Improving Text Readability through Segmentation into Rheses

Antoine Jamelot, Solen Quiniou, Sophie Hamon

Improving Text Readability through Segmentation into Rheses

Antoine Jamelot[1], **Solen Quiniou**[2], **Sophie Hamon**[1]
[1]MOBiDYS, Nantes, France, {antoine.jamelot, sophie.hamon}@mobidys.fr
[2]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France,
solen.quiniou@ls2n.fr

Abstract

Enhancing text readability is crucial for readers with challenges like dyslexia. This paper delves into the segmentation of sentences into rheses, i.e. rhythmic and semantic units. Their aim is to clarify sentence structures for improved comprehension, through a harmonious balance between syntactic accuracy, the natural rhythm of reading aloud, and the delineation of meaningful units. This study relates and compares our various attempts to improve a pre-existing rhesis segmentation tool, which is based on the selection of candidate segmentations. We also release TeRheSe (Texts with Rhesis Segmentation), a bilingual dataset, segmented into rheses, comprising 12 books from classic literature in French and English. We evaluated our approaches on this dataset, showing the efficiency of a novel approach based on token classification, reaching a F1-score of 90.0% in English (previously 85.3%) and 91.3% in French (previously 88.0%). We also study the potential of leveraging prosodic elements, though its definitive impact remains inconclusive.

Keywords: rhesis, dyslexia, readability, segmentation, BERT, SpaCy

1. Introduction

Mastering written language is an essential skill in modern society, yet a challenging endeavour for many, especially those grappling with reading disorders such as dyslexia. Digital books can provide assistance to these people through adjustable layout and tools allowing to adapt any text to the users' needs and preferences. Among the possible adjustments, Schneps et al. (2013) have shown that short lines of text are beneficial to people with dyslexia as it limits the span of attention required for reading. However, random line splits can sometimes be more confusing than helpful. Let's consider the sentence *"The vase broke after a gust slammed the window"*. When segmented as *"The vase broke / after a gust slammed the window"*, the sentence is more readily understood than if divided into *"The vase broke after a gust / slammed the window"*. Indeed, in the latter case, the word *after* could be mistakenly interpreted as a preposition, which might confuse readers when they encounter the subsequent portion of the sentence.

In this paper, as in our previous works, the segmentation of sentences into short lines is designed with the primary objective of optimizing text readability for challenged readers, on digital devices. Hence, we rely on the concept of rhesis to perform this segmentation. Rheses can be considered as short lines of text that make sentence structures more transparent, taking into account the rhythm, the grammatical structure and even the sense of the sentence. In order to ease the creation of digital book layout based on rheses, a first automatic segmentation system was proposed by Nin et al. (2016), and underwent several iterations until Houbart et al. (2019). Nevertheless, the system continued to produce erroneous rhesis segmentations, necessitating manual corrections, which represent a significant human work investment on a full book. The goal of this paper is to propose three various improvements of the original rhesis segmentation system as well as a gold standard rhesis corpus containing French and English books segmented into rheses.

The rest of this paper is organized as follows. After discussing the concept of rhesis (section 2), we describe the original version of our rhesis segmentation system as well as the the three improved systems (section 3). We then present the rhesis corpus and we evaluate our approaches on it (section 4). We finally compare the concept of rhesis with other text unit concepts (section 6).

2. Definition of Rhesis

The concept of rhesis was defined in two different research fields. In speech therapy, a rhesis is defined as a sequence of words pronounced in a single exhalatory breath (Brin-Henry et al., 2018). In linguistic discourse, it refers to the "rhythmic unit" of a statement, predominantly composed of either a verb or a noun together with their nearest modifiers (Damourette and Pichon, 1930).

Following these definitions and the objective of optimizing legibility, we define a rhesis as a text segment embodying a balance amongst four criteria. First, each rhesis should encapsulate a distinct meaningful unit, potentially evoking a specific emotion or a mental picture. Secondly, it must conform to the inherent syntactic structure of the sentence. Thirdly, it should resonate with the natural cadence

of reading aloud. Lastly, each rhesis should ideally fit within a short line, not exceeding 40 characters in length.

It is also worth noting that multiple correct rhesis segmentations may exist for a given sentence. For example, the sentence *"My adventure began in a port town filled with men of the sea"* could be segmented as: *"My adventure began in a port town / filled with men of the sea"* to follow mental images; but, the rhesis *"My adventure began in a port town"* could in turn be divided as *"My adventure began / in a port town"* to emphasize the syntactic structure.

3. Segmentation into Rheses

In this section, we first present the original version of our system and its extension based on syntax. Then, we present two alternative approaches, based either on prosody or on token classification.

3.1. Original Rhesis Segmentation Approach

Figure 1 outlines the major steps of our initial rhesis segmentation system.

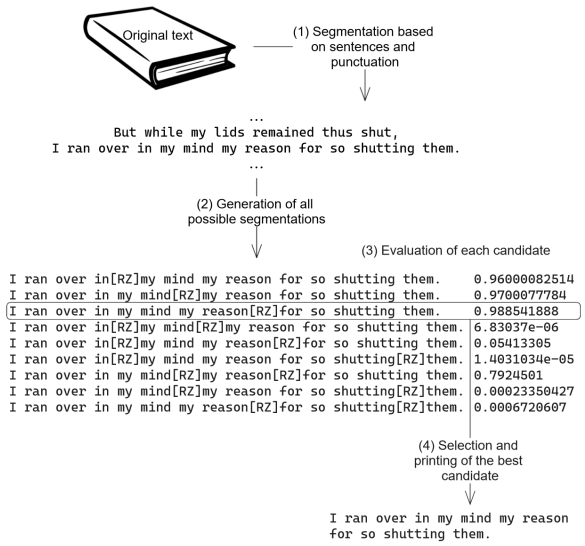

Figure 1: Overview of the original system.

Step 1 aims at segmenting the text into phrases. First, the sentence segmentation model of SpaCy is used, along with some additional rules to avoid errors around abbreviations and quotation marks. Within each sentence, the text is further divided into phrases, based on secondary punctuation marks, such as commas, colons, and parentheses. In order to avoid a too fine-grained segmentation, short phrases are merged following their syntactic relations (if any), detected by SpaCy, while ensuring that their combination does not exceed the defined span of 40 characters. In the example of Figure 1, since the short segment *"But while my lids remained thus shut,"* is bounded by a comma, it is extracted and directly recorded as a rhesis. The next phrase, being longer than 40 characters, must go through the remaining steps.

Step 2 consists in generating all possible segmentations of phrases into rheses, based on spacing characters with two conditions: no text segment must exceed the length limit of 40 characters, and the number of divisions does not exceed, by more than one, the minimum number needed to satisfy the first condition.

At step 3, a BERT-based language model, fine-tuned to recognize correctly rhesis-segmented sentences, assigns a score between 0 and 1 to each proposed segmentation. This step relies entirely on the quality of the model.

Step 4 simply consists in comparing the obtained scores to retain and write the best-segmented proposal, as judged by the model, into the output file.

3.2. Improving Syntax-Based Segmentation Rules

Our first approach targeted the initial phases of the system's operation, by preventing grammatically erroneous breaks, before the system generates possible segmentations and moves on to the phase of selecting the appropriate rheses.

Here, we first adjust the initial segmentation based on punctuation, corresponding to the first step depicted in Figure 1. Noting that conjunctions and relative pronouns – even when succeeded by a comma – relate more closely to the words that follow than to those preceding them, we shifted the division point. Instead of segmenting after the comma as per the general rule, we opted to segment immediately before the conjunction or relative pronoun. For example, in the fragment *"It was the shaft of a spear that, lunged through the opening, [...]"*, the segmentation now occurs before the conjunction *that* rather than at the comma.

Then, we focus our attention on small groups that should not be divided. Leaves in the syntactic tree constructed by SpaCy, which are words that no other word depend on, are generally tightly coupled to their syntactic head, such as a determiner to its noun, or the verb *to be* to its attribute, and so on. When a "leaf" word is found within one or two positions of its dependent word, it's considered to form an indivisible "chunk" with it, encompassing any words between them. While these chunks do not match the definition we will give in section 6, they approximate it by being continuous sequences structured around a "strong head".

From the tree displayed in Figure 2, the extracted "chunks" are: *"I ran over"* and *"my mind"*. This is why, in the diagram from Figure 1, segmentation candidates including cuts within the mentioned

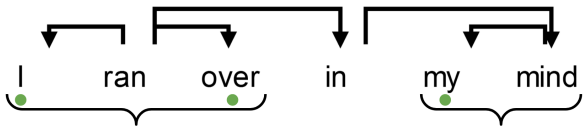

Figure 2: Detection of chunks in a syntactic dependency tree (leaves are marked with a green spot).

groups are not shown in the second step. This mechanism also offers the supplementary advantage of hastening the rhesis process by preventing the model from evaluating numerous unneeded proposals.

Despite these improvements, the original architecture still suffers from several drawbacks. One of the primary concerns is the model's reliance on fine-tuning through both positive and negative samples. Acquiring unambiguously wrong segmentations is especially challenging due to the inherent multiplicity of valid segmentations. On the other side, processing separately the shorter and longer fragments delimited by punctuation marks (step 1), while limiting the exponential number of segmentation candidates to evaluate, hinders some natural groupings and causes the neural language model to evaluate incomplete sentences. To overcome these limitations, we explored two new approaches to rhesis segmentation.

3.3. Segmentation based on prosodic features

We explored the potential of utilizing prosodic features to detect rhesis boundaries. Given the oral nature of rhesis, it was hypothesized that oral readings might offer segmentation insights that are not immediately discernible in written text.

Drawing inspiration from indices used to detect intonational periods as studied by Avanzi et al. (2008), we detected pauses in speech and studied the duration of each one, as well as the amplitude of pitch movements between two pauses (defined as the difference between the last F0 extremum and the mean F0 of an inter-pausal segment), and the pitch jump, which refers to the immediate difference in pitch before and after a pause. We used simple decision trees, trained on these three features, to determine wether each pause corresponds to a rhesis boundary or not.

3.4. Segmentation based on token classification

Finally, we reconsidered the rhesis segmentation as a token classification problem. We trained a token classification model to recognize rheses, following a BIO scheme. Each space-separated word is labeled with *B* (beginning) if it starts a rhesis, and *I* (inside) otherwise. This approach comes with several benefits. First, it does not require negative samples anymore. It also solves the issue of producing a high number of segmentation candidates, and thus allows to process a text file by entire lines.

An overview of this system is illustrated by Figure 3. Each line of the input text is tokenized by the model's tokenizer, and tokens are labeled by the fine-tuned model. A rhesis is then produced for every token (word or subword) preceded by a space and labeled with a *B*. If a rhesis exceeds the desired maximum length (typically 40 characters), it is re-segmented before the word which, according to the model, is the most likely to be a *B* rather than an *I*. If a line exceeds the capacity of the model, which amounts to 512 tokens, the last rhesis produced is ignored, and the line is processed again, beginning on the last token labeled *B*. This operation is repeated as many times as necessary to complete the line segmentation into rheses.

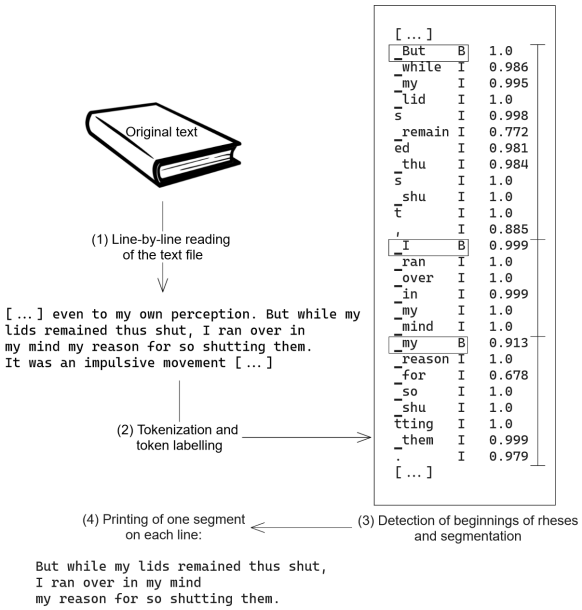

Figure 3: Overview of the system based on token classification.

4. Evaluation

In this section, we first present our corpora designed for rhesis segmentation. Then, we describe the experimental setup as well as the results obtained on our rhesis corpora.

4.1. Rhesis Corpora

In order to train and evaluate models for rhesis segmentation, we collected a corpus of 92 literary books whose texts were segmented into rheses by a previous version of the software and cor-

rected manually. 80 books where used for training, totalizing 348 534 rheses, including 285 544 in French, 40 924 in English, 17 657 in Italian and 1 409 in Spanish. Our evaluation corpus, named TeRheSe (Texts with Rhesis Segmentation), comprises 6 duty-free books in French (21 602 rheses) and 6 in English (51 681 rheses). The average length of rheses was similar among languages, namely, in particular, 24.6 characters in French and 23.4 in English. We release TeRheSe as a novel language resource.

For prosodic analysis, we needed audio versions of the considered books. To adapt the method to a particular narrator's diction, we used the French audio renditions of the first chapters of Jules Verne's *Le Tour du Monde en Quatre-Vingt Jours (Around the World in Eighty Days)* sourced from litteratureaudio.com, and the English renditions of Charlotte Brontë's *Jane Eyre*, sourced from LibriVox. For each book, the first chapter was used for training, and the second one for evaluation.

4.2. Experimental Setup

We fine-tuned the pretrained model XLM-RoBERTa base (Conneau et al., 2020) on the training corpus, on the tasks of segmentation classification (for the approaches presented in subsections 3.1 and 3.2) and token classification (for the approach presented in subsection 3.4). Each model was trained on 1 epoch, with a learning rate of 2×10^{-5}. The batch size was of 32 for the segmentation classifier, 16 for the token classifier.

For the prosodic system (presented in subsection 3.3), a forced alignment between audio files and their corresponding texts was performed with WebMAUS, a tool provided by the University of Munich (Kisler et al., 2017). Pause durations, amplitude of pitch movements and pitch jumps were then extracted algorithmically using the Python library Parselmouth. Then we trained a simple decision tree classifier to predict wether a pause in speech corresponds to a rhesis boundary. The tree depth was limited to 3 to prevent any overfitting. As Figure 4 shows, it turned out that only the pause duration and, to a lesser extent, the pitch movement amplitude, were significant to detect that a pause matches a rhesis boundary. Trained on the French selected extract, the tree followed the same scheme with barely different thresholds: a pause is classified as rhesis boundary if its duration is greater than 0.225 s, or greater than 0.165 s if the pitch movement in the preceding segment is greater than 3.151 semitones.

4.3. Experimental Results

First, we evaluated all the systems, but the prosodic one, by running them on the evaluation corpus

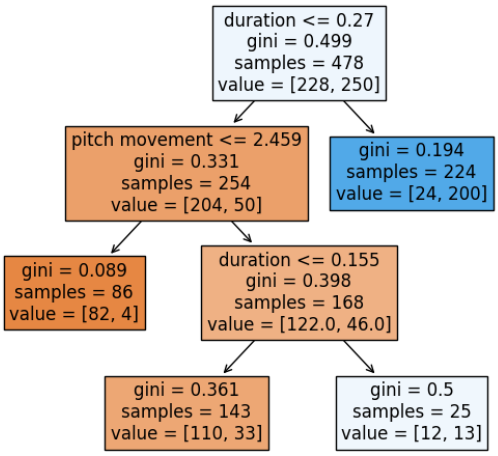

Figure 4: Decision tree trained on the first chapter of *Jane Eyre*. Duplicate leaves were pruned.

TeRheSe. Considering the task as a rhesis boundary detection task, we calculated the precision (P), recall (R), and F1 score (F) of each output. The results are shown in Table 1. Whereas incorporating segmentation rules based on syntax slightly improves the original system, rethinking the rhesis segmentation task as a token classification task greatly improves the results. We can see similar results for French and English: the precision remains quite the same between systems (and across languages) but the recall greatly increases, especially for the English books (by 8.7%).

Language	System	P	R	F
French	Original	88.0	88.2	88.0
	Syntactic	88.2	88.8	88.4
	Token	88.5	94.3	91.3
English	Original	89.6	81.5	85.3
	Syntactic	89.8	81.7	85.4
	Token	90.1	90.2	90.0

Table 1: Average scores on TeRheSe (%). *P*, *R* and *F* stand respectively for precision, recall and F_1 score.

The prosodic approach was evaluated on two extracts for each studied language. First, we ran each decision tree on the second chapter of the same book that was used for training it, read by the same person. Then, to check generalizability, we used the first chapter of another book read by another person of the same language, namely, for English, *The Adventures of Tom Sawyer* by Mark Twain, and, for French, *Au revoir là-haut (The Great Swindle)* by Pierre Lemaître, read by the author himself.

Results, shown in Figure 2, did not encourage us to keep on with this approach, although they reveal

a decent precision in some cases, and similar score whether training and inference be run on texts read by the same reader or not.

Language	Book	P	R	F
French	same	85.8	54.7	66.8
	other	77.4	63.8	69.9
English	same	89.4	54.6	67.8
	other	87.3	57.0	69.0

Table 2: Scores obtained with prosody-based segmentation on *Around the World in Eighty Days* (French) and *Jane Eyre* (English). "Same" means that the same book was used for training and evaluation, "other" that we ran the evaluation on another book of the same language.

To compare automatic and manual annotation, we asked six people, informed about rheses, to segment the French version of *The Oval Portrait*, and compare the resulting segmentations with the "official" one, which was produced beforehands by a professional annotator. The token classifier obtained similar or even better scores than the human annotators. The results are shown in table 3.

Segmentation	P	R	F
Human (average)	85.1	86.3	85.7
Human (best)	86.8	89.9	88.2
Original	84.8	85.7	85.2
Enhanced	85.3	87.1	86.2
Token	84.7	92.7	88.5

Table 3: Scores obtained on the French version of *The Oval Portrait*.

5. Discussion

We compared errors made by our systems to determine the possibility of combining them for enhanced accuracy. It appears, however, that the token classifier, by itself, outperforms other methods on every criterion. Notably, it not only "instinctively" adheres to the indivisible chunks as defined in 3.2, but also benefits from its ability to segment sentences finelier when needed. Additionally, the classifier demonstrates a more nuanced interpretation of ambiguous punctuation, such as abbreviation periods and dashes.

The role of the prosodic way remains uncertain, since we could not find an error made by the token classifier that would have been avoided by an analysis of pauses in narration. However, it is possible that elements subtler or more complex than pauses provide clues for a more accurate segmentation.

6. Related Work

The concept of rhesis can be compared to other concepts of text units. A chunk, as defined by Abney (1991), consists typically of "a single content word surrounded by a constellation of function words, matching a fixed template". Unlike rheses, chunks are purely syntactic units, that do not take into account semantic unity, and they tend to be shorter than rheses. Elementary Discourse Units (EDUs), defined as "the minimal building blocks of a discourse tree" (Carlson et al., 2001), are typically, but not always, syntactic clauses. While closer to the idea of semantic units, EDUs are often too long to constitute rheses. EDU segmentation is typically treated as a token classification task (Braud et al., 2023). Kamaladdini Ezzabady et al. (2021) reached a F1 score of 88.41% for this task on a multilingual corpus. Idea Units (Kroll, 1977) are another attempt to define information units based on syntactic rules. These units are rather close to rheses, but unlike them, they can be discontinuous. Gecchele et al. (2022) proposed a rule-based algorithm to extract idea units from a document parsed with SpaCy, with a F1 score of 81.6%.

7. Conclusion

In this paper, we endeavoured to enhance a system specifically designed for rhesis segmentation, with the aim to optimize text readability for those who face reading challenges. Our key contribution is the development of a new token classification-based method, which showcased superior performance in comparison to the original system, based on sentence classification. This new version reached a near-human performance, notably highlighted by a better respect of syntactic units and a better interpretation of ambiguous punctuation marks.

Our research highlight directions for future exploration. The potential of prosodic clues in rhesis segmentation remains untapped. Utilizing deep learning to craft a model that segments audio based on subtle prosodic elements could be a way to perfect rhesis segmentations. Other future works include exploring the cross-lingual capabilities of our system on languages other than English and French, especially for languages for which no segmented texts are available. Future work could also focus on adapting segmentation to accommodate various styles and specific preferences regarding the balance between the criteria of rhesis: semantic unity, syntactic accuracy, rhythm, and length.

8. Bibliographical References

S.P. Abney. 1991. Parsing by chunks. In *Principle-Based Parsing: Computation and Psycholinguistics*, Studies in Linguistics and Philosophy, pages 258–278. Springer Netherlands.

Mathieu Avanzi, Anne Lacheret-Dujour, and Bernard Victorri. 2008. ANALOR. A Tool for Semi-Automatic Annotation of French Prosodic Structure. In *Proceedings of the 4th International Conference on Speech Prosody, SP 2008*, pages 119–122., Campinas, Brazil.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

Frédérique Brin-Henry, Catherine Courrier, Emmanuelle Lederle, and Véronique Masy. 2018. *Dictionnaire d'Orthophonie*. Ortho-Edition.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacques Damourette and Édouard Pichon. 1930. *Essai de grammaire de la langue française : des mots à la pensée*, chapter IV. J.L.L. d'Artrey.

Marcello Gecchele, Hiroaki Yamada, Takenobu Tokunaga, Yasuyo Sawaki, and Mika Ishizuka. 2022. Automating idea unit segmentation and alignment for assessing reading comprehension via summary protocol analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4663–4673, Marseille, France. European Language Resources Association.

Jean-Claude Houbart, Solen Quiniou, Marion Berthaut, Béatrice Daille, and Claire Salomé. 2019. Automatic segmentation of texts into units of meaning for reading assistance. In *IJCAI workshop on AI and the United Nations SDGs*, Macao, China.

Morteza Kamaladdini Ezzabady, Philippe Muller, and Chloé Braud. 2021. Multi-lingual discourse segmentation and connective identification: MELODI at disrpt2021. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 22–32, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347.

Barbara Kroll. 1977. Combining ideas in written and spoken english: A look at subordination and coordination. *Discourse across time and space*, 5.

Constance Nin, Victor Pineau, Béatrice Daille, and Solen Quiniou. 2016. Segmentation automatique d'un texte en rhèses. In *23e Conférence sur le Traitement Automatique des Langues Naturelles (TALN'2016)*, Paris, France.

Matthew H. Schneps, Jenny M. Thomson, Gerhard Sonnert, Marc Pomplun, Chen Chen, and Amanda Heffner-Wong. 2013. Shorter Lines Facilitate Reading in Those Who Struggle. *PLoS One*, 8(8):e71161.