

# Gesture and Part-of-Speech Alignment in Dialogues

Zohreh Khosrobeigi, Maria Koutsombogera, Carl Vogel

Computational Linguistics Group

Trinity College, the University of Dublin

Dublin 2, Ireland

{khosrobz, koutsomm, vogel}@tcd.ie

## Abstract

This paper studies the relation between language and gesture in interaction by investigating the temporal alignment of gestures and the words they co-occur with in a corpus of task-based dialogues. Specifically, we examine the relationship of different semiotic gesture types (their quantity and duration) with the syntactic categories assigned to the words that coincide with the gesture. We observe that different types of gesture-word alignment present different distributions, as well as different associations depending on the gesture type.

## 1 Introduction

We seek to understand the interfaces between gesture and language. It has been suggested that gesture accompanying linguistic content has functions tied more closely to speakers' formulation of uttered thoughts than to facilitating hearers' grasp of a shared interpretation of expressed statements (Hadar, 1989; Tuite, 1993; McNeill, 1997; Kita, 2000). Evidence for the perspective is in the fact that people may frequently be witnessed gesturing while speaking without a visual modality of communication, via a telephone, for example, even if those gestures are reported to be smaller in extent and shorter in duration than those occurring during visual contact (Bavelas et al., 2008). Other aspects of language as used in communication are marked by economizing production effort: an introduction of an entity into discourse may be initially accompanied by adjective and prepositional phrases and relative clauses, but quickly after introduction, relegated to pronouns. In contrast, gesture consumes rather more physical energy than uttering overt nominals, and may be seen at all stages of a conversation.<sup>1</sup> However, gesture does not appear to

<sup>1</sup>We take it to be self-evident that moving arms, etc. requires more energy than the movements required by speech. However, for indirect support of this claim we not past work (Pouw et al., 2020) that shows greater variation in F0 and

be randomly distributed. Further, even if gesture is not performed primarily for the benefit of a listener, a listener may interpret gesture, partly on the basis of how they gesture themselves (to the extent that they are aware of how they gesture themselves).

Public gesture has systematic properties that enable consensus on the classification of a small number of semiotic gesture types – Beats, Symbolics, Iconics, and Deictics, for example, and further, those other movements are none of those.<sup>2</sup> Some research addresses the relationship between gesture and intonation (Steedman, 1991; Loehr, 2004; Jannedy and Mendoza-Denton, 2005; Loehr, 2012), and some work on gesture semiotic type and word meaning relation (Bernardis and Gentilucci, 2006; De Marco et al., 2015). While information structure and intonation outlines conform in a manner somewhat at odds with syntactic constituent structure, word-level categories have been modeled as carrying the relevant information (Steedman, 1991). Further, it has been found that there are strong asymmetric links between gesture,<sup>3</sup> using a more fine-grained set of gesture types than described above, and part of speech categories (Mehler et al., 2012), but without reporting on the effects of individual gesture types and part of speech types. These are among reasons it is interesting to explore word categories, starting with the part of speech categories that are addressed in this study.

Here we explore whether there are systematic facts in the distribution of gesture types in collab-

amplitude in vowel expression with gesture than without; thus, speech and gesture requires more energy than speech without gesture (the work cited explores the theory that synchrony of speech and gesture is more of a mechanical process than a cognitive one).

<sup>2</sup>Curiously, a “miscellaneous” movement may still contain private, and public meaning: scratching an eyebrow may be what one person does while puzzling on something, and what another does only when conversations are lively and the agent is confident it will not be noticed.

<sup>3</sup>They see stronger evidence of gestures being selected by words than vice versa.

orative conversations in relation to the syntactic categories of words that are nearby. One could imagine that Deictics mainly occur with nominals, Symbolics mainly with verbals, and Iconics split between those categories. However, it is necessary to be precise about what “with” means. MacNeill’s hypothesis would suggest that gesture onset is typically prior to the accompanying word onset but with some extent of co-temporality.

The next section describes related work. Section 3 reports on the datasets used. Section 4 explains the methods followed for forced alignment, text-gesture alignment as well as POS tagging. The data profiling according to the alignment categories is reported in Section 5 and analyzed and discussed further in Section 6, while Section 7 presents the conclusions.

## 2 Related Work

Co-speech gestures are visible hand actions that are produced while speaking. Several views have been reported in the literature with respect to the role of gestures in communication, i.e., as an indispensable part of the language system (McNeill, 1992, 2005; Kendon, 2004) or the overall perspective of speaking as a multimodal construct (Cienki and Muller, 2008). There is a large amount of research surrounding the relationship between gesture and speech. This relationship can be viewed from its cognitive perspective, namely how gestures are linked to our thought (McNeill, 2005) as well as from its communicative perspective, i.e., how gestures regulate the organization of the interaction (Kendon, 2004). Theoretical research suggests that speech and gesture share a common conceptual origin and that they form a single integrated system (McNeill, 1992; McNeill and Duncan, 2000; Rieser, 2015); and that both speech and gesture have communicative functions that come from the same communicative intention (de Ruiter, 2000).

In this way, hand gestures help speakers talk, think, and disclose information that cannot be verbalized (Goldin-Meadow et al., 1993); and at the same time, performing gestures helps speakers organize visuo-spatial information into units that are compatible with the format of speech (Kita and Özyürek, 2003; Hahn and Rieser, 2010).

We study the temporal relation between gesture and speech seeking insight into the nature of their links. Words that match most closely the meaning of a gesture have been termed lexical affili-

ates (Schegloff, 1985). McNeill (1992) defined three rules of synchronization between gesture and speech, namely the phonological synchrony rule, predicting that a gesture stroke should occur before the most prominent syllable; the semantic synchrony rule predicting that co-occurring gestures and speech relate to the same idea unit; and the pragmatic synchrony rule predicting that co-occurring gestures and speech have the same pragmatic function (Wagner et al., 2014). While other works in this space address gesture morphology (Hahn and Rieser, 2010; Rieser and Lawler, 2020), we focus on the semiotic type of gestures without reference to gesture-internal phases.

The present study investigates the dependencies of gestures and the grammatical categories (part-of-speech - POS) of the words co-occurring with gestures. We study this in the totality of the gestures occurring in a multimodal corpus, and we consider gestures of all semiotic types. We use the manual transcripts of the corpus dialogues as well as existing manual annotations of gestures of dialog participants. To temporally identify lexical correlates, we use timestamps from manual word and gesture annotation. The dialog transcripts were further annotated automatically with POS tags.

## 3 Resources

### 3.1 Dataset

To study the distribution of the gesture types co-occurring with syntactic categories of words, we use the MULTISIMO corpus (Koutsombogera and Vogel, 2018), a multimodal dataset of three-party, task-based dialogues which were collected to investigate different aspects of collaborative interaction. The dataset consists of 18 dialogue sessions. In each session, two players collaborate with each other in English to answer three questions and rank the answers and are aided by a moderator who provides guidance. The dataset includes the video and audio of the dialogues, as well as a set of annotations including speech transcripts, gaze, laughter, and gesture annotations. There are 39 dialogue participants, 16 of which are native English speakers.

Hand gestures were manually annotated using the ELAN editor (Brugman and Russel, 2004). The entire duration of a gesture was annotated, i.e. the transition from a neutral position to the gesture as well as the return to the neutral position once the gesture concluded; that is, including preparation, stroke and retraction, as well as gesture holds, if

applicable. Also, the annotation scheme does not distinguish between the use of one or two hands performing the gesture. Once a gesture is visible, the start and end time of that gesture is marked and assigned with one of the following semiotic types: Beat, Iconic, and Deictic and Symbolic. The semiotic categories defined in the annotation scheme are based on McNeill (McNeill, 1992), who, in turn, built on Peirce’s semiotic types (Peirce, 1931).

*Beat* gestures are utilized in rhythm with utterances in order to emphasize what is being said or to improve the coherence of the statement for the listener. *Iconic* gestures provide a pictorial representation of any concrete or abstract concept, e.g. moving the hand upwards or downwards while ranking the answers. *Deictic* gestures are usually depicted by pointing at a particular object or individual, whether they are real or imaginary. They are commonly used in the corpus by one participant to point at another participant, as if to encourage a contribution to the discourse. *Symbolic* gestures are culture-specific gestures where the relation between their shape and the accompanying speech is based on social conventions, such as the thumbs up gesture (to denote agreement) or the OK symbol.

In addition to those four types, the label N/A was used for visible hand movements, which, however, did not have a communicative function in the dialogue. Gestures were annotated by one rater and the annotation was validated by a subject matter expert, who had frequent interaction with the rater to monitor the task and to discuss, among others, difficult or ambiguous cases.

Apart from the data described above, we extracted some additional features and data from MULTISIMO: In order to analyze the gestures along with the rest of the data, at first, gesture timestamps of each player were exported from ELAN. Then, the files were pre-processed to keep the information related to the onset, offset and type of gestures. The duration of gestures, of 14 out of 18 dialogues, (mean = 1573 msec) is approximately five times greater than the duration of words (mean = 300 msec), and there are fewer than 50 gesture tokens in each session. The number of each semiotic gesture type (1004 instances of gestures), as well as the number of spoken word tokens, are counted in Table 1.

#Beat	#Iconic	#Deictic	#Symbolic	#N/A	#Word
374	251	64	15	300	12862

Table 1: Count of gestures per gesture type and count of spoken words of all dialogue players.

### 3.2 New Dataset for Text Alignment

An important aspect of dialogue analysis is understanding the factors that influence the alignment of the communication channels available – linguistic content, back-channels, social signals, laughter, gesture, gaze, and so on. All aspects convey meaning, although not always about the dialogue’s linguistic content (sometimes, about the participants’ level of engagement, sometimes about their personal relationships, etc.). Nonetheless, we take the linguistic content as the focal point of dialogue, and seek to understand alignment with respect to the linguistic form of that content. This entails requiring knowledge of the timing of the words spoken – temporal onset and offset for each item.

The onset and offset of linguistic content are important information when studying the relation of different channels in multimodal interaction. To identify the relation between audio and text in the dialogues, we use two streams of information: the audio channel of participants’ speech and the transcript of the dialogues, performed at an utterance level. To be able to align the audio and the transcript at a word level, we labeled the onset and offset of each word in speech manually using the transcripts and monophonic audio files for each speaker with the Praat software (Boersma and van Heuven, 2001). Through Praat, the audio files and their corresponding transcripts are processed to define the start and end of each word. The output CSV files include the onset and offset of words and the words. The text alignment was done for the participants that had the player role, for 14 out of the 18 corpus dialogues.<sup>4</sup> Each dialogue needed about 8 hours to label.

## 4 Methods

Our research aim here is to identify the major syntactic category that is used most frequently during, before, or after a hand gesture occurrence. We approach this using word-level rather than phrasal

<sup>4</sup>The alignment of the remaining four dialogues is currently in progress. The aligned CSV files are available from the MULTISIMO website (<http://multisimo.eu/datasets.html>).

constituent-level labeling of POS. To answer this question, we first performed temporal text and gesture alignment. Then, each word was labeled with its POS tag. Each of these steps is explained below.

#### 4.1 Temporal Gesture-Word Alignment

Players perform gestures while speaking. The gestures may be short, long, or located in any part of an utterance. They can be semantically related to words that precede, follow, or are uttered simultaneously with the gesture. Using text alignment and gestures alignment data, an alignment of gestures and words is possible at different times of occurrence in relation to each other. To align gestures and words, the occurrence of spoken words was computed in rather than gestures at various times of happening. All possibilities of occurrence are seven categories. Table 2 encapsulates the explanation of each temporal gesture-word alignment and shows a graphic view for each alignment.

#### 4.2 Part of Speech Tagging

NLTK (Bird et al., 2009) and TreeTagger (Schmid, 1994) are used to categorize words with POS labels. Applied to this data, the systems differ in many words. NLTK tags our dataset using 26 different POS labels, of which the most used is a noun. For instance, NLTK tags the verb “think” as “Noun” and the adjective “dirty” as “Noun”. On the other hand, TreeTagger tags the dataset using 51 different POS labels. TreeTagger tags the above examples correctly. The reason for having different tags is that taggers consider different types for each POS. For instance, TreeTagger has several types of a verb, such as VBZ, VBB, VBI, and VBG. For our purposes and also given the relatively small size of the dataset at hand, broader syntactic category labels seem more appropriate. As a result, similar categories are mapped to one main category. Table 3 illustrates the mapping from TreeTagger POS categories to more general category labels. In total, 3807 tags are mapped to eight categories using TreeTagger (including words and non-word vocalizations). We tag non-word vocalizations (e.g “hmm”, laughter, etc.) as NW.

Taggers work with high accuracy on well-structured and standard texts. But in natural dialogue, people do not talk solely in complete grammatical sentences – sometimes, utterances are sentence fragments or ungrammatical. As a result, we tokenized MULTISIMO transcripts, and fed each word in succession to taggers as input. Moreover,

we did not normalize tokens since normalization to lemmas can confuse automatic syntactic labeling.

Table 4 shows the number of POS instances at different alignments using TreeTagger.

To see the reliability of NLTK and TreeTagger, their error rates are estimated on the basis of 183 randomly selected items. For NLTK, 46 of 183 tags are incorrect, and the error rate is 25.12%. For TreeTagger one of 183 tags is incorrect, and the error rate is 0.5%. Considering these error rates, only TreeTagger labellings are analyzed.

When a player performs a gesture and then says a word, some POS types are used more than others. Table 5 illustrates the most used POS for each type of gesture at different alignments.

#### 4.3 Durations

With respect to gestures and the words with which they align, it is interesting to examine durations, not least because these include aspects of production time and execution time. Table 6 shows the central tendencies of durations for each gestural types.

### 5 Data profile

Each word instance participates in at most two alignment categories, falling into more than one category if the token duration overlaps with the duration of successive gestures. Of 12862 words spoken by the players, 9055 words do not align with any gesture (see Table 7 for the distribution of POS categories for these words). Also, 3707 words align with exactly one gesture, and 100 words align with two gestures and enter into two alignment categories. Table 7 illustrates the distribution of POS categories for unaligned words, words aligned with one gesture and with two gestures.

Gesture instances can also enter into more than one alignment category, for example, gest-word-with-overlap and word-gest-with-overlap, when a gesture happens with two different words at different intervals. Of 1004 gestures, 290 gestures enter exactly one category; 714 gestures enter more than one.

We emphasize that the relative frequency of these alignment categories in natural dialogue are not given, *a priori*. The fact that 714 gesture tokens and 100 words are in more than one alignment category necessitates that the instances analyzed in terms of their counts in the contingency tables below be *pairs* of gesture tokens and word tokens. Each pair is independent. For the three alignment








Alignment	Description	Pictures
short-gest	The duration of a gesture is shorter than the duration of a word and occurs within the word duration; hence, it includes only one word. The onset and offset of the gesture are inside the word timestamp.	
long-gest	The duration of a gesture is longer than the duration of a word. As a result, the gesture occurs with a few words simultaneously. The longest gesture in the dataset co-occurs with four words. The onset and offset of words are inside the gesture timestamp.	
gest-word-no-overlap	A gesture occurs before a word, and when the gesture is completed, the word is uttered. The offset of gesture is before the onset of word. The distance between the offset of gesture and onset of word is less than one millisecond.	
word-gest-no-overlap	A gesture starts immediately as soon as a word is finished. The onset of a gesture is after the offset of a word. The distance between the onset of gesture and offset of word is less than one millisecond.	
gest-word-with-overlap	A gesture starts before the beginning of a word, and it ends before that word ends. The offset of gesture is inside the word onset and offset.	
word-gest-with-overlap	A gesture starts in the middle of a word and finishes after the word. The onset of gesture is inside the word onset and offset.	
silent-gest	A person gestures without speaking.	

Table 2: Types of temporal gesture-word alignment.

Main	Abbr.	#Tags	Mapped Tags
Noun	NN	658	NN2=106, NN1=510, NN0=28, NP0=14
Verb	VRB	768	VBB=434, VBI=36, VBZ=130, VM0=79, VBD=40, VBG=29, VBN=20
Adjective	AJ	137	AJ0=134, AJC=1, AJS=2
Adverb	ADV	478	AV0=389, XX0=61, AVQ=21, AVP=7,
Determiner / Pronoun	DP	903	DT0=112, AT0=245, DTQ=25, CRD=48, ORD=39, EX0=16, DPS=24, PNI=13, PNQ=2, PNX=3, PNP=376
Preposition	PRP	195	PRP=145, PRF=46, TO0=4
Conjunction	CJ	182	CJS=46, CJC=134, CJT=2
Interjection	IJ	168	ITJ=168
Non-word	NW	315	NW=315
Sum		3807	

Table 3: Main POS categories and mapped sub-categories aligned to eight categories using TreeTagger.

categories for which the total number of observations exceeds 45, we construct a contingency table analysis to test whether there is an interaction between the alignment category and the part of speech distribution; the interaction is significant

Alignment	ADV	AJ	CJ	DP	IJ	NN	NW	PRP	VRB
short-gest	1	0	0	0	1	2	6	0	0
long-gest	344	97	145	745	115	398	144	171	619
gest-word-no-overlap	0	1	0	0	0	1	0	0	1
word-gest-no-overlap	1	0	1	0	0	0	0	0	0
gest-word-with-overlap	63	22	14	93	30	165	76	12	85
word-gest-with-overlap	69	17	22	67	22	92	89	12	63
Sum	478	137	182	905	168	658	315	195	768

Table 4: Counts of POS instances for each alignment.

Alignment	Beat	Iconic	Deictic	Symbolic	N/A
short-gest	NN=2 NW=2 ADV=1	NW=2			IJ=1 NW=2
long-gest	DP=295 VRB=264 NN=153 ADV=151	DP=284 VRB=221 NN=171	DP=73 VRB=52	VRB=7 DP=5 IJ=6	DP=88 VRB=75 ADV=56
gest-word-no-overlap	VRB=1 NN=1				AJ=1
word-gest-no-overlap	CJ=1	ADV=1			
gest-word-with-overlap	NN=74 DP=46 VRB=44	NN=60 DP=29 VRB=23	NN=18 VRB=9 DP=9	VRB, ADV, IJ=2	NW=26 NN=14 ADV=13
word-gest-with-overlap	NN=40 ADV=35 VRB=37	NN=27 DP=23 NW=22	NN=10 DP=6	NW=3	NN=14 ADV=10 VRB=10 NW=24

Table 5: The most used POS with each type of gesture at different alignments using TreeTagger.

Duration	Beat	Iconic	Deictic	Symbolic	N/A
Word mean	328.7	301.1	266.3	327.9	387.5
Word median	255.2	244.8	220.5	300.1	302.6
Word s.d.	489.0	222.1	170.7	157.5	372.0
Gest. mean	1687.6	2168.4	1609.6	1603.1	2136.0
Gest. median	1513.5	2040.0	1450.0	1640.0	1910.0
Gest. s.d.	891.8	985.3	719.8	643.8	1177.2

Table 6: Word and gesture millisecond duration statistics for aligned gesture-word pairs.

Aligned and unaligned words			
	Aligned words in one category (distinct)	Aligned words in two categories	Unaligned words
NN	627	32	1506
VRB	755	13	1562
AJ	132	5	298
ADV	470	8	1205
DP	891	14	1703
PRP	194	1	300
CJ	180	2	334
IJ	167	1	1032
NW	291	24	1115
SUM	3707	100	9055

Table 7: The distribution of POS for words which are aligned or are not aligned.

( $\chi^2 = 281.66$ ,  $df = 16$ ,  $p < 2.2^{-16}$ ).

Analysis of the Pearson residuals reveals: for long-gest alignments, there are significantly more DP ( $p < 0.05$ ), PRP ( $p < 0.05$ ), and VRB ( $p < 0.05$ ) observations and significantly fewer NN ( $p < 0.05$ ), and NW ( $p < 0.001$ ) observations than would be expected with no interaction; for gest-word-with-overlap alignments, significantly more NN ( $p < 0.001$ ), and NW ( $p < 0.001$ ) observations and significantly fewer CJ ( $p < 0.05$ ), DP ( $p < 0.05$ ), PRP ( $p < 0.05$ ), and VRB ( $p < 0.05$ ) observations than would be expected with no in-

teraction; for word-gest-with-overlap alignments, significantly more NW ( $p < 0.001$ ) and significantly fewer DP ( $p < 0.05$ ), PRP ( $p < 0.05$ ) and VRB ( $p < 0.05$ ) observations than would be expected with no interaction. Thus, there appears to be an interaction between the starting point and span of a gesture and the accompanying parts of speech – DP, PRP and VRB categories are prominent in long-duration gestures; NN and NW are prominent in shorter duration gestures, with NN being most prominent for the short gestures commencing before and ending during the aligned word (gest-word-with-overlap). Considering the token durations, note from Table 8, that for long-duration gestures, the categories significant for the extent of positive observations (DP, PRP and VRB) are also the shortest in duration for that alignment category. For short gestures commencing before and ending during the aligned word the most numerous category (NN) is the second longest in duration for the alignment category. For short gestures commencing in the middle of a token and ending after it, the categories significant in the extent of their positive count (NN and NW) are the longest in duration for the alignment category. Thus, significant counts are not always explained by shorter durations.

**short-gest, gest-word-no-overlap, word-gest-no-overlap Alignments:** Of seven alignment categories, three categories, short-gest Alignment (gestures are short and only include one word), gest-word-no-overlap Alignment (a player gestures and then says a word after the gesture), and word-gest-no-overlap Alignment (a player gestures after finishing a word) have a few instances of POS cate-

POS	Alignment					
	long-gest		gest word		gest word	
	mean	median	mean	median	mean	median
ADV	259.0	235.1	408.1	379.1	343.1	309.2
AJ	353.7	322.6	479.2	478.3	464.9	401.9
CJ	219.0	185.2	<b>320.4</b>	<b>270.6</b>	305.3	230.9
DP	166.6	140.4	<b>241.6</b>	<b>193.3</b>	<b>262.8</b>	<b>217.1</b>
IJ	281.4	270.8	387.2	346.7	255.1	254.9
NN	<b>400.5</b>	<b>382.2</b>	515.4	494.3	673.2	491.7
NW	<b>471.1</b>	<b>422.1</b>	717.5	622.5	849.4	663.6
PRP	166.5	153.4	<b>264.4</b>	<b>271.9</b>	<b>314.1</b>	<b>289.2</b>
VRB	198.1	171.0	<b>432.1</b>	<b>309.2</b>	<b>311.8</b>	<b>277.1</b>

Table 8: The millisecond *durations* of POS instances for frequent alignments. To save space word abbreviates gest-word-with-overlap, and word abbreviates word-gest-with-overlap.

The figures that are boxed or in bold indicate the cells of the contingency of counts for which Pearson residuals were significant ( $p < 0.05$ ), as described in the text: durations in cells are boxed where the corresponding count was significantly greater than would be expected with no interaction between the part of speech and alignment type; durations in cells are bold where the corresponding count was significantly less than would be expected with no interaction between part of speech and the alignment type.

gories aligned with gestures (see Table 5).

**long-gest Alignment:** In long-gest alignment, a gesture is long and co-occurs with a few words. Beat gestures are accompanied by 1092 words, and Iconics by 1037 words. Verb and DP are the most used POS categories in this temporal alignment (see Table 5). Table 9 shows the distribution of POS categories across gesture types.

long-gest Alignment					
POS	Beat	Iconic	Deictic	SYMBOLIC	N/A
ADV	151	104	33	0	56
AJ	40	39	3	1	14
CJ	51	65	12	0	17
DP	295	284	73	5	88
IJ	31	32	9	6	37
NN	153	171	25	2	47
NW	39	58	4	2	41
PRP	68	63	21	3	16
VRB	264	221	52	7	75
Sum	1092	1037	232	26	391

Table 9: The count of POS categories in long-gest alignment by gesture type.

**gest-word-with-overlap Alignment:** A player gestures and starts a word in during the gesture and finishes the gesture before the word. The most accompanied gesture types are Beats (n=258) and

Iconics (n=162). NN is the most used POS with Beat, Iconic, Deictic, and N/A (excluding NW as non-vocalized POS) (see Table 5). Table 10, shows the distribution of POS and gesture types.

gest-word-with-overlap Alignment					
POS	Beat	Iconic	Deictic	Symbolic	N/A
VRB	44	23	9	2	7
NN	74	60	18	0	14
PRP	8	4	0	0	0
DP	46	29	9	0	9
AJ	8	6	2	0	6
ADV	31	12	5	2	13
CJ	6	5	1	0	2
IJ	8	8	3	2	9
NW	33	15	2	0	26
Sum	258	162	49	6	86

Table 10: The distribution of POS categories in gest-word-with-overlap alignment by gesture type.

**word-gest-with-overlap Alignment:** Gestures that start in the middle of a word and finish after the word are categorized as word-gest-with-overlap. The most accompanied gestures in this alignment are Beat (n=208), and Iconic (n=125). NN is used the most in the alignment (excluding Symbolic and non-vocalize POS) (see Table 5). Table 11 shows the POS-gesture types distribution.

word-gest-with-overlap Alignment					
POS	Beat	Iconic	Deictic	Symbolic	N/A
VRB	37	13	3	0	10
NN	40	27	10	1	14
PRP	3	4	2	0	3
DP	29	23	6	0	9
AJ	12	2	0	0	3
ADV	35	19	4	1	10
CJ	7	9	2	0	4
IJ	10	6	2	0	4
NW	35	22	5	3	24
Sum	208	125	34	5	81

Table 11: The distribution of POS categories in word-gest-with-overlap alignment by gesture type.

**silent-gest Alignment:** The last category is silent-gesture alignment, in which a person commences and completes a gesture without accompanying vocalization. There are 128 such gestures, and miscellaneous movement (N/A) is the most frequent type (n=109). Table 12 shows all gesture type counts, and Table 13 indicates durations.

## 6 Results and Discussion

One might hypothesize that the extended gesture duration indicates cognitive processing. It is reasonable to theorize that Beats are used in a man-

Counts: gest-silent					
SUM	Beat	Iconic	Deictic	Symbolic	N/A
128	11	6	1	1	109

Table 12: The distribution of gestures accompany silence, by gesture type.

Durations: gest-silent					
	Beat	Iconic	Deictic	Symbolic	N/A
mean	803.6	1205.0	890.0	1200.0	1622.3
median	669.0	1020.0	890.0	1200.0	1450.0
sd	409.5	907.1	NA	NA	841.5

Table 13: Duration (milliseconds) statistics of gestures accompanying silence, by gesture type.

ner that punctuates completely planned speech while Symbolics are used in support of forming the thought that is being spoken. One might therefore expect words spoken during Beats to take less time than those spoken through Symbolics. This overall contrast is not significant (Wilcoxon’s  $W = 2516, p = 0.08387$ ), but it is significant when restricting attention to verbs<sup>5</sup> (Wilcoxon’s  $W = 1027, p = 0.04075$ ).

Of the six gesture-word alignments, three (long-gest, gest-word-no-overlap, gest-word-with-overlap) involve a gesture commencing before the onset of an aligned word and two of these (the exception is gest-word-no-overlap) are among the three most frequent alignment types. The third frequent alignment type, we note below, does not have significant interactions with the count of aligned POS categories, but the other two frequent alignments do. The combination of gestures commencing before the aligned word and the interaction with the distribution of POS categories of those words are suggestive of a role of the gesture in the formulation of the unfolding speech. Beat gestures are used more than other gestures in players’ conversations. Iconic, N/A, Deictic, and Symbolic are the next most used gestures, see Table 1. As Table 5 shows, short-gest, gest-word-no-overlap, and word-gest-no-overlap alignments are least frequently witnessed. The most frequently witnessed alignment is long-gest, followed by gest-word-with-overlap, and word-gest-with-overlap alignments.

Consider those gestures that have a duration that exceeds that of its first aligning word, inclusive of more words as well (long-gest). Figure 1 shows the Pearson residuals that result from the  $\chi^2$  analysis of the contingency table inherent in Table 9

<sup>5</sup>That is, we measure the contrast between Beat durations and Symbolic durations when accompanying verbs.

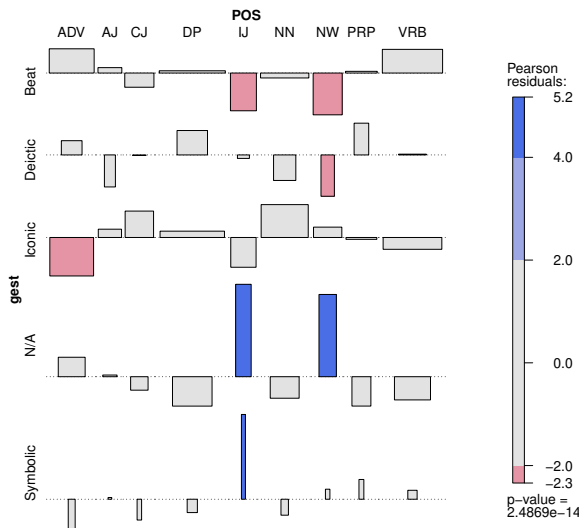


Figure 1: Residuals of interaction among gestures longer than nearby words (long-gest) and the syntactic categories of those words. The horizontal lines indicate the level of no interaction for a given row: upwards projections indicate counts in a cell that exceed what would be expected with no interaction; downwards projections indicate counts that are less than would be expected; shading indicates statistical significance ( $p < 0.05$ ).

( $\chi^2 = 134.34, df = 32, p = 2.487^{-14}$ ). In contrast to a null-hypothesis expectation of no interaction between gesture type and syntactic categories: Beats show a significant dearth with interjections and non-word vocalizations; Deictics show significant dearth with non-words; Iconics show significant dearth with adverbs; miscellaneous motions show significant co-occurrence with interjections and non-word vocalizations; Symbolics show significant co-occurrence with interjections (but we treat the effect of Symbolics with caution, given the low count of observations).

It is not surprising that Beats do not appear to be multi-modal exclamation marks for interjections or that Beats and Deictics are conspicuously missing from non-word vocalizations. It also makes sense for iconic gestures to neglect adverbs. It seems natural that miscellaneous motions accompany interjections and non-words.

Figure 2 shows the residuals of the  $\chi^2$  test of interaction between gesture types and parts of speech for the alignments in which a gesture starts before a word and ends in the middle of the word (gest-word-with-overlap;  $\chi^2 = 72.526, df = 32, p = 5.572^{-5}$ ). In comparison with the distribution of counts that would be expected if there were no interaction be-



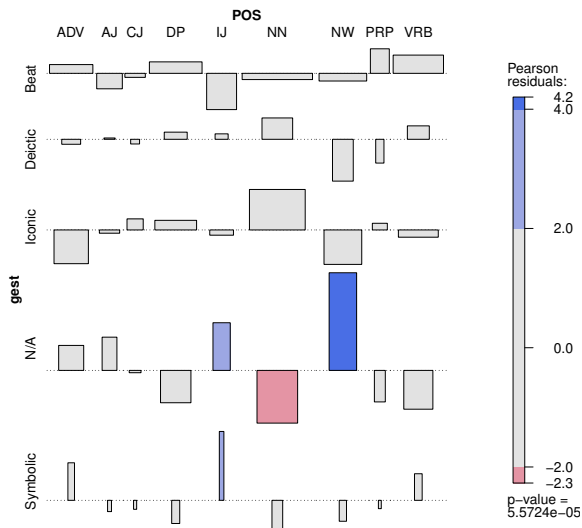


Figure 2: Residuals of interaction among gestures longer than nearby words (gest-word-with-overlap) and the syntactic categories of those words. The horizontal lines indicate the level of no interaction for a given row: upwards projections indicate counts in a cell that exceed what would be expected with no interaction; downwards projection indicate counts that are less than would be expected; shading indicates statistical significance ( $p < 0.05$ ).

tween gesture type and part of speech, there are more non-words and interjections and fewer nominals aligned with miscellaneous movements; more Symbolics aligned with interjections.

Table 4 illustrates determiners followed by verbs and nouns are used the most around the gestures which the highest ones occur in long-gest, gest-word-with-overlap and word-gest-with-overlap alignments. The interactions between gesture category and linguistic categorization of vocalizations are not significant for alignments in which the gesture starts in the middle of a word and ends afterwards (word-gest-with-overlap) –  $\chi^2 = 31.955$ ,  $df = 32$ ,  $p = 0.47$ .

Of the alignments for which there were sufficient interactions to meaningfully analyze the interaction between semiotic types and part of accompanying parts of speech, two demonstrated statistically significant interactions, and in both of those, the gesture onset preceded the linguistic content onset. The primary effects for the contentful semiotic types (i.e. not miscellaneous movements) was in a relative lack of gestures accompanying certain syntactic categories, but without systematic sensitivity to the whether the category is mainly popu-

lated by open-class or closed-class subcategories. Certain interesting trends are visible (e.g. beats occurring with relational categories; iconics with the nominal domain; deictics with nouns and verbs) but not statistically significant. While this work uses a more general typology of gesture types than (Mehler et al., 2012), we see more detail about where the relationships between gesture types and part of speech categories carry strong associations.

## 7 Conclusions

We have presented our observations of the counts and durations of gestures aligned with major syntactic categories assigned to vocalizations that accompany them, given a small number of possible alignment types. We think that the type of alignment (e.g., gesture onset prior to accompanying word onset) is revealing aspects of cognitive processing associated with the unfolding utterance. Of course, observations of different sorts than we have reported here would also be useful, but the alignments provided here will enable hypothesis testing regarding the interactions of gestures, syntactic categories, and their alignments. Of the six considered gesture-word alignment types, three are more frequently witnessed than the others, and within one of those, where gestures have a long duration from an onset before the first aligned word, there is noteworthy dearth of interjections and non-words with Beats, non-words with Deictics, adverbs with iconics; there is noteworthy coincidence of miscellaneous movement and interjections and non-words and interjections and Symbolics.

While the observations reported here are anchored in a single multi-modal dialogue corpus, the corpus involves distinct dialogues among a number of interlocutors, the dialogue settings do not impose particular constraints on gestures or part of speech categories. We intend to continue to explore gesture and word alignments in this and other multi-modal dialogue corpora.

## Acknowledgments

This work was conducted with the financial support of Science Foundation Ireland under Grant No. 18/CRT/6223 and the GEHM research network (Independent Research Fund Denmark grant 9055-00004B). We are grateful to the anonymous reviewers who provided helpful constructive feedback on an earlier draft of this work.

## References

- Janet Bavelas, Jennifer Gerwing, Chantelle Sutton, and Danielle Prevost. 2008. [Gesturing on the telephone: Independent effects of dialogue and visibility](#). *Journal of Memory and Language*, 58(2):495–520.
- Paolo Bernardis and Maurizio Gentilucci. 2006. [Speech and gesture share the same communication system](#). *Neuropsychologia*, 44(2):178–190.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O’Reilly Media, Inc.
- Paul Boersma and Vincent van Heuven. 2001. [Speak and unSpeak with PRAAT](#). *Glott International*, 5(9/10):341–347.
- Hennie Brugman and Albert Russel. 2004. [Annotating multi-media/multi-modal resources with ELAN](#). In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04)*, Lisbon, Portugal. European Language Resources Association (ELRA).
- Alan Cienki and Cornelia Muller. 2008. [Metaphor, gesture, and thought](#). In Raymond W. Editor Gibbs, Jr., editor, *The Cambridge Handbook of Metaphor and Thought*, Cambridge Handbooks in Psychology, pages 483–501. Cambridge University Press.
- Doriana De Marco, Elisa De Stefani, and Maurizio Gentilucci. 2015. [Gesture and word analysis: the same or different processes?](#) *NeuroImage*, 117:375–385.
- Susan Goldin-Meadow, Martha Wagner Alibali, and R. Breckinridge Church. 1993. [Transitions in concept acquisition: Using the hand to read the mind](#). *Psychological Review*, 100(2):279–297.
- Uri Hadar. 1989. [Two types of gesture and their role in speech production](#). *Journal of Language and Social Psychology*, 8(3-4):221–228.
- Florian Hahn and Hannes Rieser. 2010. [Explaining speech gesture alignment in MM dialogue using gesture typology](#). In *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, pages 99–111, Poznań, Poland. SEMDIAL.
- Stefanie Jannedy and Norma Mendoza-Denton. 2005. [Structuring information through gesture and intonation](#). *Interdisciplinary studies on information structure: ISIS; working papers of the SFB 632*, (3):199–244.
- Adam Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press.
- Sotaro Kita. 2000. [How representational gestures help speaking](#). In David McNeill, editor, *Language and Gesture*, pages 162–85. Cambridge University Press.
- Sotaro Kita and Asli Özyürek. 2003. [What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking](#). *Journal of Memory and Language*, 48(1):16–32.
- Maria Koutsombogera and Carl Vogel. 2018. [Modeling collaborative multimodal behavior in group dialogues: The multisimo corpus](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2945–2951, Paris, France. European Language Resources Association (ELRA).
- Daniel P Loehr. 2012. [Temporal, structural, and pragmatic synchrony between intonation and gesture](#). *Laboratory phonology*, 3(1):71–89.
- D.P. Loehr. 2004. *Gesture and Intonation*. Georgetown University. PhD Thesis.
- David McNeill. 1992. *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, Chicago.
- David McNeill. 1997. [Growth points cross-linguistically](#). In Jan Nuyts and Eric Pederson, editors, *Language and Conceptualization*, 1 edition, pages 190–212. Cambridge University Press.
- David McNeill. 2005. *Gesture and thought*. University of Chicago Press, Chicago.
- David McNeill and Susan D. Duncan. 2000. [Growth points in thinking-for-speaking](#). In David Editor McNeill, editor, *Language and Gesture*, Language Culture and Cognition, pages 141–161. Cambridge University Press.
- Alexander Mehler, Andy Lücking, and Peter Menke. 2012. [Assessing cognitive alignment in different types of dialog by means of a network model](#). *Neural networks*, 32:159–164.
- Charles S. Peirce. 1931. *The Collected Papers of Charles Sanders Peirce, Vol. I: The Principles of Philosophy*. Harvard University Press, Cambridge.
- Wim Pouw, Steven J Harrison, and James A Dixon. 2020. [Gesture–speech physics: The biomechanical basis for the emergence of gesture–speech synchrony](#). *Journal of Experimental Psychology: General*, 149(2):391.
- Hannes Rieser. 2015. [When hands talk to mouth, gesture and speech as autonomous communicating processes](#). In *SEMDIAL 2015 goDIAL: Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue*, pages 122–130.
- Hannes Rieser and Insa Lawler. 2020. [Multi-modal meaning – an empirically-founded process algebra approach](#). *Semantics & Pragmatics*, 13:1–55.

- Jan Peter de Ruiter. 2000. [The production of gesture and speech](#). In David McNeill, editor, *Language and Gesture*, Language Culture and Cognition, pages 284–311. Cambridge University Press.
- Emanuel A. Schegloff. 1985. [On some gestures' relation to talk](#). In J. Maxwell Editor Atkinson, editor, *Structures of Social Action*, Studies in Emotion and Social Interaction, pages 266–296. Cambridge University Press.
- Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK.
- Mark Steedman. 1991. [Structure and intonation](#). *Language*, 67(2):260–296.
- Kevin Tuite. 1993. [The production of gesture](#). *Semiotica*, 93(1/2):83–106.
- Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. [Gesture and speech in interaction: An overview](#). *Speech Communication*, 57:209–232.