LDK 2023

Lexical Retrieval Hypothesis in Multimodal Context

Po-Ya Angela Wang, Pin-Er Chen, Hsin-Yu Chou, Yu-Hsiang Tseng, Shu-Kai Hsieh
Graduate Institute of Linguistics, National Taiwan University
differe94nt@gmail.com, cckk2913@gmail.com,
r10142008@ntu.edu.tw,
seantyh@gmail.com, shukaihsieh@ntu.edu.tw

Abstract

Multimodal corpora have become an essential language resource for language science and grounded natural language processing (NLP) systems due to the growing need to understand and interpret human communication across various channels. This paper presents our efforts in building the first Multimodal Corpus for Languages in Taiwan (MultiMoco). Based on the corpus, we conduct a case study investigating the Lexical Retrieval Hypothesis (LRH), specifically examining whether the hand gestures co-occurring with *speech constants* facilitate lexical retrieval or serve other discourse functions. With detailed annotations on eight parliamentary interpellations in Taiwan Mandarin, we explore the co-occurrence between *speech constants* and non-verbal features (i.e., *head movement*, *facial movement*, *hand gesture*, and *function of hand gesture*). Our findings suggest that while hand gestures do serve as facilitators for lexical retrieval in some cases, they also serve the purpose of information emphasis. This study highlights the potential of the MultiMoco Corpus to provide an important resource for in-depth analysis and further research in multimodal communication studies.

1 Introduction

Over the past decades, there has been a growing interest in multimodal corpus linguistic research (Paquot and Gries, 2021), which focuses on the analysis and comprehension of information from diverse modalities, including speech, image, and gesture. To facilitate research in this field and other interdisciplinary studies, the creation of multimodal corpora, or collections of data from various modalities, has become more crucial.

We thereby introduce the Multimodal Corpus for Languages in Taiwan (the MultiMoco Corpus), a newly released multimodal corpus that includes audio, video, gestural, and textual data involving various languages and discourse contexts. The MultiMoco Corpus is comprised of recordings of realistic interactions taken in news and interpellation in parliament, where interviews and spontaneous speech take place. The synchronization of the audio, video clips, and gesture segments enables researchers to study the link between the communication modes. These data assist researchers in annotating information on the speakers, their actions, and the communication contexts. This corpus is designed for human communication and interaction-related research, such as conversation analysis, multimodal machine learning, and natural language processing.

To demonstrate the feasibility of the MultiMoco Corpus, we conduct a case study based on the parlimental interpellation clips in Taiwan Mandarin, aiming to validate the widely discussed Lexical Retrieval Hypothesis (hereafter, LRH) (Dittmann and Llewellyn, 1969; Ekman and Friesen, 1972; Butterworth and Beattie, 1978; Rauscher et al., 1996), which suggests that gesture and verbal disfluency tend to co-occur in spontaneous speech.

More specifically, we take *speech constants*, based on the framework of Voghera (2001), as indicators of potential verbal disfluency. We annotate one verbal feature (*speech constants*) as well as four non-verbal features, including three forms of non-verbal expressions (*head movement*, *face movement*, *hand gesture*) and *functions of hand gesture*. With careful annotation, we attempt to answer research questions as follows: (1) Could we observe co-occurrences between *speech constants* and gestures in the context of interpellation? (2) If there are co-occurrences with *speech constants*, do hand gestures mainly play the role of priming lexical items? And (3) Do the hand gestures serve other functions regarding interlocutors and the entire discourse context?

To provide guidance on utilizing the MultiMoco Corpus to address multimodal research problems, we first review studies on the multimodal corpus,

the multimodal annotation framework, and the LRH (Section 2). Following this, we outline the data collection and annotation framework for the case study in Section 3.2 and Section 3.3. Next, we analyze if the non-verbal features co-occur with *speech constants* (Section 4.1). The LRH mechanism is examined by identifying the co-occurrences between *speech constants* and LRH-related/ non-LRH-related functions of hand gesture (Section 4.2), along with the individual performances discussed in Section 4.3. Section 5 concludes the paper.

2 Related Works

2.1 Multimodal corpus

Communication, by nature, is multimodal (Carter and Adolphs, 2008), and thereby constructing multimodal corpora affords researchers the opportunity to get a comprehensive understanding of the cognitive mechanisms underlying communication. "Multimodal corpus" can be defined at varying degrees depending on its architecture (Allwood, 2008). Generally speaking, it refers to an online repository of language and communication-related content that contains several modalities. In a narrower sense, it can be specified with audiovisual materials accompanied by annotations and transcriptions.

Most earlier multimodal corpora are for specific purposes. For example, the Mission Survival Corpus (McCowan et al., 2003), the Multimodal Meeting (MM4) Corpus (McCowan et al., 2005), and the VACE corpus (Chen et al., 2006) are all built on conversations in meeting. Others are task-oriented corpora elicited in lab settings, such as the Fruit Carts corpus (Gallo et al., 2006), CUlture-adaptive BEhavior Generation for interactions with embodied conversational agents (CUBE-G) (Rehm et al., 2009), and the spatial task-based dialogue corpus, SaGA (Lücking et al., 2010). Still, others include dyadic conversation in academic discourse: the Nottingham Multi-Modal Corpus (NMMC) (Knight et al., 2008) and the Pisa Audiovisual Corpus project (Camiciottoli and Bonsignori, 2015)), providing domain-specific multimedia materials for English for Specific Purposes (ESP) learners in higher education.

Recent corpora attempt to be less specific and purpose-oriented. Mlakar et al. (2017) select 4 recordings of multiparty conversation in a talk show, with more spontaneous discourses and more topics. The NTHU-NTUA Chinese interactive multimodal emotion corpus (NNIME) (Chou et al., 2017) constructed a dataset with 44 subjects majoring in drama to record performed scenes for affective behaviors. In addition, the Communicative Alignment of Brain and Behaviour (CABB) (Eijk et al., 2022) builts a dataset on recordings of 71 pairs of participants discussing innovative, unconventional objects[1] (Barry et al., 2014), which provides pre-and-post behavioral and fMRI measurement information. Nevertheless, these corpora have their limitations. Certain datasets are built on less amount of data, some are restricted to conversations revolving around narrow topics, and others are collected for particular experiments.

The MultiMoco Corpus presented in this study incorporates video and audio recordings from ten public news channels and interpellation videos, which encompass a broader spectrum of languages and communication genres.[2] This renders it a more balanced resource for investigating multilingual and multimodal communication in everyday conversations, with the capacity to accommodate multidimensional annotations.

2.2 Multimodal annotation framework

Various annotation frameworks have been proposed to encode labels for gesture forms and corresponding functions (Bavelas et al., 1992; McClave, 2000; Kendon, 2004; Müller, 2004; Allwood et al., 2005; Bressem et al., 2013). According to Debras (2021)'s proposal, "articulator" (e.g., hand or head), and "configuration of articulator" (e.g., head nod, wave, or turn) should be formally annotated. Functional annotation is to indicate co-verbal intentions of gestures. The Facial Action Coding System (FACS; Ekman and Rosenberg, 1997; Clark et al., 2020), for facial expression annotations, and the Linguistic Annotation System for Gestures (LASG; Bressem et al., 2013), for hand annotations, are both well-designed but complicated annotation systems. Annotation frameworks such as these can be time-consuming and challenging to achieve annotation agreement. Debras (2021) suggests that coarse-grained annotations can benefit the onset of the research.

We here review the annotation frameworks that will be adopted in the case study. Firstly, *speech constants* will be annotated to examine the LRH

[1] "Fribbles"

[2] The collection and characteristics of MultiMoco Corpus data are described in Section 3.1.

evaluated by Trotta and Guarasci (2021), given that gestures tend to co-occur with verbal disfluency. Referring to the guidelines in Voghera (2001), four types of *speech constants* (i.e., *pause*, *repetition*, *truncation*, and *semi-lexical*) are taken as the annotation targets. Secondly, the non-verbal target features comprise forms and functions, namely *head movement*, *face movement* (*eyebrows and mouth*), *hand gesture*, and *functions of hand gesture*. Considering Debras (2021)'s suggestions for coarse-grained annotations, this study follows the concise annotation framework adopted by Camiciottoli and Bonsignori (2015), incorporating gesture form abbreviations by Julián (2011) and the gesture functions by Kendon (2004) and Weinberg et al. (2013). In Camiciottoli and Bonsignori (2015)'s framework, *head movement* include *head-nodding/tilting/jerking/moving* together with multiple directions and repetition; *face movement* involve the movement of eyebrows and mouth; *hand gesture* mark the movements of fingers, palm, and the whole hand. The comprehensive labels and definitions for each feature will be explained in Section 3.3.

2.3 Lexical Retrieval Hypothesis

As reviewed in Özer and Göksun (2020), multimodal interaction in speech production and comprehension regarding individuals' cognitive tendencies has been heatedly discussed. When a speaker cannot clarify intended thoughts, gestures are incorporated during hesitation pauses or the lexical pre-planning stage (Dittmann and Llewellyn, 1969; Butterworth and Beattie, 1978). The link between verbal, non-verbal, and conceptual aspects can be addressed by the "growth point," the smallest thought unit, comprising both utterances and gestures (McNeill, 1992). Krauss (1998) has considered the relationship between thoughts, utterances, and gestures from another perspective, specifying three parts in speech production: conceptualizing, grammatical encoding, and phonological encoding. Among these three parts, phonological encoding, the retrieval of lexical form, is the part where gestures affect the verbal modality, and limited gestures reduce speech fluency when a speaker discusses spatial information (Krauss, 1998). Later, Krauss and Hadar (1999) have further proposed that concepts in the mind are stored in various forms, so activating one idea in one modality may also activate concepts in other modalities. Thus, concepts can be fully comprehended when information from different modalities is all presented, and representations from one modality can be converted into another modality. Following the line of this discussion, the gestural modality can assist lexical retrieval in the verbal modality because of such cross-modal priming. This is termed the "Lexical Retrieval Hypothesis" (Gillespie et al., 2014; Trotta and Guarasci, 2021). Namely, LRH refers to the process that the triggered idea's lexical gestures[3] (i.e., gestures that can iconically represent meanings) can semantically prime the phonological encoding of the related words, reviewed in Gillespie et al. (2014). Gillespie et al. (2014) also specify that LRH is less applicable if the speaker can resort to alternative tactics to avoid lexical access challenges, which occur in improvisational speech production.

The Lexical Retrieval Hypothesis is tested in several tasks and contexts. Hostetter and Alibali (2007) distinguish the phonemic fluency from the semantic fluency[4], suggesting lexical access efficiency may be related to different types of gestures. Additionally, Smithson and Nicoladis (2013) have proposed that the negative association between verbal working memory and iconic gesture production in bilinguals designates gesture production's assistance in the retention and utilization of language information. Trotta and Guarasci (2021) calculate the weighted mutual information (WMI) between the hand movements and the concurrent speech disfluency features involving five kinds of *speech constants*[5]. The result concurs with the LRH since hand gestures are more related to semi-lexical features and pauses in interview contexts. It is noted that in Trotta and Guarasci (2021), *speech constants* are considered disfluency features to assess the LRH, whereas hesitation pauses may signal lexical retrieval difficulties.

As most of the studies mentioned have examined the LRH with laboratory tasks or free-form inter-

[3] Krauss (1998) refers to these lexical retrieval supporting gestures as "lexical gestures."

[4] As defined in Hostetter and Alibali (2007), movements that transmit information relevant to the content of the vocal communication are representational gestures. Beat gestures are short, rhythmic motions that accentuate terms without demonstrating what they mean."Phonemic fluency" indicates thought-organizing skills associated with representational gesture rates, whereas "semantic fluency" is less correlated with representational gesture rates but has a significant correlation with beat gestures.

[5] Five kinds of *speech constants*: pause, repetition, truncation, and semi-lexical), as specified by Voghera (2001)

views, we aim to assess the LRH in formal speaking contexts (i.e., political interpellation) as well as its applicability in less colloquial speech. Meanwhile, given that investigations in multiple modalities can provide us with more comprehensive perspectives on cross-modal interaction, we also aim to extend the hypothesis testing scope by exploring how disfluency co-occurs with more gestures: face, head, and hand. Among them, different*functions of hand gesture* co-occurring with *speech constants* are investigated to ascertain whether or not gestures assist in lexical retrieval. This case study conjectures that gestures co-occurring with *speech constants* are not just for facilitating lexical retrieval.

3 Methodology

Our study of the lexical retrieval hypothesis is based on the multimodal data made available from Multimoco. We first introduce the construction and contents of the MultiMoco Corpus (Section 3.1). Then, the data collection for our case study on the LRH is illustrated (Section 3.2), followed by the annotation framework for the target features (Section 3.3). The annotation results and analyses will be discussed in the subsequent sections.

3.1 MultiMoco Corpus

The MultiMoco Corpus is built on recorded videos and audios from 10 public television channels[6] in Taiwan, including news in multiple languages (i.e., Taiwan Mandarin, Taiwan Southern Min, Hokkien, Hakka, and Formosan languages) and the interpellation of the Taiwan Legislative Yuan (the parliament of Taiwan). While the TV news is recorded by wireless television receivers, the interpellation video clips with transcriptions in Taiwan Mandarin are retrieved directly from the Internet Multimedia Video-on-Demand System for Rebroadcasting Legislative Yuan Proceedings[7].

Figure 1 displays the data processing workflow of the MultiMoco Corpus. With 223 video clips from Taiwan public television channels and the interpellation from Taiwan Legislative Yuan, the MultiMoco Corpus provides 5,854 minutes of dialogue, accompanied by 1,485,297 characters of captions transcribed via Whisper (Radford et al., 2022) model. In addition, 22,805 gestures identified via MediaPipe (Lugaresi et al., 2019) are also included in the corpus. The multimodal nature of the corpus allows researchers to conduct cross-modality analyses, thereby broadening the understanding of the communicative potential of various modalities beyond spoken texts. That is, the MultiMoco Corpus provides us with the potential to extend communication studies to diverse linguistic and multimodal contexts.

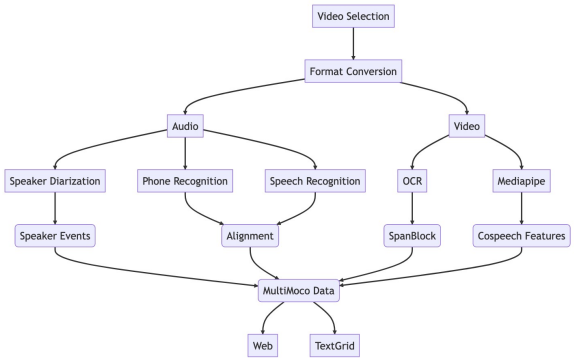

Figure 1: Establishment workflow of the MultiMoco Corpus

3.2 Data collection

Our lexical retrieval analysis data are extracted from MultiMoco Corpus, specifically focusing on spontaneous speech during interpellation involving interactions between legislators and officers. To control the gender, speech delivery performance, and speech topics of the selected data, we chose two biological females and two biological males, along with a balanced selection of speech topics. The interpellation topics are detailed in Table 1. As to speech delivery performance, we have selected interpellation clips based on the evaluation scores of 103 legislators from Citizen Congress Watch (CCW) in the 10th session of Congress [8]. we have selected interpellation clips based on the evaluation scores of 103 legislators from the Citizen Congress Watch (CCW) in the 10th session of Congress. Figure 2 shows the distribution of individually-averaged evaluation scores, with an average score of approximately 16, a minimum of 11.25, and a maximum of 17.998. After considering the evaluation score, interpellation topics, and

[6]The target channels are as follows: CTV News PTS News, PTS Taigi, Hakka TV, Taiwan Indigenous TV, TTV News, CTS News, Congress Channel I, Congress Channel II, and FTV News.

[7]https://ivod.ly.gov.tw/Demand

[8]Using the Legislative Yuan's Internet multimedia Video on Demand System, civil jurors can evaluate the performance of parliamentarians in sessions and fill out questionnaires. Then, the evaluation score of each legislator is calculated through this procedure.

political parties, we choose four legislators (two with higher evaluation scores and two with lower evaluation scores) for subsequent multimodal analyses. In the end, we collect eight interpellation clips, each lasting between 8 and 12 minutes and featuring a male and a female legislator in each pair.

Legislator	Topic of Interpellation Clips	
high_A	Social welfare	Education and culture
high_B	Finance	Communications
low_C	Finance	Judiciary and organic laws
low_D	Social welfare	Education and culture

Table 1: Topics of the interpellation clips. The prefixes (*high* or *low*) in the Legislator column are used for identifying the evaluation scores for the legislators (i.e., A, B, C, and D).

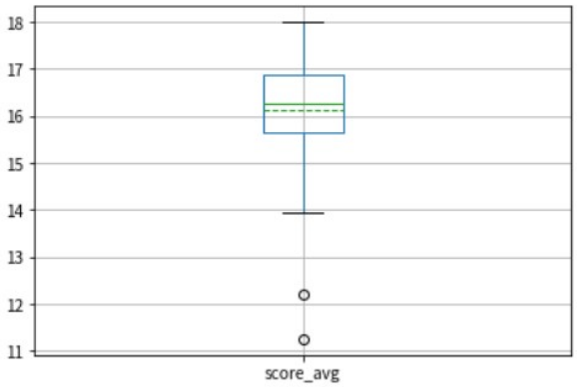

Figure 2: Descriptive statistics of citizen evaluation score

3.3 Data annotation

We investigate the functions of non-verbal features and their co-occurrence with disfluency in spontaneous speech. Three non-verbal forms (i.e., *head movement*, *face movement*, and *hand gesture*), one non-verbal function (i.e., *functions of hand gesture*), and one verbal feature (i.e., *speech constants*) are selected as our annotation targets; the latter is used to identify disfluency in speech.

Considering the specificity of each feature and the consensus in prior studies, we adopt different annotation frameworks for corresponding features. The *speech constants* are annotated based on the framework in Voghera (2001), as shown in Table 2;

Label	Definition
`Pause`	This marks a pause either between or within utterances.
`Non-lexical item`	This marks interjections (e.g., eh and ehm), or more general words that convey the meaning of an entire sentence, constituting a complete linguistic act demonstrated by their paraphrasability.
`Repetition`	This marks cases of repetition of utterances in order to give coherence and cohesion to the speech or self-repetition as a control mechanism of the speech programming.
`Truncation`	This indicates the deletion of a phoneme or a syllable in the final part of a word.

Table 2: Labels for speech constants. It is noted that the original label "semi-lexical" in Trotta and Guarasci (2021) is renamed "non-lexical item" in our study.

functions of hand gesture were annotated via Camiciottoli and Bonsignori's framework, as presented in Table 3. The three non-verbal forms (i.e., *head movements*, *face movements*, and *hand gestures*) are classified based on Camiciottoli and Bonsignori (2015)'s framework, as illustrated in Table 4. It is noted that the labels in the table are generalized to a more coarse-grained scale regarding the entailment of the original labels.

Five native speakers annotate the five verbal and non-verbal features (i.e., *head movement*, *face movement*, *hand gesture*, *function of hand gesture*[9], and *speech constants*) via ELAN (Sloetjes and Wittenburg, 2008)[10], an open-source software appropriate for multimodal annotations and linguistic analysis. Take *speech constants* for instance, the two annotators separately mark the time periods and corresponding labels of *speech constants* that occur in all eight clips. Then, the annotated pair of tiers (made by the two annotators) for each clip are segmented into units of 100 milliseconds and aligned with each other.

For annotation consistency, the annotators are asked to annotate different features from clip segments and decide on an agreed-upon criterion for disagreed annotations. For instance, the function, `Parsing`, marks situations in which a speaker intends to initiate a new discourse turn, recur the same gesture as if beating, or make some trivial

[9]For clarity, we use the *italic* form when referring to the five targets, and we use the `typewriter` font when referring to the labels under each target.

movements that have no clear reference. In terms of our Inner Annotator Agreement (IAA), we calculate the ratio of intersecting annotation segments and the agreement ratio of the intersecting segments to measure the agreement between the annotators. As shown in Table 5, *hand gesture* (.76) and *function of hand gesture* (.81) acquire a higher ratio of intersecting segments, in which the annotators are able to identify more overlapping time periods of hand movements. Conversely, the ratio of intersecting segments for the *head movement* (.26) and *face movement* (.37) is relatively low. We suggest that the lower number of intersecting segments may relate to the different scales of movements perceptualized by the annotators. Although we generalized certain categories of the labels, we found it hard to define the degree of the speakers' movements. While one annotator perceived and marked some subtle tilting periods, the other annotator may have missed the same units. The subjectivity in continuum segmentation poses a challenge for multimodal annotation, yet since the annotators have discussed their inconsistencies and reached a consensus, the annotation results of the subsequent discussion are reliable.

As we focus on the co-occurrence and association between non-verbal features and disfluency, we will not inspect the details of the annotation results within each non-verbal feature but rather discuss the general co-occurrence with *speech constants* in the following sections.

Label	Definition
`Social`	social (emphasizing a message)
`Repres`	representational (representing object/idea)
`Index`	indexical (indicating a referent)
`Parsing`	parsing (distinguishing units of speech)
`Perform`	performative (illustrating speech act)
`Modal`	modal (expressing certainty/uncertainty)

Table 3: Labels for functions for hand gesture.[11] The functions of 'beat' and 'representational' in Hostetter and Alibali (2007) are represented as `Parsing` and `Representational` in this study.

4 Results & Discussions

We first examine the non-verbal features' co-occurrence with *speech constants*, which indicate verbal disfluency (Section 4.1). Then, the potential discourse *functions of hand gesture* will be analyzed (Section 4.2). Finally, we will discuss more comprehensive gesture functions independent of verbal disfluency but related to interlocutors and the entire discourse context in Section 4.3.

Type	Label
Face	Frowning eyebrows
	Raising eyebrows
	Smile
	Other
Head	Nod
	Jerk
	Move Forward/Backward
	Tilt
	Side-turn
	Shake (repeated)
	Other
Hand	Finger pointing towards audience
	Hands sweeping sideways
	Hands rotating at center of body
	Hands wide apart moving down
	Hands clasped together in front of body
	Other

Table 4: Labels for co-speech gestures: face, head, and hand.

Target	Ratio	Agreement Rate
`Head`	.26	.78
`Face`	.37	.99
`Hand gesture`	.76	.70
`Function of hand gesture`	.81	.41
`Speech constant`	.49	.89

Table 5: Inter-annotator agreement on five targets. "Ratio" refers to "Ratio of Intersecting Segments." Intersecting segments are those existing on both annotation tiers (of the two annotators) after aligned to the timeline of each clip. "Agreement Rate" refers to the "Agreement Rate on Labels of the Intersecting Segments."

4.1 Co-occurrence overview

As we target one verbal feature (*speech constants*) and three forms of non-verbal features (head, hand, and face)[12], we calculate the co-occurrences[13] of the six patterns by modality. Figure 3 shows that *head movement* and *speech constants* co-occur most frequently, followed by *hand* and

[10] ELAN (https://archive.mpi.nl/tla/elan); Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands.

[11] The functions of hand gestures are mutually exclusive.

[12] It should be noted that one non-verbal related feature, i.e., the *functions of hand gesture*, are annotated based on the occurrence of hand gesture; thus, calculating the co-occurrences (i.e., overlapping segments) between *functions of hand gesture* and the other features would be meaningless, as it would be the same as hand gesture.

[13] The co-occurrence of one pair of features is defined as the summed number of overlapping segments; one segment is a unit of 100 milliseconds.

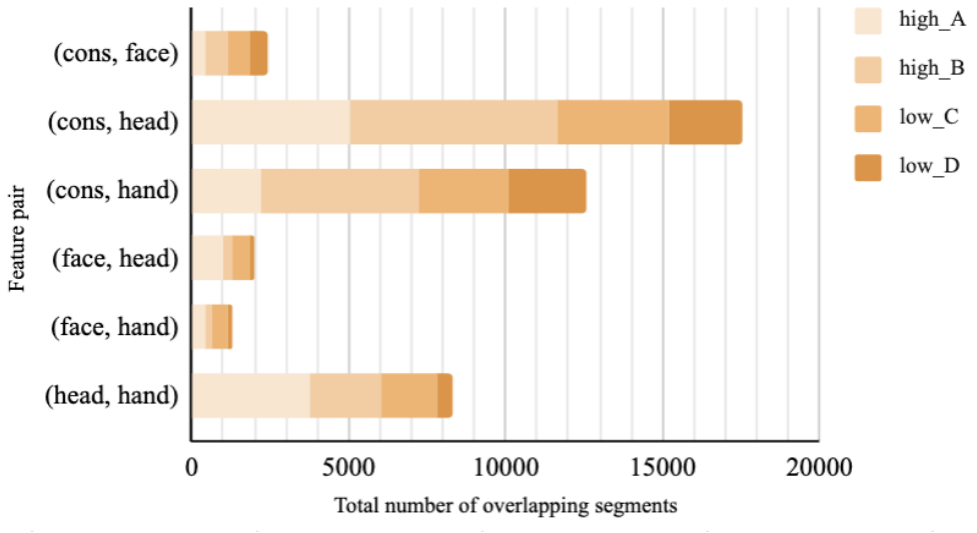

Figure 3: Co-occurrences of different feature pairs. The y-axis represent the number of overlapping segments between different pairs of annotated features.

speech constants. *Face movement* shows fewer co-occurrences with the other features (i.e., face & head, face & hand, and face & *speech constants*), which may relate to the few occurrences of face movement in all clips. In addition to mask-wearing situations, these few occurrences of facial movement are the result of the face movements being so frequent and inconsequential that the annotators reach an accord to only record the apparent ones, as some trivial ones may be the result of habitual movements. This annotation procedure illuminates considerations for future annotation frameworks. While the non-verbal features tend to co-occur with one another, the frequencies are far lower than their respective co-occurrence with *speech constants*. This may correspond to the LRH that when *speech constants* appear, i.e., during hesitation pauses or the lexical pre-planning stage, non-verbal gestures are possibly employed by the speaker as well (Dittmann and Llewellyn, 1969; Butterworth and Beattie, 1978). To sum up, the distribution illustrates that non-verbal characteristics are more likely to co-occur with disfluent situations than with other types of non-verbal movements. Furthermore, it demonstrates the significance of both the head and the hand in the research of verbal disfluency.

4.2 Co-occurring functions of hand gestures

As significant as the respective gesture co-occurrence with *speech constants* is, could we claim that the identified *speech constants* require gestures to facilitate lexical retrieval? To further understand the purposes of the hand gestures co-occurring with *speech constants*, Table 6 below presents the overall frequencies of each type of *speech constants* co-occurring with different *functions of hand gesture*. *Speech constants*, especially `non-lexical items` and `pauses`, are taken as verbal disfluency traits in the LRH evaluation (Trotta and Guarasci, 2021). We would like to argue that the intentions of performing *speech constants* are various, so the functions resulting from the interplay between verbal and non-verbal modalities are complicated. Thus, in addition to using *speech constants* as markers of the possible presence of verbal disfluency, we study the functions of co-occurring hand gestures in order to realize whether the co-occurring hand gestures are lexical retrieval facilitators or carry out other functions in speech contexts.

First, we examine the distributions of *speech constants* and their co-occurring *functions of hand gesture*. Regarding *speech constants*, `pause` is the most frequently observed category with 345 frequencies, accounting for 72.2% co-occurrences among all. `Repetition` and `non-lexical item` both rank second. `Truncation` sporadically occurs in the collected dataset. As for *functions of hand gesture*, `Social` (i.e., to emphasize a message) is the most frequent function for the *speech constants* as a whole. The rest of the ranking goes as follows: `Parsing` > `Indexical` > `Representational` > `Performative` >

(SC / FH)	Indexical	Parsing	Performative	Representational	Social	Total
Non-lexical item	10	24	2	16	9	61
Pause	59	87	20	46	133	**345**
Repetition	6	16	0	7	32	61
Truncation	8	0	0	0	3	11
Total	83	127	22	69	177	478

Table 6: Contingency table of *speech constants* and *functions of hand gesture*. SC represents *speech constants*, and FH represents *functions of hand gesture*.

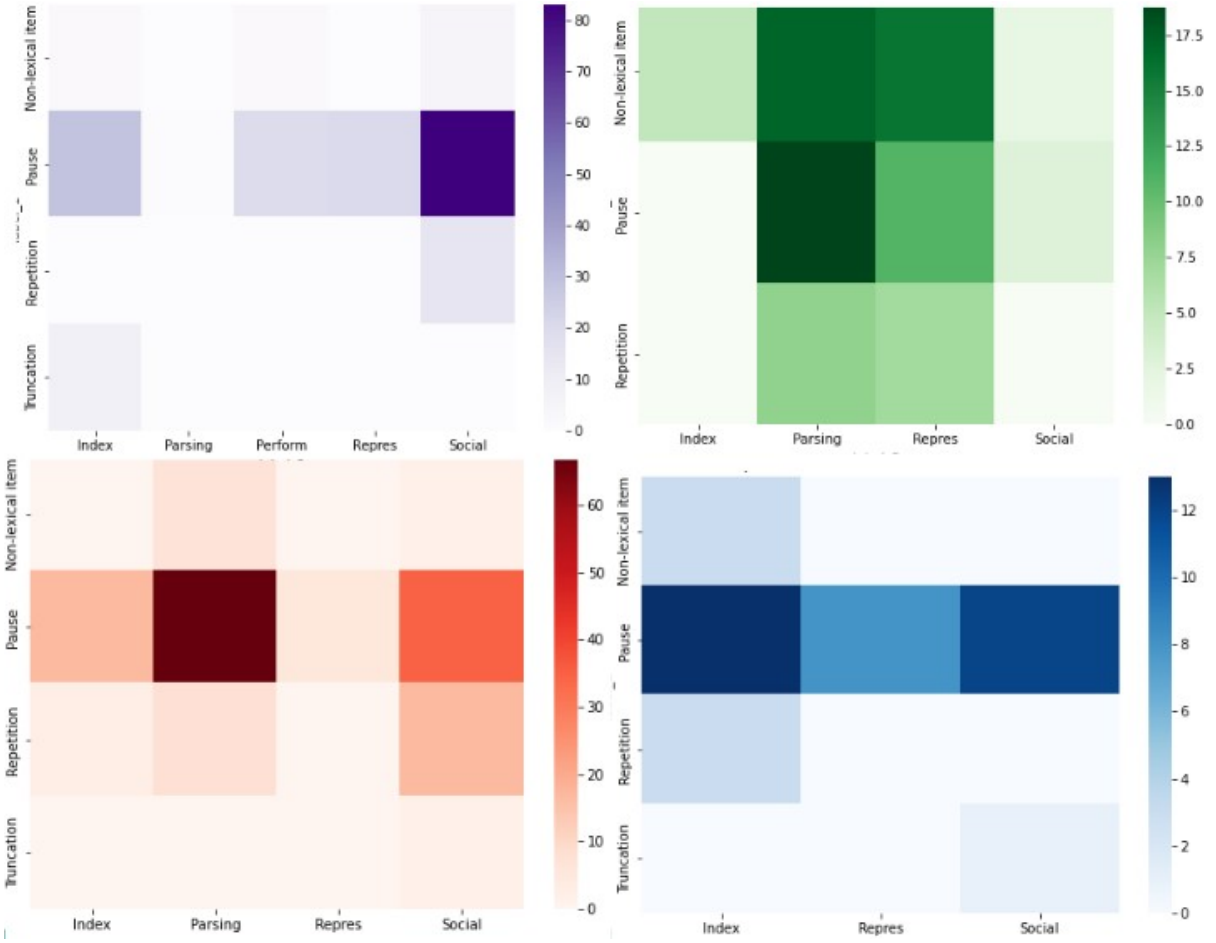

Figure 4: Heat maps of co-occurrence of *speech constants* and *functions of hand gesture* (by-legislator). The upper left image belongs to high_C, the upper right image belongs to high_B, the lower left image belongs to high_D, and the lower right image belongs to high_A.

`Modal`[14].

Trotta and Guarasci (2021) claim that more hand gestures go with semi-lexical items ("non-lexical item" in our study) and that `pauses` can confirm LRH. In this way, if we take *speech constants* as the speech disfluency indicators, then `pauses` and `non-lexical items` seem to be the focused indicator to evaluate the LRH. In the following analysis, we focus on *function of hand gesture* co-occurring with `pause` and `non-lexical item`. These functions of concurrent hand gesture can be subcategorized into LRH-related functions (`Parsing` and `Representational`) and non-LRH-related functions (`Social`, `Indexical`, and `Performative`), for beat and representational gestures receptively correlate with different types of fluency (Hostetter and Alibali, 2007).

Starting from the LRH-related functions of

[14]As there is no co-occurrence between `Modal` and *speech constants*, this label is not displayed in Table 6.

hand gesture, Table 6 shows that *functions of hand gesture* co-occurring with `pause` and `non-lexical item` account for 42.6%. `Parsing` is the second-highest intended *function of hand gesture* co-occurring with `pause`; this is noticeably consistent with the obvious correlation between semantic fluency and beat gestures (Hostetter and Alibali, 2007). Although `pauses` co-occur with hand gestures for `Representational` rank fourth, it still comprises 13.3% of total occurrences. In the case of `non-lexical items`, hand gestures for `Parsing` and `Representational` functions show higher frequencies for appearing with `non-lexical item` (65.5%), suggesting that hand gestures co-occurring with `non-lexical item` are more likely to facilitate verbal delivery in formal speech. From the discussion above, it can be concluded that `pauses` and `non-lexical items` are often accompanied by hand gestures for `Parsing` and `Representational`, which appears to correspond with the findings of how gestures prime lexical retrieval reviewed in Gillespie et al. (2014).

When it comes to non-LRH-related functions of concurrent hand gestures, the pause is highly associated with hand gestures for `Social` function. This indicates that `pauses` seem not primarily to represent hesitation pauses but rather to emphasize the primary topic of the speech in interpellation. Subsequently, `Indexical` is the ranked third *function of hand gestures* synchronizing with `pause`, implying that speakers prefer to depict the referent with visual-motion modality. `Performative` function is the least frequent one, but its occurrence is still significant compared to other *speech constants*. `Indexical` function in `non-lexical item` case is subtly higher than `Social` and `Performative`. As shown in Figure4, it can be inferred that synchronous hand gestures of `pause` and `non-lexical item` also carry out information emphasis and referent depiction functions.

To sum up, in formal speech hand gestures co-occurring with *speech constants* related to speech disfluency are not just used to iconically represent the unspoken thoughts but also serve the function of reinforcing the verbal information.

4.3 Co-occurrence of individual legislators

This research takes formal speech as a research target to reexamine the applicability of LRH in individual performance since Gillespie et al. (2014) specify that LRH is less applicable if the speaker can use alternate strategies to circumvent lexical access difficulties that arise during improvised speech. Trotta and Guarasci (2021) illustrate that LRH does not confirm in all interviewers' performances, whereas the applicability of LRH in formal speech stays unclear. Accordingly, the purpose of this section is to highlight the functions adopted by all speakers and their implications related to LRH.

According to individual speaker behaviors in Figure 4, `Social`, `Indexical`, and `Representational` are the functions employed by all of the speakers. This exemplifies that information accentuation and referent portrayal are primary functions of synchronous hand gestures despite possible variations in individual style preferences. Notably, all speakers adopt the concurrent hand gestures for the `Representational` function when pausing, indicating the widespread use of nonverbal modalities to compensate for verbal delivery difficulties in improvised speech situations. This offers a new perspective to extend the suggestions presented by Gillespie et al. (2014), highlighting the general applicability of hand gestures to serve the lexical retrieval purpose in formal spontaneous speech contexts.

5 Conclusion

In conclusion, this paper highlights the creation of a multimodal corpus of Taiwanese languages and evaluates its research potential by investigating the lexical retrieval hypothesis in gestures and speech.

The case study using the MultiMoco dataset presented in this paper examines the application of multimodal corpora in the investigation of the lexical retrieval hypothesis, indicating that hand gestures often accompany *speech constants* such as `pauses` and `non-lexical items`, priming the function of lexical retrieval. By leveraging the corpus, our finding suggests that hand gestures are not solely for retrieval struggles but can also serve as means of emphasizing information. Additionally, the outcome of individual speech performances signifies the general applicability of hand gestures for the lexical retrieval purpose.

In the subsequent investigation, our emphasis will be on examining the potential correlation be-

tween hand movements and the content of regular speech (excluding non-speech elements). Following the current study, our objective is to conduct a thorough comparison of how various gesture functions are distributed in both disfluent and fluent speech contexts. We can also investigate the issue from neurolinguistic perspectives (Weisberg et al., 2017), with active learning in annotation expansion (Gal et al., 2017), or for Multimodal Learning Analytics (MMLA) applications in education disciplines (Chen et al., 2014). We believe that the continued development and utilization of the MultiMoco Corpus will pave the way for enhancing our understanding of the intricate interplay between verbal and non-verbal communication channels.

References

Jens Allwood. 2008. Multimodal corpora. In *Corpus Linguistics. An International Handbook*, pages 207–225. Berlin: Mouton de Gruyter.

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2005. The mumin annotation scheme for feedback, turn management and sequencing. In *Proceedings from the Second Nordic conference on Multimodal Communication. Gothenburg Papers in Theoretical*.

Tom J Barry, James W Griffith, Stephanie De Rossi, and Dirk Hermans. 2014. Meet the fribbles: novel stimuli for use within behavioural research. *Frontiers in Psychology*, 5:103.

Janet Beavin Bavelas, Nicole Chovil, Douglas A Lawrie, and Allan Wade. 1992. Interactive gestures. *Discourse processes*, 15(4):469–489.

Jana Bressem, Silva H Ladewig, and Cornelia Müller. 2013. 71. linguistic annotation system for gestures. In *Volume 1*, pages 1098–1124. De Gruyter Mouton.

Brian Butterworth and Geoffrey Beattie. 1978. Gesture and silence as indicators of planning in speech. *Recent advances in the psychology of language: Formal and experimental approaches*, pages 347–360.

Belinda Crawford Camiciottoli and Veronica Bonsignori. 2015. The pisa audiovisual corpus project: a multimodal approach to esp research and teaching. *ESP Today*, 3(2):139–159.

Ronald Carter and Svenja Adolphs. 2008. Linking the verbal and visual: new directions for corpus linguistics. In *Language, People, Numbers*, pages 275–291. Brill.

Lei Chen, Chee Wee Leong, Gary Feng, and Chong Min Lee. 2014. Using multimodal cues to analyze mla'14 oral presentation quality corpus: Presentation delivery and slides quality. In *Proceedings of the 2014 ACM workshop on multimodal learning analytics workshop and grand challenge*, pages 45–52.

Lei Chen, R Travis Rose, Ying Qiao, Irene Kimbara, Fey Parrill, Haleema Welji, Tony Xu Han, Jilin Tu, Zhongqiang Huang, Mary Harper, et al. 2006. Vace multimodal meeting corpus. In *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers 2*, pages 40–51. Springer.

Huang-Cheng Chou, Wei-Cheng Lin, Lien-Chiang Chang, Chyi-Chang Li, Hsi-Pin Ma, and Chi-Chun Lee. 2017. Nnime: The nthu-ntua chinese interactive multimodal emotion corpus. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 292–298. IEEE.

Elizabeth A Clark, J'Nai Kessinger, Susan E Duncan, Martha Ann Bell, Jacob Lahne, Daniel L Gallagher, and Sean F O'Keefe. 2020. The facial action coding system for characterization of human affective response to consumer product-based stimuli: a systematic review. *Frontiers in psychology*, 11:920.

Camille Debras. 2021. How to prepare the video component of the diachronic corpus of political speeches for multimodal analysis. *Research in Corpus Linguistics*, 9(1):132–151.

Allen T Dittmann and Lynn G Llewellyn. 1969. Body movement and speech rhythm in social conversation. *Journal of personality and social psychology*, 11(2):98.

Lotte Eijk, Marlou Rasenberg, Flavia Arnese, Mark Blokpoel, Mark Dingemanse, Christian F Doeller, Mirjam Ernestus, Judith Holler, Branka Milivojevic, Asli Özyürek, et al. 2022. The cabb dataset: A multimodal corpus of communicative interactions for behavioural and neural analyses. *NeuroImage*, 264:119734.

Paul Ekman and Wallace V Friesen. 1972. Hand movements. *Journal of communication*, 22(4):353–374.

Paul Ekman and Erika L Rosenberg. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR.

Carlos A Gomez Gallo, Scott Stoness, Mary Swift, and Michael Tanenhaus. 2006. Software architectures for incremental understanding of human speech.

Maureen Gillespie, Ariel N James, Kara D Federmeier, and Duane G Watson. 2014. Verbal working memory predicts co-speech gesture: Evidence from individual differences. *Cognition*, 132(2):174–180.

Autumn B Hostetter and Martha W Alibali. 2007. Raise your hand if you're spatial: Relations between verbal and spatial skills and gesture production. *Gesture*, 7(1):73–95.

Mercedes Querol Julián. 2011. *Evaluation in discussion sessions of conference paper presentations*. LAP LAMBERT Academic Publishing.

Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

Dawn Knight, Svenja Adolphs, Paul Tennent, and Ronald Carter. 2008. The nottingham multi-modal corpus: A demonstration.

Robert M Krauss. 1998. Why do we gesture when we speak? *Current directions in psychological science*, 7(2):54–54.

Robert M Krauss and Uri Hadar. 1999. The role of speech-related arm/hand gestures in word retrieval. *Gesture, speech, and sign*, 93.

Andy Lücking, Kirsten Bergmann, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2010. The bielefeld speech and gesture alignment corpus (saga). In *LREC 2010 workshop: Multimodal corpora–advances in capturing, coding and analyzing multimodality*.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. Mediapipe: A framework for building perception pipelines.

Evelyn Z McClave. 2000. Linguistic functions of head movements in the context of speech. *Journal of pragmatics*, 32(7):855–878.

Iain McCowan, Samy Bengio, Daniel Gatica-Perez, Guillaume Lathoud, Florent Monay, Darren Moore, Pierre Wellner, and Hervé Bourlard. 2003. Modeling human interaction in meetings. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 4, pages IV–748. IEEE.

Iain McCowan, Daniel Gatica-Perez, Samy Bengio, Darren Moore, and Hervé Bourlard. 2005. Towards computer understanding of human interactions. In *Machine Learning for Multimodal Interaction: First International Workshop, MLMI 2004, Martigny, Switzerland, June 21-23, 2004, Revised Selected Papers 1*, pages 56–75. Springer.

David McNeill. 1992. Hand and mind1. *Advances in Visual Semiotics*, 351.

IZIDOR Mlakar, ZDRAVKO Kačič, and MATEJ Rojc. 2017. A corpus for investigating the multimodal nature of multispeaker spontaneous conversations–eva corpus. *WSEAS transactions on information science and applications*, 14:213–226.

Cornelia Müller. 2004. Forms and uses of the palm up open hand: A case of a gesture family. *The semantics and pragmatics of everyday gestures*, 9:233–256.

Demet Özer and Tilbe Göksun. 2020. Gesture use and processing: A review on individual differences in cognitive resources. *Frontiers in Psychology*, 11:573555.

Magali Paquot and Stefan Th Gries. 2021. *A practical handbook of corpus linguistics*. Springer Nature.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Frances H. Rauscher, Robert M. Krauss, and Yihsiu Chen. 1996. Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science*, 7(4):226–231.

Matthias Rehm, Elisabeth André, Nikolaus Bee, Birgit Endrass, Michael Wissner, Yukiko I Nakano, Afia Akhter Lipi, Toyoaki Nishida, and Hung-Hsuan Huang. 2009. Creating standardized video recordings of multimodal interactions across cultures. *Multimodal corpora*, (5509):138–159.

Han Sloetjes and Peter Wittenburg. 2008. Annotation by category - elan and iso dcr. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Lisa Smithson and Elena Nicoladis. 2013. Verbal memory resources predict iconic gesture use among monolinguals and bilinguals. *Bilingualism: Language and Cognition*, 16(4):934–944.

Daniela Trotta and Raffaele Guarasci. 2021. How are gestures used by politicians? a multimodal co-gesture analysis. *IJCoL. Italian Journal of Computational Linguistics*, 7(7-1, 2):45–66.

Miriam Voghera. 2001. Teorie linguistiche e dati di parlato. *Dati Empirici E Teorie Linguistiche*, pages 75–96.

Aaron Weinberg, Tim Fukawa-Connelly, and Emilie Wiesner. 2013. Instructor gestures in proof-based mathematics lectures. In *Proceedings of the 35th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, volume 1119.

Jill Weisberg, Amy Lynn Hubbard, and Karen Emmorey. 2017. Multimodal integration of spontaneously produced representational co-speech gestures: an fmri study. *Language, cognition and neuroscience*, 32(2):158–174.