CORMIP (Corpus Multimodale dell'Italiano Parlato): questioni intorno al trattamento del dato linguistico multimodale

Luca Lo Re

1. Introduzione

Il lavoro presentato in queste pagine propone un metodo per la costruzione di un corpus multimodale dell'italiano parlato, inserendosi nella tradizione dei *corpora* di italiano parlato compilati dal gruppo di ricerca LABLITA dell'Università di Firenze. Il lavoro di ricerca ha dato vita a un piccolo corpus pilota con l'intento di sperimentare il metodo che verrà illustrato e per porre così le basi per ricerche future. Questa ricerca nasce dall'esigenza di approcciarsi alla lingua come sistema multimodale che costruisce ed esprime i significati attraverso diversi indici e canali, in particolare gesto e parlato. Gli studi sul gesto (*gesture studies*) hanno sviluppato approcci e metodi diversi così che ad oggi manca uno standard rispetto alla raccolta e al trattamento dei dati. Inoltre, la complessità dell'oggetto di studio richiede un grande sforzo nella compilazione dei *corpora* portando a una quantità di dati ridotta (dati elicitati in laboratorio o raccolte annotate per specifici obiettivi di ricerca). Il lavoro di ricerca ha l'intento di proporre una metodologia per la compilazione dei *corpora* multimodali, basata su un metodo pragmatico e percettivo.

2. La nozione di multimodalità

Il termine 'multimodalità', coniato a metà degli anni '90, è ampiamente usato da diversi studiosi che a oggi non sono riusciti a costruire una

DOI 10.35948/DILEF/978-88-6032-750-5.15

Luca Lo Re, Università degli Studi di Firenze, luca.lore@unifi.it, 0000-0002-2937-1636

definizione condivisa e unica dell'oggetto di studio della multimodalità. In modo generico è possibile affermare che l'oggetto di interesse è l'uso di diverse modalità utilizzate per la creazione del significato (Goodwin 2000; Kress-van Leeuwen 2001).

L'interesse della presente ricerca si rivolge all'italiano parlato spontaneo, concependo la lingua come un'azione che si realizza attraverso diversi mezzi: prosodia, parola, gesto, espressioni facciali e postura; e si basa sui risultati e i riferimenti teorici degli studi sul gesto di Kendon, McNeill, Enfield. Kendon definisce il gesto come «a name for visible action when it is used as an utterance or as a part of an utterance» (Kendon 2004) e vede l'enunciato come «any unit of activity that is treated by those co-present as a communicative 'move', 'turn' or contribution. Such units of activity may be constructed from speech or from visible bodily action or from combinations of these two modalities» (*ibidem*). Invece Enfield parla di un *composite utterance* definendolo «as a communicative move that incorporates multiple signs of multiple types» (Enfield 2013). Così l'idea di un enunciato multimodale sembra essere un concetto teorico, basato su prove empiriche, ma che non può essere considerato come unità di riferimento per l'analisi linguistica. Infatti, non esiste alcuna definizione basata su caratteristiche pratiche oltre che sull'enunciato parlato.

3. Il dibattito intorno alla natura multimodale del linguaggio

La nozione di linguaggio come sistema multimodale ci porta a riconoscere che la natura del gesto è verbale (inteso come elemento della comunicazione orale) e che è un canale che esprime diversi valori linguistici, ponendosi in contrapposizione con la nozione di lingua tradizionale caratterizzata dall'arbitrarietà (Hockett 1960) e dalla doppia articolazione (Martinet 1960).

Il punto di vista multimodale sul linguaggio implica che il sistema linguistico sia dinamico e semiologicamente eterogeneo. Infatti, la significatività di un atto linguistico multimodale nasce da una continua dialettica con il contesto in cui si realizza e la concomitanza di diverse

modalità di significazione che non hanno lo stesso grado di convenzionalizzazione. Quindi il sistema lingua non può essere definito su parametri come la doppia articolazione e l'arbitrarietà, poiché farebbero escludere i gesti da questo sistema riducendoli a elementi accessori del linguaggio. La lingua appare come un sistema composto da diversi tipi di modalità semiologiche che interagiscono tra loro per realizzare un'unità globale.

La nozione di linguaggio elaborata da De Mauro 2000 rappresenta un importante contributo al dibattito. Egli ha sostenuto che non è il carattere dell'audioralità a definire le lingue umane, piuttosto esse si distinguono dagli altri codici comunicativi per la capacità di produrre nuovi significati e nuove parole. Questa possibilità è data dalla capacità, del sistema lingua, di riarticolare i propri segni sia a livello semantico che a livello del significante. L'inderteminatezza del segno permette la riformulazione della relazione tra significante e significato per mezzo del contesto d'uso; infatti, una delle caratteristiche della lingua elencate da De Mauro 2000 è il carattere locale del funzionamento dei segni e la necessità della realizzazione di intese tra i parlanti. Questa flessibilità, quindi, si realizza nell'uso del linguaggio e rende possibile la ridertiminatezza dei segni linguistici così da estendere il loro significato fino al raggiungimento dell'autoreferenzialità; ed è attraverso l'uso metalinguistico che i parlanti controllano e gestiscono la duttilità del segno nella negoziazione sociale. In questo modo il linguaggio si caratterizza per la sua indeterminatezza data dall'arbitrarietà e dall'uso sociale. Così, De Mauro ha sostenuto che il linguaggio è uno strumento sociale che riafferma continuamente la relazione tra significante e significato nell'uso attraverso l'arbitrarietà e la negoziazione sociale. Ogni sistema linguistico utilizza la sua modalità, la lingua dei segni con il segno e la lingua orale con le parole, ma come abbiamo visto sopra ogni sistema linguistico è multimodale.

In questo modo viene offerta una nozione dinamica di linguaggio in cui il gesto può vedere riconosciuto il suo ruolo e le sue funzioni. In particolare, lo studioso ne individua quattro: 1) appoggio extrafunzionale (inteso come ruolo di scansione e sottolineatura, «un ruolo di appoggio alla scansione sintattica e alla determinazione del senso complessivo di un enunciato», *ibidem*); 2) integrazione alla semantica

di singoli lessemi o gruppi sintagmatici di lessemi; 3) sostituzione semioticamente equivalente; 4) sostituzione semioticamente equivalente (linguaggi speciali degli operatori di borsa, di pescatori, degli addetti aeroportuali ecc.).

Su una prospettiva simile si collocano gli studi di Enfield 2009 che cerca di superare la staticità della nozione tradizionale di lingua attraverso un approccio socio-interazionale arrivando a concettualizzare il *compostie utterance*. L'enunciato è visto come una unità del comportamento sociale che ha una ben definita relazione causa-effetto e ogni mossa comunicativa scaturisce da determinati presupposizioni e dai *commitments* che vengono riportati nello scambio dei turni dialogici e a cui l'interlocutore è tenuto a rispondere. È proprio il riconoscimento dell'anatomia casuale/condizionale e normativa delle sequenze di interazione, in cui ogni mossa porta a un nuovo obiettivo con conseguenze per i parlanti coinvolti, che permette a Enfield il superamento della staticità del segno linguistico sassuriano. La dinamicità semantica del *composite utterance* è determinata anche dalla diversa natura dei segni che lo compongono: segni convenzionali, segni non convenzionali e segni simbolici indessicali (Enfield 2009).

La questione della multimodalità del linguaggio ha portato a rivedere la definizione di lingua da un punto di vista semiotico-linguistico, inglobando l'aspetto sociale e interazionale. A ciò è necessario aggiungere anche l'aspetto cognitivo che influenza in toto l'espressione linguistica, rispetto alla concettualizzazione e all'uso di diverse strategie espressive.

4. Il corpus CORMIP

4.1 La raccolta dei dati: metodi e strumenti

Una prima questione da affrontare per la compilazione del *corpus* CORPIM è stata la definizione di spontaneità dei dati dell'italiano parlato e la definizione di dati multimodali. Di fronte alla complessità delle questioni da affrontare, si è ritenuto opportuno iniziare a elaborare una proposta metodologica attraverso l'uso di un *corpus* pilota che includa

dati sulla gestualità, oltre che sul parlato inteso come canale, lasciando in secondo piano la rappresentatività del *corpus*.

Abbiamo progettato la raccolta dati cercando di garantire la diversificazione delle tipologie interazionali, un livello minimo di variabilità diatopica e la spontaneità degli eventi linguistici.

Per la tipologia interazionale abbiamo seguito il parametro del numero dei locutori «che determina l'ossatura dell'evento comunicativo e permette di distinguere i testi orali in monologhi, dialoghi e conversazioni» (Cresti 2000) includendo nei dati tre tipologie interazionali:

- monologico, in cui ritroviamo solamente un parlante in un contesto interattivo dove l'interlocutore non può togliere il turno in modo (es. una lazione universitaria);
- dialogico, in cui ritroviamo due parlanti che possono interagire fra loro prendendo il turno liberamente;
- conversazionale, dove troviamo più di tre parlanti che interagiscono potendo prendere il turno liberamente.

La quasi totalità delle registrazioni è ascrivibile a un contesto sociale privato tra persone con un grado di conoscenza reciproca alto, fatta eccezione di un solo brano che riporta una lezione universitaria e dunque è possibile parlare di contesto comunicativo pubblico. Tutte le interazioni sono spontanee e raccolte in un contesto naturale.

Rispetto alla variabilità diatopica, sono stati inclusi solamente due punti di raccolta dati riferibili a due città, Firenze e Catania. Per ciascuna delle due città abbiamo raccolto un brano per ogni tipologia interazionale. Questa nostra scelta intende includere una livello minimo di variazione diatopica.

Raccogliere dati spontanei porta a tener conto del paradosso dell'osservatore (Labov 1972). Per motivi etici e legali, non è stato possibile realizzare la registrazione senza l'accordo dei partecipanti portando a cercare di attenuarne gli affetti. Così, i parlanti sono stati informati (attraverso l'informativa sulla privacy) sulla finalità delle registrazioni senza però specificare il campo di studi e gli elementi di interesse. È stata utilizzata strumentazione non invasiva, nello specifico: una tele-

camera GoPro Hero 6 e un audio registratore Zoom H6 con un microfono panoramico (120°). Il setting garantiva ai parlanti libertà di movimento, non sono state date loro regole o accorgimenti da seguire e gli argomenti trattati sono totalmente spontanei.

In totale sono state registrate sei diverse situazioni comunicative, tre generi conversazionali per due diverse città. I partecipanti hanno un range di età che va dai vent'anni ai sessanta e il livello di istruzione più basso è il diploma di scuola media superiore. Nella tabella 1 viene illustrato il *dataset* del *corpus*.

Genere conversazionale	Evento comunicativo	Città	Durata
Monologo	Lezione letteratura	Firenze	5'.20"
Dialogo	Dialogo Scout	Firenze	5'.46"
Conversazione	Arbitri pallamano	Firenze	5'.42"
Monologo	Racconto di vita	Catania	5'.27"
Dialogo	Conversazione studenti	Catania	5'.52"
Conversazione	Conversazione viaggio	Catania	7'.12"

Tabella 1 Data-set del corpus CORMIP

4.2 Trascrizione e annotazione dei dati

I sistemi di annotazione del gesto sono diversi e calibrati su metodi e obiettivi propri: NEUROGES (Lausberg 2013) (Lausberg-Sloetjes 2016), CoGesT (Gibbon et al. 2003) (Trippel et al. 2004), e LASG (Bressem-Ladewig-Müller 2013).

Questa ricerca ha l'obiettivo di creare un sistema di trascrizione e annotazione che possa identificare le unità strutturali su base percettiva dando centralità alla diversità delle modalità da analizzare. Infatti, i diversi metodi di significazioni di ciascuna modalità (gesto e parlato) dipendono fortemente dalla loro fisicità, pur scaturendo entrambe da un medesimo processo cognitivo.

L'annotazione del parlato fa riferimento alla Teoria della Lingua in Atto (Cresti 2000), che identifica gli enunciati e le unità di intona-

zione della modalità parlata sulla percezione uditiva; per la trascrizione e l'annotazione gestuale il riferimento teorico è rappresentato dagli studi di Kendon 1972 e McNeill 1992. I due livelli di annotazioni sono mantenuti separati in modo da lasciare le informazioni linguistiche dei due canali indipendenti l'uno dall'altro, per poter così ridurre le possibili influenze e per poter indagare in modo dettagliato come le due modalità si correlano. La multimodalità dell'azione linguistica emerge dall'annotazione dell'illocuzione, che rappresenta l'elemento linguistico che caratterizza a nostro avviso l'uso di elementi semantici, intonativi e gestuali.

4.3 La trascrizione e l'annotazione del parlato

La Teoria della Lingua in Atto (Cresti 2000) si basa sulla teoria degli atti linguistici di Austin 1962. La proposta si poggia su due tipi di unità di riferimento individuate prosodicamente: l'*utterance* e la *stanza*. L'*utterance* è l'unità linguistica minima e principale caratterizzata da un confine prosodico terminato e compie un unico atto linguistico; la *stanza* è formata da una sequenza di *comment* deboli che non corrispondono a una sequenza di enunciati. Le unità di riferimento del discorso sono entità linguistiche basate su caratteristiche semantiche, pragmatiche e prosodiche e la loro identificazione avviene prosodicamente attraverso il riconoscimento percettivo dei confini tonali da parte dell'annotatore. La struttura informativa del parlato è costruita intorno all'unità necessaria e sufficiente chiamata *Comment* e che potrebbe essere accompagnata da altre unità opzionali con le quali forma lo schema informativo. Le unità aggiuntive assumono diverse funzioni: *Topic*, *Parenthesis*, *Appendix*, *Locutive Introducer* e *Discourse Markers*.

La centralità della prosodia all'interno della Teoria della Lingua in Atto si appoggia al modello prosodico elaborato dai lavori di IPO ('t Hart-Collier-Cohen 1990) dimostrando che tra la struttura informativa e struttura prosodica esiste una corrispondenza (Moneglia-Raso 2014).

Il quadro teorico della Teoria della Lingua in Atto, oltre a restituirci un metodo di analisi che non può prescindere dalle caratteristiche fi-

siche del parlato, ci fornisce gli strumenti utili per poter segmentare il flusso del parlato in unità prosodiche su base percettiva.

Il formato di trascrizione utilizzato è CHAT-LABLITA ed è stato creato in conformità con l'approccio teorico che implementa il formato CHAT, creato nell'ambito del progetto CHILDES, includendo l'intonazione e la sua funzione di demarcazione delle unità di enunciazione e di informazione (Cresti 2000; Cresti-Moneglia 2005). Il flusso del parlato è segmentato percettivamente in unità tonali segnate da pause prosodiche che possono essere terminate o non terminate. L'unità prosodica terminata determina i confini dell'enunciato ed è rappresentata con due barre //; mentre l'unità prosodicamente non terminata identifica le altre unità prosodiche all'interno dell'enunciato ed è rappresentata con una sola barra /. Per la trascrizione di altri fenomeni il formato fornisce un repertorio completo come è illustrato nella tabella 2.

Simbolo	Valore
//	Break prosodico terminale
?	Break prosodico terminale con intonazione interrogativa
…	Break prosodico terminale con intonazione sospensiva
+	Break prosodico terminale per sequenza interrotta
/	Break prosodico non terminale
[/]	Falsa partenza con ripetizione
&	Vocalizzazione o frammento di parola
hhh	Fenomeno paralinguistico o non linguistico come tosse o risata
xxx	Parola non comprensibile

Tabella 2 Simboli del sistema di trascrizione CHAT-LABLITA

La natura dialogica dell'evento è stata riportata nella costruzione del *template* di annotazione attraverso l'uso del software ELAN.

Per il modello di Cresti la struttura informativa è pienamente corrispondente alla struttura prosodica, quindi le unità prosodiche, delimitate dai break prosodici, esprimono un valore informativo.

Nell'ottica della Teoria della Lingua in Atto è possibile rintracciare una corrispondenza tra le unità prosodiche di tipo *root* ('t Hart-Col-

LIER-COHEN 1990), che sono necessarie e sufficienti per la realizzazione di un pattern tonale, alle unità informative di tipo *comment*. Mentre alle unità prosodiche di tipo *prefix* corrispondono le unità informative di *topic* e a quelle prosodiche di tipo *suffix* le unità informative di appendice (FIRENZUOLI 2003; MONEGLIA-RASO 2014). La corrispondenza tra *pattern tonale* e *pattern informativo* risulta rappresentate in modo efficace da una tabella presente in MONEGLIA-RASO 2014 e che riportiamo di seguito:

Prosodic pattern			**Information pattern**		
	Root			Comment	
(Prefix)		(Suffix)	(Topic)		(Appendix)
	(Introducer)			(Locutive Introducer)	
	(Parenthetical)				
(Incipit)		(Phatic)	(Incipit)		(Phatic)

Tabella 3 Corrispondenza struttura prosodica e struttura intonativa (adattata da Moneglia-Raso 2014)

Il tags-set utilizzato per l'annotazione del parlato tiene conto, pertanto, delle unità prosodiche con l'aggiunta di ulteriori tag per il discorso riportato, contrassegnando le diverse etichette con "_r". Abbiamo inoltre dedicato un'etichetta per le unità interrotte e una per le unità prosodiche non decifrabili. Nella tabella 4 riportiamo il tag-set per il parlato.

TAG	DEFINIZIONE
ROOT	unità prosodia prominente, necessaria e sufficiente per la realizzazione dell'enunciato
PREFIX	unità prosodica opzionale e subordinata, che occupa una posizione temporalmente antecedente a una Root, una Suffix o un'altra Prefix
SUFFIX	unità prosodica opzionale e subordinata, temporalmente segue le unità di Root o di Prefix

TAG	DEFINIZIONE
INCIPIT	unità prosodica intonativamente opzionale e subordinata, occorre a inizio turno o enunciato ed è lessicalmente caratterizzate
PHATIC	unità prosodiche subordinata e opzionale, occorre in qualsiasi posizione all'interno dell'enunciato e svolge la funzione comunicativa per il mantenimento dell'apertura del canale
PARENTHETICAL	unità intonative con profilo prosodico basso e realizzata con velocità maggiore rispetto al resto dell'enunciato di cui rappresenta un'inserzione contenutistica
ROOT_r	unità ROOT di parlato riportato
PREFIX_r	unità PREFIX di parlato riportato
SUFFIX_r	unità SUFFIX di parlato riportato
INCIPIT_r	unità INCIPIT di parlato riportato
PHATIC_r	unità PHATIC di parlato riportato
PARENTHETICAL_r	unità PARENTHETICAL di parlato riportato
INTERRUPTED	unità prosodica interrotta
EMPTY	unità prosodica non interpretabile

Tabella 4 Tag-set della struttura intonativa di CORMIP

Il tag-set appena riportato ci ha permesso di annotare il flusso del parlato, segmentato percettivamente, restituendo un elenco di unità di base la cui concatenazione, insieme alla sincronizzazione con gli altri canali di espressione linguistica, permette la codifica di valori informativi, illocutivi e semantici. Di seguito vedremo su quali basi abbiamo costruito la trascrizione e l'annotazione del gesto, cercando di mantenere fede all'approccio percettivo e dando dunque risalto al movimento e al canale visivo. L'annotazione sul software ELAN apparirà come nella figura 1.

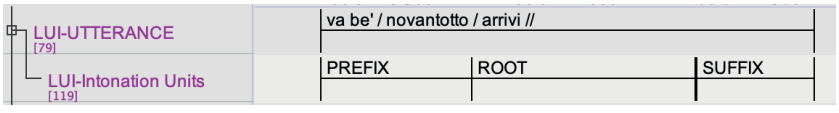

Figura 1 Esempio di annotazione del parlato di CORMIP su ELAN

4.4 Trascrizione e annotazione del gesto

In un quadro linguisticamente complesso è necessario affermare che il carattere multimodale della lingua – e dunque la sua natura polisemiotica – debba necessariamente indurci a indagare come i diversi sistemi semiotici riescano a formare un sistema unico e coerente. Alla luce di quanto affermato da De Mauro 2000, i gesti esprimono diverse funzioni in relazione al loro grado di convenzionalizzazione a cui si associano un crescente grado di iconicità e articolazione. Così, al Kendon's continuum (McNeill 1992) si potrebbe associare un continuum di funzioni che vede la corrispondenza tra un gesto iconico e una funzione lessicale e a gesti meno iconici, in cui è difficile distinguere elementi di articolazione, funzioni soprasegmentali.

Nell'attuare il principio di trascrizione e annotazione su base percettiva abbiamo scelto di utilizzare le unità di analisi di Kendon (1972, 1980, 2004). Se nel parlato il flusso fonico è percepito uditivamente, e quindi trascritto e annotato sulla base di ciò che l'annotatore percepisce[1], per il gesto entrano in campo il movimento, l'iconicità e di conseguenza il canale visivo.

La segmentazione è avvenuta su diversi livelli, secondo l'architettura del gesto illustrata da Kendon 2004. Nel livello più alto abbiamo segmentato le *Gesture Unit* (G-UNIT), l'unità gerarchicamente maggiore che ingloba l'intera escursione del movimento, ed è visivamente riconoscibile perché definita dall'inizio del movimento delle mani fino al loro ritorno in posizione di riposo. Il secondo livello di segmentazione è rappresentato dalle *Gesture Phrase* (G-PHR) ed è l'unità che suddivide la *gesture unit*. Può essere definita come l'unità a cui corrisponde un significato comunicativo e si identifica per un particolare movimento nello spazio o per una particolare configurazione delle mani. Il livello più basso è rappresentato dalle *Gesture Phases*, cioè le unità che compongono una *Gesture Phrase* e si distinguono: la *preparation* che è la fase di preparazione del movimento, l'unità di *stroke* che è la fase culminante del gesto ed è l'unita necessaria e suffi-

1 Nei passaggi più complessi abbiamo risolto i dubbi attraverso l'uso di Praat.

ciente e la *retraction*, che è la fase in cui le mani o le braccia tornano in una posizione di riposo. Inoltre, riconoscendo l'importanza assunta dall'espressione facciale nell'esprimere significati o nel modificarli attraverso l'espressione di un'attitudine o di un'azione linguistica, abbiamo ritenuto necessario aggiungere al primo livello di segmentazione un'etichetta generica che segni la presenza di un'espressione facciale.

La fase di segmentazione e trascrizione del gesto, che è stata fatta separatamente per la mano destra e per la mano sinistra, corrisponde a un'unica fase che si declina su tre livelli.

Nella tabella 5 riportiamo in sintesi i livelli di trascrizione e i simboli utilizzati.

LIVELLO	SIMBOLO	DEFINIZIONE
I	G	unità gerarchicamente maggiore e ingloba l'intera escursione del movimento
I	FACIAL EXPRESSION	espressione facciale
II	PHR	l'unità a cui corrisponde un significato comunicativo e si identifica per un particolare movimento nello spazio o per una particolare configurazione delle mani
III	STROKE	fase culminante del gesto ed è l'unita necessaria e sufficiente
III	PREPARATION	fase di preparazione del movimento
III	RETRACTION	la fase in cui le mani o le braccia tornano in una posizione di riposo

Tabella 5 Tag-set per la trascrizione gestuale di CORMIP

Mentre nella figura 2 mostriamo un estratto di trascrizione del gesto estratta dal software, in cui è possibile notare l'organizzazione dei tiers che riporta la natura gerarchica dell'architettura del gesto.

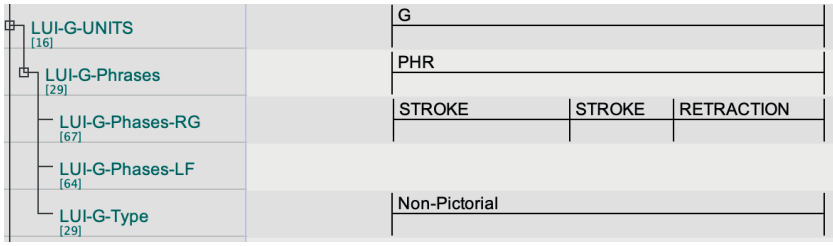

Figura 2 Esempio di annotazione del gesto in CORMIP su ELAN

Nel *corpus* non è stata aggiunta nessuna annotazione gestuale che si basi su categorizzazioni o classificazioni di funzioni di relazioni con il parlato e/o di gradi convenzionalità del gesto. Perciò sono state utilizzate etichette di natura generale per restituire minime informazioni di base utili ai ricercatori. Queste categorie si basano sull'importanza dell'iconicità per il gesto che si esprime in modo graduale anche in relazione alle funzioni espresse dal gesto[2].

Così i gesti sono stati divisi in *Pictorial*, *Non-Pictorial* e *Conventional*. Nello specifico, con l'etichetta *Pictorial* abbiamo cercato di raggruppare tutti i gesti che assumono visivamente una forma riconducibile a un'immagine, a un contorno di un oggetto o a un'azione con riferimento a oggetti e cose del mondo reale. Sotto questa etichetta rientrano tutti i gesti che possono essere classificabili come iconici, pittografici, ideografici o rappresentativi, sulla base delle strategie di rappresentazione individuate da Müller 2013. Mentre con l'etichetta *Non-Pictorial* sono annotati tutti quei gesti che assumono movimenti ritmici (come i batonici) o forme non assimilabili a oggetti nel mondo (come forme geometriche). Mentre con l'etichetta *Conventional* sono etichettati quei gesti che hanno raggiunto un grado di convenzionalità tale che, in un

2 Tra gli altri (Efron 1972; Ekman-Friesen 1969) anche Kendon 2004 ha stilato le diverse funzioni espresse ed esprimibili dai gesti distinguendo le funzioni pragmatiche (quelle di parsing, quelle modali che esprimono la modalità di interpretazione di un enunciato e quelle performative che indicano il tipo di atto linguistico), funzioni interattive o interpersonali relativi alle funzioni dialogiche interazionali e le funzioni referenziali.

determinato sistema linguistico, a quel gesto è possibile associare un valore semantico. Nella tabella 6 riportiamo il tagset del gesto.

TAG	DEFINIZIONE
PICTORIAL	gesti che riproducono una forma riconducibile a un'immagine o a un'azione
NON-PICTORIAL	gesti che riproducono movimenti ritmici o forme non assimilabili a oggetti nel mondo
CONVENTIONAL	gesti convenzionali a cui è possibile associare un valore semantico riconosciuto dalla comunità dei parlanti

Tabella 6 Tag-set per l'annotazione gestuale in CORMIP

È necessario che l'indagine sulla lingua da un punto di vista multimodale si traduca nello studio del comportamento di unità di analisi dei diversi canali per evitare categorizzazioni che rischiano di ridurre il gesto a mero ausilio del canale verbale.

4.5 Dall'azione linguistica all'analisi multimodale

La Teoria della Lingua in Atto (Cresti 2000), sulla base della simultaneità degli atti linguistici teorizzata da Austin 1962, ha derivato la possibilità di porre in relazione l'atto linguistico con l'enunciato (Cresti 2005). In questa relazione la prosodia è considerata l'interfaccia tra l'atto illocutivo e quello locutivo e rappresenta il mezzo necessario per trasdurre la concezione pragmatica in entità concreta e udibile, che è l'*utterance* (Cresti 2020). Nella struttura dell'enunciato, la forza illocutiva è espressa dall'unità informativa di *Comment* e che prosodicamente corrisponde all'unità di *Root*.

Nel corso degli studi su *corpora* portati avanti dal gruppo di ricerca LABLITA, è stato costituito un repertorio di tipi illocutivi distribuiti su cinque classi generali identificate in:

- *Rifiuto*: un atteggiamento di libertà e separazione dal parlante dall'interlocutore, che permette uno scontro con quest'ultimo, una richiesta di sua trasformazione;
- *Asserzione*: un atteggiamento di certezza del parlante nei confronti dell'interlocutore, sicurezza che consente di proporre giudizi, scoperte, valutazioni, rappresentazioni, come oggetti nuovi al mondo;
- *Espressione*: un atteggiamento di manifestazione "estetica" di stati d'animo, emozioni e credenze;
- *Rito*: un atteggiamento esterno di assolvimento di compiti linguistici che hanno effetti legali e sociali e che possono essere compiuti con la minima partecipazione affettiva.

Queste cinque classi generali sono state usate per l'annotazione dell'illocuzione, il cui valore è stato determinato dall'interpretazione dell'azione multimodale, il cui nucleo espressivo di riferimento rimane l'unità tonale di tipo Root. Nella tabella 7 riportiamo il tag-set per l'annotazione delle classi illocutive.

TAG	DEFINIZIONE
ASSERTION	Atto linguistico assertivo
DIRECTION	Atto linguistico direttivo
EXPRESSION	Atto linguistico espressivo
RITE	Atto linguistico rituale
REFUSAL	Atto linguistico di rifiuto

Tabella 7 Tag-set dei valori illocutivi in CORMIP

Nello specifico la nostra annotazione dell'illocuzione ha seguito tre principi:

- se il parlante esprime l'azione linguistica attraverso l'uso esclusivo del canale verbale, viene annotato il valore illocutivo espresso dall'unità prosodica di tipo Root;
- se il parlante esprime l'azione linguistico attraverso l'uso esclusivo del canale gestuale (compreso l'espressione faccia, o la co-occorren-

za tre gesto manuale ed espressione facciale), viene annotato il valora illocutivo espresso dall'unità gestuale corrispondente;
- se il parlante esprime l'azione linguistica attraverso l'uso concomitante del canale verbale e gestuale compreso l'espressione faccia, o la co-occorrenza tre gesto manuale ed espressione facciale, viene annotato il valore illocutivo valutando in modo complessivo espresso dalla co-occorrenza delle relative unità gestuale e verbali.

I principi appena elencati si basano su due presupposti. Il primo è che il valore illocutivo – che esprime l'intenzione comunicativa del parlante di natura affettiva – si trasforma fisicamente in azione linguistica attraverso il potenziale uso delle diverse modalità (Cresti 2020) e, aggiungiamo noi, anche attraverso l'uso dei gesti e dell'espressione facciale. Il secondo presupposto è il giudizio dell'annotatore. Infatti, se da un lato la teoria della lingua in atto ha sistematizzato secondo quali parametri pragmatici, semiologici e cognitivi si definiscono i diversi valori illocutivi (Cresti 2005; 2020) dall'altro mancano dei parametri corrispettivi per giudicare il valore illocutivo espresso dalla gestualità o dalla co-occorrenza di gesto e parlato. Pertanto, in questi casi, abbiamo fatto fede al giudizio analitico dell'annotatore in quanto, come già detto, ci approcciamo all'annotazione dell'illocuzione nel nostro *corpus* sperimentalmente.

La natura sperimentale dell'operazione mira a utilizzare l'annotazione dell'azione linguistica come mezzo per indagare l'esistenza e le caratteristiche dell'unità multimodale. L'atto linguistico non rappresenta un'unità di analisi, ma un approccio metodologico di tipo pragmatico all'intenzione comunicativa espressa dai parlanti attraverso l'uso di diversi canali e/o della loro coordinazione. Al momento non siamo in grado di poter esprimere definizioni inerenti alle nozioni di atto linguistico multimodale, ma senza dubbio il *corpus* rappresenta uno strumento utile a indagarlo.

5. Conclusioni

In queste pagine è stato riportato e discusso il metodo usato per realizzare un *corpus* multimodale pilota dell'italiano parlato in contesto spontaneo (CORMIP). L'obiettivo è stato quello di creare uno strumento per indagare la lingua come sistema multimodale. Un ruolo importante nell'annotazione è svolto dall'azione linguistica che, all'interno del template annotativo, svolge un ruolo chiave di interpretazione del carattere multimodale dell'enunciato. In conclusione, possiamo affermare che questo nostro approccio ci ha permesso di costruire un *corpus* pilota – che raccoglie sei brani di interazioni spontanee in contesti diversi e in diverse tipologie di interazioni – e di indagare le interazioni comunicative da un punto di vista multimodale senza però creare uno strumento che possa apparire centrato maggiormente su una delle due modalità. Questo è stato possibile grazie all'approccio percettivo e pragmatico al fenomeno lingua. È possibile, inoltre, sostenere la necessità di *corpora* multimodali sempre più grandi contenenti dati spontanei e realizzati attraverso l'uso di standard largamente condivisi.

Bibliografia

Austin 1962 = John Langshaw Austin, *How to do things with words*, Oxford, Oxford University Press, 1962.

Bressem-Ladewig-Müller 2013 = Jana Bressem, Silva H. Ladewig e Cornelia Müller, *Linguistic Annotation System for Gestures*, in *Body - Language - Communication. An International Handbook on Multimodality in Human Interaction*, edited by Cornelia Müller, Alan Cienki, Ellen Fricke, Silva Ladewig, David McNeill and Sedinha Tessendorf, «Handbücher Zur Sprach-Und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science (Hsk)», 38/1, Berlin-Boston, De Gruyter Mouton, 2013, pp. 1098-1124.

Cresti 2005 = Emanuela Cresti, *Per una nuova classificazione dell'illocuzione*, in *Tradizione e innovazione. Atti del VI Convegno Internazionale della SILFI – Gerhad-Mercator Universit (Duisburg, 28 giugno-2 luglio 2000)*, a cura di Elisabetta Burr, Firenze, Franco Cesati Editore, pp. 233-246.

Cresti-Moneglia 2005 = *C-ORAL-ROM. Integrated reference corpora for spoken romance languages*, eds. by Emanuela Cresti, Massimo Moneglia, Amsterdam, John Benjamins, 2005.

Cresti 2000 = Emanuela Cresti, *Corpus di italiano parlato*, Firenze, Accademia della Crusca, 2000.

Cresti 2020 = Emanuela Cresti, *The pragmatic analysis of speech and its illocutionary classification according to the Language into Act Theory*, in *In search of basic units of spoken language: A corpus-driven approach*, eds. by Shlomo Izre'el, Heliana Mello, Alessandro Panunzi, Tommaso Raso, Amsterdam, John Benjamins, 2020, pp. 181-219.

De Mauro 2000 = Tullio De Mauro, *Vocalità, gestualità, lingue segnate e non segnate*, in *Viaggio nella città invisibile*, a cura di C. Bagnara, G. Chiappini, M.P. Conte, M. Ottolini, Pisa, Edizioni del Cerro, 2000, pp. 17-45.

Efron 1972 [1941] = David Efron, *Gesture, race and culture*, The Hague, Mouton, 1972.

Ekman-Friesen 1969 = Paul Ekman e Wallace V. Friesen, The Repertoire of Nonverbal Behavior: Categories, Origins, Usage and Coding, in «Semiotica», 1, 1, 1969, pp. 49-98.

Enfield 2009 = Nick J. Enfield, *The Anatomy of Meaning: Speech, Gesture, and Composite Utterances*, Language Culture and Cognition, Cambridge, Cambridge University Press, 2009.

Enfield 2013 = Nick J. Enfield, *Composite Utterances approach to meaning*, in *Body - Language - Communication. An International Handbook on Multimodality in Human Interaction*, edited by Cornelia Müller, Alan Cienki, Ellen Fricke, Silva Ladewig, David McNeill and Sedinha Tessendorf, «Handbücher Zur Sprach-Und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science (Hsk)», 38/1, Berlin-Boston, De Gruyter Mouton, 2013.

Enfield 2009 = Nick J. Enfield, *The Anatomy of Meaning: Speech, Gesture, and Composite Utterances*, Language Culture and Cognition, Cambridge, Cambridge University Press.

Firenzuoli 2003 = Valentina Firenzuoli, *Le Forme Intonative di Valore Illocutivo dell'Italiano Parlato: Analisi Sperimentale di un Corpus di Parlato Spontaneo (LABLITA)*, Tesi di Dottorato, Università degli Studi di Firenze.

Gibbon et al. 2003 = Dafydd Gibbon, Ulrike Gut, Benjamin Hell, Karin Looks, Alexandra Thies, Thorsten Trippel, *A Computational Model of Arm Gestures in Conversation*, in Eighth European Conference on Speech Communication and Technology, Geneva, Switzerland, 2003.

Goodwin 2000 = Charles Goodwin, *Action and embodiment within situated human interaction*, «Journal of Pragmatics», 32, 2000, pp. 1489-1522.

't Hart-Collier-Cohen (1990) = Johan 't Hart, Rene Collier, Antoine Cohen, *A Perceptual Study of Intonation. An Experimantel-Phonetic Approach to Speech Melody*, Cambridge University Press, Cambridge, 1990.

Kress-van Leeuwen (2001) = Gunther Kress, Theo van Leeuwen, *Multimodal Discourse: The Modes and Media of Contemporary Communication*, London, New York, Edward, 2001.

Hockett 1960 = Charles F. Hockett (1960), *The origin of speech*, in «Scientific American», 203, 1960, pp. 88-96.

Kendon 1972 = Adam Kendon, *Some relationships between body motion and speech: an analysis of an example*, in *Studies in Dyadic Communication*, edited by Aron Wolfe Siegman e Benjamin Pope, Elmsfiord, New York, 1972, pp 69-89.

Kendon 1980 = Adam Kendon, *Gesticulation and speech: Two aspects of the process of utterance*, in *The relationship of verbal and nonverbal communication*, edited by Mary R. Key, Berlin-New York, De Gruyter Mouton, 1980, pp. 207-228.

Kendon 2004 = Adam Kendon, *Gesture: Visible Action as Utterance*, Cambridge, Cambridge University Press, 2004.

Labov 1972 = William Labov, *Sociolinguistic Patterns*, Philadelphia, PA, University of Pennsylvania Press, 1972.

Lausberg 2013 = Hedda Lausberg, *NEUROGES. A Coding System for the Empirical Analysis of Hand Movement Behaviour as a Reflection of Cognitive, Emotional, and Interactive Processes*, in *Body - Language - Communication. An International Handbook on Multimodality in Human Interaction*, edited by Cornelia Müller, Alan Cienki, Ellen Fricke, Silva Ladewig, David McNeill and Sedinha Tessendorf, «Handbücher Zur Sprach-Und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science (Hsk)», 38/1, Berlin-Boston, De Gruyter Mouton, 2013, pp. 1022-1037.

Lausberg-Sloetjes 2016 = Hedda Lausberg, Han Sloetjes, *The revised NEUROGES-ELAN: An objective and reliable interdisciplinary analysis tool for nonverbal behavior and gesture*, «Behavior Research Methoids», 48, 2016, pp. 973-993.

Martinet 1960 = André Martinet, *Éléments de linguistique Générale*, Paris, Armand Colin, 1960.

McNeill 1992 = David McNeill, *Hand and Mind: What Gestures Reveal about Thought*, Chicago, University of Chicago Press, 1992.

Moneglia-Raso 2014 = Massimo Moneglia, Tommaso Raso, *Notes on Language into Act Theory (L-AcT)*, in *Spoken corpora and linguistic studies*, edited by Tommaso Raso, Heliana Mello, Amsterdam, John Benjamins, 2014, pp. 468-495.

Müller 2013 = Cornelia Müller, *Gestural Modes of representation as techniques of depiction*, in *Body - Language - Communication. An International Handbook on Multimodality in Human Interaction*, edited by Cornelia Müller, Alan Cienki, Ellen Fricke, Silva Ladewig, David McNeill and Sedinha Tessendorf, «Handbücher Zur Sprach-Und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science (Hsk)», 38/1, Berlin-Boston, De Gruyter Mouton, 2013, pp. 1687-1702.

Trippel et al. 2004 = Thorsten Trippel, Dafydd Gibbon, Alexandra Thies, Jan-Torsten Milde, Karin Looks, Benjamin Hell, and Ulrike Gut, *CoGesT: A Formal Transcription System for Conversational Gesture*, in Proceedings of the Fourth International Conference on Language Resources and Evaluation (Lrec'04), Lisbon, Portugal, 2004.

Riassunto Questa ricerca affronta la questione metodologica per la compilazione del corpus multimodale di lingua parlata CORMIP, proponendo una possibile soluzione alle questioni legate alla gestione dei dati linguistici multimodali. I problemi riguardano tutte le fasi della compilazione, dalla raccolta dei dati fino alla loro trascrizione e annotazione.

Abstract This research addresses the methodological issue for the compilation of the multimodal spoken language corpus CORMIP. It proposes a possible solution to the issues related to the management of multimodal language data. The problems concern all stages of compilation, from data collection to data transcription and annotation.