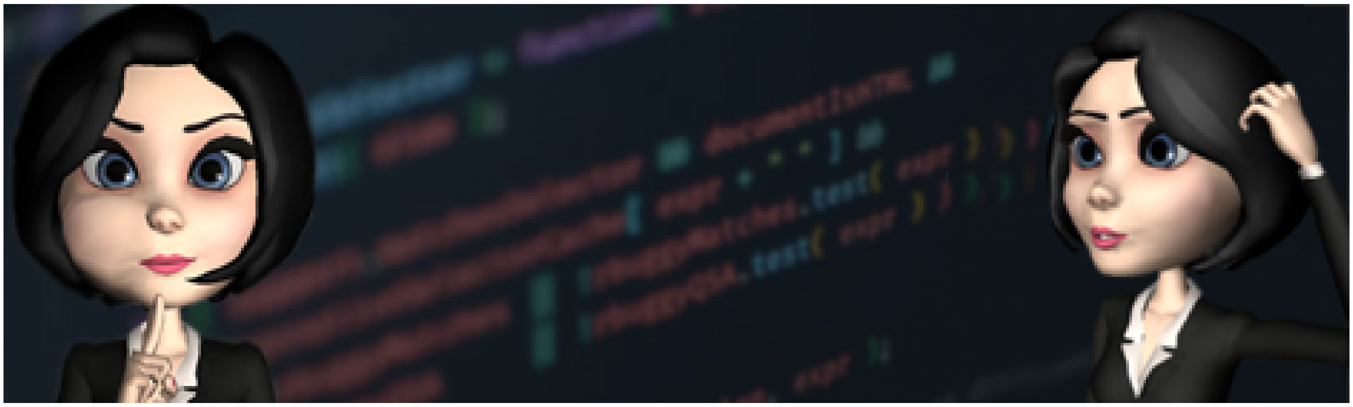
Generating Realistic Sign Language Animations

Inês Lacerda
ines.lacerda@tecnico.ulisboa.pt
Instituto Superior Técnico
Lisbon, Portugal



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 translator.

KEYWORDS

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

1 INTRODUCTION

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[6]. 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[14]. 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.

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** [11] 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.

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 [7] 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. 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.

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. 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.

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.

2 BACKGROUND

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. In Portugal, we have Portuguese Sign Language (LGP). 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.

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 [1, 5], 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.

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.

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 by the manual components, facial expressions play an important role as distinctive phonological parameters for minimal pairs.

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**.

At a **syntactic level**, facial expressions acquire roles similar to 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.

3 STATE OF THE ART

In this section, we explore some techniques that synthesize facial expressions and then, we explore the importance and synthesis of linguistic and secondary movements in Sign Languages.

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 [2–4, 10].

Blend Shape-based approaches [3, 4] 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.

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 [15, 18] have simulated the anatomical structure of the human face, including skin, muscle, soft tissue, and more, to improve the visual realism of synthetic facial expressions.

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**.

Conclusion. 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**.

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.

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.

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 a 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. 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

The core of our system that generates Sign Language animations (Section 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.

4.1 Hand Pose Editor

As shown in the following video, 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.

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 was used to study and analyze our annotated LGP corpus, 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 that 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. There are two types of interrogatives: polar and content questions. Polar questions have a slightly forward upper body and head tilt and content questions have an upward head movement without body tilt. Additionally, both questions have frowned eyebrows, narrowed eyes, and shoulders upward movement. There are two types of negatives: regular and irregular. 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. 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, 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 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.

For our translation system (Section 5) 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

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. 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).

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. We implemented a **a Restful API** where there is a server connected to the Translator process and clients connected to the Unity application. 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.

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.

5.2 Translation Process

The Translation Process was developed by Matilde Gonçalves in a previous thesis[7, 8]. This translation system is divided into two main modules. 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[13] and a syntactic analysis using SpaCy [9]. 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).

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. 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 5.7). Furthermore, we extended the system to recognize verbs with incorporated negation. Our updated system recognizes regular and irregular negatives, and polar and content questions.

We also extended the system to account for composite utterances, to transcribe numerals into corresponding numbers, and identify pauses between clauses and between sentences. The last 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 noticed 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[1], where the espeak backend is used to produce phoneme sequences described based on the International Phonetic Alphabet (IPA) transcription. After, normalization is done by encoding non-ASCII to ASCII, words are separated into their corresponding syllables using syllabification rules and then we map each phoneme into one viseme using the phoneme-viseme mapping we created. 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.

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.

5.4 Database Creation

In Unity, animation files (*.anim*) must be serialized as *AnimationClips* so that these can be loaded and played correctly in runtime.

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

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**. The Resource Folders system should be used for components that are not memory-intensive and do not need to be constantly updated, whereas the Asset Bundle system should be used for files that require continuous content updates.

5.5 Loading components from Database

As described previously, there are two systems capable of storing components: Resource folders and 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.

5.6 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.

5.7 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.

5.7.1 ***Manual Signs.*** In the animator controller, a layer was created for manual signs and dactylology signs. 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.

To animate the avatar, we implemented 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, we 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.

5.7.2 ***Dynamic Transitions.*** 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. 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. These squared magnitude values are then converted to percentages by defining a scale. 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.

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. 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.

5.7.3 ***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.

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 Inverse Kinematics system[2] that

[2] https://docs.unity3d.com/Manual/InverseKinematics.html

allows us to more easily manipulate the avatar's hands. To do so, an invisible ball was added to the scene and using the Inverse Kinematics algorithm the hand can be moved by moving the ball. 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. This video shows the avatar animating numbers and letters.

5.7.4 ***Facial Expressions**.* 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.

Body and facial movements must be carefully combined with manual signs 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.

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.

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 a blend shapes limitation, 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. 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.

5.7.5 ***Mouthing**.* 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 avatar contains 7 visemes: *A*, *B*, *C*, *E*, *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.

The 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. The synchronization between mouthing and signs is extremely important because studies[11] 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 shows the mouthing animation.

5.7.6 ***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. 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. 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 [16] revealed that eye blinks have a prosodic role in marking Intonational Phrase boundaries in four Sign Languages. 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.

In Sign Languages, typically, the head is more active than the torso. The study developed by Tyrone and Mauk[17] for American Sign Language 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[12]. Using Unity's Inverse Kinematics 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. This video shows the avatar with and without secondary head and torso movements. To rotate the spin according to the reaching of the arm, we also used the Inverse Kinematics 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 shows the avatar with and without the torso rotation.

6 EVALUATION

To evaluate our system, we have designed and executed **three experimental user studies**. 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 Section 3.2, it is necessary to separate linguistic and non-linguistic components for evaluation purposes.

All three studies were approved by the Ethics Committee of Instituto Superior Técnico, University of Lisbon. 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 ***Procedure.*** We recruited 10 participants fluent in LGP with the help of the Católica team and the snowballing sampling technique. We conducted within-subject user tests where each participant tested all conditions because we did not want individual differences to affect our results.

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 two-paired sentences where one contained 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. 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.

6.1.2 ***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 ***Procedure.*** We recruited 11 participants 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. For this user study, we only conducted a quantitative evaluation that consisted of questionnaires. 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[3]. 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 5.7.2) 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. Each participant received a different version of the questionnaire, therefore, we created eleven different versions where, in each version, the condition’s order is counterbalanced and the sections’ order is random.

6.2.2 ***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 ***Procedure.*** 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. 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

[3] https://spreadthesign.com/pt.pt

that all signs could execute mouthing. 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 was 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. 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.

6.3.2 ***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.
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.

7 CONCLUSION

In this dissertation, we presented an approach that 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. 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.

Although the results obtained are good, there are some aspects to be improved and extended, especially, in terms of naturalness. Some suggestions included adding more facial expressions, adding corporal movements, adding appropriate pauses and accelerations between signs, and creating more fluid movements.

REFERENCES

[1] Inês Rodrigues Almeida. 2014. 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* 2104 (2014).

[2] J Bento, AP Cláudio, and P Urbano. 2014. Avatares em língua gestual portuguesa. In *Proc. 9th Iberian Conference on Information Systems and Technologies.* 185–191.

[3] Zhigang Deng and Junyong Noh. 2008. Computer facial animation: A survey. In *Data-driven 3D facial animation.* Springer, 1–28.

[4] Nikolaos Ersotelos and Feng Dong. 2008. Building highly realistic facial modeling and animation: a survey. *The Visual Computer* 24, 1 (2008), 13–30.

[5] Paula Escudeiro, Nuno Escudeiro, Rosa Reis, Jorge Lopes, Marcelo Norberto, Ana Bela Baltasar, Maciel Barbosa, and José Bidarra. 2015. Virtual Sign–A Real Time Bidirectional Translator of Portuguese Sign Language. *Procedia Computer Science* 67 (2015), 252–262.

[6] Carrie Lou Garberoglio, Jeffrey Levi Palmer, Stephanie W Cawthon, and Adam Sales. 2019. *Deaf people and employment in the United States: 2019.* Technical Report. 1–20 pages. https://doi.org/10.26153/tsw/10053

[7] Matilde Gonçalves. 2020. *PE2LGP 4.0: de português europeu para língua gestual portuguesa.* Master's thesis.

[8] Matilde Gonçalves, Luisa Coheur, Hugo Nicolau, and Ana Mineiro. 2021. PE2LGP: tradutor de português europeu para língua gestual portuguesa em glosas. *Linguamática* 13, 1 (2021), 3–21.

[9] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear* 7, 1 (2017), 411–420.

[10] Hernisa Kacorri. 2015. *A Survey and Critique of Facial Expression Synthesis in Sign Language Animation.* Technical Report. Department of Computer Science, The City University of New York.

[11] Michael Kipp, Quan Nguyen, Alexis Heloir, and Silke Matthes. 2011. Assessing the deaf user perspective on sign language avatars. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility.* 107–114.

[12] John McDonald, Rosalee Wolfe, Jerry Schnepp, Julie Hochgesang, Diana Gorman Jamrozik, Marie Stumbo, Larwan Berke, Melissa Bialek, and Farah Thomas. 2016. An automated technique for real-time production of lifelike animations of American Sign Language. *Universal Access in the Information Society* 15, 4 (2016), 551–566.

[13] Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *LREC2012.*

[14] Sen Qi and Ross E Mitchell. 2012. Large-scale academic achievement testing of deaf and hard-of-hearing students: Past, present, and future. *Journal of deaf studies and deaf education* 17, 1 (2012), 1–18. https://doi.org/10.1093/deafed/enr028

[15] Eftychios Sifakis, Igor Neverov, and Ronald Fedkiw. 2005. Automatic determination of facial muscle activations from sparse motion capture marker data. In *ACM SIGGRAPH 2005 Papers.* 417–425.

[16] Gladys Tang, Diane Brentari, Carolina González, and Felix Sze. 2010. *Crosslinguistic variation in prosodic cues.* na.

[17] Martha E Tyrone and Claude E Mauk. 2016. The phonetics of head and body movement in the realization of American Sign Language signs. *Phonetica* 73, 2 (2016), 120–140.

[18] Yu Zhang. 2008. Muscle-driven modeling of wrinkles for 3D facial expressions. In *2008 IEEE International Conference on Multimedia and Expo.* IEEE, 957–960.