

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



Thomas François¹



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

Séminaire de Linguistique

28 juin 2013



Plan

- 1 **Problématique**
- 2 Lisibilité
- 3 Une formule de lisibilité pour le FLE
 - Méthodologie de la recherche
 - Résultats
- 4 Une formule pour les textes administratifs
- 5 Simplification automatique de textes
- 6 Conclusions et perspectives

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

- 1 Problématique
- 2 **Lisibilité**
- 3 Une formule de lisibilité pour le FLE
 - Méthodologie de la recherche
 - Résultats
- 4 Une formule pour les textes administratifs
- 5 Simplification automatique de textes
- 6 Conclusions et perspectives

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.

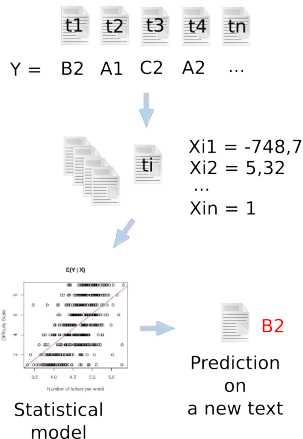
[Dale and Chall, 1949, 1]

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.

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 Définir 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



Un exemple de formule classique

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

$$\text{Reading Ease} = 206,835 - 0,846 w/l - 1,015 s/l$$

où :

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

w/l : nombre de syllables par 100 mots

s/l : 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.

Les grandes étapes en lisibilité

Il s'agit d'un domaine essentiellement anglo-saxon :

- Les formules classiques : elles sont basées sur la régression linéaire et 2 prédicteurs (un lexical, un syntaxique)
[Flesch, 1948, Dale and Chall, 1948, Kandel and Moles, 1958]
- La révolution de la closure : apparition de formules plus complexes et premières automatisations.
[Smith and Senter, 1967, Bormuth, 1966, Henry, 1975]
- Période structuro-cognitiviste : critique des formules classiques et mise en évidence des aspects organisationnels (cohérence, cohésion) ou cognitifs (inférence, densité conceptuelle).
[Kintsch and Vipond, 1979, Kemper, 1983]
→ Problème : L'approche structuro-cognitive est fondée sur de solides acquis théoriques, mais est très complexe à automatiser, pour des résultats similaires aux formules classiques ;
- La « lisibilité computationnelle » tente de synthétiser ces deux paradigmes.

La lisibilité computationnelle

La lisibilité computationnelle : définition

- Combine une extraction de variables via le TAL et des algorithmes statistiques avancés.
- Aborde la lisibilité plutôt comme un problème de classification.
- Premières études : mesure de la cohérence par LSA [Foltz et al., 1998] et premiers modèles de langue [Si and Callan, 2001].
- Première vague : modélisation de variables lexico-syntaxiques [Collins-Thompson and Callan, 2005, Schwarm and Ostendorf, 2005].
- Ensuite, apparaissent de sétudes qui prennent aussi en compte les variables sémantiques, cognitives et du discours [Crossley et al., 2007, Pitler and Nenkova, 2008, Feng et al., 2009]
- [Tanaka-Ishii et al., 2010] approchent la lisibilité comme un problème de tri (binaire) : bons résultats.
- [Vajjala and Meurers, 2012] introduisent des variables de SLA et obtiennent une bonne exactitude sur le Weekly Reader (93, 3%).

Plan

- 1 Problématique
- 2 Lisibilité
- 3 Une formule de lisibilité pour le FLE**
 - Méthodologie de la recherche
 - Résultats
- 4 Une formule pour les textes administratifs
- 5 Simplification automatique de textes
- 6 Conclusions et perspectives

Objectifs de cette recherche (thèse)

Premier objectif

Concevoir une formule de lisibilité pour le FLE, qui essaie de prendre en compte les spécificités de ce contexte.

Second objectif

Évaluer la pertinence de ce nouveau courant, la lisibilité computationnelle, pour la lisibilité du FLE.
Revient à se demander : est-ce que le TAL apporte quelque chose à la lisibilité du FLE ?

Plan

- 1 Problématique
- 2 Lisibilité
- 3 Une formule de lisibilité pour le FLE
 - Méthodologie de la recherche
 - Résultats
- 4 Une formule pour les textes administratifs
- 5 Simplification automatique de textes
- 6 Conclusions et perspectives

L'annotation du corpus

- Il faut rassembler des textes de FLE dont on connaît déjà le niveau de difficulté !
- Il faut que l'échelle de difficulté utilisée soit pratique pour les professionnels du FLE.
→ Par conséquent, nous avons extraits 2042 textes de 28 manuels de FLE, qui respectent l'échelle du CECR [Conseil de l'Europe, 2001].
- Critères : langue moderne, adultes, pas de FOS, tâches de compréhension à la lecture.

Hypothèse sous-jacente

Le niveau d'un texte sera considéré comme équivalent à celui du manuel dont il est issu.

Sélection des prédicteurs

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

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

grammaticaux : longueur de la phrase ; ratio des catégories de discours ;

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

spécifique au FLE : présence de dialogue.

Plusieurs d'entre-eux n'ont jamais été testés dans le contexte du FLE (en gras).

L'apport de la psychologie cognitive

La description des processus de lecture sous un angle psycholinguistique a permis de proposer de nouveaux prédicteurs :

lexicaux : le voisinage orthographique ; le TTR normalisé ; **nombre de sens d'un mot.**

grammaticaux : type de verbes et de modes ;

spécifique au FLE : caractéristiques des unités polylexicales, **étapes d'acquisition.**

Des pistes inexplorées subsistent à ce niveau (en gras)

Plan

- 1 Problématique
- 2 Lisibilité
- 3 Une formule de lisibilité pour le FLE**
 - Méthodologie de la recherche
 - **Résultats**
- 4 Une formule pour les textes administratifs
- 5 Simplification automatique de textes
- 6 Conclusions et perspectives

Deux étapes de résultats

Nos expérimentations ont été opérées en deux temps :

- 1 Identifier les bons prédicteurs employés seuls ;
- 2 Évaluer la capacité prédictive de combinaisons de variables (= formules).

En effet, deux bons prédicteurs peuvent apporter une information fortement redondante au modèle.

Results1

Principales variables d'intérêt

Label	Description de la variable	<i>r</i>	<i>ρ</i>	<i>F(p)</i>
Prédicteurs lexicaux				
PA_Alterego1a	Proportion d'absents de la liste de <i>Alter Ego 1</i>	/	0,65**	< 0,001
X90FFFC	90 ^e percentile de la probabilité des mots de contenu (formes fléchies)	/	-0,64**	< 0,001
PAGoug_2000	Proportion d'absents de la liste de Gougenheim (2000 premiers mots)	/	0,60**	0,017
ML3	Modèle unigramme (formes fléchies)	/	-0,55**	< 0,001
NLM	Nombre moyen de lettres par mot	0,48**	/	0,084
meanNGProb.G	Moyenne des probabilités des bigrammes de Google	0,38**	/	0,05
TTR	Type token ratio (basé sur les lemmes)	/	-0,26**	0,009
MedNeigh+Freq	Mediane du nombre de voisins plus fréquents	-0,22**	/	0,359
Prédicteurs syntaxiques				
NMP	Nombre moyen de mots par phrase	/	0,62**	0,014
LSDaoust	Pourcentage de phrases de plus de 30 mots	/	0,56**	0,008
PPres	Présence de participes présents	/	0,44**	0,003
PPasse	Présence de participes passés	/	0,39**	< 0,001
Pres_C	Proportion des verbes à l'indicatif présent	/	-0,34**	< 0,001
PRO.PRE	Rapport des pronoms aux prépositions	-0,18**	/	0,226
Prédicteurs sémantiques				
avLocalLsa_Lem	Cohésion interphrastiques moyenne	/	0,63**	0,01
PP1P2	Proportion des pron. pers. (1 ^{re} et 2 ^e)	/	-0,33**	0,008
Prédicteurs spécifiques au FLE				
BINGUI	Présence de guillemets de dialogue	/	0,46**	0,018
NAColl	Proportion d'UPs nominales de structure NA	/	0,29**	/

Résumé de l'analyse bivariable

- Chaque famille de prédicteurs comporte au moins une variable d'intérêt :
 - Proportion d'absents d'une liste tirée d'un manuel de FLE (*Alter Ego*) ;
 - Nombre de mots par phrase ;
 - Taux de cohérence moyen estimé par ASL ;
 - Le type de texte (dialogué ou non).
- idée : concevoir une formule avec ces prédicteurs ;
- Parmi celles-ci, les deux premiers sont classiques ;
- L'efficacité du premier montre l'intérêt d'adapter une formule à un contexte précis (liste spécifique au FLE) ;

Intérêt du TAL en lisibilité ?

- La présence de la LSA parmi les meilleurs prédicteurs (et ML3) confirme l'intérêt du TAL en lisibilité ;
- Mais, pas mal de variables TAL peu efficaces : n-grammes ! ; unités polylexicales.

Etape de modélisation

4 familles de modèles ont été comparées :

- Baseline (classique) : NMP + NLM ;
- 4 meilleurs prédicteurs/famille : PA_Alterego1a + NMP + avLocalLsa_Lem + BINGUI ;
- 8 meilleurs prédicteurs/famille : PA_Alterego1a + X90FFFC + NMP + PPres + avLocalLsa_Lem + PP1P2 + BINGUI + NAColl.
→ Hypothèse : prendre en compte le plus de dimensions textuelles ;
- Sélection automatisée de variables.

Algorithmes statistiques utilisés : SVM, régression logistique, Boosting, Bagging, et arbres de décision.

Résumé des résultats

Modèle	Classifieur	Paramètres	R	acc	acc – cont
Hasard	/	/	/	16,6%	44,4%
Baseline	SVM	$\gamma = 0,05; C = 25$	0,62	34%	68,2%
Expert1	RLM	/	0,70	39%	74,2%
Expert2	SVM	$\gamma = 0,002; C = 75$	0,73	41%	78%
Auto	SVM	$\gamma = 0,004; C = 5$	0,73	49%	79,6%
Modèle 2009	RLM	/	0,62	41%	71%
[Kandel and Moles, 1958]	RL	/	0,55	33%	/

Ex. de comparaison avec la littérature

[Schwarm and Ostendorf, 2005] : le gain au hasard (*acc – cont*) est de +24,5% à +29%, contre +36% en moyenne pour notre modèle.

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.*).

Results1

Exemple web (1)

Dmesure - Introduire un texte - Mozilla Firefox

http://cental.fltr.ucl.ac.be/team/tfrancois/dmesure_interface/html/introduireText.php

Dmesure - Introduire un texte

Rechercher un texte | Introduire un texte | Aide | Connexion

Dmesure

Dmesure vous offre également trois méthodes pour analyser directement un texte et en évaluer le niveau de difficulté à la lecture (sur l'échelle du CECR) pour un apprenant de français langue étrangère :

- Sélectionner la finesse de l'échelle : ☐ Echelle à 6 niveaux ☒ Echelle à 9 niveaux
- Copier-coller le texte dans le champ de saisie ci-dessous :

Sylvie est partie chez elle à cinq heures et demi avec son amie Nathalie. Les deux jeunes femmes sont allées au cinéma, à la séance de six heures, voir le film Un cœur en hiver de Claude Sautet. À huit heures, elle est sortie du cinéma et elles ont rencontré Philippe un architecte, ami de Nathalie. Tous les trois sont allés prendre un verre au café et, pendant une demi-heure, ils ont parlé de musique et de théâtre. Puis Philippe a proposé aux deux jeunes femmes de les accompagner chez elles en voiture. Il a d'abord déposé Nathalie. Ensuite, il est allé au 10 de la rue Chateaubriand où habite Sylvie. Mais Gérard, le mari de Sylvie, arrivait à ce moment-là. Il a vu sa femme descendre de la voiture de Philippe. La scène de jalousie n'a pas eu lieu parce que Sylvie a patiemment répondu aux questions de Gérard.

- Charger un texte depuis votre ordinateur : Parcourir...
- Indiquer l'URL d'un site internet :

Estimer la difficulté

Centre de traitement automatique du langage (CENTAL)
Collège Erasme, 1 place Blaise Pascal, B-1348 Louvain-la-Neuve (Belgique)

Exemple web (2)



Dmesure

Difficulté estimée : A2+ 

Votre texte : Sylvie est partie de chez elle à cinq heures et demi avec son amie Nathalie. Les deux jeunes femmes sont allées au cinéma, à la séance de six heures, voir le film Un coeur en hiver de Claude Sautet. À huit heures, elle sort du cinéma et elles ont rencontré Philippe un architecte, ami de Nathalie. Tous les trois sont allés prendre un verre au café et, pendant une demi-heure, ils ont parlé de musique et de théâtre. Puis Philippe a proposé aux deux jeunes femmes de les raccompagner chez elles en voiture. Il a d'abord déposé Nathalie. Ensuite, il est allé au 10 de la rue Chateaubriand où habite Sylvie. Mais Gérard, le mari de Sylvie, arrivait à ce moment-là. Il a vu sa femme descendre de la voiture de Philippe. La scène de jalousie n'a pas eu lieu parce que Sylvie a patiemment répondu aux questions de Gérard.

Conclusions

1. Une nouvelle formule de lisibilité pour le FLE

- Nouvelle formule de lisibilité de FLE par SVM, avec 46 variables ;
- 1^{re} formule de FLE au sein du paradigme de la lisibilité computationnelle ;

2. L'apport du TAL à la lisibilité du FLE

- Plusieurs variables « TAL » se sont révélées de bons prédicteurs employés seuls (LSA, unigramme, ratio cat. grammaticales, etc.) ;
- L'efficacité combinée est moindre (LSA n'est pas retenue) ;
→ certaines variables souffrent des approximations liées aux technologies TAL.
- l'apport des techniques d'apprentissage automatisé et celui du TAL reste flou, voir [François and Miltsakaki, 2012] pour une étude plus poussée de la question.

Plan

- 1 Problématique
- 2 Lisibilité
- 3 Une formule de lisibilité pour le FLE
 - Méthodologie de la recherche
 - Résultats
- 4 Une formule pour les textes administratifs
- 5 Simplification automatique de textes
- 6 Conclusions et perspectives

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 oculaire 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 fidèlement 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

[Home](#)

Logged as beber1

[Logout](#)

Dmesure - Testing

[Next sentence](#)

J'ai dans mes prisons beaucoup de gens du pays de Logres : des chevaliers, des dames, des jeunes filles.

Center for Natural Language Processing (CENTAL) at Louvain-la-Neuve
in collaboration with Choositol search and learn at Philadelphia.

Interface web (2)

Interface avec les questions (QCM)

[Home](#)Logged as beber1 | [Logout](#)

Dmesure - Testing

Question 1:

Quel marché propose Méléagant au roi Arthur ?

- ☐ Il veut combattre le chevalier le plus brave d'Arthur et, s'il gagne, prendre la reine Guenièvre avec lui.
- ☐ Il veut combattre le roi Arthur et, s'il perd, il rendra les prisonniers qu'il retient dans son royaume.
- ☐ Il veut échanger des prisonniers contre la reine.

[Check your answers](#)

Center for Natural Language Processing (CENTAL) at Louvain-la-Neuve
in collaboration with Choosito! search and learn at Philadelphia.

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 découper en 220 fragments.
- Difficulté des fragments est évaluée via la formule de [Kandel and Moles, 1958], pour assurer une bonne représentativité.
- 10 textes de difficulté différentes ont été testés via AMesure-Testing → guide d'annotation
- Annotation manuelle par des experts de la FWB (α de Krippendorff = 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 de mots moyen 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 corrélacionnelle.
 - Modèle basé sur 11 variables : $acc = 50\%$ et $adj - acc = 86\%$ pour 5 niveaux.

Performance meilleures que le 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

- 1 Problématique
- 2 Lisibilité
- 3 Une formule de lisibilité pour le FLE
 - Méthodologie de la recherche
 - Résultats
- 4 Une formule pour les textes administratifs
- 5 Simplification automatique de textes
- 6 Conclusions et perspectives

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

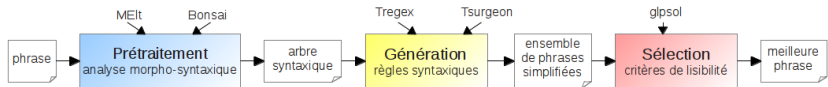
Simplificateur syntaxique

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.
→ 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.

→ En 802 arriva à Aix un éléphant blanc.

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

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

Evaluation

Evaluation manuelle sur 202 phrases :

- 113 sont simplifiées par le programme, ce qui donne 333 phrases produites
- Parmi celles-ci, 71 sont incorrectes, dont 63 à cause du prétraitement.
- Le simplificateur en lui-même ne fait donc que 2 à 3% d'erreurs.

Ex. d'erreurs

- C'est aussi depuis le XVIIIe siècle le terme en usage **pour désigner un clerc séculier ayant au moins reçu la tonsure.** (prop. inf.)
→ C'est aussi depuis le XVIIIe siècle le terme en usage.
- Ils ne sont pas caractérisés **par leur profession** comme dans la Bible : l'un pasteur, l'autre agriculteur. (agent)
→ Ils ne sont pas caractérisés : l'un pasteur, l'autre agriculteur.

Plan

- 1 Problématique
- 2 Lisibilité
- 3 Une formule de lisibilité pour le FLE
 - Méthodologie de la recherche
 - Résultats
- 4 Une formule pour les textes administratifs
- 5 Simplification automatique de textes
- 6 Conclusions et perspectives

Conclusions

- Présentation de deux formules de lisibilité, un pour le FLE et un pour les textes administratifs. // → résultats similaires et dans la ligné 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) // → permet d'envisager des approches spécialisées (public, type de textes, etc.) en lisibilité et en simplification ;

Perspectives

- Extension du système de simplification syntaxique aux niveaux lexical et sémantique.
- Couverture élargie via étude de corpus sur les simplifications (thèse qui débute).
- Intégrer davantage la lisibilité et les systèmes de simplification.
- Proposer des systèmes de diagnostic 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 :-)

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.



Bormuth, J. (1966).
Readability : A new approach.
Reading research quarterly, 1(3) :79–132.



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.

References II



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.



Collins-Thompson, K. and Callan, J. (2005).

Predicting reading difficulty with statistical language models.

Journal of the American Society for Information Science and Technology, 56(13) :1448–1462.



Conseil de l'Europe (2001).

Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer.

Hatier, Paris.



Crossley, S., Dufty, D., McCarthy, P., and McNamara, D. (2007).

Toward a new readability : A mixed model approach.

In *Proceedings of the 29th annual conference of the Cognitive Science Society*, pages 197–202.

References III



Dale, E. and Chall, J. (1948).
A formula for predicting readability.
Educational research bulletin, 27(1) :11–28.



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 IV



Feng, L., Elhadad, N., and Huenerfauth, M. (2009).
Cognitively motivated features for readability assessment.
In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pages 229–237.



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



Foltz, P., Kintsch, W., and Landauer, T. (1998).
The measurement of textual coherence with latent semantic analysis.
Discourse processes, 25(2) :285–307.



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).

References V



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.



Henry, G. (1975).

Comment mesurer la lisibilité.

Labor, Bruxelles.



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.



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.

References VI



Kemper, S. (1983).

Measuring the inference load of a text.

Journal of Educational Psychology, 75(3) :391–401.



Kintsch, W. and Vipond, D. (1979).

Reading comprehension and readability in educational practice and psychological theory.

In Nilsson, L., editor, *Perspectives on Memory Research*, pages 329–365.

Lawrence Erlbaum, Hillsdale, NJ.



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 VII



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.



Pitler, E. and Nenkova, A. (2008).

Revisiting readability : A unified framework for predicting text quality.

In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 186–195.



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.

References VIII



Schwarm, S. and Ostendorf, M. (2005).

Reading level assessment using support vector machines and statistical language models.

Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 523–530.



Si, L. and Callan, J. (2001).

A statistical model for scientific readability.

In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, pages 574–576. ACM New York, NY, USA.



Siddharthan, A. (2006).

Syntactic simplification and text cohesion.

Research on Language and Computation, 4(1) :77–109.



Smith, E. and Senter, R. (1967).

Automated Readability Index.

Technical report, AMRL-TR-66-220, Aerospace Medical Research Laboratories, Wright-Patterson Airforce Base, OH.

References IX



Tanaka-Ishii, K., Tezuka, S., and Terada, H. (2010).

Sorting texts by readability.

Computational Linguistics, 36(2) :203–227.



Vajjala, S. and Meurers, D. (2012).

On improving the accuracy of readability classification using insights from second language acquisition.

In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pages 163–173.



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.



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.

References X



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.