

Automatic Concept Explanation for Deaf Users

André Dias¹ and Nuno Escudeiro²

¹ Instituto Superior de Engenharia do Porto
1140279@isep.ipp.pt

² Instituto Superior de Engenharia do Porto
nfe@isep.ipp.pt

Abstract. Concept explanation has a great importance in languages with small lexicons, such as sign languages. In Portugal, the official language used by the deaf community and the community that surrounds them is known as Língua Gestual Portuguesa (LGP). This language, like other sign languages has a small lexicon that is not able to represent all the words of the Portuguese language. The users of LGP, when faced with words like 'Nanotechnology' have to resort to tools made for Portuguese because the ones available to them are not ideal. Those solutions either lack a translation, which is the case of online dictionaries or aren't practical, which is the case of sign language interpreters. To solve this problem, this project was created with the hypothesis of verifying if it's possible to utilize Text Mining, Information Scraping and Information Retrieval techniques to generate the explanation of a given word or expression that does not exist in the LGP lexicon. The developed solution consists in an Application Programming Interface (API) capable of generating an explanation of a given word or expression. This API will feed a web application responsible for receiving the user input and presenting the explanation in plain text and its translation to LGP using an avatar. The solution is capable of displaying explanations in plain text, generated with the mentioned techniques, and exhibiting a clear indicator of their LGP legibility for a deaf user.

Keywords: Sign language · Text summarization · Automatic concept explanation.

1 Introduction

Concepts are abstract or generic ideas, that have a common set of features that are shared across multiple situations and contexts. Concept explanation can take a fundamental part of the learning process[1], whether it takes place in an educational environment, a day-to-day or a professional context. The importance of concept explanation is even greater when it comes to languages with a small lexicon, in particular sign languages.

According to the World Health Organization [2], as of march 2018, there were 466 million people with disabling hearing loss, this represents 5% of the world population. Without being able to hear, deaf people are not able to communicate using sound, therefore they must use sign languages.

ISBN: 978-989-54758-6-5

In Portugal the official language of the deaf community is known as *Língua Gestual Portuguesa* (LGP). This language is also used by the surrounding community, such as their relatives, teachers, and so on. The lexicon of LGP is composed by multiple signs where each one represents a word or an expression in Portuguese. However, the opposite is not true. There are numerous words in Portuguese that don't have a sign that translates them.

In order for a LGP user to understand those words/expressions the explanation must be formed by the available lexicon[3], yet there are no LGP encyclopedias. The closest thing available is the LGP dictionaries but they fail to translate words that have no sign that can directly translate them.

This problem is recurrent when it comes to scientific domains. To understand a concept like 'Nanotechnology', which there are no signs for it, a LGP user has to access tools with content made for the Portuguese users.

Even though LGP is the language taught to the people that are not able to communicate using Portuguese, they are two very different languages. Each with their own lexicon and grammar, so most of the LGP users will struggle to understand Portuguese content.

A LGP user might also resort to a sign language interpreter that would explain the meaning of the word to him, but this would be an even less practical solution

To solve this problem for the users of LGP, this project was created with the hypothesis that is possible to utilize Text Mining[4], Information Scraping[5][6] and Information Retrieval[7] techniques to generate the explanation of a given word or expression in LGP that does not exist in its lexicon.

To test this hypothesis, the developed project consists of three major components. An API capable of generating explanations of a given word. A formula that calculates the LGP readability score of each expression. A web app for the users to interact with, that allows them to search for a word and displays the generated explanation in plain text, as well as a translation in LGP with the help of a previously developed avatar.

Concerning results, it's expected that this project can to generate accurate LGP explanations that were classified with a reliable LGP readability score by the developed formula.

After this introduction, the paper is divided into four more sections: Section 2 presents a deeper explanation of the partial solutions available, as well as the Portuguese readability metrics that were the foundation of the LGP readability metric. Section 3 describes the developed solution. Section 4 explains this project's contributions and how it's going to be evaluated to confirm or deny the previously mentioned hypothesis. Section 5 concludes and how it can evolve to become more complete and target a bigger audience.

2 Related Work

As mentioned in the introduction, there's a lack of solutions, capable of translating concepts from Portuguese to LGP. There are however some partial solutions.

This section will describe those partial solutions, as well as the existing Portuguese readability metrics that were analyzed in order to create the LGP readability metric.

2.1 Online Dictionaries

One of the most known online dictionaries with LGP content is the Spreadthesign [8]. In this dictionary, a user can search for a word and if there's a previously recorded video translation for that word, at the site database, it will be displayed for the user. The search results only display the video of a person performing the corresponding sign, and the possibility to look at the same word in another sign language.

Another online dictionary that provides content for the LGP users is the Infopédia [9]. Although this is a Portuguese dictionary, it also contains a section for searching words in LGP. Here, the search results, not only provide a video translation of the word, but also an explanation in Portuguese on how to reproduce the sign shown in the video.

Online dictionaries as a solution, is limited by the database of prerecorded videos, and the LGP lexicon. The latter is a problem, because they focus on a direct translation, word to sign, instead of trying to translate the meaning of the word.

2.2 Sign Language Interpreter

Concerning the sign language interpreters in Portugal, there is CTILG [10], a company that provides professional LGP translation services in workshops, classes, congresses, events and more. This company is responsible for the live translation of some morning TV shows.

A more affordable solution for a regular LGP user, is the Serviin [11] which is a service that provides an interpreter to work as a middle-man between a deaf person and a targeted service/company. This solution is available as a mobile app, with a very low cost for the deaf user, or through a web app that is free.

This solution is limited by the interpreters own knowledge and their cost. Also by utilizing this solution, the LGP user is sacrificing some of his autonomy.

2.3 Portuguese Readability Metrics

Readability metrics are used to calculate a score [12], that relates to the level of education a reader will need, to fully understand the context of a given text. There are some widely known and used metrics for the English language, such as Flesh Reading Ease, New Dale-Chall, SMOG, Flesh-Kincaid, Gunning Fog and so on.

In 2019, Antunes, H and Lopes, C published an article [13] that adapted the values of the metrics used to calculate the readability of text in English so it could be applied to Portuguese. The adapted readability metrics are presented in Table 1.

Table 1. Adjusted Portuguese formulas.

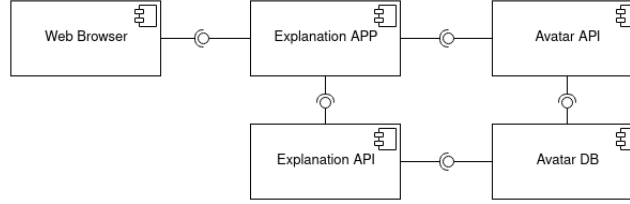
	Formula
SMOG	$16.830 \times \sqrt{CW} \times 30 \div SE - 23.809$
Flesch-Kincaid	$0.883 \times WO \div SE + 17.347 \times SY \div WO - 41.239$
ARI	$6.286 \times CH \div WO + 0.927 \times WO \div SE - 36.551$
Coleman Liau	$5.730 \times CH \div WO - 171.365 \times SE \div WO - 6.662$
Gunning Fog	$0.760 \times WO \div SE + 58.600 \times CW \div WO - 12.166$

CH - characters, CW - complex words, SY - syllables, WO - words, SE - sentences

3 Automatic Concept Explanation

To solve the previously mentioned problems, Automatic Concept Explanation was created. The main functionality of this solution, is to allow the LGP users to search for a word or expression in Portuguese and display its explanation as well as the respective LGP translation.

All the components can be seen in Figure 1.

**Fig. 1.** Component Diagram.

A user interacts with the web application that was developed in React[14], an open-source JavaScript framework for building user interfaces. Here, the user can search for the desired concept, as well as changing the page language. This is also the place where the generated explanations are shown in plain text and in LGP by the avatar. After each explanation, it is also presented their LGP Readability Score, which is presented more in depth in the following subsection, as well as feedback options. If one of the presented explanations is not acceptable, it is possible for the user to request a new one.

When a concept is searched for, the input is sent to an API that was developed in Flask[15], a lightweight open-source Python web framework. This API will generate a list of explanations using only web scraping or a combination of web crawling, web scraping and text summarization. The approach to be used is set on its configuration file, and will depend on the initial source of information. For websites where a single page can be a rich source of information like in Priberam[16], the first approach can yield good results. On the other hand, for websites where the information is scattered across many pages like Wikipedia[17], the second approach is ideal.

All the tasks that compose each approach were developed using Python libraries: Scrapy³ for web-crawling, BeautifulSoup⁴ for web scraping and Natural Language Toolkit (NLTK)⁵ for text summarization.

The list of explanations is sorted based on each explanation's LGP Readability Score, and the ones with the lowest score are selected.

The avatar and its database are projects that were previously developed by GILT and are used to enhance this project.

3.1 LGP Readability

The developed API always generates more than one explanation, and the number is even higher for words that have different meanings based on its context. Those explanations need to be sorted in order to present the users with the ones best suited for their needs.

For a regular language, there are metrics, like the readability score, that can be used to classify each expression in order to sort them accordingly. This score indicates how easily a reader would comprehend the text that he's reading.

There are multiple formulas that could be used for calculating this score in Portuguese, as shown in Table 1. However, there is none for the LGP.

When looking at the readability formulas for Portuguese, it is easy to notice that, all of them take in consideration the same variables, which are common in every written text: characters, complex words, syllables, words and sentences.

After analyzing the LGP signs, with the goal of creating a new readability calculation formula, all the shared variables were identified:

- Hand configurations (CF) - The hand shape in a particular moment.
- Moments (MT) - The position of the hand in relation to the body.
- Hands (HS) - Both hands or only the dominant hand.
- Facial expressions (FE) - Motion or position of the face muscles.

Using those variables the following formula was created:

$$(0.7 \times CF + 0.3 \times MT + 1 \times FE) \times (0.5 \times HS) \quad (1)$$

The formula was tested using the signs from the avatar database and the constant values, that were initially set to 1, were manually adjusted to produce a more compact interval of results. However, the constant value for the facial expressions was unaltered due to the current version of the avatar not supporting them. In the Table 2 is shown an example of some signs and their readability score.

³ <https://scrapy.org/>

⁴ <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

⁵ <https://www.nltk.org/>

Table 2. Readability scores example.

Sign	CF	MT	HS	Score
Javali	1	1	1	1.00
Fornecedor	2	6	1	2.09
Auxílio	2	4	2	3.59
Consumo	4	6	2	5.60
Esclarecer	7	7	2	8.00

4 Contributions and Evaluation

The main contribution is an application capable of helping in the learning process of a LGP user, by allowing him to find explanations for a word or expression in his main language. Another important contribution, is a formula that can be used to calculate the readability of LGP content. This can make it easier to produce said content, and by doing so, promoting the inclusion and equality of opportunities for the deaf community.

Although the project is not yet complete, the already developed functionality is capable of generating explanations in plain text using web crawling, web scraping and text summarization and the avatar can to translate plain text into LGP. Thus corroborating the hypothesis that is possible to utilize Text Mining, Information Scraping and Information Retrieval techniques to generate the explanation of a given word or expression that does not exist in the LGP lexicon.

5 Conclusions and Future Work

In conclusion, the project is still in development and this article presents a general view of it.

The LGP readability formula was able to produce reliable scores when there was no need for signs that required facial expression. However, a new version of the avatar, that uses facial expressions, is being released which will allow to obtain values to calibrate the formula for the signs that do require facial expressions.

At the current stage, the application can to produce explanations based on the word searched and show it to the user in plain text. The explanation to be shown is not yet selected based on the readability formula. In the cases where it was generated using only web scraping, the selection was based on the order from the source page. If it was generated using text summarization, the selection was based on the order created by its process.

Furthermore, other functionalities will still be implemented or changed to better fulfill the need of the target audience. The main feature to still be implemented is the integration of the avatar, that is responsible for translating the explanations displayed in plain text to LGP. Another feature not yet implemented, is the displaying of the LGP readability score next to each explanation.

As well as the possibility for the user to ask for a new explanation if any of those available is deemed unacceptable.

Concerning future work, the project can target an even broader audience by supporting other languages which, the avatar, is capable of performing the translation from plain text to the corresponding sign language. Additionally, the developed API may be used to further enhance the functionality of future or already existing GILT projects.

References

1. Ghorbani, A., Wexler, J., Zou, J., Kim B.: Towards Automatic Concept-based Explanations. In: Wallach, H., Larachelle, H., Beygelzimer, A., Alchebuc, F., Fox, E., Garnett, R. (eds.) *Advances in Neural Information Processing Systems* 32, pages 9277–9286. Curran Associates, Inc. (2019) <http://papers.nips.cc/paper/9126-towards-automatic-concept-based-explanations.pdf>
2. WHO Deafness and hearing loss, <https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss>. Last accessed 19 Aug 2020
3. Musselman, C.: How Do Children Who Can't Hear Learn to Read an Alphabetic Script? A Review of the Literature on Reading and Deafness. In: *The Journal of Deaf Studies and Deaf Education*, vol. 5, pages 9–31. (2000) <https://doi.org/10.1093/deafed/5.1.9>
4. Gupta, V., Lehal, G: A Survey of Text Mining Techniques and Applications. In: *Journal of Emerging Technologies in Web Intelligence*, vol.1, pages 60–76. Academy Publisher (2009) <http://www.jetwi.us/uploadfile/2014/1230/20141230112729939.pdf>
5. Young, J., Sinha, P., Lundberg, S., Kalis, J. 2004: Method and apparatus for scraping information from a website. US20060095377A1, viewed 7 Sep 2020, retrieved from Google Patents. <https://patents.google.com/patent/US20060095377A1/en>
6. Rose, R. 2012: Automatically scraping and adding contact information . US9665574B1, viewed 7 Sep 2020, retrieved from Google Patents. <https://patents.google.com/patent/US9665574B1/en>
7. Baeza-Yates, R., Ribeiro-Neto, B.: *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., USA, (1999)
8. Spread The Sign Homepage, <https://www.spreadthesign.com/>. Last accessed 26 Aug 2020
9. Infopédia Portuguese Sign Language dictionary, <https://www.infopedia.pt/dicionarios/lingua-gestual>. Last accessed 26 Aug 2020
10. Ctilg Homepage, <http://www.ctilg.pt/>. Last accessed 27 Aug 2020
11. Serviin Homepage, <http://www.portaldocidadaosurdo.pt/Serviin>. Last accessed 27 Aug 2020
12. Meyer, Bonnie J.: Text Coherence and Readability. In: *Topics in Language Disorders*, vol. 23, pages 204–224. (2003)
13. Antunes, H., Lopes, C.: Analyzing the adequacy of readability indicators to a non-english language. In: *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 149–155. Springer (2019).
14. React Homepage <https://reactjs.org/>, Last accessed 6 Sep 2020
15. Flask Documentation <https://flask.palletsprojects.com/en/1.1.x/>, Last accessed 6 Sep 2020

16. Priberam Homepage <https://dicionario.priberam.org/>, Last accessed 5 Sep 2020
17. Wikipedia Homepage <https://www.wikipedia.org/>, Last accessed 5 Sep 2020