An analysis of a French as a Foreign Language Corpus for Readability Assessment

Thomas François

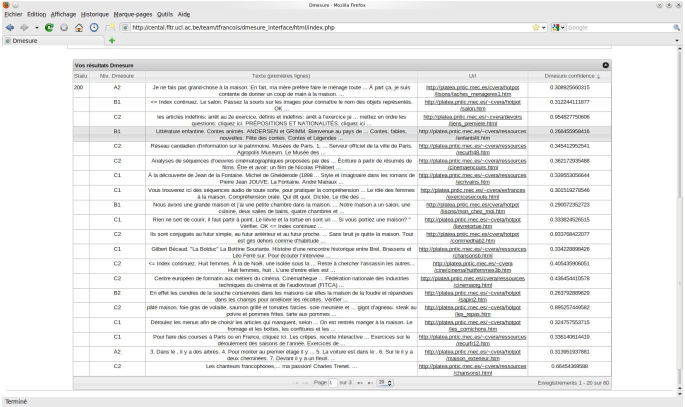
NLP4CALL, Uppsala

13-11-2014

Context

Readability models have roles to play in iCALL:

1. Find educational material of a given level, on the web


Dmesure - Mozilla Firefox
Fichier Édition Affichage Historique Marque-pages Outils Aide
http://cental.fltr.ucl.ac.be/team/tfrancois/dmesure_interface/html/index.php
Google
Dmesure
Terminé

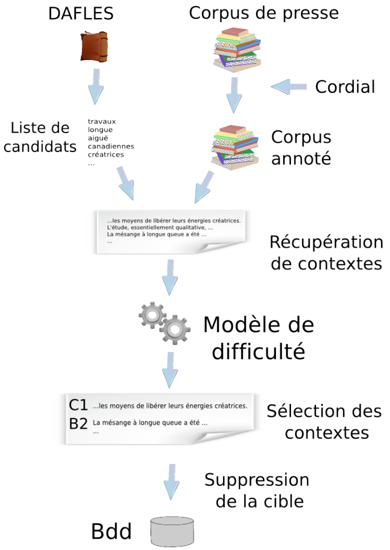
Vos résultats Dmesure

Statu	Niv. Dmesure	Texte (premières lignes)	Url	Dmesure confidence
200	A2	Je ne fais pas grand-chose à la maison. En fait, ma mère préfère faire le ménage toute ... À part ça, je suis contente de donner un coup de main à la maison. ...	http://platea.pntic.mec.es/cvera/hotpot/lisons/taches_menageres1.htm	0.308925660315
	B1	<= Index continuez. Le salon. Passez la souris sur les images pour connaître le nom des objets représentés. OK ...	http://platea.pntic.mec.es/~cvera/hotpot/salon.htm	0.312244111877
	C2	les articles indéfinis: arrêt au 2e exercice. définis et indéfinis: arrêt à l'exercice je ... mettez en ordre les questions: cliquez ici. PRÉPOSITIONS ET NATIONALITÉS. cliquez ici ...	http://platea.pntic.mec.es/~cvera/devoirs/liens_premiere.html	0.954827750606
	B1	Littérature enfantine. Contes animés. ANDERSEN et GRIMM. Bienvenue au pays de ... Contes, fables, nouvelles. Fête des contes. Contes et Légendes ...	http://platea.pntic.mec.es/~cvera/ressources/enfantslit.htm	0.266455958416
	C2	Réseau candadien d'information sur le patrimoine. Musées de Paris. 1, ... Serveur officiel de la ville de Paris. Agropolis Museum. Le Musée des ...	http://platea.pntic.mec.es/~cvera/ressources/recurfr48.htm	0.345412952541
	C2	Analyses de séquences d'oeuvres cinématographiques proposées par des ... Écriture à partir de résumés de films. Être et avoir: un film de Nicolas Philibert ...	http://platea.pntic.mec.es/~cvera/ressources/cinemaencours.html	0.362172935488
	C1	À la découverte de Jean de la Fontaine. Michel de Ghelderode (1898 ... Style et Imaginaire dans les romans de Pierre Jean JOUVE. La Fontaine. André Malraux ...	http://platea.pntic.mec.es/~cvera/ressources/ecrivains.htm	0.339553056644
	C1	Vous trouverez ici des séquences audio de toute sorte, pour pratiquer la compréhension ... Le rôle des femmes à la maison. Compréhension orale. Qui dit quoi. Dictée. Le rôle des ...	http://platea.pntic.mec.es/~cvera/exfrances/exercicesecoute.html	0.301519278546
	B1	Nous avons une grande maison et j'ai une petite chambre dans la maison. ... Notre maison a un salon, une cuisine, deux salles de bains, quatre chambres et ...	http://platea.pntic.mec.es/cvera/hotpot/lisons/mon_chez_moi.htm	0.290072352723
	C1	Rien ne sert de courir, il faut partir à point. Le lièvre et la tortue en sont un ... Si vous portiez une maison? " Vérifier. OK <= Index continuez ...	http://platea.pntic.mec.es/~cvera/hotpot/lievretortue.htm	0.333824526515
	C2	Ils sont conjugués au futur simple, au futur antérieur et au futur proche. ... Sans bruit je quitte la maison. Tout est gris dehors comme d'habitude ...	http://platea.pntic.mec.es/cvera/hotpot/commedhab2.htm	0.933768422077
	C1	Gilbert Bécaud. "La Bolduc" La Bottine Souriante. Histoire d'une rencontre historique entre Brel, Brassens et Léo Ferré sur. Pour écouter l'interview ...	http://platea.pntic.mec.es/~cvera/ressources/chansonsb.html	0.334228898426
	C2	<= Index continuez. Huit femmes. À la de Noël, une isolée sous la ... Reste à chercher l'assassin les autres... Huit femmes, huit . L'une d'entre elles est ...	http://platea.pntic.mec.es/~cvera/cine/cinema/huitfemmes3b.htm	0.405435906051
	C2	Centre européen de formatin aux métiers du cinéma. Cinémathèque ... Fédération nationale des industries techniques du cinéma et de l'audiovisuel (FITCA) ...	http://platea.pntic.mec.es/cvera/ressources/cinemaorg.html	0.436454410578
	B2	En effet les cendres de la souche conservées dans les maisons car elles la maison de la foudre et répandues dans les champs pour améliorer les récoltes. Vérifier ...	http://platea.pntic.mec.es/~cvera/hotpot/sapin2.htm	0.263792889629
	C2	pâté maison. foie gras de volaille. saumon grillé et tomates farcies. sole meunière et ... gigot d'agneau. steak au poivre et pommes frites. tarte aux pommes ...	http://platea.pntic.mec.es/~cvera/hotpot/les_repas.htm	0.895257449582
	C1	Déroulez les menus afin de choisir les articles qui manquent, selon ... On est rentrés manger à la maison. Le fromage et les boîtes, les confitures et les ...	http://platea.pntic.mec.es/~cvera/hotpot/les_cornichons.htm	0.324757553715
	C1	Pour faire des courses à Paris ou en France, cliquez ici. Les crêpes, recette interactive ... Exercices sur le déroulement des saisons de l'année. Exercices de ...	http://platea.pntic.mec.es/~cvera/ressources/recurfr12.htm	0.336140614419
	A2	3. Dans le , il y a des arbres. 4. Pour monter au premier étage il y ... 5. La voiture est dans le . 6. Sur le il y a deux cheminées. 7. Devant il y a un fleuri. ...	http://platea.pntic.mec.es/~cvera/hotpot/maison_exterieur.htm	0.313951937881
	C2	Les chanteurs francophones,... ma passion! Charles Trenet. ...	http://platea.pntic.mec.es/~cvera/ressources/chansonst.html	0.66454369588

Page 1 sur 3 · 20 · Enregistrements 1 - 20 sur 60

Context

2. Enhance systems for automatic exercise generation (proposal based on ALFALEX) [Verlinde et al., 2003]



DAFLES

Corpus de presse

Cordial

Liste de candidats

travaux
longue
aiguë
canadiennes
créatrices
...

Corpus annoté

...les moyens de libérer leurs énergies créatrices.
L'étude, essentiellement qualitative, ...
La mésange à longue queue a été ...
...

Récupération de contextes

Modèle de difficulté

C1 ...les moyens de libérer leurs énergies créatrices.
B2 La mésange à longue queue a été ...
...

Sélection des contextes

Suppression de la cible

Bdd

Problematic

- Many formulas “available” for L1, especially for English
  [Collins-Thompson and Callan, 2005, Feng et al., 2010, Vajjala and Meurers, 2012]
- Limited amount of models “available“ for L2
  [Heilman et al., 2007, Schwarm and Ostendorf, 2005]
- AND only two formulas use the CEFR scale (current standard)
  [François and Fairon, 2012, Pilán et al., 2014]

- Problem: Large amount of efforts required to collect the annotated data to train a readability model!

Objectives of the paper

1. Gathering a corpus of texts annotated for difficulty, as large as possible
    - Critical review of 5 common annotation approaches in readability
    - We opt for the extraction of texts from textbook series (with some conditions) and giving the textbook level to all texts extracted from it.
    - We report the collect process of about 2,000 texts for French as a Foreign Language (FFL)
2. Investigate the shortcomings of this criterion (which is mainstream in the field)
    - Produce unbalanced corpus
    - Homogeneity of annotations across textbook series is questionable

2 experiments

Class imbalanced effect

- We compared the results of models either based on a balanced or an unbalanced corpus
- The ordinal logistic models used includes two well-acknowledged features (mean number of words/sentence and of letters/word)
- It confirmed that majority class may deform the prediction space

Homogeneity of the annotations

- We compared the lexical and syntactic difficulty of textbook series among them (for each levels)
- We used ANOVA and MANOVA and noticed significant homogeneity issues in the corpus
- A relation between the type of pedagogical approach and homogeneity issues also appeared.

Take-home message: Test the homogeneity of annotations when using texts from educational sources (just as we would do for human annotators).

References I

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.

Feng, L., Jansche, M., Huenerfauth, M., and Elhadad, N. (2010).
A Comparison of Features for Automatic Readability Assessment.
In *COLING 2010: Poster Volume*, pages 276–284.

François, T. and Fairon, C. (2012).
An ”AI readability” formula for French as a foreign language.
In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP 2012)*, pages 466–477.

Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007).
Combining lexical and grammatical features to improve readability measures for first and second language texts.
In *Proceedings of NAACL HLT*, pages 460–467.

Pilán, I., Volodina, E., and Johansson, R. (2014).
Rule-based and machine learning approaches for second language sentence-level readability.
In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 174–184.

References II

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.

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.

Verlinde, S., Selva, T., and Binon, J. (2003).
Alfalex: un environnement d'apprentisage du vocabulaire français en ligne, interactif et automatisé.
Romaneske, 28(1):42–62.