cantiere passa aaa Giovan Giacomo di Conforto ehm che completaaa il progetto del Dosio.
2. In questa fase di ristrutturazione del complesso lavoraronooo pittori ehm che furono fra i più grandi artistiii del Seicento.
3. I lavori vennero affidati aaa Giovanni Antonio Dosio ehm che fu di fatto il primo responsabile delleee trasformazioni del complesso.
4. La facciata della chiesa trecentesca fu rimaneggiata sul finire del Cinquecentooo dal Dosio ehm a cui si deve il pronaooo a tre arcate.
5. Nella prima metà del diciottesimo secolo i lavori passarono aaa Nicola Tagliacozzi Canale ehm che farà rifare gliii spazi del priore.

No Disf_No Disf

1. Cronologicamente la Certosa di San Martino è la seconda certosa della Campania essendo nata diciannove anni dopo quella di San Lorenzo a Padula.
2. Le transenne di tutte le cappelle sonoo del Fanzago a cui si devono anche i festoni di frutta sui pilastri.
3. Nel registro inferiore della sala sono collocati alle pareti arredi mobiliari intarsiati i cui intagliatori furono Nunzio Ferraro e Giovan Battista Vigilante.
4. La chiesa delle donne era destinata ad uso esclusivo delle donne alle quali era proibito l'accesso alla certosa.
5. Le esecuzioni marmoree interne sono frutto dell'opera di Cosimo Fanzago che fu chiamato a ristrutturare la certosa dal mille seicento ventitrè al mille seicento cinquantasei.

Disf_Disf

1. Solo verso la seconda metà del Sedicesimo secolo il complesso fu dedicato aaa Martino di Tours hh probabilmente per la presenza nel luogo diii un'antica cappella preesistente a lui dedicata.
2. La terza cappella di sinistra è quellaaa dell'Assunta hh la quale presenta unaaa decorazione seicentesca.
3. Sul piazzale esterno al complesso certosino ehm è defilata sulla sinistra laaa chiesa delle Donne ehm che è opera diii Giovanni Antonio Dosio.
4. La facciata della chiesa trecentesca fu rimaneggiata successivamente daaa Cosimo Fanzago ehm che costruì nella prima metà del Seicento unaaa serliana.
5. La chiesa si compone di una navata unica e delleee cappelle laterali ehm che si succedono aiii lati della zona absidale.

Appendix 2

Test link:
https://phdmglunina.fra1.qualtrics.com/jfe/form/SV 6yAWNyk5xCDMHz0

Autori

Iolanda Alfano – Dipartimento di Studi Umanistici, Università di Salerno, Italia
ialfano@unisa.it; https://orcid.org/0000-0001-6141-327X

Cinzia Avesani – Istituto di Scienze e Tecnologie della Cognizione, sede di Padova, Consiglio Nazionale delle Ricerche (CNR), Italia
cinzia.avesani@cnr.it; https://orcid.org/0000-0001-9911-1189

Serena Bonifacio – ricercatrice indipendente
serena.bonifacio@gmail.com

Giulia Calignano – Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Università di Padova, Italia
calignano.giulia@gmail.com; https://orcid.org/0000-0002-2913-8770

Francesco Cangemi – Istituto di Linguistica, Università di Colonia, Germania
fcangemi@uni-koeln.de; https://orcid.org/0000-0003-1016-5178

Violetta Cataldo – Dipartimento di Studi Umanistici, Università di Salerno, Italia. Dipartimento di Linguistica, Università di Gand, Belgio
vcataldo@unisa.it; https://orcid.org/0000-0002-3915-1143

Claudia Crocco – Dipartimento di Linguistica, Università di Gand, Belgio
claudia.crocco@ugent.be; https://orcid.org/0000-0003-1099-956X

Valeria D'Aloia – Università degli Studi "G. d'Annunzio"Chieti-Pescara
valeria.daloia@studio.unibo.it

Sonia D'Apolito – Dipartimento di Studi Umanistici, Università del Salento, Italia
sonia.dapolito@unisalento.it; https://orcid.org/0009-0002-5593-5076

Dalila Dipino – Seminario di Romanistica, Università di Zurigo, Svizzera
dalila.dipino@gmail.com; https://orcid.org/0000-0002-9591-2111

Margherita Di Salvo – Dipartimento di Studi Umanistici, Università degli Studi di Napoli Federico II, Italia
margherita.disalvo@unina.it; https://orcid.org/0000-0002-1341-139X

Davide Garassino – Istituto di traduzione e interpretazione, Università di Scienze Applicate di Zurigo, Svizzera
davide.garassino@zhaw.ch; https://orcid.org/0000-0002-6224-4065

Barbara Gili Fivela – Dipartimento di Studi Umanistici, Università del Salento, Italia
barbara.gilifivela@unisalento.it; https://orcid.org/0000-0002-4694-8652

Martine Grice – Istituto di Linguistica, Universita di Colonia, Germania
martine.grice@uni-koeln.de; https://orcid.org/0000-0003-4973-4059

Giovanni Leo – Dipartimento di Studi Umanistici, Università di Salerno, Italia. Dipartimento di Linguistica, Università di Gand, Belgio
gioleo95y@gmail.com; https://orcid.org/0000-0002-9033-1131

Marta Maffia – Dipartimento di Studi Letterari, Linguistici e Comparati, Università di Napoli L'Orientale, Italia
mmaffia@unior.it; https://orcid.org/0000-0002-4913-374X

Giuseppe Magistro – Dipartimento di Linguistica, Università di Gand, Belgio
giuseppe.magistro@ugent.be; https://orcid.org/0000-0002-0272-741X

Giacomo Matrone – Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione, Università degli Studi di Napoli Federico II, Italia
giacomo.matrone@unina.it; https://orcid.org/0009-0001-5318-7249

Francesco Olivucci – ricercatore indipendente
francesco.olivucci@gmail.com

Antonio Origlia – Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione, Università degli Studi di Napoli Federico II, Italia
antonio.origlia@unina.it; https://orcid.org/0000-0002-8635-1623

Riccardo Orrico – Centro Interdipartimentale di Ricerca Urban/Eco, Università degli Studi di Napoli Federico II, Italia. Dipartimento di Lingue e Culture Moderne, Università Radboud, Paesi Bassi
riccardo.orrico@ru.nl; https://orcid.org/0000-0001-9260-7210

Anna Chiara Pagliaro – Dipartimento di Studi Umanistici, Università del Salento, Italia
annachiara.pagliaro@unisalento.it; https://orcid.org/0009-0001-7332-7587

Massimo Pettorino – ricercatore indipendente
mpettorino@gmail.com; https://orcid.org/0000-0001-5521-6536

Martina Rossi – Dipartimento di Linguistica e Fonetica, Università di Kiel Christian Albrechts, Germania
mrossi@isfas.uni-kiel.de; https://orcid.org/0000-0001-5970-3366

Simona Sbranna – Istituto di Linguistica, Universita di Colonia, Germania
s.sbranna@outlook.com; https://orcid.org/0000-0001-6915-7047

Loredana Schettino – Centro Interdipartimentale di Ricerca Urban/Eco, Università degli Studi di Napoli Federico II, Italia. Dipartimento di Studi Umanistici, Università degli Studi di Salerno, Italia
loredana.schettino@unina.it; https://orcid.org/0000-0002-3788-3754

Stephan Schmid – Laboratorio di Fonetica, Università di Zurigo, Svizzera
stephan.schmid@uzh.ch; https://orcid.org/0000-0002-5937-5427

Mario Vayra – Alma Mater Studiorum -Università di Bologna, Italia
mario.vayra@unibo.it; https://orcid.org/0000-0002-6198-2437

Simon Wehrle – Istituto di Linguistica, Universita di Colonia, Germania
simon.wehrle@uni-koeln.de; https://orcid.org/0000-0001-9715-9541

Claudio Zmarich – Istituto di Scienze e Tecnologie della Cognizione, sede di Padova, Consiglio Nazionale delle Ricerche (CNR), Italia
claudio.zmarich@cnr.it; https://orcid.org/0000-0003-0384-4041

Studi AISV è una collana di volumi collettanei e monografie dedicati alla dimensione sonora del linguaggio e alle diverse interfacce con le altre componenti della grammatica e col discorso. La collana, programmaticamente interdisciplinare, è aperta a molteplici punti di vista e argomenti sul linguaggio: dall'attenzione per la struttura sonora alla variazione sociofonetica e al mutamento storico, dai disturbi della parola alle basi cognitive e neurobiologiche delle rappresentazione fonologiche alle applicazioni tecnologiche. I testi sono selezionati attraverso un processo di revisione anonima fra pari e vengono pubblicati nel sito dell'Associazione Italiana di Scienze della Voce con accesso libero a tutti gli interessati.

Riccardo Orrico è ricercatore post dottorato presso la Radboud University di Nimega (Paesi Bassi) dove lavora principalmente sulla percezione dell'intonazione. Precedentemente è stato assegnista di ricerca presso l'Università degli studi di Napoli Federico II. I suoi interessi di ricerca includono la produzione e la percezione dell'intonazione, gli aspetti semantici e pragmatici dell'intonazione e la variabilità individuale.

Loredana Schettino è assegnista di ricerca post dottorato presso l'Università degli Studi di Napoli Federico II dove è anche docente a contratto per il corso di Linguistica Generale. Si occupa principalmente del valore comunicativo di fenomeni caratteristici del parlato spontaneo quali pause, autocorrezioni e la variazione nel grado di specificazione fonica di unità linguistiche con attenzione alle relative implementazioni nelle tecnologie della voce.

AISV - Associazione Italiana Scienze della Voce
sito: www.aisv.it
email: aisv@aisv.it| redazione@aisv.it
ISBN: 978-88-97657-63-7

Edizione realizzata da
Officinaventuno
info@officinaventuno.com | sito: www.officinaventuno.com
via F.lli Bazzaro, 18 - 20128 Milano - Italy