Le TAL pour l'assistance à la lecture : lisibilité et simplification automatique de textes

Thomas François[1]

(1) CENTAL, IL&C (Université Catholique de Louvain)

Séminaire du Cental

4 octobre 2013

Plan

Problèmes de lecture

La lecture reste un problème dans nos sociétés à haut niveau d'éducation :

- rapport récent de l'UE : en 2009, 19, 6% des jeunes de 15 ans sont des "low achievers" [De Coster et al., 2011, 22]
- [Richard et al., 1993] : 92 demandes d'allocation de chômage (personnes avec un faible niveau d'éducation), il manque moitié des infos, dont certaines cruciales.
- [Patel et al., 2002] : leurs sujets rencontrent des problèmes importants dans la compréhension des différentes étapes pour la bonne administration de médicaments.
- Sans compter les populations allophones, confrontées régulièrement à de l'écrit (cours, administration, web, etc.)

La lecture et la TAL

Le TAL peut intervenir à divers niveaux :

- Sélection automatique de matériaux pour l'éducation ;
- Conception automatique d'exercices de lecture ou de langue ;
- Intervention dans des logiciels d'aide à la lecture : adaptation à l'apprenant, mise à jour des textes, etc. ;
- Simplification automatisée de documents authentiques ;
- ...

Cette présentation se concentre sur deux dimensions : la lisibilité et la simplification de textes.

Plan

Qu'est-ce que la lisibilité ?

La lisibilité : définition

The sum total (including the interactions) of all those elements within a given piece of printed material that affect the success of a group of readers have with it. The success is the extent to which they understand it, read it at a optimal speed, and find it interesting.
[Dale and Chall, 1949, 1]

La **lisibilité** vise à modéliser les difficultés de textes à la lecture en référence à une population (apprenants de L2, illettrés, personnes avec des difficultés cognitives, etc.)

Les **formules de lisibilité** sont des modèles statistiques visant à associer des lecteurs à des textes de leur niveau.

Un exemple de formule classique

Un exemple de formule : [Flesch, 1948, 225] :

$$\text{Reading Ease} = 206,835 - 0,846\, wl - 1,015\, sl$$

où :

Reading Ease (RE) : un score compris entre 0 et 100

wl : nombre de syllables par 100 mots

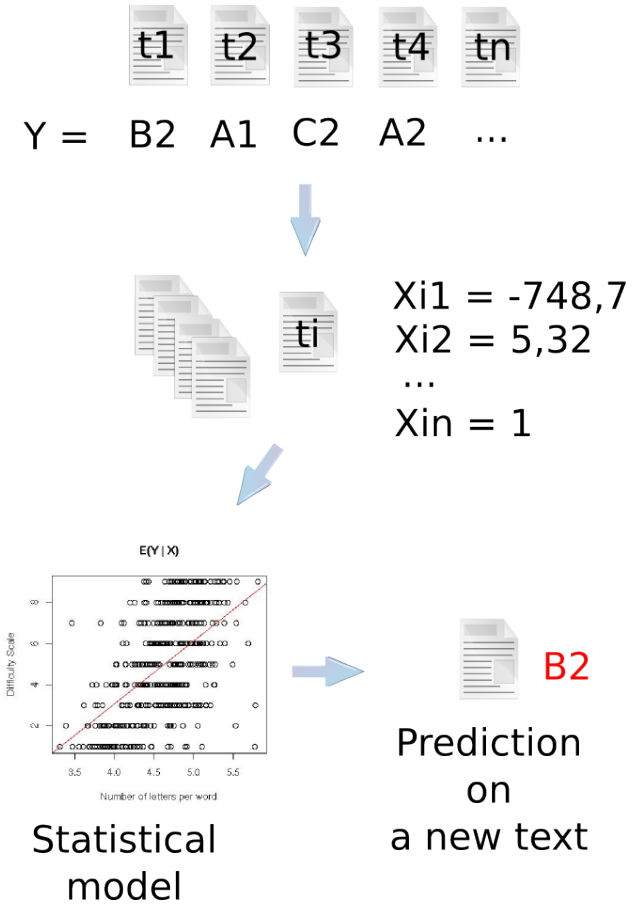
sl : nombre moyen de mots par phrase.

- Emploi de la régression linéaire et d'un nombre réduit de variables linguistiques (de surface).
- Flesch estime que sa formule peut s'appliquer à un large panel de situations.

Plan

La méthodologie pour la conception d'une formule

1. Rassembler un corpus de textes dont la difficulté a été mesurée à l'aide d'un critère tel que des tests de compréhension ou des tests de closure
2. Definir une liste de prédicteurs linguistiques de la difficulté, par ex. la longueur des phrases ou la charge lexicale
3. À partir de ces variables et du corpus, entraîner un modèle statistique (traditionnellement une régression linéaire)
4. Valider le modèle



Le corpus d'entraînement

- Critère = jugements d'experts = manuels !
  → Hypothèse : le niveau d'un texte = niveau du manuel dont il est tiré
- Nous avons extrait 2042 textes de 48 manuels de FLE, qui respectent l'échelle du CECR [Conseil de l'Europe, 2001]
  → Critères : public = adultes, langue moderne, pas de FOS + textes liés à une tâche de compréhension uniquement

L'échelle du CECR

C'est l'échelle officielle pour l'éducation en L2
Il y a 6 niveaux : A1 (plus facile), A2, B1, B2, C1, and C2 (plus difficile)

Les prédicteurs

Nous avons implémenté 406 variables, dont la plupart sont basées sur la littérature :

lexicaux : statistiques de la fréquence lexicale ; absents d'une liste de référence ; modèles n-grammes ; mesures de la diversité lexicale ; longueur des mots ; le voisinage orthographique

grammaticaux : longueur de la phrase ; ratio des catégories de discours ; type de verbes et de modes

sémantiques : taux d'abstraction et de personnalisation ; densité des idées ; taux de cohérence mesuré via LSA

spécifiques au FLE : présence de dialogue ; caractéristiques des unités polylexicales

Le modèle retenu

Il s'agit d'un classifieur par SVM, basé sur 46 variables, retenues via une analyse corrélationnelle.

	Modèle à 6 classes	**Modèle à 9 classes**
Algorithme	SVM « un contre un »	SVM « un contre un »
SVM-Type	C-classification	C-classification
Cost	5	15
Gamma	0,004	0,004
Nb. de vecteurs supports	±335	±500
Nb. de variables	46	46
Nb. de données d'entraînement	398	592
R	0,73	0,74
Exactitude	49% (9,7%)	35% (7,4%)
Exactitude contiguë	79,6% (5,4%)	65,4% (7,4%)
RMSE	1,27	1,92
MAE	0,9	1,15

Contribution de chaque famille de variables

Nous avons comparé des modèles incluant soit une seule famille de prédicteurs, soit toutes les variables sauf celles de cette famille :

	Famille seulement		**Tous sauf la famille**	
	Acc.	Adj. acc.	Acc.	Adj. acc.
Lexical	40.5	75.6	41.1	73.5
Syntaxique	39.3	69.5	43.2	78.4
Sémantique	28.8	61.5	47.8	79.2
FLE	24.9	58.5	47.8	79.6

Résultats

- les modèles basés sur les familles lexicale et syntaxique ont les meilleures performances et entraînent les pertes les plus significatives en exactitude.
- les variables lexicales sont les seules à réduire la quantité d'erreurs graves (*adj.acc.*).

Validation croisée de la formule

Nous avons rassemblé un autre corpus de FLE : des livres simplifiés
$\rightarrow$ textes narratifs surtout, non biaisés en fonction de la tâche.

Nous avons rassemblé 29 livres simplifiés :

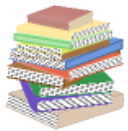
	A1	A2	B1	B2
nb. de livres	8	9	7	5
nb. de mots	41018	71563	73011	59051

Nous avons divisé les livres par chapitres et avons obtenu les données suivantes :

	A1	A2	B1	B2
nb. d'obs.	71	114	84	48
nb. de mots	41018	71528	73007	59051

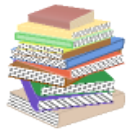
Expériences typologiques

"New" Corpus



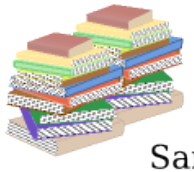
Corpus readers
A1 to B2 !!!

+



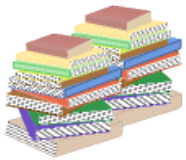
Corpus textbooks
C1 and C2

NB: sampling is different



Sampling (48/lev.)
Taking out outliers

Textbooks
(68/level)



Old Corpus

New Model

SVM(38) → 

R = 0.845; acc = 58.2%;
adj. acc. = 87.7%

SVM(41) → 

R = 0.85; acc. = 56%;
adj. acc. = 88%

SVM(41) ←

R = 0.72; acc. = 38%;
adj. acc = 81.7%

SVM(41) → 

R = 0.72; acc. = 48%;
adj. acc. = 77.9%

Not available: meanNGProb.G, NCPW, NAColl
Now constant: Infi (1) and med_nbNeighMoreFreq (0)

Old Model

Conclusions

1. Une nouvelle formule de lisibilité pour le FLE

- Nouvelle formule de lisibilité de FLE par SVM, avec 46 variables ;
- 1re formule de FLE à utiliser des variables fondées sur le TAL et de l'apprentissage automatique ;

2. Leçons et limitations

- l'apport des techniques d'apprentissage automatisé et celui du TAL reste flou (voir [François and Miltsakaki, 2012]).
- La formule ne semble pas se généraliser si bien que prédit par la procédure de validation croisée (influence du type de textes)
  → Nécessité de prévoir des formules spécialisées !

Plan

Objectifs et problématiques

- Les textes administratifs sont connus pour être difficile d'accès pour une proportion non négligeable de la population.
- Objectif : proposer une formule de lisibilité qui classe les textes administratifs sur une échelle de 1 (très facile) à 5 (très difficile).
- Problème principal : le corpus d'entraînement...
  - La technique dominante en lisibilité computationnelle = prendre des textes éducatifs, déjà annotés par les concepteurs des manuels
  - Il n'existe pas de ressources de ce type pour les textes administratifs !

Comment annoter lisibilité ?

Il existe plusieurs critères acceptés en lisibilité :

Avis d'experts : hétérogénéité, population pas testée, mais pratique
Critère principal en lisibilité

Test de compréhension : population testée, mais interaction entre questions et textes

Test de closure : test sur population, au niveau du mot, mais lien avec la compréhension douteux (redondance ?)

Vitesse de lecture :

- [Brown, 1952] compare les temps de lecture sur des textes difficile (306 mots/min.) et très difficile (235 mots/min).
- [Just and Carpenter, 1980] : temps de fixation occulaire d'un mot correspondrait au temps de traitement cognitif de celui-ci.

Avis de non experts : [van Oosten and Hoste, 2011] montrent que N (N > 10) non experts peuvent annoter aussi fiablement que des experts (jugements binaires).

Le temps de lecture comme critère : expérience

Temps de lecture est très peu utilisé et est pourtant le critère le plus fiable psychologiquement, en théorie.

Méthodologie

- 28 textes courts (100 mots) issus de livres simplifiés ont été sélectionnés pour les niveaux A1 à B2.
- Présentation phrase par phrase via un logiciel d'auto-présentation segmentée (Linger et Dmesure)
- Le temps passé sur chaque phrase est enregistré ; pas de retour en arrière possible.
- A la fin, il y a une ou deux questions de compréhension.
- Analyse des résultats avec un modèle à effets mixtes [Baayen et al., 2008] (supprime la variation inter-sujet)

Résultats

Linger			
	Min-Max RT/W	nb. sujets	Corr
Débutants II	$717ms - 78680ms$	9	$0,33$
Intermédiaire I	$747ms - 69250ms$	4	$0,32$

DMesure-Testing			
	Min-Max RT/W	nb. sujets	Corr
Débutants II	$562ms - 45351ms$	9	$0,07$
Intermédiaire I	$1296ms - 61770ms$	4	$0,29$
Natifs	$493ms - 33050ms$	6	$0,579$

On note que la méthode devient plus « fiable » en relation avec le niveau de compétence des lecteurs.

Interface web

Nous avons développé une interface web pour ce même test

Interface web (2)

Interface avec les questions (QCM)

Quelle technique d'annotation sur cette base ?

Annotation mixte : vitesse de lecture et jugements d'experts.

- 115 textes administratifs authentiques (FWB) sont numérisés (XML) et découpés en 220 fragments.
- Difficulté des fragments est évaluée via la formule de [Kandel and Moles, 1958], pour assurer une bonne représentatitivité.
- 10 textes de difficulté différentes ont été testés via AMesure-Testing $\rightarrow$ guide d'annotation
- Annotation manuelle par des experts de la FWB (α de Kripendorf = 0.37).

Au final, 115 fragments annotés en 5 niveaux de difficulté.

Conception de la formule (AMesure)

- Variables : adaptation des variables de la thèse (344). Les plus efficaces sont :
  - le nombre moyen de mots par phrase ($r > 0,64$)
  - le ratio de pronoms et de conjonctions, la proportion de mots > 8 lettres, fréquence cumulée des voisins orthographiques, taux de personnalisation des textes, proportion de participes passés, cohérence interphrastique, etc.
- Modélisation : application des SVMs avec sélection des variables sur la base d'une analyse corrélationnelle.

  $\rightarrow$ Modèle basé sur 11 variables : $acc = 50\%$ et $adj - acc = 86\%$ pour 5 niveaux.

Performances équivalentes au modèle pour le FLE, et formule adaptée à un domaine (malgré le peu de textes) !

Conclusions

- Nos travaux tendent à montrer l'intérêt de formules spécialisées (population, type de textes)
  → Application de la formule FLE sur un corpus de livres simplifiés = perte de 10% !
- Pour explorer cette question, il est vital de disposer d'un système d'annotations de textes, fiable et rapide.
  → Peu de travaux dans ce sens. Van Osten et al. suggèrent le crowd-sourcing avec des non-experts.
- Nous avons exploré une méthode alternative : la mesure du temps de lecture.

  → La fiabilité de la méthode n'est pas très élevée en FLE, mais l'est plus pour des natifs.

Notre étude montre la possibilité d'avoir une formule relativement fiable, spécialisée et à moindre coût.

Plan

La simplification automatique de texte (SAT)

Définition

Ensemble de techniques de TAL visant à rendre des textes plus faciles à lire, tout en garantissant l'intégrité de leur contenu et de leur structure.

Cela revient à :

- Identifier les termes complexes, les structures syntaxiques problématiques, etc.
- Distinguer les informations essentielles (à mettre en évidence) et les infos secondaires (à supprimer)
- Partie analyse de textes et (re)génération de textes

Travaux en simplification

Simplification comme prétraitement

- Pour améliorer la traduction automatique ou l'analyse syntaxique [Chandrasekar et al., 1996]
- Pour l'extraction de données [Lin and Wilbur, 2007]
- Pour améliorer la génération automatique de questions [Heilman and Smith, 2010]

Simplification pour les humains

- Pour personnes atteintes d'un handicap langagier [Inui et al., 2003, Carroll et al., 1999]
- Apprentissage d'une langue première [De Belder and Moens, 2010]
- Apprentissage d'une langue seconde [Siddharthan, 2006, Medero, 2011]

Travaux en simplification (2)

Symbolique VS statistique

- Règles de transformation définies manuellement
- Méthodes issues de la traduction automatique ou de l'apprentissage automatique [Zhu et al., 2010, Woodsend and Lapata, 2011]

Limitations

- Peu ou pas de prise en compte des aspects sémantiques, organisationnels [Siddharthan, 2006]
- Type de textes ou spécificités du public peu prises en compte
- Très peu de travaux pour le français

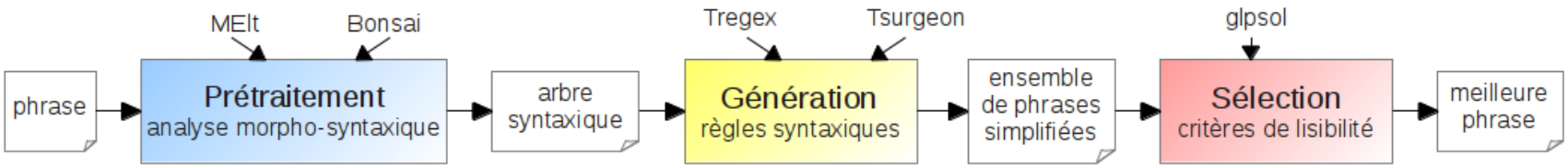
Simplificateur syntaxique

Approche pour le français, à destination des enfants : [Brouwers et al., 2012]

Objectifs :

- Etudier les simplifications sur la base d'un corpus "parallèle" : Wikipédia et Vikidia.
  $\rightarrow$ Définition d'une typologie, qui sert de base à des règles symboliques
- Système de simplification basé sur des contraintes (Programmation Linéaire en nombres entiers).

Le système



- 19 règles de simplifications syntaxiques implémentées : suppression (12), modification (3), division (4)
- Surgénération de phrases simplifiées par application récursives de toutes les règles
- Sélection de la meilleure via 4 contraintes : longueur phrase et mots, familiarité du lexique et présence de termes-clés

Exemples de simplification

Ex. d'erreurs

C'est à Aix qu'arriva en 802 l'éléphant blanc.
$\rightarrow$ En 802 arriva à Aix un éléphant blanc.

Arlette Laguiller, **née le 18 mars 1940 aà Paris**, est une femme politique française d'extrême gauche, appartenant au mouvement Lutte ouvrière.
$\rightarrow$ Arlette Laguiller est une femme politique française d'extrême-gauche, appartenant au mouvement Lutte ouvrière. Elle est née le 18 mars 1940 à Paris.

Pas de simplifications lexicales dans ce système.

Projet : un dictionnaire de synonymes gradués

Objectifs de l'étude

- Identifier les variables qui caractérisent les mots 'simples' et les paramétriser
  $\rightarrow$ On se base sur des ressources générales (Lexique 3) et des productions de patients atteints de la maladie de parkinson (Pk_corpus) (troubles du langage en découlent)
- Entraîner un modèle de difficulté pour le lexique
- Construire une ressource de mots (et de synonymes) gradués en fonction de la difficulté de leur forme (ReSyf)

Cette ressource pourra servir à sélectionner des alternatives plus simples à un mot difficile.

Variables intralexicales et psycholinguistiques

Variables intralexicales

- Nombre de lettres, de phonèmes et de syllables
- Classes de structures syllabiques (structures les plus fréquentes dans Pk_corpus : V, CVC, CV, CYV)
- Consistence entre la forme graphique et phonologique (0 = transparence, < 2 caractères, < 2 caractères)
- Patterns orthographiques complexes

Variables psycholinguistiques

- Fréquence lexicale (logarithme)
- Présence/absence de la liste de Gougenheim
- Voisins orthographiques

Entraînement du modèle sur Manulex
[Lètè et al., 2004]

Manulex, c'est...

- Une liste de 19 037 lemmes, dont les fréquences ont été calculées pour trois niveaux scolaires.

Word	**POS**	**1**	**2**	**3**
pomme	N	724	306	224
patriarche	N	-	-	1
cambrioleur	N	2	-	33

- Trois niveaux = CP, CE1, CE2 à CM2 (rassemblés en un)
- Nous avons transformé les fréquences en un niveau = première attestation à un niveau
- Répartition : 31% niveau 1, 21% niveau 2, 48% niveau 3

Premiers résultats

- Sélection des variables via une étude corrélationnelle sur les données de Manulex (Spearman)
  $\rightarrow$ Meilleures : log(frequences) (-0.51), presence dans Gougenheim (-0.41), nombres de phonèmes/lettres, voisins.
- Combinaison des 9 meilleures variables (sur 27) dans un modèle SVM
- Paramètres SVM obtenus par grid search : RBF kernel, $C = 1$, $\gamma = 0.5$

Exactitude de la classification : 62 % pour 3 classes (tâche difficile)

Plan

Conclusions

- Présentation de deux formules de lisibilité, un pour le FLE et un pour les textes administratifs.
  $\rightarrow$ résultats similaires et dans la lignée de la littérature (+/- $acc = 50\%$).
- Système de simplification syntaxique, précis, mais dont la couverture est limitée (approche symbolique)
- Méthodologie alternative d'annotation de la difficulté de textes (natifs)
  $\rightarrow$ permet d'envisager des approches spécialisées (public, type de textes, etc.) en lisibilité et en simplification ;

Perspectives

- Poursuivre l'extension du système de simplification syntaxique aux niveaux lexical et sémantique.
- Couverture élargie via étude de corpus sur les simplifications.
- Intégrer davantage la lisibilité et les systèmes de simplification.
- Proposer des systèmes de diagnostique plus précis sur les textes.

Merci

Difficulté estimée : A2

Votre texte : Merci pour votre attention.

Sachez que les questions
et les commentaires sont les bienvenus :-)

Merci à ...

Eleni Miltsakaki, Laetitia Brouwers, Núria Gala, Hubert Naets, Cédrick Fairon, Bernadette Dehottay...

References I


Baayen, R. H., Davidson, D. J., and Bates, D. (2008).
Mixed-effects modeling with crossed random effects for subjects and items.
Journal of memory and language, 59(4) :390–412.


Brouwers, L., Bernhard, D., Ligozat, A.-L., and François, T. (2012).
Simplification syntaxique de phrases pour le français.
In *Actes de la Conférence Conjointe JEP-TALN-RECITAL*, pages 211–224.


Brown, J. (1952).
The Flesch Formula 'Through the Looking Glass'.
College English, 13(7) :393–394.


Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., and Tait, J. (1999).
Simplifying text for language-impaired readers.
In *Proceedings of EACL 1999*, pages 269–270.


Chandrasekar, R., Doran, C., and Srinivas, B. (1996).
Motivations and methods for text simplification.
In *Proceedings of the 16th conference on Computational Linguistics*, volume 2, pages 1041–1044.

References II


Conseil de l'Europe (2001).
Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer.
Hatier, Paris.


Dale, E. and Chall, J. (1949).
The concept of readability.
Elementary English, 26(1) :19–26.


De Belder, J. and Moens, M.-F. (2010).
Text simplification for children.
In *Proceedings of the SIGIR workshop on accessible search systems*, pages 19–26.


De Coster, I., Baidak, N., Motiejunaite, A., and Noorani, S. (2011).
Teaching reading in europe : Contexts, policies and practices.
Technical report, Education, Audiovisual and Culture Executive Agency, European Commission.

References III


Flesch, R. (1948).
A new readability yardstick.
Journal of Applied Psychology, 32(3) :221–233.


François, T. and Miltsakaki, E. (2012).
Do NLP and machine learning improve traditional readability formulas ?
In *Proceedings of the 2012 Workshop on Predicting and improving text readability for target reader populations (PITR2012)*.


Heilman, M. and Smith, N. (2010).
Extracting simplified statements for factual question generation.
In *Proceedings of QG 2010 : The Third Workshop on Question Generation*, pages 11–20.


Inui, K., Fujita, A., Takahashi, T., Iida, R., and Iwakura, T. (2003).
Text simplification for reading assistance : a project note.
In *Proceedings of the second international workshop on Paraphrasing*, pages 9–16.

References IV


Just, M. and Carpenter, P. (1980).
A theory of reading : From eye fixations to comprehension.
Psychological review, 87(4) :329–354.


Kandel, L. and Moles, A. (1958).
Application de l'indice de Flesch à la langue française.
Cahiers Études de Radio-Télévision, 19 :253–274.


Lètè, B., Sprenger-Charolles, L., and Colè, P. (2004).
Manulex : A grade-level lexical database from French elementary-school readers.
Behavior Research Methods, Instruments and Computers, 36 :156–166.


Lin, J. and Wilbur, W. (2007).
Syntactic sentence compression in the biomedical domain : facilitating access to related articles.
Information Retrieval, 10(4-5) :393–414.


Medero, J.and Ostendorf, M. (2011).
Identifying targets for syntactic simplification.
In *Proceedings of the SLaTE 2011 workshop*.

References V


Patel, V., Branch, T., and Arocha, J. (2002).
Errors in interpreting quantities as procedures : The case of pharmaceutical labels.
International journal of medical informatics, 65(3) :193–211.


Richard, J., Barcenilla, J., Brie, B., Charmet, E., Clement, E., and Reynard, P. (1993).
Le traitement de documents administratifs par des populations de bas niveau de formation.
Le Travail Humain, 56(4) :345–367.


Siddharthan, A. (2006).
Syntactic simplification and text cohesion.
Research on Language and Computation, 4(1) :77–109.


van Oosten, P. and Hoste, V. (2011).
Readability Annotation : Replacing the Expert by the Crowd.
In *Sixth Workshop on Innovative Use of NLP for Building Educational Applications*.

References VI

Woodsend, K. and Lapata, M. (2011).
Learning to simplify sentences with quasi-synchronous grammar and integer programming.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 409–420. Association for Computational Linguistics.

Zhu, Z., Bernhard, D., and Gurevych, I. (2010).
A monolingual tree-based translation model for sentence simplification.
In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1353–1361. Association for Computational Linguistics.