Constitution de corpus parallèle pour la simplification de textes médicaux

Rémi Cardon

CNRS, UdLille, UMR 8163 - STL - Savoirs Textes Langage
remi.cardon@univ-lille.fr

M2 SDL

Plan

- La simplification automatique de textes
- Le corpus CLEAR
- La tâche d'alignement de phrases
  - Annotation
  - Calcul d'accord inter-annotateur
  - Évaluation
  - Alignement automatique
  - Analyse d'erreurs

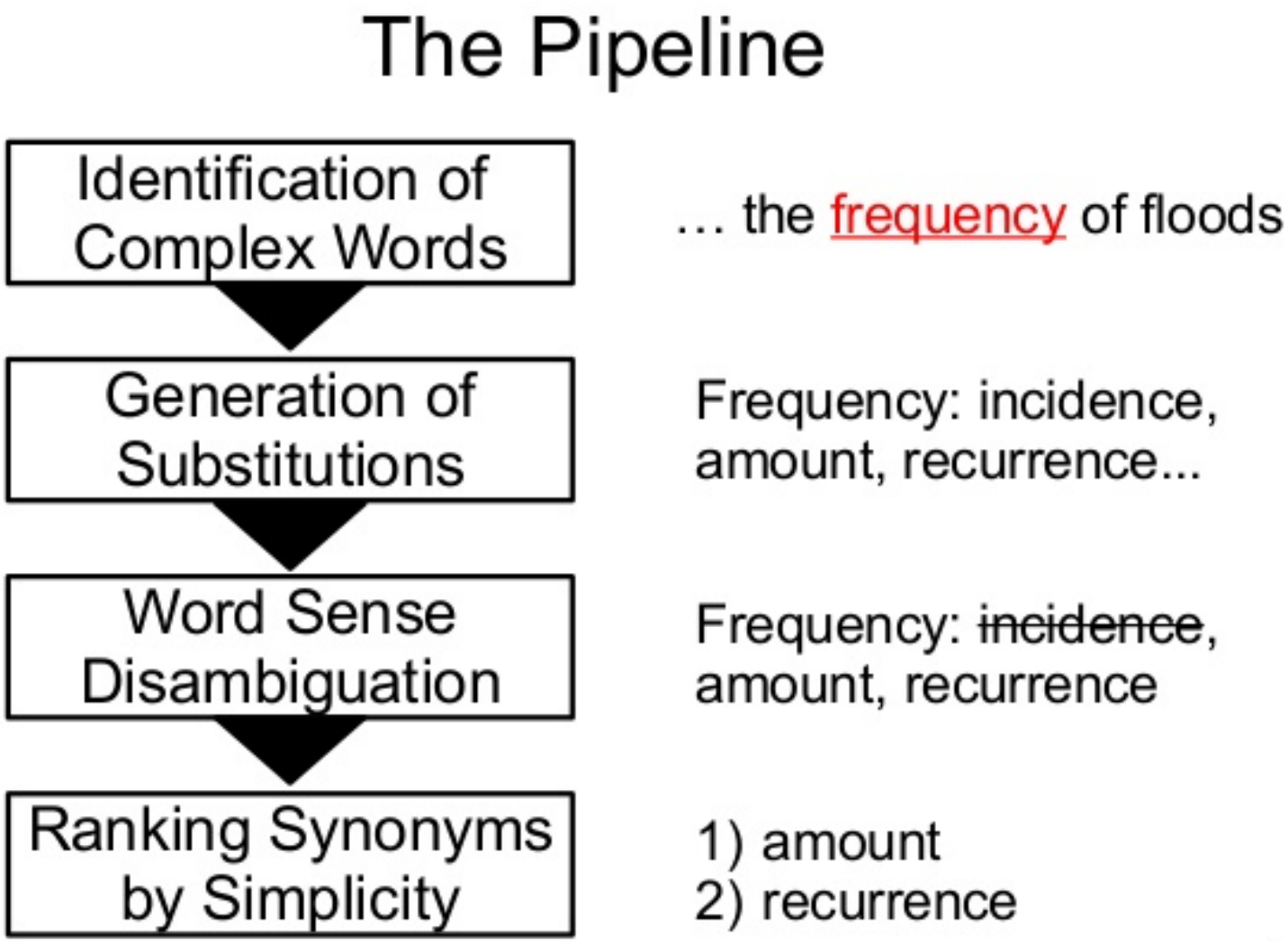
Simplification automatique de textes – Objectifs

- Rendre l'information accessible aux humains
  - Enfants ou adultes ayant des difficultés de lecture (Bosch *et al.*, 2008; De Belder *et al.*, 2010; Vu *et al.*, 2014; Paetzold & Specia, 2016a)
  - Étrangers (Paetzold & Specia, 2016b)
  - Adultes ayant des troubles neurocognitifs (Chen *et al.*, 2016)
  - Grand public (textes spécialisés) (Arya *et al.*, 2011)
  - ...
- Pré-traitement de données pour le TAL
  - Analyse syntaxique (Chandrasekar & Srinivas, 1997)
  - Résumé automatique (Blake *et al.*, 2007)
  - Traduction automatique (Stymne *et al.*, 2013; Štajner & Popović, 2016)
  - Extraction / Recherche d'information (Beigman Klebanov *et al.*, 2004)
  - ...

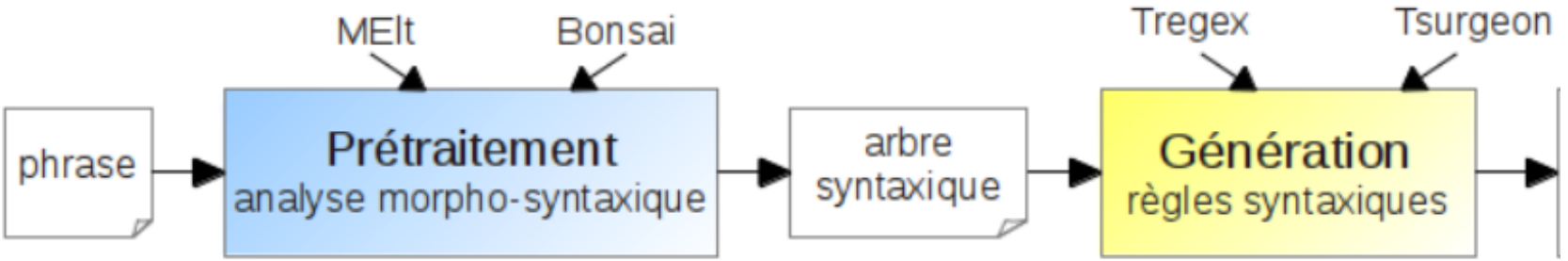
Simplification automatique de textes – Méthodes

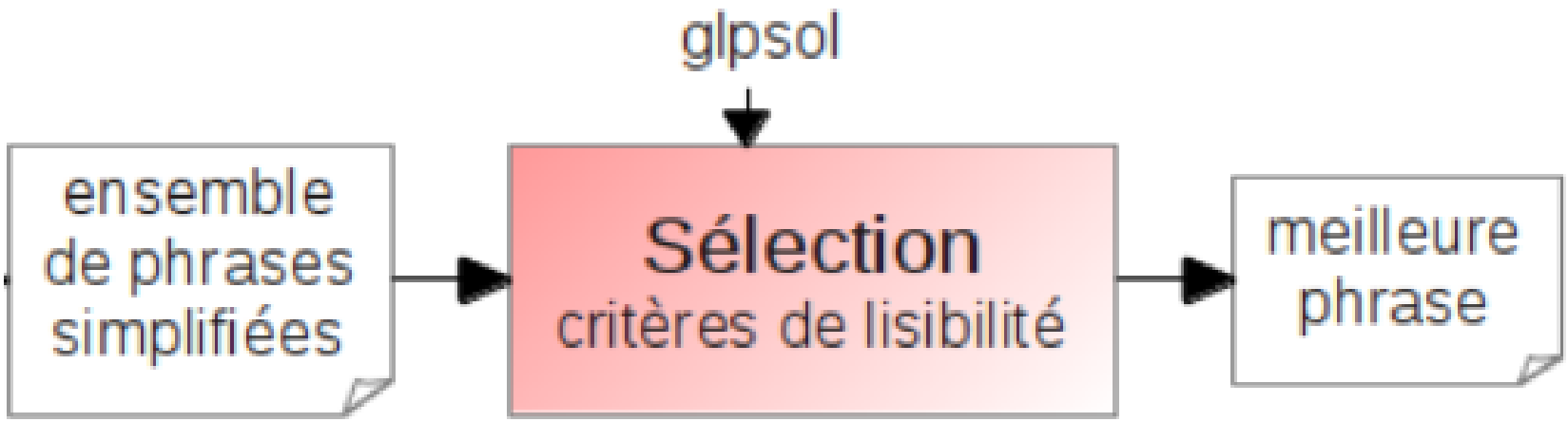
- Modifier un texte pour le rendre accessible, tout en préservant son sens
- Principalement deux versants :
  - Simplification lexicale : remplacer les termes ou expressions complexes par des équivalents accessibles
  - Simplification syntaxique : retravailler les structures de phrases

Simplification lexicale



Simplification syntaxique

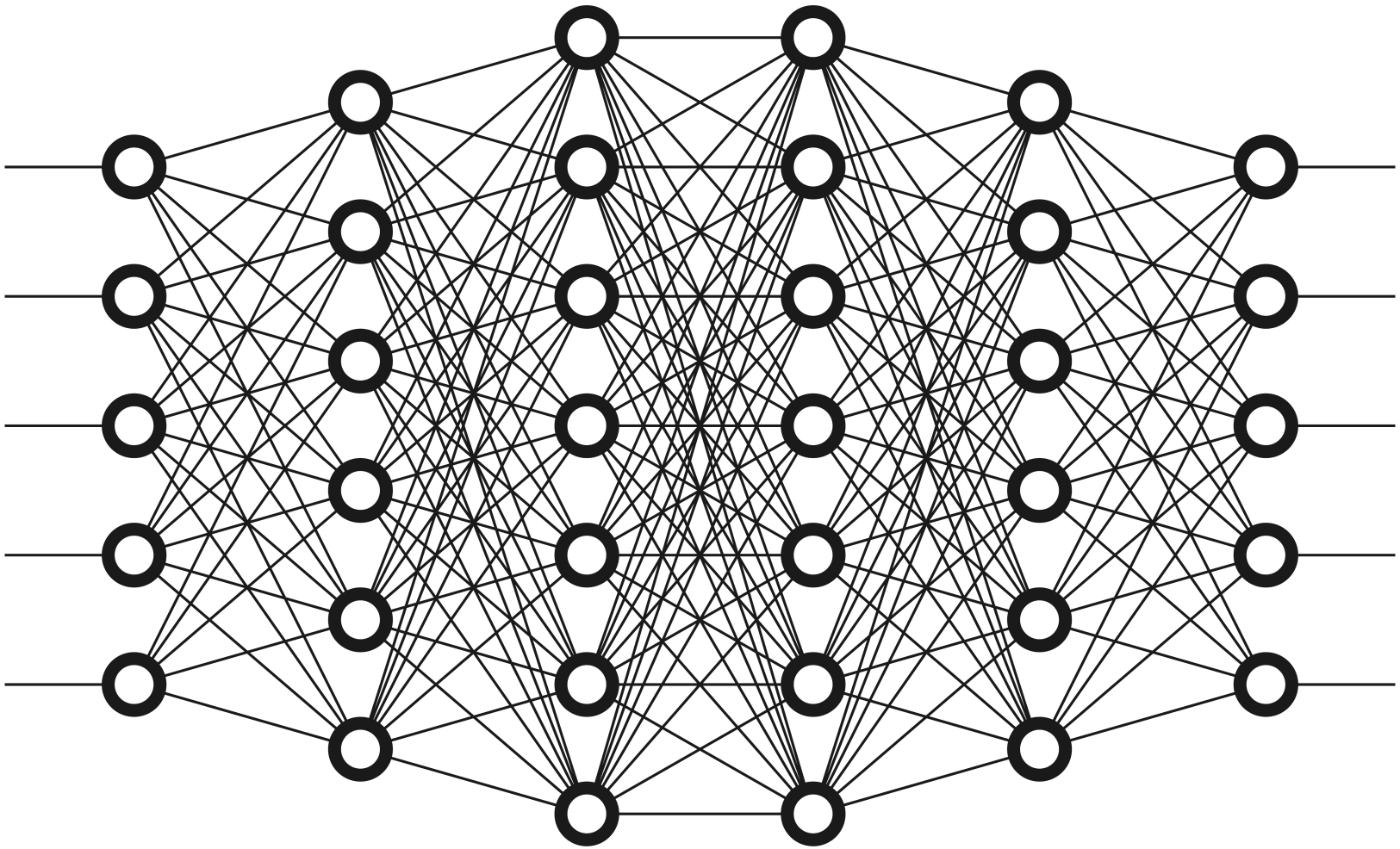




Simplification syntaxique

- Phrase d'origine :
  - Caïn, l'aîné, cultive la terre et Abel (étymologie : de l'hébreu "souffle", "vapeur", "existence précaire") garde le troupeau.
- Possibilités de substitution :
  1. Caïn, l'aîné, cultive la terre et Abel garde le troupeau.
  2. Caïn, l'aîné, cultive la terre. Abel garde le troupeau.
  3. Caïn, l'aîné, cultive la terre.
  4. Abel garde le troupeau.
  5. Caïn, l'aîné, cultive la terre. Abel (étymologie : de l'hébreu "souffle", "vapeur", "existence précaire") garde le troupeau.
  6. Abel (étymologie : de l'hébreu "souffle", "vapeur", "existence précaire") garde le troupeau.
- Simplification sélectionnée :
  - Caïn, l'aîné, cultive la terre. Abel garde le troupeau.

Réseaux de neurones



Simplification automatique de textes – Méthodes

- Modifier un texte pour le rendre accessible, tout en préservant son sens
- Principalement deux versants :
  - Simplification lexicale : remplacer les termes ou expressions complexes par des équivalents accessibles
  - Simplification syntaxique : retravailler les structures de phrases
- Schéma d'annotation pour le processus de simplification (Brunato *et al.*, 2014)
  - Découpage, fusion, réorganisation de phrases
  - Insertion / Effacement (verbes, sujets, autres)
  - Transformation (substitution lexicale, remplacement des anaphores, modification des traits verbaux...)
- Pas de consensus pour l'évaluation d'un système de simplification
- Nécessite des corpus parallèles avec différents degrés de complexité

Simplification automatique de textes – Corpus

- Corpus parallèles :
  - Anglais, Espagnol, Italien, Portugais du Brésil, Danois
    (Chandrasekar & Srinivas, 1997; Bott *et al.*, 2014; Brunato *et al.*, 2014; Caseli *et al.*, 2009; Klerke & Søgaard, 2012)
  - Pas toujours accessibles librement
- Corpus comparables :
  - Anglais : Simple English Wikipedia - English Wikipedia (SEW-EW), Newsela
  - Autres : Révisions dans Simple English Wikipedia, articles scientifiques ou romans simplifiés
- Pas de ressource en français

Le corpus CLEAR

- Corpus comparable médical en français, pour la simplification de textes biomédicaux
- Trois sous-corpus :
  - Articles encyclopédiques
  - Notices de médicaments
  - Articles scientifiques
- Paires de documents
  - Même sujet
  - Deux niveaux de complexité

Articles Encyclopédiques

- Publiés par la Fondation Wikimedia
- Wikipedia: s'adresse au grand public

  (2 186 891 tokens, 19 287 lemmes)
- Vikidia: s'adresse aux enfants de 8 à 13 ans

  (183 051 tokens, 3 117 lemmes)
- Rédigés indépendamment
- Articles du portail de la médecine
- 2*575 articles
- Exemples :
  - La luette ou uvule est un appendice conique situé au fond de la cavité buccale. La luette est un organe de 10 à 15 millimètres de long. Elle est constituée d'un tissu membraneux et musculaire.
  - La luette ou uvule est un appendice conique situé au fond de la bouche. C'est un organe fait de tissus membraneux et musculaires, d'environ 10 à 15mm de long, qui pend à la partie moyenne du voile du palais.

Notices de médicament

- Publiés par le Ministère de la Santé
- Résumés des caractéristiques produit (RCP), pour les praticiens

  (51 705 111 tokens, 43 515 lemmes)
- Notices, pour les patients

  (33 116 119 tokens, 25 725 lemmes)
- 2*11 800 notices
- Exemples :
  - hypersensibilité à l'huile de paraffine. / - ne pas utiliser chez les personnes présentant des difficultés de déglutition en raison du risque d'inhalation bronchique et de pneumopathie lipoïde.
  - si vous avez une allergie à l'huile de paraffine. / - ne pas utiliser chez les personnes présentant des difficultés pour avaler en raison du risque d'inhalation de la paraffine liquide qui entraîne une pneumopathie lipoïde.

Articles scientifiques

- Publiés par la Fondation Cochrane
- Revues de littérature médicale pour les praticiens

  (2 804 335 tokens, 11 558 lemmes)
- Versions simplifiées manuellement par la fondation pour le grand public

  (1 491 243 tokens, 7 567 lemmes)
- 2*3 815 résumés
- Exemples :
  - L'hématome aigu de l'oreille est une affection qui se caractérise par la formation d'une collection sanguine sous le périchondre du pavillon. Il est souvent provoqué par un traumatisme contondant.
  - L'hématome aigu de l'oreille est une affection qui se caractérise par la formation d'une collection sanguine dans le pavillon (oreille externe), souvent à la suite d'un traumatisme contondant.

Résumé

	docs	*tokens*$_{tech}$	*tokens*$_{acc}$	*total tokens*	*lemmes*$_{tech}$	*lemmes*$_{acc}$
Wiki	575*2	2,293,078	197,672	2,490,750	19 287	3 117
Médicaments	11,800*2	52,313,126	33,682,889	85,996,015	43 515	25 725
Cochrane	3,815*2	2,840,003	1,515,051	4,355,054	11 558	7 567
Total	16,190	57,446,207	35,395,612	92,841,819		

- 16 190 paires de documents
- Technique : 57M tokens
- Simple : 35M tokens
- Total : 92M tokens
- Plus grande diversité lexicale côté technique

Alignement automatique

- Objectif : constituer un corpus de phrases parallèles
- Alignement manuel : données d'entraînement
- Unité de traitement : paire de phrases
- Trop peu de données pour les réseaux de neurones
- Approche : tâche de classification binaire (paire alignée / non-alignée)
- Données :
  - paires alignées : résultat de l'alignement manuel
  - paires non alignées : appariement aléatoire hors paires alignées

Critères d'alignement

- Alignement :
  - Équivalence – sens identique ou presque identique :
    - Une gêne visuelle passagère peut être ressentie après instillation du collyre.
    - Une gêne visuelle passagère peut être ressentie après l'administration du collyre.
  - Inclusion – le sens d'une phrase se retrouve intégralement dans une autre :
    - La maladie de Charcot est l'autre nom de la sclérose latérale amyotrophique.
    - Il est le découvreur de la sclérose latérale amyotropique, ou maladie de Charcot, une maladie neurodégénérative.

Critères d'alignement

- Non-alignement :
  - Phrases identiques, ou qui varient seulement par la ponctuation ou les mots grammaticaux :
    - Effets sur l'aptitude à conduire des véhicules **ou** à utiliser des machines
    - Effets sur l'aptitude à conduire des véhicules **et** à utiliser des machines
  - Intersection – deux phrases qui partagent du sens mais qui apportent chacune une information qui leur est propre :
    - Une faiblesse musculaire (hypotonie axiale), des difficultés d'alimentation (troubles de la succion entrainant une faible prise de poids), une hyperexcitabilité, une agitation ou des tremblements peuvent survenir chez le nouveau-né, ces troubles étant réversibles.
    - Un traitement en fin de grossesse par benzodiazépines même à faibles doses, peut être responsable chez le nouveau-né de signes d'imprégnation tels qu'hypotonie axiale, troubles de la succion entraînant une faible prise de poids.

Critères d'alignement – Récapitulatif

- Alignement :
  - Équivalence – sens identique ou presque identique
  - Inclusion – Le sens d'une phrase se retrouve intégralement dans l'autre
- Non-alignement :
  - Phrases identiques
  - Phrases qui diffèrent seulement par la ponctuation ou les mots grammaticaux
  - Intersection – Les phrases partagent une partie de leur signification mais chacune apporte une information qui ne se retrouve pas dans l'autre
- **Format :**
  - Créer un fichier .ex et un fichier .gp (exemple : Hématome.ex et Hématome.gp). Pour chaque alignement, la phrase ex est à mettre dans le fichier .ex, et la phrase gp dans le fichier .gp, **au même numéro de ligne**

Alignement automatique

- Descripteurs des paires de phrases :
  - Nombre de mots communs hors stopwords
  - Nombre de stopwords communs
  - Pourcentage de mots d'une phrase présents dans l'autre
  - Différence de longueur entre les deux phrases (en mots)
  - Différence de longueur moyenne des mots entre les deux phrases
  - Nombre total de bigrammes et trigrammes communs (en caractères)
  - Mesures de similarité : cosine, Dice, Jaccard
  - Distance d'édition (Levenshtein) entre les deux phrases (caractères, et mots)

Arya D. J., Hiebert E. H. & Pearson P. D. (2011). The effects of syntactic and lexical complexity on the comprehension of elementary science texts. *Int Electronic Journal of Elementary Education*, **4**(1), 107–125.

Beigman Klebanov B., Knight K. & Marcu D. (2004). Text simplification for information-seeking applications. In R. Meersman & Z. Tari, Eds., *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*. Berlin, Heidelberg: Springer, LNCS vol 3290.

Biran O., Brody S. & Elhadad N. (2011). Putting it simply: a context-aware approach to lexical simplification. In *Annual Meeting of the Association for Computational Linguistics*.

Blake C., Kampov J., Orphanides A., West D. & Lown C. (2007). Query expansion, lexical simplification, and sentence selection strategies for multi-document summarization. In *DUC*.

Bosch S., Pretorius L. & Fleisch A. (2008). Experimental bootstrapping of morphological analysers for Nguni languages. *Nordic Journal of African Studies*, **17**(2), 66–88.

Bott S., Saggion H. & Mille S. (2014). Text simplification tools for spanish. In *LREC 2014*, p. 1–7.

Brunato D., Dell'Orletta F., Venturi G. & Montemagni S. (2014). Defining an annotation scheme with a view to automatic text simplification. In *CLICIT*, p. 87–92.

Caseli H. M., Pereira T. F., Specia L., Pardo T. A. S., Gasperin C. & Aluisio S. M. (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *CICLING*, p. 1–12.

Chandrasekar R. & Srinivas B. (1997). Automatic induction of rules for text simplification. *Knowledge Based Systems*, **10**(3), 183–190.

Chen P., Rochford J., Kennedy D. N., Djamasbi S., Fay P. & Scott W. (2016). Automatic text simplification for people with intellectual disabilities. In *AIST*, p. 1–9.

Côté R. (1996). *Répertoire d'anatomopathologie de la SNOMED internationale, v3.4*. Université de Sherbrooke, Sherbrooke, Québec.

De Belder J., Deschacht K. & Moens M.-F. (2010). Lexical simplification. In *ITEC*. 1-4.

Klerke S. & Søgaard A. (2012). DSim, a Danish parallel corpus for text simplification. In *LREC*, p. 4015–4018.

Paetzold G. H. & Specia L. (2016a). Benchmarking lexical simplification systems. In *LREC*, p. 3074–3080.

Paetzold G. H. & Specia L. (2016b). Unsupervised lexical simplification for non-native speakers. In *AAAI Conference on Artificial Intelligence*, p. 3761–3767.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M. & Duchesnay E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Sajous F. & Hathout N. (2015). GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of the eLex 2015 conference*, p. 405–426, Herstmonceux, England.

Shardlow M. (2014). A survey of automated text simplification. *Int J Advanced Computer Science and Applications*, **1**, 1–13.

Stymne S., Tiedemann J., Hardmeier C. & Nivre J. (2013). Statistical machine translation with readability constraints. In *NODALIDA*, p. 1–12.

Štajner S. & Popović M. (2016). Can text simplification help machine translation? *Baltic J. Modern Computing*, **4**(2), 230–242.

Vu T. T., Tran G. B. & Pham S. B. (2014). Learning to simplify children stories with limited data. In L. . Springer, Ed., *Intelligent Information and Database Systems*, p. 31–41.

Wubben S., van den Bosch A. & Krahmer E. (2012). Sentence simplification by monolingual machine translation. In *Annual Meeting of the Association for Computational Linguistics*, p. 1015–1024.