An evolutionary approach for the automatic generation of word list fluency assessment items

Rômulo C. de Mello[1,2], Gustavo Silva[1], Patrick C. de Carvalho[1], Rafaela Lopes[1], Maria Clara C. Carneiro[1], Jairo F. de Souza[1,3]

[1]LApIC Research Group – Universidade Federal de Juiz de Fora (UFJF)
Juiz de Fora – MG – Brazil

[2]Fundação CAEd – Juiz de Fora – MG – Brasil

[3]Departamento de Ciência da Computação – UFJF
Juiz de Fora – MG – Brazil

{romulo.c.mello, gustavosjn2013, patrickcarvalho448,

almeidarafaelalopes,mcosta1807}@gmail.com, jairo.souza@ufjf.br

Abstract. *This paper presents a Genetic Algorithm (GA) to automate the generation of reading fluency assessment items, reducing manual effort while meeting pedagogical constraints. Candidate solutions are sequences of words optimized by a multi-objective function that penalizes constraint violations and repetitions. Constraints include canonicity, syllabic variety, grapheme presence, and prosodic continuity. Experiments show that the GA effectively produces valid word lists, with larger populations yielding faster and more stable convergence. A 5% mutation rate was sufficient to preserve diversity. The method is flexible, scalable, and aligned with educational standards.*

Resumo. *Este artigo apresenta um Algoritmo Genético (AG) para automatizar a geração de itens de avaliação de fluência em leitura, reduzindo o esforço manual e atendendo a restrições pedagógicas. As soluções candidatas são sequências de palavras otimizadas por uma função multiobjetivo que penaliza violações de restrições e repetições. As restrições envolvem canonicidade, variedade silábica, ocorrência de grafemas específicos e continuidade prosódica. Os experimentos mostram que o AG gera listas válidas, com populações maiores alcançando convergência mais rápida e estável. Uma taxa de mutação de 5% foi suficiente para manter a diversidade. A abordagem é flexível, escalável e alinhada aos padrões educacionais.*

1. Introduction

In the educational context, the use of learning assessment tools is essential to support the planning of pedagogical interventions aimed at overcoming students' knowledge gaps. In this scenario, assessment items are a fundamental tool for analyzing students' knowledge, as they allow for the objective measurement of the educational aspects under evaluation. However, the development of assessment items is a highly complex and demanding process, as it requires methodological rigor, mastery of content, and alignment with pedagogical approaches and the specific objectives of each educational proposal [Scully 2017; Tiemeier et al. 2011]. This complexity arises both from the need to

DOI: 10.5753/sbie.2025.12314

ensure the validity and reliability of the instruments and from the specificities inherent to each application context.

With recent technological advances, solutions focused on the automatic generation of assessment items have become increasingly common. This field of research, known as Automatic Item Generation (AIG), aims to investigate and develop methodologies that enable the systematic creation of items with specific structures, suitable for particular types of assessment tasks [Circi et al. 2023]. In parallel, there has also been significant progress in the integration of technology into literacy practices, expanding the potential for applying AIG and other tecnologies in educational contexts focused on reading and writing development [Silva and Franco 2023; Da Silva and Franco 2022].

In this context, reading fluency items, which measure the ability to read a text with accuracy, speed, and appropriate prosody, are essential for assessing reading development, but are non-trivial to create. According to Kuhn et al. [2010] and Rasinski [2012], reading fluency goes beyond correct decoding; it involves the ability to read automatically, with natural rhythm, proper intonation, and simultaneous comprehension of the content. In this way, fluency assessment provides not only indicators of automatic processing of written language but also evidence of the level of reading comprehension. These items consist of word lists manually selected based on specific pedagogical criteria, designed to reflect linguistic structures aligned with the requirements of each level or stage of reading fluency.

The selection of words to compose this type of item can be seen as a combinatorial problem: the larger the size of the word list, the more difficult it is for the expert to find words that meet all the item's requirements. This difficulty is mainly due to the types of requirements: minimum frequency requirements (are there at least X words with a certain characteristic?) and maximum frequency requirements are generally used. However, a word is associated with different characteristics (such as gender, size, tonicity, grammatical class, familiarity, etc.) and the choice of a word to meet one of the requirements can influence the fulfillment of the other requirements.

In light of this, we propose a Genetic Algorithm-based (GA) approach for the automated generation of reading fluency items. This algorithm primarily aims to reduce the time and effort typically required to produce such items, while ensuring compliance with linguistic and pedagogical constraints. The proposed method models the specification of word lists used in large-scale Brazilian fluency assessments, seeking to streamline a process that is traditionally detailed and highly dependent on careful selection.

Rather than replacing established practices, this approach is intended to support and accelerate the initial stages of item creation, allowing experts to focus their attention on refinement and quality assurance. In doing so, it preserves the essential role of human judgment in validating the linguistic diversity, phonetic adequacy, and instructional value of the final word lists, ensuring that the generated material aligns with the intended educational goals.

2. Fluency Items

Several factors influence lexical processing associated with decoding during the reading process. Among the main aspects affecting the complexity of reading a word are: syllabic

complexity, word length, its frequency of occurrence in the language, and the reader's degree of familiarity with the term.

Considering these main complexity modifiers, one of the first distinctions in constructing a fluency item is based on syllabic complexity, that is, the phonological structure of the syllables composing the word. This feature is related to syllabic canonicity—the extent to which a word follows typical phonotactic patterns of the language, such as the consonant-vowel (CV) structure. Canonical words present more common and predictable syllabic structures, whereas non-canonical words include less frequent arrangements, such as consonant clusters or vowel sequences. The greater the deviation from a canonical syllabic pattern, the higher the word's complexity in the reading process [Coscarelli 2002].

Secondly, beyond syllabic complexity, it is important to analyze word length. For this, when constructing such items, words must be arranged across multiple possible lengths, based on the number of syllables (monosyllabic, disyllabic, trisyllabic, and polysyllabic). Another relevant factor in building items of this nature is the reader's familiarity with the words present in the lexicon. This familiarity is assessed according to the word's frequency in the language—the more often a reader encounters a given word, the more familiar it becomes; the reader's background knowledge—the more immersed one is in a specific context, the more exposure they will have to the vocabulary of that domain; and other aspects—words commonly found in educational materials, children's stories, or school readings tend to be more familiar [Edwards et al. 2004].

Moreover, when developing word lists for different literacy levels, it is essential to consider the multiple correspondences between graphemes and phonemes. Based on studies that support this type of assessment [Lemle 1985], it becomes necessary to incorporate, in the development of assessment items, recurring aspects of Portuguese orthography and phonology. These aspects include: variations of the phonemes [s] and [z], alternations between [s] and [r], auditory competition between words spelled with <g> and <j>, as well as confusions between the graphemes <u> and <l> in final word position. It is also important to consider the lack of phonetic value of the grapheme <h> at the beginning of words, variations involving the digraphs <lh>, <ch>, <nh>, and the broad phonological overlap among the graphemes <c>, <s>, and <ç> in representing the phoneme [s].

3. Related Works

The use of automatic item generation methods has been increasingly proposed as technological advancements take place. In this context, there are three main models for automatic item generation: template-based models, cognitive models, and deep neural network models.

First applied in 1976 by Wolfe [1976], the template model operates through a standardized and structured sequential framework, where parts of each sentence can be replaced or reordered to produce new questions that follow a very specific mold. Among the main advantages of models of this nature are structural standardization, consistency among the generated items, and low computational cost. In addition, the simplicity of these models facilitates more objective validation processes. As discussed by Willert and Thiemann [2024], template-based models offer significant advantages in the automatic

generation of multiple-choice questions, especially in the field of mathematics. However, the main structural limitation of these models is their rigidity, which leads to various issues, such as low linguistic variability. This limitation can undermine educational processes related to the textual interpretation of questions, as well as negatively affect the analytical design of pedagogical assessments themselves, since such rigidity hampers the effective evaluation of analytical reasoning.

Subsequently, starting in the 1990s, cognitive models were introduced. These methods improve upon the template model by offering greater diversity in question formulation, using schemas developed by experts to represent the reasoning required to solve each specific task. Cognitive models enable the generation of items with greater variety and conceptual validity, as they are based on a theoretical understanding of the mental processes involved in solving each task. Unlike template-based models, which are limited to substituting variables within fixed structures, cognitive models allow for the creation of items that preserve the logic and complexity of the underlying reasoning, even when applied to different contexts or domains. This approach offers greater generalization capacity, enabling the development of items that are cognitively equivalent but vary in format and wording, which contributes to more robust, adaptive assessments that are less susceptible to rote memorization [Pugh et al. 2016].

With the rapid advancement of neural networks and the growing availability of large-scale datasets, deep learning has become a viable solution for tackling complex tasks in Natural Language Processing (NLP), including applications in Artificial Intelligence in Education (AIEd). For instance, Barlybayev and Matkarimov [2024] present a machine learning approach based on transformer architectures for the automatic generation of multiple-choice questions (MCQs). In this study, the authors propose a comprehensive system that frames question generation as a sequence-to-sequence (Seq2Seq) learning task. The system leverages Google's T5 model, fine-tuned on a diverse range of question-answering datasets such as SQuAD, RACE, MS MARCO, among others. These neural network-based methods are particularly effective in tasks requiring deep textual comprehension. Unlike rule-based systems, which follow explicitly defined instructions, neural models learn patterns directly from data. This makes them well-suited for generating discursive or complex MCQs that require nuanced understanding and analysis of source texts.

However, some problems are characterized by their combinatorial nature. In such cases, evolutionary algorithms constitute a suitable approach, as they allow the exploration of large search spaces. Algorithms of this kind have long been employed to optimize educational processes [Andrade et al. 2024]. Furthermore, the potential of these algorithms for generating educational items has already been empirically demonstrated by Popescu [2025], who proposed an enhanced genetic algorithm for the automated generation of educational assessment tests, based on criteria such as population diversity and compliance with pedagogical constraints. In the study, the author presented a methodology based on varying the initial population through the combination of subpopulations ordered by fitness value, demonstrating that such an approach improves both the quality of the solutions and the diversity of the populations generated. The experimental results highlighted the superiority of the proposed algorithm over traditional genetic implementations, particularly in large-scale tasks with multiple constraints, validating its

effectiveness in the context of AIG.

4. Materials and methods

4.1. Problem definition

Let:

- $K = \{w_1, w_2, \ldots, w_m\}$: the set of available words.
- $W \subseteq K$: the subset of selected words.
- $\mathcal{U}$: the set of constraints to be satisfied.

We define $W = \{w_1, w_2, \ldots, w_n\}$ as a valid solution if:

$$\text{(R1) } \#\text{Canonical}(W) = c_{\text{total}} \tag{1}$$
$$\text{(R2) } \forall i \in \{1, \ldots, c_n\}, \text{canonical}(w_i) = \text{True} \tag{2}$$
$$\text{(R3) } \#\text{Monosyllables}(W) = m \tag{3}$$
$$\text{(R4) } \#\text{Disyllables}(W) = d \tag{4}$$
$$\text{(R5) } \#\text{Trisyllables}(W) = t \tag{5}$$
$$\text{(R6) } \#\text{Polysyllables}(W) = p \tag{6}$$
$$\text{(R7) } \forall i \in \{1, \ldots, n-1\}, \text{final}(w_i) \neq \text{initial}(w_{i+1}) \tag{7}$$

Where:

- canonical(w) returns true if the word w is considered canonical.
- final(w) returns the last syllable of w.
- initial(w) returns the first syllable of w.
- c_{total} is the desired total number of canonical words.
- c_n is the number of initial words that must be canonical.
- m, d, t, p are the desired numbers of mono-, di-, tri-, and polysyllabic words, respectively.

The goal of the algorithm is to find a subset $W \subseteq K$ that fully satisfies all constraints $\mathcal{U} = \{R1, R2, \ldots, R7\}$.

4.2. Words List

The genetic algorithm must produce a word list according to the difficulty level specified for the test. In this case, each stage of the pedagogical literacy process is characterized by an expected level of fluency, which is mainly influenced by the school grade the notebook aims to assess. Tables 1 and 2 present examples of word list specifications used in fluency assessments for the 1st and 2th grades of elementary school, respectively. The increase in the complexity of these specifications reflects the expected fluency development between school years. In this case, for first-grade elementary school students, as shown in Table 1, the lists are composed of 60 words: 6 monosyllabic, 24 disyllabic, 24 trisyllabic, and 6 polysyllabic, containing only variations of <s>, <z>, and <x>. For second-grade students, however, although the item structure is similar, variations of <g> and <j> are also added, since the correct interpretation of words containing these variants is expected at this educational stage.

Table 1. Specification of a word list commonly applied in a fluency assessment for the 1st year of elementary school

Word List (60 words)
- 6 monosyllabic - 24 disyllabic - 24 trisyllabic - 6 polysyllabic - The first 20 words must be canonical.
Expected Correspondences
- Variations of /s/, /z/, /x/
Expected Accentuations
- 40 paroxytones; - 20 mixing oxytones and proparoxytones.

Table 2. Specification of a 60-word list for fluency assessment in the 2st year of elementary school

Word List (60 words)
- 6 monosyllabic - 24 disyllabic - 24 trisyllabic - 6 polysyllabic - The first 20 words must be canonical.
Expected Correspondences
- Variations of /s/, /z/; - Sound competition between /g/ and /j/.
Expected Accentuations
- 40 paroxytones; - 20 mixing oxytones and proparoxytones.

4.3. Lexicon

Since the implemented algorithm needs to generate a list of words based on their associated features, it is essential to have a sufficient number of words, along with those features already classified in advance. To achieve this, we use a Portuguese lexicon from a publicly available online dictionary called *Dicionário Fonético da Lingua Portuguesa* (Phonetic Dictionary of Portuguese). These lexicons often included certain attributes that are computationally difficult to determine, such as the number of syllables, word stress (oxytone, paroxytone, or proparoxytone), and phonetic transcription. These features are especially important for identifying the presence or absence of specific phonemes, as discussed in Section 2. Once these sets of words and their characteristics were collected, additional relevant data for the algorithm were extracted using simple computational procedures, such as determining whether a word is canonical or identifying the presence of certain phonemes from its phonetic transcription. Finally, computational processes were used to remove stop-words through libraries such as Spacy, in addition to some manual revisions carried out by linguistics experts to eliminate words that are not appropriate for the pedagogical context, such as swear words, outdated terms no longer used in modern

Portuguese, and slang that would only be suitable in very specific contexts. In total, the lexicon comprises 62,469 words deemed appropriate for educational purposes. Table 3 shows the distribution of information present in the dataset.

Table 3. Distribution of words by lexical category, number of syllables, and frequency of graphemes representative of phonemes

Lexical Category	
Canonical words	5,233
Non-canonical words	57,236
Number of Syllables	
Monosyllabic	526
Disyllabic	7,103
Trisyllabic	16,150
Polysyllabic	38,690
Phoneme-representative Graphemes	
cc	3,446
h	937
s	24,899
z	9,571
r	29,814
rr	6,373
x	1,716
g	6,405
j	4,201
k	19,076
u	9,376
l	5,946
nh	1,321
lh	1,416
ch	1,219

However, as shown in subsection 4.2, a relevant piece of information that is absent from the phonetic dictionaries used is the expected frequency of occurrence of words in the Portuguese language, that is how common these words are in actual language use. To address this gap, a Portuguese-language dataset from WikiData was employed, consisting of words extracted from various contexts such as magazines, journalistic texts, and social media, with the aim of representing different registers and broadening linguistic coverage. Based on this dataset, the relative frequency of each word present in the lexicon was calculated, considering its occurrence in the WikiData corpus. Subsequently, the data were stratified into sextiles, with each range representing a distinct interval of relative frequency within the analyzed corpus.

4.4. Solution

We employ a genetic algorithm [Mirjalili 2019] to assemble N fluency-assessment word lists. Each candidate is a vector of words drawn from a pedagogically curated corpus. First, we define the objective function as Equation 8, where $\{B_j\}$ is a set of word lists

($j = 1, \ldots, 5$), ω_k is the weight of a constraint k, $\delta_k(B_j)$ is a penalty function. 0 if constraint k is satisfied in B_j, otherwise a penalty). In addition, let λ be the weight for cross-wordlist repetitions, and $\delta_{\text{rep}}(B_j)$ – 1 if B_j is a binary function where $\delta_{\text{rep}}(B_j)$ is equal to 1 if B_j repeats any word in another word list and is equal to 0 otherwise.

$$\max_{\{B_j\}} F = -\sum_{j=1}^{N} \Big(\sum_{k=1}^{6} \omega_k \, \delta_k(B_j) + \lambda \, \delta_{\text{rep}}(B_j) \Big). \tag{8}$$

Taken together, the six constraints in Table 4 steer the GA toward word lists that are both non-redundant and pedagogically balanced: no word is recycled across word lists, each set mixes canonical and non-canonical forms in the required proportions, and the internal syllabic profile ranges smoothly from mono- to polysyllables—with an easy disyllabic opener and a more demanding polysyllabic closer—to scaffold reading fluency. Graphemic rules inject specific orthographic targets (guaranteeing `s`/`z`, barring `x`/`j`, and limiting `g`, for example), while the first N words are forced to be canonical so learners begin with predictable patterns. Finally, by forbidding identical end–start syllables in sequence, the model avoids tongue-twisting repetitions and keeps the reading flow natural.

Table 4. Mathematical constraints enforced on each word list B_j. $\delta(\cdot)$ and $\gamma(\cdot)$ are indicator functions; $n_{\text{syl}}(w)$ counts the syllables of word w; σ_{start} and σ_{end} give the first and last syllables.

Code	Purpose	Formula
C1	No word repetition across word lists	$B_j \cap B_k = \emptyset,\ \forall\, j \neq k$
C2	Balance canonical / non-canonical	$\sum_i \delta(\text{Can}(b_{ji})) = C_j,\ \sum_i \delta(\text{NonCan}(b_{ji})) = N_j$
C3[1]	Syllable profile and position rules	$\sum_i \delta\big(n_{\text{syl}}(b_{ji}) = s\big) = S_{s,j}\ (s \in \{1,2,3,\geq 4\});\ \delta\big(n_{\text{syl}}(b_{j1}) = 2\big) = 1;\ \delta\big(n_{\text{syl}}(b_{j20}) \geq 4\big) = 1$
C4	Grapheme control	$\sum_i [\delta('s' \in b_{ji}) + \delta('z' \in b_{ji})] \geq 1;\ \sum_i [\delta('x' \in b_{ji}) + \delta('j' \in b_{ji})] = 0;\ \sum_i \delta('g' \in b_{ji}) \leq 2$
C5	First N words canonical	$\sum_{i=1}^{N} \gamma(b_{ji}) = N$
C6	No identical end/start syllables	$\sigma_{\text{end}}(b_{ji}) \neq \sigma_{\text{start}}(b_{j,i+1}),\ \forall\, i < L_j$

5. Experiments and Results

To validate the proposed solution, experiments were conducted by varying the population size between 50 and 5000 individuals. In each scenario, the algorithm was executed for 1000 generations. The experiments were carried out on a machine with a 13th generation Intel Core i7 processor and 32 GB of RAM. For each population configuration, 10 independent runs were performed, aiming to generate 10 word notebooks that meet the pedagogical constraints defined for the early years of elementary school (1st and 2nd

[1] $S_{s,j}$ is the required count of mono- (s=1), di- (2), tri- (3) and poly- (≥ 4) syllabic words in word list j.

grades). The specific constraints for each grade are detailed in Tables 1 and 2. The algorithm also implements an early stopping mechanism when the global optimum is reached, i.e., when all constraints are correctly satisfied.

Additionally, Figure 2 shows that, besides favoring solution quality, increasing the population size significantly boosts the probability of reaching the global optimum. For populations of 5000 individuals, all runs successfully satisfied all the established constraints.

In the case of configurations aimed at the 2nd grade of elementary school, the experiments indicate that the increased complexity of the constraints negatively impacts the algorithm's performance, as expected. As illustrated in Figures 3 and 4, the greater number of constraints results in a decrease in the convergence speed to the global optimum.

Execution times vary mainly according to the population size. Larger populations, as expected, result in longer execution times. However, they offer greater assurance of convergence to the global optimum, as well as generally higher convergence rates. For populations of 5000 individuals, the standard execution time was around 72 seconds.

Finally, Table 5 presents an example of a word list generated by the proposed genetic algorithm. This list satisfies all the pedagogical constraints established for the 2nd grade of elementary school.

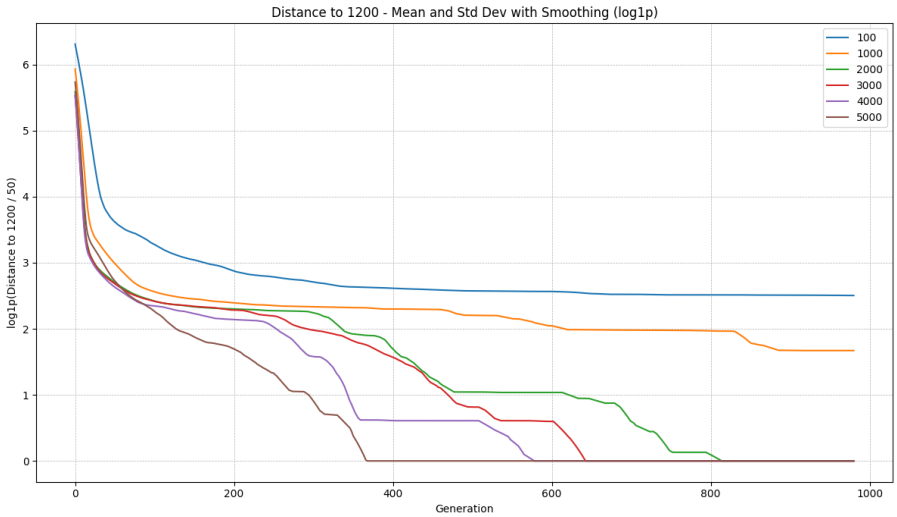

Figure 1. Convergence rate to the global optimum on a logarithmic scale, with populations of 50, 100, 250, 500, 1000, and 5000 individuals, for a 1st grade word list.

Although the generated lists meet the requirements found in manuals for large-scale reading fluency assessments, such as those adopted in the PARC evaluation, a qualitative analysis of the lists revealed that these item composition rules are still insufficient to replace the work of a specialist in item creation. The results showed the presence of vocabulary items in the booklet that are not appropriate for the material's intended purpose, for example. The apparent inadequacies are of a semantic nature, morphological complexity, and frequency of use. Words such as <mate>, for instance, carry an ambiguous meaning or connotations that are inappropriate for the target audience, which compro-

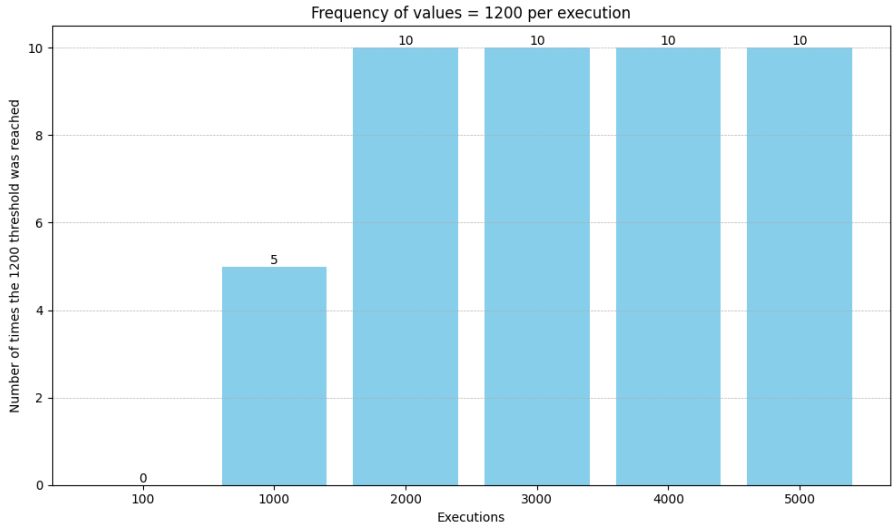

Figure 2. Number of times the global optimum was reached in 10 executions for each population size, for a 1st grade word list

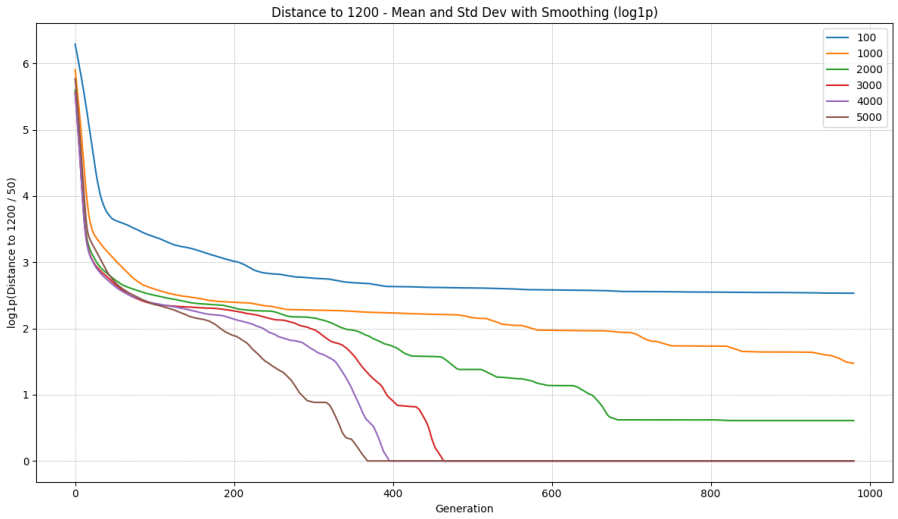

Figure 3. Convergence rate to the global optimum on a logarithmic scale, with populations of 50, 100, 250, 500, 1000, and 5000 individuals, for a 2nd grade word list

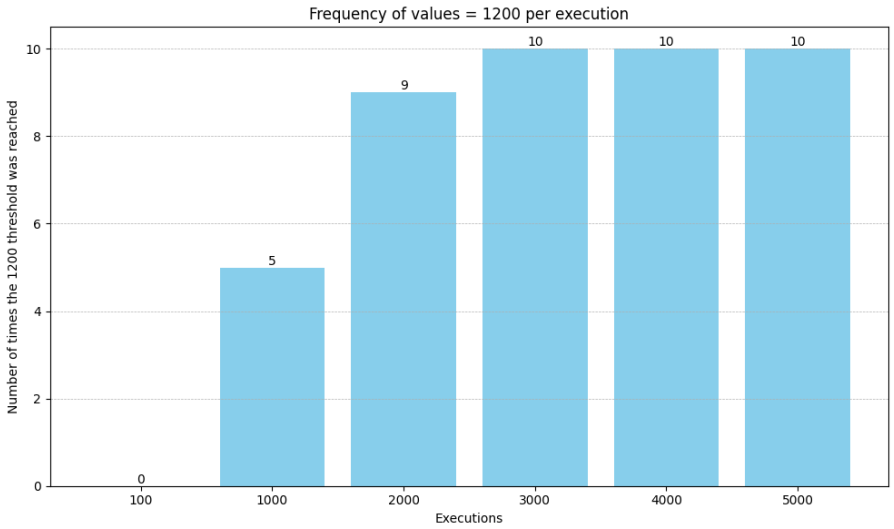

Figure 4. Number of times the global optimum was reached in 10 executions for each population size, for a 2nd grade word list

1. maçã	**2.** reno	**3.** pureza	**4.** maré	**5.** mate
6. café	**7.** mesa	**8.** batida	**9.** peru	**10.** loja
11. jogada	**12.** sabino	**13.** lo	**14.** rodapé	**15.** mi
16. subida	**17.** caneco	**18.** finito	**19.** luneta	**20.** serenidade
21. moça	**22.** má-fé	**23.** cade	**24.** imortal	**25.** tabu
26. reativação	**27.** quintal	**28.** musical	**29.** pegado	**30.** transformar
31. corporal	**32.** sofisticação	**33.** cipó	**34.** fofa	**35.** latitude
36. cidadã	**37.** adicionar	**38.** parcial	**39.** variar	**40.** robô
41. pessoal	**42.** combinar	**43.** nave	**44.** tá	**45.** pego
46. si	**47.** devedor	**48.** baixar	**49.** joga	**50.** nascer
51. crachá	**52.** esgotar	**53.** rali	**54.** popularização	**55.** cozido
56. encurtar	**57.** lá	**58.** obstruir	**59.** lã	**60.** você

Table 5. Example of words list generated by the algorithm

mises their relevance in the intended pedagogical context. Additionally, morphologically complex terms were identified, whose structure may hinder decoding and comprehension by students in the early stages of literacy. Words such as <obstruir> require a level of linguistic mastery not yet consolidated at this stage, making their inclusion inadvisable.

Another point observed concerns the frequency of use of certain words in everyday language. Terms such as <cade>, <reno>, <sabino>, and <rali> do not appear frequently in the language used by the target audience and would therefore be more appropriately placed in items that require the use of low-familiarity words, in order to allow for a more direct evaluation of the reader's decoding ability, without the interference of lexical familiarity in the observed performance.

Finally, we explored LLM-based generation as a potential baseline; however, even with prompt engineering and constrained decoding, these models failed to reliably satisfy our joint constraints (syllabic profiles, grapheme quotas, cross-booklet non-overlap, and

age/frequency appropriateness) without substantial post-hoc filtering. Because this compromises end-to-end reproducibility, we do not report head-to-head metrics against LLMs and instead focus on reproducible algorithmic baselines and GA ablations.

6. Conclusion and Future Work

This paper presented fluency assessment items and their technical specificities related to the literacy process, especially at early educational levels. Additionally, it introduced a mathematical modeling approach to the problem, aligned with the constraints typically applied to linguistic items.

The proposed method is capable of generating word lists that meet the quality criteria required for items used in large-scale reading fluency assessments. While the manual creation of such an item may take many hours of work, the algorithm can produce a valid version of a fluency booklet in a significantly shorter time.

However, the proposal does not entirely replace the role of a linguist. The construction of a fluency booklet involves not only quantitative criteria — which can be modeled computationally with relative ease — but also qualitative aspects that are fundamental to the literacy process. As discussed in Section 5, these aspects require expert judgment. Some of the challenges presented in that section could be mitigated through more rigorous curation of the lexical base used by the algorithm. This would include considering more specific characteristics, such as the morphological complexity of each word, the level of familiarity required for its decoding, and a more refined analysis of word frequency. Such an approach would allow the exclusion of terms that are rarely used in the everyday language of the target audience, thereby improving the relevance and effectiveness of the instructional material.

Another challenge to be addressed in order to improve the automated generation of large-scale assessment booklets lies in the incorporation of qualitative metrics and constraints. One potentially useful metric for avoiding the inclusion of words with inappropriate connotations for the target audience or context is the application of not only morphological but also semantic analysis to the words in the lexical base used in the proposed solution. This enhancement aims to further narrow the gap between automatically generated booklets and those produced by human experts.

Another important consideration for fluency items involves pseudowords, terms that exhibit a morphological structure similar to the language but lack actual meaning. In this study, the generation of pseudoword lists was not addressed, although they share similar constraints with those used here, such as syllabic control and the presence of specific graphemes. However, their creation requires an additional step: the construction of a pseudo-lexicon. This step can be carried out through mutation operations on real words, generating new terms that follow the orthographic rules of Brazilian Portuguese without carrying any recognizable semantic content.

References

[Andrade et al. 2024] Andrade, M. K. T., Silva, V. A., and Ferreira, H. N. M. (2024). Como Formar Grupos em Ambientes Virtuais de Aprendizagem? Uma Abordagem Híbrida Utilizando Algoritmos Genéticos e Algoritmos de Agrupamento. In *Anais*

do XXXV Simpósio Brasileiro de Informática na Educação (SBIE 2024), pages 1971–1983, Brasil. Sociedade Brasileira de Computação - SBC.

[Barlybayev and Matkarimov 2024] Barlybayev, A. and Matkarimov, B. (2024). Development of system for generating questions, answers, distractors using transformers. *International Journal of Electrical and Computer Engineering (IJECE)*, 14(2):1851–1863.

[Circi et al. 2023] Circi, R., Hicks, J., and Sikali, E. (2023). Automatic item generation: foundations and machine learning-based approaches for assessments. *Frontiers in Education*, Volume 8 - 2023.

[Coscarelli 2002] Coscarelli, C. (2002). Entendendo a leitura. *Revista de Estudos da Linguagem*, 10.

[Da Silva and Franco 2022] Da Silva, L. F. and Franco, M. H. I. (2022). Jogos Educacionais Digitais no apoio ao processo de Alfabetização e Letramento: Revisão Sistemática da Literatura. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação (SBIE 2022)*, pages 453–462, Brasil. Sociedade Brasileira de Computação - SBC.

[Edwards et al. 2004] Edwards, J., Beckman, M. E., and Munson, B. (2004). The Interaction Between Vocabulary Size and Phonotactic Probability Effects on Children's Production Accuracy and Fluency in Nonword Repetition. *Journal of Speech, Language, and Hearing Research*, 47(2):421–436.

[Kuhn et al. 2010] Kuhn, M., Schwanenflugel, P., and Meisinger, E. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly*, 45:232–253.

[Lemle 1985] Lemle, M. (1985). *O Guia Teórico do Alfabetizador*. Editora Vozes, Petrópolis, RJ.

[Mirjalili 2019] Mirjalili, S. (2019). *Genetic Algorithm*, pages 43–55. Springer International Publishing, Cham.

[Popescu 2025] Popescu, D.-A. (2025). An Enhanced Genetic Algorithm for Optimized Educational Assessment Test Generation Through Population Variation. *Big Data and Cognitive Computing*, 9(4):98.

[Pugh et al. 2016] Pugh, D., De Champlain, A., Gierl, M., Lai, H., and Touchie, C. (2016). Using cognitive models to develop quality multiple-choice questions. *Medical Teacher*, 38(8):838–843.

[Rasinski 2012] Rasinski, T. V. (2012). Why reading fluency should be hot! *The Reading Teacher*, 65(8):516–522.

[Scully 2017] Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research, and Evaluation*, 22.

[Silva and Franco 2023] Silva, L. F. D. and Franco, M. H. I. (2023). Requisitos para utilização de Jogos Educacionais Digitais na Alfabetização e Letramento sob a perspectiva docente. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação (SBIE 2023)*, pages 718–727, Brasil. Sociedade Brasileira de Computação - SBC.

[Tiemeier et al. 2011] Tiemeier, A., Stacy, Z., and Burke, J. (2011). Using multiple choice questions written at various bloom's taxonomy levels to evaluate student performance across a therapeutics sequence. *INNOVATIONS in pharmacy*, 2.

[Willert and Thiemann 2024] Willert, N. and Thiemann, J. (2024). Template-based generator for single-choice questions. *Technology, Knowledge and Learning*, 29(1):355–370.

[Wolfe 1976] Wolfe, J. H. (1976). Automatic question generation from text - an aid to independent study. *SIGCSE Bull.*, 8(1):104–112.