

# Dmesure: a readability platform for French as a foreign language



Thomas François<sup>1, 2</sup> and Hubert Naets<sup>2</sup>



(1) Aspirant F.N.R.S.

(2) CENTAL, Université Catholique de Louvain

Presentation at CLIN 21

February 11, 2011



# Plan

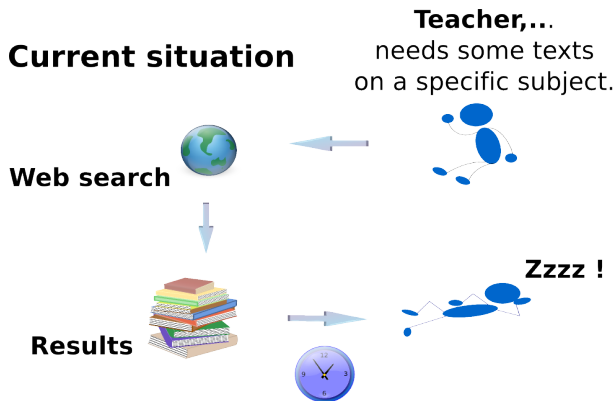
- 1 Introduction : the issue of finding texts
- 2 Current work for English
- 3 Dmesure : a web tool for FFL readability
- 4 Issues and perspectives with Dmesure
- 5 References

# Plan

- 1 Introduction : the issue of finding texts
- 2 Current work for English
- 3 Dmesure : a web tool for FFL readability
  - The one-text interface
  - Dmesure as a web crawler
  - Dmesure as a collaborative platform
- 4 Issues and perspectives with Dmesure
- 5 References

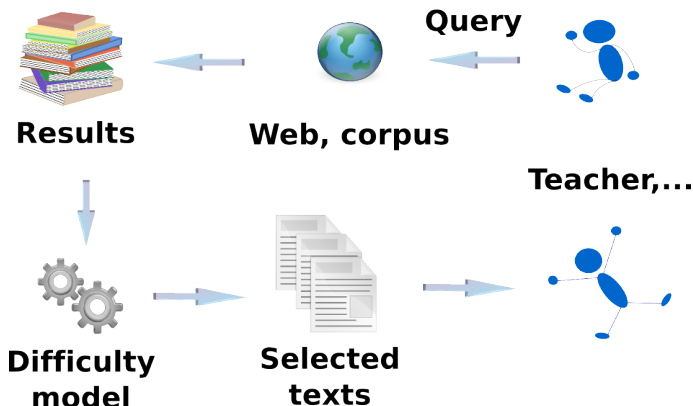
# Retrieval of web texts for FFL

Beyond search engines, there is no tool to find FFL texts at a specific level of difficulty.



# A solution : a difficulty model as a filter

## Improvement



# What is a difficulty model for reading ?

We consider that **readability formulas** are valid models of the reading difficulties in a L2.

## What is readability ?

*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.*

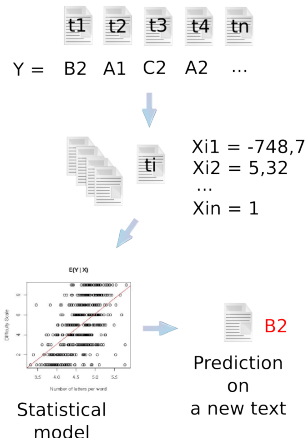
[Dale and Chall, 1949, 1]

Some of the well-known formulas :

[Flesch, 1948, Dale and Chall, 1948, Kincaid et al., 1975]

# Conception of a formula : methodological steps

- 1 Collect a corpus of texts whose difficulty has been measured using a criterion such as comprehension tests or cloze tests
- 2 Define a list of linguistic predictors of the difficulty, such as sentence length or lexical load
- 3 Design a statistical model (traditionally linear regression) based on the above features and corpus
- 4 Validate the model



# Readability : an example

## Grammar-based Reading Difficulty Prediction

**Grade level predicted: 12.0**

Accuracy generally improves with text length. The software will provide estimates for texts of any length, but a minimum length of 30 words is recommended. Also, the system is generally more accurate for grade levels above 2.

Type or paste your text into the box below and press "Submit" to obtain an estimate of the difficulty of your text.

A narrow grave-yard in the heart of a bustling, indifferent city, seen from the windows of a gloomy-looking inn, is at no time an object of enlivening suggestion; and the spectacle is not at its best when the mouldy tombstones and funereal umbrage have received the ineffectual refreshment of a dull, moist snow-fall. If, while the air is thickened by this frosty drizzle, the calendar should happen to indicate that the blessed vernal season is already six weeks old, it will be admitted that no depressing influence is absent from the scene.

Submit

An estimation of the readability of the first lines of *The Europeans* (H.James). It has been assessed by the model of [Heilman et al., 2007].

Url : <http://boston.lti.cs.cmu.edu/demos/readability/index.php>



# Plan

- 1 Introduction : the issue of finding texts
- 2 Current work for English**
- 3 Dmesure : a web tool for FFL readability
  - The one-text interface
  - Dmesure as a web crawler
  - Dmesure as a collaborative platform
- 4 Issues and perspectives with Dmesure
- 5 References

# Existing platforms

Web crawlers for the retrieval of web texts on a specific topic and at a specific readability level have been designed for English

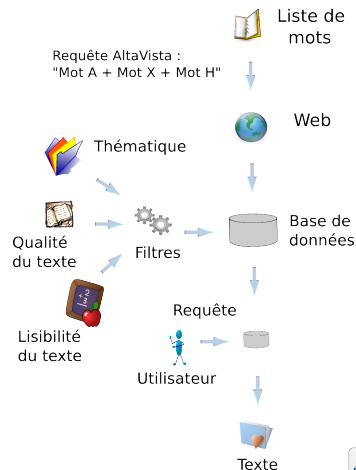
- **IR4LL** [Ott, 2009] ;
- **REAP** [Heilman et al., 2008b] ;
- **READ-X** [Miltsakaki and Troutt, 2008].

# Retrieval of web texts : an example for EFL

## ● REAP

[Heilman et al., 2008b,  
Collins-Thompson and Callan, 2004]

- REAding-specific Practice aims at improving reading comprehension abilities through practice.
- It integrates a SVM thematic classifier
- Difficulty is checked using the readability formulas described in [Collins-Thompson and Callan, 2005, Heilman et al., 2008a]
- <http://reap.cs.cmu.edu/>



# Plan

- 1 Introduction : the issue of finding texts
- 2 Current work for English
- 3 **Dmesure : a web tool for FFL readability**
  - The one-text interface
  - Dmesure as a web crawler
  - Dmesure as a collaborative platform
- 4 Issues and perspectives with Dmesure
- 5 References

# Dmesure : 3 goals

Dmesure (stands for Difficulté Mesure) aims at the 3 following objectives :

- 1 Makes available our formula for FFL [François, 2009] through a copy-cut interface.
- 2 Provides a free tool to help FFL teachers in the use of the web as a corpus for finding teaching materials
- 3 Provides a collaborative web platform where teachers can participate in assessing the difficulty of texts they collected through Dmesure and they used in their teaching

# Plan

- 1 Introduction : the issue of finding texts
- 2 Current work for English
- 3 **Dmesure : a web tool for FFL readability**
  - **The one-text interface**
    - Dmesure as a web crawler
    - Dmesure as a collaborative platform
- 4 Issues and perspectives with Dmesure
- 5 References

## The one-text interface

# Dmesure : the one-text interface

Dmesure - Introduire un texte - Mozilla Firefox

http://cental.ftr.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 : ☐ Échelle à 6 niveaux ☒ Échelle à 9 niveaux
- Copier-coller le texte dans le champ de saisie ci-dessous :

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 sont sorties 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.

- 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)

Terminé

This text comes from the textbook Panorama 2 (A2, p.26)

The one-text interface

# Dmesure : the one-text interface



## 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 cœur 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.

Terminé

The model did quite well on this one !!

Centre de traitement automatique du langage (CENTAL)  
Collège Erasme, 1 place Blaise Pascal, B-1348 Louvain-la-Neuve (Belgique)  
Contact : [dmesure@uclouvain.be](mailto:dmesure@uclouvain.be)





# The readability formulas used

The readability formulas used in Dmesure are variations of those presented in [François, 2009].

## 9-classes model characteristics

### Features :

- An unigram model based on inflected forms disambiguated using TreeTagger [Schmid, 1994] ;
- Mean number of words per sentence ;
- Proportion of personal pronouns of dialogue (1P, 2P), based on [Henry, 1975]
- 5 tense variables (binary) : Conditional & Future & Imperfect & Past part. & Subjunctive pres.

**Algorithm** : ordinal logistic regression [Agresti, 2002]

**Performance** :  $R^2 = 0,57$  (computed on 100 .632 bootstrap samples)

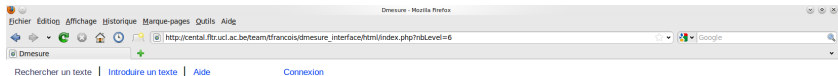


# Plan

- 1 Introduction : the issue of finding texts
- 2 Current work for English
- 3 **Dmesure : a web tool for FFL readability**
  - The one-text interface
  - **Dmesure as a web crawler**
  - Dmesure as a collaborative platform
- 4 Issues and perspectives with Dmesure
- 5 References

Dmesure as a web crawler

# The basic search interface



**Termes de recherche :**  [Recherche avancée](#)

**Difficulté du texte :** ☐ A1 ☐ A2 ☐ B1 ☐ B2 ☒ C1 ☒ C2 [Utiliser 9 niveaux](#)

# Results of previous request

Dmesure - Mozilla Firefox

http://cental.ftr.ucl.ac.be/team/francois/dmesure\_interface/ftrm/index.php

Dmesure

**Termes de recherche :**  [Recherche avancée](#)

**Difficulté du texte :** ☐ A1 ☐ A2 ☐ B1 ☐ B2 ☐ C1 ☐ C2 [Utiliser 9 niveaux](#)

**Vos résultats Dmesure**

Stat	Niv. Dmesure	Texte (premières lignes)	Url	Dmesure confidence
200	C2	La situation politique belge se dégrade. L'implosion du pays, certains dans son principe, mais incertaine quant à sa date de réalisation, inquiète ...	<a href="http://pms-europe.typepad.com/jms/2010/04/la-contestation-est-rachete-bruxelles.html">http://pms-europe.typepad.com/jms/2010/04/la-contestation-est-rachete-bruxelles.html</a>	0.904638894245
	C2	Situation politique - Informateur. Situation politique - Informateur. 17/06/2010. Voir aussi : ... savoir plus. Voir aussi : Le Roi. Lire l'entière © La Monarchie belge ...	<a href="http://www.monarchie.be/factuel/agenda/archives/situation-politique-informateur">http://www.monarchie.be/factuel/agenda/archives/situation-politique-informateur</a>	0.999999986765
	C2	La classe politique belge est divisée autour des propos de Karel de Gucht, ... le Premier ministre belge qui se dit préoccupé par la situation en RDC, où il ...	<a href="http://www.digitalcongo.net/article/51134">http://www.digitalcongo.net/article/51134</a>	0.809996125391
	C2	BELGIQUE : séjour, vacances, voyage. Climat de la Belgique, Belges, histoire belge, situation politique belge, situation économique de la Belgique, culture belge ...	<a href="http://www.vacances-sejour.ch/belgique/">http://www.vacances-sejour.ch/belgique/</a>	0.494059434514
	C2	Petit clin d'oeil sympathique envers la situation chaotique que traverse la Belgique en ... de Merkostas contre la politique d'immigration du gouvernement belge qui enferme des ...	<a href="http://www.wtbo.fr/international/europe/belgique/politique_belge/gouvernement_belge">http://www.wtbo.fr/international/europe/belgique/politique_belge/gouvernement_belge</a>	0.999015288525
	C2	Reste ça n'est en rien comparable avec la situation du militaire Belge ci dessus. ... resté la dernière à intervenir dans la situation politique Belge. ...	<a href="http://francoisquinqua.blog.lemonde.fr/2010/10/29/un-militaire-belge-detaile-le-ges-annonce-la-remise-don-de-ses-armes-video/">http://francoisquinqua.blog.lemonde.fr/2010/10/29/un-militaire-belge-detaile-le-ges-annonce-la-remise-don-de-ses-armes-video/</a>	0.957565870076
	C2	La FOTB fédérale «ses ailes wallonne, flamande et bruxelloise- est particulièrement inquiète de la situation politique belge et de ses potentielles conséquences ...	<a href="http://www.ps.be/Source/2PageContent.aspx?ContentID=5846&amp;MailID=241754&amp;EntID=1">http://www.ps.be/Source/2PageContent.aspx?ContentID=5846&amp;MailID=241754&amp;EntID=1</a>	0.963980073398
	C2	Accueil :: Bhw-attitude: un geste de contestation face à la situation politique belge. Bhw-attitude: un geste de contestation face à ...	<a href="http://www.fullymag.com/a2774.html">http://www.fullymag.com/a2774.html</a>	0.908877301847
	C2	Mise en garde : Les informations reprises dans les communiqués n'ont pas été ... BHW-attitude: un geste de contestation face à la situation politique belge ...	<a href="http://www.categorynet.com/communiqués-de-presse/politique/bhw/">http://www.categorynet.com/communiqués-de-presse/politique/bhw/</a>	0.99925781852
	C2	Source: RTBFRTBF Accroche: La situation politique belge paraît vitiée. Chacun campe sur ses certitudes. Côté flamand, le rapport unilatéral de Bart ...	<a href="http://www.rtf.be/lebelgique/politique/bart-de-voyeur-mene-le-bol-266634">http://www.rtf.be/lebelgique/politique/bart-de-voyeur-mene-le-bol-266634</a>	0.9999998095

Page 1 sur 6

Enregistrements 1 - 10 sur 60

Centre de traitement automatique du langage (CENTAL)  
 Collège Erasme, 1 place Blaise Pascal, B-1348 Louvain-la-Neuve (Belgique)  
 Contact : [dmesure@uclouvain.be](mailto:dmesure@uclouvain.be)

Terminé



Dmesure as a web crawler

# The advance search interface

Dmesure - Mozilla Firefox

http://central.fltr.ucl.ac.be/team/francois/dmesure\_interface/html/index.php?avancee=yes&nbLevel=6

Dmesure

[Rechercher un texte](#) | [Introduire un texte](#) | [Aide](#) | [Connexion](#)

## Dmesure

### Recherche avancée

[Rechercher par termes via BOSS...](#)

**Tous les termes suivants :**  [Recherche simple](#)

**L'expression suivante :**

**Aucun des termes suivants :**

[Autres types de recherche](#)

**Explorer un site :**  ⓘ

[Filtres](#)

**Difficulté du texte :** ☒ A1 ☒ A2 ☒ B1 ☒ B2 ☒ C1 ☒ C2 ⓘ [Utiliser 9 niveaux](#)

**Filtre contenu adulte :** ☐ Filtre désactivé ☒ Filtre activé ⓘ

# The advance search interface

## Various options are currently allowed

- Search a specific expression or exclude some keywords
- Allows to limit the search to a domain (useful when the teachers have their favorite site)
- Choose between the 6-classes and 9-classes scales
- Use a adult content filter

Dmesure as a web crawler

# Advance search : exemple of results

100% Edition Affichage Historique Marque-pages Outils Aide

http://cental.ftr.ul.ac.be/team/francois/dmesure\_interface/ftrm/index.php

Dmesure

Vos résultats Dmesure				
Statut	Niv. Dmesure	Texte (premières lignes)	Url	Dmesure confiance
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. ...	<a href="http://platea.primc.mec.es/cvera/hotpot/disons/taches_menageres1.htm">http://platea.primc.mec.es/cvera/hotpot/disons/taches_menageres1.htm</a>	0.308925660315
	B1	<=> Index continuez. Le salon. Passez la souris sur les images pour connaître le nom des objets représentés. OK ...	<a href="http://platea.primc.mec.es/cvera/hotpot/disons1.htm">http://platea.primc.mec.es/cvera/hotpot/disons1.htm</a>	0.31224411877
	C2	les articles indéfinis: arrêt à 2e exercice. définis et indéfinis: arrêt à l'exercice je ... mettez en ordre les questions: cliquez ici. PRÉPOSITIONS ET NATIONALITÉS. cliquez ici ...	<a href="http://platea.primc.mec.es/cvera/devoirs/1sens_premiers.htm">http://platea.primc.mec.es/cvera/devoirs/1sens_premiers.htm</a>	0.954827750506
	B1	Littérature enfantine. Contes animés. ANDERSEN et GRIMM. Bienvenue au pays de ... Contes, tables, nouvelles, Fête des contes, Contes et Légendes ...	<a href="http://platea.primc.mec.es/cvera/ressources/enfance1.htm">http://platea.primc.mec.es/cvera/ressources/enfance1.htm</a>	0.266455968416
	C2	Réseau canadien d'information sur le patrimoine. Musées de Paris. 1. ... Serveur officiel de la ville de Paris. Agropolis Museum. Le Musée des ...	<a href="http://platea.primc.mec.es/cvera/ressources/recueil88.htm">http://platea.primc.mec.es/cvera/ressources/recueil88.htm</a>	0.345412952541
	C2	Analyses de séquences d'oeuvres cinématographiques proposées par des ... Ecriture à partir de résumés de films. Être et avoir: un film de Nicolas Philibert ...	<a href="http://platea.primc.mec.es/cvera/ressources/cinemaencours.htm">http://platea.primc.mec.es/cvera/ressources/cinemaencours.htm</a>	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 ...	<a href="http://platea.primc.mec.es/cvera/ressources/decouverte.htm">http://platea.primc.mec.es/cvera/ressources/decouverte.htm</a>	0.339503056644
	C2	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 ...	<a href="http://platea.primc.mec.es/cvera/textfances/1exerciceecoute.html">http://platea.primc.mec.es/cvera/textfances/1exerciceecoute.html</a>	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 ...	<a href="http://platea.primc.mec.es/cvera/hotpot/disons/maison_chez_moi.htm">http://platea.primc.mec.es/cvera/hotpot/disons/maison_chez_moi.htm</a>	0.290072362723
	C1	Rien ne sert de courir, il faut partir à point. Le livre et la tortue en sont un ... Si vous portiez une maison? * Vérifier. OK <=> Index continuez ...	<a href="http://platea.primc.mec.es/cvera/hotpot/devistortue.htm">http://platea.primc.mec.es/cvera/hotpot/devistortue.htm</a>	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 ...	<a href="http://platea.primc.mec.es/cvera/hotpot/1commedia2.htm">http://platea.primc.mec.es/cvera/hotpot/1commedia2.htm</a>	0.933708422077
	C1	Gilbert Beicaud, "La Bolduc": La Bolduc Souhaites. Histoire d'une rencontre historique entre Brel, Brassens et Léo Ferré sur. Pour écouter l'interview ...	<a href="http://platea.primc.mec.es/cvera/ressources/1bolducsbrel.htm">http://platea.primc.mec.es/cvera/ressources/1bolducsbrel.htm</a>	0.334228986426
	C2	<=> Index continuez. Huit femmes. À la de Noël, une isolée sous la ... Reste à chercher l'association les autres... Huit femmes, huit. L'une d'entre elles est ...	<a href="http://platea.primc.mec.es/cvera/cinecinema/huitfemmes3b.htm">http://platea.primc.mec.es/cvera/cinecinema/huitfemmes3b.htm</a>	0.405439906051
	C2	Centre européen de formation aux métiers du cinéma. Cinéma-thèque ... Fédération nationale des industries techniques du cinéma et de l'audiovisuel (FITCA) ...	<a href="http://platea.primc.mec.es/cvera/ressources/1cinemaeag.htm">http://platea.primc.mec.es/cvera/ressources/1cinemaeag.htm</a>	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 ...	<a href="http://platea.primc.mec.es/cvera/hotpot/sagat2.htm">http://platea.primc.mec.es/cvera/hotpot/sagat2.htm</a>	0.263732889029
	C1	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 ...	<a href="http://platea.primc.mec.es/cvera/hotpot/1desappas.htm">http://platea.primc.mec.es/cvera/hotpot/1desappas.htm</a>	0.805257449582
	C2	Déroulez les menus afin de choisir les articles qui mangent selon ... On est rentrés manger à la maison. Le fromage et les boîtes, les confitures et les ...	<a href="http://platea.primc.mec.es/cvera/hotpot/1des_corrrections.htm">http://platea.primc.mec.es/cvera/hotpot/1des_corrrections.htm</a>	0.324757053715
	C1	Pour faire des courses à Paris ou en France, cliquez ici. Les côpes, recette interactive ... Exercices sur le déroulement des saisons de l'année. Exercices de ...	<a href="http://platea.primc.mec.es/cvera/ressources/recueil12.htm">http://platea.primc.mec.es/cvera/ressources/recueil12.htm</a>	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 chemins. 7. Devant il y a un feu ...	<a href="http://platea.primc.mec.es/cvera/hotpot/1maison_entre.htm">http://platea.primc.mec.es/cvera/hotpot/1maison_entre.htm</a>	0.313991937881
	C2	Les chanteurs francophones... ma passion! Charles Trenet ...	<a href="http://platea.primc.mec.es/cvera/ressources/1chansons1.htm">http://platea.primc.mec.es/cvera/ressources/1chansons1.htm</a>	0.06454304058

Page 1 sur 3 20

Enregistrements 1 - 20 sur 80

Terminé

## Results are already a bit more heterogeneous.

# Plan

- 1 Introduction : the issue of finding texts
- 2 Current work for English
- 3 **Dmesure : a web tool for FFL readability**
  - The one-text interface
  - Dmesure as a web crawler
  - **Dmesure as a collaborative platform**
- 4 Issues and perspectives with Dmesure
- 5 References



# Dmesure as a collaborative platform

For teachers previously recognized as experts (details have still to be defined), Dmesure offers an opportunity to contribute to further advances of the tool :

- Teachers can validate or correct the predictions of Dmesure on texts they have read or used in a teaching context
- This would allow to gather more texts, that may be assessed by more than one judge
- Then, a new readability formula can be trained
- Furthermore, as the dominant L1 of the students is saved for each text, it will allow to study L1 effects on L2 readability.

Dmesure as a collaborative platform

# Dmesure : the teacher interface

Dmesure - Validation de textes - Mozilla Firefox

Fichier Édition Affichage Historique Marque-pages Outils Aide

http://cental.fltr.ucl.ac.be/team/francois/dmesure\_interface/html/validateText.php

Dmesure - Validation de textes

Rechercher un texte | Introduire un texte | Aide | Valider un texte

Connecté : Thomas | Déconnexion

# Dmesure

**Merci de valider les textes que vous avez lus ou testés en classe. Cela permettra d'améliorer les performances de Dmesure.**  
**[en savoir plus : Rechercher un texte ; Valider un texte ; Annuler une validation précédente ]**

Les derniers textes que vous avez consultés

Validé	Date	Niv. dmesure	Texte (premières lignes)	Url	Votre niv.	Validation	L1 étudiants
✓	2011-02-10	A2+	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 cœur en hiver de Claude Sautet. À huit heures, elle sort du cinéma et elles ont rencontré Philippe un architecte, ami		A2	lu	allemand
!	2011-02-10	A1	- Bonjour M. Durant. Comment allez-vous ? - Très bien, merci. Et vous M. Durant ? - Bien aussi. Je vous remercie.				
!	2011-02-10	C2	L'institutionnel reste un prélabile aux yeux du CD&V, a rappelé le député Eric Van Rompuy jeudi sur les ondes de Twizz radio, alors que les interrogations allaient bon train sur la méthode que proposera l'informateur Didier Reynders. Pour M. Van Rompuy, il faut avant tout restaurer la confiance		B2		

Page 1 sur 1 10 Enregistrements 1 - 3 sur 3

# Plan

- 1 Introduction : the issue of finding texts
- 2 Current work for English
- 3 Dmesure : a web tool for FFL readability
  - The one-text interface
  - Dmesure as a web crawler
  - Dmesure as a collaborative platform
- 4 Issues and perspectives with Dmesure
- 5 References

# Dmesure : First conclusions

- While still needing to be debugged, the architecture seems suited to the task
- The one-text interface already gives quite good results, but the web search tool gives less usefull predictions
- This is explained by the boilerplate issue.

# The boilerplate issue

## Why boilerplate is an issue ?

- Difficulty of surrounding context (ads, news, etc.) may differ from the difficulty of the target text
- Long menus don't have full stops
- Boilerplate may incorporate more Named Entities

## Some numbers

For 30 web pages, we compared the predictions of Dmesure on the text with and without boilerplate (manually removed) :

Correlation is low :  $r = 0,56$ ;  $se = 1,73$

# A solution : automatical boilerplate remover

We tried to automatically remove the boilerplate  
[Kohlschütter et al., 2010].

## evaluation 1

For the same 30 web pages, we compare the predictions of Dmesure :

- with and without boilerplate (manually removed) :  
 $r = 0,56$ ;  $se = 1,73$
- with and without boilerplate (automatically removed) :  $r = 0,80$
- without boilerplate (manual) and without boilerplate (automatical) :  $r = 0,70$ ;  $se = 1,41$

# A solution : automatical boilerplate remover

To obtain a absolute measure, we took 180 annotated texts (20 for each level) and, for each, we generated a fake boilerplate from those of the 30 previous web pages.

## Results of evaluation 2

- 1 Results of Dmesure on those fake web pages (with boilerplate) :  $se = 4,08$
- 2 Results of Dmesure on the texts (without boilerplate) :  $se = 2,03$
- 3 Results of Dmesure on those clean web pages (boilerplate automatically removed) :  $se = 2,91$

It seems to improve somehow the predictions, but not so much

# Conclusions

- Dmesure aims to be a collaborative platform for information retrieval in ICALL. It may answer to real needs of L2 teachers.
- BUT... more work is necessary :
  - Settle the boilerplate issue or adapt the formula to the specificities of the web
  - Use a Named Entities extractor and define a way to assess the difficulty of NE.
  - Develop a filter for language checking



# The end

**Difficulté estimée :** A1 ?

**Votre texte :** Merci pour votre attention.

Les questions et les commentaires  
sont les bienvenus.

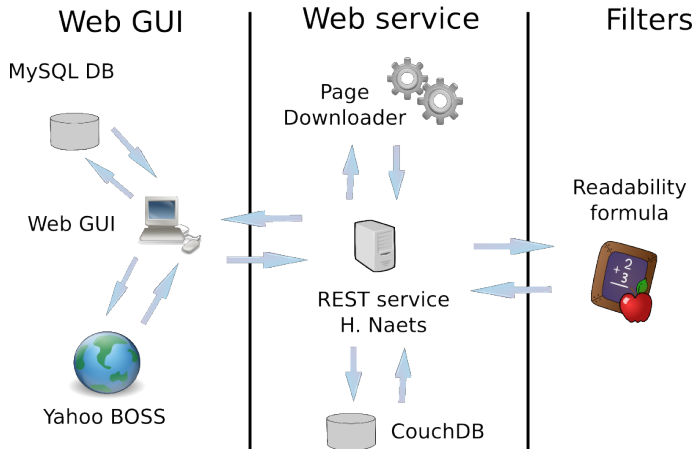
**Difficulté estimée :** C2 ?

**Votre texte :** Thanks for your attention.

Questions and commentaries are welcome !

# Dmesure : the architecture

## Architecture of Dmesure



# Plan

- 1 Introduction : the issue of finding texts
- 2 Current work for English
- 3 Dmesure : a web tool for FFL readability
  - The one-text interface
  - Dmesure as a web crawler
  - Dmesure as a collaborative platform
- 4 Issues and perspectives with Dmesure
- 5 References

# References I



Agresti, A. (2002).  
*Categorical Data Analysis. 2nd edition.*  
Wiley-Interscience, New York.



Collins-Thompson, K. and Callan, J. (2004).  
Information retrieval for language tutoring : An overview of the REAP project.  
In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 545–546.



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.



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

# References II



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



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



François, T. (2009).  
Combining a statistical language model with logistic regression to predict the  
lexical and syntactic difficulty of texts for FFL.  
*In Proceedings of the 12th Conference of the EACL : Student Research  
Workshop*, pages 19–27.



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.

# References III



Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008a).  
An analysis of statistical models and features for reading difficulty prediction.  
*In Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–8.



Heilman, M., Zhao, L., Pino, J., and Eskenazi, M. (2008b).  
Retrieval of reading materials for vocabulary and reading practice.  
*In Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 80–88.



Henry, G. (1975).  
*Comment mesurer la lisibilité*.  
Labor, Bruxelles.



Kincaid, J., Fishburne, R., Rodgers, R., and Chissom, B. (1975).  
*Derivation of new readability formulas for navy enlisted personnel*.  
Technical report, n°8-75, Research Branch Report.

# References IV



Kohlschütter, C., Fankhauser, P., and Nejdl, W. (2010).

Boilerplate detection using shallow text features.

In *Proceedings of the third ACM international conference on Web search and data mining*, pages 441–450.



Miltsakaki, E. and Troutt, A. (2008).

Real-time web text classification and analysis of reading difficulty.

In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 89–97.



Ott, N. (2009).

Information Retrieval for Language Learning : An Exploration of Text Difficulty Measures.

Master's thesis, University of Tübingen, Seminar für Sprachwissenschaft.

<http://drni.de/zap/ma-thesis>.



Schmid, H. (1994).

Probabilistic part-of-speech tagging using decision trees.

In *Proceedings of International Conference on New Methods in Language Processing*, volume 12. Manchester, UK.