Dmesure: a readability platform for French as a foreign language

Thomas François[1, 2] and Hubert Naets[2]

(1) Aspirant F.N.R.S.
(2) CENTAL, Université Catholique de Louvain

Presentation at CLIN 21

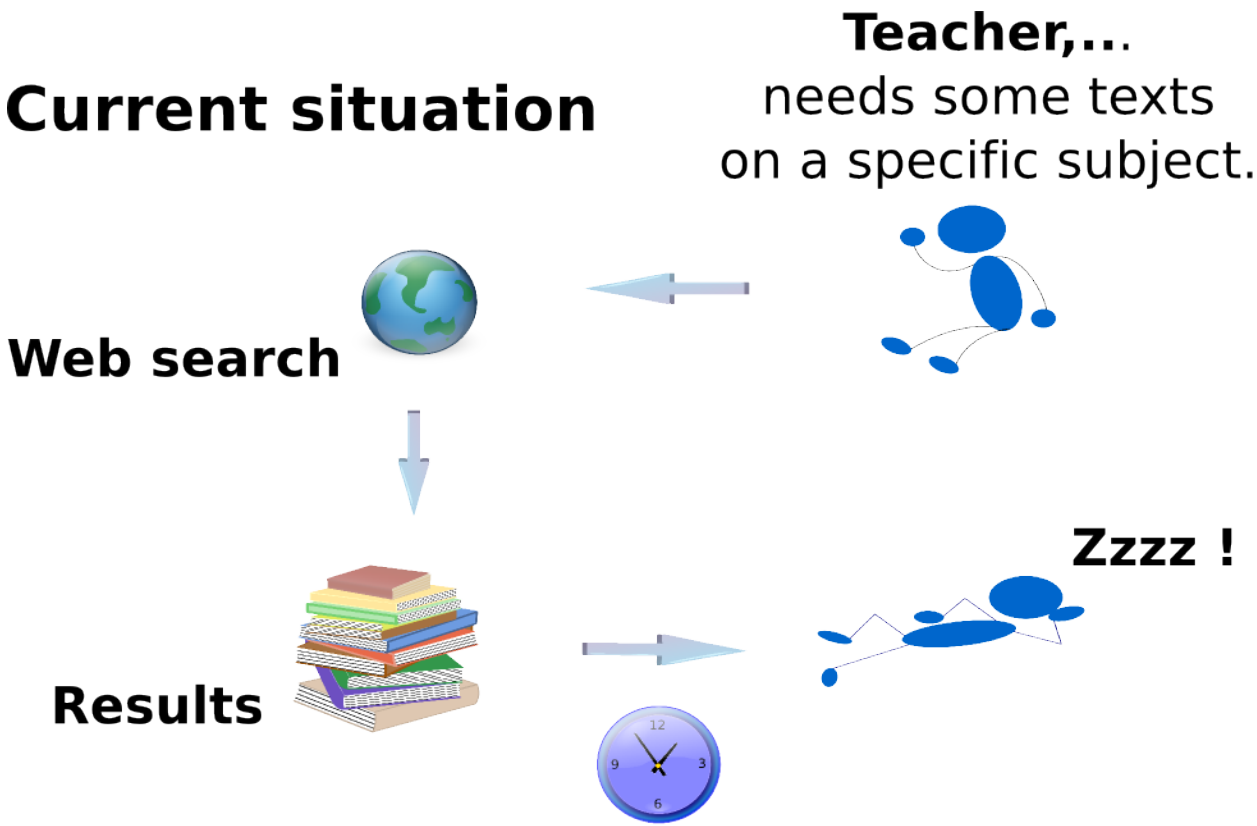
February 11, 2011

Plan

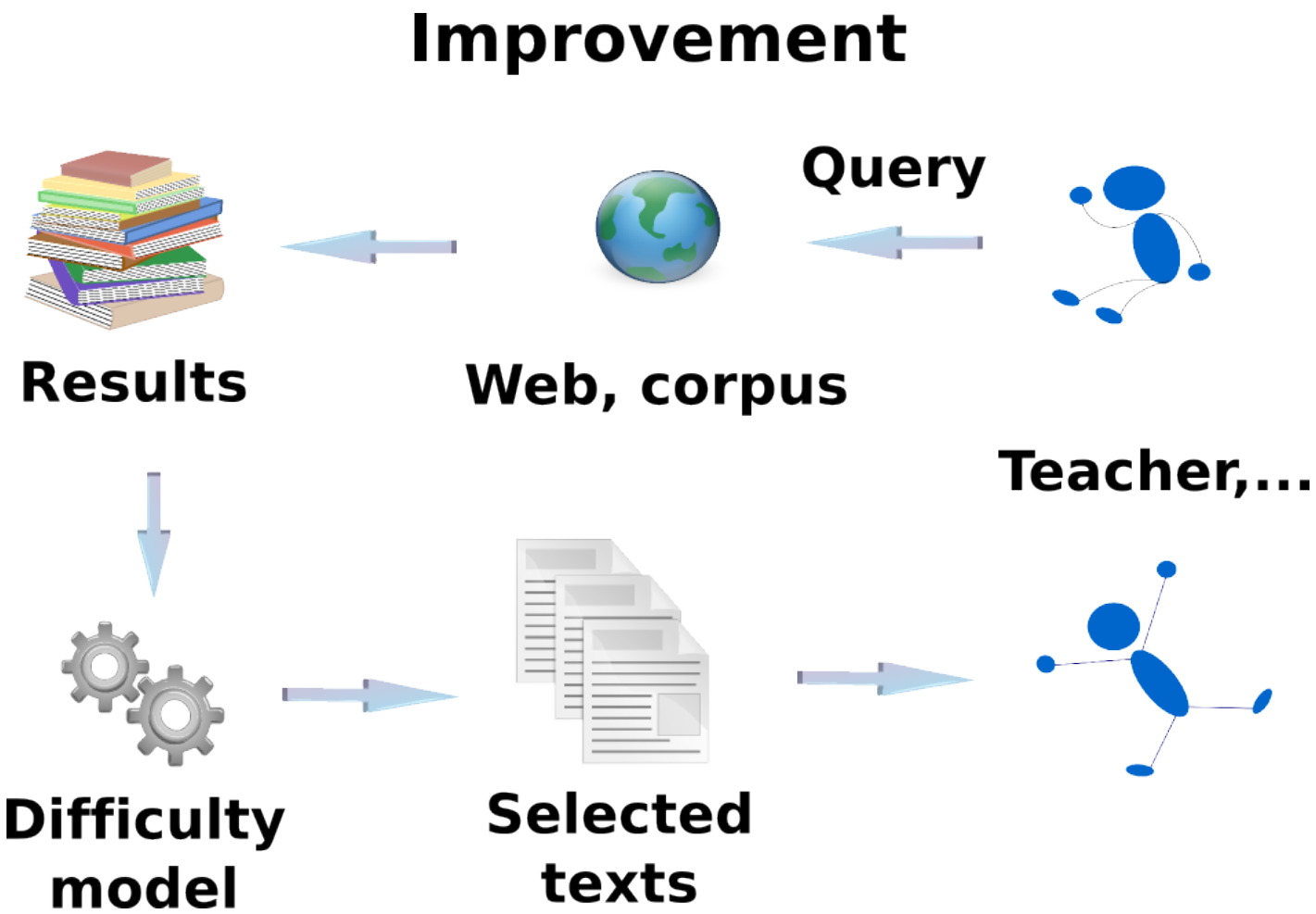
Plan

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



What is a difficulty model for reading ?

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

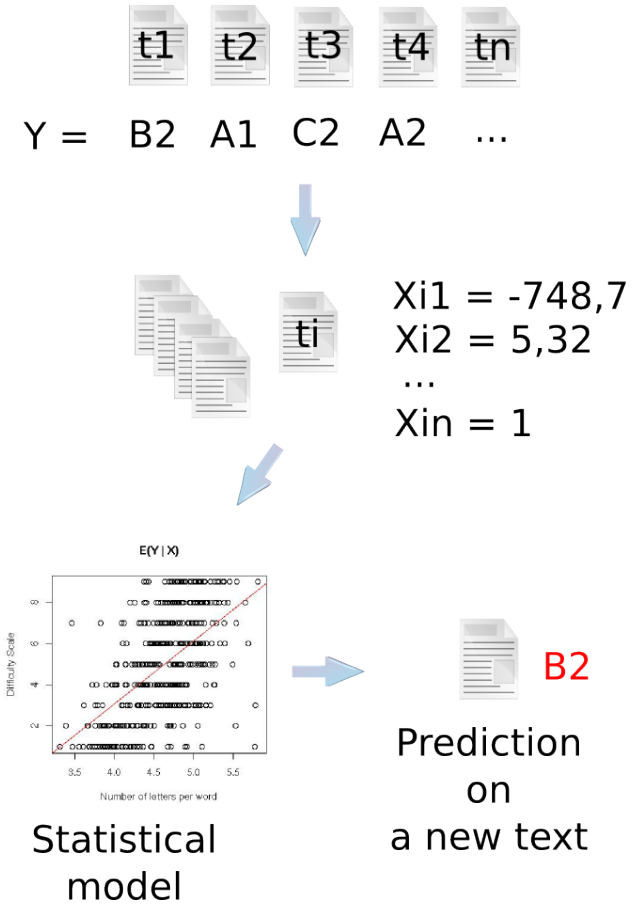
What is readability ?

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]

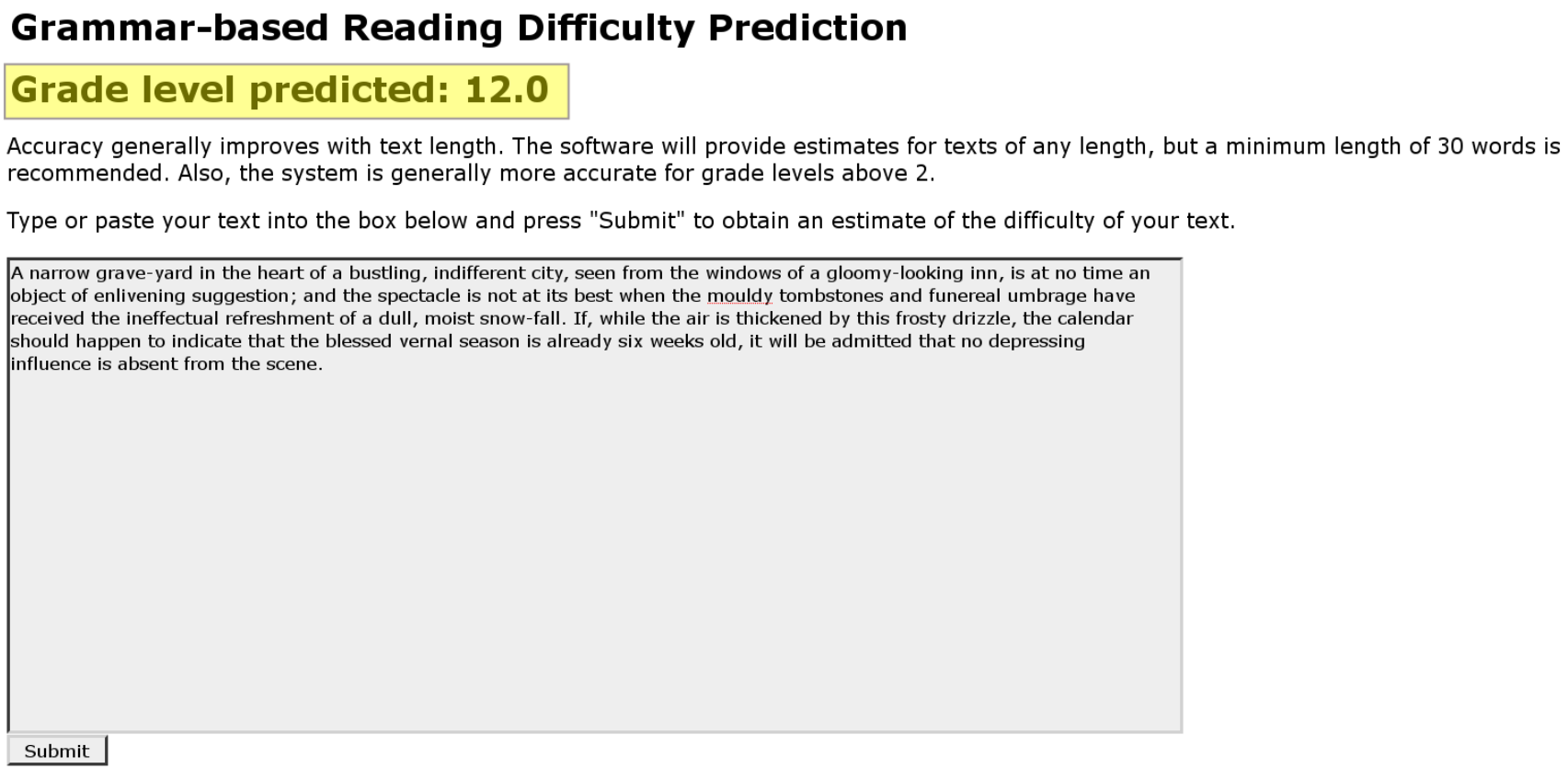
Some of the well-known formulas :
[Flesch, 1948, Dale and Chall, 1948, Kincaid et al., 1975]

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



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

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].

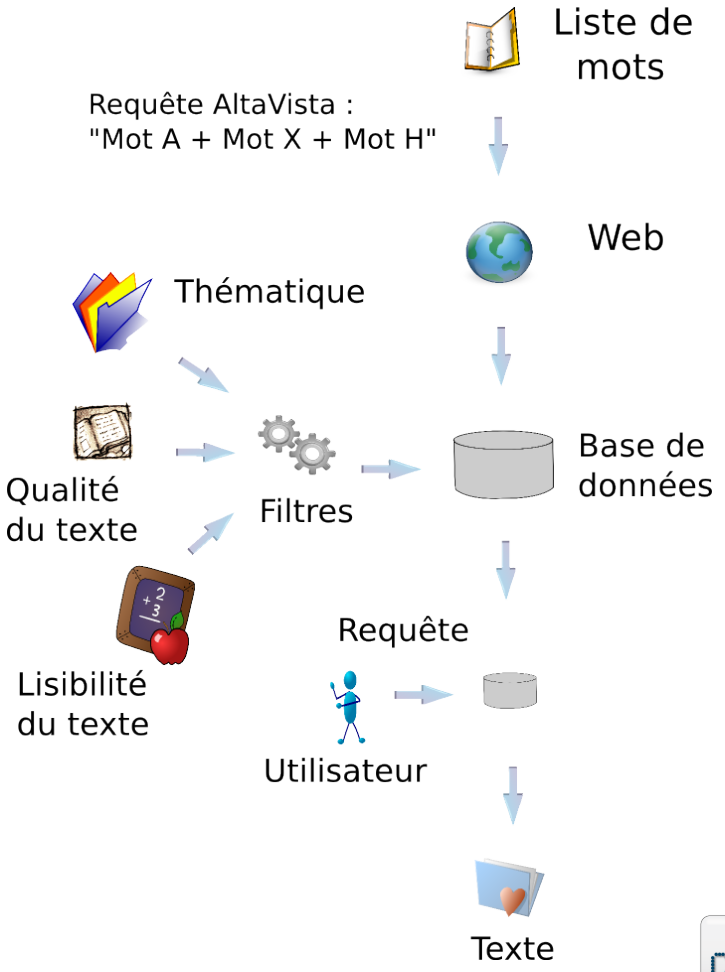
Retrieval of web texts : an example for EFL

- **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

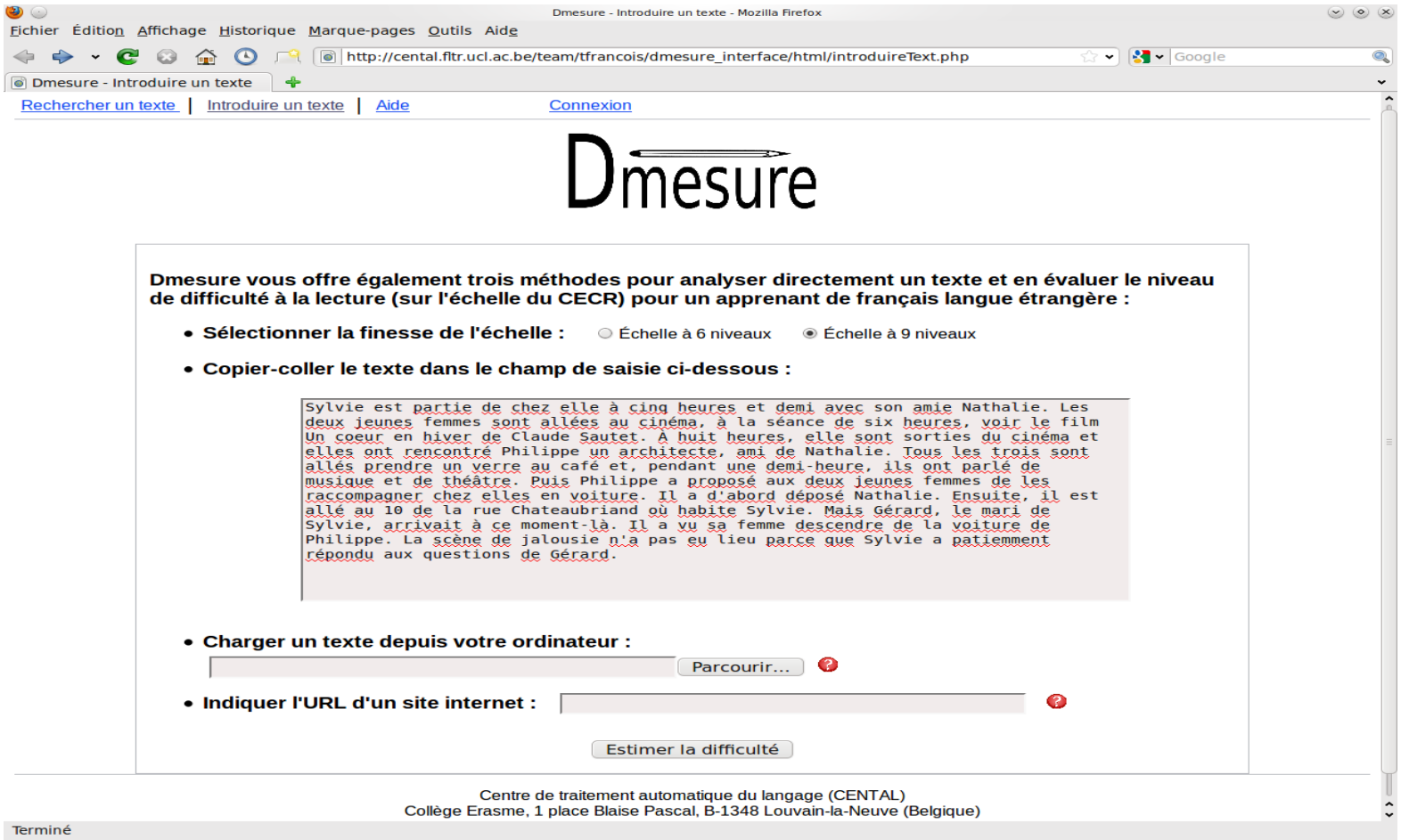
Dmesure : 3 goals

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 plateform where teachers can participate in assessing the difficulty of texts they collected through Dmesure and they used in their teaching

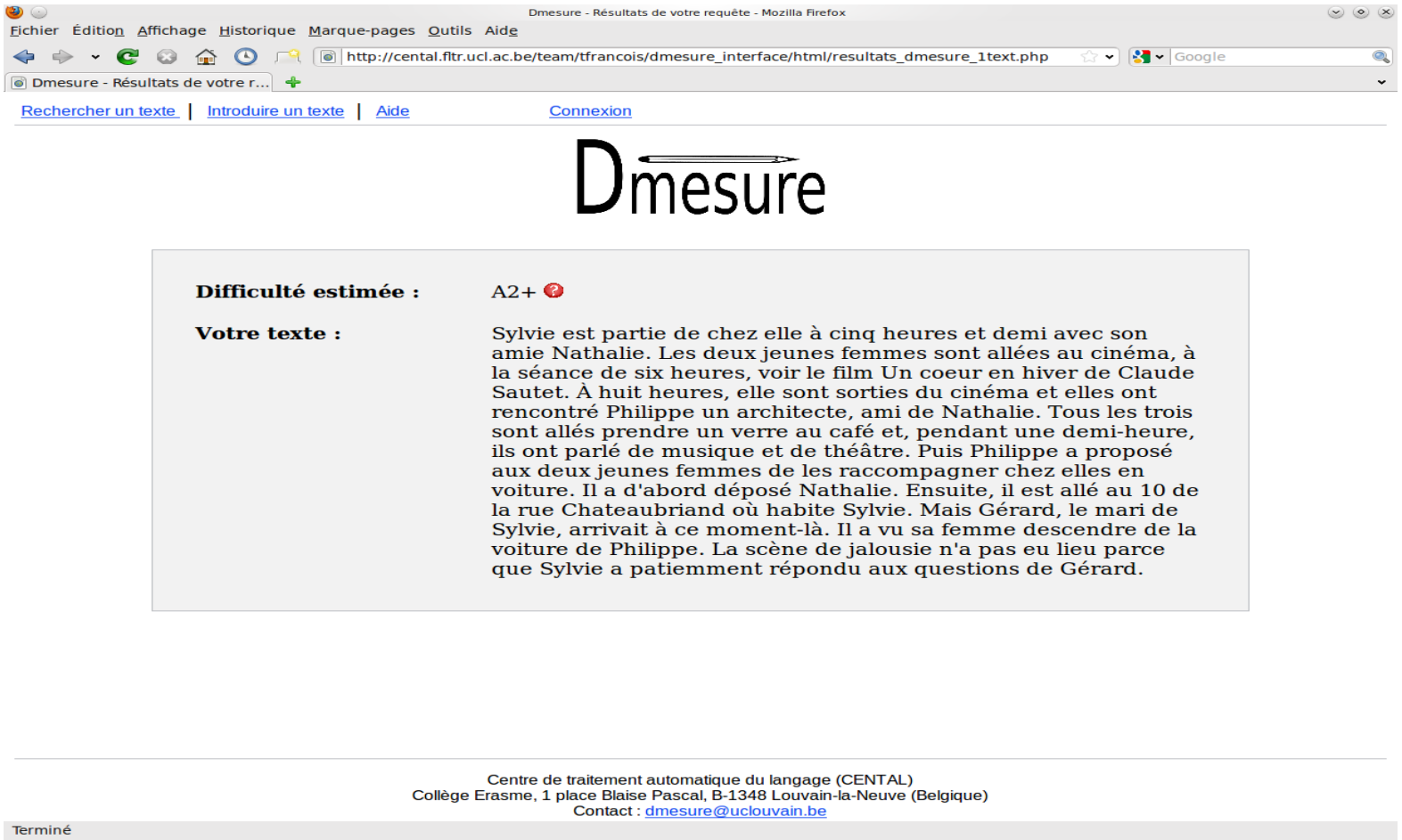
Plan

Dmesure : the one-text interface



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

Dmesure : the one-text interface



The model did quite well on this one ! !

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.

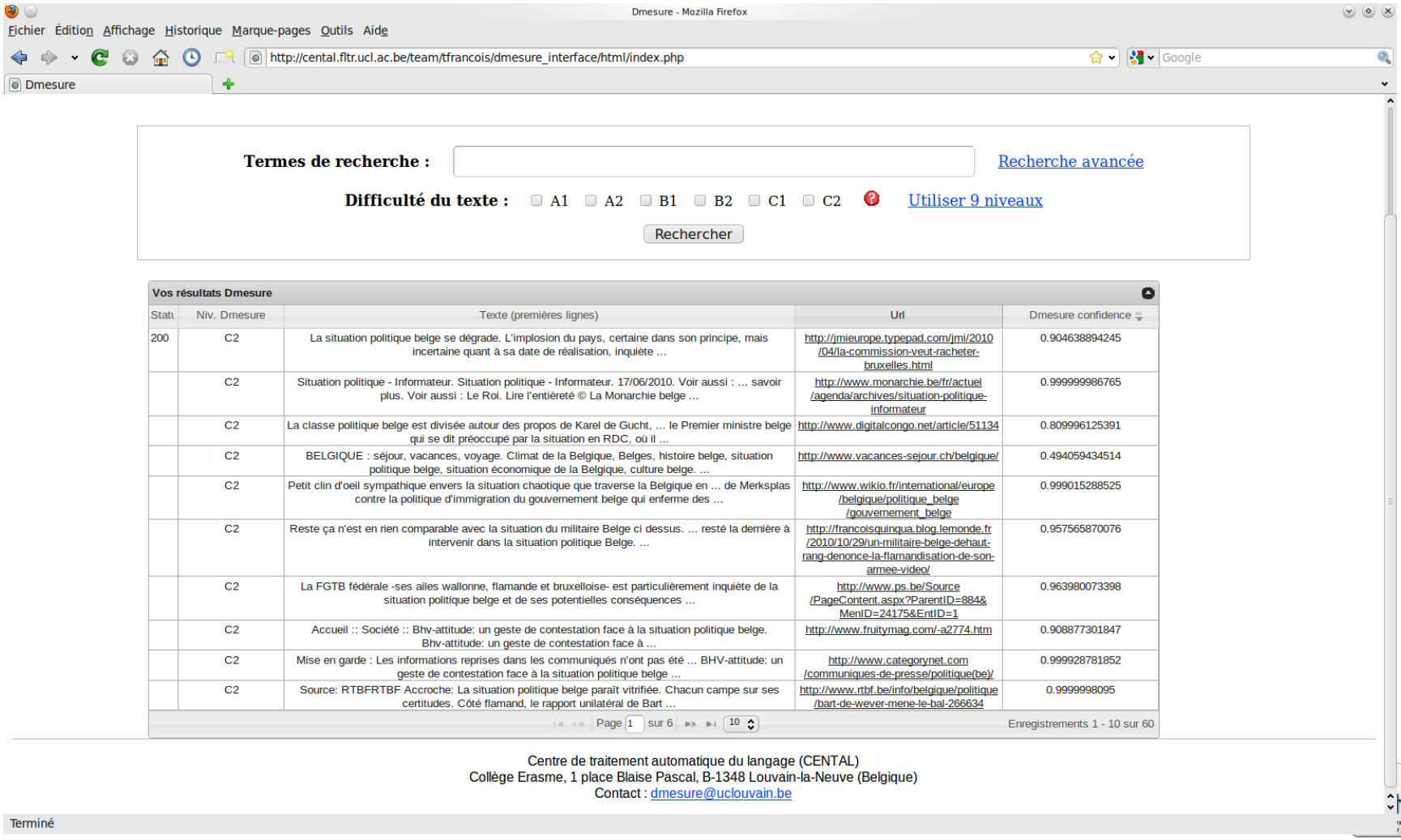
Algorithm : ordinal logistic regression [Agresti, 2002]

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

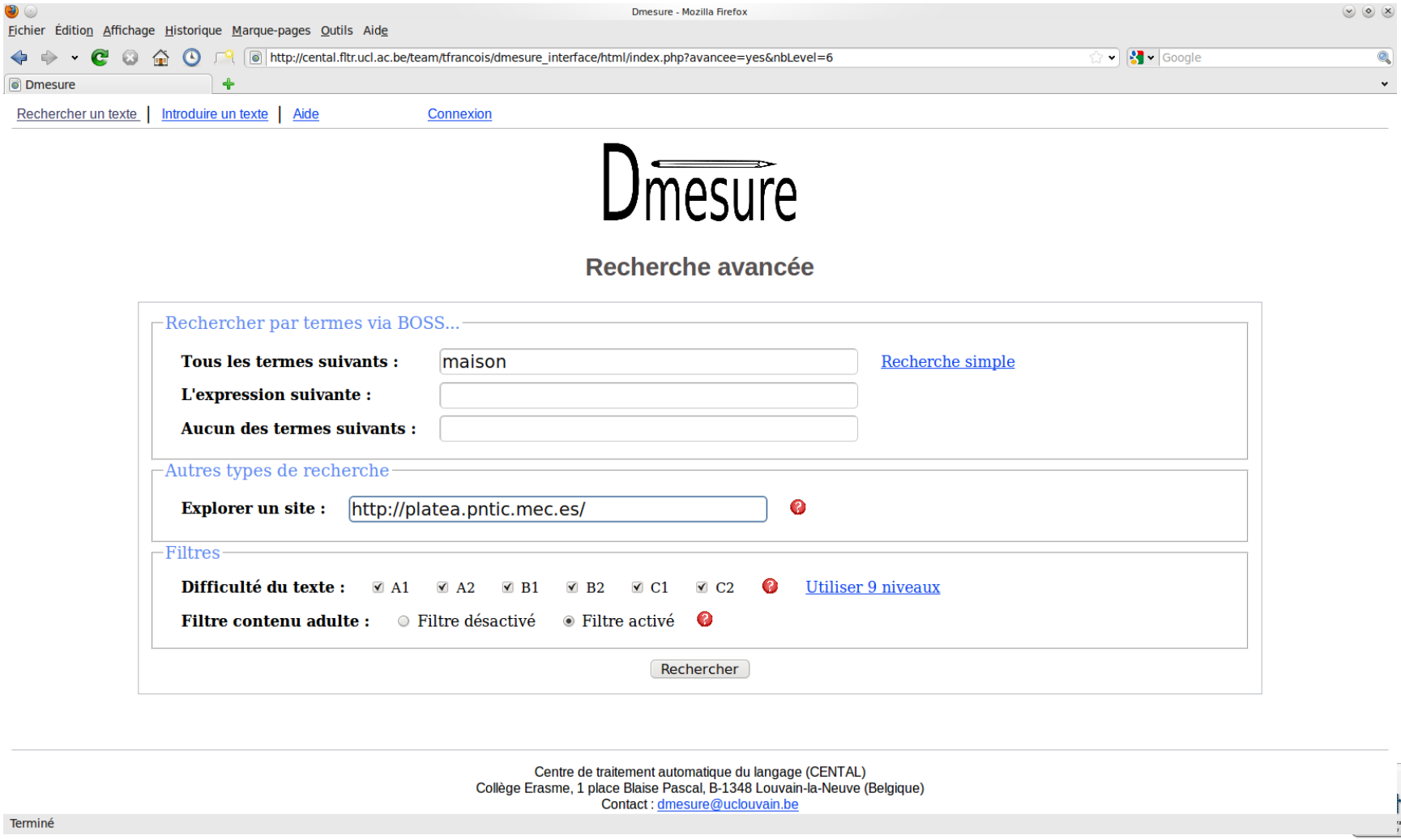
Plan

The basic search interface

Results of previous request



The advance search interface

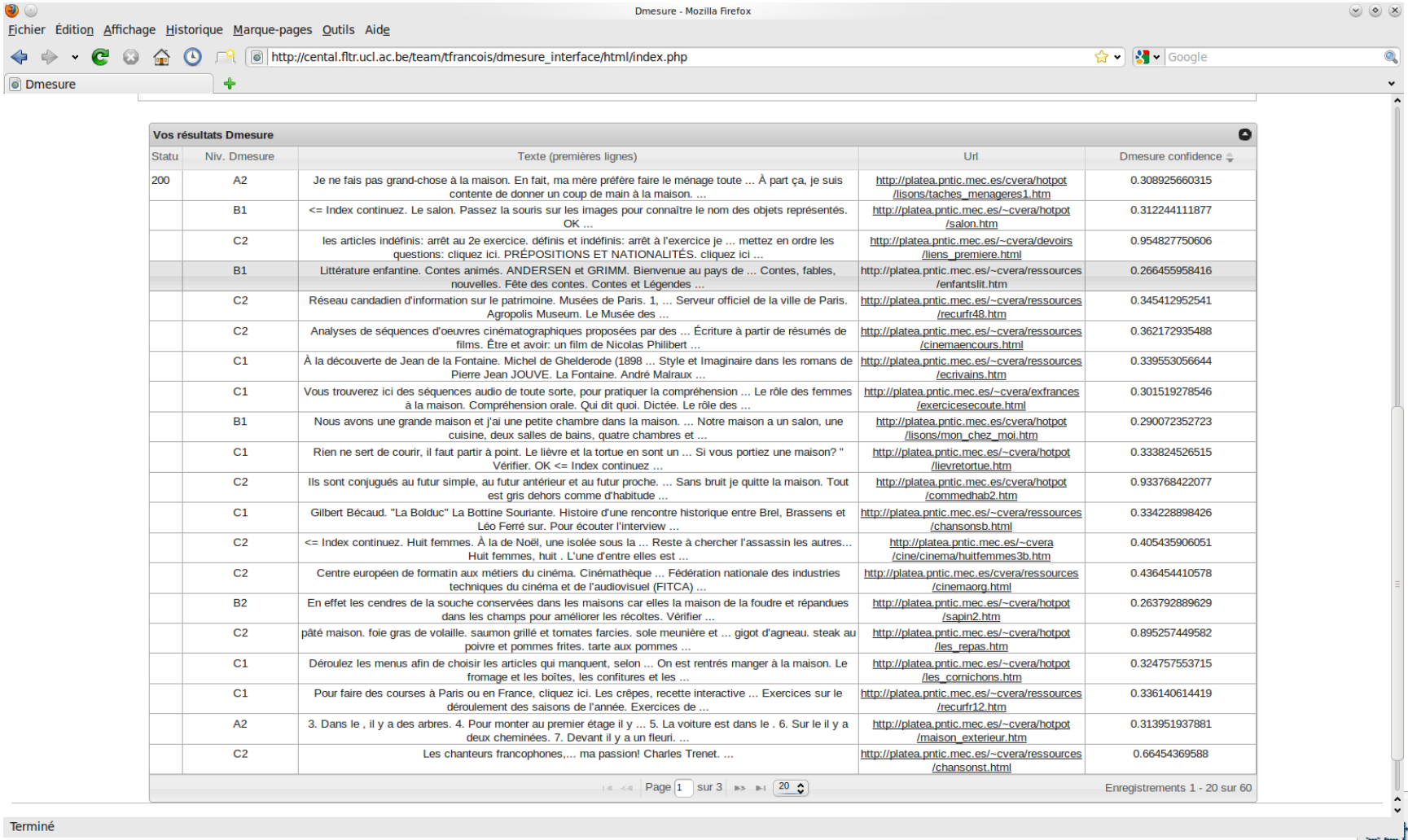


The advance search interface

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

Advance search : example of results



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

Results are already a bit more heterogeneous.

Plan

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 : the teacher interface

Plan

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) :

$$\text{Correlation is low : } r = 0,56; se = 1,73$$

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.

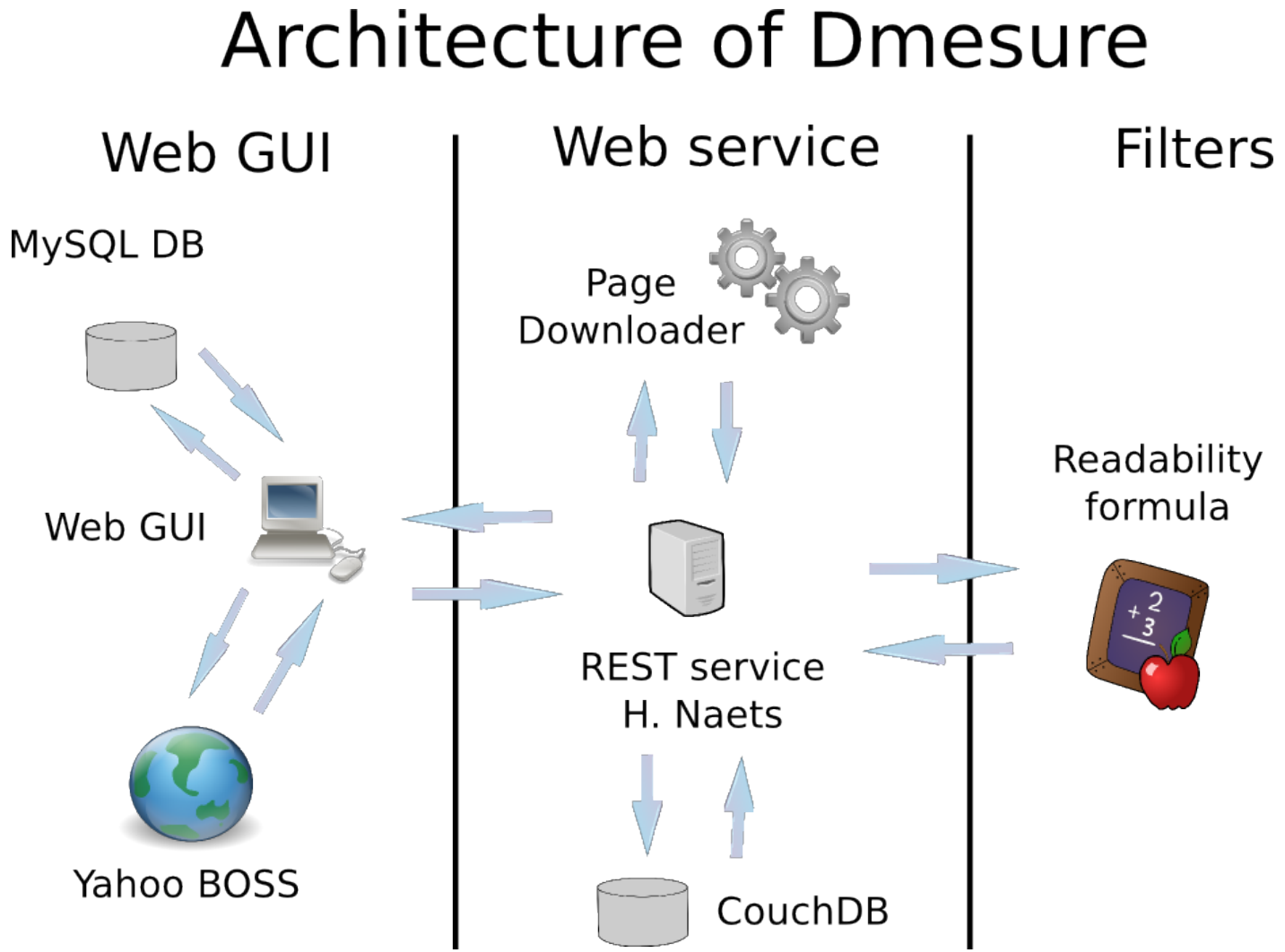
Les questions et les commentaires
sont les bienvenus.

Difficulté estimée : C2

Votre texte : Thanks for your attention.

Questions and commentaries are welcome !

Dmesure : the architecture



Plan

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.