Universitat Pompeu Fabra Barcelona

Treball de fi de màster *de Recerca*

M3D Gesture Labeling: Developing and evaluating a training program for gesture annotation – a mixed-methods study

Nom i Cognoms: *Ulya Tütüncübasi*

Màster: *Theoretical and Applied Linguistics*

Edició: 2022-2023

Directors: *Prof. Pilar Prieto, Dr. Patrick Louis Rohrer*

Any de Defensa*:* 2023

Col.lecció: Treballs de fi de màster

Departament de Traducció i Ciències del Llenguatge

Acknowledgments

First, I would like to express my deep gratitude to my supervisors, Pilar Prieto and Patrick Louis Rohrer for their guidance, support and encouragement throughout my master's thesis, but also before. Patrick, your enthusiasm and positive energy have been truly inspiring. Since 2021, it has been an absolute pleasure working with you. Not only your gestures are iconic, you are iconic, too! Pilar, I would like to thank you for your endless support, both academically and personally. From the moment I arrived at UPF, you welcomed me with open arms. I have always looked forward to our Monday meetings!

Furthermore, I would like to thank Stefanie Shattuck-Hufnagel and Ada Ren-Mitchell. I really appreciate your insightful comments, your patience and your willingness to help in the project. Patrick, Pilar, Stefanie and Ada - I really enjoyed the process of developing the *M3D Training Program* with you! Ada, thank you so much for investing your time in helping us with the web design of the training program, I really valued our Friday meetings!

I would also like to thank Núria Esteve-Gibert, Ingrid Vilà-Giménez and Júlia Florit Pons for serving as speakers for the videos, donating their time to the project and Nil Bernat Gil, Marcos Taboada Lopez and Justí Torn Poch for their essential contribution to the project with editing the videos. Alina Gregori and Clara Lombart, thank you for your valuable feedback on the training program. Your input has been incredibly helpful!

Above all, I would like to thank my parents who always stood by my side through every moment of life. I would like to thank you, Mama, for showing me how to find joy in the little things, and how to remain true to yourself and to always be independent. I would also like to thank you for always offering your help. Ačiu. Baba, I would like to thank you for your words of encouragement and wisdom, for supporting me and trusting in my decisions and for always telling me how proud you are of me. And of course, thank you for all the countless hours you spent in the kitchen preparing delicious meals for me! Teşekkür ederim.

Abstract

Various gesture labeling systems aim to establish guidelines on how to systematically assess and annotate gestures. However, there is a lack of resources that provide guidance and training on the proper use of the labeling systems. To address this gap, the present study aims to (a) collaboratively develop an online training program for the M3D gesture labeling system (Rohrer et al., 2023) using the ADDIE paradigm (Branch, 2009) and (b) to evaluate its usability, acceptability, feasibility and effectiveness. Developing a training program for gesture annotation is essential for advancing the field of gesture studies, equally important is conducting an evaluative process to ensure the program's high quality nature.

The evaluation will employ a mixed-methods approach, using the System Usability Scale (SUS) (Brooke, 2013) for usability assessment, an ADDIE questionnaire and an in-house questionnaire for acceptability, a starting-and-completion ratio for feasibility and an inter-annotator reliability assessment for effectiveness. Thirteen participants took part in the evaluation. 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 70-97.7 on a 1-100 scale. In terms of acceptability, it was found to enhance the participants' skills (M= 4.90, SD= 0.316), learning material was perceived as supporting the learning process (M= 4.90, SD= 0.316) and value was added by visual material (M= 5.00, SD= 0.00). The starting-and-completion ratio stands at 13:10. Effectiveness shows high inter-annotator reliability results for Gesture Phases and Gesture Referentiality. Overall, the projects' results represent a major advancement for gesture studies, offering for the first time an online gesture annotation training program which moves the field towards a more standardized approach to assessing gesture in multimodal corpora.

Keywords: MultiModal MultiDimensional (M3D) Labeling System; Training Program; Evaluation; Usability; Acceptability; Feasibility; Effectiveness; System Usability Scale; Reliability

Table of Contents

1. Introduction

1.1 Multimodality in language studies

Human language is multimodal in nature. In face-to-face communication, speakers convey meaning not only through a verbal mode, but also through bodily movement and spoken prosody so that the interlocutor can interpret this meaning accurately and effortlessly. At the beginning of the 20th century, linguists mainly focused on spoken and written language when studying human language, leaving gestures such as facial expressions or head movements out of the scope of linguistics (Perniss, 2018). Today, however, gestures have become an essential part of modern linguistics. Prominent researchers in the field of gestures, like David McNeill and Adam Kendon, have been key proponents of the full incorporation of gestures as a part of language. Their theoretical proposals are nowadays considered to be the foundation of gesture studies and are among the most widely accepted concepts in gesture research today (see Kendon, 2004; McNeill, 1992). McNeill adopts a semantic-oriented viewpoint on gestures, considering them a window onto thought while Kendon adopts a pragmatic approach, stating that gestures can function as utterances (as well as convey semantic meaning in speech). This underpins the following definition of gesture, which is adopted in this current work: " [Gesture is] a visible action of any body part, when it is used as an utterance, or as part of an utterance" (Kendon, 2004, p. 7).

1.2 Gesture annotation systems for labeling multimodal corpora and the importance of training material

Nowadays, research on multimodal language has gained importance not only in linguistics, but also in psychology, cognition, neuroscience, biology and computational science. As the interest in studying gestures is on the rise, various disciplines and individual labs have

developed their own gesture labeling systems to assess and study gestures in multimodal corpora. Examples of such systems include CorpAGEst (Bolly, 2016), LASG (Bressem et al., 2013), NEUROGES® (Lausberg and Sloetjes, 2009) and MUMIN (Allwood et al., 2007) (see Rohrer, 2022 for thorough review). Each labeling system brings forth their own approaches to gesture classification and annotation, which vary widely between them. Importantly, none adopt the most recent and widely accepted approaches in the field today.

One recently developed gesture labeling system bridges the approaches set forth by McNeill (1992) and Kendon (2004), integrating multiple viewpoints into a widely adopted labeling system. The Multimodal MultiDimensional (M3D) labeling system (Rohrer et al., 2023) has been developed in collaboration between the Prosodic and Gestural Studies Group at the Department of Translation and Language Sciences (Universitat Pompeu Fabra, Barcelona), the Speech Communication Group at the Research Laboratory of Electronics (Massachusetts Institute of Technology, Boston, Massachusetts) and the Faculty of Psychology and Education Sciences (Universitat Oberta de Catalunya, Barcelona).

M3D aims to be flexible enough to encompass all aspects of gesture and to adapt to a variety of research questions, while making use of a systematized annotation procedure by proposing a set of guidelines. This approach allows researchers to benefit from comparable annotations across different studies and labs. As a single gesture can be assessed in terms of its form (the form dimension), its raw organizational structure (the prosodic dimension), and the referential and/or pragmatic meaning it conveys (the meaning dimension), M3D offers a holistic perspective on gestures by considering all gestural dimensions in a non-mutually exclusive manner. The idea of non-mutual exclusivity is particularly relevant for annotating the gestural meaning, as the meanings can overlap. Researchers should not be forced to choose between one meaning over another, but rather be able to annotate multiple semantic and pragmatic meanings for one single gesture (Rohrer et al., 2023).

The multidimensional approach of M3D offers even more advantages. There is no gestural data that falls under the radar because of having to choose one category over another. Furthermore, researchers are able to assess the different "modes" of communication independently, e.g. to annotate speech prosody separately from gesture and therefore avoid bias from the different modes of communication. Apart from that, researchers may annotate only the aspects of gesture that are relevant to their specific research objectives. They also have the option to add tiers to annotate additional aspects as needed. Finally, as a result of its systematic annotation procedure and its set of annotation guidelines, it shows high levels of inter-annotator reliability (see Rohrer, 2022). Due to these novel approaches and benefits, M3D has generated much interest in the field of gesture studies.

However, despite the increasing interest in gesture annotation, there is a lack of available training material for gesture labeling systems. Research suggests that providing training material is highly effective, as actively engaging in a learning process facilitates the development of skills and the transfer of knowledge to real-life situations (Lombardi and Oblinger, 2007). It provides opportunities for learners to actively experience and apply learned knowledge, encouraging them to explore, experiment and build their understanding of the topic (see *Constructivism*, Kolb et al., 2014; Vygotsky and Cole, 2018).

Another advantage of creating training material is the fact that all learners receive the same information, thus ensuring consistency and reducing variations in knowledge and learning outcomes (Sánchez-Cabrero et al., 2021). For teaching the annotation of gestures, this can be particularly useful as one of the main criteria for having valid annotations is to have a high inter-annotator reliability, meaning that the annotations coming from different annotators consistently align well, and agree in the labels they have assigned. Having unified gesture training material helps to prevent researchers from adopting a combination of the approaches of multiple systems and implementing them without following much of a

standard protocol. Moreover, unified gesture training material ensures that the annotations are done in a way that reflects accurate theoretical and methodological approaches that are widely adopted in the field of gesture studies. Besides that, training material provides learners with clear instructions and real-life examples, helping them to accurately identify and label different aspects of gestures, resulting in a faster and more streamlined annotation workflow. Creating and openly sharing training material for a particular gesture labeling system makes it possible for others to apply the system to their own research questions or to replicate research results. Generally speaking, such resources play a crucial role in guiding potential labellers on how to use the system. Having openly accessible training material for gesture studies would lead to a major advancement for the whole field of gesture studies.

To our knowledge, the NEUROpsychological GESture coding system (NEUROGES®) (Lausberg and Sloetjes, 2009) has been found to be the only system that offers training. The training is conducted once a year in Cologne, Germany. Optionally, learners may request a special training program in their own city. Alternatively, NEUROGES® offers self-study training, with additional lecture material and tutorial videos including a final exam and feedback. In all cases, learners become a certified NEUROGES® annotator after successfully completing the training (Lausberg and Sloetjes, 2009; Lausberg and Sloetjes, 2016). Yet, all options, even the self-study training, are associated with a registration fee. Even though the NEUROGES® seminars offer a valuable opportunity for acquiring gesture annotation skills, apart from these existing fee-based training seminars, training material that is openly available cannot be found.

In our view, providing open-access gesture labeling training material in an online format would democratize learning, giving everyone the opportunity to acquire knowledge or skills regardless of their geographical location, schedule, and financial resources.

In other fields of linguistics, training material has already been available for many years. The training material for Cat_ToBI and Sp_ToBI teaches researchers the annotation of Catalan and Spanish prosody using a set of transcription conventions adapted from the ToBI (Tones and Breaks Indices) annotation system (Silverman et al., 1992). It consists of openly-accessible labeling examples and exercises (Aguilar et al., 2009). Furthermore. "Transcribing Prosodic Structure Of Spoken Utterances With ToBI" is another existing online course offered by the Massachusetts Institute of Technology (MIT)[1]. It was developed by Alejna Brugos, Stefanie Shattuck-Hufnagel and Nanette Veilleux in 2006 and provides examples and exercises for Mainstream American English (MAE_ToBI).

Taking into consideration all the aforementioned arguments, the first goal of the present MA thesis will be to actively participate in the development of the training material for the M3D gesture labeling system. Despite the fact that the M3D labeling manual (Rohrer et al., 2023) is openly-accessible online and offers extensive explanations of each dimension together with screenshots of examples, video examples, and annotation tips for the different dimensions, a training program (i.e., training material organized in a coherent, structured series of tasks with specific learning objectives) was not yet available. Thus, having an interactive and engaging training program as an additional resource on top of the already existing manual would help learners to actively engage in the learning process rather than just reading a manual. For the aforementioned reasons, the motivation for jointly developing an online gesture training program is to provide learners with an openly accessible resource for gesture annotation that promotes equal access and therefore democratizes learning and benefits various research fields.

Choosing M3D out of the available gesture labeling systems for putting this into practice assures teaching a comprehensive and flexible system that is in line with widely

[1] https://ocw.mit.edu/courses/6-911-transcribing-prosodic-structure-of-spoken-utterances-with-tobi-january-iap-2006/

accepted theoretical approaches to gestures, enabling researchers to apply it to answer a variety of research questions.

1.3 The evaluation of training material

While the development of an online training program for gesture annotation is essential for advancing the field of gesture studies, it is equally important that this program undergoes an evaluative process to demonstrate its high quality nature. Conducting an evaluation has become a standard practice due to the increasing dependence on digital learning material (Alenezi, 2020).

Many popular e-learning platforms undergo an evaluation procedure. For example, the language-learning platform *Duolingo* provides interactive lessons and gamified exercises which has been evaluated for its impact on learners' language proficiency (Vesselinov and Grego, 2012; Ye, 2014). *Khan Academy*, an online educational platform that offers free video lessons and interactive exercises on a wide range of subjects has undergone thorough evaluations to assess its impact on improving student learning outcomes (Murphy et al., 2014; Phillips and Cohen, 2015). Finally, *EdX* is a popular MOOC platform (Massive Online Open Course) that offers a wide range of online courses from universities. Studies have evaluated the courses in terms of student engagement and satisfaction (Öztürk and Gürler, 2020).

Undergoing an evaluative process increases not only the validity and transparency of the training program but it gives learners the confidence that they are participating in a program that has its place in the field.

Training programs are found to be evaluated in four main aspects, namely usability, acceptability, feasibility and effectiveness (e.g., Carter et al., 2018; Horne et al., 2022; Jackman et al., 2019). “Usability” is understood as 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” refers to how learners react to the program and the extent to which they are inclined to use it (Bowen et al., 2009; Cavallo et al., 2015; Eicher et al., 2019; Holthe et al., 2018). Assessing acceptability and usability of a training program would consequently help to analyze if it meets the learners’ needs and if they achieve their learning goals through the program. “Feasibility” is defined as the extent to which a program can be successfully used or carried out. It involves assessing the willingness of learners to use the learning resource and assessing if the implementation is viable (Guadana, 2018; Weiner et al., 2017). Finally, “effectiveness” refers to the measurement of the impact of a training program, taking into account the ability of learners to complete tasks within the program and the quality of the tasks output (Brooke, 2013; Glasgow et al., 2019). It involves investigating the reasons behind the success of the output or lack thereof (Glasgow et al., 2019).

Learners are motivated to engage in a training program when it demonstrates good usability and feasibility and receives positive learner acceptance, resulting in a good learning outcome (Holthe et al., 2018). Consequently, the second goal of the present work will be to evaluate the online *M3D Training Program* by assessing its usability, acceptability, feasibility and effectiveness.

1.4 Objectives

The goal of the present work is twofold, namely (a) to jointly develop an M3D online training program (*M3D Training Program*) together with other four members of the M3D team and (b) to evaluate the *M3D Training Program* in terms of usability, acceptability, feasibility and effectiveness.

The *M3D Training Program* will be developed using the ADDIE paradigm (Branch, 2009; see Chapter 2 of this thesis). In order to subsequently evaluate the *M3D Training Program*, a mixed-methods approach will be used (see Chapter 3 of this thesis*)*. Firstly, the evaluation will involve a quantitative post-training questionnaire using the System Usability Scale (SUS, Brooke, 2013) to assess usability. Secondly, a quantitative questionnaire provided by ADDIE (Branch, 2009) will be used to assess acceptability. By incorporating not only the theoretical approach of ADDIE in the development of the training program but also a part of its evaluation method, we will ensure the implementation of the ADDIE paradigm to be more comprehensive and well-rounded. As an additional source of feedback, a third questionnaire will be developed to assess acceptability specifically about the *M3D Training Program*. Furthermore, feasibility will be assessed by examining the starting-and-completion ratio of the training program among the participants (see Cisel, 2014).

Lastly, a quantitative post-training inter-annotator reliability assessment conducted on annotated audiovisual data will be used to gauge effectiveness. Calculating inter-annotator reliability, also known as inter-annotator agreement, is a widely used technique after linguistic annotation. It involves comparing the annotations of a single source (such as text, audio or video) made by multiple annotators. This technique is important for assessing the validity and replicability of annotations, which in turn can be used to evaluate proposed labeling schemes or identify potential ambiguities in the understanding of annotation guidelines and procedures. In the case of the training program, assessing reliability would provide insights about the effectiveness of the training as it demonstrates the quality of the learning outcome.

2. Goal I: Developing the online *M3D Training Program*

In a joint collaboration with **Pilar Prieto** (Prosodic and Gestural Studies Group, Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona), **Patrick Louis Rohrer** (SPEAC lab, Donders Institute for Brain, Cognition, and Behavior, Radboud University, Nijmegen), **Stefanie Shattuck-Hufnagel** and **Ada Ren-Mitchell** (Speech Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology) and myself, **Ulya Tütüncübasi**, the online *M3D Training Program* was developed between May 2022 and May 2023. Blending the knowledge and skills of five researchers that are experts in gesture studies ensured having a holistic and comprehensive view on the topic. Moreover, the training program also profited from expertise in user experience and graphic web design through the skills of Ada Ren-Mitchell.

The program is based on a three week in-person workshop on M3D labeling led by Patrick Louis Rohrer that took place in January 2023 at Universitat Pompeu Fabra. It is hosted as an openly-accessible, permanent website with a university internal domain (https://m3d.upf.edu/).

The development of the *M3D Training Program* followed the ADDIE paradigm (Branch, 2009; see also: Kristanto, 2021; Mayfield, 2011; Supriyatno et al., 2020; for examples of its application and Spatioti et al., 2022, for a comparative study). ADDIE is a widely accepted framework used for creating learning resources, emphasizing a student-centered and goal-oriented approach. It utilizes a method known as intentional learning which refers to the purposeful engagement in the learning process and involves learners being actively involved in constructing knowledge (Branch, 2009). It moves away from traditional, passive educational design methods and promotes a dynamic learning environment.

Table 1 summarizes the concepts and procedures as laid out by the ADDIE paradigm. The term "performance gap" used in **Table 1** refers to the discrepancy between the current performance and the desired performance in a specific field. This discrepancy can arise from a lack of resources, a lack of knowledge and skills or a lack of motivation.

	Analyze	Design	Develop	Implement	Evaluate
Concept	Identify the probable causes for a performance gap	Verify the desired performances and appropriate testing methods	Generate and validate the learning resources	Prepare the learning environment and engage the students	Assess the quality of the instructional products and processes, both before and after implementation
Common Procedures	1. Validate the performance gap 2. Determine instructional goals 3. Confirm the intended audience 4. Identify required resources 5. Determine potential delivery systems (including cost estimate) 6. Compose a project management plan	7. Conduct a task inventory 8. Compose performance objectives 9. Generate testing strategies 10. Calculate return on investment	11. Generate content 12. Select or develop supporting media 13. Develop guidance for the student 14. Develop guidance for the teacher 15. Conduct formative revisions 16. Conduct a Pilot Test	17. Prepare the teacher 18. Prepare the student	19. Determine evaluation criteria 20. Select evaluation tools 21. Conduct evaluations
	Analysis Summary	*Design Brief*	*Learning Resources*	*Implementation Strategy*	*Evaluation Plan*

Table 1: Procedure steps identified by ADDIE (Branch, 2009 p.3 Figure 2)

The implementation of ADDIE and the structure of the *M3D Training Program*

The following section of this thesis will start by outlining the implementation of the ADDIE paradigm at each developmental stage of the *M3D Training Program*. It will then proceed to provide an overview of the structure, both within each individual section of the training program and at a holistic level.

Step 1: Analyze

As the performance gap, we recognized the lack of knowledge and skills in gesture annotation as a result of the lack of learning resources. Despite the existence of numerous gesture labeling systems, in most of the cases only written guidelines are provided. The lack of available training material and annotated examples poses a challenge for individuals seeking to learn and apply these labeling systems (see Rohrer, 2022 for thorough review). As for the main goal of the learning resource, we identified it being the acquisition of skills and knowledge for the annotation of audiovisual corpora using the M3D labeling system, taking all dimensions into account, namely the prosodic, the form and the meaning dimensions (Rohrer et al., 2023).

Step 2: Design

As ADDIE suggests, our chosen training strategy was student-centered and was built to encompass both the theoretical aspects as well as the applied aspects of annotating gestures. This training strategy was then implemented in an interactive online learning platform, utilizing an approach of proximal development, where learners are constructing knowledge systematically in a stepwise fashion (Fussell and Bianco, 2021). This means, learners are guided through their learning process, progressively developing their understanding of gestures and gesture annotations. They begin by acquiring knowledge in the theoretical aspects and gradually become familiar with annotating through interactive exercises, ultimately reaching a point where they are asked to perform annotations independently.

When initially accessing the *M3D Training Program*, learners are greeted with a landing page that offers a brief description and overview of the program's properties (see **Figure 1**). As they scroll down, they are presented with a comprehensive overview of the main sections ("challenges"), including their educational goals. Moreover, a clickable button

is provided for each challenge, enabling learners to directly access the specific challenge they wish to explore.

The term “challenge” was chosen as it emphasizes the dynamic nature of each dimensional aspect of gesture. Moreover, the term implies that each learning step within the dimension may present unique difficulties or obstacles that need to be overcome. Also, it helps to convey the idea that each step needs its own specific knowledge and strategies for being successful.

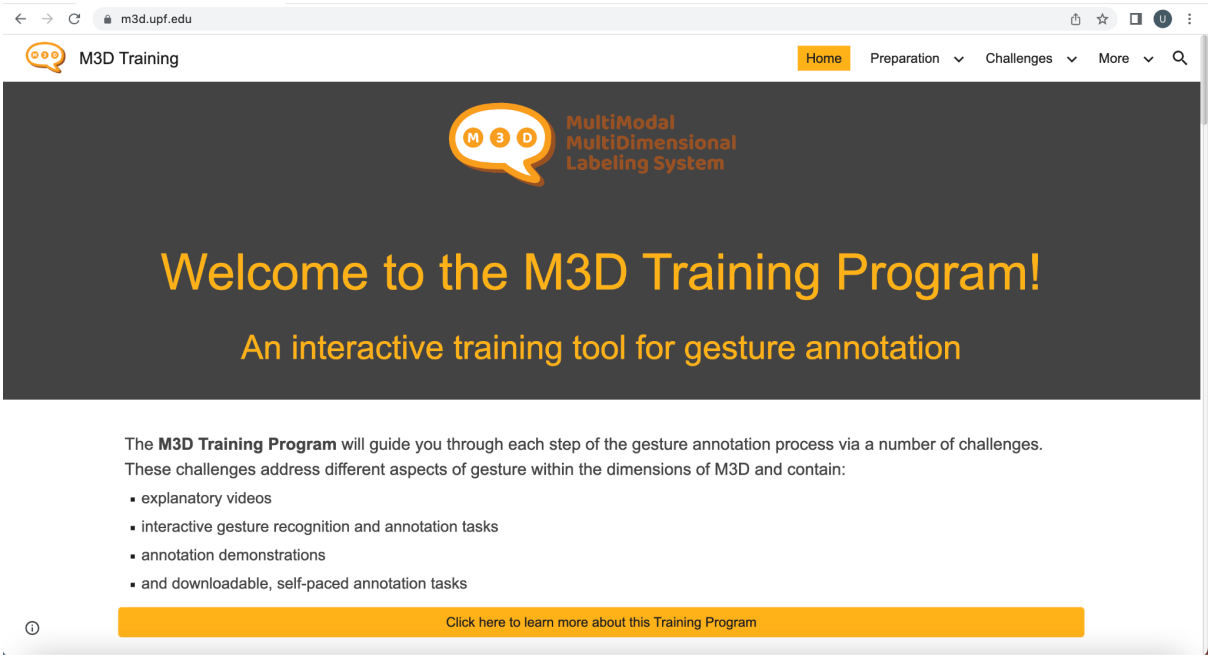

Figure 1: Screenshot of the landing page of the *M3D Training Program*

Following ADDIE, we implemented repetition when it was useful (Branch, 2009, p. 11). Therefore, the structure within the challenges always follows the same logical sequence with smooth transitions from one task to another. Each challenge begins with an explanatory video, followed by two multiple-choice tasks, an annotation tutorial and ends with an autonomous annotation task. The task inventory of each of the three tasks was thoroughly discussed in the design phase of the program creation and is presented in detail in **Table 2**, along with the other features within each challenge.

Type	Description	Properties
Explanatory video	Learners are asked to watch a video about the theoretical underpinnings and specific approaches of each dimensional aspect.	• Live-action video (a speaker explaining key concepts), speaker demonstrating gestures with own hands, example videos of gestures
Task 1: Recognize the [respective aspect] (e.g. Gesture Units)	Learners are asked to apply the theoretical knowledge they learned about in the explanatory video to identify various aspects of gesture in a series of short videos. This task is indicated by the blue M3D logo in all challenges.	• multiple-choice questions • progressive increase in difficulty • self-paced • possibility to repeat task as often as preferred • immediate feedback provided to guide self-correction
Task 2: Select the best [respective aspect] annotation (e.g. Gesture Units)	Learners are asked to examine three different potential annotations and try to match the most suitable annotation with the video. The goal here is to practice assessing actual annotations in ELAN. This task is indicated by the purple M3D logo.	• multiple-choice questions • progressive increase in difficulty • self-paced • possibility to repeat task as often as preferred • immediate feedback provided to guide self-correction
Annotation tutorial "Let's annotate together!"	Learners are asked to watch the tutorial to become familiar with how the actual annotation procedure looks like in ELAN. The annotation procedure is shown together with commentary, highlighting the key considerations.	• Screen Recording video showing live annotation demonstration (by Patrick Louis Rohrer)
Task 3: "Now it's your turn!"	Learners are provided with a video sequence and are asked to independently annotate the respective aspect. Once the annotations are completed, learners have the possibility to compare their annotations against a sample solution.	• mp4-file of the video • wav-file of the video • M3D template • eaf-file containing a sample solution

Table 2: Structure of each individual challenge (color intensity of the column indicates that knowledge was accumulated gradually)

In accordance with the ADDIE paradigm, we implemented the *advance organizer* (Branch, 2009, p.112) in the beginning of each challenge so that learners find key points outlining the learning objectives and an overview of what to expect in the respective challenge. In the end of each challenge (as well as in the preparatory sections), learners find a list of helpful

resources that were either mentioned in the explanatory video or that simply provide learners with further information.

This screen-recorded video demonstrates the practical implementation of the *advance organizer*, the incorporation of the features presented in **Table 2**, as well as the list of resources by using challenge: Identifying Gesture Units as an example. It also shows all the interactive elements within each challenge.

Step 3: Development

In line with ADDIE, our initial objective was to ensure that all learners started the training program with a basic level of knowledge regarding the theoretical framework of M3D and the technical aspects of ELAN (Wittenburg et al., 2006). To achieve this, learners have the possibility to access the preparatory sections of the training program before entering into the main tab.

The main tab of the training is organized into 6 challenges, which represent the key concepts within the different gesture dimensions set out in M3D, namely Identifying Gesture Units, Identifying Gesture Form, Identifying Gesture Phases, Identifying Rhythmic Properties - Beat-like-ness, Identifying Gesture Referentiality and Identifying Gesture Pragmatics. Within the present work, the focus has been placed on three main challenges, namely Identifying Gesture Units, Identifying Gesture Phases and Identifying Gesture Referentiality, as these are key aspects in terms of identifying and generally describing gestures.

Table 3 presents an overview of the challenge titles, together with their corresponding educational goals and the method employed to achieve them. The challenges are organized in the order of the proposed annotation workflow (i.e. the order in which M3D suggests labellers to annotate). Nonetheless, if learners are only interested in one specific aspect of gesture, they are encouraged to directly jump into the corresponding challenge. This is why

the challenges are organized in a way that they can be generally understood without having to complete the one that comes before.

Challenge title (click on hyperlink to access corresponding challenge on the website)	**Educational goals**	**Methods**
Preparation: Learning about M3D	• understand the main approaches of M3D and the gestural dimensions it consists of	Explanatory video (speaker: Pilar Prieto)
Preparation: Getting started with ELAN	• understand how to prepare the workstation in the annotation software ELAN	Tutorial video (recorded by Patrick Louis Rohrer)
Challenge: Identifying Gesture Units	• realize that Gesture Units are the first aspect of the prosodic dimension • realize that gestures are organized into larger units • identify the starting and ending points of a Gesture Unit • learn how to annotate these units in ELAN	1) Explanatory video (speaker: Patrick Louis Rohrer) 2) Task 1: Recognize the Gesture Units 3) Task 2: Select the best Gesture Unit annotation, 4) Annotation tutorial "Let's annotate together" 5) Task 3: Now it's your turn
Challenge: Identifying Gesture Form	• explore the four different aspects of the form dimension, namely hand shape, palm orientation, trajectory shape, and trajectory direction • become familiar with the various labels within each aspect of form • learn how to annotate form in ELAN	1) Explanatory video (speaker: Ada Ren-Mitchell) *2) In development* *3) In development* 4) Annotation tutorial "Let's annotate together" 5) Task 3: Now it's your turn
Challenge: Identifying Gesture Phases	• realize that Gesture Phases are the second aspect of the prosodic dimension • identify the different phases of a gestural movement • discover that the stroke phase carries the overall gestural meaning • explore the diverse phrasal structures that can exist	1) Explanatory video (speaker: Núria Esteve-Gibert) 2) Task 1: Recognize the Gesture Phases 3) Task 2: Select the best Gesture Phase annotation, 4) Annotation tutorial "Let's annotate together" 5) Task 3: Now it's your turn

	• learn how to annotate phasing in ELAN	
Challenge: Identifying Rhythmic Properties - Beat-like-ness	• realize that rhythmic properties are another aspect of the prosodic dimension • understand the concept of beat-like-ness • evaluate the level of beat-like-ness in a gesture. • learn how to annotate beat-like-ness in ELAN	1) Explanatory video (speaker: Ingrid Vilà-Giménez) *2) In development* *3) In development* *4) In development* 5) Task 3: Now it's your turn
Challenge: Identifying Gesture Referentiality	• realize that Referentiality is one aspect of the meaning dimension • identify different meanings that a gesture can convey • realize that referential meanings can overlap • learn how to annotate the gestural meaning.	1) Explanatory video (speaker: Ulya Tütüncübasi) 2) Task 1: Recognize the Gesture Referentiality 3) Task 2: Select the best Gesture Referentiality annotation 4) Annotation tutorial "Let's annotate together" 5) Task 3: Now it's your turn
Challenge: Identifying Gesture Pragmatics	• realize that pragmatic meaning is another aspect of the meaning dimension • discover the five pragmatic domains (speech act marking, operational marking, stance marking, discourse organization, interactional marking) • realize that a gesture may convey multiple pragmatic meanings • realize that pragmatic meaning can also overlap with referential meaning • learn how to annotate the pragmatic domain.	1) Explanatory video (speaker: Júlia Florit-Pons) *2) In development* *3) In development* *4) In development* 5) Task 3: Now it's your turn

Table 3: Overall structure of the *M3D Training Program*

In addition to the "Challenges" tab, learners can also access the "More" tab of the website, which includes four sections. First, in "About this training", structural information about the training program is provided, while "About us" lists the members of the *M3D Training Program* development team. "Additional resources" offer a full list of the useful references

from each challenge, as well as a list of studies that have successfully implemented M3D. Finally, there is a "frequently asked questions" (FAQ) section where learners can find common questions and answers related to each gesture dimension. Within the preparation-tab, a section called “Assessing Annotations” provides learners with valuable information about the subjective nature of annotations, as they can often be a theory-driven and subjective affair. It includes specific tips on how to develop multimodal corpora and how to achieve reliability when applying M3D.

Following ADDIE, the training program underwent two revision phases after completing development: The first revision was carried out within the *M3D Training Program* team, while the second revision was done by three volunteer reviewers. The three volunteer reviewers had participated in the hybrid M3D labeling workshop by Patrick Louis Rohrer that took place in January 2023 at UPF. Therefore, they were already familiar with the existing dimensions in M3D and knew how to annotate them. They provided feedback regarding the usability, acceptability and feasibility of the training program.

Step 4: Implement

To ensure that all learners begin the training program with a consistent baseline of knowledge, the training program consists of the two aforementioned preparatory sections, encompassing an explanatory video addressing the main approaches of M3D and the dimensions it consists of, as well as a tutorial video highlighting the key features of ELAN. In addition to ensuring a baseline of knowledge, the preparatory sections serve to gradually enter the subject at hand, as it might be overwhelming for learners to enter the program with the assessment and annotation of gestures.

Step 5: Evaluation

As suggested by ADDIE, the creation process of the *M3D Training Program* was followed by an evaluation process. The evaluation process is thoroughly described in the upcoming Chapter 3 of this thesis.

3. Goal II: The evaluation of the online *M3D Training Program*

Participants

For the recruitment of participants, an initial sign-up form was distributed, weeks before the training program started. The form was distributed among the ViCom members across German Universities, as well as among the students who participated in an ELAN workshop at University of Cologne. It was also sent to the participants of the in-person M3D workshop who showed interest in sharing M3D with their lab members at the Faculty of Communication and Humanities at the University Finis Terrae Santiago, Chile and the Max Planck Institute for Evolutionary Anthropology in Leipzig.

A total of 17 participants, from 10 different universities around the world, including Germany, Norway, Spain, Malta, Romania, Chile and the United States signed up. Among the participants were five Bachelor's students, three Master's students, four PhD candidates, three Postdocs, one associate professor, and one assistant lecturer of linguistics. Their area of study spanned topics including bilingualism, sign language and animal communication. They all shared a common interest in studying gestures. None of the participants had annotated gestures before.

From the 17 participants that signed up, 13 started the training program. The criteria we used to consider that someone started with the program was their initial engagement in the training program in the form of attending the ELAN workshop. Participants who previously indicated that they could not attend our workshop but were committed to watching the recording of it at a later time or had already attended Florence Baills' (Postdoctoral Researcher, University of Cologne) workshop were also considered as starting the program.

Pre-training procedure

Following ADDIE, we first assessed the students' current knowledge about annotating gestures through the aforementioned sign-up form. This included questions, such as: Have you used the annotation tool ELAN before? Have you studied gestures before? Have you annotated gestures before? etc. (see Appendix A for the full questionnaire). This form was administered via Google Forms.

After that, all participants were invited to our introductory one hour ELAN workshop. The purpose of this workshop was to ensure that everyone possessed an equal level of technical knowledge of the annotation software used for M3D. Since part of the evaluation procedure was to assess inter-annotator reliability, it was deemed important to ensure that participants had a thorough understanding of ELAN. The workshop encompassed a presentation highlighting the key features of ELAN, such as the definition of tiers, types and controlled vocabularies followed by a practical demonstration on how to import media and annotation templates, how to create tiers and types and how to update a controlled vocabulary as well as an annotation demonstration. The workshop served also as a measure to ensure that participants had a basic understanding of ELAN and did not lose valuable time seeking assistance in case they encountered any ELAN related issues during the two-week period. The workshop took place on Friday, the 19th of May 2023 and 9 students attended the workshop. Those who did not attend the workshop either already participated in Florence

Bailly's ELAN workshop that took place two days prior (N= 2), or they watched the recording of the workshop at a later time (N= 2).

Training Procedure

On Monday, the 22nd of May 2023 after the ELAN workshop, the participants were given two weeks to complete the training program (until the 5th of June 2023), which included the preparatory sections "Learning about M3D" and "Getting started with ELAN", as well as 3 of the main challenges: Identifying Gesture Units, Identifying Gesture Phases, and Identifying Gesture Referentiality. These three challenges were chosen as they cover the main aspects of gesture identification and description, while not overwhelming participants with too much information in a short timespan. A workload of approximately 8-10 hours in total was estimated. During the two weeks, participants had the opportunity to ask questions regarding organizational aspects of the training program or technical aspects regarding ELAN. After completing the training program, participants were directed to a confirmation page indicating the successful completion of the *M3D Training Program*.

Post-training procedure

After completing the training, participants indicated their completion via email to subsequently receive the post-training annotation task. The post-training annotation task consisted of a 20-second video sequence which served as a representative example of all the aspects of gestures covered in the training program. The video sequence was specifically chosen because of the presence of different gesture meanings and the variety of phasing configurations. Previously evaluated by the members of the M3D development team, the video sequence was deemed to be of medium difficulty.

When finished with the post-training annotation task, the participants were requested to send the annotation file back to then receive the final evaluative set of questionnaires

consisting of the ADDIE questionnaire, the System Usability Scale (SUS) and our own questionnaire (hereinafter referred to as “in-house questionnaire”; see Appendix B, C and D). The timeframe for participants to complete the post-training annotation task, send it back and receive the final set of questionnaires varied based on each participant’s individual pace. No specific time restrictions were given, except for the requirement to complete it by the end of the two weeks.

Evaluation measures

Table 4 illustrates the evaluation measures used, alongside the tools utilized for assessing them and the method employed.

Evaluation measure	Definition	Tool	Method
Usability	The degree to which a learning resource can be effectively, efficiently and satisfactorily used by learners to accomplish specific goals within the defined context of use, analyzing if the training meets the needs of learners (Eicher et al., 2019; Holthe et al., 2018).	System Usability Scale (Brooke, 2013)	Quantitative
Acceptability	How learners react to the learning resource and the extent to which learners are inclined to use the resource (Cavallo et al., 2015; Bowen et al., 2009).	ADDIE questionnaire (Branch, 2009 p.155)	Quantitative
		In-house questionnaire	4 qualitative questions; 3 quantitative
Feasibility	The extent to which training can be successfully carried out. It involves assessing the willingness of learners to use the learning resource and assessing if the implementation is viable (Guadana, 2018; Weiner et al., 2017).	Starting-and -completion ratio	Quantitative
Effectiveness	The measurement of the impact of training, taking into account the ability of learners to complete tasks within the training and the quality of the tasks output (Glasgow et al., 2019; Brooke, 2013). It involves investigating the reasons behind the success of the output or lack thereof (Glasgow et al., 2019).	Inter-annotator reliability assessment	Quantitative

Table 4: Definition of the evaluation measures along with a description of the tool and method used for the evaluation (usability, acceptability, feasibility, effectiveness)

In the following four subsections [*(I) Usability, (II) Acceptability, (III) Feasibility* and *(IV) Effectiveness*], the evaluation measures, the evaluation tool and method used will be described in more detail.

(I) Usability

Usability was explored by using the validated System Usability Scale (SUS) (Brooke, 2013). SUS is a validated tool for assessing the usability of learning resources and has been available for 30 years. Being referenced in over 1,300 articles and publications, it has gained significant recognition (Klug, 2017; see Horne et al., 2022; Moura et al., 2021, for examples of its application). Multiple studies have demonstrated its high reliability as a measurement tool (Bangor et al., 2008; Borsci et al., 2009). It is particularly useful, as it can provide reliable results even when implemented on small sample sizes (Brooke, 1996). Furthermore, it is also recognized for its flexibility, which enables a usability assessment of various systems, including interactive voice response systems (IVRs), novel hardware platforms, traditional computer interfaces and websites. Regardless of the specific technology being evaluated, the SUS can provide valuable insights into the usability of a system (Bangor et al., 2008).

SUS consists of a 10-item questionnaire in which participants assess online training material from a subjective perspective (see Appendix C for the SUS questionnaire). Each item is scored on a 5-point Likert-type scale, ranging from “Strongly Agree” to “Strongly Disagree”. Higher scores correspond to higher perceived usability. The score contribution of each item can vary between 0 and 4. When it comes to the positively worded items (1,3,5,7 and 9), their score contribution is calculated by subtracting 1 from the position on the scale. On the other hand, for the negatively worded items (2,4,6,8 and 10), the contribution is

determined by subtracting the scale position from 5. After adding up all the scores, the sum is multiplied by 2.5 to calculate the overall value of SUS. SUS scores have a range of 0 to 100. A score above 68 is said to be above average, while a score below 68 is considered below average (Brooke, 1996). Additionally to the numerical and letter-grade scale of the SUS, Bangor et al. (2009) developed an adjective rating scale ranging from "worst imaginable" to "best imaginable" (see **Figure 2**).

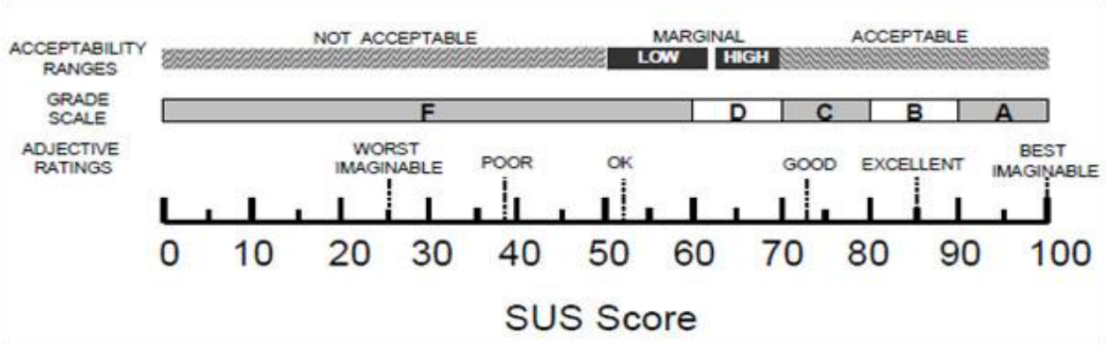

Figure 2: Grade rankings of SUS scores (taken from Brooke, 2013)

(II) Acceptability

A questionnaire that was developed within the ADDIE framework was used as an assessment tool for acceptability (Branch, 2009, Level 1: Perception, Part 1: Course Content, p.155, see Appendix B for the ADDIE questionnaire). It focuses on the practicability of the content of the training material and in the assessment of whether the material was sufficient to meet the learning objectives. The questionnaire consists of 8 items. Similarly to the System Usability Scale (SUS), each item is scored on a 5-point Likert-type scale, ranging from "Strongly Agree" to "Strongly Disagree". The mean value of the Likert scale for the 8 items was calculated using the Descriptives-function in Jamovi (The jamovi project, 2022).

Additionally, we used a 7-item in-house questionnaire as another assessment tool for acceptability. The items in the questionnaire consist of 4 open-ended questions, one yes/no

closed question, one multiple-choice question and one question in a 5-point Likert scale format (see Appendix D for the in-house questionnaire). It aims to receive concrete qualitative feedback specifically about the *M3D Training Program*. For instance, participants could say exactly which challenge and which task they found difficult, and explain why. This was key, as this information cannot be retrieved from a standardized questionnaire. The in-house questionnaire was evaluated manually.

The System Usability Scale, the ADDIE questionnaire together with the in-house questionnaire were administered in one single form via Google Forms.

(III) Feasibility

Feasibility was assessed by gathering data on the starting-and-completion ratio which involved comparing the number of participants who attended the ELAN workshop against the number of participants who completed the post-training annotation task. As previously mentioned, a total of 13 participants started the training program.

(IV) Effectiveness

Quantitative data on the effectiveness of the training program was collected by an inter-annotator reliability assessment. Following ADDIE, we compared the performance within the group of participants rather than comparing it against the desired outcome. The Gesture Units, Gesture Phases and Gesture Referentiality annotations made by the participants were compared for the video sequence that encompassed all the elements of gestures targeted in the training program.

The reliability for Gesture Units and Gesture Phases annotations was assessed by calculating Fleiss' Kappa (Fleiss, 1971) using the inter-annotator reliability function in ELAN, which makes use of the algorithm developed by Holle and Rein (2014). This algorithm functions by taking both temporal aspects as well as the annotated label into

account. It identifies overlaps in annotations, considering them as “paired” when they have a minimum overlap (set at 60% for the current analysis, the default value). This assessment thus incorporates both the segmentation (boundary detection) and the assigned label of the annotations.

For the evaluation of Gesture Referentiality, Fleiss’ Kappa was not used as a measure of reliability because labellers may assign multiple annotations to the same gesture, which would not be in line with the assumption of mutual exclusivity for Fleiss’ Kappa. Therefore, a MASI distance metric was used, which allows for the assessment of non-mutually exclusive labels (Passonneau, 2006). Secondly, the presence of a considerable number of labels for non-referential gestures introduces the possibility of encountering the “Kappa Paradox”. This paradox arises when one label occurs more frequently than any other, thus impacting the calculation of chance agreement (Gwet, 2014). Therefore, reliability for Gesture Referentiality was assessed with another measure, namely Gwet’s Agreement Coefficient 1 (Gwet’s AC1, Gwet, 2008). Gwet’s AC1 makes use of the same formula as the Kappa, but calculates the chance agreement without any interference. That means, unlike Kappa, Gwet’s AC1 is not influenced by the “Kappa Paradox” but can be interpreted in a similar way (Gwet, 2014). The calculation of the MASI distance metric and Gwet’s AC1 was carried out in R (R Core Team, 2021).

4. Results

A total of 10 participants completed the *M3D Training Program*. Among participants, there was significant variation in how long it took them to complete the training program. While some indicated their completion of the program only a few days after beginning, others made

use of the entire two-week period. Nonetheless, all participants filled out the evaluative questionnaire within a maximum of one day after sending their post-training annotation file. The results of the program evaluation will be presented based on the responses from these 10 participants. In order to maintain anonymity, the assigned numbers of the participants in the tables and figures were randomized.

(I) Usability

The results of the System Usability Scale questionnaire demonstrate consistently high ratings across all items (see **Table 5**).

Items	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Total	**SUS score/item**
Item 1(P)	2	4	4	4	4	4	3	4	4	3	36	90
Item 2 (N)	3	4	4	3	4	3	4	4	4	3	36	90
Item 3 (P)	4	4	4	4	4	4	4	4	4	3	39	97,5
Item 4 (N)	3	2	3	2	4	4	4	4	4	1	31	77,5
Item 5 (P)	4	4	4	4	3	3	3	4	4	2	35	87,5
Item 6 (N)	4	3	4	4	4	4	4	3	3	0	33	82,5
Item 7 (P)	2	4	4	1	4	4	4	4	4	1	32	80
Item 8 (N)	2	3	4	4	4	3	3	4	4	2	33	82,5
Item 9 (P)	1	3	4	2	4	2	4	3	4	1	28	70
Item 10 (N)	4	4	2	3	3	3	4	4	4	2	33	82,5
												Mean SUS score: 84

Table 5: Descriptive statistics of the System Usability Scale (SUS) (Brooke, 2013), top row P1-P10 indicating the participant number (P (participants) + nr.), P = positively worded item; N = negatively worded item **(score already subtracted from 5), click here to go to questionnaire**

The results for usability show an overall SUS score of 84, which identifies the training program having an “Excellent” usability [Bangor et al. (2009) (**Figure 2**)]. Furthermore, it is labeled with the grade B on the letter-grade scale (see **Figure 2**). Individual scores range from 70-97.7 on a 1-100 scale. Looking at the single SUS scores, it is evident that all items

achieved a score above 68, which, according to SUS, is considered “above average” (Brooke, 1996). Item 3 (“The training program was easy to use.”) received the highest score among all items with a SUS score of 97.5.

Both Item 1 (“I think that I will refer back to this learning resource in the future.”) and Item 2 (“I found the training program unnecessarily complex.”) obtained high scores, scoring 90 on the SUS. Furthermore, the lowest SUS score received item 4 (“I thought that I would need the support of a technical person to be able to use the training program.”), scoring 77.5 on the SUS as well as item 9 (“I felt very confident using the training program.”) with a SUS score of 70.

(II) Acceptability

The results for the ADDIE questionnaire demonstrate a high acceptability of the training program (see **Table 6**)

Descriptive statistics								
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
N	10	10	10	10	10	10	10	10
Missing	0	0	0	0	0	0	0	0
Mean	4.30	3.70	4.90	4.20	4.90	5.00	4.70	4.40
Median	4.50	4.00	5.00	4.50	5.00	5.00	5.00	4.50
Standard deviation	0.823	1.16	0.316	1.23	0.316	0.00	0.483	0.699

Table 6: Descriptive statistics of the ADDIE questionnaire, **click here to go to questionnaire**

The mean Likert value for the 8 items of the ADDIE questionnaire lies between 3.70 and 5.00, with item 6 “The visual material added value” receiving the highest score with a mean score of 5.00 (SD = 0.00) and item 2 “The content of the training program was relevant for my studies” receiving the lowest score with a mean score of 3.70 (SD = 1.16). The other 6 items can be found within a range of 4.20 and 4.90. Item 4 “Sufficient instructions were

given to complete the tasks successfully.” gave the highest standard deviation of all items (SD = 1.23).

When asked about their overall enjoyment of the training program in the in-house questionnaire (“Overall, I enjoyed doing the training.”), 6 participants responded with "Strongly Agree" while 4 participants responded with "Agree", indicating a high level of enjoyment among the participants. The first item (“Did you have difficulties with one specific challenge or section in the training?”) received 7 very helpful responses, particularly addressing the difficulty of identifying and segmenting Gesture Units and Gesture Phases. Two participants mentioned struggling with identifying subtle movements as strokes. Another participant expressed difficulties in differentiating between a single multidirectional stroke and two (or more) subsequent strokes. Moreover, one participant expressed challenges in applying the annotations of the provided examples to their own work and suggested adding more examples for Gesture Phases to the training program. This suggestion was also mentioned for Gesture Units. All in all, it appears that more supporting material would be beneficial in the areas of Gesture Units and Gesture Phases. No improvement suggestions were mentioned for Gesture Referentiality.

Overall, participants did not encounter significant issues with website navigation. In response to the question about technical difficulties within specific challenges, one participant reported having issues with the feedback function not working properly when selecting “1 multidirectional stroke” in exercise 8 in Task 1 in the challenge addressing Gesture Phases. Another participant found the GIF-looping of task 2 to be too fast for the identification of the number of Gesture Phases. Some participants mentioned technical doubts related to ELAN, but these were claimed to be resolved through email communication with the team of the training program. For the question about potential improvements, we received helpful suggestions, such as in the following responses:

Participant 3: “As I said above: The transfer between the examples and the own work was hard for me. Although I found the examples very well selected. Maybe you could add more cases in doubt and explain them in a separate video. It's always easy to look at examples given. But sitting in front of a video with no annotations is way harder. But it was manageable.”

Participant 5: “What I mentioned already in another answer: maybe a couple more examples in the gestural phases tasks, or even addressing directly the issue of having to pay attention to multiple factors e.g. "what do you pay attention to when segmenting gestural phases?" a) changes in speed and accentuation b) changes in direction (e.g. between preparation and stroke) c) overall feeling: what movement seems the most important? If I remember well, all of these things were mentioned (and maybe you, knowing it better, would argue they could be listed differently), but I felt like a reminder to keep them point-by-point in mind while coding would have helped me avoid the mistakes I made in the challenge.”

Participant 6: More tasks on the web to practice with, with more detail on why certain answers are wrong and others are right.

Participant 9: Gesture phases were the most difficult part. It was the one where I had most differences in my annotation compared to the solution file. Maybe it would be good to focus on recognising very subtle movements in the exercises. Also: the final video there lasts 40 minutes. It is difficult to keep the attention for so long. Better more interactive exercises and shorter videos.

(III) Feasibility

The starting-and-completion ratio stands at 13:10, with 13 participants starting the training program and 10 participants finishing it. This finding suggests that the training program is highly feasible (see Cisel, 2014). Among the three participants who did not complete it, one participant mentioned having a busy schedule as the reason for not being able to finish. Another participant could not meet the given deadline, and the third participant did not provide any specific reasons for not completing the program.

(IV) Effectiveness

The results of the inter-annotator reliability assessment for Gesture Units and Gesture Phases are presented in **Table 7**.

Fleiss' Kappa: Gesture Units	
Kappa including "Unmatched" value:	k = (0.6694 - 0.5253) / (1.0 - 0.5253) = 0.3036
Kappa excluding "Unmatched" value:	k = (1.000 - 1.000) / (1.0 - 1.000) = NaN (k = 1.0)
Fleiss Kappa: Gesture Phases:	
Kappa including "Unmatched" value:	k = (0.6497 - 0.2570) / (1.0 - 0.2570) = 0.5285
Kappa excluding "Unmatched" value:	k = (0.8070 - 0.3170) / (1.0 - 0.3170) = 0.7174

Table 7: The Fleiss' Kappa values for Gesture Units and Gesture Phases

The results of the reliability calculations for Gesture Units show an agreement of 0.3036 for the Fleiss' Kappa including the "Unmatched" value, which in literature is interpreted as "poor" agreement (Fleiss et al., 1981, p.604). While the Fleiss' Kappa value including "Unmatched" value considers everything (matched and unmatched annotations), the Kappa excluding "Unmatched" value considers only if the annotations matched (overlapped at least 60%). The Kappa excluding the "Unmatched" value for Gesture Units shows no data. This means that for the excluding "Unmatched" value for Gesture Units, only matched annotations were considered, and all participants had the same annotation label, which results in a "perfect" agreement. The results of the reliability calculations for Gesture Phases shows an agreement of 0.5285 for the Fleiss' Kappa including "Unmatched" value, which is interpreted as "fair" agreement (Fleiss et al., 1981, p. 604). The reliability calculations for the Fleiss' Kappa excluding "Unmatched" value shows "substantial" agreement with the Fleiss' Kappa standing at 0.7174 (Landis and Koch, 1977, p. 165).

The descriptive analysis of the number of stroke annotations of the 10 participants shows that most participants annotated between 11 to 15 strokes (SD = 1.59) (see **Table 8**)

Descriptives

	Stroke Count
N	9
Missing	0
Mean	12.6
Median	12.0
Standard deviation	1.59
Minimum	11.0
Maximum	15.0

Table 8: Descriptive statistics of the stroke count of 9 participants using Descriptives-function in Jamovi (1 participant with an 26 stroke annotations excluded from this calculation)

As ADDIE suggests, we did not compare the annotations of the participants against a sample. However, an experienced M3D gesture labeller identified 15 strokes, which lies within the range of what the participants annotated. **Figure 3** shows the stroke annotations of the participants in more detail.

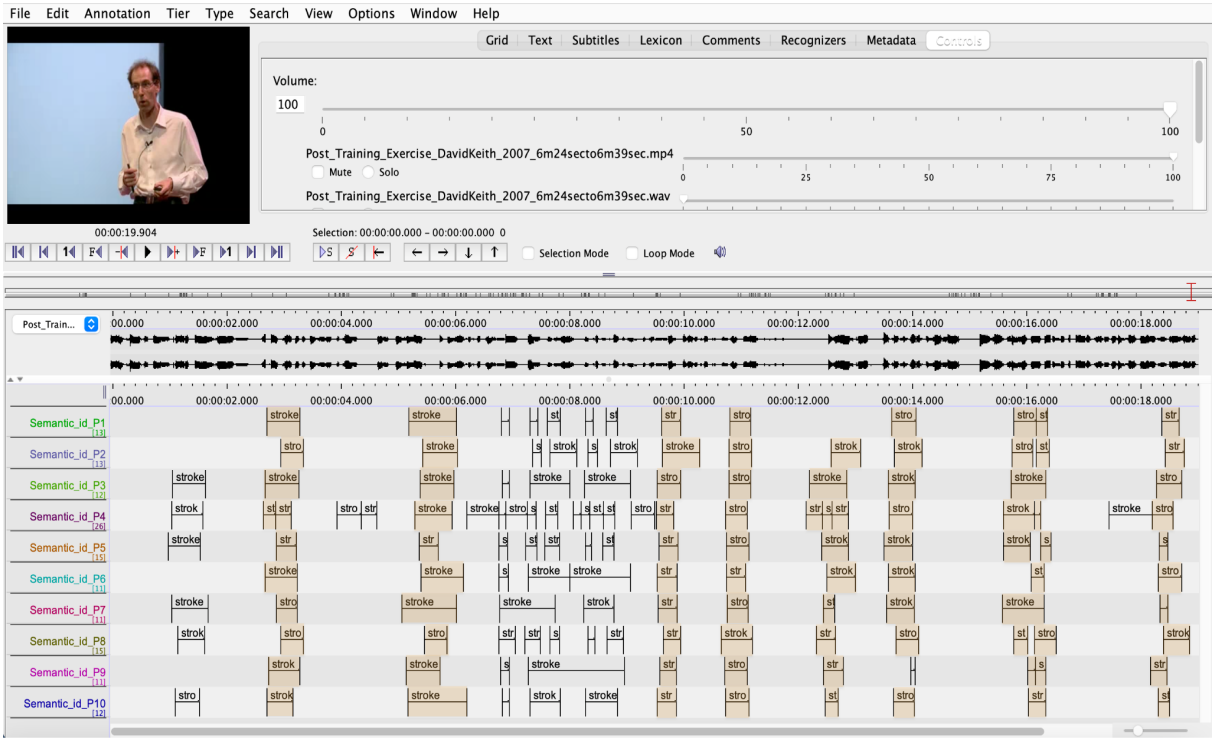

Figure 3: Stroke annotations of the 10 participants in ELAN for a 20-second video sequence (one participant's annotation per tier), annotations with a clear overlap are highlighted in orange, click here to see video sequence together with the stroke annotations

Figure 3 provides a clear visual representation where it is evident at first glance that all participants were able to roughly identify the same eight strokes with a similar onset and offset annotation in the stream of movements. However, four participants did not identify the first stroke of the video sequence. Nevertheless, except for Participant 4 (Semantic_id_P4), all participants could successfully distinguish a recoil-like movement (a movement that occurs after a stroke, which is due to biomechanics) from a meaningful movement at 00:04. Additionally, at 00:12, except Participant 1 (Semantic_id_P1), all participants were able to identify meaningful movements as strokes. The same goes for 00:14, 00:16 and 00:18. All in all, the participants show high stroke-identification abilities.

In terms of stroke-segmentation, particularly at 00:16, there is a disagreement among participants regarding the annotation of the movement as either a single long multidirectional stroke or as two (or more) subsequent strokes. This disagreement might be due to the nature of the movement itself as the two strokes seamlessly merge into each other. This opens possibilities for different interpretations and perceptions.

Furthermore, there are some discrepancies at 00:08. The participants seem to have issues in segmenting subtle beat-like movements that occur right after another. However, there are still some tendencies visible. For example, participant 1 (Semantic_id_P1), participant 5 (Semantic_id_P5) and participant 8 (Semantic_id_P8) all annotated five small strokes. On the other hand, participant 3 (Semantic_id_P3), participant 6 (Semantic_id_P6) and participant 10 (Semantic_id_P10) annotated one small stroke followed by two equally long stroke annotations. Participant 9 (Semantic_id_P9) has similar annotations, with the only difference of annotating a single long multidirectional stroke instead of two subsequent ones. In summary, participants demonstrated their ability to identify meaningful movements as strokes within the video sequence. However, they encountered difficulties when it came to segmenting these movements.

In terms of the reliability for Gesture Referentiality, the global AC1 value shows an AC1 of 0.672 [CI (0.612, 0.732), $p < .001$]. The result can be interpreted as "substantial" agreement as the AC1 value was found to be interpreted equally as Cohen's and Fleiss' Kappa (Gwet, 2014; Keener, 2020). **Table 9** shows the Gesture Referentiality annotations of the 10 participants in more detail.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
1	N	N	N	N	M	N	N	N	N	N
2	N	N	N	N	N	N	N	N	N	N
3	N	N	M	N	N	N	N	N	N	N
4	N	N	N	N	N	N	N	N	N	N
5	N	I	N	N	N	N	N	N	N	N
6	N	M	N	I, M	I, M, A	N	N	N	N	N
7	I, M	N	N	N	I, M	N	N	I, M, A	I	N
8	M	N	N	N	N	N	N	N	M	I
9	N	N	I	N	M, A	I, M	N	N	N	M
10	A	N	M, C	N	M	M	N	A	N	M, A
11	N	N	N	N	M	N	N	I, M	N	I, M
12	N		A			M	N	I	N	I, M
13			M			N	N		M	I, M
14			M			N	N			
15			N			M	N			
16							N			
17							I, M			
18							M			
19							N			
20							N			
21							N			
22							A			
23							N			
24							M			
25							I, M			
26							N			

Table 9: Gesture Referentiality annotations for the 10 participants in Excel (one participant's annotation per column; first column indicating the stroke (gesture) count), meaning dimensions: N = non-referential, I = iconic, M = metaphoric, A = abstract deictic, C = conventional, combinations thereof separated by a comma

The majority of participants agreed that the first five gestures in the sequence were non-referential, meaning they did not carry a referential meaning (i.e., they did not show degrees of iconicity, metaphoricity, or deixis). Starting from the sixth gesture, most participants agreed that the subsequent two gestures had a metaphoric dimension, and thus were referential in nature. In the case of gesture eight, 7 out of 10 participants agreed that it was non-referential again. It is noteworthy that all participants adhered to the M3D guideline of not annotating a non-referential gesture having a referential dimension as well (i.e., that

the referential/non-referential distinction is indeed mutually exclusive). This shows that participants have committed to following the annotation principles set out in the *M3D Training Program*.

5. Discussion and conclusions

Following the development of the *M3D Training Program*, it is important that the program undergoes an evaluative process to demonstrate its high quality nature. This evaluation not only enhances the program's validity and transparency, but provides learners with confidence that they are engaging in a training program that has its place in the field. Conducting an evaluation after developing a training program has become a standard practice. The *M3D Training Program* has been evaluated in terms of usability, acceptability, feasibility and effectiveness. A qualitative method together with quantitative methods were utilized to highlight the strengths of the training program as well as to point out the particular areas of improvement. The results provide a thorough understanding of how the *M3D Training Program* performs in terms of the four evaluation measures.

(I) Usability

The application of the System Usability Scale (SUS) (Brooke, 2013) provided valuable insights into the usability of the training program. With a score of 84, the program's usability is clearly above average, reaching the adjective rating "Excellent" (Bangor et al., 2009). Particularly, the training program was found to be easy to use, and many indicated that they would make use of the program in the future. The level of complexity was considered appropriate and the integration of functionalities within the program was considered well-integrated. These findings highlight the success of the training program in creating a usable environment for learners.

The SUS item to assess learner confidence received the lowest score (SUS score of 70), suggesting that the program could be improved by specifically addressing learner confidence. A possible explanation for the relatively low confidence might be the complex nature of annotating gestures. Learning how to annotate gestures requires time and patience, and the memorization of a comprehensive set of guidelines. The training phase in the current study lasted two weeks, with an estimated workload of 8-10 hours. It is understandable that aspiring gesture annotators may not feel the same level of confidence as someone who can freely go through the training program without time constraints. The limited duration of the training phase may not have provided enough time for the participants to fully develop the same level of confidence as those who are able to dedicate more time to working with the training program in other settings. Nonetheless, when considering the achieved score of 70 on a scale of 100 from a holistic perspective rather than a relative one, it is still a considerably high score.

Additionally, in the System Usability Scale (SUS), some participants expressed the need for technical assistance to be able to use the training program. Given the ambiguity of this question, the observed result could potentially be caused by the technical challenges associated with ELAN, as the participants agreed on the statement that the training program was easy to use, earlier. Although an introductory ELAN workshop and a preparatory "Getting Started with ELAN" section was provided, some questions concerning the use of ELAN remained. The obvious reason behind that is that most questions arise as learners spend more time with a program, exploring and experimenting with its functions. The development team of the *M3D Training Program* received three emails concerning technical aspects of ELAN. While all questions could be successfully resolved, it might be possible that some participants chose to not reach out for technical assistance. This could have

influenced their experience with ELAN which then potentially resulted in having issues throughout the whole training program.

Nonetheless, the objective of the *M3D Training Program* is to train learners on how to annotate gestures using the M3D gesture labeling system, and it falls beyond the scope of the training program to train individuals on the technical aspects of how to use ELAN. The training program can be applied with any other multimodal annotation software [e.g. ExMARaLDA (Schmidt and Wörner, 2014), Anvil (Kipp, 2014) etc.]. However, for further improvements, it is suggested to include more external supporting resources that specifically address the use of the annotation software ELAN, such as YouTube tutorials and links to resources available directly on the ELAN website. Alternatively, learners can be directed to the website of the Max-Planck-Institute for Psycholinguistics, where they can find comprehensive support addressing technical aspects of ELAN. Apart from the already existing forum entries, it is worth noting that the ELAN-team is highly responsive and provides rapid assistance to new inquiries.

(II) Acceptability

Furthermore, the training program was shown to be highly acceptable, as the scores of the ADDIE questionnaire for assessing acceptability indicate a highly positive outcome. All participants acknowledged the value that visual elements brought to the learning process (i.e. explanatory videos, video sequences of gestures within tasks, annotation tutorial). This consensus highlights the positive impact of these components on the overall acceptability of the training program. The answers to the in-house questionnaire support the aforementioned findings of usability and acceptability: None of the participants encountered any issues when navigating through the training program. Furthermore, participants expressed a high level of enjoyment with the training program. This overwhelmingly positive feedback provides clear

indication that the training program was highly satisfactory for the participants, painting a very positive picture of their experience with the training program.

Nevertheless, the item assessing the adequacy of instructions provided to successfully complete the tasks gave the highest standard deviation of 1.23. This result reflects the individual needs of learners regarding the clarity of the instructions provided and that some learners need additional support for successfully completing the tasks. This is in line with the qualitative feedback we received from the in-house questionnaire, stating the need for more illustrative examples of gestural phenomena. The participants suggested including more examples of gestural movement before proceeding to the tasks, as they encountered difficulties related to the identification and differentiation of gestural phenomena. These qualitative findings are supported by the results regarding the adequacy of instructions (ADDIE) and the participants' confidence (SUS).

(III) Feasibility

The starting-and-completion ratio stands at 13 to 10, demonstrating high feasibility of the training program. The program was not monetarily compensated and some participants stated that the training program was not relevant for their studies. The only motivation shared by all the students who enrolled in the training program was their genuine interest in gesture studies and gesture annotation. It is important to note that among the participants who completed the training program, there was a significant variation in the timing of completion. Some participants submitted the post-training annotation task within a few days after starting the program, while others waited until the deadline to complete. This variation can be attributed to differences in individual schedules and availability.

Correspondingly, out of the 3 participants who did not complete the training program were a Bachelor's student who stated to have a busy schedule, a Master's student who could

not complete the program by the given deadline and an assistant lecturer who did not provide any specific reasons for not completing the program. It is likely that participants who were in their Bachelor's or Master's had final exams or other graded projects that they needed to prioritize over the voluntary participation in the training program. Similarly, the assistant lecturer may have had a busy schedule due to their academic responsibilities. These factors could have contributed to their inability to complete the training program within the given deadline.

Consequently, it is possible that the chosen deadline to complete training was too demanding. While a demanding deadline may impact the number of participants completing the training program, it is important to recognize that it does not directly reflect the overall feasibility of the training program as it is more of an organizational aspect that should be taken into account when considering future evaluation processes. On the other hand, this observation might also suggest that the training appeared to be too demanding and time consuming for them to manage alongside their other responsibilities. This would align with the qualitative feedback provided by two participants in the in-house questionnaire who expressed their opinion that the annotation tutorial videos may have been too long.

To improve the already notable feasibility of the training program, it might be useful to include shorter annotation tutorial videos in addition to the current ones, explaining only the key aspects of annotating gestures. This approach would allow learners to choose between two levels of detail based on their individual learning needs.

(IV) Effectiveness

As for the evaluation of effectiveness, the agreement for Gesture Phases, shows a "fair" to "substantial" agreement. After only two weeks of autonomous training, the participants were able to successfully detect the beginning and ending of subtle movement phases within a

stream of a 20-second-long video and to assign labels to the movement phases. Furthermore, achieving “fair” and even “substantial” agreement after just one annotation round is impressive, when considering that the inter-annotator reliability procedure we use in M3D involves regular meetings where annotations are compared and discussed in a group for later revision. Here, the participants were not able to communicate with each other during the annotation process to not influence each other’s annotations and to thus have representative evaluation results. The fact that such a level of agreement was achieved without having multiple annotation rounds reflects the effectiveness of the given input in the training program.

All in all, these findings, both qualitative and quantitative, indicate that the participants had a good understanding of the concept of strokes but struggled with the finer details of segmentation. To some degree, it should be emphasized that annotations are a subjective endeavor. Certain aspects of manual movement segmentation involve a subjective judgment and are not right or wrong per se. Especially for stroke annotations there are different strategies to assess the onset of offset of a stroke within a continuous flow of manual movement. Some annotators may pay attention to kinematic aspects (i.e. amplitude, velocity), others may focus more on form configurations (i.e. changes in handshape) to segment manual movement.

The presence of 2-3 annotation patterns for the stroke annotation at 00:08 in **Figure 3**, rather than 10 completely different annotations provides additional support for this explanation. As mentioned before, this is why it is crucial to have discussions among the group of annotators to allow evaluating different perspectives, to clarify ambiguities and to try to find agreement on annotation choices. Nevertheless, it would be still beneficial to provide more concrete guidelines on what to pay attention to when annotating, as one participant qualitatively stated having difficulties in deciding on what criteria to base their

Gesture Phase annotations on. Concrete guidelines would outline the specific factors that participants should pay attention to when annotating. By doing that, the consistency of the annotation segmentation could be improved. Similarly, considering the participants' expressed uncertainties and the observed poor agreement in identifying Gesture Units during the inter-annotator reliability assessment, additional supporting material on the criteria and techniques of how to identify Gesture Units needs to be provided.

As for Gesture Referentiality, the inter-annotator reliability results demonstrate a level of substantial agreement. This is remarkable, as the video sequence used for the final annotation task included not only non-referential beat-like movements, but also a variety of referential gestures and gestures where referential meanings overlapped (e.g. metaphoric and abstract deixis). Interpreting referential meaning and distinguishing between the sometimes hard to distinguish referential subdimensions (i.e. metaphoric and iconic) requires a comprehensive understanding of Gesture Referentiality, which the participants achieved in only two weeks.

Moreover, the inter-annotator reliability procedure we use in M3D for Gesture Referentiality involves comparing and discussing Gesture Phases annotations in the first assessment. After that, any stroke annotations for which agreement could not be reached are excluded from the further Gesture Referentiality annotations procedure. Again, the participants were not able to communicate with each other during the annotation process. The calculation of referentiality reliability is limited in that it included annotations even when participants did not agree on phasing, essentially creating a glass-ceiling effect, where perfect agreement is already not possible. However, having substantial agreement despite these constraints is very remarkable. This positive reliability result goes along with the lack of qualitative feedback addressing the Gesture Referentiality, indicating that the participants did not have significant issues in this dimensional aspect.

The *M3D Training Program* is still evolving. More aspects of gesture are going to be covered in the future, adding more content to the training program. For future research, it can be considered conducting another evaluation phase. This would allow for the assessment of the new aspects introduced in the training program, as well as providing an opportunity to further evaluate and improve the changes made to the existing components based on the current evaluation. For the evaluation of effectiveness, future research could explore the possibility of implementing an initial reliability-stage of comparing and discussing the annotations, by organizing regular (virtual) meetings. Learners would have the opportunity to collectively revise their annotations and review the cases of disagreement. This would contribute to a more standardized inter-annotator reliability procedure.

In conclusion, the evaluation results of the *M3D Training Program* indicate that it achieved excellent usability, with participants finding it easy to use and navigate, with an appropriate degree of complexity and integration of functionalities, while enhancing the participants' skills, thereby demonstrating strong feasibility and a satisfactory effectiveness. There is room for improvement, particularly in providing additional support for the identification and segmentation of Gesture Units and Gesture Phases, adding more illustrative examples and shorter annotation tutorial videos.

The *M3D Training Program* provides openly-accessible training material for the annotation of gestures using the approach of the M3D gesture labeling system. It makes use of a student-centered approach and guides learners through the learning process, gradually building up knowledge, while being provided with initial feedback. It consists of a total of 6 challenges and two preparatory sections that address the different dimensional aspects of gestures. The evaluation of the *M3D Training Program* in terms of usability, acceptability, feasibility and effectiveness provided valuable insights into the high quality nature of the

program and enhanced its validity and transparency. The novel *M3D Training Program* presents a valuable resource in the field of gesture studies.

6. References

Aguilar, L., De-la-Mota, C., & Prieto, P. (2009). Cat_ToBI training materials. Prosodia UPF. http://prosodia.upf.edu/cat_tobi/

Aguilar, L., De-la-Mota, C., & Prieto, P. (2009). Sp_ToBI training materials. Prosodia UPF. http://prosodia.upf.edu/sp_tobi/

Alenezi, A. (2020). The role of e-learning material in enhancing teaching and learning behaviors. *International Journal of Information and Education Technology*, *10*(1), 48-56.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., & Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation, 41*, 273–287.

Arthur, W. B. (2009). *The nature of technology: What it is and how it evolves*. Simon and Schuster.

Artstein, R. (2017). Inter-annotator agreement. *Handbook of linguistic annotation*, 297-313.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction*, *24*(6), 574-594.

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies*, *4*(3), 114-123.

Bolly, C. T. (2016*). CorpAGEst Annotation Manual (II. Speech Annotation Guidelines)*.[Bolly CorpAGEst © 2016]

Borsci, S., Federici, S., & Lauriola, M. (2009). On the dimensionality of the System Usability Scale: a test of alternative measurement models. *Cognitive processing*, *10*, 193-197.

Bowen, D. J., Kreuter, M., Spring, B., Cofta-Woerpel, L., Linnan, L., Weiner, D., Bakken, S., Kaplan, C. P., Squiers, L., Fabrizio, C. & Fernandez, M. (2009). How we design feasibility studies. *American journal of preventive medicine*, *36*(5), 452-457.

Branch, R. M. (2009). *Instructional design: The ADDIE approach* (Vol. 722). Springer.

Bressem, J., Ladewig, S. H., & Müller, C. (2013). Linguistic Annotation System for Gestures. In C. Müller, E. Fricke, A. Cienki, D. McNeill, S. Ladewig, & S. Tessendorf (Eds.), *Handbücher zur Sprach-und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science (HSK), 38/1*, (pp. 1098–1124). De Gruyter.

Brooke, J. (1996). SUS - A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester & I. L. McClelland (Eds.), *Usability evaluation in industry*, *189*(194), 4-7.

Brooke, J. (2013). SUS: a retrospective. *Journal of usability studies*, *8*(2), 29-40.

Carter, M., Fletcher, E., Sansom, A., Warren, F. C., & Campbell, J. L. (2018). Feasibility, acceptability and effectiveness of an online alternative to face-to-face consultation in general practice: a mixed-methods study of webGP in six Devon practices. *BMJ open*, *8*(2).

Cavallo, F., Aquilano, M., & Arvati, M. (2015). An ambient assisted living approach in designing domiciliary services combined with innovative technologies for patients with Alzheimer's disease: a case study. *American Journal of Alzheimer's Disease & Other Dementias®*, *30*(1), 69-77.

Cisel, M. (2014). Analyzing completion rates in the first French xMOOC. *Proceedings of the European MOOC stakeholder summit*, *26*, 51.

Eicher, C., Haesner, M., Spranger, M., Kuzmicheva, O., Gräser, A., & Steinhagen-Thiessen, E. (2019). Usability and acceptability by a younger and older user group regarding a mobile robot-supported gait rehabilitation system. *Assistive technology*, *31*(1), 25-33.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

Fleiss, J. L., Levin, B., & Paik, M. C. (1981). The measurement of interrater agreement. *Statistical methods for rates and proportions*, *2*(212-236), 22-23.

Fussell, G., & Bianco, M. (2021). Integrating Technology in Online Education to Foster Student Success. In *Society for Information Technology & Teacher Education International Conference* (pp. 165-169). Association for the Advancement of Computing in Education (AACE).

Glasgow, R. E., Harden, S. M., Gaglio, B., Rabin, B., Smith, M. L., Porter, G. C., Ory, M. G. & Estabrooks, P. A. (2019). RE-AIM planning and evaluation framework: adapting to new science and practice with a 20-year review. *Frontiers in public health*, *7*, 64.

Guadana, R. R. H. (2018). Extent of Utilization, Viability and Effectiveness of E-Learning in the Camanava City Colleges: Basis for the Framework of E-Camanava. In *2018 4th International Conference on Education and Technology (ICET)* (pp. 40-47). IEEE.

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.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Holle, H., & Rein, R. (2015). EasyDIAg: A tool for easy determination of interrater agreement. *Behavior Research Methods*, 47(3), 837–847

Holthe, T., Halvorsrud, L., Karterud, D., Hoel, K. A., & Lund, A. (2018). Usability and acceptability of technology for community-dwelling older adults with mild cognitive impairment and dementia: a systematic literature review. *Clinical interventions in aging*, 863-886.

Horne, M., Youell, J., Brown, L., Brown-Wilson, C., Dickinson, T., & Simpson, P. (2022). Feasibility and acceptability of an education and training e-resource to support the sexuality, intimacy and relationship needs of older care home residents: a mixed methods study. *Age and Ageing*, *51*(10)

Jackman, M. M., Nabors, L. A., McPherson, C. L., Quaid, J. D., & Singh, N. N. (2019). Feasibility, acceptability, and preliminary effectiveness of the OpenMind (OM) program for pre-school children. *Journal of Child and Family Studies*, *28*, 2910-2921.

Keener, A. (2020). *A Comparison of Cohen's Kappa and Gwet's AC 1 with a Mass Shooting Classification Index: A Study of Rater Uncertainty* (Doctoral dissertation, Oklahoma State University).

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kipp, M. (2014). Anvil: The video annotation research tool. Handbook of corpus phonology. In J. Durand, U. Gut, G. Kristofferson (Eds.), *Handbook of Corpus Phonology*, (pp. 420–436). Oxford University Press.

Klug, B. (2017). An overview of the system usability scale in library website and system usability testing. *Weave: Journal of Library User Experience*, *1*(6).

Kolb, D. A., Boyatzis, R. E., & Mainemelis, C. (2014). Experiential learning theory: Previous research and new directions. In R. J. Sternberg, L. Zhang *(Eds.), Perspectives on thinking, learning, and cognitive styles* (pp. 227-248). Routledge.

Kristanto, A. (2021). Integration of a Constructivism Approach in E-learning Photography Subjects: Feasibility and Effectiveness in Learning. *Journal of Education Technology*, *5*(2), 167-174.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior research methods, 41*, 841–849.

Lausberg, H., & Sloetjes, H. (2016). The revised NEUROGES–ELAN system: An objective and reliable interdisciplinary analysis tool for nonverbal behavior and gesture. *Behavior research methods*, *48*, 973-993.

Lombardi, M. M., & Oblinger, D. G. (2007). Authentic learning for the 21st century: An overview. *Educause learning initiative, 1*(2007), 1-12.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* University of Chicago press.

Mayfield, M. (2011). Creating training and development programs: using the ADDIE method. *Development and Learning in Organizations: An International Journal*, *25*(3), 19-22.

Moura, J., Estrela, M., Almeida, A. M., Ferraz, I., Figueiras, A., Roque, F., & Herdeiro, M. T. (2021). A usability study of Pharmacists' perceptions toward an online course for respiratory infections and antibiotic use. *Procedia Computer Science*, *181*, 269-276.

Murphy, R., Gallagher,L., Krumm, A ., Mislevy, J., & Hafter, A. (2014). *Research on the Use of Khan Academy in Schools*. Menlo Park, CA: SRI Education.

Öztürk, Y. E., & Gürler, İ. (2020). Evaluation of Moodle, Canvas, Blackboard, and Open EdX. In *ICT-based assessment, methods, and programs in tertiary education* (pp. 363-382). IGI Global.

Passonneau, R. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation.

Perniss, P. (2018). Why we should study multimodal language. *Frontiers in Psychology, 9*, 1109.

Phillips, D., & Cohen, J. (2015). Learning gets personal: How Idaho students and teachers are embracing personalized learning through Khan Academy. *JKAF. http://www. jkaf. org/content/uploads/2015/07/learning-gets-personal. pdf Raths, D.(2014). Nine video tips for a better-flipped classroom. The Education Digest*, *79*(6), 15-21.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. R-project. https://www.R-project.org/.

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).

Rohrer, P. L., Vilà-Giménez, I., Florit-Pons, J., Gurrado, G., Esteve-Gibert, N., Ren-Mitchell, A., Shattuck-Hufnagel, & Prieto, P. (2023). The MultiModal MultiDimensional (M3D) labeling system.

Sánchez-Cabrero, R., Casado-Pérez, J., Arigita-García, A., Zubiaurre-Ibáñez, E., Gil-Pareja, D., & Sánchez-Rico, A. (2021). E-assessment in e-learning degrees: comparison vs. face-to-face assessment through perceived stress and academic performance in a longitudinal study. *Applied Sciences*, *11*(16), 7664.

Silverman, K. E., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., Pierrehumbert, J. B. & Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. In *ICSLP* (Vol. 2, pp. 867-870).

Schmidt, T. and Wörner, K. (2014), "EXMARaLDA", In Handbook on Corpus Phonology, pp. 402-419. Oxford University Press.

Spatioti, A. G., Kazanidis, I., & Pange, J. (2022). A comparative study of the addie instructional design model in distance education. *Information*, *13*(9), 402.

Supriyatno, T., Susilawati, S., & Hassan, A. (2020). E-learning development in improving students' critical thinking ability. *Cypriot Journal of Educational Sciences*, *15*(5), 1099-1106.

The jamovi project (2022). *jamovi* (Version 2.3) [Computer Software]. Jamovi. https://www.jamovi.org

Vesselinov, R., & Grego, J. (2012). Duolingo effectiveness study. *City University of New York, USA*, *28*(1-25).

Vygotsky, L., & Cole, M. (2018). Lev Vygotsky: Learning and social constructivism. *Learning theories for early years practice*, *66*, 58.

Weiner, B. J., Lewis, C. C., Stanick, C., Powell, B. J., Dorsey, C. N., Clary, A. S., Boynton, M. H. & Halko, H. (2017). Psychometric assessment of three newly developed implementation outcome measures. *Implementation Science*, *12*, 1-12.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). Elan: a professional framework for multimodality research. *Proceedings of the Fifth International Conference on Language Resources and Evaluation(LREC'06)*. European Language Resources Association (ELRA).

Ye, F. (2014). Validity, reliability, and concordance of the Duolingo English Test. *University of Pittsburgh School of Education*.

7. Appendix

Appendix A: Item-list sign-up form

1. Name
2. Email
3. University
4. Academic Position

   Bachelor, Master, PhD, Postdoc, Other
5. Field of studies
6. Have you used the annotation tool ELAN before?

   Yes, No
7. Have you studied gestures before?

   Yes, No
8. Have you annotated gestures before?

   Yes, No
9. Please add a few lines to describe your experience in gesture studies (including annotating gestures). If not applicable, please leave this field blank.

Appendix B: Item-list (Branch, 2009, Level 1: Perception, Part 1: Course Content, p.155)

Questions slightly altered to adapt to our context of use

1. The objectives (learning how to annotate gestures using the M3D gesture labeling system) were met.
   Strongly Disagree 1 2 3 4 5 Strongly Agree
2. The content of the training program was relevant for my studies.
   Strongly Disagree 1 2 3 4 5 Strongly Agree
3. The activities within the training program (tasks, videos etc.) enhanced my understanding of the skills.
   Strongly Disagree 1 2 3 4 5 Strongly Agree

4. Sufficient instructions were given to complete the tasks successfully.
   Strongly Disagree 1 2 3 4 5 Strongly Agree

5. The training material supported my learning.
   Strongly Disagree 1 2 3 4 5 Strongly Agree

6. The visual material added value.
   Strongly Disagree 1 2 3 4 5 Strongly Agree

7. The delivery method (web-based) was appropriate.
   Strongly Disagree 1 2 3 4 5 Strongly Agree

8. The training program met my learning needs.
   Strongly Disagree 1 2 3 4 5 Strongly Agree

Click here to go back to results

Appendix C: System Usability Scale (SUS) (Brooke, 2013)

Questions slightly altered to adapt to our context of use

1. I think that I will refer back to this learning resource in the future.
   Strongly Disagree 1 2 3 4 5 Strongly Agree

2. I found the training program unnecessarily complex.
   Strongly Disagree 1 2 3 4 5 Strongly Agree

3. The training program was easy to use.
   Strongly Disagree 1 2 3 4 5 Strongly Agree

4. I thought that I would need the support of a technical person to be able to use the training program.
   Strongly Disagree 1 2 3 4 5 Strongly Agree

5. I found the various functions (explanatory videos, tutorial videos, tasks) in the training program were well integrated.
   Strongly Disagree 1 2 3 4 5 Strongly Agree

6. I thought there was too much inconsistency in the training program.
   Strongly Disagree 1 2 3 4 5 Strongly Agree

7. I would imagine that most people would learn to use the training program very quickly.
   Strongly Disagree 1 2 3 4 5 Strongly Agree

8. I found the training program very cumbersome to use.
   Strongly Disagree 1 2 3 4 5 Strongly Agree

9. I felt very confident using the training program.
   Strongly Disagree 1 2 3 4 5 Strongly Agree

10. I needed to learn a lot of things before I could get going with the training program.
    Strongly Disagree 1 2 3 4 5 Strongly Agree

Click here to go back to results

Appendix D: In-house questionnaire

1. Did you have difficulties with one specific challenge or section in the training? If yes, please tell us more about it. If you haven't had any particular difficulties, leave the field blank.

2. Did you have any issues while navigating through the website?

3. Did you have any other technical difficulties with one specific challenge or part in the training? If yes, please tell us more about it or attach a screenshot in the box below. If you haven't had any particular difficulties, leave the field blank.

   Please attach a screenshot here. If you do not have a screenshot, leave the field blank.

4. What could be improved in the training program? Please be precise.

5. Have you completed all the three challenges (G-Unit, G-Phases, G-Referentiality) of the training program?
   Yes, No

6. If your answer was "no", please indicate if it was because of external reasons (e.g. being busy) or if it had something to do with the training itself. In case of the latter, please indicate the concrete reason into "Other".
   External Reasons, Other

7. All in all, I enjoyed doing the training.
   Strongly Disagree 1 2 3 4 5 Strongly Agree