Sguardi sul Giappone
ISBN 978-88-7543-485-4
DOI 10.48231/97888754348544
pp. 53-69 (settembre 2020)

L'uso di *corpora* orali per la ricerca sociolinguistica. Uno studio sulla variazione delle pause non silenti nel giapponese spontaneo

Giuseppe Pappalardo

Introduzione

La sociolinguistica, intesa come studio della variazione linguistica secondo vari parametri sociali, è un campo di ricerca empirico che fa generalmente uso di dati raccolti *ad hoc* per l'indagine da condurre. Tuttavia, i dati linguistici raccolti non sono sempre sufficienti per indagini di tipo statistico e quantitativo e i risultati ottenuti non possono dunque essere estesi a tutta la popolazione di riferimento. La creazione di banche dati di ampie dimensioni, soprattutto per la lingua inglese scritta, ha determinato negli ultimi decenni un forte interesse per la linguistica dei *corpora* anche da parte dei sociolinguisti (McEnery, 2001). Per essere utilizzato nella ricerca sociolinguistica, il *corpus* deve però essere stato etichettato e annotato secondo parametri sociali, come l'età, il genere e la classe sociale di chi parla o scrive. Il *Corpus of Spontaneous Japanese* (CSJ) è un *corpus* orale di giapponese spontaneo, etichettato con numerose informazioni sul parlante e sulla tipologia di parlato e consente quindi di effettuare analisi sulla variazione sociolinguistica nel giapponese contemporaneo (Maekawa, 2002). Il presente contributo si propone di illustrare l'utilizzo di questo *corpus* orale per indagini di tipo sociolinguistico presentando i risultati di un *case study* sulla variazione delle pause non silenti nel giapponese spontaneo. Le pause non silenti, contrapposte alle pause silenti, sono fenomeni macroprosodici del parlato spontaneo e svolgono numerose funzioni prosodiche e pragmatiche. Attraverso l'interrogazione del *corpus* sarà analizzata la variazione diagenerica, diacronica, diafasica e diastratica nella scelta della tipologia di pausa non silente (pause lessicali, nasalizzazioni, vocalizzazioni, ecc.). Saranno inoltre presentati i risultati di indagini quantitative come l'indice di disfluenza e il rapporto tra la tipologia di pausa e la lunghezza della frase successiva.

Sociolinguistica e linguistica dei *corpora*

Ogni lingua è varia, muta e si differenzia a ogni uso che ne fa un parlante. Noi spesso non pensiamo ai differenti modi con cui utilizziamo parole e frasi par-

lando o scrivendo. Sebbene abbiamo delle preferenze e abbiamo costruito un nostro modo originale di articolare le nostre espressioni, la lingua che usiamo tutti i giorni cambia. Quali sono le cause di questa variazione? È noto che le lingue naturali sono soggette a continui e inevitabili mutamenti dovuti non solo a fattori interni al sistema linguistico ma anche a fattori extra-linguistici relativi alle caratteristiche del parlante, come il luogo d'origine, l'età e il livello culturale, e alla situazione comunicativa, come il grado di formalità e il mezzo di comunicazione (Labov, 1963). La variazione linguistica secondo parametri extra-linguistici è l'oggetto di studio della sociolinguistica.

Il termine sociolinguistica viene spesso usato nella letteratura scientifica come termine ombrello che copre un ambito interdisciplinare molto vasto, che comprende tutti gli studi che mettono in relazione lingua e società. Coulmas (1996, p. 2), per esempio, lo usa con questa accezione e distingue al suo interno due grandi branche: la *micro-sociolinguistica* o sociolinguistica in senso stretto e la *macro-sociolinguistica* o sociologia del linguaggio. La prima viene descritta come lo studio degli usi della lingua influenzati dalla struttura sociale, mentre la seconda ha come oggetto di studio i fenomeni sociali correlati al linguaggio. Si tratta di una distinzione che ricorre spesso nei manuali teorici che tentano di delimitare gli ambiti di ricerca della disciplina. Hudson (1980, p. 5), per esempio, definisce la sociolinguistica come 'the study of language in relation to society' e la sociologia del linguaggio come, viceversa, 'the study of society in relation to language'. Si tratta di una distinzione necessaria data dalla diversa formazione degli studiosi e da una diversa metodologia di indagine: mentre la sociolinguistica in senso stretto può essere considerata a tutti gli effetti una branca delle scienze del linguaggio, i cui metodi di analisi sono quelli propri della ricerca linguistica, la sociologia del linguaggio appartiene piuttosto alle scienze della società e il sociologo che se ne occupa osserva e descrive con un approccio di tipo sociologico fenomeni sociali strettamente legati al linguaggio. Se nella letteratura di lingua inglese *sociolinguistics* è ormai diventata una denominazione assai generica, al punto che per indicare la sociolinguistica in senso stretto bisogna ricorrere all'espressione *variationist sociolinguistics* (Tagliamonte, 2012), negli studi in lingua italiana si tende a mantenere la distinzione tra ricerche di tipo linguistico e quelle di tipo sociologico, indicate rispettivamente da sociolinguistica e sociologia del linguaggio (o delle lingue) (Berruto, 1995). In questo contributo, seguendo la prassi degli studi in lingua italiana, il termine sociolinguistica sarà usato in senso stretto, per indicare quindi, senza utilizzo di ulteriori specificazioni, lo studio empirico-descrittivo della variazione linguistica secondo parametri sociali.

La sociolinguistica, così come la linguistica storica e la dialettologia, è un campo di ricerca empirico, in cui i dati e i materiali utili per l'analisi vengono raccolti *ad hoc* per l'indagine da condurre. Tuttavia, per essere adatti a una ricerca di tipo sociolinguistico, questi dati linguistici devono non solo avere un certo grado di spontanei-

tà ma devono anche essere differenziati per parametri sociali e demografici. Nel caso di studi di sociofonetica, in cui sono le variabili fonetiche ad essere oggetto dell'indagine, un parlato elicitato, letto su fogli o sullo schermo di un computer e registrato in condizioni controllate non rappresenta un campione ideale della lingua usata spontaneamente (Calamai, 2015, p. 38). Raccogliere materiale sonoro in condizioni naturali e analizzarlo con le strumentazioni della fonetica sperimentale rappresenta un problema di non facile soluzione per il sociolinguista interessato allo studio della variazione nella lingua parlata. Questo tipo di materiale, infatti, presenta molteplici difficoltà, quali i rumori di fondo, il basso numero di combinazioni fonotattiche da analizzare, la sovrapposizione tra le voci, ecc. Inoltre, i dati raccolti sono spesso insufficienti e non adatti per indagini di tipo quantitativo (McEnery, 2001, p. 115).

I *corpora* di grandi dimensioni possono offrire al sociolinguista un campione di lingua scritta o parlata che gode di un certo grado di rappresentatività e con cui è possibile ottenere risultati statisticamente significativi. Negli ultimi decenni sono state create banche dati di ampie dimensioni, soprattutto per la lingua inglese scritta, che hanno determinato un crescente interesse per la linguistica dei *corpora* anche da parte dei sociolinguisti. Piuttosto che una branca della linguistica a sé, la linguistica dei *corpora* deve essere intesa come un approccio metodologico utile per tutte le branche della linguistica descrittiva come la fonetica, la sintassi, ecc. Si tratta di un approccio empirico, che permette di osservare l'effettivo uso della lingua in una grande quantità di testi scritti e orali, chiamati appunto *corpora*. Si serve di computer e sistemi informatici che impiegano tecniche automatiche e interattive e che permettono di fare ricerche integrate selezionando filtri e parametri. La linguistica dei *corpora* offre supporto alla visione che la variazione linguistica è sistematica e può essere descritta usando metodi quantitativi e basati sulla frequenza.

Per essere utilizzati in maniera efficace per la ricerca sociolinguistica i *corpora*, scritti o orali, devono essere etichettati con informazioni relative a vari parametri sociali per permettere la consultazione sincronica e integrata di variabili analizzate in base a fattori sia linguistici che extra-linguistici. Uno dei più grandi *corpora* usati per lo studio della variazione sociolinguistica nella lingua inglese è il *British National Corpus* (BNC), la cui ultima edizione comprende circa 97 milioni di parole. I *corpora* di lingua scritta sono assai più comuni rispetto a quelli che rappresentano la lingua parlata, questo perché, per ovvie ragioni, i testi scritti sono più facili da raccogliere e informatizzare. La metodologia di campionatura usata nel BNC è stata presa come modello per la creazione di altri *corpora* di lingua scritta, come il *Freiburg-Brown Corpus of American English* (FROWN) e lo *Australian Corpus of English* (ACE), la cui analisi simultanea ha permesso la comparazione d'uso di variabili in più varietà della lingua inglese (Friginal, 2014).

Non mancano *corpora* dedicati interamente alla lingua parlata. Sali A. Tagliamonte dell'università di Toronto lavora alla creazione di *corpora* orali adatti per essere usati per la ricerca sociolinguistica, registrando e trascrivendo, se-

condo la tradizione variazionista e laboviana, la lingua spontanea di un gran numero di parlanti. Il parlato, registrato in situazioni non controllate per catturare la lingua realmente usata, è il cosiddetto 'vernacolo', lo stile di parlato utilizzato quando non si fa attenzione a quello che si dice. Con un *corpus* di questo tipo è possibile studiare la variazione sociolinguistica di variabili anche nel parlato spontaneo. Tanto per fare un esempio, Tagliamonte & Roeder (2009) hanno analizzato, usando un *corpus* orale di 1,2 milioni di parole (92 informatori) dell'inglese parlato a York, la variazione di pronuncia dell'articolo determinativo *the* secondo più parametri extra-linguistici. Mentre i *corpora* di Tagliamonte contengono principalmente trascrizioni di interviste, esistono anche *corpora* che comprendono diversi tipi di parlato, e che consentono quindi di analizzare anche la variazione diafasica. Sempre per la lingua inglese possiamo citare il *Santa Barbara Corpus of Spoken American English* (SBCSAE) e il *Wellington Corpus of Spoken New Zealand English* (WSC). Quest'ultimo è particolarmente adatto all'indagine sociolinguistica perché accuratamente annotato secondo diversi parametri sociali, come l'età, l'etnia e il genere del parlante.

È importante sottolineare che un *corpus* orale utile alla ricerca sociolinguistica e sociofonetica non è una semplice raccolta di tracce audio. Il parlato spontaneo contenuto al suo interno deve essere analizzato con gli strumenti informatici della fonetica acustica e accuratamente trascritto. Inoltre, deve essere annotato con una serie di etichette che riguardano sia i livelli di segmentazione (livello fonetico, fonologico, sillabico, ecc.) che le informazioni sul parlante e sulla tipologia di parlato. Queste etichette vanno poi informatizzate e organizzate in modo che chi interroga il *corpus* possa fare una ricerca automatica e simultanea delle variabili oggetto dell'indagine, impostando uno o più parametri. Il *Corpus of Spontaneous Japanese*, descritto nel paragrafo successivo, ha tutte le caratteristiche appena elencate e risponde perfettamente alle esigenze del ricercatore interessato allo studio della variazione nel giapponese parlato.

Il *Corpus of Spontaneous Japanese* (CSJ)

Presso il Kokuritsu Kokugo Kenkyūjo (National Institute for Japanese Language and Linguistics) sono stati sviluppati vari tipi di *corpora* relativi alla lingua giapponese, di cui il più conosciuto è il BCCWJ (*Balanced Corpus of Contemporary Written Japanese*, in giapponese *Gendai nihongo kakikotoba kinkō kōpasu*). Si tratta di un *corpus* bilanciato[1] di lingua scritta che contiene più di 100 milioni di parole. I campioni testuali contenuti nel *corpus*, organizzati in tre

[1] Il *corpus* viene definito 'bilanciato' perché contiene varie tipologie testuali (articoli di giornale, testi letterari, saggi accademici, conversazioni di chat online, ecc.) selezionate in modo che ogni genere venga rappresentato nella medesima misura.

principali *subcorpora*, ognuno dei quali contiene dai 30 ai 35 milioni di parole, sono etichettati con ricche informazioni relative al testo (autore, anno di pubblicazione, genere testuale, ecc.) e all'analisi linguistica. Il BCCWJ si presta dunque alla ricerca sociolinguistica che ha come oggetto d'indagine la lingua scritta e risulta particolarmente adatto per lo studio delle collocazioni lessicali e per il lavoro lessicografico.[2] Per lo studio della variazione diacronica è stato sviluppato invece il CHJ (*Corpus of Historical Japanese*, in giapponese *Nihongo rekishi kōpasu*), un *corpus* in continuo aggiornamento che include risorse testuali dal periodo Nara al periodo Meiji. Tutti i *corpora* sviluppati presso il Kokuritsu Kokugo Kenkyūjo possono essere consultati attraverso la piattaforma online *Chūnagon* che consente di impostare diversi parametri e filtri di ricerca.

Interamente dedicato alla lingua parlata è il CSJ (*Corpus of Spontaneous Japanese*, in giapponese *Nihongo hanashikotoba kōpasu*). Progettato per essere adatto allo studio della variazione linguistica nel parlato (Maekawa 2002), il CSJ è un *corpus* di giapponese spontaneo sviluppato dal Kokuritsu Kokugo Kenkyūjo in collaborazione con il Communications Research Laboratory (CRL) e il Tokyo Institute of Techonology (TiTech). Contiene circa 660 ore di parlato spontaneo che corrispondono a più di 7 milioni di parole, interamente trascritto secondo una trascrizione ortografica, *kanji* e *kana*, e una trascrizione 'fonetica', esclusivamente in *kana*. L'analisi morfologica applicata permette di distinguere due tipologie di parola: l'unità lessicale breve indicata con la sigla SUW (*Small Unit Word*), che risulta molto vicina a una scomposizione in morfemi, e l'unità lessicale lunga indicata con la sigla LUW (*Long Unit Word*).[3] Una parte del CSJ, chiamata *Core*, è stata oggetto di una annotazione estremamente precisa e dettagliata in modo da poter essere utilizzata per lo studio della variazione fonetica e sociofonetica. Il *CSJ-Core* contiene circa 45 ore di parlato, corrispondenti a circa 500.000 parole, di cui è stato trascritto ogni più piccolo segmento secondo un sistema di trascrizione sub-fonemica. È inoltre provvisto di etichette che indicano l'andamento prosodico, sia a livello lessicale che a livello di frase. Sono tre le tipologie di parlato contenute: monologo, dialogo e parlato riprodotto.[4] I monologhi si suddividono a loro volta in due tipologie: APS (*Academic Presentation Speech*), presentazioni a convegni di circa 9 società accademiche che coprono i campi dell'ingegneria, delle scienze sociali e degli studi umanistici, e SPS (*Simulated Public Speech*), monologhi di circa 10-12 minuti registrati presso studi di registrazione su temi assegnati come per esempio la migliore o peggiore esperienza della propria

[2] Sull'uso del BCCWJ in lessicografia si veda Calvetti (2013).

[3] Per esempio, 本研究 'la presente ricerca' corrisponde a un LUW formato da due SUW, 本 e 研究.

[4] Per parlato riprodotto (*Reproduction Speech*) si intende la lettura ad alta voce della trascrizione del monologo da parte dello stesso parlante.

vita (Maekawa 2015a). I dialoghi contenuti, in misura assai minore rispetto ai monologhi, comprendono interviste ai parlanti di APS e SPS sul contenuto della loro presentazioni, dialoghi *task-oriented* e conversazioni libere. La lingua contenuta nel CSJ è il cosiddetto giapponese standard, la varietà condivisa da persone istruite e usata in situazioni più o meno pubbliche. Sono stati, infatti, esclusi dal *corpus* tutti i parlanti che hanno mostrato delle inflessioni dialettali. Pertanto, sebbene conosciamo il luogo di nascita dei parlanti e di entrambi i loro genitori, risulta difficile compiere analisi che riguardano la variazione diatopica. Sono possibili invece analisi della variazione diacronica, diastratica, diagenerica e diafasica, in quanto è possibile ottenere, come sarà mostrato più avanti, informazioni dettagliate sul parlante e sulla tipologia di parlato.

Tutte le informazioni relative al *CSJ-Core*, come le trascrizioni e le informazioni sul parlante, sono state organizzate in un *Relational Database* (RDB) composto da tabelle che possono essere messe in relazione usando il sistema di *query* SQLite (Maekawa 2015b). Le tabelle che contengono i dati relativi a diversi livelli di analisi linguistica sono ordinate in maniera piramidale, dalla tabella che contiene i singoli foni a quella che contiene il discorso, passando per il fonema, la mora, il SUW, il LUW, il *bunsetsu* e il sintagma accentuale. Le tabelle che contengono dati linguistici possono essere collegate e fuse con le tabelle che contengono dati extra-linguistici, come le informazioni sul parlante (sesso, fascia d'età, titolo di studio, luogo di nascita, luogo di nascita del padre e della madre, ecc.) e le informazioni relative al parlato (tipologia di parlato, velocità di eloquio, numero di ascoltatori, ecc.). Per interrogare il *CSJ-Core* usando il linguaggio SQLite è necessario usare un software ponte come Navicat o DB Browser for SQlite.

La figura 1 rappresenta una porzione[5] della tabella segPhone che contiene i dati relativi al livello sub-fonemico del parlato contenuto nel *CSJ-Core*. In questa tabella il segmento fonico 発表します «faccio una presentazione» è scomposto in foni, dei quali abbiamo una serie di informazioni: (in ordine) l'ID del *talk*, l'ID del fono, il canale di comunicazione, il momento in cui inizia, il momento in cui termina, la trascrizione fonetica, la classe fonica e infine informazioni sulla desonorizzazione vocalica. Le altre tabelle contenute nel *CSJ-Core* riportano, allo stesso modo, le informazioni relative a più alti livelli di analisi linguistica, come il fonema, la mora, il morfema, ecc. La tabella che contiene le informazioni sul parlante è chiamata infoSpeaker (figura 2).[6] Sono presenti informazioni relative ai parlanti[7], tra cui il sesso, il luogo di nascita del parlante stesso e quello dei suoi genitori.

[5] La tabella segPhone contiene oltre 2 milioni di segmenti fonici.

[6] Sono 129 i parlanti presenti nel *CSJ-Core*, 79 uomini e 60 donne.

[7] Per comodità sono state omesse le colonne relative alla durata di permanenza nella prefettura metropolitana di Tokyo (durata totale e durata nel periodo della scuola dell'obbligo).

TalkID	PhoneID	Channel	StartTime	EndTime	PhoneEntity	PhoneClass	Devoiced
A01F0055	00005551385L	L	5.551385	5.632454	h	consonant	0
A01F0055	00005632454L	L	5.632454	5.698874	a	vowel	0
A01F0055	00005698874L	L	5.698874	5.760029	Q	special	0
A01F0055	00005760029L	L	5.760029	5.821184	<cl>	others	0
A01F0055	00005821184L	L	5.821184	5.837566	py	consonant	0
A01F0055	00005837566L	L	5.837566	5.907903	o	vowel	0
A01F0055	00005907903L	L	5.907903	5.978241	H	special	0
A01F0055	00005978241L	L	5.978241	6.028153	sj	consonant	0
A01F0055	00006028153L	L	6.028153	6.078064	i	vowel	1
A01F0055	00006078064L	L	6.078064	6.16653	m	consonant	0
A01F0055	00006166530L	L	6.16653	6.277931	a	vowel	0
A01F0055	00006277931L	L	6.277931	6.382126	s	consonant	0
A01F0055	00006382126L	L	6.382126	6.532801	u	vowel	0

Figura 1. Porzione della tabella segPhone contenuta del *CSJ-Core.*

SpeakerID	SpeakerSex	SpeakerBirthGeneration	SpeakerBirthPlace	FatherBirthPlace	MotherBirthPlace
3	女	45to49	東京都	秋田県	新潟県
8	男	55to59	京都府	滋賀県	京都府
12	男	55to59	東京都	東京都	群馬県
19	女	65to69	東京都	東京都	東京都
21	男	70to74	栃木県	栃木県	栃木県
26	女	60to64	東京都	東京都	東京都
27	男	45to49	東京都	神奈川県	東京都
28	男	70to74	東京都	東京都	長崎県
31	男	65to69	東京都	広島県	トルコ
47	男	55to59	新潟県	新潟県	新潟県
59	男	75to79	埼玉県	埼玉県	福島県
68	女	55to59	千葉県	千葉県	山形県
75	男	70to74	千葉県	千葉県	茨城県

Figura 2. Porzione della tabella infoSpeaker contenuta nel *CSJ-Core.*

Attraverso il linguaggio SQLite è possibile fondere queste due tabelle selezionando alcuni specifici parametri e ottenere quindi una tabella che contiene solo le informazioni utili alla nostra indagine. Supponiamo di avere la necessità di conoscere la durata media in secondi delle vocali divisa per sesso del parlante. Inoltre, vogliamo circoscrivere l'indagine a parlanti nati tra il 1940 e il 1959 e il cui luogo di nascita è la prefettura metropolitana di Tokyo. Per poter ottenere queste informazioni sarà necessario definire la *query* come in figura 3: vorrei selezionare (SELECT) il PhoneEntity con la relativa durata media (*avg*) arrotondata (*round*) e il sesso del parlante; dico al sistema che deve prendere queste informazioni dalla

(FROM) tabella segPhone che per comodità chiamo qui *t0*; devo però unire attraverso il comando JOIN la tabella segPhone, che contiene le informazioni sui foni, con la tabella infoSpeaker, chiamata qui *t2*, che contiene le informazioni sul parlante; per unire le due tabelle devo però utilizzare la tabella infoTalk, chiamata qui *t1*, che contiene informazioni in comune e che può quindi fare da ponte tra le due; con il comando WHERE dico al sistema di selezionare solo i PhoneEntity che corrispondono a vocali, di parlanti nati tra il 1940 e il 1959, e il cui luogo di nascita è la prefettura metropolitana di Tokyo; con il comando GROUP BY posso impedire la visualizzazione replicata per fono e per sesso del parlante, mostrando soltanto un unico risultato (per fono e sesso) che contiene entrambi i valori specificati; infine dico al sistema di ordinare le vocali in ordine alfabetico con il comando ORDER BY (ASC). Se la *query* è stata definita correttamente, senza errori grammaticali di linguaggio SQLite, il sistema analizzerà tutti i dati relativi alle 45 ore di parlato contenute nel *CSJ-Core* e otterrò in pochi secondi la tabella con le informazioni utili alla mia indagine (figura 4). I risultati ci dicono che la durata media espressa in secondi di tutte le vocali prodotte da parlanti di sesso femminile è sempre maggiore rispetto a quella di parlanti di sesso maschile, per la fascia di età e luogo di origine considerati. Questo è un esempio di una semplice *query* che considera solo alcuni parametri presenti in due tabelle. Tuttavia, indagini che mettono in relazione fattori linguistici, come per esempio la struttura fonotattica, e simultaneamente fattori extra-linguistici richiedono *query* molto più lunghe e articolate. Il sociolinguista che fa uso di questa tipologia di *corpus* dovrà pertanto familiarizzare con il linguaggio SQLite per poter sfruttare al meglio le potenzialità offerte da questo nuovo approccio metodologico.

```
SELECT t0.PhoneEntity, round(avg(t0.EndTime - t0.StartTime),3) AS DurataMedia,
t2.SpeakerSex
FROM segPhone AS t0
JOIN infoTalk AS t1
ON t0.talkID = t1.TalkID
JOIN infoSpeaker AS t2
ON t1.SpeakerID = t2.SpeakerID
WHERE t0.PhoneClass = "vowel"
AND t2.SpeakerBirthGeneration IN ("40to44", "45to49", "50to54", "55to59")
AND t2.SpeakerBirthPlace = "東京都"
GROUP BY t0.PhoneEntity, t2.SpeakerSex
ORDER BY t0.PhoneEntity ASC
```

Figura 3. Esempio di *query* in SQLite.

PhoneEntity	DurataMedia	SpeakerSex
A	0.086	女
A	0.083	男
E	0.089	女
E	0.086	男
I	0.061	女
I	0.059	男
O	0.08	女
O	0.076	男
U	0.058	女
U	0.053	男

Figura 4. Risultato della *query*.

Un case study sulla variazione delle pause non silenti

Obiettivo dell'indagine

Al fine di mostrare un esempio concreto di applicazione del CSJ per la ricerca sociolinguistica sarà presentato un *case study* sulla variazione delle pause non silenti nel giapponese spontaneo. L'indagine ha l'obiettivo di individuare i fattori, linguistici ed extra-linguistici, che influenzano le scelte del parlante sulla tipologia di pausa non silente da utilizzare. Si tratta di uno studio empirico e quantitativo che mette in evidenza le potenzialità di un *corpus* orale di grandi dimensioni e con un ricco sistema di annotazioni come il CSJ.

Le pause non silenti

Le pause non silenti, chiamate in inglese *filler*, si contrappongono alle pause silenti e sono fenomeni di esitazione tipici del parlato spontaneo. Servono al parlante per prendere tempo nell'elaborazione di un discorso oppure, da un punto di vista più pragmatico, per mantenere il turno conversazionale. A differenza delle pause silenti, quelle non silenti, per la loro varia tipologia, risultano difficili da classificare. Una classificazione generale prevede, ad esempio, una distinzione tra pause lessicalizzate, del tipo «diciamo, appunto, praticamente, allora», e non lessicalizzate, del tipo

«uhm, ehm, eeh» (Giannini, 2001). In letteratura viene spesso usata la dicitura *pausa piena*, in luogo di pausa non silente, in opposizione a *pausa vuota*. Tuttavia, poiché con *pause piene* si fa generalmente riferimento a un tipo particolare di pausa, vale a dire le vocalizzazioni (Giannini, 2003), è più opportuno, soprattutto parlando del fenomeno nella lingua giapponese, definirle pause silenti e non silenti.

È noto che nella lingua giapponese parlata vengono utilizzate numerose pause lessicalizzate, come per esempio *ēto, ano*, *ma*, *sono* (le cui vocali possono essere brevi o lunghe a seconda dei casi), talvolta utilizzate anche in composizione generando un tipo di pausa non silente che possiamo definire pausa lessicalizzata multipla (come per esempio *mā anō*, *ēto ma*). Tra le pause non lessicalizzate più comuni troviamo le vocalizzazioni e le nasalizzazioni, le quali possono essere abbinate a una o più pause lessicalizzate. Per la nostra indagine divideremo per comodità le pause non silenti della lingua giapponese in quattro categorie: pause lessicalizzate, vocalizzazioni, nasalizzazioni e pause multiple (da intendersi come la composizione di più tipologie di pause, lessicalizzate e non lessicalizzate).[8]

Materiali e metodi

Interrogando il *CSJ-Core* attraverso il linguaggio SQLite sarà analizzata l'occorrenza delle 32162 pause non silenti presenti nelle 45 ore di parlato.[9] I fattori extra-linguistici scelti come parametri di indagine sono il genere, l'età, il titolo di studio del parlante e la tipologia di parlato. Poiché la scelta della pausa non silente può essere condizionata non solo da fattori extra-linguistici, si è scelto di considerare tra i parametri un fattore linguistico, e cioè la lunghezza in secondi della frase di cui la pausa non silente rappresenta la testa. Sarà inoltre calcolato l'indice di disfluenza, inteso come la percentuale di pause non silenti sul totale dei *bunsetsu*[10] prodotti.[11]

Risultati e discussione

In figura 5 possiamo osservare la variazione dell'indice di disfluenza in base al sesso e alla generazione dei parlanti. Notiamo che il sesso del parlante ha un chiaro effetto solo per la fascia d'età medio-bassa dove gli uomini presentano un indice di disfluenza maggiore rispetto alle donne. Il fattore età diventa pressoché irrilevante

[8] Uno studio completo sulla variazione delle pause non silenti nel giapponese spontaneo viene presentato in Watanabe (2009). Nella sua indagine, Watanabe non considera le pause multiple come una categoria a sé stante.

[9] Sono stati tuttavia esclusi dall'indagine i monologhi presentati leggendo ad alta voce un testo scritto.

[10] Il *bunsetsu* è un costituente sintattico della lingua giapponese, composto da una parola piena a cui possono essere attaccate eventuali parole grammaticali o funzionali.

[11] Nel *CSJ-Core* le pause non silenti, singole o multiple, costituiscono un singolo *bunsetsu*.

se consideriamo solo parlanti donne, che mostrano un indice di disfluenza che varia dal 13,19% al 12,24%.

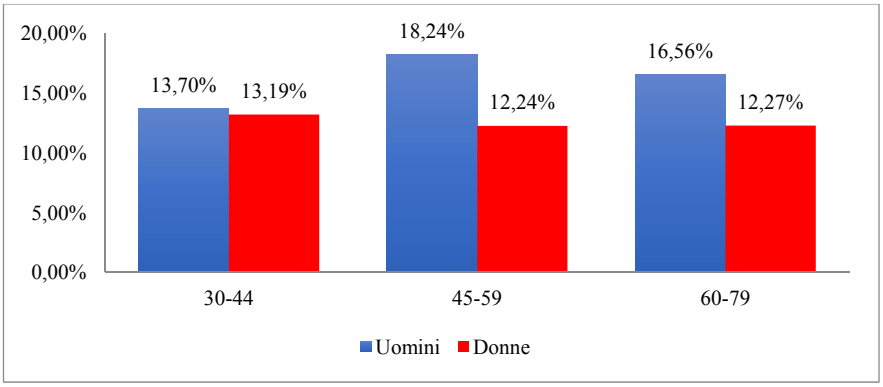

Figura 5. Variazione dell'indice di disfluenza per sesso e anno di nascita del parlante.

In figura 6 possiamo osservare l'interazione tra la variazione diagenerica e quella diafasica per quanto riguarda l'indice di disfluenza. È evidente che la tipologia di parlato ha un effetto sulla frequenza di pause non silenti, con percentuali maggiori nei dialoghi e minori nel parlato riprodotto, mentre lo scarto tra i due sessi si mantiene pressoché costante su tutte le tipologie: i parlanti di sesso maschile presentano indici di disfluenza piuttosto alti che arrivano a superare il 23% nei dialoghi.

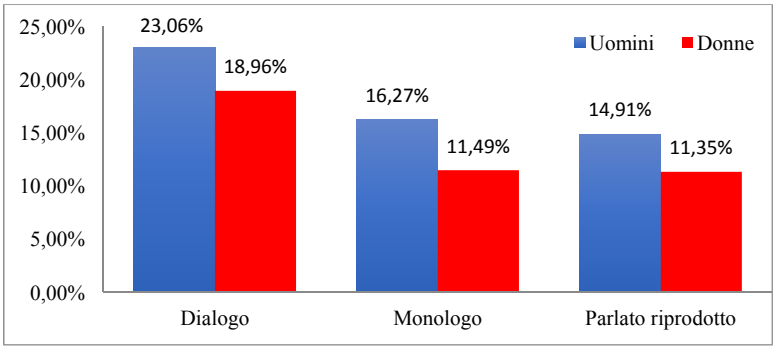

Figura 6. Variazione dell'indice di disfluenza per sesso del parlante e tipologia di parlato.

La figura 7 mostra l'interazione tra la variazione diagenerica e quella diastratica, qui indicata dal titolo di studio del parlante. Contrariamente a quanto ci si potrebbe aspettare, la frequenza delle pause non silenti prodotte aumenta quanto più è elevato il livello di istruzione del parlante. Il titolo di studio ha dunque un effetto sulla variazione delle pause non silenti, in particolare per i parlanti di sesso femminile.

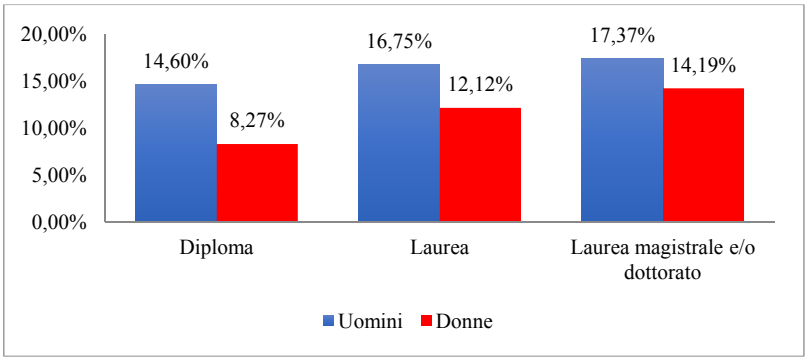

Figura 7. Variazione dell'indice di disfluenza per sesso e titolo di studio del parlante.

I grafici in figura 8 e 9 riportano le percentuali di tipologia di pausa non silente in entrambi i sessi. Le vocalizzazioni rappresentano il 50% delle pause non silenti utilizzate dagli uomini, seguite dalle pause lessicalizzate con il 38%. Le donne, invece, tendono a preferire le pause lessicalizzate che raggiungono il 47% del totale.

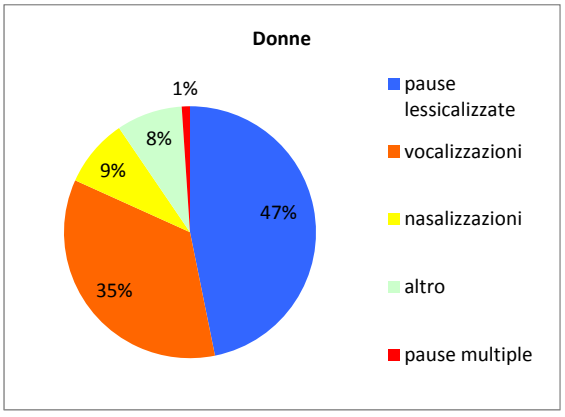

Figura 8. Percentuali d'uso di tipologia di pausa non silente nei parlanti di sesso femminile.

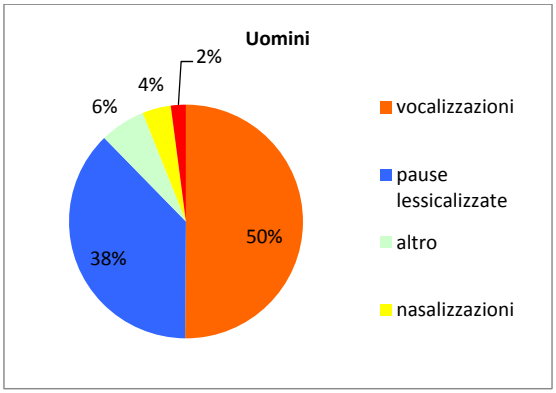

Figura 9. Percentuali d'uso di tipologia di pausa non silente nei parlanti di sesso maschile.

La variazione nella scelta della tipologia di pausa non silente può essere influenzata, oltre che dal sesso del parlante, anche dalla tipologia di parlato (figura 10). Le nasalizzazioni, quasi assenti nei monologhi, risultano essere tratti distintivi di un parlato dialogico, usate probabilmente per mostrare incertezza in risposta a una domanda.

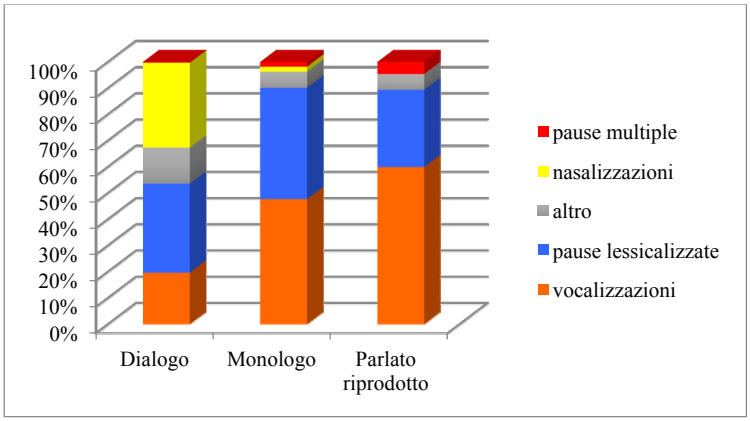

Figura 10. Percentuali d'uso di tipologia di pausa non silente per tipologia di parlato.

Le pause lessicalizzate, come abbiamo appena visto, rappresentano nella lingua giapponese una porzione importante delle pause non silenti utilizzate, sia dalle donne che dagli uomini. I grafici in figura 11 e 12[12] ci mostrano che le donne hanno una netta preferenza per *ano* (あの), mentre gli uomini tendono a preferire *ma*

[12] Per la creazione dei grafici 11 e 12, le pause lessicalizzate sono state raggruppate per struttura di base, accorpando in un'unica tipologia le pause con vocali o consonanti brevi e lunghe (per esempio *anō* e *ano*, *mā* e *ma*, *ēttō* e *ētō*, e così via); sono state inoltre escluse le pause lessicalizzate multiple.

(ま). Risulta tuttavia difficile speculare sui motivi che stanno alla base di questa variazione diagenerica.

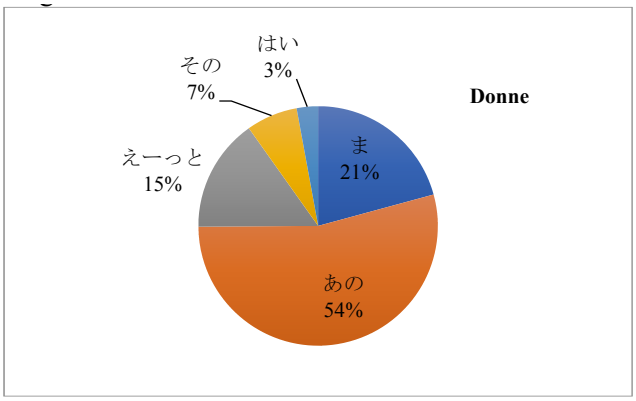

Figura 11. Percentuali d'uso di pause lessicalizzate nei parlanti di sesso femminile.

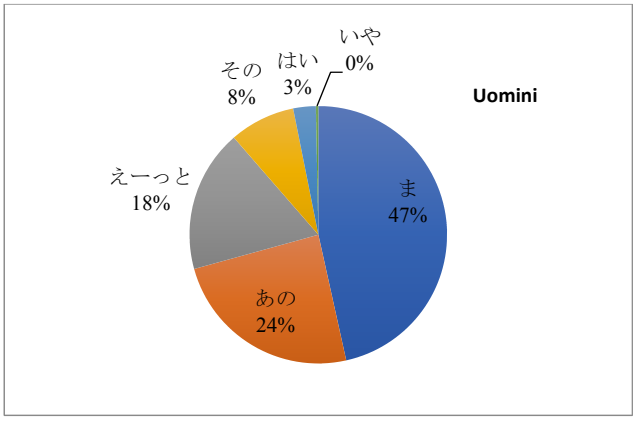

Figura 12. Percentuali d'uso di pause lessicalizzate nei parlanti di sesso maschile.

Come abbiamo visto dai risulti finora presentati, l'uso delle pause non silenti viene influenzato da vari fattori extra-linguistici, come il sesso del parlante, la tipologia di parlato e il livello di istruzione. Tuttavia, le scelte linguistiche del parlante possono essere condizionate anche da fattori pertinenti alla struttura linguistica e per questo è necessario considerare almeno un fattore propriamente linguistico. Il *CSJ-Core* ci consente di calcolare la durata in secondi della frase di cui la pausa non silente rappresenta la testa e verificare se c'è una relazione tra tipologia di pausa e la lunghezza

e la complessità della frase che si sta per elaborare. Il grafico in figura 13 presenta la durata media in secondi delle frasi introdotte da vari tipi di pausa non silente. Come possiamo osservare, una pausa multipla viene preferita quando si vuole formulare una frase piuttosto lunga. Notiamo inoltre che le pause lessicalizzate precedono frasi proporzionalmente più lunghe rispetto alle vocalizzazioni (a esclusione di *ē*) e alle nasalizzazioni.

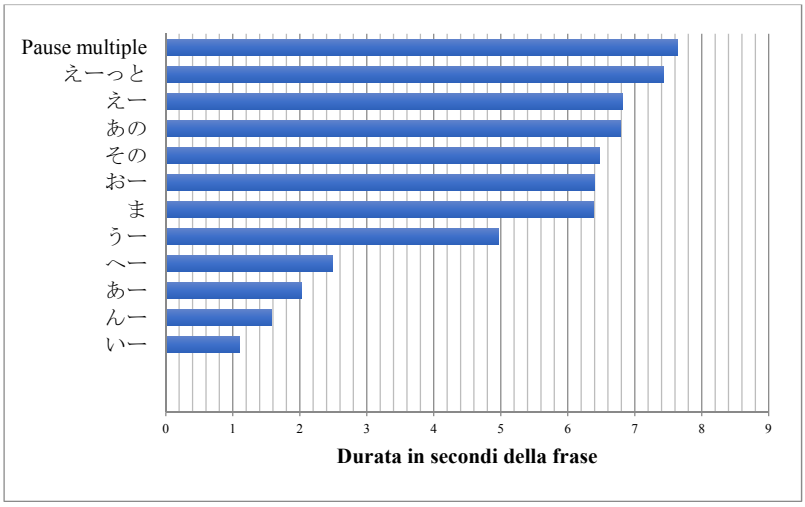

Figura 13. Rapporto tra pausa non silenti e durata media della frase che precede.

Conclusioni

Volendo trarre delle conclusioni da questa indagine preliminare sulla variazione delle pause non silenti nel giapponese spontaneo, possiamo così riassumerle:

- La variazione delle pause non silenti in giapponese è influenzata sia da fattori linguistici che extra-linguistici;
- I parlanti di sesso maschile hanno mostrato, in tutti i casi analizzati, un indice di disfluenza sempre maggiore rispetto ai parlanti di sesso femminile, in particolare per le fasce d'età più giovani;
- Il parlato dialogico, che ha registrato un indice di disfluenza maggiore rispetto alle altre tipologie di parlato considerate, è caratterizzato da una presenza consistente di nasalizzazioni, quasi assenti nei monologhi;
- La variazione diagenerica nella scelta di pausa lessicalizzata riguarda principalmente le pause *ano* e *ma*: la prima nettamente preferita dalle donne, la seconda dagli uomini;

– La lunghezza e la complessità di pausa non silente hanno una relazione diretta con la frase che si sta per elaborare: le pause lessicalizzate multiple vengono scelte dal parlante quando si vuole formulare una frase mediamente più lunga.

I dati di questa indagine hanno messo in evidenza che il CSJ risulta essere uno strumento estremamente utile per lo studio della variazione linguistica nel giapponese spontaneo. Permette non solo di mettere in relazione i dati linguistici con le informazioni sul parlante e sulla tipologia di parlato ma anche di effettuare analisi quantitative che rilevano correlazioni tra più elementi della struttura linguistica. Lo sviluppo di *corpora* orali di sempre più grandi dimensioni consentirà in futuro di ampliare gli studi sulla variazione linguistica nella lingua parlata e di fornire dati utili per la ricerca sociologica, psicolinguistica e glottodidattica.

Riferimenti bibliografici

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

Calamai, Silvia (2015). *Introduzione alla sociofonetica.* Roma: Carocci editore.

Calvetti, Paolo (2013). "L'uso dei *corpora* bilanciati nella compilazione di dizionari bilingui. Il caso del progetto del grande dizionario giapponese-italiano". In Casari, Matteo; Scrolavezza, Paola (a cura di). *Giappone, storie plurali*. Bologna: I libri di Emil, pp. 319-334.

Coulmas, Florian (a cura di) (2006). *The Handbook of Sociolinguistics*. Oxford & Cambridge, MA: Blackwell.

CSJ (The Corpus of Spontaneous Japanese) (2004) National Institute for Japanese Language and Linguistics and National Institute of Information and Communications Technology. http://www.ninjal.ac.jp/english/products/csj/.

Friginal, Eric; Hardy, Jack A. (2014). *Corpus-Based Sociolinguistics. A Guide for Students.* New York: Routledge.

Giannini, Antonella (2001). "Corpus AVIP: ehm, ehm". In Magno Caldognetto, Emanuela; Cosi, Piero (a cura di). *Multimodalità e multimedialità nella comunicazione*. *Atti delle XI Giornate di Studio del Gruppo di Fonetica Sperimentale*, XXVIII. Padova: Unipress, pp. 179-184.

Giannini, Antonella (2003). "Vocalizzazioni e prolungamenti vocalici". In Magno Caldognetto, Emanuela; Cosi, Piero (a cura di). *Voce, canto, parlato. Studi in onore di Franco Ferrero.* Padova: Unipress*,* pp. 163-172.

Labov, William (1963). "The Social Motivation of a Sound Change". *Word*, 19, pp. 273-309.

Hudson, R. A. (1980). *Sociolinguistics*. Cambridge: Cambridge University Press.

Maekawa, Kikuo (2002) "Nihongo hanashi kotoba kōpasu wo mochiita gengo hen'i kenkyū". *Onsei kenkyū*, 6 (3), pp. 48-59.

Maekawa, Kikuo (2003) "Corpus of Spontaneous Japanese: Its Design and Evaluation". *Proceedings of ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, Tokyo, pp.7-12.

Maekawa, Kikuo (2015a) "Corpus-based phonetics" In Kubozono, Haruo (a cura di). *Handbook of Japanese Phonetics and Phonology*. Berlin: De Gruyter Mouton, pp. 651-680.

Maekawa, Kikuo (2015b) "Rirēshonaru dētabēsu". *Nihongogaku* 35 (1), pp. 82-87.

McEnery, Tony; Wilson, Andrew (2001). *Corpus Linguistics*. Edimburgh: Edimburgh University Press.

Tagliamonte, Sali A. (2012). *Variationist sociolinguistics: Change, observation, interpretation*. Malden, MA: Wiley-Blackwell.

Tagliamonte, Sali A.; Roeder, Rebecca V. (2009). "Variation in the English definite article: Socio-historical linguistics in t'speech community". *Journal of Sociolinguistics*, 13(4), pp. 435-471.

Watanabe, Michiko (2009). *Features and Roles of Filled Pauses in Speech Communication.* Tōkyō: Hituzi Syobo Publishing.