




# Algorithm for Selecting Words to Compose Phonological Assessments

João Vítor B. Marques<sup>1</sup><sup>a</sup>, João Carlos D. Lima<sup>1</sup><sup>b</sup>, Márcia Keske-Soares<sup>2</sup><sup>c</sup>, Cristiano C. Rocha<sup>3</sup>,  
Fabrício André Rubin<sup>4</sup> and Raphael Vieira Miollo<sup>1</sup>

<sup>1</sup>*Centro de Tecnologia, Universidade Federal de Santa Maria, Santa Maria, Brazil*

<sup>2</sup>*Centro de Ciências da Saúde, Universidade Federal de Santa Maria, Santa Maria, Brazil*

<sup>3</sup>*Xebia Data, Eindhoven, Netherlands*

<sup>4</sup>*Petroleo Brasileiro S.A., Rio de Janeiro, Brazil*

**Keywords:** Speech Therapy, Phonological Processes, Graphs and Phonological Assessments.

**Abstract:** The phonological assessment is one of the main resources that speech-language therapist has to identify phonological disorders in children. For this, it is necessary to be composed of a words set that have a variety of phonemes in different positions of the syllable and the word, in order to obtain a representative sample of the phonological system. Thus, the present work aimed to analyze a set of 84 words from a phonological assessment instrument, with the objective of identifying and removing words with over-represented phonemes. Aiming to facilitate the phonological evaluation by making it more succinct with the reduction of the number of words, the present work describes a judicious method organized in three steps, which was implemented in *Javascript* and obtained a subset of 55 words, which have at least two occurrences of the same phonemes in the proper positions in which they appeared in the initial set, representing a 35% reduction in the number of words without losing quality.

## 1 INTRODUCTION

A thorough and comprehensive phonological assessment is one of the main tools for the speech therapist (Savoldi, 2012), as it helps in the identification and diagnosis of speech disorders. To compose a phonological assessment tool, it is necessary to select words that represent figures known to children, and that are inserted in their basic vocabulary and social context (Gomes et al., 2013).


In southern Brazil, the Child Phonological Assessment (CPA) (Yavas et al., 2001) is one of the evaluations used by speech therapists, and according to the authors, 125 words were chosen, which represent the vocabulary of children aged 3 with a balanced sample of the adult phonological system, and present, at least, three possibilities of occurrence for each consonant sound, in all possible syllable positions.


In this sense, in order to represent the adult phonological system, it is important that the set of words is


comprehensive in relation to the variety of phonemes and, also, that they are evaluated more than once in different syllable and word positions (Pagliarini, 2009). Thus, it is also possible to evaluate and identify the phonemes present in the child's phonetic inventory, that is, those that he can reproduce spontaneously (Stoel-Gammon, 1985).

The selection of the best words to compose a phonological assessment has been the subject of studies such as (Savoldi et al., 2013), where 116 words were selected from an initial set of 722, after validation with experts in the field. This number was reduced to 84 in the study of (Ceron et al., 2020), when obtaining a balanced sample of words in an attempt to avoid the over-representation of one phoneme or the under-representation of another. However, it is observed that many phonemes in this set still repeat in the same position, with a frequency greater than the minimum necessary for a quick and effective assessment.

It was then sought to develop a method to evaluate a set of words through their phonetic transcriptions, in order to determine the smallest sub-set that contains a minimum number of occurrences for each

<sup>a</sup> <https://orcid.org/0009-0007-3206-725X>

<sup>b</sup> <https://orcid.org/0000-0001-9719-3205>

<sup>c</sup> <https://orcid.org/0000-0002-5678-8429>

phoneme in the different positions of the syllable and the word, respecting the due restrictions that make a reliable phonological assessment instrument.

The work is organized as follows. In the next section, the basic concepts and related works that served as the basis for this research will be presented. In Section 3, the logic behind the choice of words that make up a phonological assessment instrument will be presented, and then the algorithm will be detailed. In Section 4, the results obtained by this work will be presented and discussed. Finally, in Section 5, the conclusion is presented, and in the appendix all the words of the initial set are found, and the crossed-out words signal that these have been removed from the final set.

## 2 BACKGROUND AND RELATED WORK

In this section, some theoretical concepts that underlie the use of structures such as graphs in this study will be addressed, as well as concepts from speech therapy such as Phonological Processes (PPs) and how they are connected to the present work.

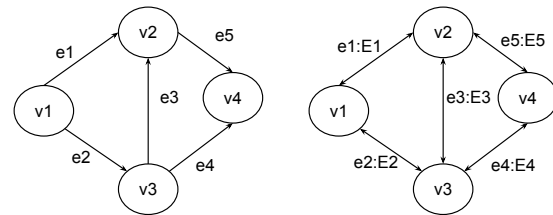
### 2.1 Important Concepts

The reader will be briefly introduced to some terms from speech therapy, that are important for understanding the present work. In cases of phonological disorders, only consonant phonemes are observed (Shriberg et al., 1997). In addition, they can appear in different positions in the syllable and word (beginning, middle, and end), and the production of each phoneme in a pronunciation must necessarily be in one of these positions seen below.

- (OI) Onset Initial: beginning of syllable, word beginning - ca.sa [*house*];
- (OM) Onset Medial: beginning of syllable, middle of the word - ca.va.lo [*horse*];
- (CM) Coda Medial: end of syllable, middle of the word - ca.dar.ço [*shoelace*];
- (CF) Coda Final: end of syllable, end of the word - a.mor [*love*];
- (OCI) Onset Complex Initial: beginning of syllable, beginning of word - **Bra**.sil [*Brasil*];
- (OCM) Onset Complex Medial: beginning of syllable, middle of the word - bi.**bl**io.te.ca [*library*].

### 2.2 Graphs

A graph  $G = (V, E)$  is a structure in which  $V$  is a finite and non-empty set of  $n$  vertices, and a set  $E$  of  $m$  edges, which are pairs of vertices of  $V$ . They are classified according to the nature of the connection between their vertices, being able to be “undirected” when their weights do not have direction, or “directed”, as seen in Figure 1.



(a) Directed

(b) Double Directed

Figure 1: Types of Graphs.

In this work, a Double Directed Graph structure was used, where each edge has, necessarily, two weights: one for each direction. By analyzing the example of Figure 2, we can notice that, given the vertex  $p(OI)$  the edge that connects it to the vertex  $t(OM)$  has different weights depending on the choice of the central vertex, highlighting the importance of centrality in graphs to determine the influence of one vertex on another.

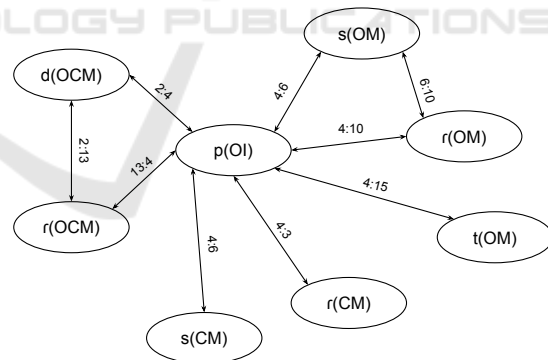


Figure 2: Partial mapping of the entry data.

The weights of each vertex are determined according to the number of words in the set in which a phoneme can occur in a certain position. It is also worth of notice that words which have more than one phoneme influence the weight of more than one vertex. Therefore, when removing or adding words to a vertex, the weights of its edges will be updated as well.

### 2.3 Graph Centrality

Introduced by (Bavelas, 1948) while studying the communication of individuals and influence in small social groups, the concept of centrality in graphs is associated with the degree of importance and influence that each vertex has on another in the graph, and what bottlenecks may exist in their connections. (Freeman, 1978) also works with the same concepts of centrality in social networks, investigating the quantitative measures capable of defining the importance of each vertex.

In this context, this work seeks to identify the importance that each phoneme has in the set of analyzed words, considering possible phonetic transcriptions where a phoneme can appear in more than one position of the syllable and the word. In this way, it is possible to identify the most influential phonemes in the set, that is, those that are over-represented, and make them less important by removing some of their words, making the set more balanced.

Each vertex has a list of words in which a phoneme occurred in a certain position. With that, it is possible to identify the importance that each word has in the vertices, in order to list which could or could not be removed from the graph without the same ending with under-represented phonemes.

### 2.4 Phonological Processes

In the context of speech therapy, phonological processes have a great influence on a child’s language acquisition process. It is expected that during this stage, she applies several phonological processes, such as replacing one phoneme with another or omitting them. Such substitutions and omissions are considered in speech therapy as Phonological Processes (PP), and some examples are presented in Table 1.

However, if a PP persists for a long time, it can become a phonological disorder and remain in the child’s speech, accompanying her in school during her literacy process, bringing harm to her social life (Goulart and Chiari, 2014). Therefore, some works are dedicated to identifying possible phonological disorders through voice recognition in phonological evaluations (Franciscatto et al., 2019), so that the diagnosis and treatment is given early.

In the work of (Franciscatto et al., 2019), Machine Learning (ML) techniques were used to classify the pronunciations of 84 words as correct or incorrect and to recognize phonological processes through them. In (Franciscatto. et al., 2018), a case-based method, commonly used in the health field and in ML techniques (Tavana et al., 2022), is developed and is

able to have good learning while allowing new cases to be stored in a database without complications (Husain and Pheng, 2010). The method works as an extra validation layer after the pronunciation classification by ML, registering new cases and validating them with an expert.

However, despite studies using Artificial Intelligence (AI) techniques (Iliya and Neri, 2016) and ML (Franciscatto et al., 2019) in speech therapy, they are only used to identify phonological processes and classify (correct/incorrect) the pronunciation of spoken words in phonological evaluation. But, before that, a set of words must be chosen by a specialist to be pronounced by the children, and such a set must analyze all the phonemes of the language in different positions of the word. Thus, the choice of words in the set must follow specific criteria, addressed in Section 3.1, because in a phonological evaluation, for example, all phonemes must be analyzed at least twice (Stoel-Gammon, 1985). So, to define the smallest subset of words that meets the same criteria as the initial set, it is a matter of quantifying the “least effort”, discussed in Section 3, rather than learning from errors and successes.

## 3 DEVELOPMENT

In this section, the operation of the algorithm will be presented and detailed, from the basic input structures to the final result.

The set of 84 words from (Ceron et al., 2020) was used as the database. Additionally, all the words in the set should be phonetically decomposed in order to detail in which positions the phonemes that compose them appear. For this, the JSON structure, developed in the work of (Marques, 2022), was used, which is synthesized in Figure 3.

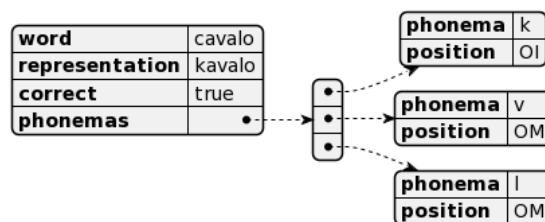


Figure 3: Example of the word “cavalo [horse]” [horse] decomposed in consonant phonemes.

The main idea of the algorithm is to avoid as much as possible that the same phoneme is over-represented in the set of words, being analyzed in the same position with a frequency equal to or greater than the min-

Table 1: Phonological processes in Portuguese acquisition. (Yavas et al., 2001).

Phonological Process (PP)	Definition	Example
Cluster reduction	Reduction of a consonant within the same syllable.	bruxa (witch) [brúsa]→[búsa]
Final fricative deletion	Deletion of phoneme /s/ in both syllable and word final positions	lápiz (pencil) [lápiz]→[ápi]
Liquid substitution	Substitution of one liquid for another.	zero (zero) [zeru]→[selu]
Plosivation	Substitution of a fricative consonant for a plosive	vaca (cow) [vaka]→[baka]
Intervocalic liquid nasalization	Substitution of a liquid by a nasal in intervocalic position	carro (car) [káχ u]→[kámu]

imum necessary, avoiding the under-representation as well.

In order for a phonological assessment to be reliable and the obtained set can be used, it is necessary to follow some rules of speech therapy, which will be discussed in Section 3.1.

### 3.1 Phonological Assessment Rules

In order for an assessment to be reliable and comprehensive, phonemes need to be analyzed in different positions of the syllable and word with a defined minimum frequency, in order to contain a balanced sample of the adult phonological system (Savoldi, 2012).

Also, to determine the presence or absence of the sound in the phonetic inventory, a minimum of two occurrences of the segment can be considered, regardless of the position in the word (Stoel-Gammon, 1985). Studies such as (Yavas et al., 2001) consider, for phonological assessment, a minimum of three possibilities of occurrence for each consonant phoneme in all possible syllable positions. In this work, a minimum of two occurrences of the same phoneme evaluated in a certain position was considered, according with (Stoel-Gammon, 1985). Aiming at the flexibility of the algorithm for its applicability in different scenarios from this research, the minimum number of occurrences is one of the inputs provided by the user.

It should be noted that not all phonemes appear in all allowed positions, due to a limitation of the language itself (Savoldi, 2012). In the input data, it was observed that the phoneme \*n always appears in word productions in the diminutive form in OM and therefore should not have an impact on the proposed method, as it would not occur in words in their natural form.

In short, the method introduced by this work must follow the criteria:

- C.1** The phoneme /ŋ/ evaluated in OM is disregarded — words in diminutives;
- C.2** Each phoneme must continue to occur at least 2 times in the same position after deleting any word.

Next, in Section 3.2 the first step of the algorithm will be detailed.

### 3.2 Phoneme Mapping and Graph

As a first step, the words that have a certain phoneme in a certain position are counted on a map. The pair “phoneme (position)” will be the key of each entry in the map, which will point to a list of words in which the phoneme appears in the right position. As a result, a graph structure is obtained, shown in Figure 4.

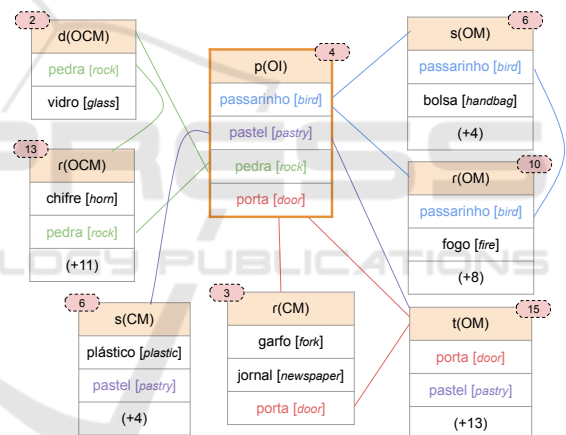


Figure 4: Sample of the graph generated after phoneme mapping.

The algorithm must receive a set of words broken down into their consonant phonemes, as shown in Figure 3. In this study, the abstract set from the work of (Marques, 2022) was used as input, which decomposed the 84 words proposed by (Ceron et al., 2020) into consonant phonemes with the validation of experts in the field.

From the analysis of Figure 4, we observe that some phonemes (/t/) occur in the same position in a greater number of words (15) than the minimum necessary established by criterion C.1. Therefore, since the phoneme repeats a lot in the set, it is thought that some words can be excluded from the evaluation, since they would be generating a super-representation of the phoneme.

But which ones could be excluded? It is noted that the phoneme /s/ occurs in Onset Medial (OM) in 6 words. According to the minimum value of occurrences criterion, 4 of these words could be disregarded, but it is not that simple. Before removing a word from the evaluation, it is necessary to remember that it is composed of other phonemes that are evaluated in different positions. And, if the word that we are willing to exclude is one of the few that evaluates some other phoneme in a certain position? All these issues are considered by the algorithm proposed by this work.

### 3.3 Word Analysis

As a second step, through the graph shown in Figure 4, we will analyze each node and the words contained in it. As we visit each node, we can remove words, as long as the graph remains valid according to the criteria established in Section 3.1.

Given a node of the graph, we need to determine if the phoneme is over-represented in a certain position or not. As an example, we will work with the node highlighted in Figure 4, where the phoneme /p/ appears in OI in 4 different words. Thus, it can be said that this pair is over-represented, since it occurs in more words than the minimum necessary.

Now, we need to check if it is possible to remove any of the words, without compromising the final set with under-represented phonemes. We do this by analyzing each word of the visited node separately, as shown in Figure 5.

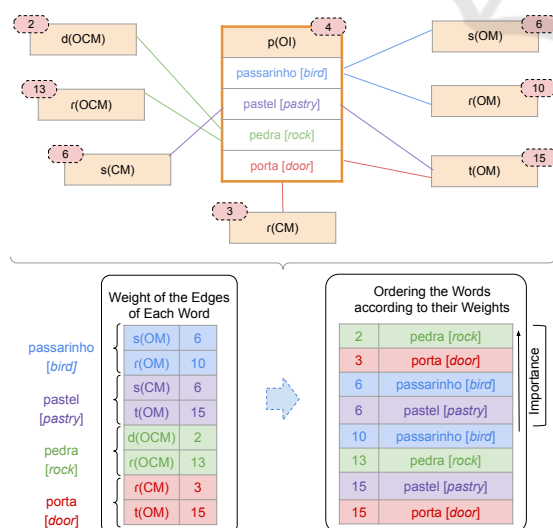


Figure 5: Analysis and ordering of each word of the visited node.

The calculation of the weights of each word is done through the counting of the weight of each node in the remaining graph in which it appears. In Figure 4, the word “passarinho [bird]” appears in the nodes “s(OM)” and “l(OM)”, which have weights of 6 and 10, respectively. With the ordering of the weights, we can evaluate the importance of each word in that node, and it would be enough to keep the first two and exclude the others, while still keeping the graph valid. However, when removing words, it is important to be careful not to exclude any that have a weight lower than the established minimum, as we will impact other nodes, and these may become under-represented. Figure 6 shows the steps followed so far, with the removal of words and subsequent updating of the graph, finishing the first cycle of the algorithm.

To better visualize the impact that the first cycle of exclusion had on the graph, we can check in Figure 7 the updated weight of each node after the removal of words.

However, the algorithm does not have a rule for choosing the next node to be visited, being chosen by order of insertion into the graph’s data structure. In each visited node, words can be removed and the graph must be updated, making the algorithm possibly reproduce different results depending on the ordering of the graph.

## 4 RESULTS AND VALIDATION

All the stages of the algorithm described in Section 3 were implemented in the *Javascript* language. As input for the method, a set of correct phonetic transcriptions in *JSON* of the 84 words proposed by (Ceron et al., 2020) was used. After all the data in the set were analyzed by the algorithm, a reduction of 29 words was obtained, resulting in a subset of 55 words, which satisfies the criteria established in Section 3.1.

However, to validate how the reduction of words proposed by this work would behave in a real scenario, data from 1611 phonological evaluations applied to 1357 children from April/2013 to January/2017 were analyzed. Some evaluations were incomplete in the database used, so that some words were not spoken in certain evaluations. For this reason, it was decided to consider only evaluations with 42 or more words spoken by the children, which represent half of the initial set of words, resulting finally in 1587 evaluations as a base for validation.

The PCC-R index (Percent of Consonants Correct-Revised), developed by (Shriberg et al., 1997), was used as a metric to validate the new



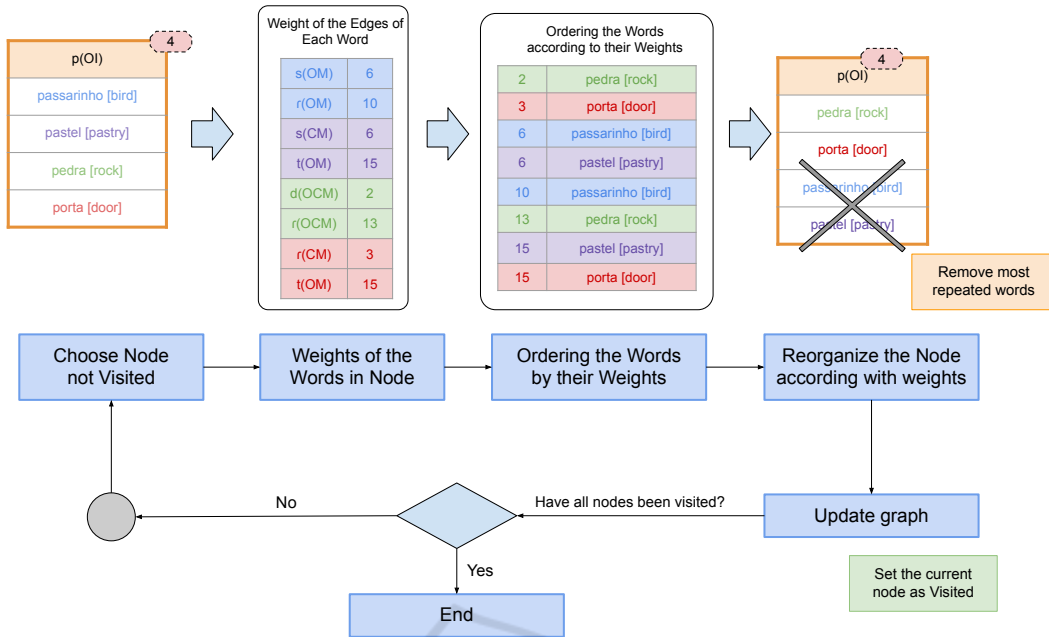


Figure 6: Flowchart of the analysis of each phoneme represented in the graph.

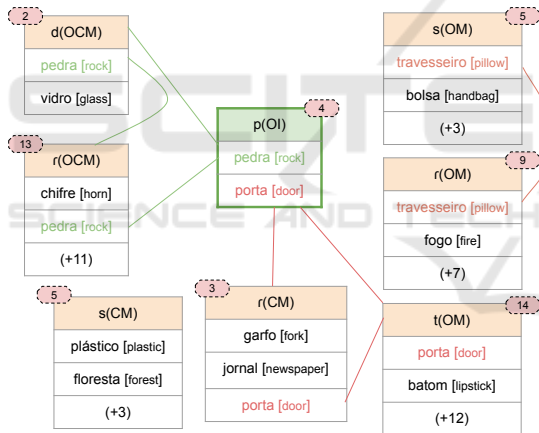


Figure 7: Graph sample after removing two words from the visited node.

dataset. This index is also used as a basis for recommending therapeutic activities in the system proposed by (Franciscatto et al., 2021). The PCC-R value is calculated using Equation 1, based on the number of correct consonants (CC) produced by the child and the total expected productions (TEP). This allows us to determine the severity of the phonological disorder, as shown in Table 2. Additionally, it is directly related to the presence of phonological processes in the child’s speech, as found in the study by (Ghisleni et al., 2010).

$$PCC-R = \frac{CC}{TEP} \times 100 \quad (1)$$

Table 2: Indication of speech disorder according with PCC-R value (Shriberg et al., 1997).

PCC-R Value	Indication of Disorder
Less than 50%	High
Between 51% e 65%	Moderate-High
Between 66% e 85%	Low-Moderate
Greater than 85%	Low

Thus, the PCC-R of each assessment was calculated by first considering all 84 words of the initial set. Then, another PCC-R was calculated in the same way, but this time considering only the 55 words of the new proposed set. The goal was to determine if the degree of disorder associated with the first PCC-R remained in the second case. The result of this calculation can be seen in the graph of Figure 8.

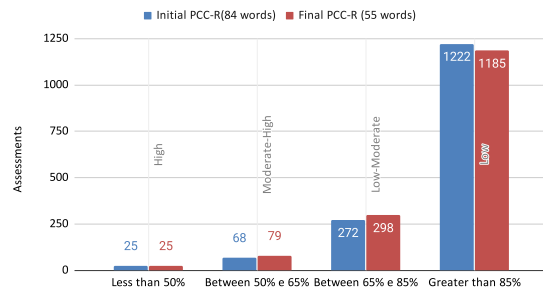


Figure 8: Comparative of PCC-R Classifications.

However, in order to identify if there were changes in the disorder classification of an assessment when only the 55 words of the proposed subset were considered, the two PCC-R results in each assessment were compared, and each change was calculated. After that, we arrived at the graph shown in Figure 9, where it is possible to identify that 96% of the assessments had no change in the Degree of Disorder associated with the PCC-R when the new subset was considered.

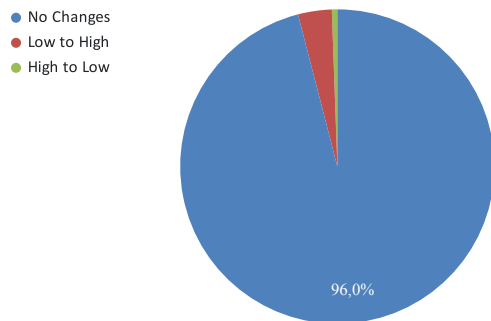


Figure 9: Changes in Indication of Speech Disorder when only the 55 selected words were considered.

Nevertheless, 64 evaluations, which represent 4% of the evaluations analysed, had changes in the classification of the Degree of Disorder. Such changes are mapped in the graph of Figure 10, in which it is possible to identify that the changes are mainly concentrated in lower degrees.

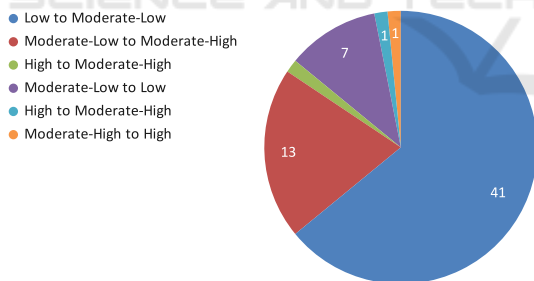


Figure 10: Changes in Classification of 64 assessments.

## 5 CONCLUSION

This work aimed to perform a computational analysis of the 84 words proposed by the work of (Ceron et al., 2020) to compose a phonological assessment tool. The analysis focused on identifying words that generate a super-representation of their phonemes in the set and to perform a critical removal of them. To do this, it was necessary to establish rules for the validation of removals, in order to not generate under-represented phonemes in the set.

The method was described in stages in Section 3, and the algorithm was implemented in *JavaScript* and validated with the use of real data in Section 4. The algorithm was efficient in reducing the number of words used by speech therapists in phonological evaluations of children, also representing a reduction in the spent time applying the evaluations. In addition, the reduction of the words did not bring negative impacts to the quality of the evaluation, since the algorithm was implemented following pre-established criteria in Section 3.1. The validation was done using the PCC-R value of (Shriberg et al., 1997), which indicates the degree of phonological disorder based on the number of correct consonants produced in the evaluation.

The set proposed by (Ceron et al., 2020) was used as input for the algorithm, and at the end of the analysis of the 84 initial words, it was identified that 55 of them would already be sufficient to analyze the same phonemes in the same positions that the initial set evaluated, with at least 2 occurrences for each phoneme in each position, representing a reduction of 35% in the set size. Therefore, this work proposes a method that can be followed manually or implemented in any programming language to analyze the words that compose a phonological assessment tool, with the aim of making it more succinct without losing quality.

## REFERENCES

Bavelas, A. (1948). A mathematical model of group structures. *Human Organization*, 7:16–30.

Ceron, M. I., Gubiani, M. B., Oliveira, C. R. d., and Keske-Soares, M. (2020). Phonological assessment instrument (infono): A pilot study. *CoDAS*, 32(4).

Franciscatto, M. H., Del Fabro, M. D., Damasceno Lima, J. C., Trois, C., Moro, A., Maran, V., and Keske-Soares, M. (2021). Towards a speech therapy support system based on phonological processes early detection. *Computer Speech & Language*, 65:101130.

Franciscatto, M. H., Lima, J. a. C. D., Trois, C., Maran, V., Soares, M. K., and Rocha, C. C. d. (2019). Applying situation-awareness for recommending phonological processes in the children’s speech. SAC ’19, page 739–746, New York, NY, USA. Association for Computing Machinery.

Franciscatto., M. H., Lima., J. C. D., Moro., A., Maran., V., Augustin., I., Keske Soares., M., and Cortez da Rocha., C. (2018). A case-based system architecture based on situation-awareness for speech therapy. In *Proceedings of the 20th International Conference on Enterprise Information Systems - Volume 1: ICEIS*, pages 461–468. INSTICC, SciTePress.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239.

- Ghisleni, M. R. L., Keske-Soares, M., and Mezzomo, C. L. (2010). The use of repair strategies considering the severity of the evolutionary phonological disorder. *Revista CEFAC: Atualização Científica em Fonoaudiologia e Educação*, 12(5):766–771.
- Gomes, A. M., Feltin, T. D., Savoldi, A., and Keske-Soares, M. (2013). Phonological assessments: analysis of the most frequent words considering the vocabulary and sound class. *Distúrbios da Comunicação*, 25(1).
- Goulart, B. N. G. d. and Chiari, B. M. (2014). Speech disorders and grade retention in elementary. *Revista CEFAC: Atualização Científica em Fonoaudiologia e Educação*, 16(3).
- Husain, W. and Pheng, L. T. (2010). The development of personalized wellness therapy recommender system using hybrid case-based reasoning. In *2010 2nd International Conference on Computer Technology and Development*, pages 85–89.
- Iliya, S. and Neri, F. (2016). Towards artificial speech therapy: A neural system for impaired speech segmentation. *International Journal of Neural Systems*, 26(06):1650023. PMID: 27354188.
- Marques, J. V. B. (2022). Módulo de análise contrastiva para avaliações fonológicas de e-fono.
- Pagliariin, K. C. (2009). A abordagem contrastiva na terapia fonológica em diferentes gravidades do desvio fonológico. Mestrado em distúrbios da comunicação humana, Universidade Federal de Santa Maria, Santa Maria.
- Savoldi, A. (2012). Instrumento de avaliação fonológica: validação de conteúdo. Mestrado em distúrbios da comunicação humana, Universidade Federal de Santa Maria, Santa Maria.
- Savoldi, A., Ceron, M. I., and Keske-Soares, M. (2013). What are the best words to compose an evaluation phonological instrument? *Audiology-Communication Research*, 18(3):194–202.
- Shriberg, L. D. et al. (1997). The speech disorders classification system (sdcs). *Journal of Speech, Language and Hearing Research*, 40(4):723–740.
- Stoel-Gammon, C. (1985). Phonetic inventories, 15-24 months: a longitudinal study. *Journal of Speech and Hearing Research*, 28:505–512.
- Tavana, M., Nazari-Shirkouhi, S., Mashayekhi, A., and Mousakhani, S. (2022). An integrated data mining framework for organizational resilience assessment and quality management optimization in trauma centers. *SN Operations Research Forum*, 3:17.
- Yavas, M., Hernandorena, C., and Lamprecht, R. (2001). *Avaliação fonológica da criança: reeducação e terapia*. Biblioteca Artmed. Fonoaudiologia. Artmed Editora.



