electronics

MDPI

Article

Phonological Proximity in Costa Rican Sign Language

Luis Naranjo-Zeledón [1,2,†], Mario Chacón-Rivas [1,†], Jesús Peral [2,*,†] and Antonio Ferrández [2,†]

[1] Inclutec, Instituto Tecnológico de Costa Rica, Cartago 30101, Costa Rica; lnaranjo@itcr.ac.cr (L.N.-Z.); machacon@itcr.ac.cr (M.C.-R.)

[2] Department of Software and Computing Systems, University of Alicante, San Vicente del Raspeig, 03690 Alicante, Spain; antonio@dlsi.ua.es

* Correspondence: jperal@dlsi.ua.es

† These authors contributed equally to this work.

Received: 28 June 2020; Accepted: 11 August 2020; Published: 13 August 2020

Abstract: The study of phonological proximity makes it possible to establish a basis for future decision-making in the treatment of sign languages. Knowing how close a set of signs are allows the interested party to decide more easily its study by clustering, as well as the teaching of the language to third parties based on similarities. In addition, it lays the foundation for strengthening disambiguation modules in automatic recognition systems. To the best of our knowledge, this is the first study of its kind for Costa Rican Sign Language (LESCO, for its Spanish acronym), and forms the basis for one of the modules of the already operational system of sign and speech editing called the International Platform for Sign Language Edition (PIELS). A database of 2665 signs, grouped into eight contexts, is used, and a comparison of similarity measures is made, using standard statistical formulas to measure their degree of correlation. This corpus will be especially useful in machine learning approaches. In this work, we have proposed an analysis of different similarity measures between signs in order to find out the phonological proximity between them. After analyzing the results obtained, we can conclude that LESCO is a sign language with high levels of phonological proximity, particularly in the orientation and location components, but they are noticeably lower in the form component. We have also concluded as an outstanding contribution of our research that automatic recognition systems can take as a basis for their first prototypes the contexts or sign domains that map to clusters with lower levels of similarity. As mentioned, the results obtained have multiple applications such as in the teaching area or the Natural Language Processing area for automatic recognition tasks.

Keywords: sign language; phonological proximity; similarity measures; clustering; recognition

1. Introduction

Since 19 July 2012, Law 9049 declared Costa Rican Sign Language (LESCO) as the mother tongue of the deaf community in Costa Rica [1]. Some other laws have strengthened its status, such as Law 20767, in recognition and promotion of LESCO, as a cultural and linguistic heritage of the deaf community [2].

The research carried out in sign languages is presented in a systematic mapping study [3], making it clear that the recognition of these languages still presents many challenges, particularly in real-time software systems.

The fundamental components of the phonology of each sign in LESCO are the form of the hands, location of the hands in space, and orientation of the hands. These phonological components are very standard and have been studied mainly for American Sign Language (ASL) [4,5]. Therefore,

throughout this study, we will refer to phonological proximity taking into account whether these elements are considered in a consolidated way or separately.

Automatic sign recognition tasks require adaptations to infer signs that can only be partially determined. When the system decides which sign it is, it can use phonological proximity techniques to disambiguate this sign, determining if the characteristics it managed to identify make it a candidate to be part of a density neighborhood, in accordance with a context. The concept of context consists of a previous classification that was made on the sign database. This previous classification was not based on phonological or semantic proximity, but on the areas or domains from which the signs were taken. Thus, for example, the context "CoopeAnde" corresponds to a savings and loan cooperative of educators with specific signs of the financial sector, or the context "CDPCD", which corresponds to the Convention on the Rights of Persons with Disabilities, with the proper signs of a bill of rights. Table 1 shows the eight contexts and their meanings.

Table 1. Contexts and their meaning.

Context	Meaning
CDPCD	Convention on the Rights of Persons with Disabilities
CENAREC	National Resource Center for Inclusive Education
CoopeAnde	Cooperative of the National Association of Educators
HimnoNacional	National Anthem of Costa Rica
SENATON	Validation of the PIELS platform with the Costa Rican deaf community, in 2018
SICID	Costa Rican Disability Information System
TEC	Costa Rica Institute of Technology
TSE	Supreme Court of Elections

On the other hand, a clear characterization of the phonological proximity is particularly relevant because it has become evident that the processes that make up phonological awareness, while having different degrees of psycholinguistic complexity, regarding those of being able to identify differences between words require only certain phonological perceptual acuity and normality in language development, to establish similarities and differences between words, which favors the development of language teaching programs [6].

Despite its legal recognition as the mother tongue of the Costa Rican deaf community, currently, the authors have not found evidence of any study on phonological proximity in LESCO, so this is the first formal proposal of its kind.

By determining the phonological proximity, there is a basis of applicable knowledge in different areas that can receive important inputs to improve their processes. For example, in the area of sign language teaching, the lesson planning mechanisms can be strengthened, emphasizing the similarities and differences between sign groups, which can be subtle and yet lead to important semantic differences.

Another application is in the field of automatic sign recognition. As in many other natural language processing (NLP) tasks, these systems usually require a disambiguation module, to produce more accurate results, which can also be favored by having neighborhoods of similar signs, from which the system can choose the one that best suits the context in question, and hence solve a sign that has been only partially recognized.

The production of sign language thesauri can also benefit from discoveries in phonological proximity, facilitating the work of curators by suggesting clusters that can be evaluated by humans and, in this way, organizing the thesauri in order of transition from one sign to another (in a non-alphabetical order).

Phonological proximity clusters can also be used in the future to compare them with semantic proximity clusters. It may seem obvious to connoisseurs of sign language that many signs that are similar to each other have very different meanings and that, conversely, many signs with close meanings can take very different forms. The generation of phonological proximity clusters is an

important input to compare in the future with semantic similarity clusters and thus determine if there exist high correlation domains that can be exploited for didactic purposes.

We have clearly identified the research questions underlying this work, namely:

RQ1. What methods can be used to measure the similarity between signs?
RQ2. What relevant information do these methods provide once applied?

In order to answer the research questions, several experiments were carried out obtaining the following main contributions of this paper:

- Establishment of a scheme to study the phonological proximity in LESCO, nonexistent until now, that can be useful for decision-making in educational programs design, linguistic analysis, or development of automatic recognition software
- A method of generating quantitative measures for decision-making for recognition systems.
- Based on well-known linguistic theories for spoken languages, we have made an adaptation for sign languages, which can be a very promising line of research for the future.
- We have proved the suitability of classical similarity measures, as well as dimensionality reduction and clustering algorithms, for the characterization of the phonological proximity in LESCO.
- We showed that LESCO has a high degree of phonological proximity, particularly in the orientation and location components, but considerably lower in form.
- We have analyzed the impact of context in the resulting phonological proximity.

Our case study is LESCO, but what is stated in this paper would also be applicable to other languages, provided that the methodology explained in our approach and methods be followed by those interested in conducting similar research.

The remainder of the paper structure is as follows: Section 2 covers the literature background on similarity measures, cognitive and psycholinguistic foundations, and phonological proximity, to provide a strong conceptual basis to our study. Section 3 explains the approach and methods in detail. Section 4 presents the experimentation and discussion. Finally, Section 5 presents the conclusions and Section 6 provides some insights on future work.

2. Background

In this section, we make a classification of the related literature by subject, namely: similarity measures, cognitive and psycholinguistic foundations, and phonological proximity. It has been divided into these three subsections for clarity: (1) It has been taken into account that the similarity measures are central when designing the experiments. (2) On the other hand, cognitive and psycholinguistic foundations are important to have a theoretical foundation on what is meant by similarity and how it applies to our object of study. (3) Finally, the topic of phonological proximity is addressed to determine what has been done in this field, which precisely gives its name to this paper, and which demonstrates the valuable contribution that has been made for languages in general in contrast to the scarce specific contribution for sign languages.

2.1. Similarity Measures

According to [7,8], similarity measures can be categorized into five dimensions: Edit-based, Token-based, Hybrid, Structural (Domain-dependent), and Phonetic. There is a wide variety of measures of similarity within this classification, as shown in Figure 1 (adapted from Nauman et al. [7] and from Bisandu et al. [8]). Table 2 shows a comparison of the similitude categories and their general approach, as well as their best known proposals and their methods.

To measure the phonological proximity in a sign language, one can resort to some of the similarity measures that we have explained, taking into account the characteristics of the data used as phonological parameters. In this way, if the parameters have a string data type, the edit-based

measurements can adequately satisfy the measurement objective. If the parameters are token sets (as in the case of this work), token-based measures can be used. Hybrid approaches seek to balance the speed of response of known measures with the robustness of a complete comparison between all tokens, to find the best matches, mainly to deal with named entities or solve problems of misspelled words in big data contexts, which does not outline them as the best options for sign languages. Phonetic measures are used for spoken languages, so they are ruled out for sign languages. Finally, domain-dependent measures make use of very particular characteristics of the data, which are hardly going to form the primary basis of a corpus of sign languages.

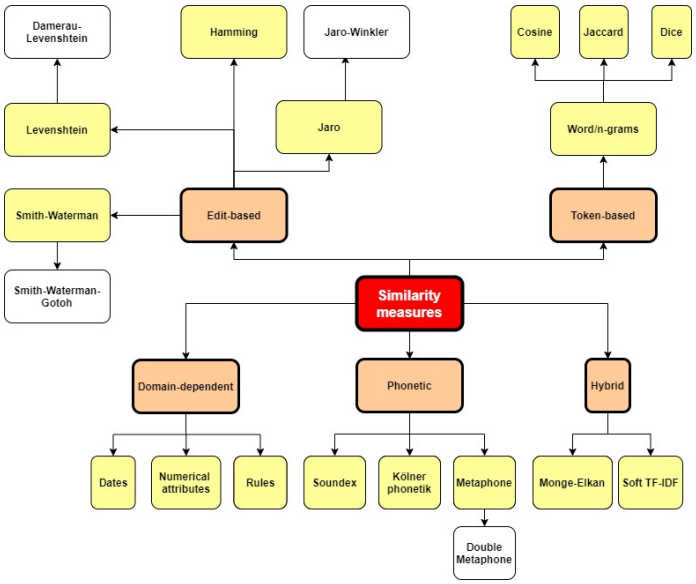

Figure 1. Classification of similarity measures, based on Naumann et al. [7] and Bisandu et al. [8].

Table 2. Similarity measures comparison.

Basis	**General Approach**
Edit-based	Calculation of the changes necessary to produce one string from another, weighing the amount of changes necessary (insertions, deletions or modifications) to produce the new string
Best known proposals	**Method**
Hamming distance	Compares bit strings (0 s and 1 s).
Levenshtein distance	Compares alphanumeric strings.
Token-based	Measure the number of matches between two sets of parameters, (n-gram tokens), where tokens are words or numbers
Best known proposals	**Method**
Jaccard distance	Weighs unsorted sets.
Cosine distance	Respects order of the parameters, preferred when order is relevant.
Hybrid	Compare strings, using an internal similarity function (Jaro or Levenshtein, for instance)
Best known proposals	**Method**
Monge–Elkan	Measures the best scores of analyzed tokens, then adds these scores and averages them to obtain the final result.
Soft TF-IDF	Combines cosine with TF-IDF (term frequency-inverse document frequency) weighted vectors, and Jaro–Winkler. Works best for named entities [9].
Structural (Domain-dependent)	Focus on data particularities
Best known proposals	**Method**
Dates	Relying highly on rules, for instance, if there are dates that differ only by one month and one character in the year (from month 12 of the year to month 1 of the following year), the algorithm assigns a high similarity to both dates, following the rule of the new year.
Phonetic	Mapping similar sounds in spoken languages. In English, for example, they give maximum qualification to pairs of words such as ("feelings", "fillings"), applying pre-established rules of similar sounds
Best known proposals	**Method**
Soundex	Applying rules to similar sounds in English.
Kölner–Phonetik	Applying rules to similar sounds in German.

2.2. Cognitive and Psycholinguistic Foundations

Since our study concentrates on the parametric similarity, that is, the components that characterize each sign, thorough discussions of a philosophical or psychological nature about what is considered similar are not part of the scope of this research. For further reference in this regard, the interested reader can resort to classical literature on the subject [10–12]. We will, however, cover some salient matters about these subjects.

Briefly, we find evidence in [10] that people assume what the relevant context factors are. The authors give the example of a comparison between the USA and Russia, indicating that they generally assume that the relevant context is the set of countries and that the relevant frame of reference includes political, geographical, and cultural characteristics. The relative weights assigned to these characteristics, however, may differ for different people. With stimuli such as countries, people, colors, and sounds, there is relatively little ambiguity regarding the contextual characteristics. However, with artificial, separable stimuli, such as figures with different colors and shapes or lines with different lengths and orientations, subjects sometimes experience difficulties in assessing similarity, and occasionally tend to assess similarity with respect to one factor only, or otherwise change the relative weights of attributes with a contextual change.

2.3. Phonological Proximity

The literature on phonological proximity in sign languages is not prolific. Although phonetic and phonological proximity has been a topic of exploration for linguists for decades for spoken languages, signed languages have seen much less effort [13]. In particular, there is a lack of such studies for LESCO. This sign language can offer many examples of proximity, just by looking at the database used in this study and comparing 26 phonological parameters (thirteen for each hand)—for example, the words "autonomy" and "Braille" (sixteen equal parameters), "library" and "each" (twenty-one equal parameters), or "year" and "change" (twenty-three equal parameters).

After a thorough inspection in academic repositories, we have determined three specific references: Hildebrandt et al. [14], Williams et al. [15], and Keane et al. [16]. In an attempt to broaden the critical apparatus of this work, we have resorted to apply the forward snowballing and backward snowballing techniques [17] to the aforementioned three references to our object of study, to some extent. This is because the base studies of these references, as well as the new studies that cite them, cover topics that are beyond the scope of our research. However, the use of snowballing techniques proved useful to accomplish our task of filling a literature gap detected at first try in conventional search engines. In Table 3, the findings of our literature review are schematized, tabulated in different categories.

> Subjects categorize signs that share transition from an initial form to a final form and location as highly similar, regardless of handshape. Nevertheless, varying degrees of previous linguistic knowledge of deaf signers influenced their perception of similarity [14].

From the ideas previously exposed, it can be concluded that the different works that have been carried out present the study of the phenomenon mainly for ASL, as well as a confirmation that, in sign identifying, the components of the initial form usually have a greater weight than the transition movement or the final shape. After carrying out this review of the literature related to the topic, it is evident that our research is the first of its kind for LESCO, and the results of the writing of this paper derive directly from the stages of data gathering, data quality assessment, and an ulterior analysis, thus providing a meaningful first contribution to our object of study.

Table 3. Phonological proximity literature review.

Category	Findings
Sign categorization	Subjects categorize signs that share transition from an initial form to a final form and location as highly similar, regardless of handshape. Nevertheless, varying degrees of previous linguistic knowledge of deaf signers influenced their perception of similarity [14].
Sign language lexical access	In [15], research found out that the way which individuals access words in the mental lexicon [18] are vastly unknown. Phonological proximity facilitates lexical retrieval in sign languages. Lexical access in sign language is facilitated by phonologically similar lexical representations in memory.
Fingerspelling proximity	In [16], the researchers concentrate their efforts in two methods for calculating phonetic form similarity for fingerspelled words only. Experimentation was validated with psycholinguistic evidence based on similarity ratings by deaf signers. The positional similarity method best predicts native signer intuition judgments about form similarity.
Analytical model of proximity	Alkoby [19] proposed an analytical model of similarity. The nature of experimentation is very costly, whenever it involves human subjects and many comparisons. The proposal seeks to take a step forward with respect to the experimentation of Richards and Hanson [20], by means of an analytical model, instead of costly perceptual tests.
Most relevant component	The studies of Brentari [21] and Valli and Lucas [22] have determined that the form is the most apparent and complex parameter of a sign. Hence, this is the first one that people think of when preparing to sign.
Phonology in sign languages	Phonological systems not only pertain to spoken languages; in fact, every sign language exhibits clear phonological patterns [23–26], and signers acquire their native phonology spontaneously [27].
Speech production	Vitevitch [28] determined the influence of phonological proximity neighborhoods on the speed and accuracy of speech production for spoken language. The results indicated that more errors were elicited for words with few similar sounding words. This shows that multiple word forms are activated at once, improving speech production.
Prosodic models	The rhythmic and intonational aspect of language (prosody) has been studied by Brentari and Padden [29], by representing the form through a branching feature system. It consists of indications on which fingers are selected, as well as the flexion–extension of the base joints and the non-base joints. Their main focus of attention is on ASL.
Fingerspelling as lexicon	Hendriks and Dufoe [30] discuss an extension of [29] lexicalized fingerspelling and initialization for Mexican Sign Language (LSM). They analyze phonological restructuring processes that take place when fingerspelled words become a part of the lexicon, arguing that the lexicalized signs are part of the foreign/non-native lexicon of LSM.
Early learners and late learners	Keck and Wolgemuth [31] point out that non-native or ASL Late-Learners attempt to use oral phonological similarities to choose ASL words that are semantically similar although not phonologically related, while ASL Early-Learners were able to focus much more on ASL phonological similarities to select correct words. Early learners distinguish phonological similarities to a larger extent than late learners.
Parameters for codification	Caselli et al. [32] report on a very important finding regarding the choosing of parameters for sign codification. The calculation of the neighborhood density may be incomplete, which means that, though neighbors overlap in the properties coded in their database, they may or may not differ phonological properties not coded yet. Incomplete codification may lead to error.

Table 3. *Cont.*

Neuroscience approach	In a study from the neuroscience perspective, Meade et al. [33] demonstrate that phonological overlap of two parameters facilitates processing, using a measured brain response that is the result of a sensory, cognitive, or motor event.
Perceptual differences	According to Williams et al. [34], a bifurcation of neighborhood density based on sublexical features reveals perceptual differences for the form and location. They mention Carreiras et al. [35], and their findings in lexical access of Spanish Sign Language and its density as perceived by the deaf. They found that signs in dense location neighborhoods are harder to identify than those in sparse location neighborhoods.
Production versus learning	Gahl et al. [36] demonstrate in their study for spoken languages that words that are part of dense phonological neighborhoods tend to be difficult to recognize, but easy to produce. In other words, a high phonological density can pose greater challenges both for recognizing and learning the language, but a greater ease of creating new words. This dichotomy between ease of production and difficulty of recognition and learning is a fact of great importance, and will be addressed later in this paper.
Hierarchy of parameters	Steinbach et al. [37] found a hierarchy of parameters and their impact on processing. The location parameter is prioritized in sign recognition, followed by form and movement. However, the effects of combined parameters give greater insight on the subject: the overlap in location and movement leads to higher similarity rates.
Environment and iconicity	Cardin et al. [38] report on the findings of Thompson and et al. [39], pointing out that for adult learners of a second language the environment plays a role in the effectiveness of iconicity in sign learning. Learning signs with greater location proximity is more difficult.
Native bimodal bilinguals	Villameriel et al. [40] determined that, for sign recognition, native bimodal bilinguals (native speakers both of spoken and signed languages) showed an earlier competition from location than form, whereas they achieved stronger competition in form than location.

3. Approach and Methods

We have used a combined approach of a widely used similarity measure, such as cosine, along with Principal Component Analysis (PCA) techniques [41] in preparation to run the k-means clustering algorithm [42].

In our experiments, we used Python deployed on a Jupyter Notebook to produce clusters for both hands, using refined data. The resultant display, after reducing dimensionality via de PCA algorithm, is a three-dimensional graphic, for ease of interpretation.

We also used Python code deployed on a Jupyter notebook to run the Word2Vec algorithm, in order to map a text to a n-dimensional space and then display it in two dimensions. The Word2Vec algorithm allows for a way to visually determine clusters, receiving a corpus of texts pre-categorized by contexts. Li [43] explains in detail how to implement this algorithm using Gensim in Python and her suggestions were the basis for our implementation.

The contexts that Word2Vec use are categorizations of the texts that serve as input. For example, a text may contain the contexts tech, business, sport, politics, and entertainment, along with relative full paragraphs. Thus, for instance, an excerpt of an input file could contain the following entries, as taken from [43] (notice that the lack of punctuation marks comes from the source, as a means to make processing easier):

> ***sport***, *moya fights back for indian title carlos moya became the first man to successfully defend the chennai open title by beating four-times finalist paradorn srichaphan 3-6 6-4 7-6 (7/5). the spaniard then donated his £28,000 prize money to relief efforts for the victims of the asian tsunami . . .*

***entertainment**, career honor for actor dicaprio actor leonardo dicaprio s exceptional career has been honored at the santa barbara international film festival. the star was presented with the award by martin scorsese who directed him in oscar-nominated movie the aviator . . .*

***tech**, mobile gig aims to rock 3 g forget about going to a crowded bar to enjoy a gig by the latest darlings of the music press. now you could also be at a live gig on your mobile via the latest third generation (3g) video phones . . .*

Our fundamental input has been the International Platform for Sign Language Edition (PIELS) database [44] to have a stable and official database of signs for research purposes. The data had to be curated in order to eliminate duplicates and test data, as well as to eliminate the use of two parameters that currently store a constant, and were considered superfluous for our study since they do not add any significant difference to the phonology of the right hand nor left. A decision also had to be made to choose the eight relevant domains (contexts) to be used in this research and thereafter.

Several pre-defined contexts already existed, some of which were outdated or did not entail a benefit to the PIELS platform. Because of this, we decided to choose the definitive official contexts in Table 1.

The phonological parameters for the hands are divided into three equal groups for each hand: the form group with six, the orientation group with three, and the location group with four, totaling thirteen. The parameters are the same for each hand, so the grand total is twenty-six. The parameter list for each hand is as follows:

- Form:
    1. Index
    2. Medium
    3. Annular
    4. Little finger
    5. Separation
    6. Thumb
- Orientation:
    1. Rotation
    2. Wrist posture
    3. Intentionality
- Location:
    1. Space laterality
    2. Space height
    3. Depth
    4. Arm contact

A spreadsheet was used to calculate the similarity measures, by using the conventional formulas already included therein. The dataset was included in this spreadsheet for tabulation and data analysis purposes. The dataset consists of twenty-nine columns. The first column contains alphanumeric characters with one of the eight contexts described above. The second column contains alphanumeric characters with the name of the sign, that is, the concept it represents. The third column contains the twenty-six phonological numerical parameters of the signal, separated by semicolons. The next 26 columns consist of these phonological parameters in a disaggregated way. Therefore, columns 4 through 15 contain the numerical form parameters, 16 through 21 have the numerical orientation parameters, and 22 through 29 contain the numerical location parameters. Table 4 depicts how many

signs belong to each context, with very different sign counts, ranging from 46 to 1000, that is, the nature of these contexts varies considerably.

Table 4. Sign count by context.

Context	Sign Count
CDPCD	436
CENAREC	1000
Coope Ande	207
Himno Nacional	46
SEÑATON 2018	259
SICID	300
TEC	197
TSE	120
Grand total	**2565**

Typical entries in the dataset look like this:

TSE,For,1;2;2;2;1;5;2;3;2;4;9;2;2;1;1;1;1;1;1;1;3;3;2;4;22;1;2;1;1,1,2,2,2,1,5,1,1,1,1,1,1,2, 3,2,3,3,2,4,9,2,2,4,22,1,2

TEC,South,2;2;2;2;1;5;4;3;2;4;15;1;2;1;1;1;1;1;1;1;3;3;2;4;22;1;2;1;1,2,2,2,2,1,5,1,1,1,1,1,1,4, 3,2,3,3,2,4,15,1,2,4,22,1,2

SICID,CR,4;4;4;4;1;4;3;3;2;4;15;1;2;1;1;1;1;1;1;1;3;3;2;4;22;1;2;1;1,4,4,4,4,1,4,1,1,1,1,1,1,3,3, 2,3,3,2,4,15,1,2,4,22,1,2

We used a Jupyter Notebook to implement the k-means clustering algorithm, as well as Word2Vec to map words to a vector space through a neural network and thus determine Word embeddings. Gensim was the choice to implement Word2Vec and t-SNE was used for visualization purposes.

Finally, we proceeded to carry out a detailed analysis of the results, by interpreting the graphic displays produced, as well as the raw data groupings, and we compared the contribution of the standard measures, the clustering algorithm, and the previous embeddings by contexts of relevance.

We suggest as a general methodology the following steps, which apply to other sign languages, as a useful instrument for researchers and linguists:

- it is necessary to populate an initial database, with the signs
- it is necessary to decide which contexts each of these signs belongs to, considering the possibility that they belong to several contexts must be refined, as some may be evidence or lose relevance over time, or subsume within others that are more relevant
- the phonological parameters must be determined, which are naturally subdivided in location, orientation and shape
- a mapping about the possible parameter settings to numerical values must be made, in order to execute the algorithms
- this mapping of parameters must be properly documented, for future reference
- a measure of similarity should be used to produce homogeneous ranges of signs, and thus determine their accumulation and dispersion
- a clustering algorithm must be used to determine the gain of carrying out the study across contexts or taking the contexts as a whole
- the data obtained through these similarity measurement strategies should be analyzed, mainly with the support of graphic displays
- as a result of this analysis, it must be concluded if the language studied has high, medium, or low levels of phonological proximity

- possible courses of action based on this proximity should be suggested, for example, prioritizing the development of recognition systems when the similarity is high or low, or developing these systems basically in parallel when the similarity is highly distributed.

4. Experimentation and Discussion

The experimentation has been divided into two groups: (1) embeddings using the cosine similarity formula and (2) generation of clusters using the k-means algorithm. Experiments have been designed in this way to analyze the level of homogeneity of the LESCO language in terms of phonologically close signs. The benefits that this entails consist in determining if the homogeneity in these clusters is high and therefore there is a balance between the ease of producing new signs contrasted with learning or automatically recognizing the signs. Additionally, we will be able to determine whether the clustering produced by k-means is worthwhile to be run across contexts or taking the whole database as its input.

4.1. Embeddings Using Cosine Similarity

After calculating the similarity measures among the signs and applying the clustering algorithm, a significant amount of data has been obtained, allowing for an analysis of the phonological proximity in LESCO.

From our database, we have a mapping of each sign to a vector of 26 numerical parameters, each with a precise meaning (13 parameters for each hand). These parameters follow this order: (left index, left middle, left ring finger, left finger, left finger separation, left thumb, right index, right middle, right ring, right finger, right finger separation, right thumb, left rotation, left wrist posture, left internality, rotation right, right wrist posture, right internality, left laterality, left height, left depth, left arm contact, right laterality, right height, right depth, right arm contact). Thus, for example, the array [1,3,2,2,1,3,**1,3,2,2,1,3**,3,3,2,**3,3,2**,4,15,3,2,**4,15,3,2**] contains the 26 parameters that phonologically describe the sign for "protection", and the sub-array [1,3,2,2,1,3,3,3,2,4,15,3,2] represents the parameters for the left hand.

In accordance with the comparison of the similarity measures previously exposed, for this research, the cosine was selected (a token-based approach) because it is the most natural for a numerical array of parameters, as is the case of this work. If we were to use an edit-based similarity measure, arrays would have to be converted to strings beforehand, with special care that some parameters have only one digit and others two. Clearly, domain-dependent measures are meaningless to our data, and hybrid approaches present unnecessary complications. Finally, it should be clarified that the phonetic similarity measures are designed specifically for spoken languages, with the similarities that these present at the auditory level.

In order to identify the degree of similarity within the sign repository, we used the strategy of comparing each sign against a neutral position of the signer, that is, one in which the hands are not signaling. The cosine similarity measure also had another advantage in its wide diffusion and its consistency with the principle that governs other clustering strategies, such as Word2Vec, discussed later.

For each embedding (form, orientation, location and consolidated), the grouping was parameterized in the spreadsheet, in such a way that each group contains around 52 signs, the approximate square root of 2665 signs in our database. This was done in this way to determine if the signs were grouped by degree of similarity in a homogeneous way. In an ideal scenario, each of the 52 groups would contain around 52 signs, which would be an indication that there is this homogeneity in the data and, consequently, there would be no groups with too many or too few signs. In this way, it would be in line with what was stated by Gahl et al. [36], hence achieving a balance between ease of production versus recognition and learning.

In the case of LESCO, with this study, we have managed to determine that this homogeneity does not exist, neither for the individual components (form, location, orientation) nor for the consolidated

one, as we will explain in detail later. This finding makes a fundamental contribution, since it allows making decisions using this high degree of dispersion, both for didactic purposes and for the development of automatic sign recognition systems. For example, a decision can be made to first develop a recognition system in a domain whose signs do not generally have a high degree of similarity, mapping a reference corpus to ranges of similarity. After building this first functional system, it is possible to progress towards domains with more similar signs and, therefore, difficult to recognize, and with a greater propensity for error.

Our first experiment consisted of determining if the cosine formula allowed to group the signs in packages that, on average, had a size of approximately 52 signs for later decisions on automatic recognition or even linguistic analysis and didactic purposes.

After this, we proceeded to test this same formula for the separate components of form, location and orientation, trying to determine if the dispersion of the data was lower, to facilitate decision-making on the use of separate clusters in these components. However, we determined that neither the separate components nor the consolidated one is close to a normal distribution, which imposes a limit on us in using the standard deviation. Both for the form and location components, an average of 49.33 signs per group was reached, while for orientation and the consolidated the average was 50.29. All of this means that there is no reliable measure of dispersion to establish a comparison because their distribution does not closely follow a normal distribution.

Table 5 shows three examples of sign comparisons for high, medium, and low consolidated similarities, as detected in our experimentation. This comparison shows examples of signs with high, medium, and low similarities, in order to illustrate the concept of similarity within phonological parameter arrays. The signs "authority" and "self-esteem" show high similarity, with just four different parameters. With medium similarity, the signs "mature" and "serve" appear with ten different parameters, and finally, with low similarity, the signs "limited" and "benefit" have fourteen different parameters.

Table 5. Sign comparison examples for high, medium, and low similarities. Differences appear in bold.

High similarity (99,7644)		
Context	Sign	Parameters
CDPCD	Authority	2;2;2;2;1;2;1;1;1;1;1;1;**3**;**3**;2;3;3;2;**1**;16;**3**;2;4;22;1;2
CDPCD	Self esteem	2;2;2;2;1;2;1;1;1;1;1;2;2;2;3;3;2;2;16;2;2;4;22;1;2
Medium similarity (73,1547)		
Context	Sign	Parameters
CRPD	To mature	**3;3;3;2**;1;**5**;1;1;1;1;1;1;**1;2**;3;2;3;3;2;**2;7**;1;2;4;**22**;1;2
CRPD	To attend	**1;1;1;1**;1;**2**;1;1;1;1;1;1;**2;3**;3;2;3;3;2;**4;4**;1;2;4;**4**;1;2
Low similarity (42,5490)		
Context	Sign	Parameters
CRPD	Limited	**3**;3;**3;3**;1;**1;3;3;3;3**;1;1;**3**;3;2;3;3;2;**4;6**;1;**3**;4;**6**;1;**3**
CRPD	To benefit	**4**;3;**1;1**;1;**7;1;1;1;1**;1;1;**5**;3;2;3;3;2;**2;16**;1;**2**;4;**22**;1;**2**

Figure 2 shows the embeddings using the cosine formula, for the sign form. The data are reasonably close to a normal distribution, according to the Kolmogorov–Smirnov formula [45], which in this case is 0.22509 and the *p*-value is 0.00849. This allows for computing only the average (49.33) and the variance (3202.11). It is notorious that the values are quite scattered across the ranges of similarity. This dispersion of the phonological parameters of the shape component indicates that its contribution is greater for learning effects than for the production of new signs.

Figure 3 shows the embeddings using the cosine formula, for the sign location. The data are not reasonably close to a normal distribution, according to the Kolmogorov–Smirnov formula, which in this case is 0.3811 and the *p*-value is less than 0.00001. This allows for computing only the average (49.33) and the variance (13,865.2). Such a distribution, far from a normal one, can occur because there

are extremely large values (about 80% of the data) accumulated in very high similarity ranges, ranging from 95% to 100%. This high level of skewness implies that new sign production will usually recur to little variation in location, basically what is known as the signing vector, which can be thought of as a cube coming out of the front of the chest with approximately half meter in height, width, and depth.

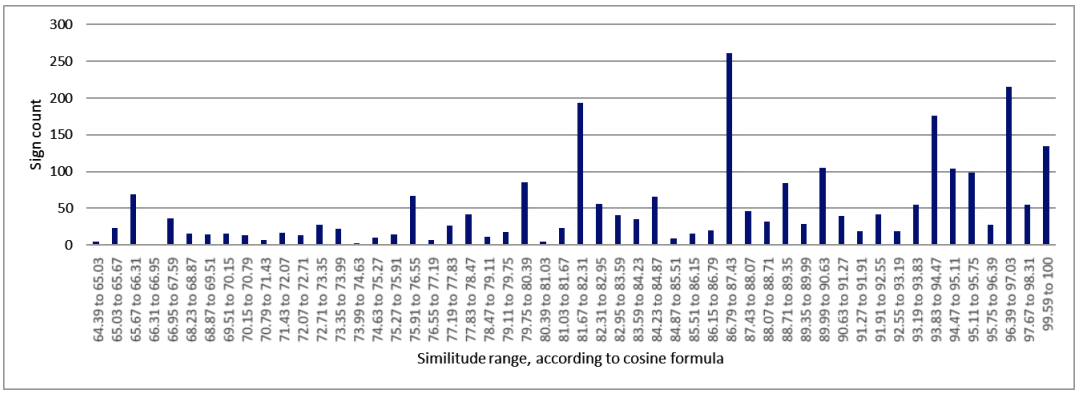

Figure 2. Form embeddings, according to cosine formula.

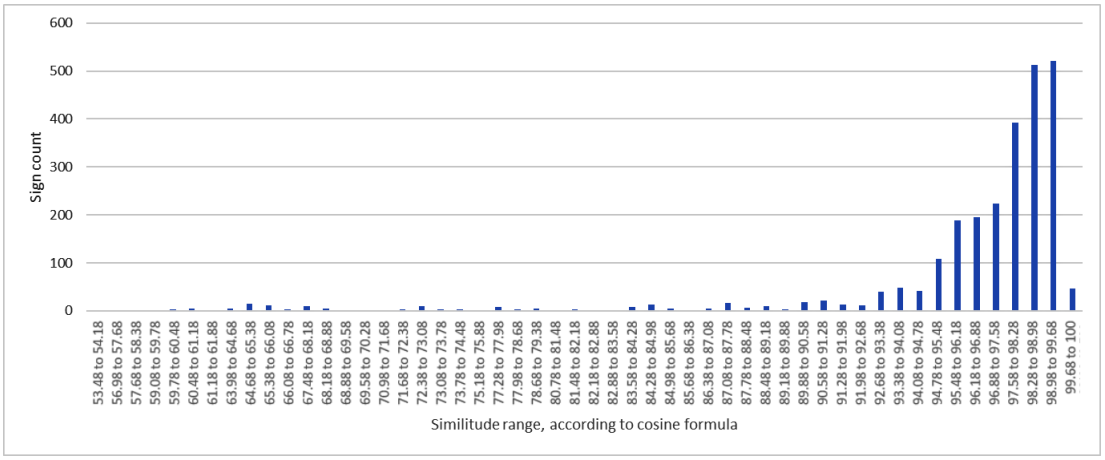

Figure 3. Location embeddings, according to cosine formula.

Figure 4 shows the embeddings using the cosine formula, for the sign orientation. The data are not reasonably close to a normal distribution, according to the Kolmogorov–Smirnov formula, which in this case is 50.29 and the *p*-value is less than 0.00001. This allows for computing only the average (50.29) and the variance (11,419.77). It is evident that the values are quite accumulated in the highest similarity ranges. Again, here is an indication that the orientation component contributes more to generating new signs than to learning already existing ones.

Figure 5 shows the embeddings using the cosine formula, for the consolidated sign. The data are not reasonably close to a normal distribution, according to the Kolmogorov–Smirnov formula, which in this case is 0.32419 and the *p*-value is 0.00003. This allows for computing only the average (50.25) and the variance (10,303.03). It is also evident that the values are quite accumulated in the highest similarity ranges. Therefore, analogously to the orientation and location components, the consolidation of phonological parameters goes in the direction of facilitating the production of new signs, rather than in the direction of facilitating their learning.

Although it is not the main focus of this work, we found it interesting to try the Word2Vec algorithm, in order to determine possible parallels between phonological and semantic proximity. We ran two experiments, where the dimension chosen for the model was 100, after testing the standard measurements that are usually suggested for this algorithm (100, 200, 300). Semantic clustering with more obvious results used 100 (the size or dimensionality of the feature vectors).

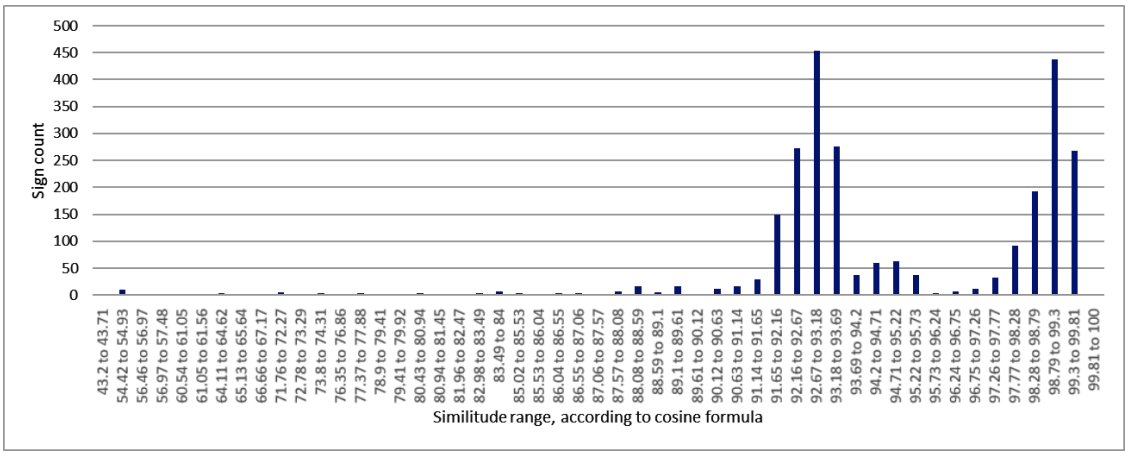

Figure 4. Orientation embeddings, according to cosine formula.

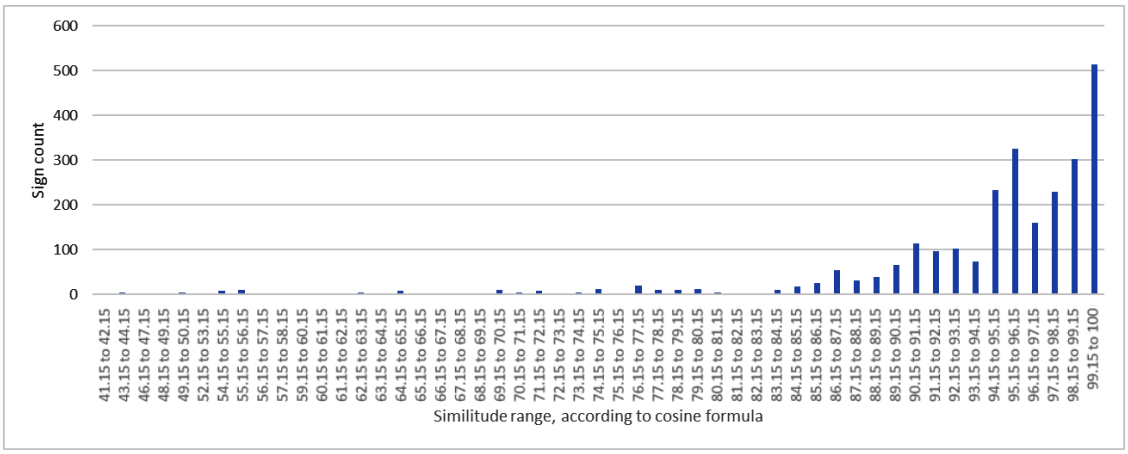

Figure 5. Consolidated embeddings, according to cosine formula.

Figure 6 shows semantic relationships using Word2Vec algorithm in a coherent text, namely the Convention on the Rights of Persons with Disabilities [46]. The differences results are highly noticeable, since by using the text of the CRPD it was possible to easily identify three clusters.

It is also interesting to note the divergence between the semantic closeness detected by the algorithm and the phonemic proximity of the signs. For instance, although the concepts "disability" and "include" appear very near in Figure 6, they have very disparate parameter arrays, hence they reside in different similarity clusters. Their corresponding parameter arrays are [1;1;1;1;1;5;2;3;3;2;15;1;2;1;1;1;1;1;1;1;5;3;3;1;2;15;2;2;1;1] and [1;1;1;1;8;2;5;3;2;4;15;1;2;1;1;4;4;4;4;1;7;2;3;2;4;15;1;2;1;1], and their similarity scores are 93.35 and 97.92. The same applies to "discrimination" and "equal", with very different parameter arrays and residing in different similarity clusters. Their parameter arrays are [1;4;4;4;1;2;4;3;2;4;15;1;2;1;1;1;1;1;1;1;2;3;3;2;4;16;3;2;1;1] and [3;3;3;3;1;1;3;3;2;4;18;3;2;1;1;3;3;3;3;1;1;3;3;2;4;18;3;2;1;1], while they reside in clusters with similarities of 98.19 and 99.81. At a simple glance, Figure 6 shows at least three well-defined clusters with an evident semantic closeness between their concepts, although there is no phonological proximity, as explained above. The signs pairs corresponding to ("disability", "include") and ("discrimination", "equal") are depicted in Figure 7. It becomes quite evident that the data in the parameter arrays correspond to phonologically dissimilar pairs of signs, albeit they portray semantic proximity.

First, a small cluster appears with the concepts "free" and "freedom". Second, there is the cluster made up of the concepts "society", "community", "discrimination", and "equal". The third cluster is made up of the concepts "social", "education", "respect", "disability", "accessible", and "include".

Other small clusters could also be identified as “information” and “communication”, or “national” and “international”.

The worth of these experiments is to confirm that several categorized texts are required to feed the Word2Vec neural network and thus produce more clusters that are more clearly identifiable. Additionally, it was possible to determine that the semantic proximity of words in English does not necessarily correspond to important phonological similarities in LESCO. The expansion and exact measurement of this divergence is considered future work, as explained later.

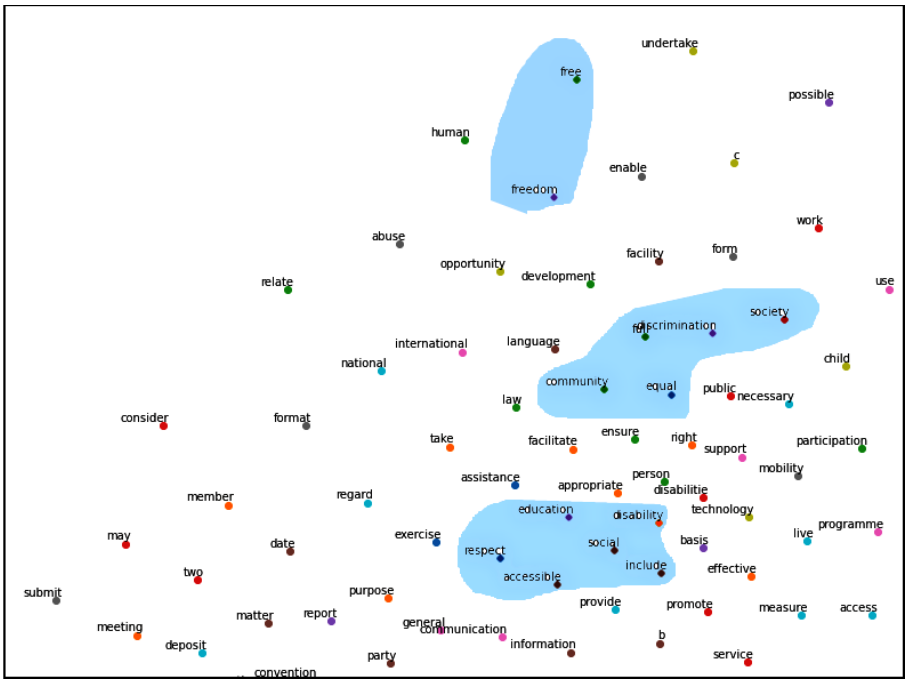

Figure 6. Semantic relationships through Word2Vec in a coherent text: Convention on the Rights of Persons with Disabilities (CRPD).

(**a**) Discrimination

(**b**) Equal

(**c**) Disability

(**d**) Include

Figure 7. LESCO signs for pairs (“discrimination”, “equal”) and (“disability”,“include”).

4.2. Clustering Generated by k-Means

In this section, we will explain the experimentation done using the k-means clustering algorithm. At first, we used the database with each separate sign with its parameters for the left hand and for the right hand. This produces unnecessary clumping and misrepresents the true underlying clustering, as discussed below. In a second experiment, the parameters of both hands were consolidated for each sign and the algorithm was run again, which now reflected a more realistic scenario. Lastly, we decided to eliminate those signs that had duplicates across the different contexts, in order to have the most refined version, and that reflected as closely as possible on how signs are grouped.

Having experimented with classical measures of similarity, we decided to use k-means to detect clusters through another theoretical perspective, namely, the grouping of observations with respect to the closest mean value. This also had a practical reason that was to determine if there was any correlation between the number of contexts identified in our reference database and the number "k" chosen for the algorithm.

To optimize the value of "k" used by k-means, techniques such as elbow or silhouette are usually used. In our case, we resorted to the elbow method since it was necessary to determine if the value of "k" could be at least approximate to the eight contexts previously identified; this in order to check if the clusters generated and the number of pre-established contexts were at least approximate numbers. Therefore, the value of "k" was progressively increased, until determining that a value of 10 reflected the optimum choice to run the algorithm, while it turned out to be very close to the number of contexts.

Figure 8 shows the cluster produced using the k-means algorithm, with k = 10, for each hand and each sign, that is, the left and right hemispheres of each sign. Figure 9 shows the cluster produced using the k-means algorithm, with k = 10, for both hands in each sign, that is, the left and right hemispheres considered together. Finally, Figure 10 shows the more refined version of the data, that is, a cluster produced using the k-means algorithm, with k = 10, for both hands in each sign (the left and right hemispheres considered together), but excluding repeated signs (some signs can mean the same thing and yet appear in more than one context). From Figure 8 to Figure 9, significant changes are noted in the clusters; for example, cluster 1 becomes more dispersed and cluster 5 becomes much more agglutinated. Similarly, from Figure 9 to Figure 10, a greater dispersion of cluster 9 and a greater agglutination of cluster 1 can be seen (even greater than in Figure 8). It is evident that the quality of the design of the experiments and the quality of the input data definitively influence the clustering algorithm.

We found out that an in-cluster standard deviation is 45% smaller when studying both hands signaling at the same time. Refined data without duplicates are even 4% smaller. This means that a good choice of data produces clusters that are more evenly distributed upon parameterization. The cleanest version of clustering is achieved, therefore, by using the parameters of both hands and eliminating in advance the signs that are repeated across different contexts.

It is of great importance to note that these figures show that clusters do indeed exist, but, in order to zoom in, it is necessary to go directly to the data and present them in tables, in case an analysis with particular signs is required. This is because the graphics produced by the algorithm are based, as indicated, on a previous stage of reduction of dimensions (the PCA algorithm), in order to go from 26 dimensions (the phonological parameters) down to three dimensions so that data can be properly plotted.

Table 6 presents a detailed zoom-view into the cube displayed in Figure 10, with a count of signs grouped first by cluster (as determined by k-means) and then by the aforementioned official contexts, as well as a grouping in the opposite direction, that is, first by official context and then by cluster. To display this information adequately, this table is made up of eight columns. The first four columns present the dispersion of signs for each context within each cluster, ordering the columns first by "Cluster" and then by "Context", to present their "Count" and "Variance". The second set of four columns presents the dispersion for each cluster within each context, and, to achieve this, the columns are listed in the order of "Context", "Cluster", "Count", and "Variance". As it can be seen, all clusters

include signs from all contexts, with the only exception of the TSE context which does not contribute signs to cluster 1. This fact should be interpreted with care because it does not mean that there are many repeated signs, but that the algorithm managed to detect important phonological similarities across the contexts, without generating any clear partition between them.

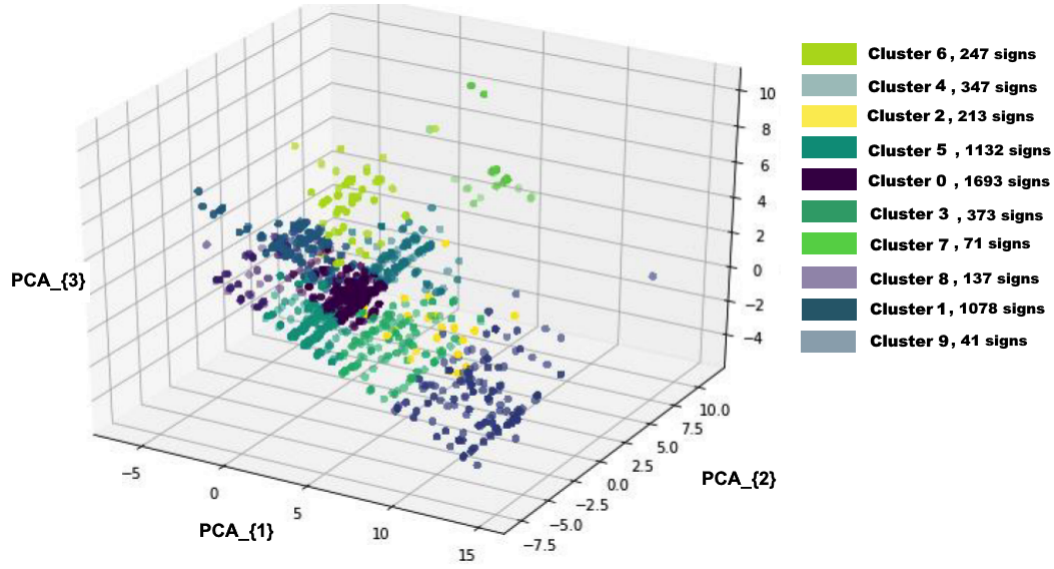

Figure 8. Separate hands cluster, more compact clustering.

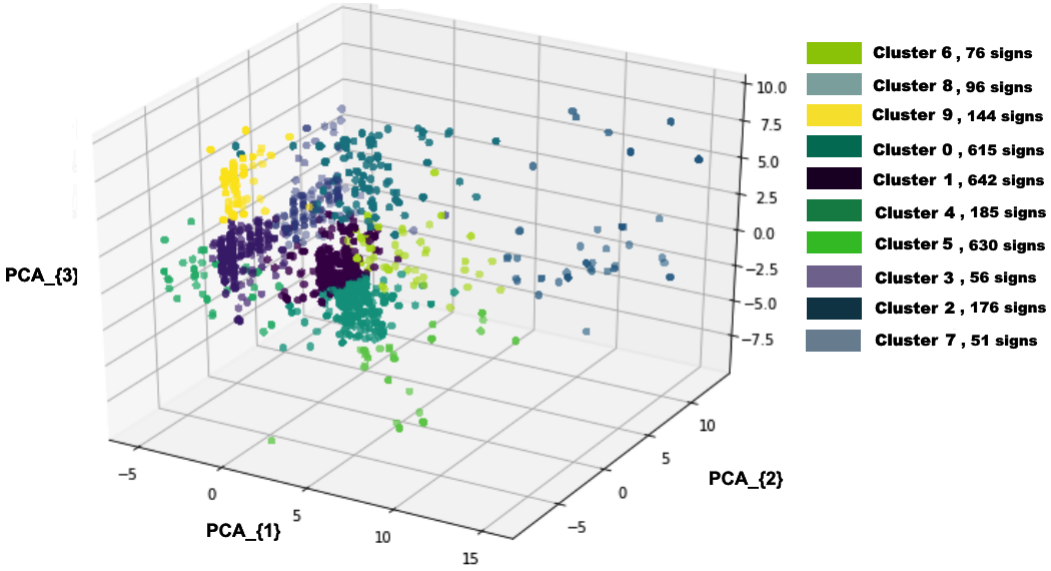

Figure 9. Both hands cluster, less compact clustering.

An extremely interesting observation comes from the fact that grouping the signs first by cluster and then by context produces almost the same average variance than grouping them for each context separately (approximately 4% difference). This means that applying the k-means algorithm produces almost equally compact sign packets across contexts. This is highly relevant, since the time needed to run the algorithm with all the signs from all contexts in a single run is almost the same as it would be separately.

In order to automatically detect hands (a mandatory task conducive to sign language recognition), our findings can greatly facilitate a harsh decision on which contexts the interested researchers should prioritize. We strongly suggest that their contexts be predefined, as in our case and that, by following the steps presented in this paper, they identify those contexts with lower similitude as candidates for a

first proof of concept or a prototype. The specific techniques used for automatic recognition are not the focus of this paper, but the use of deep learning, for example convolutional neural networks, has been strongly suggested [47].

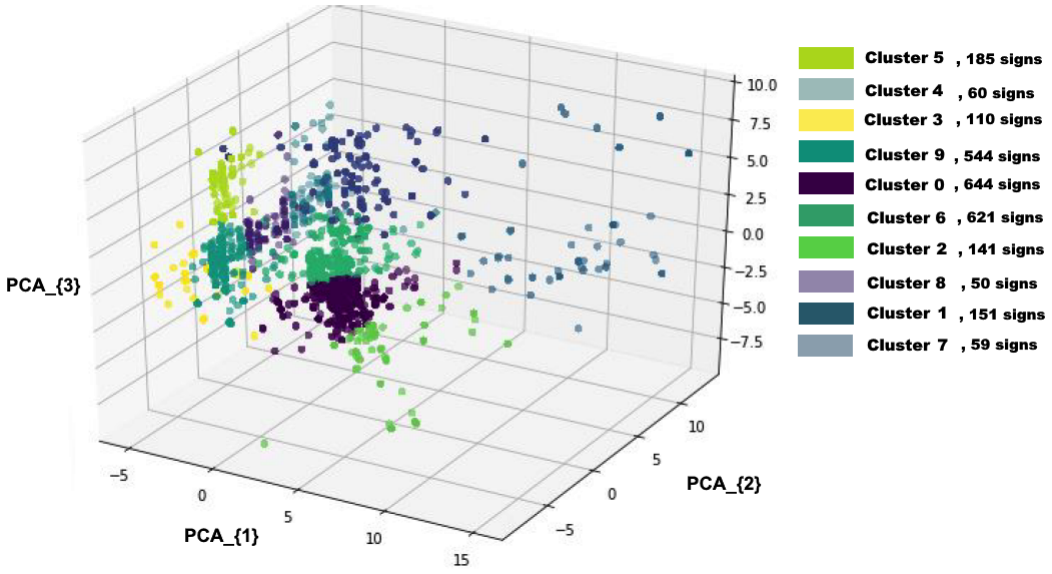

Figure 10. Both hands refined-data, a more accurate clustering.

Table 6. Grouping by cluster and official context versus grouping by official context and cluster.

Cluster	Contexts Count	Sign Count	Variance	Context	Clusters Count	Sign Count	Variance
Cluster 0	8	644	4538.5	CDPCD	10	436	1526.6
Cluster 1	7	151	888.5	CENAREC	10	1000	7922.2
Cluster 2	8	141	201.2	CoopeAnde	10	207	350
Cluster 3	8	110	286.7	HimnoNacional	10	46	12.4
Cluster 4	8	60	44.8	SENATON	10	259	641.7
Cluster 5	8	185	327.6	SICID	10	300	967.2
Cluster 6	8	621	3650.7	TEC	10	197	417.1
Cluster 7	8	59	50.2	TSE	9	120	178.2
Cluster 8	8	50	22.9				
Cluster 9	8	544	4425.3				
		Avg.Var.:	**1443.6**			**Avg.Var.:**	**1501.9**

For instance, for our eight contexts, prioritization would be based on the data shown in Table 7. Thus, the HimnoNacional context yields numbers that turn it into preliminary candidate number one, with 65% high similarity, but it is a static context (the National Anthem of Costa Rica), which is why its translation would be performed only once. With preliminary priorities two and three, the strongest candidate contexts appear, namely TEC and CENAREC, with high similarities of 79% and 81%, respectively. Although these are near levels of similarity, the total count of signs across domains differs widely, which would make the disambiguation tasks more difficult for higher sign counts, therefore giving a higher priority to the TEC context and then to CENAREC. In a very similar scenario, the CoopeAnde, SICID, and TSE contexts follow at priority level. As explained above, HimnoNacional, CDPCD, and SENATON should have the lowest levels of final assigned priorities.

On the other hand, by examining the database, it is possible to determine that the phonological proximity is not usually associated with a semantic similarity, as it is deduced from our experiments by the use of Word2Vec, as previously explained. Calculating the degree of divergence between the two is considered important future work, but, in order to provide an example, let us consider the following

signs, all belonging to cluster 3, as determined by k-means. They are clustered together, and they even belong in the same contexts; nevertheless, they do not exhibit any obvious semantic relationship:

- CRPD, blind
- CRPD, person
- CRPD, facilitate
- ...
- CoopeAnde, say
- CoopeAnde, reduce
- CoopeAnde, refrigerator
- ...
- TEC, permanent
- TEC, who
- TEC, will.

Conversely, signs with an apparently strong semantic relationship may belong in different clusters of phonological proximity. For instance:

- King: cluster 1
- Queen: cluster 3
- ...
- Father: cluster 3
- Mother: cluster 1.

Table 7. Context prioritization for the development of sign language recognition systems.

Context	Sign Count	Sign Count over 90% Similitude	High Similitude Percentage	Preliminary Priority	Assigned Priority
TEC	197	156	79%	2	1
CENAREC	1000	806	81%	4	2
CoopeAnde	207	182	88%	5	3
SICID	300	264	88%	5	4
TSE	120	107	89%	6	5
HimnoNacional	46	30	65%	1	6
CDPCD	436	347	80%	3	8
SENATON	259	234	90%	7	7

Table 8 shows a comparison of the clustering results using the k-means algorithm (k = 10), for one hand with duplicates, for two hands with duplicates, and for two hands without duplicates (refined version). The last row shows the standard deviation for one hand only, both hands and both hands refined because all data in each column is distributed in a way that is very close to a normal distribution. The Kolmogorov–Smirnov formula is 0.30356 with a p-value of 0.2584 for one hand only, 0.33335 with a p-value of 0.32117 for both hands, and 0.20393 with a p-value of 0.17157 for both hands refined. By making a distinction between separate hands, both hands with duplicates, and both hands without duplicates (refined), the data quality is progressively improved, an essential condition for the algorithm to reliably reflect the reality of the contents in the sign database.

The use of k-means revealed the importance of a judicious and considered choice of the input data. Having clusters that reliably represent the reality of the starting database is very important because it allows for exploiting the benefits of applying a clustering algorithm, without introducing unnecessary data that may artificially populate some clusters.

Table 8. In-cluster standard deviation is 45% smaller when studying both hands signaling at the same time. Refined data without duplicates is 4% even smaller.

One Hand Only		Both Hands			Both Hands Refined	
Cluster	**Signs**	**Percent**	**Signs**	**Percent**	**Signs**	**Percent**
0	1078	20.18%	615	23.03%	644	25.11%
1	116	2.17%	642	24.04%	151	5.89%
2	1690	31.64%	176	6.59%	141	5.5%
3	213	3.99%	56	2.1%	110	4.29%
4	1,132	21.19%	185	6.93%	60	2.34%
5	71	1.33%	630	23.59%	185	7.21%
6	394	7.38%	76	2.85%	621	24.21%
7	247	4.62%	51	1.91%	59	2.3%
8	350	6.55%	96	3.59%	50	1.95%
9	51	0.95%	144	5.39%	544	21.21%
	5342	100%	2671	100%	2565	100%
	Standard deviation:	**533.95**	**Standard deviation:**	**240.94**	**Standard deviation:**	**231.73**

5. Conclusions

We have tackled the problem of designing a scheme that allows for characterizing the LESCO phonology by computational means. The importance of addressing this problem is to conform to this scheme, which was not available in Costa Rica prior to this work, and carrying out experiments that demonstrate concentration and dispersion among the signs that make up the lexicon of this sign language. The benefits of working on the subject of this article are practical and of immediate application, mainly in LESCO automatic recognition systems.

Additionally, our experiments have been documented in detail and can be replicated, so the presented scheme is easy to adapt to other sign languages, as long as the interested researchers have identified the contexts, sign names, and a parameterization of the sign language phonology that they wish to study in the form of an array of numerical values.

From the literature review, it is clear that this research takes a step forward, not only by treating the phonological proximity of a particular sign language, but also because emphasis has been placed on providing a perspective beyond linguistic theories and to allow experimentation based on these theories. In a complementary way, the theoretical basis we have used does not explicitly mention sign languages, so our results also look very promising as a new line of research in linguistics.

With regard to the research questions proposed at the beginning of this work, we can conclude that they were successfully answered. Our first research question (RQ1) was stated as "What methods can be used to measure the similarity between signs?" In order to answer this question, we recurred to a thorough theoretical revision on the subject, as well as giving our own insight from an expert point of view on the phonological parameterization of LESCO and previous knowledge on general-purpose similarity measures and clustering techniques. Then, we decided to ponder these measures and justifying the selection of cosine similarity, applied to an array of numerical parameters. In addition, we chose well-proven dimensionality reduction (PCA) and clustering techniques (k-means).

Our second research question (RQ2) was stated as "What relevant information do these methods provide once applied?" To answer this question, we designed and ran a series of experiments that made several things clear.

First, the similarity measures helped to construct uniform ranges with similarity levels expressed in numerical terms, as well as the number of signs belonging to each range. This helped to have a clear idea of the levels of accumulation and dispersion in density neighborhoods, which is vital to align it with the linguistic theory of new sign production versus language learning. The results clearly demonstrate that our LESCO sign database in PIELS is characterized by high degrees of phonological proximity, particularly in the orientation and location components, but noticeably lower in the form component. This fact is most likely reinforced by the fact that our database reflects official, formal contexts, with a clear indication that the production of new signs is easier than learning already

existing ones. This fact also has important connotations in the field of automatic recognition systems, since a high similarity degree makes testing and correction in these systems more demanding.

The results of this study demonstrate that orientation and location have the highest degrees of similarity, while the form has the lowest. For orientation, the similarity accumulates in two tracts, one from 91.65 to 95.73 and the other tract from 97.77 to 99.81. For the location, there is a tract from 94.78 to 100. Finally, the form accumulates in ten tracts: from 65.67 to 66.31, from 75.91 to 76.55, from 79.75 to 80.39, from 81.67 to 82.95, from 84.23 to 84.97 , from 86.79 to 87.43, from 88.71 to 89.35, from 89.99 to 90.63, from 93.19 to 95.75 and the last one from 96.39 to 100.

Second, we used a clustering algorithm that yielded similar results in terms of data dispersion, with clusters of very inconsistent sizes. In addition, the application of the Word2Vec algorithm made it easier to appreciate that, when there are predefined contexts, as in our case, it is basically indistinct to use all the data available for its execution, instead of executing it separately for each context.

Always starting from the linguistic theory that underlies our study, we suggest using our findings weighing the interest of producing new signs versus teaching or automatic recognition. For the community with a direct interest in LESCO, the data provided serves as the basis for making decisions to develop their educational programs, linguistic analyses, or development of automatic recognition software from sign clusters that are prioritized based on the similarity levels.

6. Future Work

As future work, it will be valuable to quantify the divergence between phonological proximity and semantic similarity. In this study, we present some examples of the clear divergence between both types of similarity, but it is a clear option to quantify and characterize in a subsequent study. If it is possible to determine some clusters in which there is a higher correlation, this can serve as an input to teaching and interactive editions of discourses in PIELS, with a new module of suggestions. We also propose as a future possibility the development of an automatic recognition system that makes use of our findings, comparing the stages of tests and corrections across the contexts or domains that we currently have, based on the slogan of developing first those with lower levels of similarity that are therefore less prone to error.

Author Contributions: Conceptualization, L.N.-Z. and M.C.-R.; methodology, L.N.-Z., M.C.-R., J.P., and A.F.; software, L.N.-Z.; validation, L.N.-Z. and M.C.-R.; formal analysis, L.N.-Z.; investigation, L.N.-Z.; resources, M.C.-R.; data curation, L.N.-Z.; writing—original draft preparation, L.N.-Z.; writing—review and editing, L.N.-Z., M.C.-R., J.P., and A.F.; visualization, L.N.-Z.; supervision, M.C.-R., J.P., A.F.; project administration, J.P., A.F.; funding acquisition, M.C.-R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Spanish Ministry of Science, Innovation and Universities through the Project ECLIPSE-UA under Grant RTI2018-094283-B-C32, the Project INTEGER under Grant RTI2018-094649-B-I00, and partly by the Conselleria de Educación, Investigación, Cultura y Deporte of the Community of Valencia, Spain, within the Project PROMETEO/2018/089.

Acknowledgments: The authors thank the School of Computing and the Computer Research Center of the Technological Institute of Costa Rica for the financial support, as well as CONICIT (Consejo Nacional para Investigaciones Científicas y Tecnológicas), Costa Rica, under grant 290-2006. The support of our partners from the design and development departments at Inclutec has been crucial to achieve high quality graphic displays and to gather appropriate and timely data for experimentation, respectively. The feedback received from doctoral student Juan Zamora, from Aspen University, regarding adequate form and concept allowed for conceiving a definitive version of the paper. Our colleagues at Inclutec, Víctor Romero, Johan Serrato, and Demetrio Alvarado provided valuable access to data and feedback on the use of various tools. We greatly appreciate their support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. The Legislative Assembly of the Republic of Costa Rica. Law for the Recognition of Costa Rican Sign Language (LESCO) as a Mother Tongue. Available online: http://www.mtss.go.cr/seguridad-social/discapacidad/Ley_9049.pdf (accessed on 19 July 2012).

2. The Legislative Assembly of the Republic of Costa Rica. Law 20767 on the Recognition of Costa Rican Sign Language (LESCO). Available online: http://www.aselex.cr/boletines/Proyecto-20767.pdf (accessed on 22 May 2018).
3. Naranjo-Zeledón, L.; Peral, J.; Ferrández, A.; Chacón-Rivas, M. A Systematic Mapping of Translation-Enabling Technologies for Sign Languages. *Electronics* **2019**, *8*, 1047. [CrossRef]
4. Luchkina, T.; Koulidobrova, E.; Palmer, J. When you can see the difference: The phonetic basis of sonority in american sign language. In Proceedings of the Annual Meetings on Phonology, Santa Cruz, CA, USA, 18–20 September 2020; Volume 8.
5. Taylor, B. Towards the Automatic Translation of American Sign Language. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, USA, November 2016.
6. Bravo, L. The phonological awareness as a zone of proximal development for the initial learning of reading. *Estudios Pedagógicos* **2002**, *28*, 165–177.
7. Felix Naumann and Melanie Herschel. An introduction to duplicate detection. *Synth. Lect. Data Manag.* **2010**, *2*, 1–87. [CrossRef]
8. Bisandu, D.; Prasad, R.; Liman, M. Data clustering using efficient similarity measures. *J. Stat. Manag. Syst.* **2019**, *22*, 901–922. [CrossRef]
9. Gali, N.; Mariescu-Istodor, R.; Fränti, P. Similarity measures for title matching. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancún, Mexico, 4–8 December 2016; pp. 1548–1553.
10. Tversky, A. Features of Similarity. *Psychol. Rev. Am. Psychol. Assoc.* **1977**, *84*, 327–352. [CrossRef]
11. Gentner, D. Structure-Mapping: A Theoretical Framework for Analogy. *Cogn. Sci.* **1983**, *7*, 155–170. [CrossRef]
12. Medin, D.; Goldstone, R.; Gentner, D. Respects for similarity. *Psychol. Rev. Am. Psychol. Assoc.* **1993**, *100*, 254–278. [CrossRef]
13. Keane, J. Similarity of handshape: An articulatory model. In Proceedings of the ICPhS 2015, International Congress of Phonetic Sciences, Glasgow, UK, 10–14 August 2015; Volume 18.
14. Hildebrandt, U.; Corina, D. phonological proximity in american sign language. *Lang. Cogn. Process.* **2002**, *17*, 593–612. [CrossRef]
15. Williams, J.T.; Stone, A.; Newman, S.D. Operationalization of sign language phonological proximity and its effects on lexical access. *J. Deaf. Stud. Deaf. Educ.* **2017**, *22*, 303–315. [CrossRef]
16. Keane, J.; Sehyr, Z.S.; Emmorey, K.; Brentari, D. A theory-driven model of handshape similarity. *Phonology* **2017**, *34*, 221–241. [CrossRef]
17. Wohlin, C. Second-generation systematic literature studies using snowballing. In Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering, Limerick, Ireland, 1–3 June 2016; pp. 1–6.
18. Field, J. *Psycholinguistics: A Resource Book for Students*; Routledge: England, UK, 2003.
19. Alkoby, K. Toward True asl Dictionaries: New Developments in Handshape Similarity. 2008. Available online: http://asl.cs.depaul.edu/papers/DST2008.pdf (accessed on 20 July 2020).
20. Richards, J.; Hanson, V. Visual and production similarity of the handshapes of the American manual alphabet. *Percept. Psychophys.* **1985**, *38*, 311–319. [CrossRef] [PubMed]
21. Brentari, D. *A Prosodic Model of Sign Language Phonology*; MIT Press: Cambridge, MA, USA, 1998.
22. Valli, C.; Lucas, C. *Linguistics of American Sign Language: An Introduction*; Gallaudet University Press: Washington, DC, USA, 2000.
23. Berent, I.; Dupuis, A. The unbounded productivity of (sign) language: Evidence from the stroop task. *Ment. Lex.* **2017**, *12*, 309–341. [CrossRef]
24. Emmorey, K.; Lane, H. (Eds.) *The Signs of Language Revisited: An Anthology to Honor Ursula Bellugi and Edward Klima*; Psychology Press: East Sussex, UK, 2013.
25. Sandler, W.; Lillo-Martin, D. *Sign Language and Linguistic Universals*; Cambridge University Press: Cambridge, UK, 2006.
26. Stokoe, W. Sign language structure: An outline of the visual communication systems of the American deaf. *J. Deaf. Stud. Deaf. Educ.* **1960**, *10*, 3–37. [CrossRef] [PubMed]
27. Petitto, L.; Holowka, S.; Sergio, L.; Ostry, D. Language rhythms in baby hand movements. *Nature* **2001**, *413*, 35–36. [CrossRef]

28. Vitevitch, M. The Influence of phonological proximity Neighborhoods on Speech Production. *J. Exp. Psychol. Learn. Mem. Cogn.* **2002**, *28*, 735–747. [CrossRef]
29. Brentari, D.; Padden, C. Native and foreign vocabulary in american sign language: A lexicon with multiple origins. In *Foreign Vocabulary in Sign Languages: A Cross-Linguistic Investigation of Word Formation*; Lawrence Erlbaum Associates: New Jersey, NJ, USA, 2001; pp. 87–119.
30. Hendriks, B.; Dufoe, S. Non-native or native vocabulary in mexican sign language. *Sign Lang. Linguist.* **2014**, *17*, 20–55. [CrossRef]
31. Keck, T.; Wolgemuth, K. American sign language phonological awareness and english reading abilities: Continuing to explore new relationships. *Sign Lang. Stud.* **2020**, *20*, 334–354. [CrossRef]
32. Caselli, N.K.; Sehyr, Z.S.; Cohen-Goldberg, A.M.; Emmorey, K. Asl-lex: A lexical database of american sign language. *Behav. Res. Methods* **2017**, *49*, 784–801. [CrossRef]
33. Meade, G.; Midgley, K.J.; Sehyr, Z.S.; Holcomb, P.J.; Emmorey, K. Implicit co-activation of american sign language in deaf readers: An erp study. *Brain Lang.* **2017**, *170*, 50–61. [CrossRef]
34. Williams, J.T.; Newman, S.D. Spoken language activation alters subsequent sign language activation in l2 learners of american sign language. *J. Psycholinguist. Res.* **2017**, *46*, 211–225. [CrossRef]
35. Carreiras, M.; Gutiérrez-Sigut, E.; Baquero, S.; Corina, D. Lexical processing in Spanish sign language (LSE). *J. Mem. Lang.* **2008**, *58*, 100–122. [CrossRef]
36. Gahl, S.; Yao, Y.; Johnson, K. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *J. Mem. Lang.* **2012**, *66*, 789–806. [CrossRef]
37. Steinbach, M.; Mani, N.; Wienholz, A.; Nuhbalaoglu, D.; Herrmann, A. Phonological Priming in German Sign Language: An Eye Tracking Study Using the Visual World Paradigm. 2019. Available online: https://psyarxiv.com/x5pts/ (accessed on 21 May 2020).
38. Cardin, V.; Campbell, R.; MacSweeney, M.; Holmer, E.; Rönnberg, J.; Rudner, M. Neurobiological insights from the study of deafness and sign language. *Underst. Deaf. Lang. Cogn. Dev. Essays Honour Bencie Woll* **2020**, *25*, 159.
39. Thompson, R.L.; England, R.; Woll, B.; Lu, J.; Mumford, K.; Morgan, G. Deaf and hearing children's picture naming: Impact of age of acquisition and language modality on representational gesture. *Lang. Interact. Acquis.* **2017**, *8*, 69–88. [CrossRef]
40. Villameriel, S.; Costello, B.; Dias, P.; Giezen, M.; Carreiras, M. Language modality shapes the dynamics of word and sign recognition. *Cognition* **2019**, *191*, 103979. [CrossRef]
41. Karl Pearson, L., III. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572. [CrossRef]
42. MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*; Lucien, M.L.C., Jerzy N., Eds.; University of California Press: Berkeley, CA, USA, 1967; Volume 1, pp. 281–297.
43. Li, S. Understanding Word2vec Embedding in Practice. Available online: https://towardsdatascience.com/understanding-word2vec-embedding-in-practice-3e9b8985953 (accessed on 21 May 2020).
44. Serrato-Romero, J.; Chacón-Rivas, M. Traductor LESCO: Un esfuerzo puntual en el apoyo al proceso de aprendizaje de estudiantes con discapacidad auditiva. *Investiga.TEC* **2016**, *27*, 4.
45. Massey, F.J., Jr. The Kolmogorov–Smirnov test for goodness of fit. *J. Am. Stat. Assoc.* **1951**, *46*, 68–78. [CrossRef]
46. United Nations. Convention on the Rights of Persons with Disabilities and Optional Protocol. Available online: https://www.un.org/disabilities/documents/convention/convoptprot-e.pdf (accessed on 13 April 2020).
47. Zamora-Mora, J.; Chacón-Rivas, M. Real-Time Hand Detection using Convolutional Neural Networks for Costa Rican Sign Language Recognition. In Proceedings of the 2019 International Conference on Inclusive Technologies and Education, San Jose del Cabo, Mexico, 30 October–1 November 2019; pp. 180–186.