

Antoine Jamelot, Solen Quiniou, Sophie Hamon

Improving Text Readability through Segmentation into Rheses



LREC-COLING  2024

Optimizing short lines for readability

Let's consider this sentence:

"The vase broke after a gust slammed the window."

We want to display it on short lines, easier to read (Schneps et al., 2013).

But random line splits can sometimes be more confusing than helpful...

Default line wrapping:

The vase broke after a gust slammed the window.



Rhesis segmentation:

The vase broke
after a gust slammed the window.



What is a rhesis?

In our study: **A segment that optimizes readability** balancing four criteria:

- encapsulating a **distinct meaningful unit**, ideally a mental picture,
- conforming to the inherent **syntactic structure** of the sentence,
- resonating with the **natural cadence** of reading aloud,
- fitting within a **short line**, not exceeding 40 characters in length.

For a given sentence, **several segmentations may be correct**, depending on particular perceptions and prioritizations.

The Task: rthesis boundary detection

Input: raw text.

My adventure began in a port town filled with men of the sea. It was there I met Queequeg, a harpooner from the South Seas, and we soon became fine friends.



Output: 1 rthesis per line.

My adventure began in a port town
filled with men of the sea.
It was there I met Queequeg,
a harpooner from the South Seas,
and we soon became fine friends.

The data

	Training dataset		Evaluation set (TeRheSe)	
Language	# books	# rheses	# books	# rheses
French	54	285,544	6	21,602
English	19	40,924	6	51,681
Italian	4	17,657	—	
Spanish	3	1,409		
Total	80	345,534	12	73,283

Original approach



(1) Segmentation based on sentences and punctuation

...

But while my lids remained thus shut,
I ran over in my mind my reason for so shutting them.

...

Original approach

I ran over in my mind my reason for so shutting them.

(2) Generation of all possible segmentations

I ran over in[RZ]my mind my reason for so shutting them.
I ran over in my mind[RZ]my reason for so shutting them.
I ran over in my mind my reason[RZ]for so shutting them.
I ran over in[RZ]my mind[RZ]my reason for so shutting them.
I ran over in[RZ]my mind my reason[RZ]for so shutting them.
I ran over in[RZ]my mind my reason for so shutting[RZ]them.
I ran over in my mind[RZ]my reason[RZ]for so shutting them.
I ran over in my mind[RZ]my reason for so shutting[RZ]them.
I ran over in my mind my reason[RZ]for so shutting[RZ]them.

Original approach

(3) Evaluation of each candidate segmentation by a BERT fine-tuned sentence classifier

I ran over in[RZ]my mind my reason for so shutting them.	0.960
I ran over in my mind[RZ]my reason for so shutting them.	0.970
I ran over in my mind my reason[RZ]for so shutting them.	0.989
I ran over in[RZ]my mind[RZ]my reason for so shutting them.	0.000
I ran over in[RZ]my mind my reason[RZ]for so shutting them.	0.054
I ran over in[RZ]my mind my reason for so shutting[RZ]them.	0.000
I ran over in my mind[RZ]my reason[RZ]for so shutting them.	0.792
I ran over in my mind[RZ]my reason for so shutting[RZ]them.	0.000
I ran over in my mind my reason[RZ]for so shutting[RZ]them.	0.001

Original approach

I ran over in[RZ]my mind my reason for so shutting them.	0.960
I ran over in my mind[RZ]my reason for so shutting them.	0.970
I ran over in my mind my reason[RZ]for so shutting them.	0.989
I ran over in[RZ]my mind[RZ]my reason for so shutting them.	0.000
I ran over in[RZ]my mind my reason[RZ]for so shutting them.	0.054
I ran over in[RZ]my mind my reason for so shutting[RZ]them.	0.000
I ran over in my mind[RZ]my reason[RZ]for so shutting them.	0.792
I ran over in my mind[RZ]my reason for so shutting[RZ]them.	0.000
I ran over in my mind my reason[RZ]for so shutting[RZ]them.	0.001

(4) Selection and printing
of the best candidate

“I ran over in my mind my reason
for so shutting them.”

Original approach: Does it work?

- Model: XLM-RoBERTa-base (Conneau et al., 2020) fine-tuned on our training data set (1 epoch, initial learning rate of 2×10^{-5} , batch size of 32).
- Evaluation on TeRheSe.

The results seem quite good, but in practice many rheses remain to be corrected manually.

Language	Precision (%)	Recall (%)	F1-score (%)
French	88.0	88.2	88.0
English	89.6	81.5	85.3

Improving syntax-based segmentation rules

(1/2) Exception to the punctuation-based segmentation:

When a conjunction or a relative pronoun is followed by a comma, the split occurs before this word rather than after the comma.

Example:

It was the shaft of a spear **that**,
either thrown
or lunged through the opening,
had caught him in the side, [...]

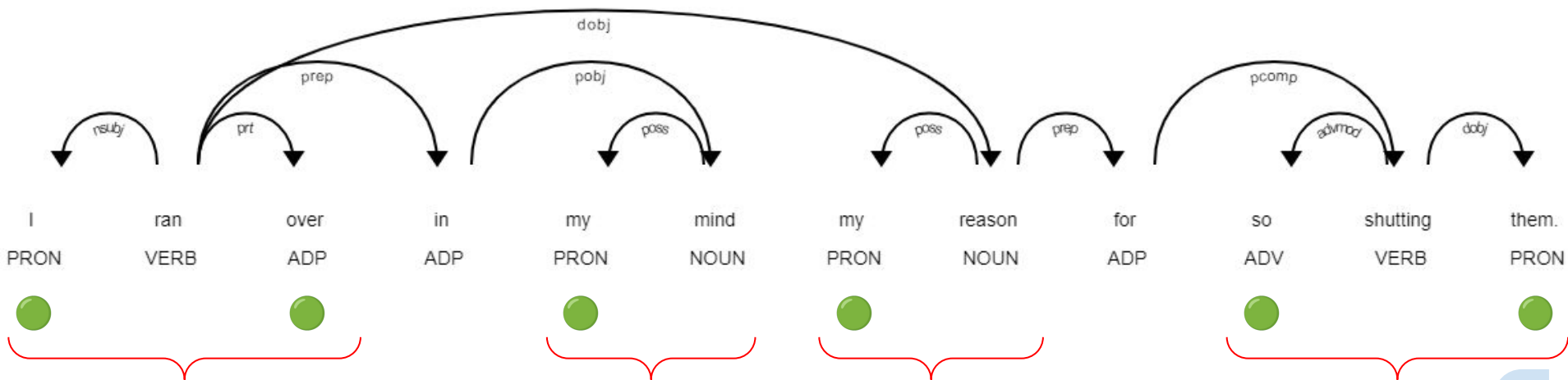


It was the shaft of a spear
that, either thrown
or lunged through the opening,
had caught him in the side, [...]

Improving syntax-based segmentation rules

(2/2) Leaf words (green spots) form indivisible chunks with their head.

The most aberrant splits are thus avoided.



Improving syntax-based segmentation rules: Results

Language	Precision (%)	Recall (%)	F1-score (%)
French	88.2 (+ 0.2)	88.8 (+ 0.6)	88.4 (+ 0.4)
English	89.8 (+ 0.2)	81.7 (+ 0.2)	85.4 (+ 0.1)

Same model and evaluation as before.

Slight but well targeted improvement: the new rules eliminate the most aberrant mistakes that could occur previously.

Process hastened by the reduction of the number of candidate segmentations.

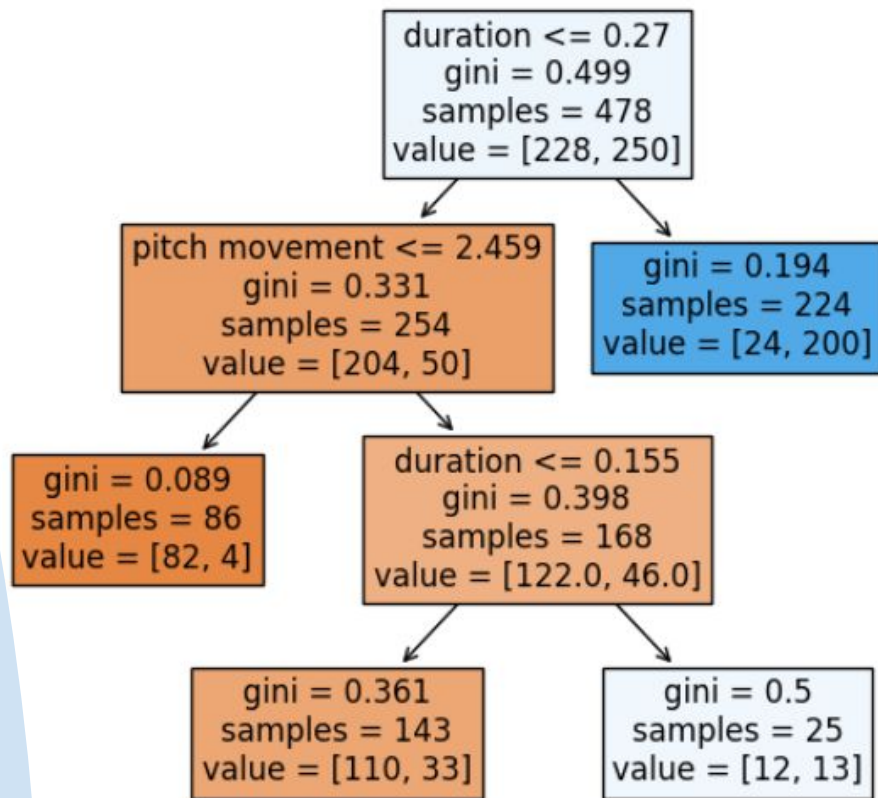
Segmentation based on prosodic features

- Audio-text alignment with WebMAUS (Kisler et al., 2017).
- Rthesis boundary detection based on three prosodic features (Avanzi et al., 2008):
 1. the **duration** of each pause (in seconds);
 2. the **pitch jump** (in semitones),
i.e. the difference of the voice pitch (F0) just before and just after a pause;
 3. the **amplitude of pitch movements** (in semitones),
i.e. the difference between the last F0 extremum and the mean F0 in an inter-pausal segment.

In practice, only pause duration and amplitude of pitch movements were significant to recognize rthesis boundaries.

Segmentation based on prosodic features

- **Decision trees** were trained upon the three features to detect silences that match rthesis boundaries.
- Training data:
 - **French:** the first chapter of *Le Tour du monde en quatre-vingt jours* by Jules Verne
 - **English:** the first chapter of *Jane Eyre* by Charlotte Brontë



Segmentation based on prosodic features: Results

Language	Text	Precision (%)	Recall (%)	F1-score (%)
French	<i>Le Tour du monde en quatre-vingt jours</i> , chapter 2	85.8	54.7	66.8
	<i>Au revoir là-haut</i> , chapter 1	77.4	63.8	69.9
English	<i>Jane Eyre</i> , chapter 2	89.4	54.6	67.8
	<i>The Adventures of Tom Sawyer</i> , chapter 1	87.3	57.0	69.0

Segmentation based on token classification

- New system based on a **token classifier**, applying a **B-I-O scheme**: *B* for a word beginning a phrase, *I* for any other word.
- Simple line-by-line processing, without any preprocessing based on punctuation or syntax.

I	B	0.999
ran	I	1.000
over	I	1.000
in	I	0.999
my	I	1.000
mind	I	1.000
my	B	0.913
reason	I	1.000
for	I	0.678
so	I	1.000
shutting	I	1.000
them.	I	0.999

Segmentation based on token classification:

Results

- Training: XLM-RoBERTa-base token classifier, same hyperparameters as before, but with a batch size of 16.
- Evaluation on TeRheSe.

Significant improvement when compared to the original approach, confirmed by a manual exploration.

Language	Precision (%)	Recall (%)	F1-score (%)
French	88.5 (+ 0.5)	94.3 (+ 6.1)	91.3 (+ 3.3)
English	90.1 (+ 0.5)	90.2 (+ 8.7)	90.0 (+ 4.7)

Human vs Machine

- 6 people, informed about rheses, were asked to segment the French version of *The Oval Portrait* by Edgar Allan Poe
- Comparison against the “official” one.

Our token classifier apparently reaches a near-human performance.

Segmentation	Precision (%)	Recall (%)	F1-score (%)
Human (average)	85.1	86.3	85.7
Human (best)	86.8	89.9	88.2
Token classifier	84.7	92.7	88.5

Conclusion

- We developed a **new rthesis segmentation method** based on token classification, showcasing a significantly superior performance in comparison with the previous system, based on sentence classification.
- We released **TeRheSe**, our **evaluation data set**, comprising 12 copyright-free books, in French and English, meticulously segmented into rheses.
- The **potential of leveraging prosodic elements** remains uncertain, and would require further investigation.

References

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.

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.

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

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.

Our paper is available on HAL:
<https://hal.science/hal-04566523>