Multidimensional Labeling of Gesture in Communication: the M3D Proposal

Corpus Pragmatics
https://doi.org/10.1007/s41701-025-00197-2

ORIGINAL PAPER

Multidimensional Labeling of Gesture in Communication: the M3D Proposal

Patrick Louis Rohrer[1] · **Ulya Tütüncübasi**[2] · **Júlia Florit-Pons**[2] · **Ingrid Vilà-Giménez**[3] · **Núria Esteve-Gibert**[4] · **Ada Ren-Mitchell**[5] · **Stefanie Shattuck-Hufnagel**[5] · **Pilar Prieto**[6]

Received: 20 February 2025 / Accepted: 26 May 2025

Abstract
Communication is multimodal in that speakers use not only their voices, but also co-speech gestures to communicate. Recent insights suggest that gestural behavior has a strong association with prosodic structure and that a single gesture can communicate various semantic and pragmatic meanings. This highlights the importance of developing a comprehensive, flexible, and transparent approach to gesture annotation that accounts for multiple dimensions of gesture, including a gesture's form, prosodic properties, and semantic and pragmatic contributions. To address this need for an increasingly dimensionalized approach to multimodal data annotation, the main goal of this paper is to present and describe a novel labeling system for manual gestures. The MultiModal MultiDimensional (M3D) system consists of an open access package that has been developed in coordination with five different labs working on gesture and its interaction with speech. The package includes a set of reliable annotation guidelines, a validated training program, and two annotated audiovisual corpora that represent over 60 minutes of lecture-style speech.

Keywords Annotation · Gesture · Speech · Multimodality

✉ Patrick Louis Rohrer
patrick.rohrer@donders.ru.nl

1 Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, The Netherlands

2 Department of Translation and Language Sciences, Pompeu Fabra University, Barcelona, Catalonia, Spain

3 Department of Specific Didactics, University of Girona, Girona, Catalonia, Spain

4 Department of Psychology and Educational Sciences, Open University of Catalonia, Barcelona, Catalonia, Spain

5 Speech Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA

6 Institució Catalana de Recerca i Estudis Avançats (ICREA), Department of Translation and Language Sciences, Pompeu Fabra University, Barcelona, Catalonia, Spain

Published online: 26 June 2025

Introduction

Researchers today acknowledge that communication is a multimodal phenomenon whereby multiple "modes of communication" come together to signal structure and to express meaning. This view has been supported by the fields of sign language, cognitive neuroscience, language evolution, multilingualism, development, and learning (Perniss, 2018; Vigliocco et al., 2014). Thus, oral communication is not only the expression of its propositional part via morphosyntax and segmental phonology, but can also include superimposed layers of meaning conferred via spoken prosody or the visual (gestural) mode. In fact, the study of gesture ("a visible action of any body part, when it is used as an utterance, or as part of an utterance", Kendon, 2004, p. 7) has been instrumental in widening our lens of investigation regarding the study of multimodal communication. Recent insights suggest that gestural behaviors have strong parallels with prosodic structure in speech and communicate both semantic and pragmatic meaning (Brown & Prieto, 2021; Lopez-Ozieblo, 2020; McNeill, 2006; Prieto et al., 2018; Shattuck-Hufnagel & Prieto, 2019) highlighting the importance of developing an approach to the study of gesture and speech that accounts for the multimodal nature of communication.

Essential work by Kendon (1980) and McNeill (1992) established how speech and speech-accompanying hand gestures are integrated temporally, semantically, and pragmatically, and developed practical ways of categorizing different gesture types. However, such categorizations are often based on varying criteria in their relationship to speech. For example, McNeill's (1992) classification of hand gestures identifies 'iconic', 'metaphoric', and 'deictic' gestures in terms of how they are semantically referential to propositional speech content, while 'beat' gestures are defined by their relationship with speech prosody as "highlighters" of prominent syllables. However, this dichotomy between gestures that contribute to meaning and gestures that are solely prominence-markers is challenged by research showing how all gesture types tend to associate with prominent syllables (e.g., Florit-Pons et al., 2023; Rohrer et al., 2023; Shattuck-Hufnagel & Ren, 2018) and that beat-like movements can be produced with referential gestures (McNeill, 1992; Parrill, 2008). In 2006, McNeill clarified his stance, saying:

> I wish to claim, however, that none of these 'categories' is truly categorical. We should speak instead of *dimensions* and say: iconicity, metaphoricity, deixis, 'temporal highlighting' (for beats), social interactivity, or some other equally unmellifluous (but accurate) terms conveying dimensionality. The essential clue that these are dimensions and not categories is that we often find iconicity, metaphoricity, deixis, and other features mixing in the same gesture. Beats often combine with pointing, and many iconic gestures are also deictic. We cannot put them into a hierarchy without saying which categories are dominant, and in general this is impossible. A practical result of dimensionalizing is improvement in gesture coding, because it is no longer necessary to make forced decisions to fit each gesture occurrence into a single box. (Italics in original, McNeill, 2006 p. 60)

Following McNeill's view of dimensionalizing gesture, we observe the pressing need for tools to describe three separate dimensions (i.e., form, prosody and meaning) of gesture. As such, the aim of the current paper is to present and describe the M3D (for MultiModal MultiDimensional) labeling system. M3D allows for the independent assessment of gestures across three dimensions in a non-mutually exclusive manner, offering a more flexible alternative to systems that require classification into discrete types. Specifically, a FORM DIMENSION captures aspects such as hand shape, handedness, trajectory shape, location with respect to the body, etc. As aspects from speech prosody such as prominence and chunking of words into groups also surfaces in gesture (e.g., Kendon, 1980), a PROSODIC DIMENSION captures the prominence and grouping aspects of gestural movements. Finally, a MEANING DIMENSION captures both a gesture's semantic referentiality to propositional speech content, and its pragmatic contribution to speech. Crucially, M3D comprehensively accounts for these three overlapping dimensions of gesture in an independent non-mutually exclusive fashion.

Operationalizing the annotation and analysis of gesture across these three dimensions offers both theoretical and practical advantages. First, by not assuming a 1:1 form-to-function relationship (e.g. iconic gestures purely convey semantic meaning, and beat gestures are only up-and-down movements which highlight prominence in speech), researchers are able to comprehensively assess any gesture across all dimensions (visual form, visual prosody, visual meaning). Second, it facilitates the study of interactions between language modalities. For example, researchers can understand how the prominence of any gesture regardless of "type" may co-vary as a function of auditory prominence in speech. By not forcing researchers to isolate gestures into a set of categories, richer datasets will then afford a much more nuanced view of gesture and its interactions with other aspects of speech communication.

M3D is openly accessible and includes a variety of resources: a detailed annotation manual, an ELAN template (Wittenburg et al., 2006), two M3D-TED corpora (comprising over 1 hour of annotated speech in English and French)[1] and, crucially, an interactive online training program[2] that has been validated by users for usability, acceptability, and feasibility. This is the first gesture annotation system to our knowledge that offers such a vast amount of openly-accessible learning resources. By drawing from these resources, researchers are able quickly to gain an understanding of the system via a large variety of examples that differ in complexity and obtain technical hands-on training on how to annotate gestures according to M3D guidelines without needing direct supervision or prior experience in gesture research.

All in all, M3D offers an annotation system that is in line with and builds upon current theoretical approaches in gesture studies, and advances the field by offering a transparent toolkit that can be applied to address a variety of research questions. Its open-access model, integrated training program, and multidimensional non-exclusive approach make M3D a versatile and accessible choice for researchers

[1] Accessible at https://osf.io/ankdx/.

[2] Accessible at https://m3d.upf.edu/.

across disciplines[3]. Further, the system is particularly well suited not only to assess the three dimensions of any gesture, but also to compare different modalities in communication, allowing researchers to better understand language as an integrative multimodal phenomenon with complex interactions that arise among different modes. It is worth noting that recent innovations in the field of gesture studies (such as motion tracking technologies, automatic tools for analysis, etc.) as well as the broad interdisciplinary nature of the field suggest a need for a system that remains up-to-date with the latest technologies and approaches to the study of gesture. Consequently, the M3D system is open to further contributions and development from the research community (e.g., more thorough guidelines for the annotation of non-manual articulators).

The system presented in this article reflects version 2.1 of the labeling manual. The following Section "Main features of the MultiModal MultiDimensional (M3D) labeling system" of this paper will describe the main features of the M3D system. Section "M3D resources" will describe the various resources available for researchers to learn and implement M3D. Finally, Section "Assessment of the M3D system" will describe the results of the inter-rater reliability tests carried out on key aspects of gesture annotation, as well as the feasibility of the M3D training program as a tool to teach gesture annotation.

Main features of the MultiModal MultiDimensional (M3D) Labeling System

A Multidimensional View

The perspective adopted here follows Kendon's definition of gesture as communicative movements which can be produced by the hands, the head, facial expressions, or any other body part (1972, p. 204–205; 2004). Following this definition, the M3D system is grounded in the idea that any gesture (manual or non-manual) can be assessed according to three dimensions: (1) the Form dimension, which refers to the description of the articulator configuration and kinematic profile of gesture; (2) the Prosodic dimension, which refers to hierarchical structure of movement in gesture units and phases, as well as prominence properties of gesture, and 3) the Meaning dimension, which captures the semantic (i.e. referential) and pragmatic meanings that can be expressed by gestures. Figure 1 shows a schematic representation of the three dimensions as well as their main sub-features. While M3D contains guidelines to annotate both manual and non-manual gestures, for simplicity the current description will focus on the annotation of manual gestures. In the following subsections, we briefly describe each dimension of the system. For more details about specific labeling guidelines, please refer to the labeling manual available on the Open Science Framework (OSF).[4]

[3] For a detailed comparison between M3D and other established annotation frameworks, see Rohrer (2022, Chapter 2).

[4] Accessible at https://osf.io/ankdx/.

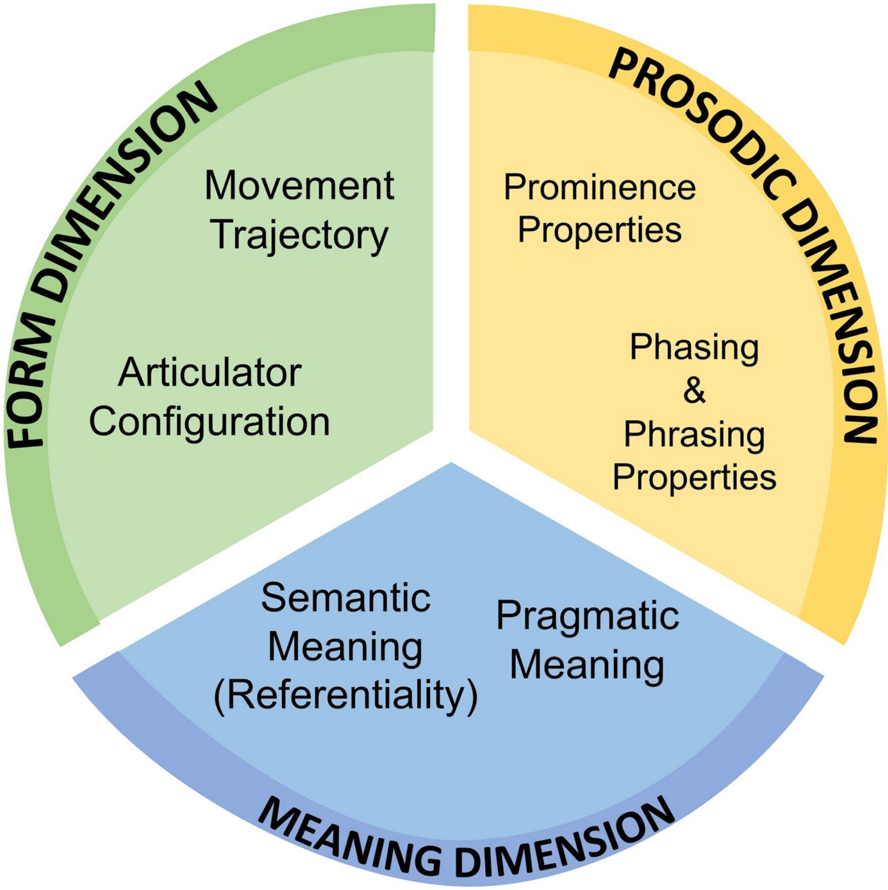

Fig. 1 Overview of the M3D system, including the 3 dimensions and the sub-features of each dimension

The Form Dimension

This dimension describes the physical nature of the movement, including a qualitative description of movement for multiple independent articulators. The manual articulator tier set includes aspects such as articulator configuration (i.e., the handedness of the gesture [e.g., right hand, left hand, both hands], handshape [e.g., fist, relaxed] and the orientation of the palm [e.g., up, forward]) as well as a qualitative description of the movement trajectory (i.e., its shape [straight, curved, looping] and its direction [e.g., up, down, left, right, etc.]). The form labels are largely based on the MIT Gesture Coding Manual[5] (Ren-Mitchell & Shattuck-Hufnagel, 2012).

It is important to note that this tier set is a description of the physical movement, and no interpretation of the communicative function of those movements should be

[5] Accessible at https://scg.mit.edu/gesture/coding-manual.html.

directly assumed. For example, gestures with a "palm-up open-hand" form are often associated with specific pragmatic meanings (e.g., Cooperrider et al., 2018; Ferré, 2011). However, M3D's form tier set is only concerned with annotating the physical description of the articulator and movement, and meaning is to be coded independent of form in the meaning dimension. Similarly, beat gestures have been traditionally described as up-down or in-out movements. However, prominence-lending properties and semantic/pragmatic meanings are assessed in the prosodic and meaning dimensions, respectively.

The Prosodic Dimension

It has long been attested that spoken prosody and gesture are closely linked—not only do their prominent parts generally co-occur in time, but recent studies show they may also be connected in space, with larger gestures matching more prosodically prominent syllables (Ambrazaitis & House, 2023). This close link between movement and speech even serves as the basis for the Articulatory Phonology framework, whereby gestures of the mouth (including lip and tongue movements) correlate with auditory prosodic modulations (Parrell et al., 2014; Krivokapic et al., 2017). The M3D proposal aligns with the conceptualization of prosody as articulated by Prieto, Esteve-Gibert, and Shattuck-Hufnagel (2025), which posits that prosody operates as a unified, modality-independent structure, coordinating both spoken and gestural elements of communication.

Prosodic structure has been described more abstractly as the raw organizational structure within which an utterance is hierarchically organized into phonologically defined constituents with heads (Beckman, 1996, p. 19). Two main aspects of prosody (that is, prominence and phrasal grouping) can be expressed through bodily visual means as well as through acoustic cues in speech. In fact, Kendon's (1980) description of the hierarchical organization of gesture movements parallel the hierarchical structure of speech prosody, where large gesture units can be broken down into one or a series of gesture phrases (i.e., individual gestures), which in turn can be divided into even smaller gesture phases (i.e., different stages in the execution of a gesture), with stroke phases being the only obligatory and most prominent phase.

Following this line of research, M3D proposes a system to assess how visual bodily properties are temporally organized and relate to the prosodic dimension of language by specifically annotating two main aspects of gesture, namely gestural movement organization into constituents (i.e., phasing and phrasing properties), as well as their prominence properties. Importantly, these prosodic characteristics refer specifically to gesture and not to acoustic prosody.

> The phasing and phrasing properties of gesture specifically refer to the hierarchically organized structure of gestural movements, where multiple gestural movements may come together to form larger gesture units (i.e., PHRASING), and alternatively any gesture may be broken down into component phases, such as the preparation, stroke, apex, or recovery (i.e., PHASING, as per Kendon, 1980; Kita

et al., 1998; Loehr, 2004). Thus, M3D has two tiers for annotating gesture units and gesture phases.

In terms of the prominence properties of gestures, we adopt a basic understanding of prominence as "a linguistic element that stands out from neighboring elements" (Grice & Kügler, 2021, p. 253). From this perspective, prominence surfaces in gestural movements in a number of ways. First, the stroke portion of the gesture is generally defined as the most kinematically salient movement in the gesture phrase. We can also assess perceptually the degree of saliency of a gestural movement in relation with neighboring gesture movements or neighboring strokes (i.e., degree of gestural prominence). Other aspects of prominence in gesture include assessing if a movement seems to be a particularly 'beat-like' or staccato-like movement reflecting an underlying rhythmic pulse (i.e., beat-like-ness), or if multiple subsequent gestures are produced in a temporally rhythmic fashion, giving the impression of marking a particular tempo (i.e., Rhythmic Groups of Gestures). M3D thus has the following 3 tiers that may be annotated: gestural prominence, beat-like-ness, and rhythmic groups of gesture.

As previously mentioned, the prosodic dimension in M3D specifically refers to the prosody of gesture. However, research has shown interest in looking at the relationship between acoustic prosody and gesture. M3D thus proposes various options that differ in complexity to assess this relationship. The simplest is a perceptual approach in which annotators watch a gesture while listening to the concomitant speech and judge if gestural prominence aligns with prosodic prominence (termed the Prominence Association Component), which may be suitable for labelers who lack any sort of prosodic annotation training. Alternatively, M3D espouses selecting an established method from the field of speech prosody to obtain independent acoustic annotations of the prosody of speech, such as ToBI annotation or Rapid Prosodic Transcription (Cole & Shattuck-Hufnagel, 2016). The choice of method depends on the researcher's experience level and specific research questions. For data comparing the prosodic association component to independent prosodic annotations of speech, see Rohrer et al. (2019).

The Meaning Dimension

The meaning dimension captures how a gesture is *meaningfully* integrated with speech via its semantic and/or pragmatic contributions. Unlike the form and prosody dimensions—which are always present—different aspects of meaning may be present or absent, that is, gestures may or may not convey clear semantic content or serve pragmatic functions. For example, traditional "beat gestures" are typically considered semantically empty, instead contributing at the discourse-pragmatic level (McNeill, 1992), though some research suggests that they may convey spatial semantics (Yap & Casasanto, 2018). Similarly, miming a steering wheel conveys semantic meaning (driving) when accompanied by the utterance "Dave drove here". The same gesture may additionally serve a pragmatic function if produced while asking "Is Dave coming?", reinforcing the illocutionary intent of the utterance and implying that the real question is whether Dave is coming by car. M3D's approach sees any gesture as potentially contributing both semantic and pragmatic meaning to an utterance in a non-mutually exclusive manner (that is, a single gesture of any kind may represent semantic meaning, which does not automatically preclude it from also serving one or more pragmatic functions, and vice versa). As a result, the meaning dimension is divided into two non-mutually-exclusive parts: one to assess the semantic meaning, and another to assess the various pragmatic functions of gesture.

Semantic Meaning: Gesture Referentiality

Largely based on McNeill (1992, 2006), M3D adopts a broader view in terms of assessing a gesture's semantic contribution (or lack thereof). Referred to as GESTURE REFERENTIALITY, M3D proposed making a first distinction as to whether gestures convey semantic meaning or not, and in the case they do convey semantic meaning, a secondary assessment of what sort of semantic meaning is expressed.

REFERENTIAL GESTURES are those which integrate aspects of semantic meaning based on a clear referent, by representing these aspects via different degrees of iconicity, metaphoricity, or deixis (or any combination thereof). Using individual tiers in ELAN for each aspect of referential meaning, labelers can annotate any potential combination of those meanings present in a superimposed manner. NON-REFERENTIAL GESTURES, on the other hand, are those which do not have a clear and direct link to the semantic content of speech, regardless of its form or relationship with spoken prosody (in other words, they do not convey referential meaning via iconicity, metaphoricity, or deixis). A third type of gesture, EMBLEMS, are also coded within this tier set. This term refers to gestures that have specific conventionalized meanings widely-shared within a community, such as a "thumbs up" to indicate approval. Such gestures are highly constrained in terms of form (doing a "thumbs up" gesture with the thumb bent will not be clearly understood as a sign of approval) and can generally be understood without corresponding speech[6] (see McNeill, 1992).

Pragmatic Function of Gestures

The literature has started to establish a wide range of pragmatic functions that gestures carry out. However, most studies restrict the analysis of pragmatic functions to gestures that have a specific form, for example, 'palm-up open hand' gestures (Cooperrider et al., 2018; Ferré, 2011), 'hand flips,' 'finger bunch,' or 'ring' gestures (Kendon, 1995, 2004). M3D proposes to assess the pragmatic contribution of all gestures independently of their form and their referential properties (i.e., both referential and non-referential gestures can contribute to pragmatic meaning).

[6] Unlike emblems, recurrent gestures (Ladewig, 2014) rely on accompanying speech for their specific meaning, despite having relatively stable forms. In M3D, they are not treated as emblems but are annotated based on referential and pragmatic functions. Their recurrence can be noted in the form dimension or on an optional "Recurrent gesture" tier

Table 1 The five pragmatic domains, subfields of pragmatics involved, and examples of individual pragmatic functions

Pragmatic domain	Subfields	Example functions
Discourse Organization	Information structure, Discourse structure	Focus marking, Listing
Operation	Operation	Affirmation, Negation
Stance	Politeness, Epistemicity, Affective stance, Irony	Positioning, Alignment
Speech Acts	Speech Acts	Request, Assertion
Interaction	Turn management	Turn-request, Turn-hold

For definitions of these specific functions, please refer to the labeling manual

The assessment of the pragmatic functions of gestures proposed by M3D has been developed based on a review of the pertinent literature on gesture pragmatics (see below). Specifically, previously identified pragmatic functions of gesture are grouped according to related subfields in pragmatics. M3D describes five main PRAGMATIC DOMAINS (shown in Table 1) for which an initial subset of functions have been described.

The domain of DISCOURSE ORGANIZATION refers to how speakers may use gestures to organize their discourse, and thus relies on two subfields of pragmatics: Information structure (e.g., Krifka, 2008) and discourse structure (e.g., Grosz & Sidner, 1986). Functions include focus marking (Ebert et al., 2011; Rohrer & Florit-Pons et al., 2022), topic-comment or theme-rheme structure marking (Ferré, 2014), and marking the information status of referents (Baills & Baumann, 2025). Within discourse structure, labels include, for example, gesturally marking the start or end of sequences, parenthetical digression, listing, or discourse-cohesion functions (Bolly & Boutet, 2018; Kendon, 2017; Ladewig, 2014). Multimodal affirmation or negation would fall under the domain of OPERATION (e.g., Kendon, 2017; Prieto & Espinal, 2020), where affirmation refers to any gesture that expresses a positive interpretation, while negation refers to any gesture that expresses a negative interpretation. The STANCE domain embodies a broad view of stance as "personal feelings, attitudes, value judgments, or assessments" (Biber et al., 1999, p. 966, as cited in Freeman, 2015). An initial subset of labels have been proposed to account for areas such as epistemic, affective, and politeness stance (see, e.g., Bolly & Boutet, 2018; Brown & Prieto, 2021, among many others). The SPEECH ACT domain refers to how speakers may use gestures to perform speech acts; thus this domain identifies the type of speech act that is being produced and the illocutionary force of the gesture (Directives, Representatives, etc., as per Searle, 1975; see also Kendon, 1995; Payrató, 1993). Finally, the INTERACTIONAL domain refers to when speakers use gestures to manage discourse in interaction with interlocutors, particularly in turn-taking (e.g., Levinson & Torreira, 2015; Sacks et al., 1974).

Just as referential gestures can be characterized by different degrees of iconicity, metaphoricity, or deixis, a single gesture can fulfill multiple pragmatic functions simultaneously (see, e.g., Lopez-Ozieblo, 2020). M3D proposes to annotate these multiple simultaneous pragmatic functions in a superimposed manner. It is also important to reiterate that referential gestures may also contribute pragmatic meaning to speech. Kendon (2017) describes how the handshape of deictic gestures may change as a result of how the pointing is being used in discourse (for example, to distinguish two different objects vs. to comment on an object). Additionally, pragmatic functions are not limited to particular hand forms. This is evidenced in the often overlapping and even contradictory functions that form-based analyses have shown (for example, a palm-up open hand form has been linked to both lacking knowledge as well as showing obviousness; see Cooperrider et al., 2018).

M3D Resources

In order to provide a standard and transparent method of gesture labeling, the system has been explicitly made openly accessible and further developed with additional resources to facilitate its implementation in multimodal corpora by other researchers. The M3D project and its various resources are hosted on OSF,[7] an open-access permanent platform which can be cited in publications. Table 2 lists the resources available on the OSF website.

The M3D Labeling Manual contains a detailed description of the system (along with theoretical justifications and a bibliography). It also offers detailed step-by-step instructions that can be adapted according to each individual researcher's objectives. The manual also contains many examples linked to actual real-world data to ensure clarity for novel coders, and to highlight cases of ambiguity that have come up when applying M3D to the M3D-TED corpus. A set of "tips" to overcome such difficulties have been included in the manual. Because such a multidimensional approach is highly interdisciplinary, it is designed to be flexible and adaptable to the individual researcher's goals. To this end, the M3D labeling manual includes recommendations for many different levels of analysis, so that annotators may choose the specific levels they will focus on for addressing their own research questions.

In addition to the manual, an ELAN annotation template file is also available. This contains tiers for every aspect of coding described in the M3D labeling manual, organized in hierarchical order (i.e., organized as parent/child tiers for coding various aspects of the form, prosodic, and meaning dimensions). The tiers are associated with controlled vocabularies, to facilitate consistent use of M3D labels. As such, researchers can download the template and begin working with M3D right away, selecting tiers to match their interests.

The English and French M3D-TED corpora are the result of applying M3D to real data, specifically TEDTalks. The annotations were carried out by a two-person team composed of the first two authors. Only a select number of M3D dimensions/tiers were annotated, namely gesture phasing, apex, gesture referentiality, and the pragmatic domain (English corpus). In addition to annotations for gesture, speech

[7] Accessible at https://osf.io/ankdx/.

Table 2 List of M3D resources available online and a brief description

Resource	Description
M3D Labeling Manual	A detailed description of the M3D system, including theoretical justifications, step-by-step instructions, workflow tips, and examples.
M3D ELAN annotation template	Includes Tier Hierarchies and Controlled Vocabularies used in M3D
English and French M3D-TED corpora	Time-aligned dataset of TED talks in English and French, annotated following M3D (5 speakers of each language, totaling over 60 minutes of multimodal speech)
	•Selection of M3D gesture annotations:
	Prosodic dimension
	Gesture units, phases, apex
	Meaning dimension
	Gesture referentiality
	Pragmatic domain annotations
	•Additional speech annotations
	Time-aligned orthographic transcription
	Information Structure
	Information status of referents, contrast
	ToBI prosodic annotations
M3D Training Program (external link hosted at https://m3d.upf.edu)	A series of lessons which include short theoretical and practical tutorial videos, as well as interactive exercises to train M3D annotation.
Resources for post-annotation data management (in R)	OSF links to other sub-projects (components) using M3D or the M3D-TED corpora

prosody (ToBI) and information structure (information status of referents, contrast) were also annotated in Praat (Boersma & Weenink, 1992) and imported into ELAN. Combined, the M3D-TED corpora contain a total of 61 minutes and 14 seconds of annotated multimodal American English and French speech (5 speakers in each language), with a total of 2,676 annotated manual co-speech gestures. Table 3 shows the descriptive statistics of gesture production across languages and speakers in terms of number of gesture strokes, as well as gesture rate in terms of words per gesture stroke and gesture strokes per minute.

The M3D-TED corpora has served multiple functions—it is a valuable openly available dataset for research (e.g., Rohrer, 2022; Rohrer et al., 2023), it has been a resource for examples for both the labeling manual as well as the training program, and it was used to assess the reliability of the M3D system (see Section "Assessment of the M3D system").

The OSF page also links to the M3D training program for manual gesture annotation.[8] The program was developed in accordance with the ADDIE paradigm

[8] Accessible directly at http://m3d.upf.edu.

Table 3 Descriptive statistics of manual co-speech gesture annotation in the M3D-TED corpus for each speaker, including overall total and average values

Speaker (Sex)	Language	Amount of multimodal speech (hh:mm:ss)	Number of Gesture Strokes	Words per Gesture Stroke	Gesture Strokes per Minute
AS (M)	English	00:04:02	163	4.61	40.45
EG (F)	English	00:05:00	282	3.79	56.4
ES (F)	English	00:05:48	275	2.92	47.41
MS (F)	English	00:03:05	159	3.62	51.62
SJ (M)	English	00:05:47	277	4.71	47.92
DL (M)	French	00:05:36	323	3.65	57.68
FB (F)	French	00:09:13	359	5.33	38.94
JP (M)	French	00:05:04	230	4.04	44.75
KF (M)	French	00:07:22	197	7.21	26.77
MD (F)	French	00:10:17	411	4.48	50.99
TOTAL	–	01:01:14	2676	–	
Average (± Standard Deviation) per speaker		00:06:07	267.6 (± 78.4)	4.44 (± 1.13)	46.29 (± 8.71)

hours:minutes:seconds

(Branch, 2009), a widely accepted framework for creating learning resources, which emphasizes a learner-centered and goal-oriented approach. The program is housed on an institutional website, where, in addition to introductory video-recorded tutorials, individual CHALLENGES or modules relate to the various dimensions and aspects of M3D. Each challenge includes video-recorded tutorials explaining the conceptual approach used in M3D, and labeling tutorials so that researchers can watch first hand actual gesture annotation demonstrations and better understand the workflow with M3D. The challenges also contain interactive practical exercises so that learners can practice identifying different aspects of gesture and annotating them, receiving automatic feedback and explanations. This resource is the first of its nature in the field of gesture studies, offering an openly accessible tool to teach novice labelers to identify and correctly annotate manual gestures.

Multidimensional annotation offers flexibility in data analysis. For example, when researchers want to "count the number of gestures by type", they can choose to count stroke annotations and categorize them as referential or non-referential, tally all referential dimensions that surface, or label a "predominant" referential dimension to approximate traditional gesture categories. The annotation manual provides guidance on these options. Additionally, the OSF page also contains links to sub-project pages (i.e., 'components') that host studies that have made use of M3D or the M3D-TED corpus. By having access to these pages, researchers can see how M3D has been applied in practice, as well as access code scripts for different approaches to data analyses. Such information may aid researchers in post-annotation data management.

Assessment of the M3D System

Inter-Annotator Reliability

Reliability between two experienced M3D coders (i.e., the first two authors) on key aspects of M3D was evaluated using the English M3D-TED corpus. Following Kita et al. (1998), a general protocol involving two passes was followed, where coders independently annotated approximately 1 minute's worth of gesture annotations per TEDTalk speaker (representing approximately 20% of the entire corpus). Initial reliability (i.e., the first pass) was then assessed based on the independent coding. Then, the two coders compared and revised their annotations together so that cases of disagreement could be reviewed. A second assessment of reliability was then carried out over the same dataset (i.e., the second pass) to understand the extent to which the coders could not resolve their disagreements.

Due to the nature of the coding system, reliability was assessed through different means depending on which aspect was to be evaluated. Four specific aspects of the annotations were assessed for reliability, namely gesture phasing, apex location, gesture referentiality and pragmatic domain. These aspects were chosen as the former two represent aspects that are traditionally annotated in multimodal corpora, and the latter two represent some of the main contributions of M3D to the field, namely

Table 4 The reliability metrics (Cohen's Kappa) obtained across speakers for individual gesture phases, as well as the global kappa value from 2 coders

	Cohen's Kappa values	
	First pass	Second pass
Preparation Identification	0.768	0.891
Stroke Identification	0.771	0.885
Incomplete or Abandoned Stroke Identification	0.799	1.0
Hold Identification	0.782	0.877
Recovery Identification	0.82	0.902
Global phasing (including annotations that did not have a "linked" counterpart)	0.65	0.808

the non-mutually exclusive manner in which gesture can convey different aspects of meaning.

Gesture Phasing Annotation Reliability

Gesture phasing was assessed using ELAN's 'Calculate Inter-Annotator Reliability' function to calculate Cohen's Kappa. This ELAN function is based on the algorithm created by Holle and Rein (2015), which works by 'linking' annotations that overlap a minimum of 60% and then calculating the Cohen's Kappa value, taking both the segmentation of the annotations and the assigned value of those annotations into account. In other words, linked annotations refer to annotations by each coder that overlap by more than 60% and thus suggest that both coders saw and annotated some phenomenon. Table 4 shows the Cohen's Kappa values obtained across speakers for each potential gesture phase (preparation, stroke, incomplete stroke, hold, and recovery) as well as the global Kappa for all phasing annotations including 'unlinked' annotations (i.e., cases where only 1 coder annotated a phenomenon). This allows for a more general view (the global kappa reflecting what each annotator observes and annotates independently), as well as a more precise view (to what degree coders were in agreement in identifying each type of gesture phase).

The results of the reliability calculations showed substantial levels of agreement already at the first pass for all gesture phases as well as globally in terms of gesture phasing. As expected, the Kappa values increase after revising the annotations, showing very high levels of agreement in all aspects of gesture phasing. This suggests that the coders' independent annotations showed substantial agreement at the first pass, and furthermore, they were able to resolve most[9] of their disagreements in terms of segmenting and identifying continuous streams of gestural movement.

[9] A small number of cases remained where the two annotators simply could resolve their differing opinions on how the gesture should be segmented into phases across time, leading to less than perfect rates of agreement.

	-99 ms	-66 ms	-33 ms	0 ms	33 ms	66 ms	99 ms	123 ms	
	3 frames	2 frames	one frame		2 frames	3 frames	4 frames	5+ frames	Total
Raw Count	4	8	44	110	63	33	20	23	305
Percent	1.3%	2.6%	14.4%	36.1%	20.7%	10.8%	6.6%	7.5%	100%
			50.5%						
		73.8%							

Figure 2 The distribution of coder 2's apex annotations relative to coder 1's apex annotations across time, binned in 33 ms groups (i.e., one frame)

Apex Annotation Reliability

The apex refers to a single point in time when the gesture has reached its maximum extension, suddenly stopped, or changed direction. Its annotation is based on a frame-by-frame analysis in ELAN which consequently cannot be assessed in terms of overlap. Thus, a qualitative assessment of reliability was used which evaluates the distance (in terms of 33 ms video frames) coders placed apexes within linked (agreed upon) gesture strokes (as per Kita et al., 1998). As such, the aim of this reliability measure is not to test whether both annotators saw the same point as an apex (indicating omission/commission errors), but rather when both coders saw a stroke and coded an apex, how closely in time did they code that apex.

Figure 2 shows the distribution of coder 2's apex annotations relative to coder 1's annotations across time, binned in 33 ms groups (i.e., one frame). In other words, the figure shows the distribution of the distance between the two coders' apex annotation in frames. Of the 305 total apex pairs that were analyzed, the majority of the apexes coded by the second coder fall within one frame (33 ms) of the first coder's apex annotation (50.5%) and almost 73.8% of apexes were within 2 frames (66 ms) of each other. None of coder 2's annotations occurred more than 3 frames before coder 1's, and only 23 annotations occurred more than 4 frames after coder 1's. As such, the qualitative assessment of the reliability of coding the apex seems quite high, especially considering Loehr (2004) included distances of up to 6 frames as being acceptable for coder agreement.

Gesture Referentiality Reliability

In order to assess the reliability of annotating gesture referentiality, Cohen's Kappa is not a suitable measure for two reasons: labelers could have multiple annotations for the same gesture (violating the Cohen's Kappa assumption of mutual exclusivity), and particularly for the current dataset, the large number of labels for non-referential gestures could potentially result in the "Kappa Paradox" (a scenario where one label is observed significantly more than any other, affecting the calculation of chance agreement [e.g., Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990]).

To overcome this, Gwet's Agreement Coefficient 1 (AC1; Gwet, 2008) was calculated in R (R Core Team, 2023) with MASI distance as the distance metric (see, e.g., Artstein & Poesio, 2008; Passonneau, 2006). MASI distances quantify the relative degree to which labelers agree for a set of non-mutually exclusive combinations of labels, while Gwet's AC1 uses the same formula as the Kappa yet calculates the chance agreement taking this bias into account. Thus it is resistant to the Kappa Paradox, yet can be interpreted in a similar fashion (Dettori & Norvell, 2020). In terms of the reliability for gesture referentiality, the global AC1 value across all five speakers resulted in very high rates of agreement (AC1 = .895, CI (.856, .933), $p < .001$).

Pragmatic Domain Reliability

The calculation of reliability for the identification of the five non-mutually exclusive pragmatic domains followed the same analysis laid out for gesture referentiality, that is, Gwet's AC1 with MASI distance as the distance metric. The resulting global AC1 value again reveals a high rate of agreement between the labelers (AC1 = .78, CI (.726, .825), $p < .001$).

Summary

Based on various reliability metrics presented in this section, it seems that it is possible for coding teams to work together to achieve acceptable levels of agreement when using M3D. In particular, this is the case for labeling gesture strokes and phasing, and for precisely pinpointing the placement of the gesture apex, as well as for understanding the semantic and pragmatic contribution of gestures to speech. While the reliability metrics presented here do not cover all aspects of M3D, the results from key novel aspects of M3D seem promising so far.

Assessment of the Training Program

In addition to the assessment of inter-annotator reliability of the labeling system between two experienced M3D coders, the training program was also assessed in terms of its usability (i.e., the degree to which a training program can be effectively, efficiently and satisfactorily used by learners to accomplish specific learning goals within the defined context of use), acceptability (i.e., how learners react to the program and the extent to which they are inclined to use it), and feasibility (i.e., the extent to which a program can be successfully used or carried out) [see Tütüncübasi (2023) and references therein].

In order to assess the program's usability, the participants evaluated the program according to the System Usability Scale, a validated, reliable, domain-general tool that assesses the usability of learning resources (Brooke, 2013). In order to assess the effectiveness of the program to train gesture annotation (a domain-specific aspect of usability), participants were asked to carry out a short annotation task on a 20-second video clip containing multiple gestures. In order to assess the program's acceptability, participants were asked to complete a questionnaire containing items from the evaluation questionnaire within the ADDIE paradigm (Branch, 2009, p. 155) as well as items from an in-house questionnaire to receive qualitative feedback specific to the M3D training program. Finally, feasibility was assessed through the starting-and-completion ratio (i.e., the number of participants that were able to complete the program in the given time span).

Thirteen participants were recruited to take part in the training program and subsequent analysis. The participants varied widely in terms of education level (from bachelors students to post-doctoral researchers) and field of linguistic study (including bilingualism, sign language, and animal communication). Importantly, none of the participants had any previous knowledge or experience regarding the study of gesture, yet indicated an interest in learning more about the topic. The participants completed the two introductory challenges as well as three main challenges in the training program (out of the 5 currently available) over the span of two weeks. Upon completion of the training program, the participants completed the aforementioned questionnaires.

The results show that the program has achieved a remarkable degree of usability, especially in terms of ease of use, complexity and integration of functionalities, reaching an overall score of 'Excellent', with individual scores ranging from 70 to 97.7 on a 1–100 scale. Specifically regarding the annotation task, we found high rates of inter-annotator reliability for the annotation of both gesture phasing and referentiality. These results suggest that people naive to the study of gestures were able to reliably annotate gestures according to M3D after completing the training program, and many participants found the program to be an easy-to-use and useful resource. In terms of acceptability, participants reported feeling that the program enhances their skills, that learning material supported the learning process and value was added through the use of visual material. The starting-and-completion ratio stands at 13:10, indicating that over 75% of the participants that began the program were able to complete it.

Conclusion

The MultiModal Multi Dimensional (M3D) labeling system offers a set of guidelines for reliably annotating gestures in multimodal speech corpora. It accounts for multiple dimensions of gesture, namely in terms of gestures' form, prosody, and meaning. M3D represents a major contribution to the field of gesture studies on multiple fronts. From a theoretical perspective, the multidimensional view of manual gesture espoused here essentially bridges the gap between two of the most prominent theoretical approaches in the field of gesture studies by moving from categorical approaches to a multidimensional approach. It combines the views from researchers such as McNeill (1992, 2006) regarding the conveyal of semantic meaning in speech and gesture with those of Kendon (1980, 2004, 2017) regarding the pragmatic meanings of gesture and their prosodic aspects.

Independently analyzing gesture form separate from its prosodic characteristics and its semantic and pragmatic meanings offers many conceptual and practical advantages. The chief conceptual advantage is that researchers no longer have to limit each gesture to a set of predefined categories. Rather, the dimensional approach allows for a more nuanced analysis of the different properties of gesture and how they relate to speech. In terms of practicality, independently assessing the different dimensions makes it easier to make labeling decisions because it allows the labeler to look at one aspect at a time instead of navigating implicitly co-occurring cues. This is particularly important considering that such cues do not always co-occur in the way that categories suggest, so by unpacking them through independent labeling is key to accuracy, agreement and eventual understanding. In terms of analysis, labeling dimensions instead of categories supports new research possibilities. For example, in the case of a palm-up, open-hand gesture, separating out a description of form (hand shape and palm orientation) from meaning and gestural prosody broadens approaches to interpreting meaning in context, similar to the way that deictic gestures have become defined as more than a specific handshape to indicate deixis (Kita, 2003).

Though the current description of the M3D system has focused on manual gestures, the tri-dimensional approach also applies to non-manual gestures. For example, in the case of the meaning dimension, head movements have been shown to convey pragmatic meaning via the marking of newer information in speech (e.g., Baills & Baumann, 2025; Esteve-Gibert et al., 2022), and can also be described in terms of their prosodic dimension (e.g., the degree of prominence of head movements). Such an approach may be suitable for the study of sign languages, where particularly the presence of head movements have been considered as markers of prosodic phrases in sign languages (Dachkovsky & Sandler, 2009; Nespor & Sandler, 1999). Non-manual cues like body, eyebrow, and head movements are produced more often (and are also more prominent) when signaling contrast than in non-contrastive speech in both signed and spoken languages (Lombart, 2021[10]; 2022), and different functions of head nods (e.g., affirmation vs. feedback) in both signed and spoken languages differ in amplitude, an aspect of the prosodic dimension (Bauer et al., 2024[11]). These findings lead us to believe that the dimensional approach may be useful for the study of bodily movements beyond manual gesture.

M3D is a flexible system, in that researchers are given multiple options to choose among for coding different aspects of language, in accordance with the individual researcher's experience level and the relevance to the research question at hand. Such an approach facilitates exploration of how a range of gestural meanings can be expressed, in a more standard and comparable way. Further, researchers are free to supplement M3D annotations with additional aspects of gesture, such as modes of representation, gesture viewpoint, or redundancy with

[10] Lombart, C. (2021, 21–23 June). *Exploring the interaction between co-speech gestures and speech prosody in the marking of contrastive focus in French: a pilot study* [Conference presentation abstract]. Phonetics and Phonology in Europe (PaPE), Barcelona, Catalonia.

[11] Bauer, A., Kuder, A., Schulder, M., & Schepens, J. (2024a, 25–27 September). *The phonetics of addressee's head nods in signed and spoken interaction using a computer vision solution* [Conference presentation]. The Second International Multimodal Communication Symposium (MMSYM), Frankfurt, Germany.

speech, as well as additional aspect of speech (e.g., discourse structure annotations, annotations of morphosyntax, etc.).

Furthermore, it is the first gesture labeling system to provide such rich resources for its application to multimodal data. M3D is highly transparent, as it offers a detailed manual with video examples, tips for complicated cases, and suggestions for workflow. Through inter-annotator reliability measures, we have shown that key parts of the labeling system can be used to provide reliable coding of multimodal corpora. The system also offers an ELAN template so coders can efficiently begin coding with M3D right away. The annotated M3D-TED corpora are openly accessible and can be used both as a tool for further independent training on how to annotate gesture, as well as a resource of scientific value. Researchers may download the corpus and enrich it with further annotations to address novel research questions and contribute to the field of multimodal communication. Finally, M3D is the first system of its kind to offer a free online training program that has been verified through standard assessments of usability, acceptability, and feasibility.

We encourage the flexible adoption of M3D and the collaboration of other labs in developing further different components of the system. So far within our team, M3D has been applied to the M3D-TED corpus for the purpose of assessing reliability, developing training materials, and for research. It would be beneficial to see how the system could be applied in other speech genres and styles, settings with multiple interlocutors, etc. Further reliability assessments could be implemented with other corpora and on aspects of M3D which have not been assessed as of yet. Not only do we encourage its application to more multimodal corpora, but the system is also open to further contributions and development. Additional aspects could still be further developed such as establishing more thorough guidelines for the annotation on non-manual gestures. By organizing periodic workshops and discussions on how the community is currently using the system and offering areas of improvement, M3D aims to be a 'living' annotation system that can adapt as research in the field progresses.

All in all, we believe that M3D represents an annotation system that is particularly well-suited to better understand how gesture holistically contributes to communication. This is because it provides independent assessments of its form, prosodic characteristics, and contribution to meaning, as well as of how these aspects interact with the other modes of communication. Importantly, using M3D will enable researchers across the domain of multimodal communication to have reliable and comparable results, making use of standard annotation practices that are readily available in open access. As such, we invite researchers in the field of multimodal communication to make use of the M3D system and its various resources, and hope that M3D can help to move the community towards a more multidisciplinary and multidimensional approach to gesture annotation, as we build on and evolve from each others' theoretical approaches and research findings.

Acknowledgements The authors would kindly like to thank the two anonymous reviewers as well as various researchers who have given us feedback at various conferences including GESPIN 2020 and ISGS 2022. We would also like to thank the ViCom DFG priority program, the Department of Translation and Language Sciences at UPF, and more specifically Renia Lopez-Ozieblo, Wim Pouw, James Trujillo, Hans

Sloetjes, Gabriella Vigliocco, Lieke van Maastricht, Marieke Hoetjes, and Anastasia Bauer for helping promote M3D in the field through organizing/funding workshops, lectures, and online promotion. We would also like to thank the students who participated in the assessment of the M3D training program. Finally, we would like to thank our funding entities, the Spanish Ministry of Science, Innovation and Universities, the Generalitat de Catalunya, and the GeHM research network.

Author Contributions All authors contributed to the theoretical conception and design of the annotation system. Data annotation was performed by PLR, UT, JFP, and IVG. Subsequent data analysis was performed by PLR. The first draft of the manuscript was written by PLR and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work was partially funded by the Spanish Ministry of Science, Innovation and Universities (Grants numbers PID2021-123823NB-I00, PID2020-115385GA-I00, PGC2018-097007-B-100), the Pla de Recerca de Catalunya from the Generalitat de Catalunya (Grant numbers 2017 SGR_971 and 2021 SGR 00922) and the GEHM (Gesture and Head Movements in Language) Research Network, funded by the Independent Research Fund Denmark (grant number 9055-00004B).

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Ambrazaitis, G. & House, D. (2023). The multimodal nature of prominence: some directions for the study of the relation between gestures and pitch accents. In *Proceedings of the 13th International Conference of Nordic Prosody* (pp. 262-273). Sciendo. https://doi.org/10.2478/9788366675728-024

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555–596. https://doi.org/10.1162/coli.07-034-R2

Baills, F., & Baumann, S. (2025). Prosody and head gestures as markers of information status in French as a native and foreign language. *Language and Cognition, 17*, e42. https://doi.org/10.1017/langcog.2025.11

Bauer, A., Kuder, A., Schulder, M., & Schepens, J. (2024). Phonetic differences between affirmative and feedback head nods in German Sign Language (DGS): A pose estimation study. *PLoS ONE, 19*(5), e0304040. https://doi.org/10.1371/journal.pone.0304040

Beckman, M. E. (1996). The parsing of prosody. *Language and Cognitive Processes, 11*(1–2), 17–68. https://doi.org/10.1080/016909696387213

Biber, D., Johansson, S., Leech, G. Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.

Boersma, P. & Weenink, D. (1992). Praat: doing phonetics by computer. https://www.praat.org.

Bolly, C. T., & Boutet, D. (2018). The multimodal CorpAGEst corpus: Keeping an eye on pragmatic competence in later life. *Corpora, 13*(3), 279–317. https://doi.org/10.3366/cor.2018.0151

Branch, R. M. (2009). Instructional design: The ADDIE approach. *Springer*. https://doi.org/10.1007/978-0-387-09506-6

Brooke, J. (2013). SUS: A retrospective. *JUX - the Journal of User Experience, 8*(2), 29–40.
Brown, L., & Prieto, P. (2021). Gesture and prosody in multimodal communication. In M. Haugh, D. Z. Kádár, & M. Terkourafi (Eds.), *The Cambridge handbook of sociopragmatics* (pp. 430–453). Cambridge University Press. https://doi.org/10.1017/9781108954105.023
Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II: Resolving the paradoxes. *Journal of Clinical Epidemiology, 43*(6), 551–558. https://doi.org/10.1016/0895-4356(90)90159-M
Cole, J., & Shattuck-Hufnagel, S. (2016). New methods for prosodic transcription: Capturing variability as a source of information. *Laboratory Phonology, 7*(1), 8. https://doi.org/10.5334/labphon.29
Cooperrider, K., Abner, N., & Goldin-Meadow, S. (2018). The palm-up puzzle: Meanings and origins of a widespread form in gesture and sign. *Frontiers in Communication*. https://doi.org/10.3389/fcomm.2018.00023
Dachkovsky, S., & Sandler, W. (2009). Visual intonation in the prosody of a sign language. *Language and Speech, 52*(2–3), 287–314. https://doi.org/10.1177/0023830909103175
Dettori, J. R., & Norvell, D. C. (2020). Kappa and beyond: Is there agreement? *Global Spine Journal, 10*(4), 499–501. https://doi.org/10.1177/2192568220911648
Du Bois, J. W. (2007). The stance triangle. In R. Englebretson (Ed.), *Stancetaking in discourse: Subjectivity, evaluation, interaction* (pp. 139–182). John Benjamins Publishing Company. https://doi.org/10.1075/pbns.164.07du
Ebert, C., Evert, S., & Wilmes, K. (2011). Focus marking via gestures. In I. Reich, E. Horch & P. (Eds.), In *Proceedings of Sinn & Bedeutung 15*, (pp. 193–208). Universität des Saarlandes. https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/372/305
Esteve-Gibert, N., Lœvenbruck, H., Dohen, M., & D'Imperio, M. (2022). Pre-schoolers use head gestures rather than prosodic cues to highlight important information in speech. *Developmental Science*. https://doi.org/10.1111/desc.13154
Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low Kappa: I: the problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549. https://doi.org/10.1016/0895-4356(90)90158-L
Ferré, G. (2011). Functions of three open-palm hand gestures. *Multimodal Communication, 1*(1), 5–12. https://doi.org/10.1515/mc-2012-0002
Ferré, G. (2014). A multimodal approach to markedness in spoken French. *Speech Communication, 57*, 268–282. https://doi.org/10.1016/j.specom.2013.06.002
Flecha-García, M. L. (2010). Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English. *Speech Communication, 52*(6), 542–554. https://doi.org/10.1016/j.specom.2009.12.003
Florit-Pons, J., Vilà-Giménez, I., Rohrer, P. L., & Prieto, P. (2023). Multimodal development in children's narrative speech: Evidence for tight gesture–speech temporal alignment patterns as early as 5 years old. *Journal of Speech, Language, and Hearing Research, 66*(3), 888–900. https://doi.org/10.1044/2022_JSLHR-22-00451
Freeman, V. (2015). *The Phonetics of Stance-taking* [unpublished doctoral dissertation]. University of Washington.
Grice, M., & Kügler, F. (2021). Prosodic prominence: A cross-linguistic perspective. *Language and Speech, 64*(2), 253–260. https://doi.org/10.1177/00238309211015768
Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics, 12*(3), 175–204.
Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61*(1), 29–48. https://doi.org/10.1348/000711006X126600
Holle, H., & Rein, R. (2015). EasyDIAg: A tool for easy determination of interrater agreement. *Behavior Research Methods, 47*(3), 837–847. https://doi.org/10.3758/s13428-014-0506-7
Kendon, A. (1972). Some relationships between body motion and speech. In A. W. Siegman & B. Pope (Eds.), *Studies in dyadic communication* (pp. 177–210). Pergamon. https://doi.org/10.1016/B978-0-08-015867-9.50013-7
Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key (Ed.), *The relationship of verbal and nonverbal communication* (pp. 207–228). De Gruyter Mouton. https://doi.org/10.1515/9783110813098.207
Kendon, A. (1995). Gestures as illocutionary and discourse structure markers in Southern Italian conversation. *Journal of Pragmatics, 23*(3), 247–279. https://doi.org/10.1016/0378-2166(94)00037-F

Kendon, A. (2004). Gesture: Visible action as utterance. *Cambridge University Press*. https://doi.org/10.1017/CBO9780511807572
Kendon, A. (2017). Pragmatic functions of gestures: Some observations on the history of their study and their nature. *Gesturez, 16*(2), 157–175. https://doi.org/10.1075/gest.16.2.01ken
Kita, S. (2003). *Pointing: Where language, culture, and cognition meet*. Lawrence Erlbaum Associates.
Kita, S., van Gijn, I., & van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.), *Gesture and sign language in human-computer interaction* (pp. 23–35). Springer. https://doi.org/10.1007/BFb0052986
Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica, 55*(3–4), 243–276. https://doi.org/10.1556/ALing.55.2008.3-4.2
Krivokapić, J., Tiede, M. K., & Tyrone, M. E. (2017). A kinematic study of prosodic structure in articulatory and manual gestures: Results from a novel method of data collection. *Laboratory Phonology, 8*(1), 1–26. https://doi.org/10.5334/labphon.75
Ladewig, S. H. (2014). Recurrent gestures. In C. Müller, A. Cienki, E. Fricke, S. Ladewig, D. McNeill & J. Bressem (Eds.), *Handbücher zur sprach- und kommunikationswissenschaft/Handbooks of linguistics and communication Science (HSK) 38/2*, (pp. 1558–1574). De Gruyter. https://doi.org/10.1515/9783110302028.1558
Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2015.00731
Loehr, D. P. (2004). *Gesture and Intonation* [unpublished doctoral dissertation]. Georgetown University.
Lombart, C. (2022). Prosodic marking of contrast in LSFB (French Belgian Sign Language): An investigation of manual cues and their relations to prominence. *Belgian Journal of Linguistics, 36*(1), 108–144. https://doi.org/10.1075/bjl.00073.lom
Lopez-Ozieblo, R. (2020). Proposing a revised functional classification of pragmatic gestures. *Lingua, 247*, Article 102870. https://doi.org/10.1016/j.lingua.2020.102870
McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.
McNeill, D. (2006). Gesture and communication. In K. Brown (Ed.), *Encyclopedia of language & linguistics* (pp. 58–66). Elsevier. https://doi.org/10.1016/B0-08-044854-2/00798-7
Nespor, M., & Sandler, W. (1999). Prosody in Israeli Sign Language. *Language and Speech, 42*(2–3), 143–176. https://doi.org/10.1177/00238309990420020201
Parrell, B., Goldstein, L., Lee, S., & Byrd, D. (2014). Spatiotemporal coupling between speech and manual motor actions. *Journal of Phonetics, 42*, 1–11. https://doi.org/10.1016/j.wocn.2013.11.002
Parrill, F. (2008). Form, meaning, and convention: A comparison of a metaphoric gesture with an emblem. In A. Cienki & C. Müller (Eds.), *Metaphor and gesture* (pp. 195–217). John Benjamins Publishing Company. https://doi.org/10.1075/gs.3.11par
Passonneau, R. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk & D. Tapias (Eds.), In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC '06)* (pp. 831–836). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2006/pdf/636_pdf.pdf
Payrató, L. (1993). A pragmatic view on autonomous gestures: A first repertoire of Catalan emblems. *Journal of Pragmatics, 20*(3), 193–216. https://doi.org/10.1016/0378-2166(93)90046-R
Perniss, P. (2018). Why we should study multimodal language. *Frontiers in Psychology, 9*, 1109. https://doi.org/10.3389/fpsyg.2018.01109
Prieto, P., Cravotta, A., Kushch, O., Rohrer, P., & Vilà-Giménez, I. (2018). Deconstructing beat gestures: A labelling proposal. In *Speech Prosody 2018* (pp. 201–205). ISCA. https://doi.org/10.21437/SpeechProsody.2018-41
Prieto, P., Esteve-Gibert, N., & Shattuck-Hufnagel, S. (2025). Towards a novel conceptualization of prosody that accounts for spoken and gestural signals: the modality-neutral prosodic framework hypothesis. OSF preprint. https://osf.io/preprints/osf/npt49_v1
Prieto, P., & Espinal, M. T. (2020). Negation, prosody, and gesture. In V. Déprez & M. T. Espinal (Eds.), *The Oxford handbook of negation* (pp. 677–693). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198830528.013.34
R Core Team. (2023). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/
Ren-Mitchell, A. & Shattuck-Hufnagel, S. (2012). *SCG gesture coding manual*. https://scg.mit.edu/gesture/coding-manual.html. (accessed 14 February, 2024).

Rohrer, P. L., Prieto, P., & Delais-Roussarie, E. (2019). Beat gestures and prosodic domain marking in French. In S. Calhoun, P. Escudero, M. Tabain & P. Warren (Eds.), In *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 1500–1504). Australasian Speech Science and Technology Association Inc.

Rohrer, P. L. (2022). *A Temporal and Pragmatic Analysis of Gesture-Speech Association: A Corpus-based Approach Using the Novel MultiModal MultiDimensional (M3D) Labeling System* [Doctoral dissertation, Universitat Pompeu Fabra & Nantes Université]. Tesis Doctorals en Xarxa. http://hdl.handle.net/10803/687534

Rohrer, P. L., Delais-Roussarie, E., & Prieto, P. (2023). Visualizing prosodic structure: Manual gestures as highlighters of prosodic heads and edges in English academic discourses. *Lingua, 293*, 103583. https://doi.org/10.1016/j.lingua.2023.103583

Rohrer, P. L., Florit-Pons, J., Vilà-Giménez, I., & Prieto, P. (2022). Children use non-referential gestures in narrative speech to mark discourse elements which update common ground. *Frontiers in Psychology, 12*, 661339. https://doi.org/10.3389/fpsyg.2021.661339

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4), 696–735. https://doi.org/10.2307/412243

Sandler, W. (2018). The body as evidence for the nature of language. *Frontiers in Psychology, 9*, 1782. https://doi.org/10.3389/fpsyg.2018.01782

Searle, J. R. (1975). A taxonomy of illocutionary acts. *Minnesota Studies in the Philosophy of Science, 7*, 344–369.

Shattuck-Hufnagel, S., Ren, P. L., & Tauscher, E. (2010). Are torso movements during speech timed with intonational phrases? In *Proceedings of Speech Prosody 2010*. https://doi.org/10.21437/SpeechProsody.2010-133

Shattuck-Hufnagel, S. & Prieto, P. (2019). Dimensionalizing co-speech gestures. In S. Calhoun, P. Escudero, M. Tabain & Paul Warren (Eds.). In *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 1490–1494). Australasian Speech Science and Technology Association Inc.

Shattuck-Hufnagel, S., & Ren, A. (2018). The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech. *Frontiers in Psychology, 9*, 1514. https://doi.org/10.3389/fpsyg.2018.01514

Swerts, M., & Krahmer, E. (2010). Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions. *Journal of Phonetics, 38*(2), 197–206. https://doi.org/10.1016/j.wocn.2009.10.002

Tütüncübasi, U. (2023). *M3D Gesture Labeling: Developing and Evaluating a Training Program for Gesture Annotation - A Mixed-methods Study* [unpublished Master's thesis]. Universitat Pompeu Fabra.

Vigliocco, G., Perniss, P., & Vinson, D. (2014). Language as a multimodal phenomenon: Implications for language learning, processing and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1651), 20130292. https://doi.org/10.1098/rstb.2013.0292

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk & D. Tapias (Eds.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)* (pp. 1556–1559). European Language Resources Association (ELRA). https://hdl.handle.net/11858/00-001M-0000-0013-1E7E-4

Yap, D. & Casasanto, D. (2018). Beat gestures encode spatial semantics. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 1211.