Détection automatique de phrases en domaine de spécialité en français

Arthur Boyer, Aurélie Névéol

Détection automatique de phrases en domaine de spécialité en français

Arthur Boyer[1] Aurélie Névéol[1]
(1) LIMSI, CNRS, Université Paris-Saclay, rue John von Neumann, Campus Universitaire, F-91405 Orsay
prénom.nom@limsi.fr

RÉSUMÉ

La détection de frontières de phrase est généralement considéré comme un problème résolu. Cependant, les outils performant sur des textes en domaine général, ne le sont pas forcement sur des domaines spécialisés, ce qui peut engendrer des dégradations de performance des outils intervenant en aval dans une chaîne de traitement automatique s'appuyant sur des textes découpés en phrases. Dans cet article, nous évaluons 5 outils de segmentation en phrase sur 3 corpus issus de différent domaines. Nous ré-entrainerons l'un de ces outils sur un corpus de spécialité pour étudier l'adaptation en domaine. Notamment, nous utilisons un nouveau corpus biomédical annoté spécifiquement pour cette tâche. La detection de frontières de phrase à l'aide d'un modèle OpenNLP entraîné sur un corpus clinique offre une F-mesure de .73, contre .66 pour la version standard de l'outil.

ABSTRACT

Sentence boundary detection for specialized domains in French

Sentence boundary detection is generally considered as a solved problem. However, tools that perform well on standard text do not necessarily deal well with specialized corpus, which may degrade the analysis of other natural language processing tools intended to process sentence-segmented text. In this paper, we conduct a benchmark evaluation of 5 standard sentence boundary detection tools on 3 corpora covering different domains and subdomains. We then retrain one of the tools on domain-specific data and show that this leads to improved performance. In particular, we experiment with the clinical domain using a new clinical corpus annotated for gold-standard sentence boundaries. Sentence boundary detection with an openNLP model trained on the clinical data achieves an F-measure of .73, vs. .66 for standard openNLP distribution.

MOTS-CLÉS : Segmentation en phrases, domaine de spécialité, évaluation.

KEYWORDS: Sentence boundary detection, specialized corpus, benchmark evaluation.

1 Introduction

La segmentation en phrases, aussi appelée "détection de frontière de phrase" (DFP) ou "sentence boundary detection" en anglais, est l'une des premières étapes des chaînes de traitement du langage naturel, sur laquelle repose les étapes suivantes tel que la segmentation en mots (ou *tokenisation*), l'étiquetage morpho-syntaxique, la reconnaissance d'entités nommées. Les performances élevées obtenues pour les corpus journalistiques en anglais font que la segmentation en phrases est globalement considérée comme un problème résolu (Kiss & Strunk, 2006). Cependant, les bons résultats obtenus sur le domaine général ne se maintiennent pas toujours sur des domaines spécialisés, ce qui

a des répercutions sur l'ensemble des étapes postérieures dans une chaîne de traitement. De plus, les performances des outils de DFP varient probablement entre différentes langues à cause de différences morphologiques ou des ressources annotées disponibles. Nous nous sommes particulièrement intéressés à l'application d'outils de traitement automatique de la langue au domaine biomédical en français. À notre connaissance, il n'existe pas d'évaluation des outils ou méthodes de segmentation en phrase en français qui pourrait guider le choix des chercheurs selon les caractéristiques du corpus ou l'utilisation voulue. Pour combler ce vide, nous avons conduit une étude comparative de quatre outils de segmentation appliqués sur trois corpus en français.

2 Travaux proches

Dans les textes de domaine général et journalistique, le principal obstacle de la segmentation en phrases est lié aux abréviations. Les systèmes doivent identifier si un point indique une fin de phrase ou marque une abréviation, puis reconnaitre si cette marque d'abréviation est une fin de phrase (Gillick, 2009). De nombreuses méthodes reposent sur les signes de ponctuation comme le point de ponctuation, le point d'exclamation, le point d'interrogation, les points de suspension et les deux-points pour identifier les fins de phrase, soit avec un jeu de règles soit en apprenant à classer ces marqueurs à l'aide d'un corpus d'entrainement (Agarwal *et al.*, 2005; Urieli, 2013).

Les textes en domaine de spécialité apportent de nouveaux défis, touchant à la sémantique avec par exemple des abréviations absentes des dictionnaires non-spécialisés (comme les noms d'organismes tels que `E. Coli` ou `A. Thaliana` dans les textes biomédicaux) à la syntaxe (c'est-à-dire le manque de signes de ponctuation ou leur utilisation de façon non-conventionnelle) ou encore à la structure globale du document (abondance des listes à points, de titres de sections qui constituent des segments assimilables à une phrase). Les problèmes liés aux lettres capitales et aux ponctuations ont été longuement étudiés dans le contexte du traitement de l'oral et de la transcription de parole, où les signes de ponctuation ne sont pas disponibles (Treviso *et al.*, 2017).

Un autre domaine où les détections de frontières de phrases ont reçu une attention particulière est le domaine de la traduction automatique, qui s'appuie sur des corpus alignés au niveau des phrases pour l'entrainement de modèles statistiques. Quelques travaux ont évalué l'impact de performance de la segmentation dans la traduction automatique (Collados, 2013). D'autres travaux ont revisité la notion de "phrase" pour préparer les textes en segments plus courts, assimilables à des phrases (Kuang & Xiong, 2016).

Newman-Griffis *et al.* (2016) évaluent les outils de détection des frontières de phrases en anglais sur des corpus de domaines différents dont des textes journalistiques, des transcriptions d'appels téléphoniques, des résumés d'articles scientifiques et des textes cliniques. Dans leur travaux, ils mettent en avant les difficultés liées au traitement des textes cliniques avec des outils non-entrainés pour cette tâche spécifique. Miller *et al.* (2015) présentent des expériences sur des textes cliniques en anglais avec un modèle statistique entrainé sur des caractères et montrent que des résultats satisfaisant peuvent être obtenus avec une quantité limitée de données annotées. Kreuzthaler & Schulz (2015) ont abordé, avec des résultats positifs, le problème de la détection des abréviations et de la segmentation des phrases pour les textes cliniques en allemand.

3 Segmentation en phrases en français

3.1 Problématique

La segmentation en phrase est un problème qui est peu abordé en traitement automatique de la langue, car il est considéré comme résolu pour des textes de langue générale où les marqueurs de fin de phrase sont facilement identifiables et font partie d'une liste fermée restreinte. Pour d'autres types de texte issus des réseaux sociaux comme twitter, les travaux portent sur l'ensemble du segment de 140 caractères quil n'est pas nécessaire de redécouper. Cependant pour des textes de spécialité comme les textes du domaine biomédical, la question de la segmentation constitue un réel problème. En TAL les chaines de traitement commencent par une segmentation des textes en unités : des phrases, puis des syntagmes et des mots. C'est par exemple le cas de CTakes, un outil d'analyse des textes cliniques en anglais(Savova *et al.*, 2010), que nous souhaitons adapter à d'autres langues dont le français.

La segmentation en phrase présente plusieurs difficultés, d'ordre définitoire, méthodologique et technique. En effet, la définition d'une "phrase" est principalement accessible au travers des quelques corpus segmentés en phrases disponibles, comme le French Tree Bank (FTB) et Sequoïa. Par ailleurs, la rareté des corpus disponibles rend difficile l'évaluation de différentes méthodes et outils. En effet, la plupart des outils implémentant des méthodes d'apprentissage (par exemple, OpenNLP ou Talismane) sont entrainés sur le corpus FTB. Enfin, on constate également en pratique que la segmentation en phrases des outils s'accompagne d'une transformation du texte original : typiquement, Talismane propose une sortie au format coNLL tandis qu'OpenNLP présente le texte segmenté avec une phrase par ligne. Dans les deux cas des insertions, suppression ou substitutions de caractères (espaces ou ponctuation) posent des problèmes techniques supplémentaires pour l'alignement de deux versions d'un document, afin d'évaluer la segmentation proposée par rapport à une segmentation standard.

3.2 Contribution

Dans cet article nous proposons une contribution qui permet d'apporter des éléments de réponse à l'ensemble de ces difficultés. D'une part nous présentons deux nouveaux corpus annotés en phrases dans le domaine biomédical, ce qui permet de proposer une caractérisation de la phrase pour une variété de textes en français. Puis, nous nous appuyons sur ces nouvelles ressources pour faire des expérimentations sur la segmentation en phrases à l'aide des divers corpus et outils français disponibles. Nous présentons également un outil d'alignement de textes au niveau des phrases afin d'évaluer la segmentation.

3.3 Présentation des corpus et outils utilisés

Dans cette partie, nous présentons les corpus et outils utilisés pour réaliser notre étude. Le tableau 1 offre une description synoptique des corpus utilisés, que nous décrivons brièvement ci-dessous.

Afin d'élargir le nombre et la diversité des corpus français disposant d'une segmentation en phrase de référence, nous avons annoté deux corpus du domaine biomédical avec des frontières de phrases. Le **corpus EDP** est une collection de 338 titres et résumés d'articles biomédicaux. Il a été développé pour la tâche de traduction automatique dans le domaine biomédical dans le cadre de WMT 2017 (Jimeno Yepes *et al.*, 2017). Le **corpus MERLoT** (Campillos *et al.*, 2017) est un corpus composé de

Corpus	Type de Texte	Nombre de phrases	Long. moy. phrases
EDP	Articles scientifiques	3368	19.18
MERLoT	Textes Cliniques	7 836	6.34
French Treebank	Presse Nationale	21 564	24.41
Sequoia	Mixte	3 204	18.67
- Annodis	Presse Régionale	529	18.18
- EMEA	Notice de Médicament	1 118	16.39
- Europarl	Débat du Parlement Européen	561	23.24
- frwiki	Article d'encyclopédie	996	18.90

TABLE 1 – Description statistique des corpus utilisés.

documents cliniques désidentifiés issus des dossiers électroniques patient d'un groupe hospitalier français. Dans le cadre de ce travail sur la segmentation en phrases, nous avons annoté manuellement une partie du corpus (160 documents sur 500).Nous avons utilisé l'outil BRAT (Stenetorp *et al.*, 2012). Les documents ont été pré-annotés automatiquement en considérant comme fin de phrase chaque fin de ligne dans le corpus MERLoT, et chaque pontuation forte ou semi-forte pour le corpus EDP. Une série de six ensembles de cinq documents ont ensuite été corrigées par deux annotateurs, avec une réunion de consensus pour chaque série afin de discuter des désaccords et de finaliser le guide d'annotation. Une fois l'accord inter-annotateur stabilisé au-delà de .95 en F-mesure, chaque annotateur a travaillé indépendemment sur une partie des corpus restant.

Nous avons également utilisés les corpus existant **French TreeBank** (Abeillé *et al.*, 2003) pour l'entrainement d'outils statistiques, et le **corpus Sequoia** (Candito & Seddah, 2012) qui rassemble des textes issus de domaine différents. Pour cette étude, nous distinguerons le corpus médical EMEA (1 118 phrases) et les autres corpus : Annodis, Europal et frwiki, notés par la suite *Sequoïa-G* (2 086 phrases).

Les expériences de segmentation ont été réalisées avec quatre suites d'outils standard en TAL, dont une dédiée au français. Ces outils, que nous décrivons brièvement ci-dessous, ont été comparés à une baseline à base de règles qui marque une fin de phrase après les signes de ponctuation forts "." , " !", " ?" ou semi-forts " ;", " :" et les retours à la ligne.

- Stanford CoreNLP s'appuie sur des règles pour effectuer la segmentation en phrase à la suite de l'étape de tokenisation (Manning *et al.*, 2014).
- La suite OpenNLP d'Apache repose sur un classifieur MaxEnt pour la segmentation en phrases. Cet outil est intégré à la plateforme Ctakes. Pour notre travail nous l'avons entrainé sur le French TreeBank.
- L'analyseur Talismane (Urieli, 2013) met en oeuvre une segmentation en phrase par classification binaire d'une liste de signes de ponctuation ; entrainé sur le French Tree Bank.
- NLTK (Natural Language ToolKit) est un librairie python en open source. Le modèle français a été entrainé sur un corpus du journal *Le Monde*.
- Unitex (Paumier, 2016) est un outil d'analyse linguistique qui offre un outil de segmentation en phrases à base de règles (Friburger *et al.*, 2000).

Les performances des outils de segmentation ont été évaluées en termes de précision, rappel et F-mesure à l'aide d'un script permettant l'alignement du texte segmenté et de la référence au format une phrase par ligne.

4 Caractérisation des fins de phrase dans les corpus français

Afin d'illustrer les particularités des différents corpus du point de vue de la segmentation en phrases, le tableau 2 présente la distribution des marqueurs de fin de phrase observés. On constate que pour les corpus FTB ainsi que la partie non médicale du corpus Sequoïa, les marqueurs de fin de phrase sont majoritairement des pontuactions fortes. Les autres marqueurs sont des chiffres et lettres, indiquant la présence de phrases de type "titre". Les corpus médicaux (EDP, EMEA et Merlot) sont intermédiaires et présentent une proportion importante de marqueurs de fin de phrase à l'aide de ponctuations semi-forte ou de marqueurs inhabituels.

Corpus	**Ponctuation forte** (. ? ! . . .)	**Ponctuation semi-forte** (; :)	**Chiffres et lettres**	**Autres marques**
FTB	90%	<1%	8%	1,6%
Sequoia	78,6%	3,2%	16,5%	1,6%
- non médical	84%	2,3%	12,9%	<1%
- médical (Emea)	68,4%	5%	23,3%	3,3%
EDP français	82,8%	16,9%	<1%	<1%
Merlot	27,3%	8%	25,3%	39,4%

TABLE 2 – Distribution des caractères de fin de phrase.

Par ailleurs, nous proposons ci-dessous quelques exemples représentatifs des cas difficiles que nous avons pu rencontrer dans les corpus médicaux. Dans chaque exemple, nous marquons la segmentation de référence par des crochets en gras, avec un numéro de segment en indice sur le crochet fermant. Dans (1) on observe une phrase contenant les caractères " :" et " ;" ne marquant pas une fin de phrase, alors que dans (2) les deux points marquent une fin de segment (titre) et le point virgule est utilisé comme une ponctuation forte marquant une fin de segment. (3) Illustre le cas de tableaux convertis. (4) illustre le cas d'une section de compte-rendu clinique rapportant des résultats d'analyse sous forme de liste non structurée.

(1) **{**Dans le cadre d'une dentisterie moderne, le praticien doit être à même d'apporter des solutions efficaces conjuguant : satisfaction du patient, en dissimulant ces défauts ; et économie tissulaire, avec l'approche la moins dommageable, laissant idéalement possible et aisée toute ré-intervention.**}**$_1$ **EDP - Actual. Odonto-Stomatol. 2014 ;269 :36-41**

(2) **{**Discussion :**}**$_1$ **{**La localisation et la teinte de la dyschromie concordaient avec la prise du traitement ;**}**$_2$ **{**tout ceci était étayé par l'absence de coloration chez la sœur jumelle.**}**$_3$ **EDP - Med Buccale Chir Buccale 2014 ;20 :279-283**

(3) {|IR DANS LE PARENCHYME |}$_1$ {POLE SUP |}$_2$ {MEDIAN |}$_3$ {POLE INF |}$_4$ {POLE SUP |}$_5$ {MEDIAN |}$_6$ {POLE INF |}$_7$
{| |0,78 |}$_8$ {0,81 |}$_9$ {0,79 | |}$_{10}$ {0,82 |}$_{11}$ {0,82 | }$_{12}$ **Extrait du corpus MERLoT**

(4) {EXAMENS COMPLEMENTAIRES : }$_1$ {Biologie : GB : 5,9 g/l. }$_2$ {PN : 3,0 g/l. }$_3$ {Plaquettes : 177 g/l. }$_4$ {Hb : 14,7 g/dl. }$_5$ {Créat. : 8,0.}$_6$ **Extrait du corpus MERLoT**

	Sequoia-EMEA			Sequoia-G			EDP			MERLoT		
	P	R	F	P	R	F	P	R	F	P	R	F
Stanford	.72	.49	.58	.87	.74	.80	.74	.81	.77	.66	.19	.29
OpenNLP	.78	.52	.63	.90	.83	.87	.81	.75	.78	.72	.61	.66
NLTK	.77	.53	.63	.91	.84	.87	.81	.74	.77	.56	.66	.61
Talismane	.78	.52	.63	.91	.84	.88	.81	.74	.77	.76	.62	.68
Unitex	.54	.73	.62	.75	.69	.72	.75	.81	.78	.63	.77	.69
Baseline	.68	.59	.63	.63	.81	.72	.92	.95	.93	.64	.68	.65

TABLE 3 – Evaluation d'outils de détection de phrases en français. Les performances sont mesurées en termes de précision (P), rappel (R) et F-mesure (F).

5 Expérimentations en segmentation

Le tableau 3 présente le résultat de l'application des outils de détection de phrase disponibles pour le français sur nos corpus de travail. On constate que les meilleures performances sont obtenues sur le corpus non médical Sequoïa-G qui présente le plus de ressemblance avec le French TreeBank, sauf dans le cas de la baseline qui s'avère bien adaptée au corpus EDP. Ces résultat reflètent la nature des corpus caractérisée par la distribution des caractères de fins de phrase présentées dans le tableau 2. En effet, une baseline fondée sur les ponctuations fortes et semi-fortes est particulièrement adaptée pour le corpus EDP dans lequel 98% des fins de phrases sont marquées par ce type de ponctuation. Le corpus MERLoT est celui qui présente le plus de diversité de fins de phrases avec presque 40% de marqueurs inhabituels, ce qui explique le faible rappel pour un outil comme Stanford, et conduit à des performances médiocres. A l'inverse, l'outil Unitex, également à base de règles, offre une bonne couverture ce qui se traduit par un rappel élevé. Néanmoins la précision reste faible et la F-mesure globale est similaire à celle des autres outils.

Nous avons également réalisé une série d'expériences plus spécifiques aux corpus du domaine biomédical (tableau 4). Nous avons entrainé des modèles statistiques fondés sur le maximum d'entropie (implémenté dans l'outil OpenNLP) sur plusieurs configurations des corpus d'entrainement :

1. dans la première configuration (Test1), on cherche à construire un modèle le plus ciblé au corpus de test, c'est à dire qui utilise le maximum de document du même corpus. L'entrainement est effectué sur deux tiers du corpus et le test sur le tiers restant.
2. dans la deuxième configuration (Test 2), on cherche à construire un modèle qui utilise le plus gros volume de données d'entrainement disponible pour chaque corpus. L'entrainement est effectué sur deux tiers des trois corpus et le test sur chacun des tiers restant des corpus médicaux.
3. dans la troisième configuration (Test 3) on cherche à construire un modèle qui utilise le plus gros volume de données d'entrainement disponible pour chaque corpus, tout en s'assurant que le corpus cible représente au moins un tiers des données d'entrainement. Le corpus d'entrainement est constitué d'une proportion égale des trois corpus. La disparité de taille entre les corpus fait que pour la composition du corpus d'entrainement de MERLoT il a été nécessaire d'utiliser l'intégralité des corpus EMEA et EDP.

La taille de ces corpus d'entrainement explique l'écart entre les resultats du tableau 3 et reportés pour OpenNLP dans le tableau 4 (rappelons que le corpus d'entrainement French Tree Bank utilisé comporte 21 564 phrases, soit plus du double de notre plus gros corpus d'entrainement spécialisé

utilisé dans les expériences du tableau 3). Les deux dernières stratégies semblent particulièrement adaptées pour le corpus clinique MERLoT, malgré la petite taille des corpus d'entrainement en comparaison avec le French Tree Bank.

	Sequoia-EMEA T=373				**EDP** T=1 123				**MERLoT** T= 2 612			
	E	P	R	F	E	P	R	F	E	P	R	F
Test 1	745	.21	.42	.28	2 245	.25	.62	.36	5 224	.31	.74	.43
Test 2	8 214	.71	.44	.54	8 214	.75	.60	.67	8 214	.78	.67	.72
Test 3	2 235	.73	.44	.55	5 608	.75	.60	.67	9 710	.78	.68	.73

TABLE 4 – Evaluation de modèles fondés sur le maximum d'entropie (OpenNLP) entrainés sur des corpus médicaux français. La taille de chaque corpus d'entrainment (E) et de test (T) est indiquée en nombre de phrases. Les performances sont mesurées en termes de précision (P), rappel (R) et F-mesure (F).

6 Conclusion et perspectives

Une conclusion assez surprenante de cette étude est que la performance des outils de segmentation en phrase pour le français est globalement modeste, en particulier en comparaison avec les performances sur l'anglais qui se situent bien au delà de .90 de F-mesure pour des corpus de langue générale. Concernant la segmentation en phrases pour les textes du domaine biomédical, il semble que le développement d'outils dédiés soit à base de règles soit statistiques reposant sur des corpus du domaine soit indispensable. Dans la suite de ce travail, nous prévoyons d'expérimenter avec des modèles statistiques reposant sur une segmentation en caractères, et d'évaluer l'impact de la segmentation en phrases sur des tâches d'extraction d'information comme la reconnaissance d'entités nommées ou l'extraction de relations.

Remerciements

Nous remercions le Service d'Informatique Biomédicale (SIBM) ainsi que l'équipe CISMeF du CHU de Rouen qui nous ont permis d'utiliser le corpus LERUDI pour cette étude. Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence CABeRneT ANR-13-JS02-0009-01.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). *Building a Treebank for French*, In A. ABEILLÉ, Ed., *Treebanks : Building and Using Parsed Corpora*, p. 165–187. Springer Netherlands : Dordrecht.

AGARWAL N., FORD K. H. & SHNEIDER M. (2005). *Sentence Boundary Detection Using a MaxEnt Classifier*. Rapport interne, Natural Language Processing Group, Stanford University.

CAMPILLOS L., DELÉGER L., GROUIN C., HAMON T., LIGOZAT A.-L. & NÉVÉOL A. (2017). A french clinical corpus with comprehensive semantic annotations : development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources and Evaluation*.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012 - 19e conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, France.

COLLADOS J. C. (2013). Splitting complex sentences for natural language processing applications : Building a simplified spanish corpus. *Procedia - Social and Behavioral Sciences*, **95**(Supplement C), 464 – 472. Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions : Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013).

FRIBURGER N., DISTER A. & MAUREL D. (2000). Améliorer le découpage des phrases sous Intex. *Revue Informatique et Statistique dans les Sciences Humaines*, **36**(1-4), 181–200.

GILLICK D. (2009). Sentence boundary detection and the problem with the u.s. In *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume : Short Papers*, NAACL-Short '09, p. 241–244, Stroudsburg, PA, USA : Association for Computational Linguistics.

JIMENO YEPES A., NEVEOL A., NEVES M., VERSPOOR K., BOJAR O., BOYER A., GROZEA C., HADDOW B., KITTNER M., LICHTBLAU Y., PECINA P., ROLLER R., ROSA R., SIU A., THOMAS P. & TRESCHER S. (2017). Findings of the wmt 2017 biomedical translation shared task. In *Proceedings of the Second Conference on Machine Translation, Volume 2 : Shared Task Papers*, p. 234–247, Copenhagen, Denmark : Association for Computational Linguistics.

KISS T. & STRUNK J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, **32**(4), 485–525.

KREUZTHALER M. & SCHULZ S. (2015). Detection of sentence boundaries and abbreviations in clinical narratives. *BMC Medical Informatics and Decision Making*, **15**(2), S4.

KUANG S. & XIONG D. (2016). *Automatic Long Sentence Segmentation for Neural Machine Translation*, In C.-Y. LIN, N. XUE, D. ZHAO, X. HUANG & Y. FENG, Eds., *Natural Language Understanding and Intelligent Applications : 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2–6, 2016, Proceedings*, p. 162–174. Springer International Publishing.

MANNING C. D., SURDEANU M., BAUER J., FINKEL J., BETHARD S. J. & MCCLOSKY D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, p. 55–60.

MILLER T. A., FINAN S., DLIGACH D. & SAVOVA G. K. (2015). Robust sentence segmentation for clinical text. In *AMIA Annu Symp*.

NEWMAN-GRIFFIS D., SHIVADE C., FOSLER-LUSSIER E. & LAI A. (2016). A quantitative and qualitative evaluation of sentence boundary detection for the clinical domain. **2016**, 88–97.

PAUMIER S. (2016). UNITEX 3.1 Manuel d'utilisation. http ://unitexgramlab.org/releases/3.1/man/Unitex-GramLab-3.1-usermanual-fr.pdf. Université de Marne la Vallée.

SAVOVA G., MASANZ J., OGREN P., ZHENG J., SOHN S., KIPPER-SCHULER K. & CHUTE C. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) : architecture, component evaluation and applications. *J Am Med Inform Assoc*, **17**(5), 507–13.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). BRAT : a Web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, p. 102–7.

TREVISO M. V., SHULBY C. & ALUÍSIO S. M. (2017). Sentence segmentation in narrative transcripts from neuropsychological tests using recurrent convolutional neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1 : Long Papers*, p. 315–325.

URIELI A. (2013). *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, Université de Toulouse II le Mirail.