Segmentation automatique d'un texte en rhèses

Constance Nin Victor Pineau
Béatrice Daille Solen Quiniou
LINA, Université de Nantes, 2 rue de la Houssinière, BP92208, 44322 Nantes Cedex 3
constance.nin@etu.univ-nantes.fr, victor.pineau@etu.univ-nantes.fr,
beatrice.daille@univ-nantes.fr, solen.quiniou@univ-nantes.fr

RÉSUMÉ
La segmentation d'un texte en rhèses, unités-membres signifiantes de la phrase, permet de fournir des adaptations de celui-ci pour faciliter la lecture aux personnes dyslexiques. Dans cet article, nous proposons une méthode d'identification automatique des rhèses basée sur un apprentissage supervisé à partir d'un corpus que nous avons annoté. Nous comparons celle-ci à l'identification manuelle ainsi qu'à l'utilisation d'outils et de concepts proches, tels que la segmentation d'un texte en chunks.

ABSTRACT
Automatic segmentation of a text into rhesis

The segmentation of a text into rhesis, parts of a sentence that are meaningful by themselves, allows the creation of tools to help dyslexic people to read. In this paper, we offer an automatic method to identify rhesis based on supervised learning on a corpus we annotated. We then compare it to manual identification as well as to similar tools and concepts, such as chunks identification.

MOTS-CLÉS : rhèse, chunk, apprentissage supervisé, dyslexie, guide d'annotation.

KEYWORDS: rhesis, chunk, supervised learning, dyslexia, annotation guide.

1 Introduction

La démocratisation des livres sur supports numériques permet d'envisager de nouvelles méthodes d'assistance pour les personnes atteintes de troubles de la lecture. Ces difficultés interviennent dans l'apprentissage de la lecture et de l'écriture sans qu'aucun désordre sensoriel (vue, ouïe), intellectuel ou social ne soit responsable. L'Institut National de la Santé Et de la Recherche Médicale (Inserm) identifie au sein de ces troubles la dyspraxie (trouble du développement moteur et de l'écriture), la dyscalculie (trouble des activités numérique), la dysphasie (trouble du langage oral), les troubles de l'attention, et la dyslexie. L'Inserm considère la dyslexie comme un trouble spécifique de la lecture dont la caractéristique essentielle est une altération spécifique et significative de l'acquisition de la lecture (Pull, 1994). La correspondance entre les graphèmes et morphèmes, ainsi que dans le rôle des mots dans la phrase (Ramus *et al.*, 2003) est le problème majeur rencontré par les enfants atteints de ce trouble phonologique (Snowling, 2012). Les maisons d'édition de littérature jeunesse ont su proposer des adaptations pour ces lecteurs. Nous pensons notamment à l'emploi de polices textuelles spécifiques, de marges et espacements plus importants, ou à l'épuration des illustrations pour une meilleure concentration. L'ère du numérique promet non seulement une économie de

production, mais surtout une automatisation de ces stratégies, et ce grâce à la tablette, permettant l'achat et la consultation de livres créés spécifiquement pour ce public. Ainsi, les éditeurs ont pu proposer des fonctionnalités pour les lecteurs dyslexiques, faisant cependant généralement abstraction des contraintes typographiques, en proposant par exemple des interfaces audio, de type dictée vocale et lecteurs d'écran (Sitbon *et al.*, 2007). La société Mobidys, quant à elle, souhaite offrir diverses fonctionnalités relatives à la typographie et sa prise en charge dans la lecture numérique sur tablette : l'aération de la mise en page, l'utilisation de polices spécifiques, la coloration syllabique ou phonétique, ou l'accès au dictionnaire. Ces stratégies ont déjà prouvé leur efficacité sur un navigateur (Parilova *et al.*, 2016) et permettent au lecteur dyslexique de mobiliser au maximum sa mémoire à court terme et, donc, ses ressources attentionnelles sur le texte et son sens. Néanmoins, ce travail typographique n'est qu'un des éléments à prendre en compte. La lisibilité, facilité à s'approprier l'information d'un texte, est aussi liée à la taille du texte, sa cohérence, et son découpage (Sitbon *et al.*, 2007). Cet article s'attachera donc à décrire une méthode de découpage automatique de textes de littérature jeunesse, adaptés aux lecteurs dyslexiques, et disponibles en version numérique. Le dégagement du concept de rhèse et son application dans le domaine du TALN s'inscrit dans un projet de partenariat avec l'entreprise Mobidys.

Dans la section 2, nous nous accorderons sur une définition de notre unité de découpage textuel : la rhèse. Nous discuterons ensuite, dans la section 3, de la constitution du corpus de référence. La section 4 détaillera enfin les différentes méthodes utilisées, qui seront évaluées en section 5.

2 Définitions de la rhèse

Nos recherches ont permis de dégager trois grands domaines contemporains : l'orthophonie, la linguistique et le théâtre. Nous ne nous intéressons ici qu'aux 2 premiers domaines, la définition théâtrale étant liée à des impératifs de diction spécifiques. Tout d'abord, en orthophonie, la rhèse s'emploie pour désigner la quantité de discours prononçable dans un souffle expiratoire (Brin *et al.*, 2011). La rhèse joue alors un rôle tonique, où le schéma vocalique transmet l'émotion voulue. En linguistique, la rhèse est « unité de cadence (...) groupement formé d'ordinaire par un factif ou un substantif accompagnés de leurs compléments les plus proches. On peut dire par exemple : *J'ai parlé | au roi*, en deux rhèses, ou : *J'ai parlé au roi*, en une seule rhèse. » (Damourette & Pichon, 1936). Enfin, les travaux de (Cartier, 1978) dans le domaine des TICE introduira notre définition de travail. La rhèse y apparaît comme l' « unité-membre de la phrase, ou petite phrase ayant une signification par elle-même, capable de former une unité de pensée. (Cette dernière) dégagée est purement intuitive et empirique, (sans) aucun critère psycholinguistique précis (...) la segmentation en rhèses (conduirait) à une segmentation perceptive. » (Ehrlich & Tardieu, 1985). La rhèse, dans son acceptation en orthophonie, est à rapprocher du concept d'unité prosodique. Plusieurs travaux visant à identifier automatiquement les frontières de ces unités existent. La plupart effectue cette tâche en analysant en parallèle un texte parlé et sa transcription à l'écrit (Avanzi *et al.*, 2008). Black et Taylor (Black & Taylor, 1998) proposent une méthode d'identification automatique de ces frontières basée uniquement sur des données textuelles. Leur approche, appliquée sur des textes en anglais, est basée sur une méthode d'apprentissage à base de modèles de Markov.

Les chunks (*analyse syntaxique de surface*) sont également proches des rhèses, puisqu'un chunk est « la plus petite séquence d'unités linguistiques possible formant un groupe avec une tête forte, et qui n'est ni discontinue, ni récursive » (Abney, 1991). Comme les chunks « définissent la structure

syntaxique superficielle des phrases » (Constant *et al.*, 2011), nous pourrions considérer chaque chunk comme une rhèse. Cependant, le chunking découpera trop finement les énoncés, notamment sur les mots grammaticaux, ce qui gênera fortement la lecture. Par exemple, le texte *La pensée qu'il était temps de chercher le sommeil m'éveillait ;* sera découpé en chunks[1] de la façon suivante : *La pensée* | *qu'il* | *était* | *temps* | *de* | *chercher* | *le sommeil* | *m'éveillait* | *;* tandis que nous souhaiterions obtenir le découpage en rhèses suivant : *La pensée* | *qu'il était temps* | *de chercher le sommeil* | *m'éveillait ;*. L'entraînement d'un « rhéseur », à partir d'un corpus segmenté en rhèses, est alors nécessaire pour découper un texte en rhèses de la façon souhaitée. C'est ce que nous verrons en section 4.

3 Constitution du corpus de référence

L'entreprise Mobidys nous a fourni un premier corpus restreint, comprenant deux ouvrages de littérature jeunesse, aux lectorats différents : *L'Arbre et le Bûcheron* comporte 2 560 mots (titres des chapitres exclus) et est destiné aux enfants de 11 ans et plus ; *Ali Baba et les 40 voleurs* comporte 1 426 mots (titres des chapitres exclus) et est destiné aux enfants de 8 ans et plus. Ces deux textes sont pré-traités avec un découpage en rhèses réalisé par les experts orthophonistes, tel qu'ils souhaitent le voir fait automatiquement. Ces deux ouvrages constituent notre corpus de référence.

Comme le font remarquer (Damourette & Pichon, 1936), pour un même énoncé[2], plusieurs segmentations en rhèses sont acceptables. Il convient donc d'observer l'accord moyen entre annotateurs humains, pour cette tâche de segmentation en rhèses, afin de fournir un référentiel pour interpréter les scores obtenus par les méthodes automatiques présentées ci-après. Deux annotateurs familiarisés à la tâche (une étudiante en français-langues étrangères et un étudiant en traitement automatique du langage) ont ainsi manuellement annoté en rhèses chacun des textes et comparé la segmentation obtenue avec celle proposée par les orthophonistes. Pour ce faire, les mesures utilisées sont :

- Le Kappa de Fleiss : nous considérons la tâche comme une tâche de classification des interstices entre les tokens, pouvant être classifiés en tant que frontières de rhèse ou non.
- La F-mesure moyenne : nous considérons la tâche comme une tâche de détection des frontières de rhèses, et comparons les annotations deux à deux en considérant temporairement l'une d'elle comme celle de référence.

Texte	Kappa de Fleiss	F-mesure
L'Arbre et le Bûcheron	0,83	0,87
Ali Baba et les 40 voleurs	0,81	0,86

Table 1: Accords inter-annotateurs sur l'identification manuelle des rhèses

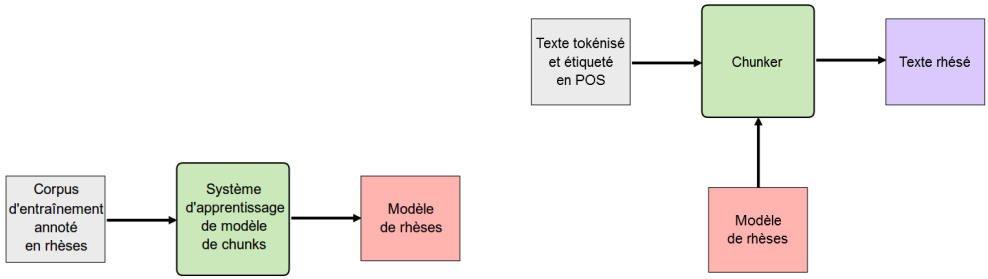
La table 1 présente les valeurs obtenues pour chacune des mesures, sur les 2 corpus de référence, entre les trois annotateurs (une orthophoniste et les deux étudiants). Les scores obtenus pour ces deux mesures dénotent d'un accord presque parfait entre les annotateurs.

[1]Ce découpage en chunks est obtenu en utilisant le chunker d'OpenNLP, à l'aide d'un modèle appris sur le Free French Treebank (Hernandez & Boudin, 2013).

[2]Un énoncé est défini par (Siouffi & Van Raemdonck, 2012) comme « Fragment de discours, inférieur ou supérieur à la phrase. Réalisation d'une phrase dans une situation déterminée. »

4 Apprentissage supervisé de rhéseurs à partir de corpus segmenté en rhèses

L'utilisation d'un chunker fournit une approximation d'une segmentation en rhèses, mais nous pensons pouvoir améliorer l'identification de ces dernières en apprenant un modèle spécifique de segmentation en rhèses. Les modèles de rhèses seront générés par un outil d'apprentissage statistique à l'aide d'un corpus d'entraînement annoté manuellement en rhèses, comme indiqué dans la figure 1a. Le texte servant à l'apprentissage du modèle est d'abord tokenisé, puis étiqueté grammaticalement à l'aide d'OpenNLP et d'un modèle d'étiquetage appris sur le Free French Treebank (Hernandez & Boudin, 2013), les étiquettes étant utilisées pour l'apprentissage du modèle de rhèses. Pour ensuite segmenter un texte en rhèses, en utilisant le modèle de rhèses précédemment appris, le texte à segmenter en rhèses est lui aussi préalablement tokenisé et étiqueté grammaticalement à l'aide des mêmes outils (voir figure 1b).



(a) Apprentissage du modèle (b) Utilisation du modèle

Figure 1: Processus d'apprentissage et de segmentation en rhèses d'un texte

L'apprentissage d'un modèle de rhèses nécessite ainsi la mise à disposition d'un corpus manuellement segmenté en rhèses. Cette segmentation manuelle peut être réalisée soit de manière intuitive, soit à l'aide d'un guide d'annotation. C'est ce que nous présentons dans la suite de cette section.

4.1 Apprentissage à partir d'un corpus d'entraînement rhésé intuitivement

Le corpus d'entraînement, *Emporté par le vent* (ouvrage jeunesse de 6 794 mots, également fourni par Mobidys), est annoté en rhèses de façon intuitive par une étudiante linguiste. Cette annotation nous permet ensuite d'apprendre un modèle de segmentation en rhèses à l'aide d'un outil d'apprentissage initialement prévu pour la segmentation en chunks.

Du fait de la quantité restreinte de données d'entraînement, en ayant à disposition un unique corpus d'entraînement, les constructions sémantiques ou locutives sont difficiles à prendre en compte par une approche statistique et détériorent l'apprentissage. Une autre stratégie a alors été envisagée : la création d'un guide d'annotation. Cela permet non seulement d'augmenter la taille du corpus grâce à l'annotation simultanée par plusieurs personnes, mais aussi de faciliter un apprentissage statistique en basant la segmentation en rhèses sur des principes plus formels.

4.2 Apprentissage à partir d'un corpus rhésé à l'aide d'un guide d'annotation

Le guide d'annotation, que nous avons élaboré, propose une définition de travail des rhèses, basée sur des règles grammaticales. Le guide d'annotation à été élaboré de façon itérative, en dégageant des règles formelles de l'observation des segmentations intuitives. Ce dernier prescrit aux annotateurs les règles grammaticales de découpage en rhèses, le but étant de respecter un nombre maximal de caractères par ligne égal à 30 (empan visuel maximal souhaité par les experts orthophonistes). Une première étape découpe l'énoncé sur la ponctuation, puis - si l'on excède toujours l'empan maximal - sur les propositions, ensuite sur les conjonctions et les prépositions. Si cela ne suffit pas, le guide prescrit le découpage des syntagmes, avant d'en arriver à une segmentation à l'échelle du mot.

Texte	Kappa de Fleiss	F-mesure
L'Arbre et le Bûcheron	0,94	0,94
Ali Baba et les 40 voleurs	0,93	0,94
Emporté par le vent	0,90	0,93

Table 2: Accords inter-annotateurs obtenus pour l'annotation à l'aide du guide d'annotation

Les scores d'accord inter-annotateurs présentés dans la table 2 sont calculés avec les mesures détaillées en section 3. Ils sont obtenus pour l'annotation d'un même texte par les deux mêmes annotateurs familiarisés à la tâche que ceux de la table 1 ; ils sont calculés sur le corpus d'entraînement ainsi que sur les deux corpus de référence. Ils sont sensiblement plus élevés que les scores d'accord obtenus lors de la segmentation intuitive en rhèses (voir table 1). Cette amélioration des accords, grâce à l'utilisation d'un guide d'annotation, permet de nuancer les problèmes liés à la faible quantité de données pour une méthode d'apprentissage statistique, en amenant une forte régularité dans la segmentation en rhèses.

5 Résultats et discussion

La tokenisation et l'étiquetage grammatical préalables sont effectués à l'aide d'OpenNLP, sur des modèles entraînés sur le Free French Treebank (Hernandez & Boudin, 2013). L'outil de chunking utilisé est OpenNLP, basé sur une technologie de modèles de Markov tels qu'utilisés par Schmid et Atterer (Schmid & Atterer, 2004). Le modèle de chunking utilisé est également appris sur le Free French Treebank (Hernandez & Boudin, 2013). Les modèles de segmentation d'un texte en rhèses sont obtenus par apprentissage, respectivement sur un corpus d'entraînement créé à partir d'*Emporté par le vent* segmenté intuitivement en rhèses (voir section 4.1), et sur un corpus créé à partir du même texte mais segmenté à l'aide du guide d'annotation (voir section 4.2). Les modèles sont évalués en comparant la segmentation en rhèses produite par l'utilisation de chacun d'eux avec celle résultant du travail des orthophonistes, à l'aide de la F-mesure.

Les scores présentés dans la table 3 indiquent que l'apprentissage d'un modèle sur un corpus annoté spécifiquement en rhèses apporte une nette amélioration de leur identification par rapport à la simple utilisation d'un outil de chunking. Ils révèlent également l'apport de l'utilisation d'un guide d'annotation, qui permet, par la régularité de ses règles, de réduire le bruit lors de l'apprentissage, améliorant encore sensiblement la segmentation. Ces scores peuvent être comparés aux accords inter-annotateurs présentés dans la section 3, utilisant la même F-mesure, et permettant d'évaluer la

Texte	Chunker (*baseline*)	Rhéseur 1 (à partir du corpus segmenté en rhèses de façon intuitive)	Rhéseur 2 (à partir du corpus segmenté selon le guide)
L'Arbre et le Bûcheron	Précision : 0,35 Rappel : 0,92 F-mesure : 0,51	Précision : 0,90 Rappel : 0,70 F-mesure : 0,79	Précision : 0,86 Rappel : 0,78 F-mesure : 0,82
Ali Baba et les 40 voleurs	Précision : 0,43 Rappel : 0,92 F-mesure : 0,59	Précision : 0,94 Rappel : 0,61 F-mesure : 0,74	Précision : 0,92 Rappel : 0,70 F-mesure : 0,79

Table 3: Évaluation des différents outils de segmentation en rhèses sur les corpus de référence

faisabilité de la tâche : $0,87$ d'accord sur la segmentation manuelle intuitive en rhèses pour *L'Arbre et le Bûcheron*, et $0,86$ pour *Ali Baba et les 40 voleurs*. Ils peuvent également être comparés aux scores de F-mesure obtenus par Schmid et Atterer (Schmid & Atterer, 2004) pour une tâche similaire sur de l'anglais : leur meilleur outil obtient des scores allant de 0,78 à 0,85 selon les corpus d'évaluation.

Segmentation automatique avec le chunker	Segmentation de référence	Observations
Les arbres | *sont* | *roux* | *et* | *dorés*	*Les arbres sont roux et dorés*	Granularité trop fine

Segmentation automatique avec le rhéseur 1	Segmentation de référence	Observations
Les arbres sont roux | *et dorés*	*Les arbres sont roux et dorés*	Granularité trop fine
Il regarde autour | *de lui*	*Il regarde autour de lui*	Segmentation détériorant la lisibilité
Il reste quelques mûres et de petites pommes sauvages | *dans les haies*	*Il reste quelques mûres* | *et de petites pommes sauvages* | *dans les haies*	Granularité pas assez fine

Segmentation automatique avec le rhéseur 2	Segmentation de référence	Observations
Cela dura mille | *et une nuits !*	*Cela dura* | *mille et une nuits !*	Locutions et entités nommées non prises en compte
L'Homme est très | *en colère*	*L'Homme est* | *très en colère*	Adverbes souvent mal segmentés lorsqu'ils sont présents entre un syntagme verbal et son objet

Table 4: Évaluation qualitative des résultats produits par les différents outils de segmentation

L'analyse qualitative des segmentations produites est également intéressante car elle permet de mieux identifier les types d'erreurs les plus courantes ainsi que de juger plus finement de la qualité de l'identification proposée. La table 4 présente des exemples de comparaison des segmentations de référence par rapport aux segmentations obtenues avec les différentes méthodes automatiques. Les rhéseurs appris permettent d'obtenir une granularité moins fine que celle obtenue avec le chunker,

ce qui se rapproche de la segmentation de référence souhaitée. Les segmentations produites à l'aide des rhéseurs peuvent cependant être encore trop fine par rapport aux segmentations de référence. Les segmentations non-satisfaisantes pour le rhéseur 2 sont le plus souvent dues à deux causes. La première cause est le mauvais rattachement d'un adverbe lorsque celui-ci est présent entre un syntagme verbal et son objet. Une analyse du cas des adverbes révèle que le rattachement de ceux-ci au syntagme verbal, ou à son objet, dépend généralement de leur catégorie. Par exemple, un adverbe de quantité sera préférentiellement rattaché à l'objet qui le suit *L'Homme est | très en colère*, tandis qu'un adverbe de temps sera généralement rattaché au syntagme verbal le précédant *Cassim se retrouve bientôt | dans la grotte au trésor !*. Une solution envisageable pour corriger ces cas est l'utilisation d'un modèle d'étiquetage grammatical différenciant les adverbes selon leurs catégories. La deuxième cause est le découpage d'une locution ou d'une entitée nommée. La correction de ces cas nécessiterait la détection de ces entitées en amont du processus de segmentation.

6 Conclusion et perspectives

Nous avons présenté dans cet article une méthode d'identification automatique des rhèses. Les différentes approches utilisées montrent la faisabilité de la tâche par une méthode d'apprentissage statistique, ainsi que l'apport d'un guide d'annotation : permettre l'apprentissage d'un rhéseur sur un volume plus important de données plus homogènes. La dernière version de cette méthode aboutit à des résultats satisfaisants, comparables à ceux produits par un expert humain.

Afin d'améliorer l'identification des rhèses, il serait intéressant de prendre en compte les diverses locutions, en les rendant insécables. Ce traitement pourrait être effectué en amont de la segmentation en rhèses, à l'aide d'un dictionnaire dédié. Il est également envisageable de détecter les tournures prenant un caractère locutif par leur répétition dans un texte donné, courantes dans le genre du conte (*Le grand méchant loup*, *Le vilain petit canard*), à l'aide d'une méthode de repérage des cooccurrences.

Références

ABNEY S. P. (1991). Parsing by chunks. In *Principle-Based Parsing: Computation and Psycholinguistics*, p. 257–278. Kluwer.

AVANZI M., LACHERET A. & VICTORRI B. (2008). Analor, un outil d'aide pour la modélisation de l'interface prosodie-grammaire. In *Actes du colloque CERLICO*, p. 27–46.

BLACK A. & TAYLOR P. (1998). Assigning phrase breaks from part-of-speech sequences. *Computer Speech and Language*, **12**(2), 99–117.

BRIN F., COURRIER C., LDERLÉ E. & MASY V. (2011). *Dictionnaire d'orthophonie*. Ortho Edition.

CARTIER M. (1978). Le caractère de l'édition de texte. Non publié.

CONSTANT M., TELLIER I., DUCHIER D., DUPONT D., SIGOGNE A. & BILLOT S. (2011). Intégrer des connaissances linguistiques dans un crf : application à l'apprentissage d'un segmenteur-étiqueteur du français. In *Actes de la conférence TALN*.

DAMOURETTE J. & PICHON E. (1936). Des mots à la pensée, essai de grammaire de la langue française. In ÉDITIONS D'ARTHREY, Ed., *Collection des linguistes contemporains*, chapter VII. CNRS.

EHRLICH M.-F. & TARDIEU H. (1985). Lire, comprendre, mémoriser les textes sur écran vidéo. *Communication et langages*, **65**(1), 91–106.

HERNANDEZ N. & BOUDIN F. (2013). Construction automatique d'un large corpus libre annoté morpho-syntaxiquement en français. In *Actes de la conférence TALN-RECITAL*.

PARILOVA T., MRVAN F., MIZIK B. & HLDKA E. (2016). Emerging technology enabling dyslexia users to read and perceive written text correctly. In *Actes de la conférence CiCLING*.

PULL C. (1994). *Troubles spécifiques du développement des acquisitions scolaires*. Classification internationale des maladies : dixième révision. O.M.S.

RAMUS F., ROSEN S., DAKIN S., DAY B., CASTELOTTE J., WHITE S. & FRITH U. (2003). Theories of developmental dyslexia : insights from a multiple case study of dyslexic adults. *Brain*, **4**(126), 841–865.

SCHMID H. & ATTERER M. (2004). New statistical methods for phrase break prediction. In *Actes de la conférence COLING*.

SIOUFFI G. & VAN RAEMDONCK D. (2012). *100 fiches pour comprendre la linguistique, 4ème édition*. Bréal.

SITBON L., BELLOT P. & BLACHE P. (2007). Éléments pour adapter les systèmes de recherche d'information aux dyslexiques. *Revue TAL*, **48**(3), 1–26.

SNOWLING M. (2012). Early identification and interventions for dyslexia : a contemporary view. *Journal of Research in Special Educational Needs (JORSEN)*, **13**(1), 7–14.