TÉCNICO LISBOA

Generating Realistic Sign Language Animations

Inês Moutinho de Gouveia Correia de Lacerda

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. Hugo Miguel Aleixo Albuquerque Nicolau
Prof. Maria Luísa Torres Ribeiro Marques da Silva Coheur

Examination Committee

Chairperson: Prof. João António Madeiras Pereira
Supervisor: Prof. Hugo Miguel Aleixo Albuquerque Nicolau
Member of the Committee: Prof. Tiago Vieira Guerreiro

November 2021

Acknowledgments

I would like to thank my dissertation supervisors, Prof. Luísa Coheur and Prof. Hugo Nicolau, for giving me the opportunity to participate in this unique project. Thank you for the support, guidance, and motivating me to always do my best. You made the whole process enjoyable with our relaxed meetings that helped me through the moments of greatest distress.

Thank you to the research group of the "LGP Corpus & Avatar" project (Refª PTDC/LLT-LIN/29887/2017) from the Institute of Health Sciences of Universidade Católica Portuguesa. In particular, I would like to thank, Mara, Neide, and Sebastião, for always being available to help me, for collaborating with me, and for all you taught me about Portuguese Sign Language and Deaf culture.

Thank you, Pedro, my partner in this challenging journey, for always being available to help me, for all the meetings and advise, for all the laughs and all the bloopers. This dissertation would have not existed, or would have not been with the same quality, without your work and help.

I would like to thank my parents, my sisters, my grandmother, my two furry babies, James and Bond, my boyfriend, and my closest friends, especially, Maria and Jin, for your unconditional support, for believing in me, even when I would not believe in myself, and for the undeniable patience when I would burden you with my worries.

Last but not least, I would like to thank all my participants for taking the time to join my studies. Thank you for collaborating with me and all the knowledge you taught me. It means a lot to see the impact this project has and the good feedback it received.

To each and every one of you, thank you for making this project possible. It was a truly challenging and enriching experience. This project helped me grow as a person and I got to learn about a language and a community I knew nothing about. I met so many incredible people and learned so much, thank you.

Abstract

The synthesis of Sign Language animations, in real-time, is a difficult task because signing avatars must account not only for multiple linguistic processes but also the naturalness of the movements. Most avatars are described as unnatural, emotionless, and stiff because they cannot accurately reproduce all the subtleties of synchronized body behaviors of a human signer. Our approach consists of the synthesis and simultaneous animation of manual and non-manual components, and secondary facial and corporal movements. The manual and non-manual components account for the morphosyntactic motions needed in Sign Languages and the secondary movements account for the naturalness of the avatar. This dissertation provides a pipeline that can be used for multiple digital applications. Animations produced by the new system were tested with 34 participants. The overall good performance and positive feedback indicate that the generated animations show great potential in the field of synthetic animation of signing avatars. In this dissertation, we introduce components that can be applied not only for Portuguese Sign Language but also for other Sign Languages. For instance, a pipeline for the synthesis of co-occurring facial expressions, a dynamic approach for transitions in-between signs, the generation of automatic secondary facial and corporal movements, and the integration and synthesis of mouthing animations. This breakthrough brings the state of the art one step closer to an automatic Portuguese to Portuguese Sign Language (LGP) translator.

Keywords

Portuguese Sign Language, Synthetic Animation, Computational Linguistics, Natural Language Processing

Resumo

A síntese de animações em Língua Gestual, em tempo real, é uma tarefa difícil porque avatares têm de reproduzir vários processos linguísticos e movimentos que sejam naturais. A maioria dos avatares são descritos como não naturais, sem emoção, e rígidos porque não conseguem reproduzir, com precisão, todas as subtilezas corporais de um gestuante humano. A nossa abordagem consiste na síntese e animação simultânea de componentes manuais e não-manuais, e movimentos secundários faciais e corporais. As componentes manuais e não manuais são responsáveis pelos movimentos morfossintácticos necessários nas Línguas Gestuais e os movimentos secundários são responsáveis pela naturalidade do avatar. Esta dissertação fornece um sistema que pode ser utilizado em várias aplicações digitais. As animações produzidas pelo novo sistema foram testadas por 34 participantes. De um modo geral, o bom desempenho e o feedback positivo indicam que as animações geradas mostram um grande potencial na área da animação sintética de Línguas Gestuais. Nesta dissertação, introduzimos componentes que podem ser aplicadas não só para a Língua Gestual Portuguesa (LGP) mas também para outras Línguas Gestuais. Por exemplo, um método para a síntese de expressões faciais simultâneas, uma abordagem dinâmica para as transições entre gestos, a geração automática de movimentos faciais e corporais secundários, e a integração e síntese de animações de labialização/mouthing. Este sistema traz o estado da arte um passo mais perto de um tradutor automático de Português para LGP.

Palavras Chave

Língua Gestual Portuguesa, Animação Sintética, Linguística Computacional, Processamento de Linguaguem Natural

Contents

List of Figures

List of Tables

Acronyms

LGP	Portuguese Sign Language
FACS	Facial Action Coding System
AUs	Action Units
ASL	American Sign Language
CORS	Cross-Origin Resource Sharing
IK	Inverse Kinematics
FK	Forward Kinematics
FE	Facial Expressions

1

Introduction

Contents

Spoken/written language and sign language are extremely different: one is an audio-oral language while the other is a spatial-visual language. Moreover, the sentence construction, the grammatical rules, and the vocabulary are also quite different. These differences lead to a language barrier between Deaf and hearing people, which unfortunately can lead to injustice and discrimination. In 2017, in the United States, there was a significant employment gap of 22.5% between deaf and hearing people [5]. A Portuguese to Portuguese Sign Language (LGP) translator could facilitate the communication between hearing and Deaf, thus, contributing to the social inclusion of the Deaf community and promoting equal opportunities.

Only in 1997, was LGP acknowledged as a teaching language for Deaf people, and together with the fact that there is still no official grammar, contributed to the lack of LGP linguistic resources, scientific knowledge, and teaching materials. The translator could provide an efficient way of learning Sign Language for both Deaf and hearing, therefore, allowing Deaf people better access to higher education and bridging the gap between the two communities.

The differences between the two languages and the fact that Sign Language is the main form of communication for Deaf people can also bring significant difficulties in their ability to read and comprehend Portuguese text. Studies, in the United States, have shown that many deaf students, from age 8 through age 17, do not exceed the fourth-grade reading comprehension equivalent [6]. Deaf people face daily hardship in accessing general and specialized information and services (e.g., health services) because most communication technologies are designed to support written or spoken language and not Sign Language. The development of a translator that could assist daily communications in schools, websites, and public services, overall, could potentially overcome the barriers Deaf people face when accessing sources of information.

1.1 Problem

An automatic written/spoken to sign translation system requires two components: a translator and a signing avatar. The translator converts written text into a sequence of glosses (i.e., lexical units that represent each gesture or sign in Sign Languages) and then the avatar displays the synthesized glosses and additional linguistic processes as signing animations. The synthesis of Sign Language animations in real-time is a difficult task because signing avatars must account not only for multiple co-occurring linguistic processes but also the naturalness of the movements.

Most avatars are described as unnatural, emotionless, and stiff [7] because they cannot accurately reproduce all the subtleties of synchronized body behaviors of a human signer. Building successful and understandable signing avatars requires expertise in many domains such as computer graphics, animation, biomechanics, and computational linguistics. This raises the following question: **"Is an**

automatic text-to-sign translator effective in generating realistic and natural Portuguese Sign Language animations?"

Research regarding hand signs and facial expression in Sign Language animations is scarce, and in a synthetic context, the blending of the two is still an open challenge. Existing solutions rely on Sign Language Annotations [8], Keyframe Animations [9], and Motion Capture methods [4]. Each approach provides advantages and disadvantages but all require a balance between quality and cost. The more accurate and natural the animations are, the more costly they are to be generated.

1.2 Approach

With the previous problem analysis in mind, our main goal is to **automatically generate realistic Sign Language animations**. Our approach is the continuation of past work that includes two components: a Portuguese to LGP translator [10] and a database with synthesized signs (i.e. animations) by a 3D avatar. Our implementation connects the two existing components while generating natural Sign Language animations.

An important component of Sign Language communication is facial expression; its use affects the meaning of a sign as well as its naturalness. This raises the following question: **Does the inclusion of non-manual components (e.g., facial expressions) enhance linguistic comprehension of Sign Language animations?** Our approach consists of the synthesis and animation of manual and non-manual components (e.g. facial expressions) that account for the morphosyntactic motions needed in Sign Languages, and also secondary facial and corporal movements that make the avatar seem more natural. This system provides a pipeline that can be used for multiple digital applications, for instance: an automatic text-to-sign language translator, a dictionary, a book translator, a virtual assistant, and a browser add-on.

Planning and scripting the facial and body movements of a signing avatar to correctly perform Sign Language is a difficult task. Minor variations in timing and speed parameters can lead to significant differences in the quality and understandability of sign animations [11, 12]. The transitions between signs rely heavily on the phonology of the previous and following signs and determine the movement fluidity that allows sign streams to be intelligible. Therefore, transitions can have an impact on the comprehension and naturalness of sign animations. To the best of our knowledge, we introduce a new approach for the interpolation of signs consisting of dynamic transitions. This raises the following question: **Do dynamic transitions have an impact on linguistic comprehension, optimal transition speed, naturalness, and preference of Sign Language animations?**

In addition to the contributions related to Sign Language generation, we introduce a solution that aims to maintain and feed the sign's database by non-tech experts (e.g., Linguists). The system facilitates the

process of adding, changing, and removing signs from the database.

Building successful and understandable Sign Language translation systems requires an understanding of sign languages to account for their complex linguistic aspects, and an understanding of Deaf culture to create systems that align with user needs and desires. To gain a deeper knowledge of Deaf Culture and Portuguese Sign Language, I read multiple articles and theses, participated in two Portuguese Sign Language courses and attended the "Portuguese Sign Language and Deaf Education" Master in which we learned about the Phonology, Morphology, Syntax, and psycholinguistics of LGP. Furthermore, this dissertation is part of the "LGP Corpus & Avatar" project being developed by the Institute of Health Sciences from Universidade Católica Portuguesa, in partnership with INESC-ID, and funded by Fundação para a Ciência e a Tecnologia (FCT) (Refª PTDC/LLT-LIN/29887/2017). This project relies on an interdisciplinary team that includes deaf native LGP signers, linguists with knowledge in LGP, computational linguists, natural language processing, and human-computer interaction researchers. Throughout the entire development process, this interdisciplinary approach was considered in which we would have weekly meetings and would develop together with, rather than for, the Deaf Community that is too often excluded from design processes [13].

1.3 Contributions

The main contributions of this dissertation are: (1) the synthesis of realistic Sign Language animations that can be used in multiple digital applications, (2) the development of the first automatic Portuguese to LGP translator that contains both manual and non-manual components based on linguistic information extracted from a corpus, (3) to the best of our knowledge, a new approach for the interpolation of signs consisting of dynamic transitions, (4) to the best of our knowledge, a new approach for the synthesis of co-occurring facial expressions, (5) a solution that aims to maintain and feed the sign's database by non-tech experts, (6) three user studies with people fluent in LGP and beginners to assess the linguistic comprehension and perceived quality of the animations.

1.4 Thesis Outline

In this dissertation, we present an approach for the synthesis of Sign Language animations. In Chapter 2, we provide an overview of the background work describing detailed notions of LGP's grammar and components. In Chapter 3, we present a state of the art analysis on the different techniques for the synthesis of facial expressions, and for the synthesis of linguistic and secondary movements that are incorporated in signing animations. In Chapter 4, we depict the process of creating signs and facial expressions that are the basis of our tool described in Chapter 5. In Chapter 6, we present the evaluation

methodology and the analysis of the results, and finally, in Chapter 7, we deliberate on our current achievements and suggestions for future work.

2

Background

Contents

In this section, we describe fundamental concepts related to sign languages, more specifically, some detailed notions of sign languages components and structure. Sign language is not a universal language. Sign languages are natural languages that differ from country to country. Generally, each country has its own native sign language and some have more than one. In Portugal, we have Portuguese Sign Language. The first studies on LGP appeared in the '90s, so there is not much research and knowledge about this language, and even across Portugal, there are some lexical variations according to the area in the country. For instance, regarding the dialectal variety, the study developed by Martins [14] found that in Porto the signs for "Bolo" ("Cake") and "Amigo" ("Friend") are different from those in Lisbon. These examples demonstrate that, similarly to other sign languages, LGP has lexical variations across the country.

2.1 Portuguese Sign Language Grammar

Since there is still no official grammar, there is no consensus on various linguistic aspects, including the basic order or canonical order of sentences. Some consider that the basic sentence structure in LGP is Object - Subject - Verb (OSV) while others believe it is Subject - Verb - Object (SVO). Perhaps due to the linguistic challenges, the state-of-the-art regarding translation to LGP is still rather limited and the few computational works that exist [15–18], don't focus on linguistic components. These works only rely on a small set of manual rules and exclude facial expressions, which result in signed Portuguese (i.e., directly mapping a word into a sign), and not LGP.

2.2 Portuguese Sign Language Components

LGP is a language that takes advantage of three-dimensional space and possesses a grammatical structure as rich as any oral language. Similarly to oral languages, sign languages have their own: phonetics, phonology, syntax, semantics, morphology, and prosody. LGP and spoken/written Portuguese are different in all these aspects.

Unlike spoken languages, which combine sounds sequentially, LGP combines linguistic units simultaneously that consist of **manual** and **non-manual components** in order to produce meaning.

2.2.1 Manual Components

Manual components are those regarding hands, which include: hand configurations, orientations, locations, and movements. The phonology in LGP is characterized by the combination of these manual components with non-manual components which will be described in the next section.

Hand configuration refers to the shape that the hands assume, which may involve dactylology (i.e., manual alphabet). The hand configuration can remain the same or change throughout the execution of a sign. According to Patrícia do Carmo, there are 76 hand configurations in LGP [14].

Hand orientation refers to where the palm of the hand is turned to when the sign is performed (e.g., right, left, up, or down). The orientation of the hand helps to identify the meaning of the sign. For instance, the signs "Entrar" ("Enter") and "Sair" ("Leave") have opposite hand orientations, the former has the hand turned towards the body and the latter, the other way around [19]. The hand configurations and orientations characterize the internal movements in sign languages.

Hand location refers to the place where the configured hand performs the sign. This component is considered one of the main categories of sign language's phonology and can assume two characteristics: **contact point** (i.e., head, forehead, temples, eye, nose, cheek, ear, mouth, lower lip, chin, neck, shoulder, sternum, trunk, middle stem, abdomen, arm, forearm, and leg) and **contact mode** (high, low, contralateral, distal, proximal, and medial).

Hand movement refers to the direction or movement of the hands/fingers. This component can be analyzed taking into account the **type**: 1) movement variations of the hands, wrists, and forearms, 2) internal movement of the wrists or hands, 3) the movement of fingers; the **direction**: 1) unidirectional, 2) bidirectional and 3) multidirectional; **mode**: describes the quality, intensity, and speed; **frequency of the sign**: movements are simple or repeated. According to phonology, movements have lexical and morphological contrasts and there are two types of movements: external movement and internal/local movement. The external movement can be a straight, arched oblique or circular motion that changes the hand location, whereas the internal/local movement results from changing the configuration and orientation of the hand without changing location (e.g., opening and closing of the fingers, finger flicks, finger wiggles, hooking, twisting or rubbing, bending and extending of wrists) [14].

2.2.2 Non-Manual Components

Non-manual components correspond to body and face components without considering the hands. These include: shoulder, body and head movements, eye gaze and facial expressions. The facial expressions are suprasegmental variations that relate to various articulators such as eyebrows, eyes, cheeks, and lips, and can occur simultaneously or independently, performing one or more functions. While most phonological properties of signs relate to the articulation done by the manual components, facial expressions play an important role as distinctive phonological parameters for minimal pairs.

Minimal pairs refer to signs that only differ in one parameter from the five existing ones (i.e., hand configurations, hand orientation, hand location, hand movement, and non-manual component). Changing one of these parameters can change the entire meaning of a sign. For instance, the signs "Perder" ("Lose") and "Morrer" ("Die") have the same configuration, orientation, location and movement, and only

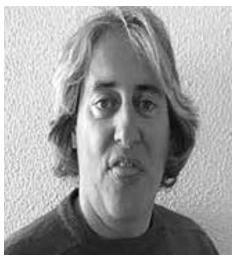
differ in the mouth gesture [20].

Facial and corporal expressions in Sign Languages are essential to convey feelings, similarly to any oral language, but are also used as morphological and syntactic parameters. Regarding **morphology**, facial expressions are used as markers for grammatical forms such as **adverbial**, **adjectival**, and **additive modifiers**.

An **adverbial** is a modifying term that independently expresses a circumstance (of place, time, mode, intensity, condition, among others) and performs the function of adjunct adverbial in a sentence. In LGP, the morphological flexion of a verb can be done, for instance, through the addition of non-manual adverbial expressions to the verb sign [21]. Analogous to oral languages, in LGP it is possible to detect different forms to produce verb tenses: Past, Present, and Future. The use of non-manual components has fundamental importance for each verb tense as these can be used to dictate the temporality in speech. For instance, during the negative sentence "Não há" ("There isn't"), the production of the phoneme "ua" can be used to represent the Present tense, as shown in Figure 2.1(a). The production of the "shh" phoneme (slightly closed mouth with lips puckered) can be used to represent the Future tense, as shown in Figure 2.1(b). The production of the "thh" phoneme can be used to represent the Past tense, as shown in Figure 2.1(c). The phoneme "va va" can also be used to represent Future tense, as well as non-manual components such as raised eyebrows, opened eyes, and the eye gaze directed to the space in front and away from the signer [21].

(a) Signer signing the phoneme "ua".

(b) Signer signing the phoneme "shh".



(c) Signer signing the phoneme "thh".

Figure 2.1: Mouth phonemes that produce verb tenses when executing the verb sign [1].

Facial expressions in LGP can also be used to mark **adverbials** and **adjectives** in terms of quantity (e.g., a lot, little) and quality (e.g., good, bad) distinctions. When referencing large quantities in LGP, that are not countable, a determinant must be added, for instance, "Trabalhar + muito" ("Work + a lot"). This sign can be performed by the same manual components as the sign "Trabalhar" ("Work") and by simultaneously performing the following facial expression: eyebrows frowned + eyes narrowed + small permanent air blow with puffed cheeks [20].

Regarding **additive modifiers**, facial expressions can be used to express the degrees of **augmen-**

tative and **diminutive** size in LGP. These indicate the intensity and the degree of size of a sign. The study developed by Gonçalves and Raposo [22] investigated the degree of augmentative and diminutive sizes used by 20 deaf adults from Azores and mainland Portugal fluent in LGP. They identified different expressions for the diminutive and augmentative degrees from both locations. These differences can be due to linguistic dialect with some lexicon variation, similarly to oral languages. To conclude, they identified the prevalence of signs dominating overall in both locations, with the expression "Mouth in the shape of a kiss and producing the phoneme "xinhoo" (which transcribes to the International Phonetic Alphabet (IPA) "ʃinu" in English) for the diminutive degree and expression "Higher teeth bites lower lip" for the augmentative degree. This investigation proved once again the importance that facial expressions have in the production of signs because removing this parameter would change the meaning of the sign completely.

2.3 Prosody

At a **syntactic level**, facial expressions acquire roles similar to the prosody of oral languages and are used as markers for sentence construction (i.e., negative, interrogative, and more). Sign Languages, thus, have prosodic systems that involve pragmatic, semantic, and syntactic information. Analogous to oral languages, LGP's prosody also refers to Intonation and Rhythm. Intonation consists of facial expressions portrayed by the face, eyes, eyebrows, head, and torso, and the Rhythm is described by the movement and pauses portrayed by the hands.

2.3.1 Intonation

Intonation is a fluctuation of the fundamental frequency curve at the sentence level that is responsible for the distinction of communicative and expressive intentions in Portuguese. In LGP, the intonation curve varies according to whether we want to express questions, exclamations, negatives, or even express doubts, certainties, and other reactions inherited from speech. In LGP, contrary to what happens in oral languages, suprasegmental variations relate to several facial and corporal articulators that can occur simultaneously or independently, performing one or several functions. In LGP, there are four different sentence types: **interrogative, exclamatory, declarative, and negative**.

There are two types of **interrogatives** in oral languages: **Yes/No questions**, also called polar questions, and **wh-questions**, also called content questions. The former refers to questions expecting a yes (affirmative) or no (negative) answers whereas the latter refers to questions that contain interrogative adverbials or pronouns such as "Aonde queres ir?" ("Where do you want to go?") or "O que compraste?" ("What did you buy?"). In LGP, a study developed by Cruz et al. [23] described the visual production of

yes-no (polar) questions as eyebrows frowned along with head nods which are different from most Sign Languages. Therefore, in LGP, the polar questions only differ from content questions in terms of head movements and not eyebrow movements.

Exclamatory sentences in LGP can be made from the combination of several expressions: opening of the mouth which slowly closes simultaneously with the eyes, and a movement of the torso and head backward which ends with a slight repeated movement of the head forward [24]. Furthermore, they tend to increase the intensity of the sign, thus, exclamatory sentences are signed faster.

Declarative sentences or **statements** in LGP are mainly produced with manual components without the use of non-manual components. Therefore, normally, signers have a neutral corporal and facial expression. However, there can be some head movements, mostly, up-down head nodding and some eyebrow-raising [23].

Negatives in LGP can be marked by two components: **1) manual components**, for example, the signs "Não" ("No") and "Nada" ("nothing") as shown in Figures 1 and 2 from the paper [1]; and **2) non-manual components**. The latter can be divided into: **1) an expression of negation** that refers to changes in facial expressions which result in negative form; and **2) a headshake** that refers to the lateral head movement.

In LGP, although facial expressions have an important role in most negative sentences, they may or may not be present, without compromising the grammar of negative sentences when not present. The study developed by Carmo et al. [1] demonstrated that facial expressions are only present and have fundamental importance in negative sentences with emotional nature and that the headshake is the most common negative marker. Two **manual signs** can be used for negative sentences: 1) the manual component **"Não" ("No")** which is the most common negative component, and 2) the manual component **"Não há" ("There is not")**. These manual components can be associated with **non-manual components** that: 1) dictate the **temporality of the speech**, as described in the adverbial subsection in Section 2.2.2, 2) are associated with a **negative marker** represented by the right cheek filled with air, or 3) are **intensifiers** (e.g., puffed cheeks) used to intensify the negation, for instance, when having a negative and exclamatory sentence.

Furthermore, there are two types of negative sentences: regular and irregular. **Regular negatives** are formed by adding one or more negative grammatical markers to a neutral sentence, without changing the morphological elements present in the sentence. The negative markers can be added simultaneously (synchronous) or simply at the end of the sentence (asynchronous). **Irregular negatives** are formed by reflecting the negation through a complete morphological change that derives from the affirmative form. An example is the verb "Querer" ("Want") and the verb "Saber" ("Know").

2.3.2 Rhythm

In Sign Languages, rhythmic structure is of great relevance to prosody and syntax interactions. **Rhythm** in Sign Languages is described by the movement and pauses portrayed by the hands. Some work regarding modeling timing and pausing parameters for manual components has been done for American Sign Language (ASL) [11, 12]. However, to the best of our knowledge, research in the area has not yet been published for LGP.

Although the timing and pauses parameters of manual components in LGP have not been studied, some investigation has been developed regarding the role of manual and non-manual components in prosodic connections.

In Sign Languages, the **prosodic connections** are not as easily identifiable or explicit as in written/spoken language, because these tend to be mostly undertaken by non-lexical elements. The study developed by Martins and Mata [25] analyzed the different sentence connectors used in LGP and based on the results concluded that most sentences (63%) used **non-manual components as prosodic connectors**. Based on their findings, there are three distinct prosodic connectors in LGP: **1) raised eyebrows** for conditional and inferential sentences accounting for 70% and 67%, respectively, and the remaining accounting for manual components; **2) frowned eyebrows** for contrastive sentences accounting for 34% and the remaining accounting for manual components; and **3) a neutral expression** for additive and temporal sentences accounting for 79% and 77%, respectively, and the remaining accounting for manual components. Furthermore, the most used manual connectors are “Mas” (“But”), “Se” (“If”), and “Então” (“Then”).

2.4 Annotated LGP Corpus

In Universidade Católica Portuguesa, an annotated corpus for Portuguese Sign Language is being developed by a six-member group composed of two linguists with knowledge of LGP, three LGP specialists (two of them deaf), and an interpreter. The corpus consists of 70 hours of videos of Portuguese deaf people from different age groups (from 10 to 60 years old), with social diversity and clear representation of the dialectal geography of LGP with signers from various parts of the country. This corpus contains formal, informal, spontaneous, and induced discourses. The annotations were made with the ELAN software, a tool that allows the creation of several layers of video and audio annotations which are aligned and synchronized temporally with the video or audio. In this corpus, the videos are translated into Portuguese, and the signs are transcribed into glosses, including their corresponding grammatical classes and the sentence constituents (subject and object). The annotation of glosses follows conventions to identify the grammatical information and the different linguistic phenomena.

3

State of the Art

Contents

In this section, we first explain two essential steps that must occur before being able to animate an avatar, then we explore some techniques that synthesize facial expressions, and finally, we explore the importance and synthesis of linguistic and secondary movements in Sign Languages.

Before being able to animate an avatar, two main steps are required: **Rigging** and **Skinning** [26]. **Rigging** is the process of setting the avatar's skeleton, which includes hierarchically linking bones, setting constrains on the joints, and creating control modes that the animator uses to move the joints. Rigging is what allows the motion of the animated avatar.

Skinning consists of binding the rigged model to the surface mesh. **Weight painting** is part of the skinning process, which consists of assigning the influence that each joint has on the mesh. Skinning creates a model that moves accurately with the rigged joints. Figure 3.1 describes the process of modeling an avatar.

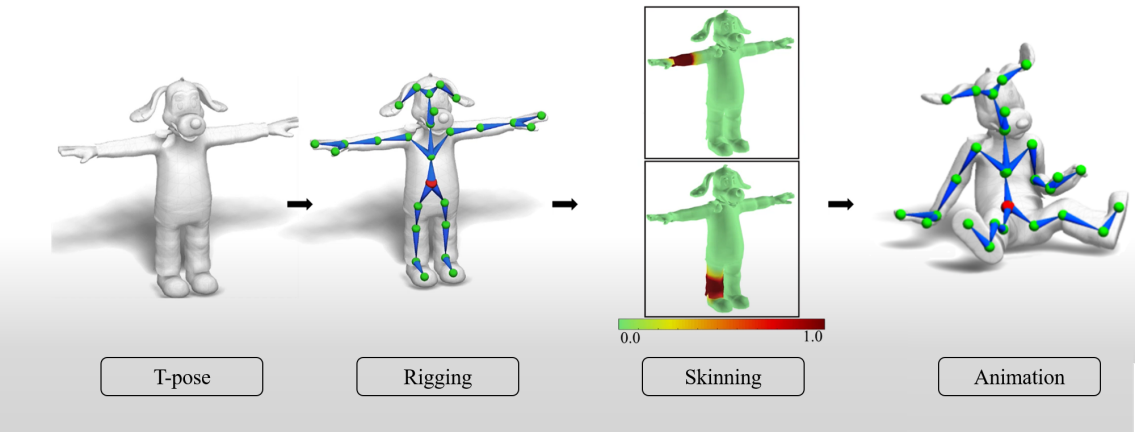


Figure 3.1: Modeling process [2].

3.1 Synthesis of Facial Expressions

For many years, facial modeling and animation have been a research focus and challenge. There are many approaches to synthesize facial expressions that can be categorized as: **Blend shape-based approaches**, **Simulation-based approaches**, and **Performance-driven approaches**. The following sections describe the most important and main approaches used for Facial Expression synthesis. For further explanations and other approaches, check these papers [27–30].

3.1.1 Blend Shape-based Approach

Blend Shape-based approaches [27, 28] are the most commonly used techniques in facial animations. A Blend-shape approach synthesizes facial expressions through the combination of a set of existing facial models. This approach involves blending different polygonal meshes of 3D face geometry known as morph targets or blend shapes to create human facial approximate expressions. Morph targets or blend shapes are a set of facial deformations applied to each frame of the animation in which each frame specifies the amount of each morph applied. The principle of this approach is that facial expressions are interpolated by specifying smooth motion between key-frames, over a normalized time interval [31].

Normally, linear interpolation [31] is often employed as it is simple, but other interpolations can also be used for higher quality animations. A cosine interpolation or spline interpolation can generate acceleration and deceleration effects between key-frames [32].

Many movies have used this technique, for instance, “Stuart Little” and “Star Wars”. An example of multiple blend shape facial expressions is shown in Figure 3.2.

Figure 3.2: Blend Shape Facial Expressions.[1]

Blend Shape Interpolations are easy and fast to synthesize facial animations, however, the ability to create a large number of facial expressions is restricted. The generation of highly detailed blend shape expressions requires animators to specifically adjust settings to each face model which can be very time-consuming.

3.1.2 Simulation-based Approach

Simulation-based approaches create synthetic facial expressions by employing simulated methods that mimic the contraction of facial bones/muscles. They require the specification of functionalities (i.e., their influence on the face) and locations of pseudo muscles such as muscles associated with mouth areas, eye areas, eyebrow areas, and more. Many multi-layer models [33–35] have simulated the anatom-

[1] https://www.artstation.com/artwork/WAJYN

ical structure of the human face, including skin, muscle, soft tissue, and more, to improve the visual realism of synthetic facial expressions.

One Simulation-based approach is **Parameterization** [29, 36]. This technique overcomes some of the restrictions and limitations of simple interpolations. Facial parameterization can be done manually, similar to rigging, in which each value has a pre-determined effect on a set of vertices that belong to a region of the geometric mesh. This way, a facial expression is created by defining a subset of vertices using the parameterization controls. Alternatively, parameterization could also be done automatically, by learning the weights of the extracted features and their deformations to create facial expressions [4].

Unlike simple interpolations, Parameterizations allow control of specific facial configurations, thus, it is more flexible. However, it has some drawbacks, for instance, there is no systematic way to blend expressions [37], which results in unnatural human expressions. These limitations led to the development of diverse techniques such as Pseudo-Muscle based approaches.

A **Pseudo Muscle approach** is the combination of multiple pseudo muscle contractions used to synthesize facial expressions [38]. Each pseudo muscle is a geometric deformation operator linked to a particular area in the face, as shown in Figure 3.3. These muscles, normally, influence either the lower or upper face. The lower muscles are responsible for the neck, chin, ears, and lips, and the upper muscles for the eyebrows and eyes. The synthesis of facial expressions is done by simulating the contractions of real muscles. For instance, the deformation of the mouth region is simulated by contracting the mouth sphincter muscle around the center of a simplified parametric ellipsoid.

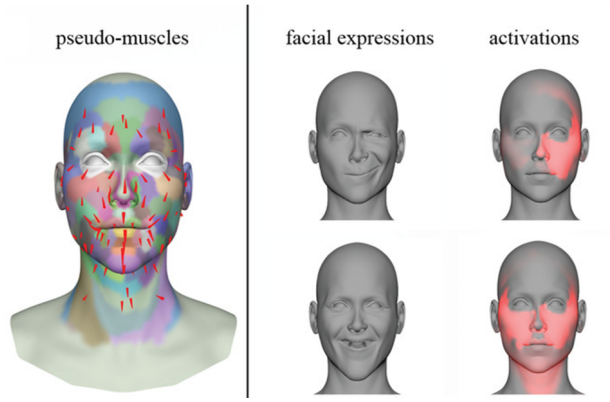


Figure 3.3: Synthesis of Facial Expressions using Pseudo-muscles [3].

This work uses the **Facial Action Coding System (FACS)** model to determine which muscles should be activated to generate specific facial movements. Pseudo muscles are classified according to their functionalities in terms of Action Units (AUs) which represent single or clusters of muscles that when combined describe a facial expression. For instance, Table 3.1 shows the combination used to create a sad expression. This muscle model is widely used because of its independence and compact representation of the facial structure. An example is the baby in the “Tin Toy” 1988 movie that uses 47 pseudo

muscles on his face.

Sad expression	
Action Unit 1	Inner Brow Raiser
Action Unit 4	Brow Lowerer
Action Unit 15	Lip Corner Depressor

Table 3.1: AUs combination for sad expression[2].

3.1.3 Performance-based Approach

Performance-based approaches create facial expressions by learning from recorded videos or by capturing facial movements using motion capture techniques and applying them to a synthetic face. Motion capture techniques are commonly used for Sign Languages not only for the study and analysis of facial and corporal movements but also for the synthesis of digital animations. These can be divided into two categories: **markerless** and **marker techniques**.

3.1.3.A Markerless Motion Capture

Markerless Motion Capture consists of affordable depth cameras that establish the relationship between a point on the image and the distance it is from the sensor. In other words, the depth sensor gives the Z coordinate from the three-dimensional space complementary to the X and Y coordinates obtained by the traditional RGB color cameras. These cameras opened doors to the investigation of algorithms capable of capturing body, hand, finger, and facial movements.

An example of a Markerless Motion Capture sensor is **Microsoft Kinect**. This system is equipped with an infrared sensor, an infrared light emitter, and four microphones that allow the recognition of the user's joints, and facial and voice recognition. It is capable of recognizing 20 joints of the human body and also supports the capture of facial expressions.

Kinect has been widely used to accurately replicate facial expressions into a synthetic facial model. The work developed in [39] captured facial motion in 3D using a Kinect v2 sensor and a Software Development Kit. As the sensor captured the real-time motion, facial movements were reconstructed into a motion tracking software called *FaceShift*. Using *FaceShift*, actors elicited 23 training facial expressions and then these were scanned and a personalized avatar was created. A total of forty-eight blend shape parameters were tracked which exhibited both non-rigid (i.e., expression changes) and rigid (i.e., head translations and rotations) motion patterns, as well as eye-gaze movements, natural speech movements, and emotions of all magnitude. Based on the analysis of the results obtained, participants were able

[2] https://imotions.com/blog/facial-action-coding-system/

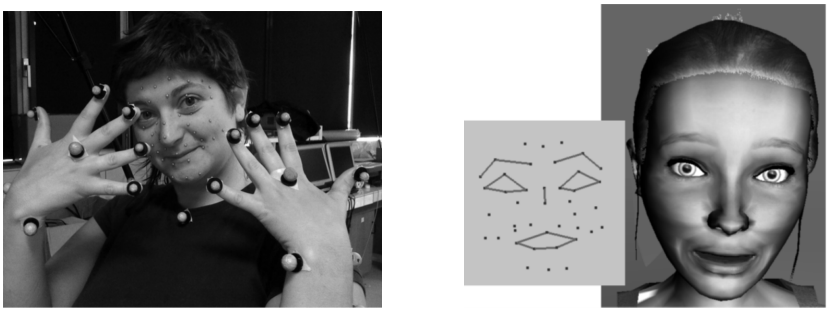
to identify different individual faces and distinguish between different facial animation videos. This work shows that markerless motion capture technology can accurately generate realistic facial motion stimuli.

3.1.3.B Marker Motion Capture

Marker Motion Capture is used in a wide range of areas, for instance, medical, military, and entertainment. Many movies and games have also used motion capture to accurately synthesize facial and corporal movements such as the movie “Avatar” and the video games “Guitar Hero” and “Grand Theft Auto IV”. This type of technology is capable of capturing the position of specific body parts, even complex hand, finger, and facial movements. Motion Capture is a complex process composed of several sequential steps: (1) studio Preparation, (2) calibration of sensors and capture volume, (3) capture or recording the movement, (4) data processing, and (5) application of data to bi-dimensional or three-dimensional spaces.

This approach is commonly used for Sign Language animations and uses data obtained from different data sources that drive the avatar’s skeleton. Markers on the signer’s body and gloves can be potential sources of data. However, finger tracking and modeling using motion capture is complex and costly due to the subtle finger movements in Sign Language and expensive equipment. The SignCom project developed by Gibet et al. [4] presented an animation system that produced French Sign Language by a virtual character. It used decomposed motion capture data for each different body part (e.g., torso, arms, hands, facial features, and head) from human signers. The facial expressions were synthesized by mapping 43 facial markers that resulted in 123 features when considering the 3D space of the marker’s values as shown in Figure 3.4(a). These features were mapped to 50 blend shape values in the avatar’s geometrical model considering probabilistic inference and by learning the corresponding blend shape weights using a Gaussian Process Regression. Figure 3.4(b) shows an example of one blend shape that was generated.

Motion Capture can improve the visual realism of synthetic facial expressions that often lack in other approaches, in particular, the synchronization of complex body and facial behaviors. Therefore, the main advantage of this technique is the realistic result that brings to animations in comparison to manual approaches such as Blend shape-based and Simulation-based approaches. However, this approach has some limitations because the data obtained from facial performances are peculiar to the signer’s facial movement style and often the data captured needs to be filtered so that it is accurate and does not contain noise. Even though many motion capture cameras are used, marker-occlusion may occur which limits the quality of the retrieved data. This approach is also costly since it requires expensive equipment, software, and specialized personnel, and sometimes lacks in accuracy which results in time-consuming manual corrections.



(a) Native signer with motion capture sensors on her hands and face.

(b) Facial expression generated in avatar with corresponding markers' positions in 2D space.

Figure 3.4: Motion Capture Technique [4].

3.1.4 Discussion

The synthesis of facial expressions is still one of the biggest challenges in computer graphics because of the subtle emotions and dynamic properties that faces have.

Blend Shape-based and Simulation-based approaches are commonly used in animation movies to create facial expressions on characters' faces. Normally, these are exaggerated to create a comical or to emphasize the character's emotions which, for an animation movie, is ideal and still believable. However, trying to achieve realistic human facial animations is harder to generate than a fantasy 3D character due to the curves and complex construction of human faces. To achieve realistic animations, animators must be careful around the edges of facial regions and specifically adjust settings to each face model in order to generate highly detailed and natural animations. Performance-based approaches are the ones with the most potential for achieving visual realism since these learn facial and corporal movements from real performance, thus, offering an animation as close to the real world as possible.

The difference between Blend shape-based and Simulation-based approaches is that the former needs to create many blend shape poses so that they can be smoothly blended, while the latter, requires a bone/muscle structure for the entire face and then animators can control these structures to generate animations. Simulation-based approaches provide more freedom and control when generating facial expressions as opposed to blend shaped-approaches. The greatest advantage of Simulation-based approaches is that the animation can be applied to different characters, as long as the facial physiology is similar. On the contrary, blend shape-based approaches present a higher level of fidelity in facial expressions compared to Simulation-based approaches, however, these techniques involve much more manual work to create morph targets and these are only specific to a certain character.

Performance-based approaches synthesize facial expressions by mapping the data acquired to the

face model which is difficult if the sizes do not match. It is also a time-consuming approach for recording and synthesizing facial expressions for a big corpus since each facial expression must be recorded separately. Blend Shape-based approaches provide an advantage if the set of facial expressions is small and repeatable because once these are synthesized they can be used multiple times in an avatar for different animations, and the creation of new blend shapes can even be facilitated by copying existing similar blend shapes and making the necessary changes. Therefore, once the blend shape facial expressions are created, the interpolated animation is much easier and faster than Simulation-based and Performance-based approaches.

Table 3.2 provides a performance summary of the comparison between the three major approaches described.

	Animation Natural Flow	**Cost**	**Equipment setup**	**Flexibility**	**Animation Synthesis Difficulty**
Blend Shape	*	*	*	*	*
Simulation	*	*	*	**	***
Markerless Motion Capture	**	**	**	***	**
Marker Motion Capture	***	***	***	***	**

Table 3.2: Qualitative evaluation of the different approaches. *Low, **Medium, ***High.

A study developed by Adamo-Villani [9] aimed at determining the most effective animation technique for ASL in terms of accuracy, readability, and closeness to signing. Twenty animated clips of ten finger-spelled words were produced using a Keyframe Animation and a Motion Capture Animation (10 Keyframed Animation + 10 Motion Capture Animation). A Keyframe Animation consists of manual approaches such as Blend Shape and Simulation approaches where animators set values to various objects' parameters (e.g., rotations of fingers or position of hands) and then interpolate these values between key frames. The realism of Keyframe Animations depends on the animator's ability to create believable key frames and then control the interpolation between them. Both animation techniques were applied to the same 3D avatar and then 71 subjects participated in the study.

Surprisingly, based on the results gathered, the Keyframe technique produced the most accurate and legible animations that were the closest to real signing. The reasons behind this are: 1) the data acquired using Motion Capture can be imprecise due to the differences between the signer's hands and the gloves or inaccuracies with the system calibration; 2) Motion Capture systems capture secondary body movements or body jitters which may distract the viewer's attention from the signing motion; 3) These systems captured the nuances of the signer's style which is not necessarily understandable for

all viewers.

The advantages and disadvantages of each approach led to the development of hybrid techniques that combine multiple approaches. Currently, there are many mixed and grouped techniques as well as many animation systems using multiple forms of these techniques. For instance, the accurate timing and motion information gathered by performance-based techniques, either by recorded videos or motion capture data, can be used to create facial animations by employing underlying blend shapes [40] and muscle structures [41]. A new approach [42] was also developed that integrated a Blend Shape Interpolation with FACS to create expressive and realistic facial animations. Furthermore, the movie "The Lord of the Rings" used a hybrid approach that combined motion capture with FACS and a keyframe animation to bring a new expressive and realistic leverage point to facial animation. This hybrid approach was also used in the movies "Monster House" and "King Kong".

Conclusion. After reviewing and discussing the related work we highlight the following focal points for our proposal: (1) importance of balancing the advantages and disadvantages of the different techniques for the synthesis of facial expressions, (2) the correlation between quality and cost for facial expression synthesis, and (3) opportunity of leveraging a combination of multiple techniques.

3.2 Animation in signing avatars

In Sign Languages, movements can greatly impact the signing quality and the way the thought or feeling is conveyed. There are two types of movements that have been widely adopted in sign animations, from now on we will call them: **linguistic movements** and **secondary movements**.

3.2.1 Linguistic Movements

Linguistic movements refer to those that are used in a phonological, syntactic, and morphological grammatical level for manual and non-manual components in Sign Languages, as described in Section 2.2. Regarding non-manual components, syntactic non-manual components determine the sentence type (i.e., declarative, exclamatory, interrogative, affirmative, and negative) and morphological non-manual components indicate the grammatical modifiers such as adverbials, adjectives, and additives.

Some work in the area of generating facial non-manual components in avatars has been done for ASL. However, to the best of our knowledge, research in the area has not yet been published for LGP. The work developed by Schnepp et al. [43] presented a method designed for generating co-occurring non-manual components in ASL based on linguistic processes rather than just a series of facial poses. This approach maps linguistic processes to anatomical movements with timing and intensity information

that controls the animation. To account for co-occurrence, they computed a matrix that combines the weighted transformation for each track that simultaneously influence the face. For instance, given a sequence of glosses, the synthesizer would automatically create an initial animation draft that displays syntactical linguistic information, lexical modifiers, as well as emotions and mouthing.

3.2.2 Secondary Movements

Secondary movements represent those that are added to improve the naturalness of the avatar and are not part of the morphosyntactic structure of Sign Languages. These include: eye blink, mouthing, and facial and corporal movements.

3.2.2.A Mouthing

Mouthing refers to the production of visual morphemes or syllables that derive from spoken language. Some believe that mouthings are incorporated into the morphosyntactic structures of Sign Languages and some believe they are not [44, 45].

The work developed by Crasborn et al. [44] studied the mouth actions from a cross-linguistic perspective for three European Sign Languages: Sign Language of the Netherlands (NGT), British Sign Language (BSL), and Swedish Sign Language (SSL). Based on the results gathered, Mouthing is the category with the largest mouth actions in all three languages, accounting for 57% in SSL, 51% in BSL, and 39% in NGT. It is also interesting to note that mouthings can also be combined with manual signs to create complex signs with a composite meaning. For example, the manual sign “mouse” in BSL can be accompanied with the mouthing “baby”, forming the composite meaning “baby mouse”.

In the study, they also confirmed the hypothesis that mouthings spread in an analogous way to ‘native’ mouth gestures, thus, mouthings have indeed a grammatical function in Sign Languages. Based on all results gathered from the study, mouthing occurs for all three Sign Languages and without it a signing avatar would look unnatural and could omit important information, thus, resulting in incomprehensible utterances. Therefore, we can conclude that an avatar capable of producing mouthing is an essential part of any automatic written/spoken to sign translation system.

Mouthing or lip-sync appeared first in the 1920s with the advent of sound cartoons. Speech can be discretized as a sequence of sounds also known as phonemes. Each phoneme is associated with a facial pose, however, not all vocal articulations are visible and some are irrelevant in the visual domain, for instance, nasality and voicing. Phonemes usually have many-to-one relationships with visemes (i.e., facial and oral poses of phonemes) because different phonemes can have the same facial pose. Normally, animators use between 7 to 12 visemes to represent the 33 phonemes that exist in the Portuguese language, in which fourteen are vowel phonemes and nineteen consonant phonemes.[3] It is also impor-

[3] https://www.youtube.com/watch?v=pMhHlfZqAYY

tant to note that visual speech cannot be directly generated by concatenating visemes, because it will over-articulate the produced animation.

Mouthing animations can be produced manually or automatically [45]. A manual approach requires animators to draw each viseme by hand and later use an interpolation scheme that concatenates the visemes according to the animated utterances. This method is a time-consuming process in which animators are responsible for the viseme selection and timing. These limitations led to the development of automatic techniques that synchronize audio with visemes. In automation approaches, visemes are collections of 3D data and artists can rely on muscle-based systems or blend shapes expressed as polygon meshes to model avatar's lip positions to depict visemes.

Automated techniques depend on the source of dialog to generate animation. If it is a pre-recorded voice track, a speech recognition system must be used, otherwise, if it is a text containing a dialog, a text to speech system must be used. Both techniques require the same process: detecting the phonemes and then selecting the corresponding visemes that can be interpolated between keyframes in the avatar. No matter the technique, the best mapping between phonemes and visemes is still a debatable issue. Many studies have been developed to understand the best Phoneme-Viseme mappings. This study [46] examined 120 mappings and analyzed their effect on visual lip reading using hidden Markov model (HMM) recognizers. Based on the results, they concluded that Lee's mapping performed the best for both vowels and consonants and that the most common viseme in all mappings is [/p/ /b/ /m/]. Although phoneme-viseme mappings are not universal among languages or within a language, some phoneme-viseme mappings have overlapping sets. For instance, similarly to English, Amazon Polly's Phoneme-Viseme mapping[4] for European Portuguese also contains the viseme [/p/ /b/ /m/] as a set, even though these are two different languages.

Some projects have been exploring the possibility of incorporating mouthings in Sign Language animations. The ViSiCAST project [8] developed the Signing Gesture Markup Language (SiGML) which is an XML-compliant representation of signs based on HamNoSys. In this project, phonemes are described based on the International Phonetic Alphabet (IPA) transcription and then visemes are mapped using the Speech Assessment Methods Phonetic Alphabet (SAMPA) encoding conventions [47]. Some work in the area of visual speech animation has been done for ASL and Swiss German Sign Language [45]. However, to the best of our knowledge, research in the field has not yet been published for LGP but research for European Portuguese has been developed.

The work developed by Serra et al. [48] is the first automatic visual speech system for European Portuguese based on viseme concatenations. This project used two phoneme-viseme mappings: one mapping with 14 different viseme classes and another mapping with 10 different viseme classes. Both Phoneme-Viseme mappings resulted in slightly different vowel classifications, but the number of vocalic

[4] https://docs.aws.amazon.com/polly/latest/dg/ph-table-portuguese.html

viseme classes remained unchanged. Each viseme class was then created in the avatar by an experienced digital artist. After creating the mappings and the visemes, the system was divided into two main components: a speech processing component and a 3D animation engine. The speech process component processed the data (e.g., text, audio, or both) and obtained the phonetic transcriptions using an EP phonetic lexicon developed by Microsoft together with Microsoft Speech API (SAPI)[5] as an automatic speech recognition (ASR) model. The SAPI system used recognition events to detect the different utterances from the audio and stored a list of words and their corresponding IPA formats. The SAPI does not provide the phonemes' duration and timing, therefore, the EP phonemes' duration were gathered from a database of 100 hours of Portuguese speech provided by Microsoft. The 3D animation engine encapsulated the data obtained from the speech process component and translated it into the 3D animation. The cartoon character relied on a bone-based rig and each viseme was interpolated using the timing obtained by the speech process component and the animation curves defined by the animator.

3.2.2.B Facial and Corporal Movements

In real life, no part of the human face and body is truly stationary, therefore, an avatar without the subtle motions of humans can appear highly robotic. Secondary facial and corporal movements are extremely important since these determine the naturalness of the avatar.

The work explored in [49], addressed two of the many reasons for robotic motion in ASL animations. The first one is the lack of spine motion as the avatar's arms move. To account for this, they developed a model that cues the spinal motion by the reaching of the arm, meaning that, greater the distance of the hand reach, the more spine will be bent to assist the arm movement. This system automatically rotates and bends the torso of an avatar by computing angles based on the targeted movement of the hand and the position of the shoulders. The second reason is the lack of motion in held joints of the shoulder. A human shoulder is composed of two articulations: the sternoclavicular and the acromioclavicular joints. These are simulated in an avatar by using a rotational joint placed at the intersection of the neck and the spine. The anteroposterior rotation allows the shoulders to move forward and backward, whereas, the vertical rotation moves the shoulders up and down.

Based on the analysis of the user tests gathered, their system has the potential to increase the naturalness of the avatar with results particularly strong for clarity and understandability measures. Overall, the naturalness ratings were lower than for the other measures, which may be due to naturalness being the most demanding criterion.

[5] https://docs.microsoft.com/en-us/previous-versions/windows/desktop/ms723627(v=vs.85)

3.2.3 Discussion

The synthesis of Sign Language animations in real-time, for instance, an automatic written/spoken to sign translation system, differs from the kind of animation used in the film industry. In a film, animations are scripted and refined until they are good enough, whereas a sign translation system requires animations to be automatically synthesized. Automatically synthesizing signing animations is an extremely difficult task, because signing avatars must account not only for multiple co-occurring linguistic processes but also the naturalness of the movements.

Although motion capture approaches provide realistic results to animations, these are not useful for automatically synthesizing facial animations. The reason behind this is that these approaches map captured data with a specific duration, for instance, the eyebrow movements for questions, normally, occur during the whole sentence and these approaches would capture the data during this specific sentence. This creates a limitation because sentences with different lengths will have different duration, which these approaches do not account for.

In an automatic signing system, the separation of linguistic movements from secondary movements is absolutely critical if the animations are to be used for linguistic testing, analysis, and verification but also if the synthesized signs must change according to morphological rules. An automatic signing system should incorporate both movements. Linguistic movements to determine the morphosyntactic motions needed in Sign Languages and secondary movements to determine the naturalness of the avatar. Therefore, the goal of an automatic signing translator is to infer secondary movements based on human kinematics as much as possible that adhere to the linguistic movements so that animations are understandable, realistic, and natural.

Conclusion. After reviewing and discussing the related work we highlight the following focal points for our proposal: (1) necessity of an automatic written/spoken to sign translation system that incorporates both linguistic and secondary movements, (2) importance of a flexible and dynamic facial animation approach, and (3) necessity of a separation between linguistic and secondary movements for evaluation purposes.

4

Creation of Signs and Facial Expressions

Contents

The core of our system that generates Sign Language animations (Chapter 5) is the transitions between individual signs that are synthesized in a database, therefore, the creation of signs and the process of continuously feeding the sign database is extremely important. The process of creating signs is divided into three modules: the Hand Pose Editor, the Facial Expression Editor, and the Sign Editor. The Hand Pose Editor (Section 4.1) allows users to create and modify hand configurations that are used in the Sign Editor. The Facial Expression Editor (Section 4.2) allows users to create phonological and syntactic facial expressions that are used in the Sign Editor and in the Translator (Section 5.3). The Sign Editor (Section 4.3) allows users to create new signs and modify existing ones that are used in the Translator (Chapter 5.3). The Hand Pose Editor, the phonological facial expressions, and the Sign Editor were created by Pedro Cabral, a member of our team. In this dissertation, I had the chance to work with all modules by creating new hand configurations with the Hand Pose Editor, creating new phonological facial expressions and all syntactic facial expressions with the Facial Expression Editor, and creating and modifying signs with the Sign Editor.

4.1 Hand Pose Editor

As shown in the following video[1], the Hand Pose Editor allows users to select each finger and modify its position in multiple ranges of motion (e.g., distal, mid, abduction, and opposition). Currently, there are 151 configurations created (variations of configurations are also included) following a phonetic table. This table is organized based on the number of fingers selected, the fingers' position (i.e., extended, flattened, bent, hooked), and the thumb opposition (i.e., open, semi-open, semi-closed, closed).

As described in Section 2.2.1, signs can have external and internal movements. The Hand Pose Editor poses a limitation for internal movements since it only allows the creation of static poses. To mitigate this problem, new configurations were created that are not official configurations according to phonology, but rather, variations of existing configurations (e.g., the configuration representing the fingers opening in the signs "sixteen" and "nineteen", as shown in figure 4.1). Moreover, animations were created using Unity's animation features that contain the combination of multiple configurations for some internal movements (e.g., finger wiggles).

4.2 Facial Expression Editor

Three different techniques were described in Section 3.1 for the synthesis of facial expressions. Based on the previous analysis and discussion, our approach consists of a combination of two approaches: A performance-based approach and a blend shape-based approach. A performance-based approach

[1] https://youtu.be/ZOGpjKddG4U

Hand configuration that represents the fingers closing.

Variation of a hand configuration that represents the fingers opening.

Figure 4.1: Hand configurations for the internal movement in the “sixteen” sign.

was used to study and analyze our annotated LGP corpus (Section 2.4), as well as use it as reference footage to create facial expressions and body movements as realistic as possible. A blend shape-based approach was used for the synthesis of facial expressions. Therefore, the main approach used for the synthesis was a blend shape approach, not only because the modeled avatar already contained several blend shapes implemented, but also because these are static and can be interpolated with the correct timing and duration values.

The avatar contains 39 blend shapes for the face which include eyebrows, eyes, mouth, and cheeks movements, 9 blend shapes for the tongue, and 3 blend shapes for the hair. The blend shapes for the hair are important because, for instance, when the cheeks are puffed, the hair goes inside the cheeks which is not realistic, therefore, when the cheeks are puffed or interfere with the hair, we need to use a hair blend shape as well. Some blend shapes are shown in Figure 4.2. The avatar and its blend shapes were created in Autodesk 3ds Max[2] by artist Denys Almaral[3].

These blend shapes were used, alongside Unity’s animator, to create phonological and syntactical facial expressions. Phonological facial expressions refer to those that are incorporated in a sign and change its entire meaning, whereas syntactical facial expressions refer to those that are used as markers for sentence construction. To create phonological and syntactic facial expressions and movements as realistic as possible, we used reference footage from LGP native signers. Pedro created sixty animations for the phonological facial expressions and added them in the Sign Editor (Section 4.3) so that these could be used to create signs that incorporate one or multiple facial expressions.

On a syntactical level, facial expressions can combine multiple blend shapes and incorporate shoulder, body, and head movements. I created facial expressions for interrogatives and negatives using reference footage from LGP native signers. As described in Section 2.3.1, there are two types of interrogatives: polar and content questions. According to previous studies, these two interrogatives have

[2] https://www.autodesk.pt/products/3ds-max/overview
[3] https://denysalmaral.com/

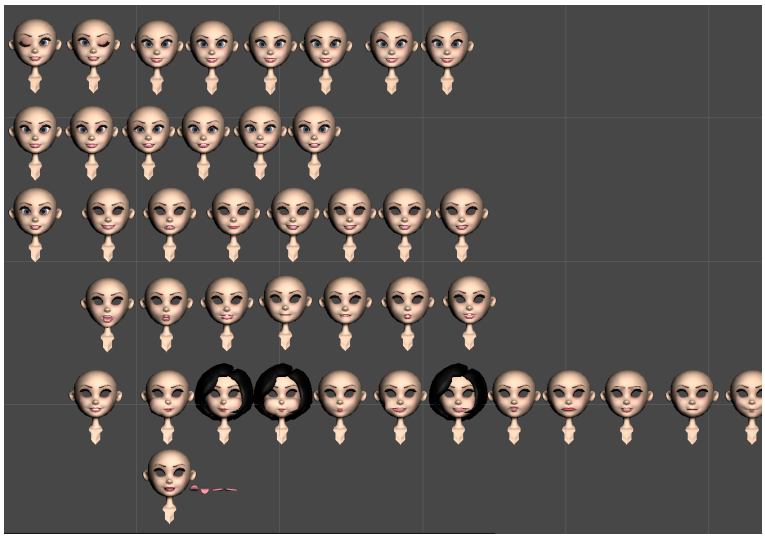


Figure 4.2: Some blend shape targets already implemented.

different facial expressions and a study developed by Cruz et al. [23] determined that, in LGP, polar questions only differ from content questions in terms of head movements and not eyebrow movements. After analyzing several videos of deaf native LGP signers, we can confirm that polar and content questions have indeed the same facial expression but different head and body movements. Polar questions (Figure 4.3(a)) have a slightly forward upper body and head tilt and content questions (Figure 4.3(b)) have an upward head movement without body tilt. Additionally, both questions have frowned eyebrows, narrowed eyes, and shoulders upward movement, as shown in Figure 4.3.

(a) Facial expression for polar questions.

(b) Facial expression for content questions.

Figure 4.3: Facial expressions for interrogatives.

As described in Section 2.3.1, there are two types of negatives: regular and irregular. After analyzing our LGP annotated corpus and several videos of deaf native LGP signers, we have noticed that regular negative is, normally, formed by adding the "Não" ("No") manual component after the negated verb, without changing its morphological elements. Irregular negative, on the other hand, is formed by reflecting the negation through a complete morphological change that derives the verb sign from its affirmative form. Additionally, both negatives are also accompanied by a headshake and an expression of negation (i.e., eyebrows frowned and eyes slightly narrowed). Furthermore, since the irregular negative does not contain the "Não" ("No") manual sign, normally, to reinforce the negation, negated verbs also incorporate negative markers (e.g., right cheek filled with air, for instance, when the sign "poder" ("can") is negated) or intensifiers (e.g., puffed cheeks, for instance, when the sign "saber" ("know") is negated). We created the sign "Não" that is used in regular negatives, some irregular negatives for the verbs "Querer" ("Want"), "Saber" ("Know"), "Haver" ("There is/are"), and "Ter" ("to Have"), and the negation adverb "Ainda não" ("Not yet").

4.3 Sign Editor

The Sign Editor component uses pre-made hand poses created with the Hand Pose Editor (Section 4.1) and pre-made phonological facial expressions created with the Facial Expression Editor (Section 4.2). In the Sign Editor, users can select hand configurations for the right and the left hands and select phonological facial expressions. Furthermore, Forward Kinematics (FK) is used by the system to allow users to rotate the avatar's joints and Inverse Kinematics (IK) is used to move the avatar's joints. Users can move and rotate the avatar's neck, wrists, elbows, and shoulders, and must define key poses to create a sign. The Sign Editor uses a Key-frame approach in which signs are animations that consist of one or several key poses throughout a time span that can be adjusted using the timeline tool. These key poses are interpolated and all in-between frames are automatically generated to create animations. This video[4] shows a sign being created by rotating and moving the avatar's joints and setting several key poses.

The linear interpolation between keyframes creates abrupt and unnatural changes in velocity which leads to a robotic motion that is extremely noticeable especially in circular motions. To improve the naturalness of these movements, Pedro and I implemented smooth tangents for each keyframe by making the final smooth slope an average of the in and out tangents. This way we replaced linear animation curves with smooth animation curves that make more natural movements. The difference between linear animation curves and smooth animation curves can be seen in this video[5].

After a sign is created and saved, the Sign Editor generates two files: (1) An *.anim* file that represents

[4] https://youtu.be/A7igDESK73w
[5] https://youtu.be/eTw-BDogBgQ

the Animation Clip structure in Unity. This file contains animation properties such as rotations, the timeline, and keyframes that specify anchor points of the animation, (2) A *JSON* file that stores all relevant information such as the facial expressions and hand locations in each frame. *JSON* files are much faster to read than *.anim* files which makes *JSON* files extremely useful in the Translator (Section 5.3.2.B).

As described in the previous section, an irregular negative is formed by a complete morphological change of the sign from its affirmative form. Therefore, signs that have an irregular negative were also created using the Editor and saved as "Não_verbo" ("No_verb"). For our translation system to work completely, we have to ensure that the Editor, Translation, and Animation processes are all following the same naming scheme. Therefore, in the translation process (Section 5.2.2), the glosses that have an irregular negative need to have the same name as the ones in the Editor. The main goal of the Sign Editor is to create an animation database that can be used by the Translator and the Dictionary component also created for this project. The dictionary component displays all signs stored in our database and contains a bilingual search by allowing users to search signs through text input or by selecting hand configurations.

5

Synthesis of Sign Language Animations

Contents

Following the conclusions taken from the literature review in Chapter 3, our approach consists of the synthesis and animation of manual and non-manual components, and secondary movements in the already modeled 3D avatar. This approach provides a pipeline for the synthesis of LGP animations that can be used for multiple digital applications. In this dissertation, we used a text-to-sign language translator to demonstrate our generated animations. To the best of our knowledge, this is the first automatic Portuguese to LGP translator that contains manual and non-manual components based on linguistic information. This system is divided into two main modules, as shown in Figure 5.1. The first module, Translation Process (Section 5.2), consists of the translation of text from Portuguese to LGP, in which the LGP sentence is represented by a sequence of glosses and additional morphosyntactic information. The second module, Animation Process (Section 5.3), consists of an avatar that animates the LGP translated message received from the first module. The communication between these two modules is described in the next section (Section 5.1).

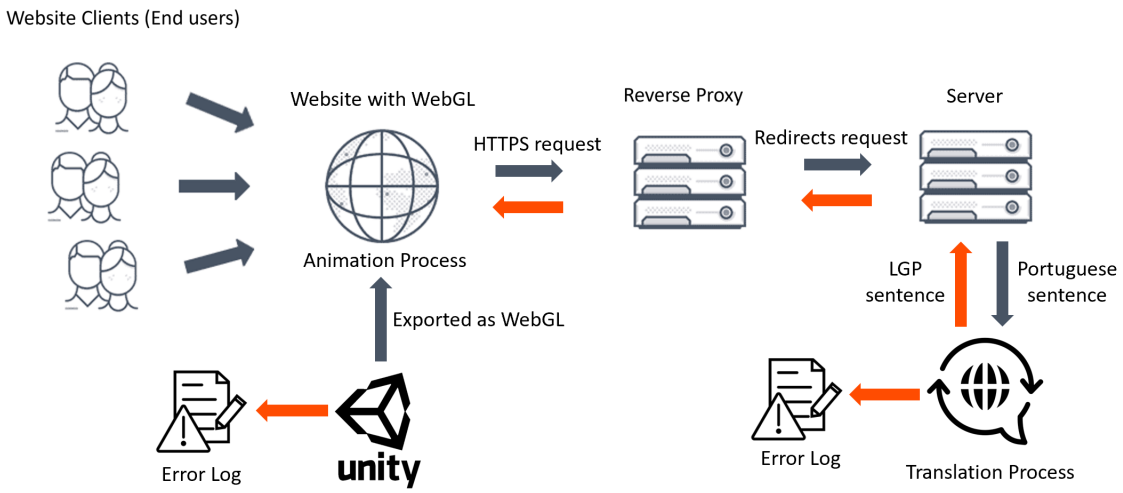


Figure 5.1: Overall architecture of the text-to-sign translator.

5.1 Communication

An automatic written-to-sign translation system requires two components: a translator and an avatar. These two components are the core of an automatic sign translation system and must be connected. The translator (Section 5.2) was written in Python3 and the animation (Section 5.3) produced by the avatar was written in C# in Unity. As far as we know there is no tool capable of running a custom Python3 environment in Unity, therefore, we used a synchronous TCP communication to connect both components. Data is sent and received over Stream sockets in blocking mode so that the message sent

is received and returned correctly. To account for parallel messages, we implemented threads in which each thread responds to one request, allowing a parallel communication. This solution worked properly while playing the animation in Unity and after exporting the project as an executable file. However, for the Unity project to be deployed on a website, it must be exported to WebGL.

WebGL[1] uses web standards that publish Unity content as Javascript programs. Due to constraints in the platform, however, not all features of Unity are available in WebGL builds. Security concerns caused WebGL to not have direct access to IP sockets to implement network connectivity; thus, the .NET networking classes are not functional in WebGL[2] and neither is the previous implemented solution. As a result, we implemented a different solution for the communication: **a Restful API** where there is a server connected to the Translator process and clients connected to the Unity application (Figure 5.1). Restful API is an architectural style for web services that uses HTTP requests to access data and defines a set of constraints to be used in the communication.

As shown in Figure 5.1, the overall architecture of the text-to-sign language translator consists of Unity being exported to WebGL and hosted in a website that is accessed by users. Users write a Portuguese text that is sent to a Reverse Proxy, which in turn, redirects the request to a server connected to the translation process. The sentence is translated, the server sends it back to the reverse proxy which redirects it to the website where users can visualize the corresponding animation. Additionally, the system also reproduces error logs for both processes which contain descriptions of errors that occur during run-time.

On the **client** side, the only classes in Unity that are supported in WebGL and allow Networking are the *WWW* or the *UnityWebRequest*. For this reason, we decided to create a Web Request object using the **UnityWebRequest**[3] class that allows Unity projects to compose HTTP requests and handle HTTP responses. However, this class uses the *XMLHttpRequest* class in JavaScript that handles WWW requests through the browser. As a result, there are security restrictions on accessing cross-domain resources, in other words, web applications using *XMLHttpRequest* or Fetch API can only send and receive HTTP requests to/from a server hosted in the same domain.

The server was implemented as an **asynchronous HTTP server** using the ***aiohttp* framework**[4] that concurrently handles hundreds of requests per second. This server is deployed in a virtual machine running Ubuntu20 and is hosted in the INESC-ID's Human Language Technology (HLT) server. Apache2, the cross-platform web server running on HLT is used as a reverse proxy that receives HTTPS requests coming from the Unity application and redirects them to the server running on the virtual machine. Essentially, the proxy server acts as an intermediary between Unity and the virtual machine, as shown in Figure 5.1.

[1] https://docs.unity3d.com/Manual/webgl-gettingstarted.html
[2] https://docs.unity3d.com/Manual/webgl-networking.html
[3] https://docs.unity3d.com/Manual/UnityWebRequest.html
[4] https://docs.aiohttp.org/en/stable/

Since the server and the client are hosted in different domains and due to the security restrictions mentioned previously, the server must configure Cross-Origin Resource Sharing (CORS)[5] so that clients can send and receive HTTP requests. CORS is an HTTP-header-based mechanism that supports secure cross-origin requests and data transfers between servers and browsers in different domains by letting servers describe which origins are permitted. To set up CORS in the server, first, we configured the server application with aiohttp, then we stored the CORS configuration for the application with the *aiohttp_cors* library[6], and lastly, we enabled CORS for all routes defined. We specified a GET route that handles GET requests from clients to check if the server is functional and a POST route that handles POST requests containing a Portuguese sentence and returns the corresponding translation done in the Translation Process (Section 5.2).

5.2 Translation Process

The Translation Process was developed by Matilde Gonçalves in a previous thesis [10, 50]. This translation system is divided into two main modules, as shown in Figure 5.2. The first module, the Translation Rules Construction, consists in extracting linguistic information from our annotated LGP corpus, and based on this information, creating translation rules and a bilingual dictionary of Portuguese and LGP. We wanted to extend the previous system by creating more translation rules but, unfortunately, it was not possible to gather new data from the corpus because the newer parts of our corpus did not have the necessary annotations (i.e., the definition of each sentence constituent). The second module, the Machine Translation, consists in the translation of text from Portuguese to LGP, in which the LGP sentence is represented by a sequence of glosses with markers that identify facial expressions and fingerspelled words. This translation system is based on the translation rules and the bilingual dictionary created in the first module, and also manual rules that capture linguistic phenomena related to morphology, such as feminine forms and facial expressions. We extended the already implemented system to account for additional linguistic processes and morphosyntactic components. The most relevant and significant changes will be described.

5.2.1 Pre-processing Phase

In this phase, Portuguese sentences undergo a morphosyntactic analysis using the Freeling tool [51] and a syntactic analysis using SpaCy [52]. The Freeling tool identifies grammatical classes and subclasses (possessive determiners, demonstrative determiners, etc.), as well as aspects of inflection (in gender, number, tense and mood, etc.), and lemmas of words in Portuguese sentences (and of signs in LGP).

[5] https://developer.mozilla.org/en-US/docs/Web/HTTP/CORS
[6] https://github.com/aio-libs/aiohttp-cors

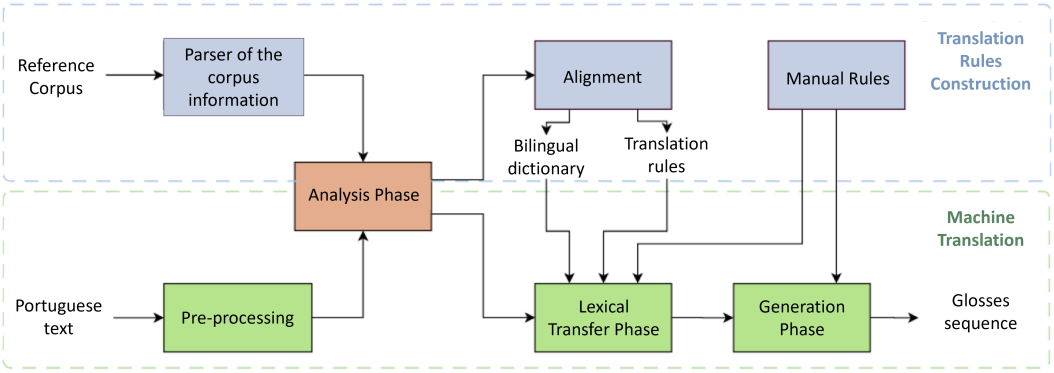


Figure 5.2: Overall architecture of the Translation process.

In this phase, we changed the previously implemented system to account for the analysis and generation of separate clauses by dividing sentences into separate clauses that have at least one verb. This step is important because the lexical transfer and generation phases must be done for each clause individually so that the order of sentence elements and the order of constituents of facial expressions is done correctly. For instance, the sentence “A Maria não gosta de escrever mas gosta de desenhar” (“Maria does not like to write but likes to draw”), using the previous system is translated into: “MARIA GOSTAR ESCREVER MAS GOSTAR DESENHAR NÃO” (“MARIA LIKE WRITE BUT LIKE DRAW NO”), which is incorrect. Using our improved system, the sentence is translated into: “MARIA ESCREVER GOSTAR NÃO MAS DESENHAR ELA GOSTAR” (“MARIA WRITE LIKE NO BUT DRAW SHE LIKE”), which is correct.

We further extended this system to identify the constituents of facial expressions by updating the labels produced by the Freeling tool. These updated labels are then used in the generation phase to order the constituents of negatives (i.e., negation adverbs and negated verbs) and content interrogatives (i.e., interrogative pronouns and adverbs). In addition to this, the system was also extended to: (1) identify the adjectival verbs/modifiers that the mode adverb “muito” (“very”) is applied to, (2) identify the adverb of conditional adverbial clauses, and (3) identify the object of transitive verbs based on the dependency relationships recognized by SpaCy. The last item is important for classifiers.

5.2.2 Generation Phase

In this phase, manual rules related to the morphology of LGP are applied and the lexicon is converted into glosses. We changed the previously implemented system to separate glosses from their corresponding facial expressions, so that facial expressions now contain the type (i.e., negative, interrogative) of each facial expression and the indices of glosses they cover. Using the indices of glosses makes it easier to animate the various simultaneous linguistic processes in the Animation process (Section 5.3.5).

Furthermore, we extended the system to recognize verbs with incorporated negation, for instance, the verb “Querer” (“Want”) which has an irregular negation. To do so, we created a list with all verbs that are in our database and have an irregular negation. Using this list, we check whether the negative sentence contains an irregular negation, and if so, we convert the gloss to have the naming scheme “NÃO_VERBO” (“NO_VERB”), we remove the gloss “NÃO” (“NO”) and add a negative facial expression containing the index of this gloss (i.e., “NO_VERB”). Our updated system recognizes regular and irregular negatives, and polar and content questions. The regular negative facial expression is applied to the “NÃO” gloss, the irregular negative facial expression is applied to the “NO_VERB” gloss, the polar question facial expression is applied to the interrogative verb, and the content question facial expression is applied to the interrogative pronoun or adverb. Furthermore, we also added a slight “narrowed eyes” facial expression throughout the negative or interrogative clause which is more intensified in the indices that contain syntactic facial expressions.

In LGP, verbs are always applied in their infinitive form, therefore, marking the verb tense can be done in three ways [53]: (1) by adding facial expressions (e.g., morphemes) to the neutral form of the verb (i.e., infinitive mode of the verb) [1], (2) by adding adverbs of time (e.g., yesterday, tomorrow, etc.) at the beginning of the sentence, (3) in the absence of the first two ways, verb tenses can be produced by resorting to 3 imaginary points in which the space behind the signer’s shoulder marks the past, the space in front of the signer marks the present and the space further away from the signer’s body marks the future [19].

In the absence of time adverbs in a sentence, the previous system would add the signs “PAST” or “FUTURE” accordingly. However, after further meetings with the Católica team and analyzing of our LGP corpus, the previous solution was not verified in our corpus since signers would either use a time adverb or the verb tenses would be understood from the given context. Further research must be done to understand if rotating and moving the signer’s body could indeed be used to represent verb tenses. Furthermore, we also extended the previous system to add a personal pronoun if the subject in a sentence is omitted, which is common in Portuguese sentences. This step is important because, in LGP, verbs are not conjugated, therefore, if the subject is omitted we cannot know who the subject is.

We also extended the system to account for composite utterances. To do so, we created a list containing all 146 composite utterances from our database and we additionally convert feminine words into their corresponding feminine composite utterances. This way when a sentence contains any composite utterance (i.e., feminine or other), we convert that gloss into its corresponding composite utterances and update their tags, which are later used to identify the indices of composite utterances. Furthermore, after converting the lexicon into glosses, we transcribe numerals into corresponding numbers. For instance, we convert “mil quinhentos e cinquenta e seis” (“one thousand five hundred and fifty-six”) into 1556. This step is important because numerals can only be animated in the Animation process if these are

transcribed into numbers, otherwise, we would fingerspell them using letters, which is not correct. After converting all glosses and additional morphosyntactic information, we extended the system further to identify pauses between clauses and between sentences. This step is important to account for prosodic properties that were missing in the previous system.

5.2.3 Phonetic Transcription

The previous system was also extended to create mouthing animations. First, we had to research whether mouthing in LGP is done with the words in Portuguese or their lemmas. After analyzing multiple videos from SpreadTheSign[7] and our LGP corpus, we can conclude that mouthing should be done with the words in Portuguese and not their lemmas. For instance, verbs are not conjugated while signing, but these should be conjugated while mouthing. Therefore, we extended the system to gather all words in Portuguese and afterwards, combine them into a sentence so that we consider the assimilation between words when executing the phonetic transcription. The phonetic transcription is done by employing the phonemizer tool[8], where the espeak backend is used to produce phoneme sequences described based on the International Phonetic Alphabet (IPA) transcription. The phoneme sequences generated by phonemizer contain some non-Portuguese letters, therefore, we use Amazon Polly's Phoneme-Viseme mapping to normalize IPA letters into corresponding Portuguese letters (e.g., "ɛ" letter is normalized to "e"). For letters that are not contained in the Amazon Polly's table and remain non-ASCII characters (e.g., "ɨ" character which should be transcribed to "i"), further normalization is done by encoding non-ASCII to ASCII.

After normalizing all letters, we separate words into their corresponding syllables using syllabification rules and then we map each phoneme into one viseme using the phoneme-viseme mapping we created (Table 5.1). While mapping visemes, we need to be careful not to over-articulate as it would generate unnatural mouthing animations. We prevented the over-articulation problem by removing visemes that are irrelevant in the visual domain. For instance, we remove viseme consonants that are at the end of a syllable and visemes that have equal consecutive visemes.

From the generation phase we get: (1) a sequence of glosses, (2) a sequence of visemes separated by syllables for each gloss, (3) sequence that identifies the indices of composite utterances, (4) sequence that identifies pauses in-between clauses and in-between sentences, (5) sequence that identifies the indices of an adverbial conditional facial expression, (6) syntactic facial expressions that contain their type and indices of glosses they cover.

[7] https://media.spreadthesign.com/video/mp4/6/224626.mp4, Video that contains the signs "FACULDADE EU IR" where the verb is conjugated while mouthing

[8] https://github.com/bootphon/phonemizer

Viseme	Phonemes
A	a, e
E	i
O	o
U	u
B	b, m, p
F	f, v
C	other consonants

Table 5.1: Phoneme-to-viseme mapping.

5.3 Animation Process

The Animation process allows users to write a Portuguese sentence and view the corresponding animation in LGP signed by the avatar. This is where all components are connected: manual signs, non-manual components, mouthing, and secondary movements. Therefore, this process is where the most complex implementation takes place as it accounts for the synchronization of multiple co-occurring linguistic and non-linguistic processes. Every week, the animations generated would be shown to the Católica team and these would be improved based on their feedback.

5.3.1 Architecture

The overall architecture of the Animation process is divided into three parts, as shown in Figure 5.3. The first part consists of two start functions that are executed simultaneously before the user interface loads: (1) One function loads signs and facial expressions stored in a database (Section 5.3.2.B), (2) the other function checks the server's connection. As described in Section 5.1, the communication between the server and the Web browser (i.e., Website where the Translator is hosted) is done by a Restful API. While the main interface is being loaded, we check the server's connectivity by sending a GET request using the *UnityWebRequest* class and waiting for a response. If the server is connected we receive a "200 OK" status code which is a standard response for successful HTTP requests, otherwise, we will receive a connection error which means the server is down.

The second part is after the user interface is loaded. This part consists of a user writing a text which is then sent to the server (Section 5.3.3) that is connected to the translation process (Section 5.2). When the translation process finishes, the server sends the response back to the animation process, which in turn, searches for the corresponding animations using our database search algorithm (Section 5.3.4). The third part consists of the avatar animating all co-occurring linguistic and non-linguistic processes (Section 5.3.5).

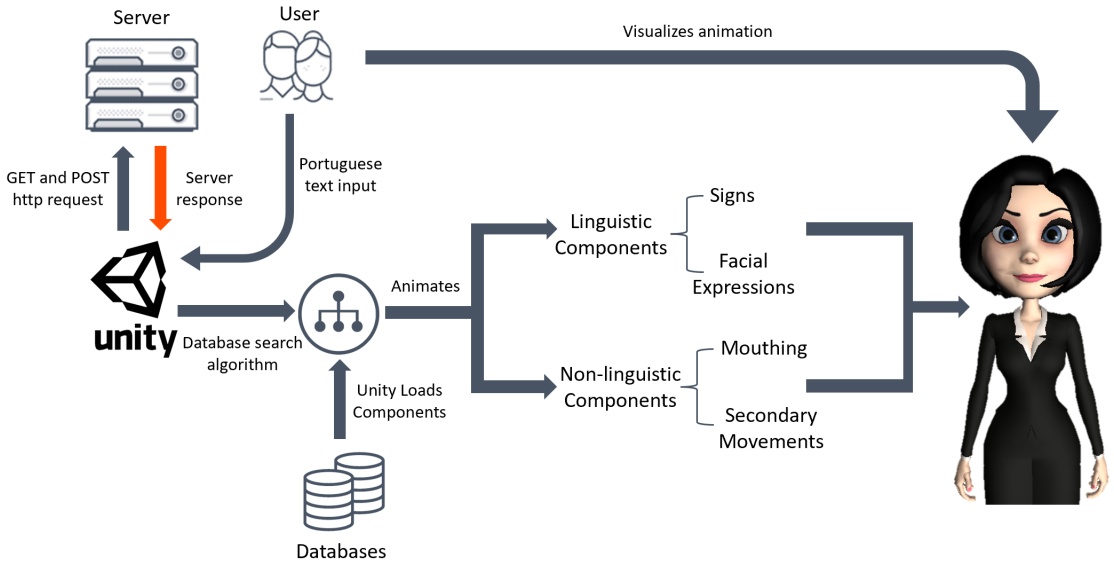


Figure 5.3: Overall architecture of the Animation process.

5.3.2 Database

5.3.2.A Database Creation

In Unity, animation files (*.anim*) must be serialized as *AnimationClips* so that these can be loaded and played correctly in runtime. There are only two ways to load animations in runtime in which these are serialized correctly as Unity structures (i.e., *AnimationClips*). One way is by storing animation files in the **Resource folders** and another way is by creating **Asset Bundles**. Further information about the Resource Folders and Asset Bundles can be seen in this link [9].

We stored all facial expressions and signs in the **Resource folders**. When the project is exported to WebGL, these assets and objects are combined into a single serialized file, and using the Resources API, assets can be loaded in runtime. The Resources folder's simplicity makes it an excellent solution for quickly prototyping, however, since the folders are compressed when the project is built as a WebGL, these cannot be continuously updated and upgraded. This limits the project's maintenance because it reduces a project's ability to deliver custom content and precludes the possibility of incremental content updates in real-time. Using the Resources folder, the person responsible for creating signs would have to download the Unity project, add the new signs, rebuild the project and upload the built files to the website, which is time-consuming.

Towards the end of the project, we wanted to introduce a solution that eases the process of adding/-modifying/deleting signs in the database. This was when we were introduced to **Asset Bundles**. The Asset Bundle system allows files to exist externally and provides a method for files to be stored in an

[9] https://learn.unity.com/tutorial/assets-resources-and-assetbundles#5c7f8528edbc2a002053b5a7

archival format that Unity can index and serialize. Using this system, we can dynamically access files in an external database that can be continuously updated without needing to download the project, rebuild it and upload it to the website every time a new sign needs to be added. We developed a new script that allows users to create Asset Bundles by clicking on a new menu button in the Unity project. After clicking on this button, all signs created or modified will be loaded, then the animation and *JSON* files corresponding to a sign are combined and a new Asset Bundle is created. This way we create individual Asset Bundles for each sign, rather than one large Asset bundle for all signs, which not only avoids incurring performance issues, but is also more memory efficient, and eases the process of adding, modifying, and deleting individual signs.

The Resource Folders system should be used for components that are not memory-intensive and do not need to be constantly updated: dactylology signs (i.e., alphabet and numbers) and syntactic facial expressions. The Asset Bundle system should be used for files that require continuous content updates: manual signs. The Resource Folders are compressed with the built files and accessed in the WebGL application, however, Asset Bundles are stored externally and can only be accessed in WebGL through APIs (i.e., Communication described in Section 5.1). The Firebase Storage is a good solution to store the Asset Bundles, because it can be easily integrated with Unity, does not have any CORS problem when Unity tries to access it, and does not overload the website the Translator is hosted on, since the files are stored externally. To integrate Firebase with Unity, the Firebase SDK must be installed and added to the Unity project by following these steps[10]. Using our script, when the users click on a new menu button, we automatically store the Asset Bundles in a Firebase Storage and save the sign names in a text file. The text file is useful for listing all signs that were added to the Firebase Storage so that these can be easily loaded in Unity.

5.3.2.B Loading components from Database

As described previously, there are two systems capable of storing components: Resource folders and a Firebase Storage that contains the Asset Bundles. Currently, we have around 1010 manual signs that were mostly created by the Católica team using the Sign Editor (Section 4.3). Unfortunately, using APIs to gather the large amount of manual signs in the Firebase Storage overpowers Unity. Thus, we decided to load them from the Resource folders. The Asset Bundles approach should, however, be further explored as it provides a great solution for the project's maintenance.

Before the user interface is loaded, we retrieve all signs and facial expressions stored in the Resource folders and save the most relevant information in dictionaries for a faster search. An overview of the process of loading components from the databases can be seen in Figure 5.4. Using the Resources API, we load all dactylology signs, manual signs, and syntactic facial expressions, and store the animations

[10] https://firebase.google.com/docs/storage/unity/start

in three dictionaries, one for dactylology signs, one for manual signs, and another for facial expressions. Furthermore, while loading each dactylology and manual sign animation, the corresponding JSON file is also deserialized into an instance (i.e., a class that contains the data structures required), and the position of the Avatar's hands in the first and last keyframes is retrieved and saved. Saving the hands' positions is useful for the dynamic transitions introduced in Section 5.3.5.B. Moreover, while loading each manual sign animation, an Aho-Corasick trie is constructed (Trie construction and search algorithm explained in Section 5.3.4) and utilities for each sign are stored. These utilities represent the number of glosses per sign (e.g., "Fim_semana" has 2 as utility and "semana" has 1) which is what allows the database search algorithm to find signs that are composed of two or more glosses. An additional dictionary is also created to store whether mouthing can be executed or not for each sign depending on the facial expression incorporated. If a sign contains a facial expression that requires the mouth (e.g., cheeks puffed, tongue touching the chin, morphemes) then mouthing cannot be executed.

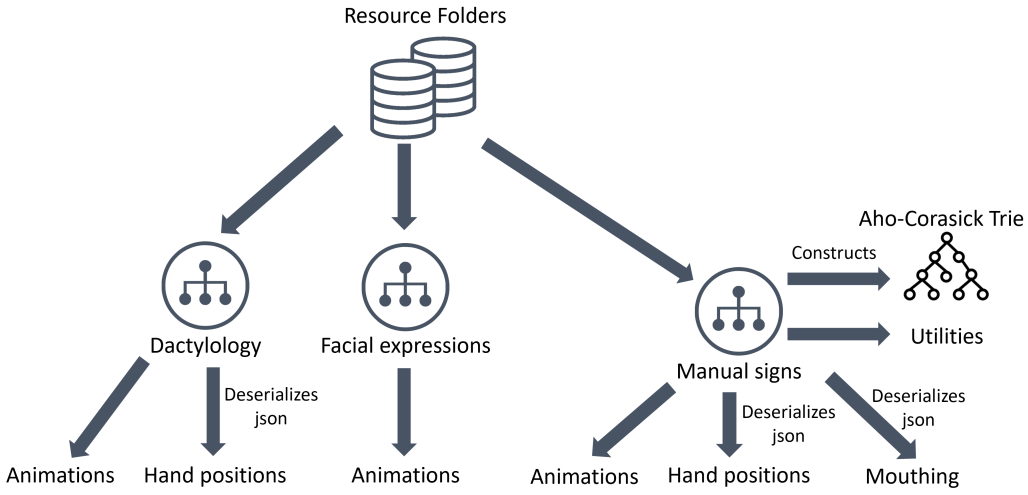


Figure 5.4: Overview of the process of loading components from the databases.

5.3.2.C Json Deserialization in WebGL

There is a problem with the *IL2CPP* serialization in WebGL[11]. The deserialization of *JSON* files in development mode works well but in a WebGL build, the *System.Text.Json* class only works with reference types (e.g., string, an instance of a class) and not generic functions with value types (e.g., float, integer, boolean). This bug has been reported to Unity for many years and has not yet been fixed. We found a workaround for this issue by converting all value types in the *JSON* file to strings. The correspond-

[11] https://issuetracker.unity3d.com/issues/system-dot-text-dot-json-dot-jsonserializer-dot-deserialize-throws-an-error-when-deserializing-a-json-to-a-class-that-has-2-or-more-properties

ing class instance that the *JSON* file is deserialized into, must also only contain strings or instances of strings (i.e., List<string>, Dictionary<string>, etc). Only after deserializing the *JSON* file, can all strings be converted into their corresponding value types.

5.3.3 Sending the Portuguese Sentence

After the user interface loads and if the server is connected, a user can write a sentence and click on the Submit button. When the submit button is clicked, the sentence is encoded in the UTF-8 format and stored as a byte array. This step is important to preserve all accented characters so that natural language pre-processing tools used in the Translation component (Section 5.2) can process and tag each word correctly. After the sentence is encoded, the byte array is sent in a POST request using the *UnityWebRequest* class and the system waits for a response. If the Translation process goes well and the server is running correctly, then a JSON message will be received as a response. Otherwise, we receive either a translation error if there was a problem in the Translation process or a connection error if the server is down. After receiving a JSON message sent by the Translation process, the message is deserialized into appropriate data structures.

5.3.4 Database Search Algorithm

To convert the glosses received into their corresponding animations, the Aho-Corasick algorithm is used. This step is important because some signs might be composed of two or more glosses (e.g., “Casa de banho”, “Boa tarde”, “Até amanhã”) and an exact match between gloss-animation would not consider this. The Aho-Corasick is an algorithm that searches multiple patterns simultaneously to locate all occurrences of strings in a text. This algorithm consists of building a finite state automaton from pre-defined patterns and then using this automaton to process the text string and return all matches. While loading signs from the database (Section 5.3.2.B), we build a trie (i.e., Keyword Tree) by creating branches for each sign. Each branch consists of nodes that represent the letters in a sign, and these nodes are connected by edges. After creating all branches, we extend the trie into an automaton that supports linear time matching by finding the longest proper suffix for each node using the Breadth first search algorithm.

After submitting a text and receiving the translated glosses, we use the Aho-Corasick algorithm to find all glosses that match signs from our database. For instance, a user submits “A Maria precisa de ir à casa de banho” which translates to “CASA”, “BANHO”, “MARIA”, “PRECISAR” and if all glosses exist in our database except for “MARIA”, the Aho-Corasick algorithm will return “CASA”, “BANHO”, “CASA BANHO”, “PRECISAR”. The sign “CASA BANHO” is different from the signs “CASA” and “BANHO” individually, hence, the importance of the utilities explained in the previous section. After receiving the matches from the algorithm, we iterate through each gloss and check if a gloss has no

match, one match, or multiple matches. If the gloss has no match, then we add it to the final results because it means we will fingerspell it (Section 5.3.5.C). If the gloss only has one match, we simply add it to the final results. If the gloss has multiple matches, we add the match that has the highest utility value, therefore, for the matches “CASA”, “BANHO” and “CASA BANHO” we would only add “CASA BANHO”. The final results would then be “CASA BANHO”, “MARIA”, “PRECISAR”.

5.3.5 Animation

In this process, we animate multiple components simultaneously: manual signs, facial expressions, mouthing, and secondary movements. To do so, we use Unity’s animator controller that maintains and arranges multiple animation layers. Each animation layer manages complex state machines that can be applied to different body parts and with different blending modes.

5.3.5.A Manual Signs

In the animator controller, a layer was created for manual signs and dactylology signs. This layer contains six states: (1) idle state, (2) thinking state, (3) temporary animation state, (4) another temporary animation state, (5) pause state for in-between clauses, (6) pause state for in-between sentences.

When the Unity application loads, the Avatar is in an idle state which is a neutral pose. When the user writes a text and submits it, the avatar goes into a thinking pose to inform the user that the translation is being processed. After the translation process (Section 5.2) ends and the Aho-Corasick algorithm finishes picking the final glosses, the avatar transitions from the thinking pose to the animation of signs. A real-time overview of the animation of signs with a pause in-between clauses can be seen in the following video[12].

To animate the avatar we need to manipulate animation resources at runtime, however, the scripts Unity provides for this task are included in the package *UnityEditor*, which cannot be included in the WebGL build. This limitation led us to the following solution: a recursive function that goes through each gloss and transitions from one sign to another alternating between two machine states. Using these two machine states and the *animatorOverrideController* tool[13], we can substitute the temporary animations that are in these states with the sign animations and consequently transition between signs by transitioning between states.

The transition between states is done by a recursive function that is called every time the current sign finishes its animation. In Unity, there are two ways to call a function after a specific amount of time: (1) Animation Events, (2) *WaitForSeconds* function. Animation Events[14] are instances that allow a

[12] https://youtu.be/TMxXhLC4fVM
[13] https://docs.unity3d.com/ScriptReference/AnimatorOverrideController.html
[14] https://docs.unity3d.com/Manual/script-AnimationWindowEvent.html

function to be called directly from an Animation Clip after a given amount of time. The *WaitForSeconds* function[15] suspends the coroutine execution for a specific amount of seconds. The *WaitForSeconds* function is extremely useful for functions that need to be stopped at a specific point in their execution, however, for the transitions of signs we decided to use Animation Events. The reason behind this is that the recursive function that animates signs also executes other components and cannot be stopped at any point of its execution. Therefore, when we are iterating through each gloss, an Animation Event is created that has the length of the current animation clip and after the animation clip finishes playing, it calls a specific function depending on three conditions:

1. If there is a pause after the current sign, the *animatePause* function is executed. This function checks whether a pause is either in-between clauses or sentences, and animates it accordingly. In this function, the *WaitForSeconds* function is used to wait for the pause animation to finish and when it does, we return to the main recursive function.

2. If there is another sign after the current one, the main recursive function is executed again and there is a transition between the current animation state and the next animation state.

3. If the current sign is the last gloss, meaning we have reached the end of the input sentence, the *StopAnim* function is executed. This function stops all animations for signs and facial expressions, and the avatar transitions to an idle state. Furthermore, all game objects (i.e., buttons and input fields that are in the interface, described in Section 5.3.5.H) are also reset.

5.3.5.B Dynamic Transitions

At first, to transition between signs, the state transitions[16] in the animation layers were used, however, their duration and offset values are constant and cannot be changed dynamically in run-time without the use of the *UnityEditor* package. As described in Section 1.2, the transitions between signs rely heavily on the phonology of the previous and following signs and determine the movement fluidity that allows sign streams to be intelligible. Therefore, dynamic transitions need to be considered as these can have an impact on the comprehension and naturalness of sign animations.

To the best of our knowledge, we created a new contribution to the state-of-the-art for the interpolation of signs through dynamic transitions that change according to the previous and following signs. To do so, we use the dictionary created in Section 5.3.2.B that stored the position of the avatar's hands in the first and last keyframes for all signs. While we iterate over each gloss in run-time, the differences between hand positions in the last keyframe of the previous sign and the first keyframe of the following sign are calculated and then the squared magnitude of these vectors is computed. Calculating the

[15] https://docs.unity3d.com/ScriptReference/WaitForSeconds.html
[16] https://docs.unity3d.com/Manual/class-Transition.html

squared magnitude of a vector is much faster than using the magnitude property, since it does not require a slow square root operation that makes the magnitude property take longer to execute[17]. These squared magnitude values are then converted to percentages by defining a scale. To decide this scale, we checked all signs created to find two signs that have the closest hand position differences (e.g., signs "EU" and "TER") and two signs that have the furthest hand position differences (e.g., signs "ELE" and "TER"). Based on our findings, we defined two scales: one that includes both hands (if the left hand has movement), and another that only considers the right hand (if the left hand has no movement). Using these scales, the squared magnitude values are converted to percentages that range between 0% and 100%. Finally, to find the duration value used in the transition between signs, we use the percentage calculated to linearly interpolate between two duration values. These two duration values correspond to the lowest and highest values that the duration of transitions can take. We defined these values by analyzing the lowest and highest transition duration in multiple videos of our LGP corpus. Furthermore, two empirical studies developed by Sedeeq [54, 55] found that ASL signers prefer slower transitions than the timing of human signers and that they prefer animations with an average transition time of 0.5 seconds. Based on the analysis of our corpus and the studies we read, we decided that the duration of transitions would range between 0.3 seconds and 1.1 seconds because this range would include 0.5 seconds as the average transition time and these are slightly slower than the human signing transitions in our LGP corpus.

Using the calculated duration values in the process previously described, we can create an interpolation between the current sign and the next sign using dynamic transitions by defining a duration value and an offset value. The first keyframe of every sign in the database starts at 1 second, which is what allows transitions between signs to be executed without cutting the signs shorter, because without it the transition would overlap the beginning of each sign. Using the offset value, we can adjust the timing until the first keyframe to match the transition duration time, therefore, the offset value is 1 second minus the transition duration value. Transitions must be seen as a continuous stream of motion without being too paused because co-articulation, similarly to oral languages, also constitutes an important part of Sign Languages. To create transitions that are fluid and not too paused between signs, we decided to define the offset value as 1.2 seconds minus the transition value, instead of 1 second, because this way signs would be slightly overlapped and transitions would be more fluid.

Another aspect taken into consideration was the phonological assimilation processes of composite utterances. Composite utterances are utterances that have meanings derived from the composition of multiple signs (e.g., "VERMELHO" + "MELÃO" means "MELÂNCIA"). Since multiple signs can be combined for one sole meaning, the transitions between these must be smaller than transitions between signs that have separate meanings. This is another reason why dynamic transitions are so important.

[17] https://docs.unity3d.com/ScriptReference/Vector3-sqrMagnitude.html

These can have an impact on the perception of composite utterances if the phonological assimilation processes are not taken into consideration. Based on the videos from our LGP corpus, we defined 0.2 seconds as the transition duration in-between all signs that comprise a composite utterance. Using the indices that define composite utterances obtained from the Translation process (Section 5.2), we transition between signs that comprise composite utterances with a transition value of 0.2 seconds, making the transitions for composite utterances faster than transitions for other signs.

5.3.5.C Dactylology

For glosses that are not in our database, we employ dactylology (i.e., fingerspelling). These are commonly used to represent names of people, places, numbers, and technical vocabulary when there is no direct translation from Portuguese to LGP. To animate the glosses received, we either animate the manual sign if it exists in our database or we fingerspell it. The process of animating dactylology is essentially the same as animating manual signs, we also use the method of substituting the temporary states. However, now, rather than substituting the temporary states by sign animations, we substitute them with the animation of each letter or number contained in glosses.

This process starts by normalizing letters in the glosses so that accents can be removed because each letter should be animated without accents. We were told by the Católica team that, while performing dactylology, there is also a horizontal hand movement when animating numbers or when there is a letter repetition in a word. We used Unity's IK system[18] that allows us to more easily manipulate the avatar's hands. To do so, an invisible ball was added to the scene and using the IK algorithm the hand can be moved by moving the ball. To animate numbers, the hand starts moving from left to right in a straight line where the hand position for each number is calculated by a fixed distance divided by the length of numbers. We wanted the hand to move as smoothly as possible, so a mathematical equation was used to smoothly interpolate between the current hand position and the ball position by gradually increasing the hand speed. Furthermore, a mathematical equation was also used to smoothly move the hand horizontally while animating numbers. Since the hand position between numbers is equal, the transition between these must also be equal. Furthermore, the differences between hand positions for letters are also not significant, therefore, we decided to employ constant transitions between letters and numbers. Dactylology in Sign Languages is typically slower and more paused than manual signs so that each letter and number can be visualized correctly. Therefore, we defined 0.5 seconds for the transition duration between numbers and letters and an offset value of 1 minus the transition value so that transitions have a slight pause to facilitate comprehension. Moreover, since there can be multiple dactylology signs consecutively, we added a pause in-between dactylology signs so that users can understand when a sign ends and the next one starts. This video[19] shows the avatar animating numbers and letters.

[18] https://docs.unity3d.com/Manual/InverseKinematics.html
[19] https://youtu.be/j4pFPnRcHv0

5.3.5.D Facial Expressions

In the animator controller, four layers were created for facial expressions: (1) syntactic blend shapes used for facial expressions, (2) facial and body movements for syntactic facial expressions, (3) facial and body movements for simultaneous syntactic facial expressions, (4) facial expression for conditional adverbial clauses.

On a syntactic level, facial expressions and body movements and not incorporated in, but rather combined with signs. These must be carefully added to not override or change, even if slightly, any sign's components (i.e., hand configurations, orientations, locations, movements, and non-manuals) because they could affect its entire meaning. For this reason, we separated the blend shape animations from the body and head movements that were created with the Facial Expression Editor (Section 4.2). Blend shape animations are in a layer that is only applied to the face and has an override blending mode, while facial and body movements are in a layer that is applied to the face and the body and has an additive blending mode. The blending mode is what allows animations to either override the current animation or be combined with it. Unfortunately, in Unity, when blend shapes are in an additive mode, these are combined without being able to define how much they should be combined. This can cause the avatar's face to be deformed, which is not ideal. We found a workaround for this issue, we only override the top part of the face when animating syntactic facial expressions. This way the avatar can combine phonological and syntactic blend shapes where the bottom part of the face is a phonological facial expression and the top part is a syntactic facial expression. On the other hand, if body and facial movements are in an additive layer, these can be combined with manual signs without any problem. However, these must be carefully combined because, for instance, shoulder movement affects the arm's position, which can have an impact on the manual sign's components. Therefore, while animating interrogatives, we also added an arms' movement to balance the shoulders' movements so that the hands' positions in manual signs are in the correct location. The difference between having and not having the arms' movement to balance the shoulders' movement can be seen in Figure 5.5.

While signs are being animated, syntactic facial expressions are simultaneously animated by calling an additional function while executing the recursive function. Every time a sign is animated, we check whether a facial expression must be animated or stopped by iterating through all facial expressions in the *JSON* message received from the Translation process (Section 5.2). Facial expressions are animated at the same time as the signs they cover, therefore, they follow the same transition duration as signs. The headshake in negatives is animated continuously until the signs they are applied to finish playing. Furthermore, in this step, we can also animate the body movements of two syntactic facial expressions at the same time by simultaneously animating two different layers with an additive blending mode. Moreover, in the Translation Process (Section 5.2) we also identified the glosses that should have facial expressions in a conditional adverbial clause. Based on the study developed by Martins

Interrogative with arms' movement and hands in the correct location.

Interrogative without arms' movement and hands in the wrong location.

Figure 5.5: The difference between having and not having the arms' movement.

and Mata [25] and videos from *SpreadTheSign*[20], we concluded that in conditional sentences, eyebrows must be raised while animating the "SE" sign ("IF" in English) and the verb it is applied to.

To the best of our knowledge, research in the field has not yet been published regarding the animation of co-occurring syntactic facial expressions (i.e., a negative and interrogative sentence) and simultaneous phonological and syntactic facial expressions. Based on the analysis of videos from native LGP signers, in co-occurring syntactic facial expressions applied to the same sign (i.e., polar interrogatives and negatives), the facial and body movements of both expressions are animated. However, only the blend shape of one of these expressions can be applied due to the blend shapes limitation described previously, therefore, we decided to only animate the interrogative blend shape because this facial expression is required for users to be able to identify interrogatives but not negatives. In a simultaneous phonological and syntactic facial expression, the phonological is applied to the lower part of the face and the syntactic to the top part of the face as described previously. These facial expressions have been through many iterations according to the feedback provided by the Católica team. Some facial expressions can be seen in this video[21].

5.3.5.E Mouthing

As described in Section 3.2.2.A, mouthing is an essential part of any automatic written-to-sign translation system and without it, a signing avatar would look unnatural and could omit important information. Our

[20] https://spreadthesign.com/pt.pt
[21] https://youtu.be/oddh6Qp1txU

avatar contains 7 visemes: *A*, *B*, *C*, *E*[22], *F*, *O*, and *U*; and since these are the most common visemes used, we decided they would be enough to animate the 33 phonemes that exist in the Portuguese language.

To create visemes as close as possible to human visemes, animations for each viseme were created by adjusting the weights of blend shapes. In the translation process (Section 5.2.3), words are translated into phonemes, separated into syllables, and then mapped into visemes. In the animation process, when the manual signs are being animated, mouthing is animated by using an interpolation scheme that concatenates the visemes according to the animated signs. To do so, we check whether mouthing can be animated for each sign resorting to the dictionary that stored this information in Section 5.3.2.B. If mouthing can be animated, a duration value for the mouthing is defined based on the duration of the sign it is applied to and based on the number of syllables for that sign. The reason behind this is that we do not want mouthing to either overlap the duration of a sign or be too slow if the duration of a sign is too large. Based on a study developed by Greenberg, Carvey, Hitchcock, and Chang [56], the average duration of syllables per utterance ranges between 0.2 seconds to 0.4 seconds. Following these findings, we defined a mouthing duration of 0.4 seconds multiplied by the number of syllables and if this duration is higher than the duration of the sign, we define the duration of mouthing to be the same as the duration of the sign. After defining the mouthing duration, this value is divided by the number of syllables and then each syllable duration is divided by the number of visemes in that syllable. This way we get a more accurate viseme duration and more natural mouthing animation. The synchronization between mouthing and the corresponding sign was taken into account by using the mouthing duration value described previously and by using the in-between sign transitions as the in-between mouthing transitions. The synchronization between mouthing and signs is extremely important because studies [7] have reported that a mismatch between the duration of signs and their corresponding mouthings can provoke a disturbing oscillation of the user's visual focus from hands to face. The following video[23] shows the mouthing animation.

5.3.5.F Secondary Movements

The linguistic processes are the most important actions in signing animations to determine the morphosyntactic motions, but secondary actions are equally important to create realistic and natural animations. Our goal in this component was to infer secondary movements based on human kinematics as much as possible that adhere to the linguistic movements. However, secondary movements must be carefully added to not interfere with the subtleties in linguistic motions or be too exaggerated, which leads to unrealistic movements.

[22] The *E* viseme represents the *I* vowel in Portuguese.
[23] https://youtu.be/oSYE16K_XmU

In real life, no part of the human face and body is truly stationary, therefore, an avatar without the subtle motions of humans can appear highly robotic. We created an idle animation using Unity's animator that contains subtle facial and corporal movements. This idle animation was added in a layer that adheres to the linguistic motions without exaggerating facial and corporal movements, but rather, bring subtle additional movements that allow the avatar to be less static. Not exaggerating movements is important because these could change, even if slightly, any sign's components or could add body jitters that distract the viewer's attention from the signing aspects.

Eye blinking has been observed in several Sign Languages to play a role as a marker in prosodic boundary cues. The study developed by Tang, Brentari, González and Sze [57] explored the use of eye blinks as prosodic cues from a crosslinguistic perspective for four Sign Languages: Hong Kong Sign Language, Japanese Sign Language, Swiss German Sign Language, and American Sign Language. This study revealed that eye blinks indeed have a prosodic role in marking Intonational Phrase boundaries in all four languages consistently. Intonational Phrases are phrases that contain syntactical coherent elements, for instance, clauses, topicalized structures, and parentheticals. Breathing in between intonational phrase boundaries in spoken languages can be seen as the eye blinking equivalent in Sign Languages. Just as speakers time breathing to coincide with intonational phrase boundaries, so do signers time blinking to coincide with those same boundaries [58]. In LGP, no studies have yet been developed to analyze whether eye blinking has prosodic properties. Due to the lack of time, we decided to include a constant blinking animation for now. Further research is required to understand the timing of eye blinking, therefore, for now, this blinking animation plays every 2 seconds since this is the average human blinking time. This animation plays throughout the signing movements except when there is a syntactic facial expression because it could interfere with the narrowed eyes expression.

In Sign Languages, typically, the head is more active than the torso. The study developed by Tyrone and Mauk [59] for ASL found that the head moves to facilitate convergence with the hand for signs with a lexical movement towards the head, whereas, the torso does not move to facilitate convergence with the hand, but rather, bend and rotate to accommodate the reaching of the arm [49]. While signing, the head tends to follow the movements of the hands, hence, the up-down nodding described in the declarative sentences paragraph in Section 2.3.1. Following these findings and using Unity's IK system, we manipulated the head and torso joints to follow the movement of both hands. The weight for the head movement is higher than for the torso because in LGP was also noticeable that the head is more active. The IK system provides a solution that is not only intuitive and easy to use but also allows the avatar to produce more natural movements. This video[24] shows the avatar with and without secondary head and torso movements. This approach allows the head and torso to move according to the hands' movement, however, for signs further away from the body we also wanted the spine to rotate to accommodate the

[24] https://youtu.be/USGAalJKWfg

reaching of the arm. To rotate the spin according to the reaching of the arm, we also used the IK system and the same approach as the one described for the secondary facial and torso movements, but now, the target position is the difference between the hands' position and the spine position. The following video[25] shows the avatar with and without the torso rotation. Unfortunately, the IK system poses a limitation because we can only either produce the head and torso movements or produce the torso rotation since these use the same approach.

5.3.5.G Head-Hands Collision

As opposed to humans, avatars know no limits in their movements which can lead to inhuman movements such as hands going through the head or chest or even going through each other. The reason behind this problem is that Unity interpolates keyframes following the shortest path in-between poses and does not take into consideration if there is an obstacle in the way. A solution for this problem is to implement collisions. To do so, first, we added capsule colliders in the head and hands as can be seen in Figure 5.6. These are used to identify collisions between the head and hands. If there is a collision, using Unity's IK system, we smoothly move the hand to the point of collision by gradually increasing the speed in the inverse corresponding direction. This way if the hand is going to intersect the head at any point, it will return to the point of collision without making the collision noticeable. To create collisions between the hands and the chest, we would have to add a capsule collider in the chest and follow the same procedure. However, for the collision between hands, the process is not as straightforward because it is not enough for either hand to move to the collision point. One hand would have to go around the other which makes this collision harder to implement. Unfortunately, we only had time to create the head and hands collision.

5.3.5.H User Interface - Additional Features

For greater clarity of animations, we made some optimizations to the avatar for real-time display: 1) The face and hands of the avatar were enlarged; 2) The avatar was dressed in dark colors so that the hands were more easily visible while signing in front of the body; 3) The avatar contains bumped specular smoothing shaders that create bright highlights by reflecting the incident light. These shaders support the Blinn-Phong Tessellation model that provides more realistic and accurate-looking results as opposed to diffuse shaders. 4) two directional lights were also positioned in a way that enhances facial and corporal features, and one of the lights casts shadows to enhance the 3D perspective much needed in Sign Languages.

In the Translator user interface, besides being able to write a Portuguese text and see the translated animation, users have additional features:

[25] https://youtu.be/MN27RgV9cPg

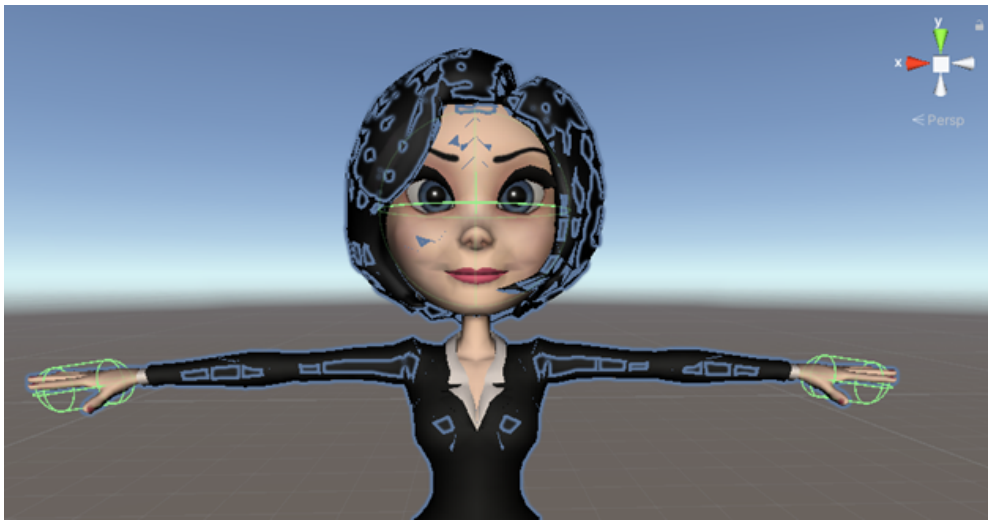


Figure 5.6: Capsule colliders in the avatar's head and hands that are used to detect collision.

1. Users can rotate the avatar by dragging the mouse cursor. This feature is important because the frontal perspective of the camera can lead users to miss imperceptible movements for instance the hands touching the face or chest, or the slight forward upper body movement and head tilts in interrogatives.
2. Users can zoom in and zoom out by pinching the cursor with two fingers. This feature is important for users to see closer facial and corporal movements.
3. Users can choose to see or not the mouthing animation.
4. Users can choose to see or not the glosses as the avatar is animating them.
5. Users can replay the previous animated sentence.
6. Users can choose the avatar's dominant hand. This feature is important because we want to consider not only right-handed signers but also left-handed signers. For learning purposes, it is also easier for a user to be able to switch the dominant hand and visualize the avatar in a mirror perspective.

An overview of the Translator user interface can be seen in this video[26].

[26] https://youtu.be/6Izj7QWjeqc

6

Evaluation

Contents

In order to evaluate our system, we have designed and executed **three experimental user studies**. In Section 6.1, we describe a quantitative and qualitative evaluation used to assess the quality of the generated signing animations and analyze the impact of non-manual components on linguistic comprehension. In Section 6.2, we describe a quantitative evaluation used to analyze the impact transitions have on linguistic comprehension, naturalness, and preference of Sign Language animations. In Section 6.3, we describe a quantitative evaluation used to analyze the impact mouthing has on linguistic comprehension, naturalness, and preference of Sign Language animations. The first user study is used to evaluate linguistic components that determine the morphosyntactic motions needed in Sign Languages, whereas the last two user studies are used to evaluate non-linguistic components that determine the naturalness of the avatar and can have an impact on the comprehension of animations. As described in 3.2.3, it is necessary to separate linguistic and non-linguistic components for evaluation purposes.

The signing animations of all three user studies were generated using our Translator. Although our main goal was not to evaluate grammar, but rather the animations generated, we wanted to assess the efficacy of our Translator in order to answer the following research question: Is an automatic text-to-sign translator effective in generating realistic and natural Portuguese Sign Language animations? Furthermore, since the sentence structure in LGP is still a debatable issue, we decided to follow the "OSV" structure for all sentences because this is the structure taught by LGP teachers and the structure most widely accepted. In this chapter, we will go into detail about each user study, describing the research questions, the participants, the procedure, the analysis and finally we conclude with our findings and suggestions for the improvement of our system.

A text-to-sign language translator is a bit of a polemic theme in the Deaf community. Some believe that these technologies could replace current accessibility services (e.g., professional interpreters) and that these systems have lower quality than professional services, thus, researchers in this field should evaluate the usability of these systems and carefully communicate their limitations and potentials. The goal of our project is, in no way, to replace the high-quality accessibility services provided by interpreters, but rather provide an LGP Avatar that can be used in multiple digital applications, and highlight the importance of learning LGP and the inclusion of the Deaf community.

All three studies were approved by the Ethics Committee of Instituto Superior Técnico, University of Lisbon. The approval is in Appendix A. One of our concerns with these studies was the Portuguese literacy level of our participants as some participants are Deaf and their native language is Portuguese Sign Language rather than Portuguese. We were assured that all participants involved had a sufficient level of Portuguese literacy to understand the consent forms and questionnaires. Furthermore, these were strategically written in simplified Portuguese and reviewed by both our teams at Instituto Superior Técnico and Católica. If any participant did not possess the level of Portuguese literacy required, we had the mitigation strategy of using a LGP interpreter to communicate in the participants' native language,

for instance, in answering any questions or concerns they had while reading the consent form, or by recording the corresponding translation in LGP.

6.1 Linguistic Components Evaluation

The first user study was conducted to answer the following research questions:

- **RQ1:** Does the inclusion of non-manual components enhance the linguistic comprehension of Sign Language animations?
    1. How effective are non-manual components in conveying different types of interrogatives?
    2. How effective are non-manual components in conveying different types of negatives?
- **RQ2:** Does the sequential or co-occurrence of facial expressions have an impact on linguistic comprehension?

6.1.1 Participants

We recruited 10 participants fluent in LGP with the help of the Católica team and the snowballing sampling technique. The demographic information about our participants can be seen in Table B.1. Given that we are designing technology for a community we are not part of and a language we are not fluent in, it was important for us to work with a team fluent in LGP, to base our work on linguistic information, and to obtain feedback from people who are fluent in LGP.

6.1.2 Procedure

We conducted within-subject user tests where each participant tested all conditions because we did not want individual differences to affect our results. The years of fluency in LGP and the region where participants are from can have an impact on the comprehension of animations because LGP has lexical variations across the country as explained in Section 2.

For this user study, we conducted a quantitative evaluation that consisted of questionnaires and afterwards, a qualitative evaluation that consisted of remote semi-structured interviews to clarify, discuss and expand on the results obtained in the questionnaires. Prior to participating in the user studies, each participant was handed a thorough consent form which they had to sign to participate in the study and allow video and audio recording for the interviews.

For the questionnaires, we created twenty sentences based on videos from our LGP corpus and SpreadTheSign. To do so, we created new signs and modified existing ones using the Sign Editor (Section 4.3), and we also added phonological facial expressions that were missing. We used the

Hand Pose Editor (Section 4.1) to create new hand configurations and the Facial Expression Editor (Section 4.2) to create new phonological facial expressions, if needed. After all signs were created and facial expressions were added to each sign, we entered each sentence in our Translator and recorded the corresponding generated animations. The process of creating sentences took about two/three weeks as we had to choose our sentences, create signs and facial expressions when needed, and fix grammar or other errors in the translation process.

The twenty sentences we created were composed of 5 sections: (1) 4 declarative sentences where 2 were simple sentences (i.e., each with 1 clause) and 2 were complex sentences (i.e., each with 3 clauses), (2) 4 interrogatives where 2 were polar questions and 2 content questions, (3) 4 negatives where 2 were regular negatives and 2 irregular negatives, (4) 4 sentences with sequential facial expressions: 1 sentence contained a content question followed by a polar question, 1 sentence contained a polar question followed by a negative, 1 sentence contained a polar question with an "OU" ("OR") conjunction, and 1 sentence contained a negative in one clause and affirmative in the second clause, (5) 4 sentences with co-occurring syntactic facial expressions: polar question + regular negative sentence, content question + regular negative sentence, polar question + irregular negative sentence, content question + irregular negative sentence.

The first three sections (i.e., declarative sentences, interrogatives, and negatives) contained two-paired sentences where one had facial expressions and the other did not. Overall, we had 6 sentences with facial expressions and 6 sentences portraying the same sentence type but without facial expressions. To mitigate experimental bias, the content of these sentences was different but both had similar number of glosses and a similar difficulty level. The remaining 8 sentences, from the last two sections, were composed of 4 sentences with sequential facial expressions and 4 sentences with co-occurring syntactic facial expressions. All sentences contained co-occurring phonological and syntactic facial expressions. The main goal of this user study was to evaluate the importance of individual facial expressions but also to understand how facial expressions are affected by the preceding or succeeding facial expressions, as well as co-occurring ones.

Each participant received a different version of the questionnaire, therefore, we created ten different versions where, in each version, the condition's order is counterbalanced and the sections' order is random. The questionnaire is composed of twenty two sections in which the first section is regarding demographic information about participants and the following twenty sections contained each sentence created for this study. In these sections, participants had to visualize a video, write the content understood, and select the type of sentence it is. Furthermore, participants had to describe whether the sentence contains an error, evaluate the facial expression, and if desired, provide additional comments either in Portuguese or by submitting a video in LGP. The last section of this questionnaire is an overall evaluation of the avatar in terms of many aspects. We asked participants to evaluate the speed, transi-

tions, and pauses in a 1-5 Likert scale with 1 as too slow and 5 as too fast. This section also contained an evaluation of the avatar in terms of naturalness in a 1-5 Likert scale with 1 as robotic and 5 as natural, and in terms of comprehension in a 1-5 Likert scale with 1 as confusing and 5 as easy to understand. Furthermore, we also asked for an evaluation of the general quality, grammatical correctness, signs quality, and facial expressions quality in a 1-5 Likert scale, with 1 as terrible and 5 as perfect.

After finishing the questionnaires, we conducted remote semi-structured interviews so that we could clarify, discuss and expand on the results obtained in the questionnaires. The structure of each interview was adjusted depending on the results obtained for each participant, and the time and duration of each interview was adapted according to the availability of each participant and a LGP interpreter (if necessary). For each interview, participants would visualize the videos, and after, we would tell them the correct answer so that we could understand some results that were wrong or misunderstood. We would go over each phonological and syntactic facial expression to understand their thoughts and how facial expressions could be improved by listening to the participants' feedback and visualizing their demonstrations in LGP. Besides the specific questions we asked each participant, we also asked them the following general questions: (1) Overall, what did you think of the animations? What aspects do you think need to be improved? (2) Overall, what did you think about the facial expressions? (3) What suggestions do you have for improving the naturalness of the avatar? (4) What is your opinion regarding the avatar producing mouthing while signing? This last question was important for us to understand the participants' opinions towards integrating mouthing in the avatar because this component has received mixed feedback in regards to its incorporation in sign languages and acceptance as it derives from spoken languages.

6.1.3 Data Analysis and Findings

In this user study, our main focus was the impact of facial expressions on linguistic comprehension and to evaluate the quality of facial expressions.

To compare the comprehension scores between conditions (i.e., with and without facial expressions), we only evaluated the comprehension scores for the first 3 sections (i.e., declarative sentences, interrogatives, and negatives) since these were the only sections that contained sentences with and without facial expressions. For each sentence in the questionnaires, we had to consider both the comprehension of glosses and the comprehension of sentence types because we have phonological and syntactic facial expressions that have an impact on the perception of signs and the perception of sentence intonation. Therefore, we separated the measurement of comprehension into two: 1) percentage of glosses understood, 2) percentage of sentence types understood that correspond to syntactic facial expressions.

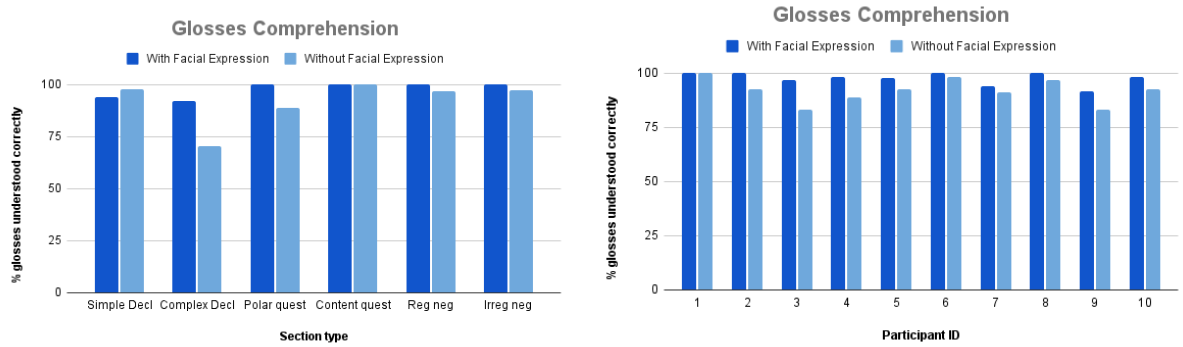
6.1.3.A Glosses Comprehension

We measured the percentage of glosses understood by checking the number of glosses correctly described with 100% as all glosses correctly understood by a participant. This process had to be done manually as synonyms of signs also counted as correct. We reviewed this process three times to make sure all calculations were correct. Overall, the **average glosses comprehension scores for all participants, with both conditions, was 95.04% (**$SD = 12.26$**)**, which is surprisingly high considering this is the first user study we are conducting to evaluate these signs and facial expressions. We divided the evaluation of glosses comprehension in two parts: 1) glosses comprehension per section, 2) glosses comprehension per participant.

First, we evaluated the **glosses comprehension per section**. According to a Shapiro-Wilk test, we rejected the null hypothesis of population normality, therefore, we conducted a Kruskal-Wallis H test to compare differences between sections for both conditions. For sentences with facial expressions, there was a **statistically significant difference in comprehension scores between sections** ($X^2(5) = 19.561, p = 0.002$), with a mean rank comprehension score of 19.65 for complex declarative sentences, 25.35 for simple declarative sentences, and 34.50 for the remaining sections (i.e., interrogatives and negatives). For sentences without facial expressions, there was also a **statistically significant difference in comprehension scores between sections** ($X^2(5) = 28.247, p = 0.000033$), with a mean rank comprehension score of 11.10 for complex declarative sentences, 25.45 for polar questions, 34.15 for regular negatives, 36.30 for irregular negatives, 36.50 for simple declarative sentences and 39.50 for content questions. It is interesting to note that polar questions scored the highest in sentences with facial expressions whereas in sentences without facial expressions scored the second lowest.

Based on the results from the Kruskal-wallis H test and clearly shown in Graph 6.1(a), the **complex declarative section had the lowest comprehension scores** in sentences with and without facial expressions. This is not surprising because this section contained the most complex sentences that were composed of three clauses and was noted by participants that "longer the sentence, the harder it is to understand". Furthermore, these sentences also consisted of classifiers. Classifiers are considered morphemic structures that behave as signs, even though they are not. They can replace, describe, specify and qualify animate and inanimate beings by incorporating actions into these referents [60]. An example of a classifier is the action of "collision between people or collision between a car and a wall" where the collision between people is a different sign than the collision between a car and a wall. In classifiers, the hand configuration, position, movement, and orientation can change completely according to the subject and object an action is applied to, hence, the linguistic complexity of sign languages.

We also evaluated the **glosses comprehension per participant**. As shown in Graph 6.1(b) and Table C.1, 9 participants had higher comprehension results in sentences with facial expressions and 1 participant had equal comprehension results in both conditions ($M = 100, SD = 0$). According to a



(a) Comprehension scores per section.

(b) Comprehension scores per participant.

Figure 6.1: Glosses comprehension scores between animations with and without facial expressions.

Shapiro-Wilk test, we retained the null hypothesis of population normality ($p = 0.020, p = 0.449$), therefore, we conducted a Paired samples T-test to compare differences in comprehension scores between our conditions. Based on the results, the **glosses comprehension scores per participant for sentences with facial expressions were statistically significantly higher** than for sentences without facial expressions ($t(9) = -4.351, p = 0.002$).

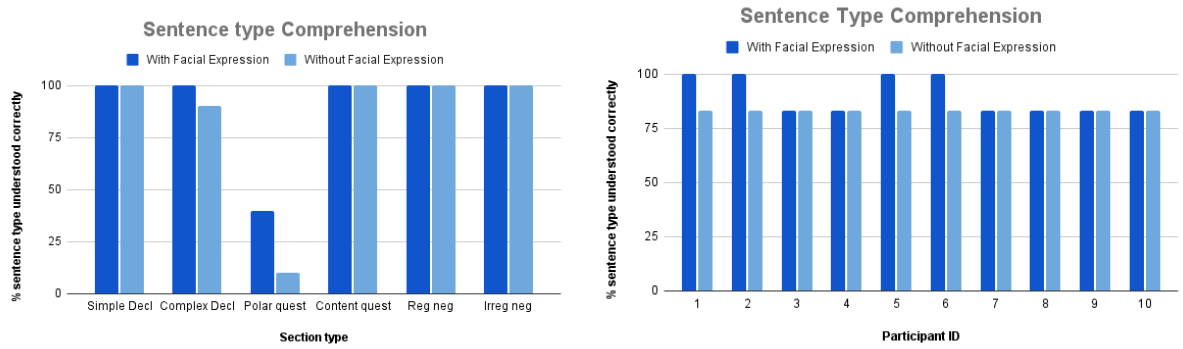
In the questionnaires and interviews, participants also commented on the unnatural fluidity of transitions and how these had an impact on comprehension. It was noted the "slow timing between signs that are composite utterances", and the slow and too paused transitions overall causing "the connection between signs to still seem unnatural". Furthermore, while the overall speed of signs was considerably good with an average score of 76.67% ($SD = 27.44$), the speed of transitions was considered too slow with an average score of 63.33% ($SD = 47.14$). This user study contained a constant transition approach with a constant value of 1.2 seconds. Based on the feedback given by participants, we can understand the **importance of a dynamic transition approach** that contains transitions that are not too paused and slow, but more fluid.

6.1.3.B Comprehension of Sentence Types

We measured the percentage of sentence types understood by checking the number of sentence types correctly chosen with 100% as all sentences type correctly understood by a participant. We reviewed this process three times to make sure all calculations were correct. Overall, the **average comprehension scores of sentence types for all participants, with both conditions, was 87.75% (**$SD = 29.44$**).**

As shown in Graph 6.2(a), the **polar questions section had the lowest comprehension scores** in sentences with and without facial expressions. In the interviews, participants commented that the 2-dimensional field of view in the videos hindered the comprehension of interrogatives as they could not

see properly the facial and corporal movements. This shows the importance of a rotation tool like the one described in section 5.3.5.H, which allows participants to get a 3-dimensional view much needed in Sign Languages.



(a) Comprehension scores per section.

(b) Comprehension scores per participant.

Figure 6.2: Sentence types comprehension scores between animations with and without facial expressions.

As shown in Graph 6.2(b), 4 participants had higher comprehension results in sentences with facial expressions (With Facial Expressions (FE): $M = 100, SD = 0$, without FE: $M = 83.33, SD = 40.82$) and 6 participants had equal comprehension results in both conditions ($M = 83.33, SD = 40.82$). According to a Shapiro-Wilk test, we rejected the null hypothesis of population normality, therefore, we conducted a Wilcoxon signed-rank test to compare differences in comprehension scores between our conditions. Based on the results, the **comprehension scores of sentence types for sentences with facial expressions were statistically significantly higher** than for sentences without facial expressions ($Z = -2.000, p = 0.046$). Therefore, we can conclude that **non-manual components were effective in conveying different types of negatives and interrogatives**, however, the facial expression of interrogatives still needs to be improved.

6.1.3.C Sequential and Co-occurring Facial Expressions

In this user study, we also wanted to understand how the perception of facial expressions is affected by the preceding or succeeding facial expressions, as well as co-occurring ones. Graph 6.3 and Table C.2, show the average comprehension scores of glosses, comprehension scores of sentence types and quality scores of facial expressions for all sections with facial expressions. We conducted a Spearman's rank to assess the relationship between glosses comprehension scores and sentence types comprehension scores. There is **no significant bivariate association between the comprehension of glosses and the comprehension of sentence types** ($r_s = -0.123, p = 0.676$). Based on these results, the glosses comprehension was not affected by the comprehension of sentence types which means that the com-

prehension of phonological facial expressions was not affected by the comprehension of syntactic facial expressions. This demonstrates that **our approach for combining co-occurring phonological and syntactic blend shapes, as described in Section 5.3.5.D, was effective**. Furthermore, as also shown in the graph and table, the sections that had lower scores for comprehension of sentence types and facial expressions quality were all sections that contained interrogatives, in particular, sections that contained polar questions.

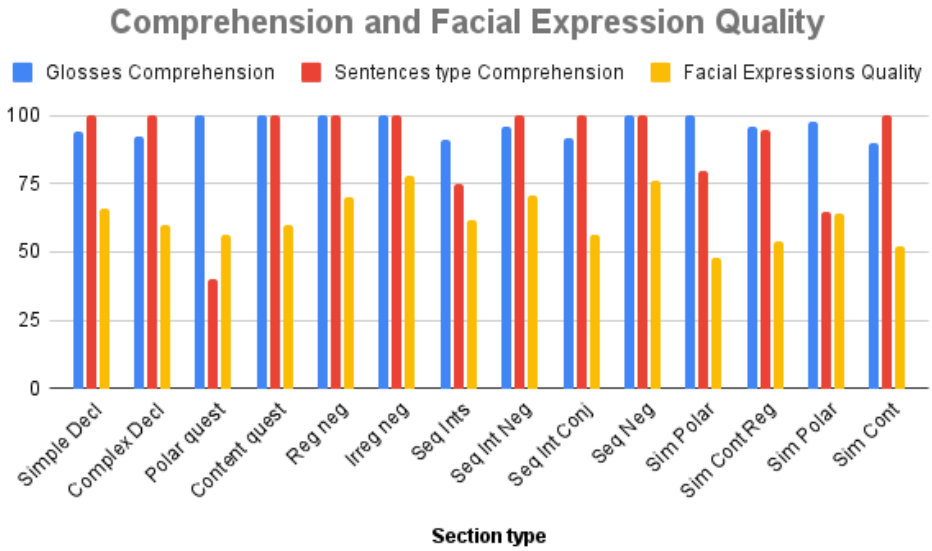


Figure 6.3: Average comprehension scores and facial expression quality for all sections.

Looking closely at each participant's results individually, most participants commented on the **lack of facial expression in polar and content questions**, and that content questions were understood by the syntactic order and not by the syntactic facial expression. Based on these results, we could clearly understand that interrogatives needed to be improved but we could not understand how, hence, the importance of the interviews. In the interviews, all participants unanimously commented that interrogatives lacked an intense facial expression, in particular, the avatar should have a more exaggerated "narrowed eyes" expression. Furthermore, the **frontal perspective of the camera also hindered the comprehension of interrogatives** as most participants did not notice the facial and corporal movements. However, participants commented that even though the **facial expressions should be more exaggerated**, facial and corporal movements should not be exaggerated as to not create unrealistic movements.

As shown in Table C.2 and looking closely at each participant's results individually, we can also notice that the comprehension of sentence types in sections that contain co-occurring syntactic facial expressions (i.e., the last four sections) was not significantly lower than other sections, and scores were solely affected by the perception of interrogatives. This means that **our approach for combining co-**

occurring syntactic facial expressions (i.e., negative and interrogative sentences), as described in Section 5.3.5.D, was effective. With this user study, we can also conclude that in co-occurring syntactic facial expressions, the blend shape of interrogatives should always be animated as this is required for users to identify interrogatives.

6.1.3.D Mouthing

Opinions regarding integrating mouthing in signing animations were mixed. Some believe that a signing avatar should not produce mouthing while signing because it derives from spoken languages, whereas others believe it should be incorporated as most Deaf signers use it. However, all agreed that **mouthing does enhance comprehension** and should be incorporated in some cases, for instance, signs that require mouthing (sign "NÃO_HAVER" and sign "BÁSICO" since it is what differentiates it from sign "BASE") and while fingerspelling. In the questionnaires, one participant noted in animations, a lack of mouthing while fingerspelling names, and in the interviews one participant commented that when interpreting LGP, most Deaf signers tell her to incorporate more mouthing while signing as it makes it easier to understand.

6.1.4 Discussion

Based on the previously reported findings, we can make the final conclusions:

1. **Does the inclusion of non-manual components enhance the linguistic comprehension of Sign Language animations?**

   Sentences that incorporated facial expressions had higher comprehension scores than sentences without facial expressions. Therefore, our study suggests that non-manuals can indeed enhance linguistic comprehension at a phonological and syntactic level, and can effectively convey different types of interrogatives and negatives. However, it was noted that facial expressions for interrogatives should be more exaggerated to enhance comprehension, but facial and corporal movements should not be exaggerated as to not create unrealistic movements.

2. **Does the sequential or co-occurrence of facial expressions have an impact on linguistic comprehension?**

   To the best of our knowledge, this was the first study that analyzed the synthesis of simultaneous phonological and syntactic facial expressions, and co-occurring syntactic facial expressions (i.e., a negative and interrogative sentence). Based on our results, the glosses comprehension was not affected by the comprehension of sentence types which means that the comprehension of phonological facial expressions was not affected by the comprehension of syntactic facial expressions.

This demonstrates that our approach for combining co-occurring phonological and syntactic blend shapes was effective. Furthermore, also based on our results, comprehension of sentence types for sequential and co-occurring syntactic facial expressions was not significantly lower than other sections and scores were solely affected by the perception of interrogatives. This demonstrates that our approach for combining co-occurring syntactic facial expressions was effective.

Our study suggests that in co-occurring syntactic facial expressions, body and facial movements of both expressions should be animated but only the blend shape expression of interrogatives must be animated as without it participants cannot identify interrogatives. The same process applies to simultaneous phonological and syntactic facial expressions, where all facial and body movements are combined, and the phonological expression is applied to the lower part of the face and syntactic to the top part of the face because without the "narrowed eyes" expression, participants cannot identify interrogatives.

Our study provides a pipeline not only for Portuguese Sign Language but also for other Sign Languages because even though syntactic and phonological facial expressions might differ for other languages, these also incorporate polar questions that cannot be understood from the syntactic order or syntactic constituents, but rather from syntactic facial expressions. Therefore, the synthesis of signing animations for all languages should prioritize the facial expression of interrogatives in co-occurrence situations.

6.2 Transitions Evaluation

The second user study was conducted to answer the following research question:

- **RQ1:** Do dynamic transitions have an impact on linguistic comprehension, optimal transition speed, naturalness, and preference of Sign Language animations?

6.2.1 Participants

We recruited 11 participants fluent in LGP with the help of the Católica team and the snowballing sampling technique. For this study, we needed people fluent in LGP that have the necessary knowledge of LGP's prosody to be able to identify the impact transitions can have on linguist comprehension. Seven of these participants are the same as the ones in the first user study, the demographic information for the new four participants is shown in Table B.2.

6.2.2 Procedure

For this user study, we only conducted a quantitative evaluation that consisted of questionnaires. We conducted within-subject user tests where each participant tested all conditions because we did not want individual differences to affect our results. Prior to participating in the user studies, each new participant was handed a thorough consent form which they had to sign to participate in the study.

The questionnaire consisted of thirteen sentences created based on videos from our LGP corpus and SpreadTheSign. For this user study, we improved some phonological facial expressions and signs based on the feedback received from participants in the previous user study. The level of complexity and difficulty in this second user study is harder than the previous user study because now the duration of transitions between signs is faster and now all sentences contain composite utterances. In this second user study, we wanted to evaluate the impact transitions could have on the phonology of signs, especially, on the phonological assimilation of composite utterances. Therefore, we created 10 sentences that contained one or more composite utterances where some sentences had composite utterances composed of three signs which increases the complexity of sentences. Each two paired sentences had the same composite utterances where one sentence had our dynamic transition approach (Section 5.3.5.B) and the other sentence had a constant transition approach with a constant value of 0.5 seconds. Overall, we had 5 sentences with dynamic transitions and 5 sentences with constant transitions. To mitigate experimental bias, the two-paired sentences were different but contained the same composite utterance, both sentences had similar number of glosses and a similar difficulty level. Based on the feedback from the Católica team and the results from the previous user study, the duration of transitions between signs has been one of the most criticized aspects in our generated animations, therefore, this study was important to evaluate the impact transitions can have on comprehension, naturalness and optimal speed, by comparing dynamic transitions versus constant transitions. To evaluate the participants' preference for the transitions approach, we also created three additional sentences where participants would see two different versions of the same sentence side-by-side, one with dynamic transitions and the other with constant transitions. To mitigate experimental bias, the positions of these videos were varied where sometimes the dynamic approach would appear on the right side and others on the left side.

After the questionnaire was created, each participant received a different version, therefore, we created eleven different versions where, in each version, the condition's order is counterbalanced and the sections' order is random. The questionnaire is composed of fifteen sections in which the first section is regarding demographic information about participants and the following thirteen sections contained each sentence created for this study. In the first 10 sections of the thirteen, participants had to visualize a video, write the content understood, and describe whether the sentence contains an error. Furthermore, for each sentence, they also had to evaluate the transitions' speed in a 1-5 Likert scale with 1 as too slow and 5 as too fast, and evaluate the avatar's naturalness in a 1-5 Likert scale with 1 as robotic and

5 as natural. In the last 3 sections of the thirteen, participants had to select which video they preferred between the two side-to-side videos. In the last section, participants had to evaluate the avatar in terms of many aspects. Similar to the previous user study, we also asked participants to evaluate the speed, transitions, pauses, naturalness, comprehension, general quality, grammatical correctness, signs quality, and facial expressions quality in the same Likert scales as described in the first user study. Additional to these, we also asked participants "How the naturalness of the avatar could be improved", whether they think "transitions between signs affect naturalness" and whether they think "transitions between signs affect comprehension".

6.2.3 Data Analysis and Findings

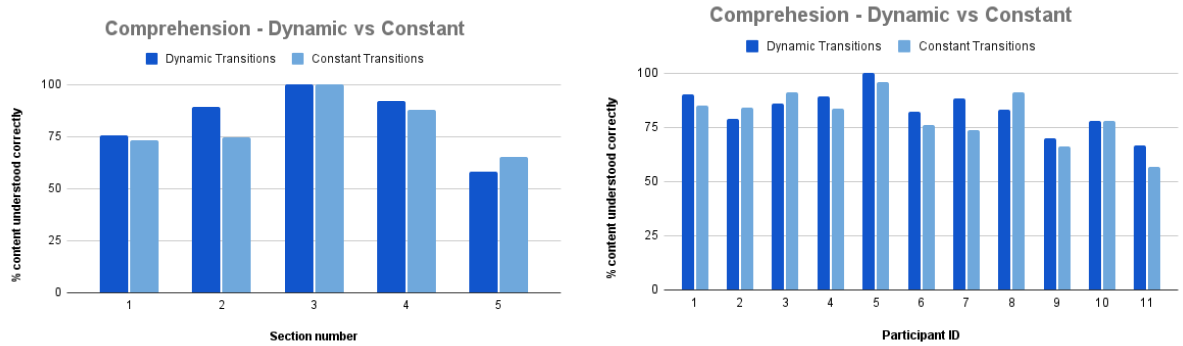
In this user study, we focused on analyzing four metrics: comprehension, transitions' speed, naturalness, and preference.

6.2.3.A Comprehension

For each sentence in the questionnaires, we measured the percentage of content understood by calculating the number of glosses correctly described with 100% as all glosses correctly understood by a participant. This process had to be done manually as synonyms of signs also counted as correct. We reviewed this process three times to make sure all calculations were correct. Overall, the **average comprehension scores for all participants with both conditions was 81.56% (**$SD = 23.29$**)** which we found surprisingly high considering the complexity and difficulty of the sentences.

As shown in Graph 6.4(a) and Table C.3, dynamic transitions had higher or equal comprehension results than constant transitions in all sections except for section 5. Looking at the individual results of participants, we have noticed that one of the sentences in section 5 had a slightly higher difficulty level than the other because there is one sign that could have multiple meanings when used in different contexts. There was only one participant that understood correctly all glosses in this sentence and only four participants that understood the correct meaning of this sign. Therefore, we can conclude that in this section, the comprehension of participants could have been influenced by the context of the sentence and how participants interpreted the meaning of this sign.

As shown in Graph 6.4(b) and Table C.4, 7 participants had higher comprehension results in sentences with dynamic transitions, 3 participants had higher comprehension results with constant transitions and 1 participant had equal comprehension results in both transitions. According to a Shapiro-Wilk test, we retained the null hypothesis of population normality ($p = 0.901, p = 0.722$), therefore, we conducted a Paired samples T-test to compare differences in comprehension scores between our conditions. Based on the results, **there was no significant difference** ($t(10) = -1.379, p = 0.198$) in the scores for dynamic transitions ($M = 82.97, SD = 9.43$) and constant transitions ($M = 80.15, SD = 11.55$).



(a) Comprehension scores per section.

(b) Comprehension scores per participant.

Figure 6.4: Comprehension scores between dynamic transitions and constant transitions.

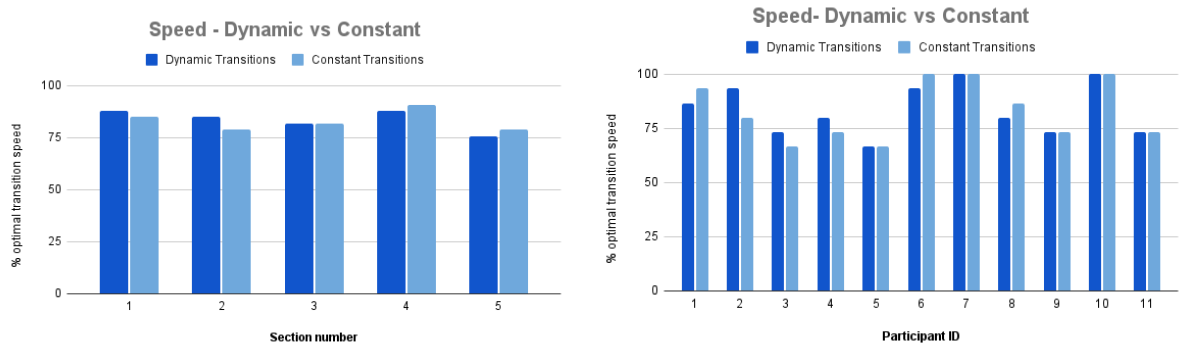
Looking closely at each participant's results individually, we wanted to see if transitions had an impact, in particular, on the comprehension of composite utterances or the comprehension of any other sign that was displayed in both approaches. In almost all cases, participants would either understand a sign or not, independently of the transition approach, which could be explained by the fact that the difference between transition values of both approaches is not significant. However, there were 4 cases in the two-paired sentences (i.e., 8 sentences) where the same sign was perceived correctly with a dynamic approach and not the constant approach but, there were no cases where a sign was perceived correctly with a constant approach and not with a dynamic. Furthermore, 7 participants believe transitions between signs do indeed have an impact on comprehension, whereas only 4 participants believe they do not.

6.2.3.B Transitions Speed

For each sentence in the questionnaires, we measured the percentage of optimal transition speed by using the scores submitted in the Likert scale (i.e., 1 as too slow and 5 as too fast) as 3 being the optimal speed with 100% and decreasing the percentage value according to the closeness to the limits of the scale with 2 and 4 as 66.67% and 1 and 5 as 33.33%. This process was done manually and reviewed three times to make sure all calculations were correct. Overall, the average optimal transition speed scores for all participants with both conditions was 83.64% ($SD = 17.36$) and the average **overall quality of transitions** given at the end of the questionnaire by all participants was 81.82% ($SD = 17.41$) which **was significantly higher compared to the average overall quality of transitions scores given in the first user study** (i.e., 63.33%). Furthermore, even though the speed of signs remained the same compared to the previous user study, the average overall quality of signs speed also increased with an average 87.88% ($SD = 18.82$) which was also significantly higher compared to the one in the first user

study (i.e., 76.67%).

As shown in Graph 6.5(a) and Table C.5, the optimal transition speed scores were distributed evenly throughout all sections, containing 2 sections with higher scores for dynamic transitions, 2 sections with higher scores for constant transitions, and 1 section with equal scores. It is interesting to note that the section with the lowest optimal transition speed scores corresponds to section 5 which is also the section with the lowest scores on comprehension.



(a) Optimal transition speed scores per section.

(b) Optimal transition speed scores per participant.

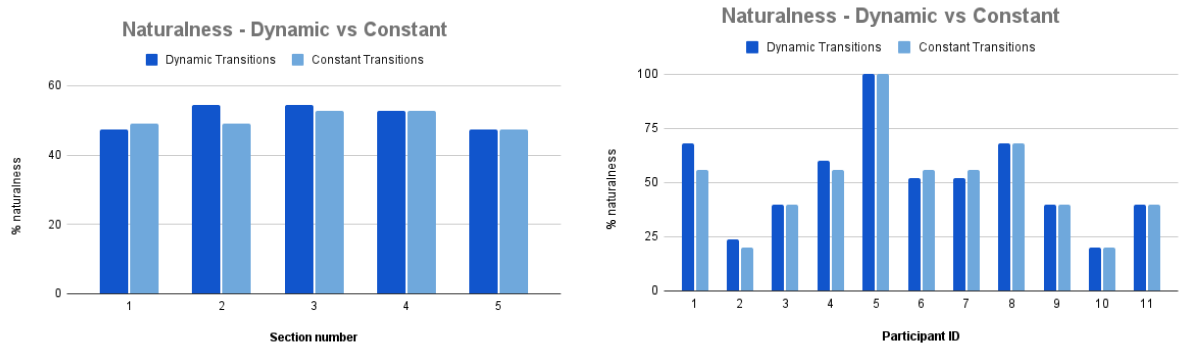
Figure 6.5: Optimal transition speed scores between dynamic transitions and constant transitions.

As shown in Graph 6.5(b) and Table C.6, 3 participants had higher optimal transition speed results in sentences with dynamic transitions, 3 participants had higher optimal transition speed results with constant transitions and 5 participants had equal optimal transition speed results in both transitions. According to a Shapiro-Wilk test, we retained the null hypothesis of population normality ($p = 0.283, p = 0.064$), therefore, we conducted a Paired samples T-test to compare differences in optimal transition speed scores between our conditions. Based on the results, **there was no significant difference** ($t(10) = -0.319, p = 0.756$) in the scores for dynamic transitions ($M = 83.64, SD = 11.68$) and constant transitions ($M = 83.032, SD = 13.45$). However, three participants commented on the importance of faster transitions in-between signs that comprise one sole meaning and noted that constant transitions were too slow for composite utterances, and surprisingly, in negatives. The latter is one aspect we did not take into consideration but coincidentally our dynamic approach produced faster transitions between the negated verb and the “NÃO” sign because the difference between hand locations of these signs is quite small. This difference is what allowed participants to note that constant transitions were too slow for transitions between the negated verb and the “NÃO” sign while our dynamic approach was optimal. We can conclude that dynamic transitions might have a positive impact on the optimal speed and that **signs that comprise one sole meaning** (i.e., composite utterances and negation of verbs) **should have faster transitions than other signs**.

6.2.3.C Naturalness

For each sentence in the questionnaires, we measured the percentage of naturalness by using the scores submitted in the Likert scale (i.e., 1 as robotic and 5 as natural) with 5 being 100%. This process was done manually and reviewed three times to make sure all calculations were correct. Overall, the **average naturalness scores for all participants with both conditions was 50.73% (**$SD = 22.78$**)** and the average overall naturalness given at the end of the questionnaire by all participants was 50.91% ($SD = 25.87$). The scores for naturalness were significantly lower than scores for the other two measures which is not surprising because **naturalness is the most demanding criterion of all**.

As shown in Graph 6.6(a) and Table C.7, dynamic transitions had higher naturalness scores in 2 sections, constant transitions had higher naturalness scores in 1 section and there were 2 sections with the same scores for both approaches. It is interesting to note that again section 5 had the lowest scores as it was noticed on the other measures.



(a) Naturalness scores per section.

(b) Naturalness scores per participant.

Figure 6.6: Naturalness scores between dynamic transitions and constant transitions.

As shown in Graph 6.6(b) and Table C.8, there were **large discrepancies between naturalness scores throughout our participants** with 20% as the lowest average score and 100% as the highest score. Furthermore, 3 participants had higher naturalness results in sentences with dynamic transitions, 2 participants had higher naturalness results with constant transitions and 6 participants had equal naturalness results in both transitions. According to a Shapiro-Wilk test, we retained the null hypothesis of population normality ($p = 0.548, p = 0.215$), therefore, we conducted a Paired samples T-test to compare differences in naturalness scores between our conditions. Based on the results, **there was no significant difference** ($t(10) = -0.820, p = 0.432$) in the scores for dynamic transitions ($M = 51.27, SD = 22.61$) and constant transitions ($M = 50.18, SD = 22.51$). However, 7 participants believe transitions between signs have an impact on naturalness, whereas only 4 participants believe they do not.

When asked “How the naturalness could be improved”, One participant said: “Do not evaluate and correct naturalness only by the execution of signs, but also see the sentence as a whole and add small pauses or accelerations between signs, depending on the sentence and its meaning. Since LGP is a visual language, everything you see counts”. The answers of participants regarding suggestions to improve naturalness were unanimous and revolved on the following aspects: (1) Add more facial expression throughout the sentence and not only when this is applied to signs individually (2) There is a lack of corporal movement, (3) Add small pauses in appropriate places (this step is linked to prosody and we were suggested to incorporate topicalization), (4) improve grammar and signs, (5) add more fluidity in the movements, (6) one participant commented that “the robotic appearance can be caused by the disproportional body of the avatar”.

The first aspect could be improved by adding emotions to the avatar. The second aspect could be improved by incorporating the secondary facial and body movements we described in Section 5.3.5.F, but unfortunately, we did not have time to conduct a fourth user study. Based on the comments made from participants, we conducted a Spearman’s rank to assess the relationship between facial expressions and naturalness, and between comprehension and naturalness. There is a **statistically significant bivariate association between quality of facial expressions and naturalness** ($r_s = 0.782, p = 0.004$) with a strong magnitude and positive correlation at the 0.01 level. There is also a **statistically significant bivariate association between comprehension and naturalness** ($r_s = 0.621, p = 0.042$) with a strong magnitude and positive correlation at the 0.05 level.

6.2.3.D Preference

We conducted a Chi-Square test to analyze which transitions approach was preferred on the three trials each participant had, therefore, there were 33 trials overall. We found a **statistically significant relation between participants and the transition approach** ($X^2(1, N = 33) = 6.818, p = .009$), as participants **preferred more dynamic transitions** ($N = 24$) than constant transitions ($N = 9$).

6.2.4 Discussion

Based on the previously reported findings, we can make the final conclusions

1. **Do dynamic transitions have an impact on linguistic comprehension, optimal transition speed, naturalness, and preference of Sign Language animations?**

   The null hypothesis was reattained in the evaluation of comprehension, transitions’ speed, and naturalness, therefore, we can conclude that the results were similar for both transition approaches. Nevertheless, we found particular cases where the same signs with the dynamic approach were perceived correctly and with the constant approach perceived incorrectly, but the opposite was

not found. Therefore, dynamic transitions could enhance linguistic comprehension, in particular, for signs that comprise one sole meaning (i.e., composite utterances and negatives) and require faster transitions. The dynamic transitions approach was also the approach most preferred by our participants which shows the positive impact they can have on animations.

Regarding naturalness, neither approach had a significant impact and this criterion is still the most demanding of all. We found a positive association between facial expressions and naturalness, and between comprehension and naturalness. It is interesting to note that participants tend to relate naturalness to the comprehension of animations, having the sections with the lowest scores in comprehension also the sections with the lowest scores in naturalness. Furthermore, it is also interesting to note that naturalness is not only linked to comprehension but also to syntax, because sentences that were completely understood but were not correct in terms of grammar, also scored lower in naturalness. The reasoning behind this is that errors in grammar make the translator still seem signed Portuguese and not LGP which makes it an unnatural reading for participants.

6.3 Mouthing Evaluation

The third user study was conducted to answer the following research question:

- **RQ1:** Does mouthing have an impact on linguistic comprehension, naturalness, and preference of Sign Language animations?

6.3.1 Participants

We recruited 20 participants that are learning LGP because we want to create a system that is inclusive for all and can be used as a learning tool. Recruiting beginners for this third user study was essential because we wanted people that had sufficient knowledge to understand some signs but not all so that we could evaluate whether mouthing could indeed have an impact on comprehension. These participants were recruited with the help of a LGP teacher that participated in our other two user studies, we also used the snowballing sampling technique and contacted participants that enrolled in LGP courses.

6.3.2 Procedure

For this user study, we conducted a quantitative evaluation that consisted of questionnaires. We conducted within-subject user tests where each participant tested all conditions because we did not want individual differences to affect our results. Prior to participating in the user studies, each participant was handed a thorough consent form which they had to sign to participate in the study.

The questionnaire consisted of thirteen sentences created based on videos from our LGP corpus and SpreadTheSign. For this user study, we removed all phonological facial expressions from signs so that all signs could execute mouthing. Furthermore, we created some new signs and improved the interrogative facial expression based on the feedback received from participants in the first user study. When recording all sentences, we also strategically lowered the overall speed of signs and transitions and added more paused transitions so as not to hinder the comprehension of animations. The level of complexity and difficulty in this third user study were lower than the previous user studies but not too easy so that we could see the impact of mouthing.

We created 10 sentences where each two-paired sentences contained one sentence with mouthing and the other without. Overall, we had 5 sentences with mouthing and 5 without. To mitigate experimental bias, the two-paired sentences were different but contained some signs in common, both sentences had similar number of glosses and a similar difficulty level. Based on the opinions from our participants in the first user study, there was some mixed feedback on the incorporation of mouthing in the avatar but all agreed that mouthing could indeed have an impact on comprehension. To the best of our knowledge, research in the field has not yet been published on whether mouthing can improve comprehension, therefore, this user study can give valuable input into this topic. To evaluate the participants' preference on animations with or without mouthing, we also created three additional sentences where participants would see two different versions of the same sentence side-by-side, one with mouthing and the other without. To mitigate experimental bias, the positions of these videos were varied where sometimes the mouthing animation would appear on the right side and others on the left side.

After the questionnaire was created, each participant received a different version, therefore, we created twenty different versions where, in each version, the condition's order and phrases' order is counterbalanced and the sections' order is random. The questionnaire is composed of fifteen sections in which the first section is regarding demographic information about participants and the following thirteen sections contained each sentence created for this study. In the first 10 sections of the thirteen, participants had to visualize a video, write the content understood, and evaluate the avatar's naturalness in a 1-5 Likert scale with 1 as robotic and 5 as natural. In the last 3 sections of the thirteen, participants had to select which video they preferred between the two side-to-side videos. In the last section, participants had to evaluate the avatar in terms of many aspects. Similar to the previous user study, we also asked participants to evaluate the naturalness, comprehension, general quality, signs quality, and facial expressions quality in the same Likert scales as described in the second user study. Additional to these, we also asked participants whether they think "mouthing affects naturalness", whether they think "mouthing affects comprehension", and "If yes, in which situations and why?"

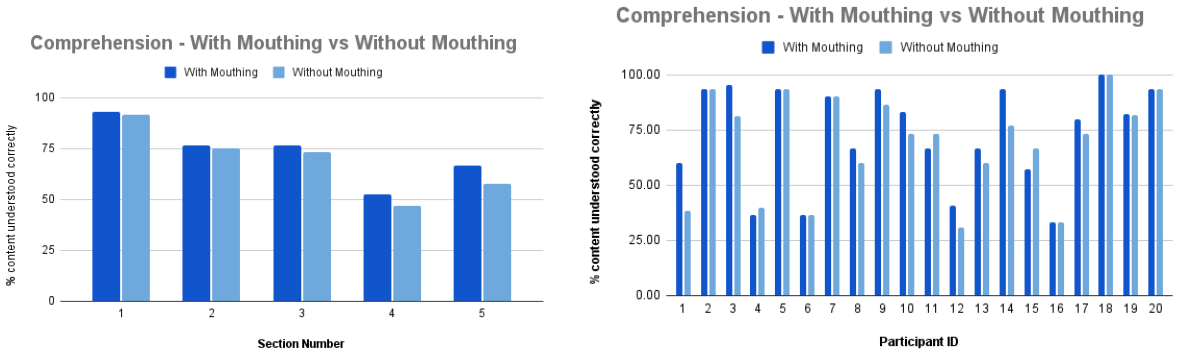
6.3.3 Data Analysis and Findings

In this user study, we focused on analyzing three metrics: comprehension, naturalness, and preference.

6.3.3.A Comprehension

For each sentence in the questionnaires, we measured the percentage of content understood by calculating the number of glosses correctly described with 100% as all glosses correctly understood by a participant. This process had to be done manually as synonyms of signs also counted as correct. We reviewed this process three times to make sure all calculations were correct. Overall, the **average comprehension scores for all participants with both conditions was 70.94% (**$SD = 37.88$**)** which we found surprisingly high considering that participants were beginners and sentences had a level of complexity and difficulty higher than beginner level with some sentences composed by interrogatives, one composite utterance (i.e., sign “IRMÃ”) and dactylology words comprised of numbers with 2 digits and names with 7 letters.

As shown in Graph 6.7(a) and Table C.9, **sentences with mouthing had higher comprehension results than sentences without mouthing in all sections**. Furthermore, sections 4 and 5 were the ones with the lowest scores which is not surprising considering these were the most difficult ones with section 4 containing composite utterances (i.e., sign “IRMÃ”) and section 5 containing interrogatives.



(a) Comprehension scores per section.

(b) Comprehension scores per participant.

Figure 6.7: Comprehension scores between animation with mouthing and without mouthing.

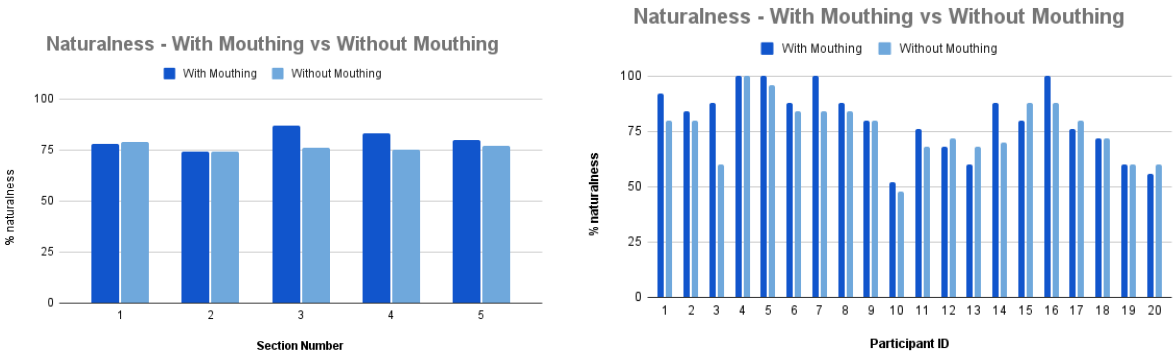
As shown in Graph 6.7(b) and Table C.10, there were large discrepancies between comprehension scores throughout our participants with 33.33% as the lowest average score and 100% as the highest score. Furthermore, 10 participants had higher comprehension results in sentences with mouthing, 3 participants had higher comprehension results without mouthing and 7 participants had equal comprehension results in both. According to a Shapiro-Wilk test, we rejected the null hypothesis of population

normality ($p = 0.012, p = 0.050$), therefore, we conducted a Wilcoxon signed-rank test to compare differences in comprehension scores between our conditions. Based on the results, the **comprehension scores for sentences with mouthing were statistically significantly higher** than for sentences without mouthing ($Z = -2.029, p = 0.043$). Furthermore, 16 participants believe mouthing does indeed have an impact on comprehension, whereas only 4 participants believe it does not. Additionally, many comments were made by participants throughout the questionnaires noting that mouthing makes it easier to understand the sentences. We can, therefore, conclude that **mouthing can indeed have a positive impact on comprehension**.

6.3.3.B Naturalness

For each sentence in the questionnaires, we measured the percentage of naturalness by using the scores submitted in the Likert scale (i.e., 1 as robotic and 5 as natural) with 5 being 100%. This process was done manually and reviewed three times to make sure all calculations were correct. Overall, the **average naturalness scores for all participants with both conditions was 78.29% (**$SD = 16.91$**)** and the average overall naturalness given at the end of the questionnaire by all participants was 78.95% ($SD = 15.60$). The scores for naturalness were significantly higher than scores given in the first and second user studies which can be explained by the fact that participants in this study are still beginners and are still not sensible to all subtleties of Sign Languages, therefore, do not notice aspects that might still be missing in our avatar.

As shown in Graph 6.8(a) and Table C.11, there were 3 sections where sentences with mouthing had higher naturalness results, 1 section where sentences without mouthing had higher naturalness results and 1 section with equal results.



(a) Naturalness scores per section.

(b) Naturalness scores per participant.

Figure 6.8: Naturalness scores between between animation with mouthing and without mouthing.

As shown in Graph 6.8(b) and Table C.12, 11 participants had higher naturalness results in sen-

tences with mouthing, 5 participants had higher naturalness results without mouthing and 4 participants had equal naturalness results in both. According to a Shapiro-Wilk test, we retained the null hypothesis of population normality ($p = 0.160, p = 0.793$), therefore, we conducted a Paired samples T-test to compare differences in naturalness scores between our conditions. Based on the results, the **naturalness scores for sentences with mouthing** ($M = 80.40, SD = 15.24$) **were statistically significantly higher** ($t(19) = -2.094, p = 0.050$) than for sentences without mouthing ($M = 76.10, SD = 13.11$). Furthermore, 18 participants believe mouthing has an impact on naturalness, whereas only 2 participants believe it does not. We can, therefore, conclude that **mouthing can indeed have a positive impact on naturalness**.

6.3.3.C Preference

We conducted a Chi-Square test to analyze which animations were preferred on the three trials each participant had, therefore, there were 60 trials overall. There was a **statistically significant relation between participants and the mouthing approach** ($X^2(1, N = 60) = 15, p = 0.000108$), as participants **preferred more animations with mouthing** ($N = 45$) than animations without ($N = 15$). One participant said that "mouthing can distract the participant from the signs" as being the reason for not choosing animations with mouthing.

6.3.4 Discussion

Based on the previously reported findings, we can make the final conclusions:

1. **Does mouthing have an impact on linguistic comprehension, naturalness, and preference of Sign Language animations?**

   Sentences that incorporated mouthing had higher comprehension and naturalness scores than sentences without mouthing. Therefore, our study suggests that mouthing can indeed enhance linguistic comprehension and naturalness, and participants prefer Sign Language animations with mouthing. It is interesting to note that based on results from this user study and interviews from our first user study, there are specific cases where mouthing supports comprehension and where most signers incorporate mouthing: (1) signs that require mouthing/visual morphemes, for instance, signs "NÃO_HAVER" and "BÁSICO", (2) while fingerspelling. It was noted by one participant, the lack of mouthing while fingerspelling a name and noted by another participant, that the lack of mouthing hindered comprehension while fingerspelling the conjunction "OU", (3) interrogative pronouns/adverbs should also incorporate mouthing.

   Our user study demonstrates not only the impact mouthing has on signing animations but also that the quality of our mouthing approach was good enough to improve comprehension.

6.4 Final Conclusions

Based on the reported findings for all three user studies, we can make the final conclusions:

1. **Is an automatic text-to-sign translator effective in generating realistic and natural Portuguese Sign Language animations?**

   These were the first user studies conducted to evaluate our signing animations, so we were surprised by the overall good performance and positive feedback from our 34 participants. The average comprehension score for all three user studies was 83.83% which shows that our translator was effective in generating Sign Language animations that were understood by not only people fluent in LGP but also beginners. Overall, the quality of our animations had an average score of 69.82% for all three studies, and an average naturalness score of 60.64%, so even though these scores were above 50%, our animations still need improvements. Some suggestions included adding more facial expressions, adding corporal movements, adding appropriate pauses and accelerations between signs, and creating more fluid movements.

   Our translator shows great potential in the field of synthetic animation of signing avatars and demonstrates components that can be applied not only for Portuguese Sign Language but also for other Sign Languages. For instance, a pipeline for the synthesis of co-occurring facial expressions, a dynamic approach for transitions in-between signs, the generation of automatic secondary facial and corporal movements, and the integration and synthesis of mouthing animations.

7

Conclusions

Contents

7.1 Achievement and Limitations

In this dissertation, we presented an approach that consists in the synthesis and simultaneous animation of manual and non-manual components, and secondary facial and corporal movements. The manual and non-manual components account for the morphosyntactic motions needed in Sign Languages and the secondary movements account for the naturalness of the avatar. Our approach provides a pipeline that can be used for multiple digital applications, for instance: an automatic text-to-sign language translator, a dictionary, a book translator, a virtual assistant, and a browser add-on. In this dissertation, we used a text-to-sign language translator to demonstrate our generated animations.

We conducted three user studies with a total of 34 participants to evaluate our generated signing animations. The overall good performance and positive feedback indicate that the generated animations by our translator show great potential in the field of synthetic animation of signing avatars. In this dissertation, we introduced components that can be applied not only for Portuguese Sign Language but also for other Sign Languages. For instance, a pipeline for the synthesis of co-occurring facial expressions, a dynamic approach for transitions in-between signs, the generation of automatic secondary facial and corporal movements, and the integration and synthesis of mouthing animations.

The WebGL platform has some limitations that restrict the overall potential of our tool: (1) *.NET* networking classes are not functional in WebGL, (2) the scripts Unity provides for the manipulation of animation resources in runtime are included in the package *UnityEditor*, which cannot be included in the WebGL build, (3) In a WebGL build, the *System.Text.Json* class only works with reference types and not generic functions with value types.

7.2 Future work

The proposed system is the first version of the synthesis of signing animations generated by a translator based on linguistic information extracted from a corpus. This approach brings the state-of-the-art one step closer to an automatic Portuguese to LGP translator. Although the results obtained are good, there are some aspects to be improved and extended, especially, in terms of naturalness. These are some of the points identified:

1. Conduct a new user study with people fluent in LGP to assess the perception of the **improved interrogative facial expression**.

2. Add more **facial expression** throughout the entire sentence and not only in certain signs, for instance, by creating an algorithm that identifies the emotion in a sentence and adds that corresponding emotion in the animation.

3. Conduct a new user study to evaluate the implemented **secondary facial and corporal movements** (Section 5.3.5.F).

4. The **torso rotation** (Section 5.3.5.F), is only applied to the right hand because the IK system lacks control over the spine rotation. To rotate the spine correctly according to the reach of both arms we would have to replace the IK system with the FK system and rotate directly the spine joint. This method would also allow the head and torso movements and the torso rotation to be simultaneously produced.

5. Create the **collision between hands and chest**, and the **collision between hands** that were described in Section 5.3.5.G.

6. Research the **prosody in LGP** to identify prosodic properties (e.g., eyes blinking, head tilts), appropriate pauses within clauses, and accelerations between signs (perhaps related to topicalization).

7. Further research the **usage of body rotation** in LGP to **mark verb tense** and **role shift** (i.e., reenacting the subject and object in sentences by shifting the body and switching between dominant hands).

8. Extend the Translation process to identify **classifiers, prepositions, and agreement verbs** that have an impact on the hand configurations, orientations, and locations of signs. A possible implementation in the Animation process would be to receive these identified components and then change the hand configurations of a sign in run-time by using a new layer in the animator controller and change the hand locations of a sign in run-time by using the IK system.

9. Create **more translation rules** in the Translation Rules Construction module (Section 5.2) by extracting new data from the corpus. The more rules created, the better as it prevents grammatical phenomena from not being considered in the appropriate rules.

Bibliography

[1] H. Carmo, V. M. da Silva, and E. Martins, “Os verbos em negação na língua gestual portuguesa,” *Cadernos de Saúde*, vol. 9, pp. 15–25, 2017.

[2] Z. Xu, Y. Zhou, E. Kalogerakis, C. Landreth, and K. Singh, “Rignet: Neural rigging for articulated characters,” *arXiv preprint arXiv:2005.00559*, 2020.

[3] M. Romeo and S. Schvartzman, “Data-driven facial simulation,” in *Computer Graphics Forum*, vol. 39, no. 6. Wiley Online Library, 2020, pp. 513–526.

[4] S. Gibet, N. Courty, K. Duarte, and T. L. Naour, “The signcom system for data-driven animation of interactive virtual signers: Methodology and evaluation,” *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 1, no. 1, pp. 1–23, 2011.

[5] C. L. Garberoglio, J. L. Palmer, S. W. Cawthon, and A. Sales, “Deaf people and employment in the united states: 2019,” Tech. Rep., 2019.

[6] S. Qi and R. E. Mitchell, “Large-scale academic achievement testing of deaf and hard-of-hearing students: Past, present, and future,” *Journal of deaf studies and deaf education*, vol. 17, no. 1, pp. 1–18, 2012.

[7] M. Kipp, Q. Nguyen, A. Heloir, and S. Matthes, “Assessing the deaf user perspective on sign language avatars,” in *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, 2011, pp. 107–114.

[8] R. Elliott, J. R. Glauert, J. Kennaway, and I. Marshall, “The development of language processing support for the visicast project,” in *Proceedings of the fourth international ACM conference on Assistive technologies*, 2000, pp. 101–108.

[9] N. Adamo-Villani, “3d rendering of american sign language finger-spelling: a comparative study of two animation techniques,” *International journal of human and social sciences*, vol. 3, no. 4, p. 24, 2008.

[10] M. Gonçalves, “PE2LGP 4.0: de português europeu para língua gestual portuguesa,” Master’s thesis, 2020.

[11] M. Huenerfauth, “A linguistically motivated model for speed and pausing in animations of american sign language,” *ACM Transactions on Accessible Computing (TACCESS)*, vol. 2, no. 2, pp. 1–31, 2009.

[12] S. Al-khazraji, L. Berke, S. Kafle, P. Yeung, and M. Huenerfauth, “Modeling the speed and timing of american sign language to generate realistic animations,” in *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, 2018, pp. 259–270.

[13] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef *et al.*, “Sign language recognition, generation, and translation: An interdisciplinary perspective,” in *The 21st international ACM SIGACCESS conference on computers and accessibility*, 2019, pp. 16–31.

[14] T. M. M. d. M. Martins, “A letra e o gesto: estruturas linguísticas em língua gestual portuguesa e língua portuguesa,” Master’s thesis, 2011.

[15] I. R. Almeida, “Exploring challenges in avatar-based translation from european portuguese to portuguese sign language,” *Diss. Master’s thesis, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal*, vol. 2104, 2014.

[16] L. R. Gaspar, “lf2lgp-intérprete automático de fala em língua portuguesa para língua gestual portuguesa,” Ph.D. dissertation, 2015.

[17] R. Ferreira, “Pe2lgp 3.0: from european portuguese to portuguese sign language,” 2018.

[18] P. Escudeiro, N. Escudeiro, R. Reis, J. Lopes, M. Norberto, A. B. Baltasar, M. Barbosa, and J. Bidarra, “Virtual sign–a real time bidirectional translator of portuguese sign language,” *Procedia Computer Science*, vol. 67, pp. 252–262, 2015.

[19] I. Mesquita and S. Silva, “Guia prático de língua gestual portuguesa: Ouvir o silêncio,” *Nova Educação, Braga*, 2007.

[20] A. Morais, J. C. Jardim, A. Silva, and A. Mineiro, “Para além das mãos: elementos para o estudo da expressão facial (ef) em língua gestual portuguesa (lgp),” *Cadernos de Saúde, Vol 4, nº 1, 2011*, vol. 4, pp. 37–42, 2011.

[21] M. A. Amaral, A. Coutinho, M. R. D. Martins, and R. Johnson, *Para uma gramática da língua gestual portuguesa*, 1994.

[22] E. Gonçalves and M. J. C. Raposo, “Expressões faciais gramaticais na morfologia da língua gestual portuguesa: expressões dos graus de tamanho diminutivo e aumentativo na lgp,” *Cadernos de Saúde*, vol. 6, pp. 78–83, 2013.

[23] M. Cruz, M. Swerts, and S. Frota, “Do visual cues to interrogativity vary between language modalities? evidence from spoken portuguese and portuguese sign language,” in *Proc. The 15th International Conference on Auditory-Visual Speech Processing*, pp. 1–5.

[24] C. S. C. Valadares, “Gestuar a história: terminologia específica e interpretação em língua gestual portuguesa,” Master’s thesis, 2012.

[25] M. Martins and A. I. Mata, “Conexões interfrásicas manuais e não-manuais em lgp: Um estudo preliminar,” *Linguística: Revista de Estudos Linguísticos da Universidade do Porto*, vol. 11, pp. 119–138, 2017.

[26] J.-M. Lu, X.-S. Chen, X. Yan, C.-F. Li, M. Lin, and S.-M. Hu, “A rigging-skinning scheme to control fluid simulation,” in *Computer Graphics Forum*, vol. 38, no. 7. Wiley Online Library, 2019, pp. 501–512.

[27] N. Ersotelos and F. Dong, “Building highly realistic facial modeling and animation: a survey,” *The Visual Computer*, vol. 24, no. 1, pp. 13–30, 2008.

[28] Z. Deng and J. Noh, “Computer facial animation: A survey,” in *Data-driven 3D facial animation*. Springer, 2008, pp. 1–28.

[29] H. Kacorri, “A survey and critique of facial expression synthesis in sign language animation,” Department of Computer Science, The City University of New York, Tech. Rep., 2015.

[30] J. Bento, A. Cláudio, and P. Urbano, “Avatares em língua gestual portuguesa,” in *Proc. 9th Iberian Conference on Information Systems and Technologies*, 2014, pp. 185–191.

[31] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin, “Synthesizing realistic facial expressions from photographs,” in *ACM SIGGRAPH 2006 Courses*, 2006.

[32] K. Waters and T. Levergood, “An automatic lip-synchronization algorithm for synthetic faces,” in *Proceedings of The second ACM international conference on Multimedia*, 1994, pp. 149–156.

[33] E. Sifakis, I. Neverov, and R. Fedkiw, “Automatic determination of facial muscle activations from sparse motion capture marker data,” in *ACM SIGGRAPH 2005 Papers*, 2005, pp. 417–425.

[34] D. Terzopoulos and K. Waters, “Physically-based facial modelling, analysis, and animation,” *The journal of visualization and computer animation*, vol. 1, no. 2, pp. 73–80, 1990.

[35] Y. Zhang, “Muscle-driven modeling of wrinkles for 3d facial expressions,” in *2008 IEEE International Conference on Multimedia and Expo*. IEEE, 2008, pp. 957–960.

[36] N. M. Patel and M. Zaveri, “Parametric facial expression synthesis and animation,” *International Journal of Computer Applications*, vol. 3, no. 4, 2010.

[37] K. Waters and J. Frisbie, “A coordinated muscle model for speech animation,” 1995.

[38] K. Waters, “A muscle model for animation three-dimensional facial expression,” *Acm siggraph computer graphics*, vol. 21, no. 4, pp. 17–24, 1987.

[39] C. Girges, J. Spencer, and J. O’Brien, “Categorizing identity from facial motion,” *Quarterly Journal of Experimental Psychology*, vol. 68, no. 9, pp. 1832–1843, 2015.

[40] Z. Deng, P.-Y. Chiang, P. Fox, and U. Neumann, “Animating blendshape faces by cross-mapping motion capture data,” in *Proceedings of the 2006 symposium on Interactive 3D graphics and games*, 2006, pp. 43–48.

[41] B. Choe, H. Lee, and H.-S. Ko, “Performance-driven muscle-based facial animation,” *The Journal of Visualization and Computer Animation*, vol. 12, no. 2, pp. 67–79, 2001.

[42] M. H. Alkawaz, D. Mohamad, A. H. Basori, and T. Saba, “Blend shape interpolation and facs for realistic avatar,” *3D Research*, vol. 6, no. 1, p. 6, 2015.

[43] J. Schnepp, R. Wolfe, J. McDonald, and J. Toro, “Generating co-occurring facial nonmanual signals in synthesized american sign language,” 2013.

[44] O. A. Crasborn, E. Van Der Kooij, D. Waters, B. Woll, and J. Mesch, “Frequency distribution and spreading behavior of different types of mouth actions in three sign languages,” *Sign Language & Linguistics*, vol. 11, no. 1, pp. 45–67, 2008.

[45] R. Wolfe, T. Hanke, G. Langer, E. Jahn, S. Worseck, J. Bleicken, J. C. McDonald, and S. Johnson, “Exploring localization for mouthings in sign language avatars,” in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2018.

[46] H. L. Bear, R. W. Harvey, B.-J. Theobald, and Y. Lan, “Which phoneme-to-viseme maps best improve visual-only computer lip-reading?” in *International Symposium on Visual Computing*. Springer, 2014, pp. 230–239.

[47] J. Glauert, R. Kennaway, R. Elliott, and B.-J. Theobald, “Virtual human signing as expressive animation,” in *Symposium on Language, Speech and Gesture for Expressive Characters, University of Leeds*, 2004, pp. 98–106.

[48] J. Serra, M. Ribeiro, J. Freitas, V. Orvalho, and M. S. Dias, “A proposal for a visual speech animation system for european portuguese,” in *Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2012, pp. 267–276.

[49] J. McDonald, R. Wolfe, J. Schnepp, J. Hochgesang, D. G. Jamrozik, M. Stumbo, L. Berke, M. Bialek, and F. Thomas, “An automated technique for real-time production of lifelike animations of american sign language,” *Universal Access in the Information Society*, vol. 15, no. 4, pp. 551–566, 2016.

[50] M. Gonçalves, L. Coheur, H. Nicolau, and A. Mineiro, “Pe2lgp: tradutor de português europeu para língua gestual portuguesa em glosas,” *Linguamática*, vol. 13, no. 1, pp. 3–21, 2021.

[51] L. Padró and E. Stanilovsky, “Freeling 3.0: Towards wider multilinguality,” in *LREC2012*, 2012.

[52] M. Honnibal and I. Montani, “spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing,” *To appear*, vol. 7, no. 1, pp. 411–420, 2017.

[53] S. Nascimento and M. Correia, “Um olhar sobre a morfologia dos gestos,” *Universidade Católica de*, 2011.

[54] S. Al-khazraji, B. Dingman, and M. Huenerfauth, “Empirical investigation of users’ preferred timing parameters for american sign language animations,” in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–7.

[55] S. Al-khazraji, B. Dingman, S. Lee, and M. Huenerfauth, “At a different pace: Evaluating whether users prefer timing parameters in american sign language animations to differ from human signers’ timing.” *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS’21)*, 2021.

[56] S. Greenberg, H. Carvey, L. Hitchcock, and S. Chang, “Temporal properties of spontaneous speech—a syllable-centric perspective,” *Journal of Phonetics*, vol. 31, no. 3-4, pp. 465–485, 2003.

[57] G. Tang, D. Brentari, C. González, and F. Sze, *Crosslinguistic variation in prosodic cues*. na, 2010.

[58] W. Sandler, “Prosody and syntax in sign languages,” *Transactions of the philological society*, vol. 108, no. 3, pp. 298–328, 2010.

[59] M. E. Tyrone and C. E. Mauk, “The phonetics of head and body movement in the realization of american sign language signs,” *Phonetica*, vol. 73, no. 2, pp. 120–140, 2016.

[60] H. Carmo, M. Moita, and A. Mineiro, “Os classificadores da língua gestual portuguesa (lgp): estudo piloto,” *Leitura*, vol. 1, no. 58, pp. 26–46, 2017.

A

Ethics Document

This appendix contains the approval issued by the Ethics Committee of Instituto Superior Técnico, University of Lisbon.

TÉCNICO LISBOA

Ethis Commission (CE-IST)

Ref. n.º 25/2021 (CE-IST)
Date: 29/09/2021

Name of IR: Hugo Miguel Aleixo Albuquerque Nicolau
Name of the project: Generating Realistic Sign Language Animations.

Prof. Hugo Miguel Aleixo Albuquerque Nicolau

The Ethics Committee of Instituto Superior Técnico (EC-IST) reviewed your application to obtain ethical assessment for the above mentioned project. The following documents have been reviewed:

Ref.	Documents	Version & date
1056517	12345C.pdf consentimento_estudo3.pdf consentimento_estudos1_2.pdf guiao_entrevistas.pdf pedido-parecer-signed.zip pedido-parecer.pdf	23-07-2021

The following members of the EC-IST participated in the ethical assessment:

Name	Role in Ethics Committee	Qualification	Gender	Affiliation to IST (Yes/No)
Mário Gaspar da Silva	President	Professor	M	Y
António Pinheiro	Member	Professor	M	Y
Isabel Sá Correia	Member	Professor	F	Y
Fernando Borges Araújo	Member	Professor	M	N

This EC-IST is working accordance to ICH-GCP, Schedule Y and ICMR guidelines, the EC-IST regulation and other applicable regulation.

None of the researchers participating in this study took part in the decision making and voting procedure for this assessment.

Based on the review of the above mentioned documents, the EC-IST states a an unanimous favourable ethical opinion about the request / trial as submitted.

The EC-IST expects to be informed about the progress of the study, any Serious Adverse Events occurring in the course of the study, any revision in the protocol and in the participants' information/informed consent, and requests to be provided a copy of the final report.

Prof. Mário Gaspar da Silva
President of Ethics Committee of
Instituto Superior Técnico (CE-IST)

B

Demographic Information of Participants

This appendix contains tables from the user studies with the demographic information of participants.

ID	Gender	Relation to LGP	Regional variant
1	M	Deaf native signer	Center
2	F	LGP interpreter	South
3	F	Deaf native signer	North
4	F	Deaf native signer	Center
5	M	native signer + Phd in LGP linguistics	South
6	F	Deaf native signer + Phd in Sign languages	South
7	F	LGP interpreter	Center
8	M	LGP interpreter	Center and South
9	F	LGP teacher	Center
10	F	LGP interpreter	North

Table B.1: Demographic information of participants in user study 1.

ID	Gender	Relation to LGP	Regional variant
8	F	LGP interpreter	North
9	F	LGP interpreter	North
10	F	LGP interpreter	North and Center
11	F	LGP interpreter	North

Table B.2: Demographic information of new participants in user study 2.

C

User Studies

This appendix contains tables from the three user studies with the average scores per section or per participant, discussed in Section 6.

C.1 Linguistic Components Evaluation

Participant ID	With Facial Expression	Without Facial Expression
2	$M = 100, SD = 0$	$M = 92.68, SD = 11.833$
3	$M = 96.82, SD = 4.94$	$M = 83.33, SD = 40.82$
4	$M = 98.48, SD = 3.72$	$M = 88.79, SD = 17.83$
5	$M = 97.73, SD = 5.57$	$M = 92.80, SD = 11.36$
6	$M = 100, SD = 0$	$M = 98.48, SD = 3.72$
7	$M = 93.94, SD = 14.85$	$M = 91.16, SD = 15.05$
8	$M = 100, SD = 0$	$M = 96.97, SD = 7.43$
9	$M = 91.48, SD = 16.05$	$M = 83.06, SD = 15.36$
10	$M = 98.33, SD = 4.08$	$M = 92.80, SD = 9.43$

Table C.1: Glosses comprehension scores per participant.

Section	Glosses Comprehension	Sentence types Comprehension	FE quality
1	$M = 94, SD = 12.65$	$M = 100, SD = 0$	$M = 66, SD = 16.47$
2	$M = 92.07, SD = 11.40$	$M = 100, SD = 0$	$M = 60, SD = 24.94$
3	$M = 100, SD = 0$	$M = 40, SD = 51.64$	$M = 56, SD = 20.66$
4	$M = 100, SD = 0$	$M = 100, SD = 0$	$M = 60, SD = 18.86$
5	$M = 100, SD = 0$	$M = 100, SD = 0$	$M = 70, SD = 19.44$
6	$M = 100, SD = 0$	$M = 100, SD = 0$	$M = 78, SD = 17.51$
7	$M = 91, SD = 19.12$	$M = 75, SD = 26.35$	$M = 62, SD = 22.01$
8	$M = 95.71, SD = 9.64$	$M = 100, SD = 0$	$M = 71, SD = 19.12$
9	$M = 92, SD = 13.98$	$M = 100, SD = 0$	$M = 56, SD = 15.78$
10	$M = 100, SD = 0$	$M = 100, SD = 0$	$M = 76, SD = 18.38$
11	$M = 100, SD = 0$	$M = 80, SD = 25.82$	$M = 48, SD = 21.49$
12	$M = 96, SD = 8.43$	$M = 95, SD = 15.81$	$M = 54, SD = 21.19$
13	$M = 98, SD = 6.32$	$M = 65, SD = 24.15$	$M = 64, SD = 22.71$
14	$M = 90, SD = 12.91$	$M = 100, SD = 0$	$M = 52, SD = 16.87$

Table C.2: Average comprehension scores and facial expression quality for all sections with facial expressions.

C.2 Transitions Evaluation

Section	Dynamic Transitions	Constant Transitions
1	$M = 75.45, SD = 19.42$	$M = 73.18, SD = 23.69$
2	$M = 89.09, SD = 16.40$	$M = 74.55, SD = 15.72$
3	$M = 100, SD = 0$	$M = 100, SD = 0$
4	$M = 91.99, SD = 10.00$	$M = 87.88, SD = 15.99$
5	$M = 58.33, SD = 31.84$	$M = 65.15, SD = 32.45$

Table C.3: Comprehension scores per section.

Participant ID	Dynamic Transitions	Constant Transitions
1	$M = 90, SD = 13.69$	$M = 85, SD = 22.36$
2	$M = 79.14, SD = 23.07$	$M = 84.14, SD = 17.12$
3	$M = 86, SD = 12.94$	$M = 91, SD = 12.45$
4	$M = 89.29, SD = 14.72$	$M = 83.43, SD = 22.71$
5	$M = 100, SD = 0$	$M = 96, SD = 8.94$
6	$M = 82.14, SD = 20.82$	$M = 76.29, SD = 14.62$
7	$M = 88.33, SD = 16.24$	$M = 73.67, SD = 28.30$
8	$M = 83, SD = 17.18$	$M = 91, SD = 12.45$
9	$M = 70, SD = 44.72$	$M = 66, SD = 42.19$
10	$M = 78.14, SD = 18.31$	$M = 78.14, SD = 18.31$
11	$M = 66.67, SD = 42.49$	$M = 57, SD = 29.73$

Table C.4: Comprehension scores per participant.

Section	Dynamic Transitions	Constant Transitions
1	$M = 87.88, SD = 16.82$	$M = 84.85, SD = 17.41$
2	$M = 84.85, SD = 17.41$	$M = 78.79, SD = 22.47$
3	$M = 81.82, SD = 17.41$	$M = 81.82, SD = 17.41$
4	$M = 87.88, SD = 16.82$	$M = 90.91, SD = 15.57$
5	$M = 75.76, SD = 15.57$	$M = 78.79, SD = 16.82$

Table C.5: Optimal transition speed scores per section.

Participant ID	Dynamic Transitions	Constant Transitions
1	$M = 86.67, SD = 13.69$	$M = 93.33, SD = 22.36$
2	$M = 93.33, SD = 14.91$	$M = 80, SD = 29.82$
3	$M = 73.34, SD = 14.91$	$M = 66.67, SD = 0$
4	$M = 80, SD = 18.26$	$M = 73.34, SD = 14.91$
5	$M = 66.67, SD = 0$	$M = 66.67, SD = 0$
6	$M = 93.33, SD = 14.91$	$M = 100, SD = 0$
7	$M = 100, SD = 0$	$M = 100, SD = 0$
8	$M = 80, SD = 18.26$	$M = 86.67, SD = 18.26$
9	$M = 73.34, SD = 14.91$	$M = 73.34, SD = 14.91$
10	$M = 100, SD = 0$	$M = 100, SD = 0$
11	$M = 73.34, SD = 14.91$	$M = 73.34, SD = 14.91$

Table C.6: Optimal transition speed scores per participant.

Section	Dynamic Transitions	Constant Transitions
1	$M = 47.27, SD = 22.40$	$M = 49.09, SD = 22.56$
2	$M = 54.55, SD = 25.44$	$M = 49.09, SD = 22.56$
3	$M = 54.55, SD = 22.07$	$M = 52.73, SD = 24.12$
4	$M = 52.73, SD = 24.12$	$M = 52.73, SD = 24.12$
5	$M = 47.27, SD = 25.73$	$M = 47.27, SD = 22.40$

Table C.7: Naturalness scores per section.

Participant ID	Dynamic Transitions	Constant Transitions
1	$M = 68, SD = 10.95$	$M = 56, SD = 8.94$
2	$M = 24, SD = 8.94$	$M = 20, SD = 0$
3	$M = 40, SD = 0$	$M = 40, SD = 0$
4	$M = 60, SD = 14.14$	$M = 56, SD = 8.94$
5	$M = 100, SD = 0$	$M = 100, SD = 0$
6	$M = 52, SD = 17.89$	$M = 56, SD = 8.94$
7	$M = 52, SD = 10.95$	$M = 56, SD = 8.94$
8	$M = 68, SD = 10.95$	$M = 68, SD = 10.95$
9	$M = 40, SD = 0$	$M = 40, SD = 0$
10	$M = 20, SD = 0$	$M = 20, SD = 0$
11	$M = 40, SD = 0$	$M = 40, SD = 0$

Table C.8: Naturalness scores per participant.

C.3 Mouthing Evaluation

Section	With Mouthing	Without Mouthing
1	$M = 93.05, SD = 18.94$	$M = 91.71, SD = 21.58$
2	$M = 76.67, SD = 32.17$	$M = 74.99, SD = 38.43$
3	$M = 76.67, SD = 34.37$	$M = 73.33, SD = 33.51$
4	$M = 52.50, SD = 43.69$	$M = 46.67, SD = 39.22$
5	$M = 66.67, SD = 39.22$	$M = 57.89, SD = 42.81$

Table C.9: Comprehension scores per section.

Participant ID	With Mouthing	Without Mouthing
1	$M = 60, SD = 54.77$	$M = 38.29, SD = 52.51$
2	$M = 93.33, SD = 14.91$	$M = 93.33, SD = 14.91$
3	$M = 95.52, SD = 7.25$	$M = 81.33, SD = 27.24$
4	$M = 36.67, SD = 50.55$	$M = 40, SD = 54.77$
5	$M = 93.33, SD = 14.91$	$M = 93.33, SD = 14.91$
6	$M = 36.67, SD = 50.55$	$M = 36.67, SD = 50.55$
7	$M = 90, SD = 22.36$	$M = 90, SD = 22.36$
8	$M = 66.67, SD = 47.14$	$M = 60, SD = 43.46$
9	$M = 93.33, SD = 14.91$	$M = 86.67, SD = 18.26$
10	$M = 83.33, SD = 23.57$	$M = 73.33, SD = 27.89$
11	$M = 66.67, SD = 33.34$	$M = 73.33, SD = 27.89$
12	$M = 40.67, SD = 41.39$	$M = 30.67, SD = 33.86$
13	$M = 66.67, SD = 40.82$	$M = 60, SD = 36.52$
14	$M = 93.33, SD = 14.91$	$M = 77.03, SD = 22.90$
15	$M = 57.14, SD = 40.55$	$M = 66.67, SD = 47.14$
16	$M = 33.33, SD = 33.34$	$M = 33.33, SD = 47.14$
17	$M = 80, SD = 44.72$	$M = 73.33, SD = 43.46$
18	$M = 100, SD = 0$	$M = 100, SD = 0$
19	$M = 82.19, SD = 28.15$	$M = 81.62, SD = 27.87$
20	$M = 93.33, SD = 14.91$	$M = 93.33, SD = 14.91$

Table C.10: Comprehension scores per participant.

Section	With Mouthing	Without Mouthing
1	$M = 78, SD = 19.36$	$M = 79, SD = 17.74$
2	$M = 74, SD = 18.47$	$M = 74, SD = 14.65$
3	$M = 87, SD = 14.90$	$M = 76, SD = 12.31$
4	$M = 83, SD = 18.67$	$M = 75, SD = 12.77$
5	$M = 80, SD = 19.47$	$M = 76.84, SD = 17.97$

Table C.11: Naturalness scores per section.

Participant ID	With Mouthing	Without Mouthing
1	$M = 92, SD = 10.95$	$M = 80, SD = 0$
2	$M = 84, SD = 8.94$	$M = 80, SD = 0$
3	$M = 88, SD = 17.89$	$M = 60, SD = 0$
4	$M = 100, SD = 0$	$M = 100, SD = 0$
5	$M = 100, SD = 0$	$M = 96, SD = 8.94$
6	$M = 88, SD = 10.95$	$M = 84, SD = 8.94$
7	$M = 100, SD = 0$	$M = 84, SD = 8.94$
8	$M = 88, SD = 10.95$	$M = 84, SD = 8.94$
9	$M = 80, SD = 0$	$M = 80, SD = 0$
10	$M = 52, SD = 17.89$	$M = 48, SD = 10.95$
11	$M = 76, SD = 8.94$	$M = 68, SD = 10.95$
12	$M = 68, SD = 22.80$	$M = 72, SD = 10.95$
13	$M = 60, SD = 14.14$	$M = 68, SD = 10.95$
14	$M = 88, SD = 17.89$	$M = 70, SD = 11.55$
15	$M = 80, SD = 0$	$M = 88, SD = 10.95$
16	$M = 100, SD = 0$	$M = 88, SD = 10.95$
17	$M = 76, SD = 8.94$	$M = 80, SD = 0$
18	$M = 72, SD = 10.95$	$M = 72, SD = 10.95$
19	$M = 60, SD = 14.14$	$M = 60, SD = 0$
20	$M = 56, SD = 16.73$	$M = 60, SD = 14.14$

Table C.12: Naturalness scores per participant.