Réutiliser des textes dans un contexte éducatif

L'évaluation de la difficulté

Elie Abou Zeid¹, Muriel Foulonneau², Talar Atéchian¹

1. Faculté d'Ingénieurs en Informatique, Multimédia, Systèmes,
Réseaux et Télécommunications
Baabda – Liban
elie_abouzeid@hotmail.com, talar.atechian@upa.edu.lb

2. Centre de Recherche Public Henri Tudor
29, av. John F. Kennedy
L-1855 Luxembourg
muriel.foulonneau@tudor.lu

RÉSUMÉ. Les collections de ressources numériques peuvent être exploitées dans un cadre éducatif, par exemple comme stimuli dans des questions de test ou comme des illustrations dans des objets pédagogiques. De nombreux détenteurs de collections de ressources média affichent de plus en plus l'ambition que leurs ressources soient utilisées dans un contexte éducatif. Néanmoins, les métadonnées associées à ces ressources ne sont souvent pas directement utilisables comme critères de sélection. Des extraits de romans ou des articles de l'actualité économique, s'ils ne sont pas des ressources pédagogiques en eux-mêmes présentent des caractéristiques qui les rendent plus ou moins intéressants et accessibles pour un profil d'apprenant et dans un contexte pédagogique particulier. Nous présentons dans cet article les mesures de difficulté des textes qui peuvent permettre d'annoter une collection et de qualifier le contexte pédagogique dans lequel elle pourra être valorisée. Nous avons mis en œuvre des mesures statistiques, syntaxiques et sémantiques dans un système unique afin d'opérer des combinaisons en fonction de l'importance relative que chaque mesure doit avoir dans un contexte et pour un type de texte particuliers. Nous avons utilisé une approche classique basée sur des calculs statistiques, ainsi qu'une approche de web mining et des technologies sémantiques. Nous avons testé le système avec des jeux de données précalibrés. Nous montrons les possibilités offertes par un outil ouvert pour l'évaluation de la réutilisabilité des collections de ressources multimédias dans un environnement éducatif.

ABSTRACT. Collections of digital resources can be used in a learning context, for instance as stimuli in test questions or as illustrations in learning objects. Many curators of digital media collections aim to address learning audiences. Nevertheless, metadata associated with digital resources are often not directly usable for the selection of digital media in education. Although a novel or a business news article for instance are not learning resources, they can

be more or less interesting and accessible for a specific learner profile and in a particular learning context. In this article, we present text difficulty metrics that can help annotate a collection and specify the learning context in which the resource can be most valuable. We have implemented statistical, syntactic, as well as semantic metrics, in a unified system in order to allow combinations according to the relative importance of each measure in a particular context and for a particular type of text. We have used a classical approach based on statistical metrics, as well as a Web mining approach and semantic technologies. We have tested the system with pre-calibrated datasets. We show the potential of an open tool for the evaluation of the reusability of multimedia collections in an educational environment.

MOTS-CLÉS : *réutilisation des ressources multimédia, éducation, métadonnées, fouille de texte, Web mining, WordNet, PIMO.*

KEYWORDS: *reuse of multimedia resources, text mining, education, metadata, Web mining, WordNet, PIMO.*

DOI:10.3166/DN.15.3.119-142 © Lavoisier 2012

1. Introduction

De nombreuses collections patrimoniales et bibliothèques numériques multimédias affichent l'ambition d'être utilisées dans un contexte éducatif (ex. Europeana ou la Bibliothèque nationale de France). Il est cependant rare que des pédagogues soient impliqués au niveau de la valorisation de la collection ou qu'une stratégie spécifique soit développée afin de rendre ces collections davantage utilisables dans un contexte éducatif.

Or, une ressource éducative est typiquement un cours ou une portion de cours (*IEEE Learning Object Metadata*). Elle a été conçue avec un objectif pédagogique et pour un contexte éducatif spécifique. Les contextes éducatifs sont représentés notamment sous la forme d'un niveau (le niveau des élèves auxquels une ressource s'adresse) ou d'un domaine enseigné (ex. l'histoire de l'Europe).

Une représentation de la Joconde ou un article de journal ne sont pas en eux-mêmes des ressources éducatives. En revanche, ils peuvent être intégrés dans une ressource à caractère éducatif, telle qu'un cours d'histoire de l'art ou un module sur les évolutions du marché des technologies (Schmidt *et al.*, 2011). Ce sont alors des ressources media externes incluses dans une ressource éducative. Le rôle des ressources media peut être extrêmement variable. Néanmoins, leur choix est très important, car elles ont une influence significative à la fois sur la difficulté des ressources (ex. Grunwald *et al.*, 2006), sur ses aspects cognitifs (impact sur la charge cognitive Mayer, 2005) et sur la capacité des apprenants à acquérir une compétence (Van de Vijver, 2004).

Ce sont ces médias qui peuvent provenir des collections patrimoniales et des bibliothèques numériques. Pourtant, leurs métadonnées ne sont pas spécifiques au contexte éducatif. Or un avis du Conseil national du numérique en France constatait en mars 2012 que « les enseignants se heurtent aux difficultés d'identifier les

ressources qui leur conviennent, d'évaluer leur qualité, de comprendre leur modalité d'utilisation, et finalement de les acquérir»¹.

Afin d'évaluer l'utilisabilité d'un texte dans le domaine éducatif, des outils ont été développés. Ils évaluent la difficulté de lecture pour un apprenant et cherchent pour certains à prédire les publics qui peuvent être ciblés. Heilman, Zhao *et al.* (2008) ont même proposé d'annoter une collection numérique de textes en fonction de leur difficulté. Ce type de mesures est particulièrement intéressant mais il doit être rapporté au contexte d'utilisation. La lisibilité d'un texte utilisé pour l'apprentissage d'une langue ne dépend pas nécessairement des mêmes mécanismes et paramètres qu'un texte utilisé en littérature ou en sciences, l'importance des mots rares, par exemple, peut varier en fonction du contexte. Afin de rendre possible ce type d'annotations non pas sur des textes individuels mais sur des corpus entiers et dans des contextes éducatifs différents, nous avons intégré plusieurs approches et outils dans un service web qui permet de fournir des indicateurs de la difficulté d'un texte dans un cadre éducatif. Il devra permettre de s'intégrer à des outils externes, tels que TAO², une plate-forme sémantique en libre accès pour la création et la diffusion de tests (Ras *et al.*, 2010) et un système de génération de questions de tests (Foulonneau *et al.*, 2012).

Dans une première partie, nous décrivons les différentes approches de la mesure de difficulté des textes. Nous présentons ensuite le système mis en place, ainsi qu'une expérimentation à petite échelle pour rendre compte des potentialités de ce type de mécanisme. Enfin, nous illustrons la manière dont des textes avec différentes caractéristiques sont réutilisés dans des ressources éducatives.

2. Les mesures de la difficulté d'un texte

Nous définissons la difficulté – et à l'inverse sa lisibilité – d'un texte par rapport à sa pertinence dans le cadre d'un parcours éducatif de niveau débutant à expert. Les textes peuvent être intégrés à des ressources éducationnelles variées, pour étudier la littérature, les sciences ou une langue étrangère. Ils présentent des difficultés de natures diverses et généralement des caractéristiques différentes, à la mesure de la différence entre un article de journal et un roman du XIX^e siècle. Les éléments qui rendront la lecture difficile n'impactent pas de la même manière le lecteur selon qu'il devra comprendre le texte pour réaliser un exercice ou ne l'utilisera que comme support ou illustration (voir une description des rôles des ressources multimédia dans Foulonneau *et al.*, 2011 et Suvorov, 2009).

Les hypothèses qui fondent les travaux sur la difficulté de lecture d'un texte sont principalement liées au vocabulaire qu'il utilise (commun ou rare, connu ou inconnu

1. Avis n° 10 du 6 mars 2012 relatif au numérique à l'École

Permettre le choix du numérique à l'école <http://www.cnnumerique.fr/avis10/>

2. <http://www.tao.lu>

du lecteur), à sa structure et à son style. Elles conduisent à définir des métriques sémantiques, syntaxiques, de cohérence et de cohésion d'un texte.

2.1. Les standards « Common Core »

La question de la difficulté de lecture des textes (ou lisibilité) a été identifiée comme un composant majeur de l'éducation aux États-Unis. Les « Common Core State Standards for Reading »³ sont des normes américaines visant à harmoniser les niveaux de difficulté des textes assignés aux élèves. Ils doivent permettre tant aux professeurs qu'aux parents et aux élèves eux-mêmes d'adapter les textes à une grille commune de niveaux.

La difficulté des textes est définie d'une part par des mesures quantitatives (ex. statistiques sur la longueur de mots), d'autre part, par des mesures qualitatives (ex. les éléments implicites ou inférés ou les connaissances requises pour la compréhension du texte).

2.2. Les mesures syntaxiques et sémantiques

Les formules de lisibilité ont débuté dès les années 1920 avec Lively et Pressey et plus de 200 formules ont été proposées depuis (Beaudet, 2001 ; Dubay, 2004). Pour la grande majorité, ces formules sont basées sur la syntaxe et la sémantique des textes. La difficulté syntaxique est mesurée notamment grâce à la longueur des phrases, alors que la difficulté sémantique prend pour référence la longueur des mots, par exemple en fonction du nombre de syllabes qu'ils contiennent, et parfois leur rareté (Fry, 2002). Des formules telles que *New Dale-Chall*, *Fry Readability*, *Gunning-Fog*, *Automated Readability* ou *Flesch Kincaid Reading Ease* utilisent toutes ce type d'approches. Encore aujourd'hui, la mesure Flesch-Kincaid, développée en 1975 est la mesure de difficulté la plus utilisée (Heilman, Collins-Thompson *et al.*, 2008). Elle est fondée sur la moyenne de syllabes par mot et la moyenne de mots par phrase. La section 3.2 détaille les métriques de Flesch.

Avec l'accroissement des capacités computationnelles, sont apparues des mesures plus complexes, exploitant notamment des corpus. Développée par Smith *et al.* (1989), la *Lexile Scale* utilise une mesure syntaxique, la moyenne de la longueur des phrases *Mean Sentence Length* et une mesure sémantique, la *Mean Log Word Frequency*, basée sur la fréquence relative d'un mot par rapport à d'autres mots dans une base de données. La mesure *Lexile Scale* est aujourd'hui largement utilisée aux États-Unis (Sheehan *et al.*, 2010).

Les formules de lisibilité de texte de ce type tendent à exagérer la difficulté des textes informationnels qui utilisent davantage de termes techniques, précis et rares. Ce type de mesures fournit souvent un score unique pour l'ensemble d'un texte,

3. http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf

alors que des disparités importantes peuvent exister entre les parties du texte. Surtout, des études ont montré qu'en diminuant la longueur des phrases par exemple, la difficulté n'était pas nécessairement inférieure (Sheehan *et al.*, 2010).

2.3. La cohérence d'un texte

La cohérence est le niveau d'organisation des concepts et relations dans un texte (Carrell, 1982). Elle est typiquement calculée à partir d'indicateurs tels que la similarité sémantique entre des parties de texte voisines.

Afin d'analyser la cohérence d'un texte, Hearst (1997) propose de le segmenter en blocs homogènes. Cette approche est appelée segmentation thématique ou *Topic Segmentation*. Elle vise à détecter un changement de sujet ou thème. Les caractéristiques utilisées dans les algorithmes de segmentation, comprennent notamment la distribution mot/concept (fréquence des mots, score de similarité lexicale, vocabulaire nouveau, relations sémantiques, répétitions des entités, etc.), les caractéristiques liées au discours et les caractéristiques multimédia. Youmans (1991) par exemple, détecte un changement de sujet lorsque le nombre des nouveaux termes introduits dans un intervalle de texte dépasse un certain seuil.

La *Latent Semantic Analysis* (LSA) permet de calculer la similarité sémantique entre deux mots en fonction de la probabilité que ces mots soient trouvés dans le même contexte ou qu'un mot apparaisse dans deux contextes différents. La similarité sémantique entre deux mots, un mot et un contexte ou bien entre deux contextes est prise comme mesure de référence. Le modèle LSA pour mesurer la similarité sémantique entre des sections de textes est implémenté dans plusieurs applications, notamment dans l'*Intelligent Essay Assessor* (Folz *et al.*, 1999) qui compare deux textes (celui écrit par un étudiant et un texte idéal) et génère un indicateur de leur similarité sur une échelle de 0 à 1.

Foltz *et al.* (1998) ont démontré l'effet de la cohérence sur la compréhension du lecteur. Le rôle de la cohérence dans la difficulté d'un texte est cependant débattu, notamment l'importance relative de la cohérence locale (une partie du texte) et de la cohérence globale du texte et l'impact du type de texte et du lecteur (Dubay, 2004).

2.4. La cohésion d'un texte

La cohésion représente les relations entre les composants d'un texte. Halliday et Hasan en 1976 distinguent cohésion lexicale et cohésion grammaticale. La cohésion grammaticale est évaluée par la présence de références, substitutions, ellipses et conjonctions. La cohésion lexicale inclut, par exemple, l'hyponymie (relation sémantique vers un élément plus spécialisé) et la référence cataphorique (référence à un élément postérieur), alors que la cohésion grammaticale est identifiée par la référence anaphorique (référence à un élément antérieur), la substitution et l'ellipse (omission d'éléments).

L'algorithme de cohésion lexicale dite du thésaurus de Roget détecte les relations sémantiques des mots à travers un texte et fournit une mesure globale de cohésion. Après un filtrage des mots fonctionnels, chaque mot restant est localisé dans le thésaurus et les informations liées à la catégorie et aux références croisées en sont extraites (Morris *et al.*, 1991). McNamara *et al.* (2010) ont testé des mesures de cohésion et démontré leur importance.

2.5. Les systèmes actuellement disponibles

Des outils pour mesurer la difficulté des textes ont été développés en particulier à destination des enseignants. NetTrekker⁴ est un outil commercial qui permet de sélectionner des liens vers des pages web, qui ont été organisés en fonction d'un niveau scolaire. L'évaluation de la difficulté est réalisée manuellement par des enseignants sur la base de différentes métriques syntaxiques et sémantiques (Huff, 2008). Largement adopté dans les écoles américaines, NetTrekker n'utilise cependant pas d'approche automatisée. Coh-Metrix⁵ développé par l'université de Memphis (McNamara *et al.*, 2010) évalue la cohérence et la cohésion du texte et fournit une équivalence entre la cohésion d'un texte et les compétences en lecture d'un utilisateur. Read-X (Millsakaki *et al.*, 2007) classifie des textes du web en fonction d'une échelle de lisibilité grâce à un corpus de référence. Toréador prédit les difficultés de vocabulaire en fonction de la fréquence du vocabulaire spécifique à un thème ou à un niveau. Ces outils sont intégrés et développés par l'université de Pennsylvanie. REAP⁶ développé par l'Université Carnegie-Mellon recommande des textes généralement issus du web en fonction des termes de vocabulaire connus par l'utilisateur (Brown et Eskenazi, 2004).

Sheehan *et al.* (2010) soulignent l'absence de prise en considération dans les outils existants du genre des textes, alors que des différences significatives existent sur le nombre de propositions, les relations de cohésion et la répétition des mots, mais aussi sur les stratégies déployées par les lecteurs pour inférer des éléments implicites, par exemple. SourceRater⁷, développé par ETS aux Etats-Unis, adopte une méthode de différenciation entre les textes de genres différents à travers 10 caractéristiques textuelles dont la difficulté de vocabulaire et deux mesures de difficulté syntaxique. SourceRater se conforme au standard Common Core.

Nelson *et al.* (2012) détaillent d'autres outils tels que ATOS et DRP Analyzer qui combinent plusieurs indicateurs. Ils concluent que les méthodes traditionnelles basées sur l'analyse statistique des textes (y compris du vocabulaire) fonctionnent toujours bien sur des corpus relativement uniformes, alors que les autres métriques permettent d'améliorer les performances des mesures classiques.

4. <http://www.nettrekker.com>

5. <http://cohmetrix.memphis.edu>

6. <http://reap.cs.cmu.edu/>

7. http://www.ets.org/research/topics/as_nlp/educational_applications

Les différentes mesures de la difficulté ont donc dans un premier temps concerné surtout les approches statistiques sur les phrases, les mots et les syllabes. L'amélioration des capacités informatiques a permis l'émergence de métriques plus complexes, telles que la cohérence et la cohésion. Finalement, les évolutions plus récentes tentent d'affiner les métriques en fonction du contexte éducatif ou du type de texte et investiguent la mise en œuvre d'approches plus individuelles (profilage) (Fry, 2002 ; Sheehan *et al.*, 2010). Les solutions les plus élaborées cherchent néanmoins à combiner les approches et proposent différents indicateurs, démontrant qu'aucun indicateur unique ne permet de traiter tous les types de textes et tous les contextes éducatifs.

3. Présentation du système et des métriques mises en œuvre pour évaluer la difficulté d'un texte

Le système adopte l'approche des standards du Common Core, qui recommandent la prise en considération de mesures quantitatives, de mesures qualitatives, ainsi que des caractéristiques du public cible. Il est divisé en trois modules, pour l'analyse syntaxique et sémantique, l'analyse statistique, enfin la prise en considération du profil de l'apprenant. Il doit permettre d'étendre les modules et d'expérimenter des combinaisons en fonction de l'évaluation de différents types de textes pour des contextes éducatifs divers.

Dans cette section, nous présentons l'architecture du système puis les métriques qu'il propose et la manière dont elles ont été mises en œuvre.

3.1. L'architecture du système

La plupart des systèmes (voir section 2.5) ne sont pas disponibles open source, ce qui empêche d'expérimenter avec de nouvelles approches pour permettre l'annotation de collections de textes. Nous avons donc mis en œuvre plusieurs mesures parmi les plus importantes pour l'anglais et les avons enrichies avec des approches liées au web sémantique.

Notre système (figure 1) analyse un texte (qu'il soit sous forme de fichier Word, de fichier PDF ou sous forme d'entrée HTML). Afin de gérer les formats de document, le système utilise différentes API. La PDFBox™ est une librairie Java open source qui permet de manipuler les documents PDF sous la License Apache v2.0. Le projet POI est utilisé comme API Java pour la manipulation de formats de fichiers différents tels que MS Word et MS Excel.

Les indicateurs sont ensuite présentés dans un fichier XML. Un score global peut être calculé en assignant des poids à chacun des indicateurs dans un fichier de paramètres.

Figure 1. Description du système

Figure 2. Interface représentant différentes métriques

Le service est accessible via une interface SOAP développée avec JAX-WS, l'API de Java pour le design de services web. Une interface web est également disponible et permet de charger un fichier Word ou PDF ou de copier un texte dans un formulaire. Les résultats fournis par l'interface web peuvent être sauvegardés au format XML. Contrairement à la plupart des outils existants, il permet de fournir différentes mesures (figure 2) individuellement et de les combiner en fonction des besoins avec un accès en service web.

3.2. Approches statistiques de la lisibilité : les mesures de Flesch

Dans cette section, nous présentons les mesures de Flesch, puis notre choix de mise en œuvre. Il existe deux mesures de Flesch : *Flesch Reading Ease* et *Flesch-Kincaid Grade Level*. Elles utilisent des critères similaires et sont corrélées négativement. Plus un texte obtient un score *Flesch Reading Ease* élevé, plus son score *Flesch-Kincaid Grade Level* est faible.

La mesure “Flesch Reading Ease”

Plus le score est élevé, plus le texte est facile à comprendre. Soient le nombre total de mots m , le nombre total de phrases p , le nombre total de syllabes s , pour un texte t , la mesure de *Flesch Reading Ease* se calcule comme suit :

$$FRES(t) = 206,835 - \left(1,015 * \frac{m}{p}\right) - \left(84,6 * \frac{s}{m}\right)$$

Cette mesure est employée dans des logiciels tels que Microsoft Office Word⁸.

La mesure “Flesch-Kincaid Grade Level”

Le score de la mesure *Flesch-Kincaid Grade Level* est compris dans un intervalle [0-100] et correspond à un niveau du système éducatif américain. Il reflète le nombre d’années d’éducation nécessaires pour la compréhension d’un texte donné. Il est calculé comme suit :

$$FKGL(t) = 0,39 * \left(\frac{m}{p}\right) + 11,8 * \left(\frac{s}{m}\right) - 15,59$$

Un score de 7,4 par exemple, indique que le texte est compréhensible par les étudiants de niveau scolaire 7 selon le système éducatif américain, ce qui correspond à des enfants d’environ 12-13 ans.

La mise en œuvre

Afin de calculer la mesure *Flesch Kincaid Grade Level*, nous avons isolé les syllabes du texte. Une syllabe est définie comme une unité basique du langage écrit et lu. Elle consiste en un son non interrompu qui peut être utilisé pour constituer des mots. Pour pouvoir compter les syllabes, il est nécessaire de :

- compter le nombre de voyelles,
- soustraire le nombre de voyelles muettes (ex. un ‘e’ muet à la fin des mots),
- soustraire une voyelle de chaque diphtongue.

Le nombre de voyelles restants correspond alors au nombre de syllabes.

8. <http://www.instructionalsolutions.com/blog/bid/61880/How-To-Enable-Readability-Measures-in-Microsoft-Word-2007>

Le comptage des syllabes a été comparé aux résultats fournis sur deux sites : <http://www.wordcalc.com/> et <http://www.syllablecount.com/>. Les tests ont montré que les deux sites calculent les syllabes de façons différentes. Les facteurs qui les différencient sont : l'omission du « e » muet à la fin, et l'omission des diphtongues. En prenant en considération ces différences, notre mise en œuvre aboutissait à des résultats similaires à ceux fournis par les outils proposés.

L'échelle de mesure *Flesh-Kincaid Grade Level* est directement liée à un système éducatif et aux modèles de description des ressources éducatives de type *IEEE Learning Object Metadata* ou *Dublin Core Education*⁹. Cependant *Flesh Reading Ease* fournit une mesure normalisée, qui permet donc d'être prise en compte plus facilement dans une mesure de difficulté combinant plusieurs métriques. Nous avons donc mis en œuvre la mesure *Reading Ease*.

3.3. La cohésion grammaticale

La cohésion grammaticale telle que définie par Halliday et Hasan (1976) comporte la référence, l'identification, l'ellipse et la conjonction. Nous avons mis en œuvre des métriques de Référence, Identification et Conjonction.

Les références designent les pronoms à la troisième personne (*he, she, it* et *they*) (ex. *David is a lawyer. He studied in Illinois*) et les pronoms démonstratifs (*this, these, that, those*). Les références peuvent être anaphoriques ou cataphoriques (références à un élément postérieur). Dans notre approche, nous prenons en considération les références, en détectant les pronoms mentionnés ci-dessus et en les rapportant au nombre de mots du texte.

Les identifications sont reconnues grâce à des déterminants (*the, this, that*) qui indiquent qu'un nom a été utilisé précédemment. Par exemple : *I read a book. This book was written by a famous author*. En comparant les termes qui suivent les déterminants et ceux qui suivent les articles indéfinis, il est possible d'évaluer le nombre d'identifications.

La détection des conjonctions et des adverbes conjonctifs permet également d'indiquer la difficulté d'un texte, comme suggéré par Innajih (2007) qui souligne l'impact de l'enseignement des conjonctions sur la compréhension de textes en langues étrangères. Notre système utilise un fichier externe contenant 74 conjonctions, telles que *and, but, because, however, thus* et *although*. Nous identifions le taux de conjonctions dans un texte.

9. <http://www.dublincore.org/groups/education/>

3.4. La cohésion lexicale

La cohésion lexicale telle que définie par Halliday et Hasan (1976) comporte la répétition et la collocation. Les répétitions exactes sont des occurrences multiples d'un mot dans un texte. Nous calculons les répétitions exactes de mots, hors mots vides (mots non significatifs).

La collocation est l'utilisation de synonymes au sein du même passage d'un texte, représentant une cohésion locale importante. Nous avons étendu l'analyse de la collocation en fonction non seulement des liens de synonymie entre les mots mais également en utilisant des relations sémantiques plus riches. La détection des relations sémantiques s'appuie sur la base lexicale WordNet¹⁰. Les noms, verbes, adjectifs et adverbes y sont groupés en ensembles de synonymes cognitifs (*SynSet*), chacun exprimant un concept distinct. Les *SynSets* sont liés par des relations sémantiques et des relations lexicales. Les informations lexicales y sont hiérarchisées. WordNet est utilisé pour comparer les noms entre les couples de phrases consécutives du texte. Au lieu de comptabiliser les synonymes, nous avons calculé la similarité sémantique entre les noms contenus dans des phrases contiguës.

$$Ref(t) = \frac{\sum sim(p1, p2)}{pp}$$

$sim(p1, p2)$ représente la similarité entre chaque paire de phrases contiguës du texte (pp). Elle est le résultat de la moyenne de la similarité entre chaque nom de la première phrase ($n(p1)$) avec chaque nom de la deuxième phrase ($n(p2)$). Nous calculons ensuite la moyenne des similarités trouvées pour chaque paire de phrases du texte.

Nous utilisons l'API RiTa.WordNet pour communiquer avec WordNet. Nous recherchons d'une part le type d'un mot, c'est-à-dire s'il s'agit d'un nom ou d'un pronom par exemple et, d'autre part, la similarité sémantique entre deux mots. La similarité sémantique est calculée en utilisant l'algorithme *Lowest Common Ancestor* (LCA) issu de la théorie des graphes. LCA retrouve le parent commun entre deux mots dans le graphe lexical, la distance minimale entre le mot dans la première phrase et le parent commun (dtp) et la distance minimale entre le parent commun et la racine de l'ontologie (dpr) pour calculer la similarité sémantique (SS) entre deux mots (t).

$$SS(t1, t2) = 1 - \left(\frac{dtp}{(dpr + dtp)} \right)$$

10. <http://wordnet.princeton.edu/>

3.5. Approche sémantique : l'analyse statistique de la rareté des mots

L'objectif de l'analyse statistique est d'identifier le taux de mots rares dans le texte. Nous avons investigué deux approches : Google N-Grams et Google Custom Search.

Un n-gram est défini par une séquence contigüe de n éléments. Il peut s'agir de lettres, de syllabes, de mots, ou de n'importe quelle autre entité linguistique. Une séquence de 1 élément est un *unigram*, une séquence de deux éléments est un *bigram*. Google a créé des jeux de données contenant des mots et leurs occurrences à partir d'environ 5 millions de livres. Google N-Grams fournit une date à laquelle les mots ont été utilisés. Cela permet de définir la rareté des mots en fonction de dates et non de prendre en considération l'ensemble des ressources disponibles. Il est en effet possible que des personnes de générations différentes par exemple aient pu se familiariser avec des mots différents. Néanmoins, les jeux de données sont très larges et il est nécessaire d'installer une infrastructure additionnelle pour supporter l'accès aux données de Google N-Grams.

Google Custom Search en revanche est plus léger à mettre en œuvre. Il donne accès via une API aux résultats de recherches de Google sur une liste prédéfinie de sites. Google Custom Search ne permet pas d'adapter le corpus à partir duquel sont identifiés les mots rares et les mots communs en fonction de dates. En revanche, il est possible d'adapter la liste des sites dans lesquels Google Custom Search recherche, notamment en fonction du contexte éducationnel. Les principales limitations du système auxquelles nous avons été confrontés concernent le coût des requêtes, la limitation à un nombre maximal de requêtes et la nécessité de réinitialiser le système de manière régulière.

3.6. Le profil de l'apprenant

L'objectif de ce module est de comparer les mots dans le texte avec les mots de vocabulaire déjà appris par l'utilisateur, suivant l'approche décrite par Heilman et Eskenazi (2008).

Pour représenter les termes connus, nous avons étendu l'ontologie PIMO (*Personal Information Model*) (Sauermaun *et al.*, 2007), issue du projet NEPOMUK¹¹. L'ontologie PIMO est utilisée pour représenter les concepts que manipule un utilisateur donné dans son environnement informationnel, par exemple ses projets, ses tâches, ses contacts, ses courriels, ses fichiers et d'autres ressources. L'objectif de PIMO est aussi de représenter le « modèle mental » d'un utilisateur. Or le modèle mental d'un utilisateur inclut aussi des mots et le sens spécifique qui leur est donné par chaque utilisateur (Naudet *et al.*, 2010). Les modèles de description des profils

11. <http://nepomuk.semanticdesktop.org/nepomuk/>

d'apprenants (ex., *IMS Learner Information Package*¹²) n'offrent pas actuellement de mécanisme spécifique pour représenter ce type de connaissances informelles.

```
<rdf:RDF xmlns="http://[domaine]/people/john#"
  xml:base="http://[domaine]/people/john#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:pimo="http://www.semanticdesktop.org/ontologies/2007/11/01/pimo#"
  xmlns:john="http://[domaine]/people/john#"
  xmlns:pkm="http://[domaine]/learning/pkm#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">

  <pimo:PersonalInformationModel rdf:about="JohnPIM"/>
    <pimo:Person rdf:about="John">
      <pimo:isDefinedBy rdf:resource="JohnPIM"/>

      <pkm:hasKnowledgeOfWord>Donkey</pkm:hasKnowledgeOfWord>
      <pkm:hasKnowledgeOfWord>Fish</pkm:hasKnowledgeOfWord>
    </pimo:Person>

</rdf:RDF>
```

Figure 3. Illustration du modèle RDF de PIMO étendu

Un mot doit pouvoir être représenté soit comme un littéral, soit comme un concept, c'est-à-dire une URI identifiant le mot dans un contexte global (ex., URI utilisée dans WordNet). Nous avons dans un premier temps implémenté la liste des mots comme littéraux (figure 3).

La liste de l'ensemble des mots ou expressions connues de l'utilisateur est utilisée dans le cadre de l'apprentissage des langues étrangères (Brown *et al.*, 2005 ; Heilman, Eskenazi, 2008). Néanmoins, il est également possible de travailler sur des listes de termes et expressions issues d'un parcours de lecture et/ou de l'utilisation de référentiels probables de termes et expressions connus en fonction d'un profil d'apprenant (par exemple, par rapport aux élèves d'un certain niveau ou d'une catégorie d'âge) (Miltakaki *et al.*, 2008).

Le système met donc en œuvre 8 métriques pour prédire la difficulté des textes (tableau 1). Certaines sont simplifiées et peuvent être améliorées facilement.

12 <http://www.imsglobal.org/profiles/>

D'autres sont déjà élaborées et proposent une approche plus riche de la difficulté des textes.

Tableau 1. Métriques utilisées pour prédire la difficulté de lecture d'un texte

Mesure	Indicateur
Flesh Kincaid Reading Ease	lisibilité sémantique et syntaxique
Similarité des mots utilisés dans des phrases contiguës	cohérence
Références	cohésion grammaticale
Conjonctions	cohésion grammaticale
Identifications	cohésion grammaticale
Répétitions exactes	cohésion lexicale
Analyse statistique de la rareté des mots	lisibilité sémantique personnalisable
Mots connus	lisibilité sémantique personnalisée

4. Calibration et validation des métriques

Le système met en œuvre des mesures classiques de la difficulté des textes pour l'enseignement dans une seule interface, mais aussi des outils pour permettre d'inclure des approches modernes de la difficulté, liées à la cohésion des textes (en étendant l'approche traditionnelle avec la prise en considération de l'environnement sémantique des concepts), fondées sur le contexte éducatif (typiquement la rareté des mots dans les lectures d'élèves d'une certaine tranche d'âge) et le profil de chaque apprenant (via la traçabilité des mots connus, qui pourra être étendue à des éléments grammaticaux acquis ou non acquis en particulier dans le cadre de l'apprentissage des langues). Les mesures fournies par le système doivent être ensuite combinées et adaptées en fonction du contexte éducatif afin d'optimiser l'efficacité des mécanismes de détection de la difficulté. Nous avons évalué les mesures mises en œuvre avec deux jeux de textes précalibrés de styles différents.

4.1. Les jeux de test

Nous avons rassemblé deux jeux de textes en langue anglaise. Ils représentent des styles différents et leur référentiel de calibration diverge. Le jeu de textes de la NSDL est relativement homogène et classifié par rapport à des niveaux scolaires américains, ce qui correspond au cas pour lequel ont été conçues les métriques de lisibilité. L'autre jeu de textes est en revanche plus hétérogène.

La National Science Digital Library

La *National Science Digital Library* (NSDL)¹³ est une initiative américaine qui vise à créer des ressources numériques pour l'éducation scientifique (sciences, technologie, ingénierie et mathématiques). De nombreux projets ont vu le jour dans le cadre des programmes de financement de la NSDL. Les ressources qui ont été créées furent ensuite intégrées dans une bibliothèque numérique distribuée avec des métadonnées éducatives, notamment l'indication des propriétés LOM *Audience* ou Dublin Core *Audience*. Nous utilisons ces propriétés créées manuellement dans la plupart des cas par les contributeurs des ressources, comme référentiels de leur calibration. Le public cible des ressources est défini en fonction des catégories : *elementary school, middle school, high school* et *higher education*.

Nous avons utilisé 32 textes issus de ressources éducationnelles répertoriées dans la NSDL, majoritairement des sites. Les textes comprennent entre 250 et 500 mots.

Livres pour enfants

Nous avons rassemblé des extraits de livres pour enfants classés selon une échelle similaire à celle utilisée dans le *Common Core State Standard for Reading*. Nous avons évalué 10 extraits de textes de niveaux 0-3 ans à 14 ans et plus. Les extraits furent choisis manuellement afin d'éliminer les textes qui ne comporteraient pas ou peu de phrases mais plutôt des suites de mots ou onomatopées, notamment parmi les textes destinés aux enfants de 0 à 3 ans. Ce sont des textes de fiction. Les extraits sont de longueur différente (entre 32 et 523 mots).

Les échantillons de textes utilisés ne permettent pas de valider l'ensemble des contextes, par exemple celui de l'apprentissage des langues. Néanmoins, ils représentent des méthodes de calibration et des styles très différents.

4.2. La calibration des mots rares

L'analyse statistique de la rareté des mots prend en considération la distribution des mots dans un corpus qui permet la calibration de l'outil. Afin de définir ce qu'est un mot rare, nous avons configuré Google Custom Search Engine avec quatre sites qui ont des caractéristiques linguistiques différentes : le site Google Books qui contient des millions d'ouvrages, le site Facebook qui correspond davantage à la langue très informelle utilisée sur les réseaux sociaux, ainsi que le site éducatif de la BBC Learning pour l'anglais britannique et le site sur l'enseignement de l'anglais du département d'État américain¹⁴.

13. <http://nsdl.org/>

14. <http://books.google.com/books> ; <http://www.facebook.com/>
<http://www.bbc.co.uk/learning/> ; <http://exchanges.state.gov/englishteaching/resources-et.html>

Le jeu de la NSDL a été divisé en 2 parties. Un premier jeu de 12 textes a été utilisé pour la calibration du mécanisme de calcul des mots rares (jeu NSDL-A). Un second jeu de 20 textes a été utilisé pour la validation de la pertinence de cette approche (jeu NSDL-B).

Nous avons calculé la distribution des mots dans les textes du jeu NSDL-A (figure 4). Après avoir éliminé les mots vides et dédoublé les mots utilisés dans chacun des textes, nous avons déterminé un premier quartile relatif aux mots rares. Il correspond à une prédiction de 163.000 réponses de Google Custom Search. Nous définissons donc un mot rare en fonction du premier quartile des mots utilisés hors mots vides dans un corpus de calibration.

Figure 4. Distribution de la fréquence des mots hors mots vides dans le jeu NSDL-A

Nous avons ensuite vérifié que la fréquence des mots reflétait effectivement le niveau de difficulté tel qu'encodé dans les textes de la NSDL. tableau 2 présente les ratios de mots rares pour les textes du jeu NSDL-B. Le nombre moyen de mots rares augmente effectivement avec le niveau scolaire encodé dans les métadonnées de la NSDL.

4.3. Les corrélations et l'efficacité des mesures

Les métriques non personnalisées ont été testées avec les deux jeux de textes. Puisque la mesure *Flesh-Kincaid Grade Level* est conçue pour identifier des niveaux scolaires, il n'est pas surprenant que ses résultats soient bien corrélés dans le cas du jeu de textes de la NSDL. Le tableau 3 présente la moyenne des différentes métriques pour le jeu de textes de la NSDL. Il comprend 9 textes classés *elementary*, 8 classés *middle school*, 7 classés *high school* et 8 classés *higher education*. Les moyennes pour les métriques *Flesh Kincaid*, Répétitions et Références sont les plus cohérentes avec les annotations des textes de la NSDL. Le coefficient de corrélation

PEARSON est très élevé pour Flesh Kincaid (-0,84) et relativement élevé pour Références (-0,57). Concernant l'ensemble de textes *Livres pour enfants*, seule la mesure de Répétition est corrélée au niveau identifié sur le site (-0,60), alors que pour Références, Identifications et Répétitions les moyennes des scores pour les différentes tranches d'âges sont cohérentes. En dépit du nombre limité de textes, cela suggère que la mesure la plus pertinente est différente en fonction de l'ensemble de données. Néanmoins, la moyenne des scores pour les diverses tranches d'âge est significativement différente de celles obtenues pour les textes NSDL, par exemple concernant la similarité.

Tableau 2. Ratios de mots rares dans les textes du jeu NSDL-B

Niveau	Nombre de mots rares	Nombre de mots moyennement rares	Nombre de mots non rares	Total	Pourcentage de mots rares	Nombre moyen de mots rares
Element.	36	72	138	246	15 %	0,16
	35	58	57	150	23 %	
	14	24	36	74	19 %	
	20	49	68	137	15 %	
	5	21	40	66	8 %	
Middle	11	33	49	93	12 %	0,18
	9	20	16	45	20 %	
	26	69	47	142	18 %	
	6	14	19	39	15 %	
	22	33	40	95	23 %	
High	9	46	42	97	9 %	0,22
	25	47	46	118	21 %	
	33	54	35	122	27 %	
	37	44	35	116	32 %	
	23	57	44	124	19 %	
Higher education	23	46	19	88	26 %	0,24
	23	39	35	97	24 %	
	25	50	33	108	23 %	
	8	19	12	39	21 %	
	35	55	54	144	24 %	

Il est également possible que les métadonnées aient été créées avec un objectif différent, éducatif dans le cas de la NSDL, commercial dans le cas de *Livres pour Enfants*, pouvant conduire à inclure des ressources plus faciles pour des enfants plus âgés. Enfin, la comparabilité entre les deux jeux de textes est limitée par la différence de genre des textes.

Tableau 3. Moyennes des scores obtenus pour les textes NSDL

	Flesch	Simil.	Ref.	Conj.	Ident.	Repet.
NSDL Elementary	60,82	0,49	3,79	8,55	0,02	0,13
NSDL Middle	49,55	0,45	3,66	9,54	0,57	0,15
NSDL High School	34,34	0,46	2,82	7,69	0,50	0,16
NSDL Higher Ed.	18,33	0,50	1,87	9,09	0,26	0,17
LE 3+	72,84	0,26	5,85	7,85	0,21	0,38
LE 9+	76,16	0,41	4,61	7,89	0,13	0,11
LE 14+	71,98	0,40	3,76	7,85	0,00	0,08

La performance des métriques de cohésion pour prédire la lisibilité d'un texte est débattue. La cohésion lexicale est relativement performante dans le cas du groupe de textes *Livres pour Enfants* et la cohésion grammaticale l'est dans le cas des textes de la NSDL. Une étude avec des groupes de textes plus larges serait nécessaire pour affiner ces résultats. Néanmoins les groupes de textes utilisés permettent de démontrer les possibilités offertes par la combinaison des différentes métriques et l'hétérogénéité des résultats pour différents types de textes.

5. L'utilisation des textes dans l'éducation et l'étude PISA

De nombreux textes ont été créés pour un environnement qui n'est pas éducatif mais sont réutilisés en environnement éducatif. C'est le cas de tous les textes illustratifs des manuels scolaires (documents ou ressources documentaires). Or, l'utilisation de manuels numériques et la conception de manuels scolaires qui orientent vers des documents publiés sur Internet (Séré *et al.*, 2010) conduit à de nouvelles difficultés (textes dont la lisibilité n'a pas nécessairement été évaluée par un pédagogue, même de manière informelle) mais aussi à de nouvelles opportunités d'inclure des textes de nature différente. Le contexte le plus critique pour l'inclusion de documents dans les ressources éducatives est celui des examens ou tests. En effet, le choix d'un document peut y avoir un impact direct sur la réussite d'un examen.

Le test PISA (*Programme for International Student Assessment*) est conduit tous les 3 ans sous l'égide de l'OCDE afin d'évaluer et comparer les performances des élèves de 15 ans. Il comporte 3 parties, Lecture, Mathématiques et Sciences. Nous

avons donc examiné l'utilisation de documents textuels dans l'échantillon de questions PISA 2009. L'échantillon PISA 2009 contient 267 questions, dont 65 visent à mesurer les compétences en lecture. Néanmoins, plusieurs questions se rapportent aux mêmes textes. Par exemple la Reading Unit 2 montre le texte d'un flyer pour inciter les employés d'une entreprise à se faire vacciner contre la grippe. La question 2.1 porte sur l'identification d'un des éléments du programme de vaccination, la question 2.2 porte sur le style du texte, la question 2.3 sur le message du texte, enfin la question 2.4 extrait deux lignes du texte et demande d'évaluer l'utilité et la confusion potentielle que la seconde ligne peut créer.

Certaines des questions de tests sont précalibrées avec des échantillons d'élèves. La *One Dimensional Item Response Theory* (Reise *et al.*, 2005) est utilisée pour garantir la validité des questions et l'absence de biais. Elle permet de définir la difficulté d'une question, laquelle peut se rapporter à un texte. Elle ne mesure cependant pas la difficulté du texte lui-même.

Parmi les ressources textuelles incluses dans les questions de PISA 2009 (tableau 4), se trouvent notamment 10 articles de journaux ou de newsletter, 3 lettres et des recettes de cuisine.

Tableau 4. Nombre de ressources textuelles incluses dans chaque section de PISA 2009

	Nombre de textes
Lecture	17
Science	14
Mathématiques	0
Total	31

Les textes ont des rôles différents dans les questions de test. Nous avons classé ces différents rôles pour les ressources réutilisées comme documents dans une question de test (Foulonneau *et al.*, 2011) en nous basant notamment sur le modèle proposé par Suvorov (2009) : *stimulus* lorsque la question porte sur le document et/ou que le document est indispensable pour répondre à la question, *supporting* lorsque le document peut aider à répondre à la question mais qu'il est possible de répondre à la question sans le document, *context/illustration* lorsque le document n'apporte pas d'élément direct sur la réponse, enfin *réponse*, lorsque la réponse doit être fournie dans ou à l'aide du document lui-même (typiquement lorsqu'il fait partie des options). La partie Lecture propose des textes mais aussi des tableaux ou des figures. Néanmoins, tous les textes ont une fonction de stimulus dans au moins une question de l'unité. En revanche, les rôles assignés aux textes présentés dans l'unité Science sont beaucoup plus variés : *stimulus*, *supporting* et le plus souvent *context*.

Nous avons extrait 6 textes de la section de lecture utilisés dans des questions calibrées. Nous avons utilisé comme référentiel un score de difficulté correspondant à la moyenne de la difficulté des questions se rapportant au texte. Dans le cas de la Reading Unit 3, elle contient 2 textes, mais l'une des questions de l'unité ne se rapporte qu'à un seul des textes, les autres se rapportant aux deux textes.

Les textes reflètent des contextes de lecture variés. Ils peuvent être littéraires (ex. théâtre), journalistiques (ex. articles) ou informationnels issus de la communication d'entreprise (ex. note d'information interne). Les textes sont de longueurs différentes. Néanmoins, les questions ont été conçues pour des élèves d'environ 15 ans et calibrées avec cette population. La difficulté des textes peut être différente mais ils s'adressent à des publics similaires. La difficulté calculée en fonction de la théorie de réponse à l'item se rapporte à la question et non au texte lui-même. Les textes de PISA devraient donc avoir soit un niveau de difficulté similaire, c'est-à-dire être adaptés à des personnes de 15 ans.

Tableau 5. Amplitude des métriques sur les différents jeux de données

Textes Métrique	Flesch	Simil.	Ref.	Conj.	Ident.	Repet.
NSDL	91,66	0,37	5,66	7,65	3,19	0,16
Livres pour enfants	37,74	0,56	6,08	10,42	0,86	0,56
PISA	31,23	0,10	3,71	2,14	0,82	0,13

L'ensemble des textes du jeu de textes PISA étant calibrés pour des élèves du même âge, les métriques qui permettent de distinguer des niveaux différents de lecture sont présumées moins discriminantes. Les métriques sont plus homogènes (amplitude inférieure entre le score minimal et le score maximal obtenu par chacun des textes) que pour les autres ensembles de textes, comme illustré dans le tableau 5. Il est important de noter que le choix d'un texte comme document dans une question de test est critique car il peut inclure un biais vers une population particulière, par exemple, via des références culturelles spécifiques ou accroître la charge cognitive de la question, entraînant un risque pour la validité de la question. La difficulté ne constitue donc qu'une des dimensions à prendre en considération dans le choix d'un document pour une ressource éducative.

6. Conclusion

Les outils d'évaluation de la difficulté des textes sont aujourd'hui orientés vers un usage par les enseignants ou les écoles, éventuellement par les élèves eux-mêmes. Nous proposons d'intégrer ce type d'approches pour les détenteurs de collections qui souhaiteraient en faciliter la réutilisation dans un contexte éducatif,

voire la mise en œuvre d'un service qui permettra l'annotation à vocation éducative de larges collections sur internet.

Le système présenté doit être évalué avec des collections plus larges et des textes plus longs. Néanmoins, nous avons illustré les capacités des algorithmes de détection de la difficulté des textes et les potentialités de leur extension, d'une part par rapport à la personnalisation des métriques et, d'autre part, avec l'intégration de ressources du web sémantique et du web. Ce type d'outils permet d'annoter des textes en fonction d'un contexte éducatif, au moment où les ressources numériques et la capacité tant des élèves que des parents à s'orienter dans les ressources éducatives est de plus en plus sollicitée.

De nombreux travaux ont mis en évidence l'intérêt d'une approche personnalisée de l'éducation, des parcours d'apprentissage mais aussi des modes d'interaction et de l'ensemble des paramètres qui peuvent correspondre au profil cognitif de l'apprenant. Dans ce cadre, l'évaluation de la difficulté et donc de la pertinence ou de la valeur formative d'un texte doit être mesurée par rapport aux acquis précédents de l'apprenant (ex., en estimant le vocabulaire connu en fonction des textes auxquels il a déjà été exposé). Ce travail devra donc déboucher sur la mise en œuvre concrète des mesures de difficulté des textes pour la sélection semi-automatique et/ou la recommandation de textes grâce à une annotation sémantique spécifique au contexte éducationnel.

Ces travaux s'inscrivent dans le cadre d'un ensemble d'études effectuées sur la réutilisation des ressources multimédia dans un contexte éducatif, notamment dans le domaine de l'évaluation des compétences (Foulonneau *et al.*, 2011) et de la génération semi-automatique de questions de tests (Foulonneau *et al.*, 2012). Alors que l'impact des ressources multimédia sur les ressources éducatives a été étudié (ex. Van de Vijver, 2004 ; Suvorov, 2009), il n'existe pas de cadre global concernant leur rôle et surtout leurs critères de sélection.

C'est à partir de la définition d'un cadre clair sur la sélection des ressources pour l'éducation, comme les standards du *Common Core* qu'il est possible de créer des métadonnées éducatives pour des ressources multimédia et faciliter ainsi leur réutilisation. La réflexion sur la lisibilité des textes comme ressources éducatives est plus développée dans le système américain que dans les systèmes européens. François *et al.* (2012) proposent d'utiliser la *Common European Framework of Reference for Languages* publiée par le Conseil de l'Europe en 2001 comme échelle pour évaluer la difficulté des textes dans le cadre de l'apprentissage des langues. Néanmoins, l'approche de l'éducation comprend aujourd'hui des modes plus informels. Les approches de la difficulté des textes ne sont pas seulement pertinentes pour évaluer si un texte présentera des difficultés pour un lecteur particulier ou permettra de se familiariser avec certaines formes grammaticales en apprentissage des langues étrangères par exemple. L'ensemble des textes que nous lisons sur internet, dans un cadre professionnel ou de loisirs appartiennent aussi à un cadre d'apprentissage informel (ex. Le Deuff, 2010). L'observation des caractéristiques des textes publiés sur internet ou dans les journaux doit permettre d'optimiser l'apprentissage informel.

Ces travaux seront poursuivis, d'une part en améliorant les métriques proposées avec des outils sémantiques (ex. pour la reconnaissance d'entités nommées) et la définition de corpus types pour la calibration de catégories d'apprenant en fonction des mots qu'ils utilisent ou lisent, et d'autre part en les adaptant à différentes langues, comme suggéré pour le chinois par Fry (2002) et pour le français par François *et al.* (2012). Finalement, nous évaluerons les métriques sur différentes collections, par exemple des collections de textes proposées dans Europeana, mais aussi des ressources orales avec donc un niveau de langage différent, par exemple dans le cas d'une vidéo utilisée pour l'apprentissage d'une langue étrangère.

Le service développé sera par la suite mis en ligne afin de fournir un service à la communauté pour évaluer la pertinence de collections de ressources multimédia dans un cadre éducatif et éventuellement intégrer de nouvelles métriques. Surtout, il doit permettre d'ajouter des métadonnées à des collections de ressources textuelles afin de faciliter leur réutilisation dans un environnement éducatif.

Bibliographie

- Beaudet C. (2001). Clarté, lisibilité, intelligibilité des textes : un état de la question et une proposition pédagogique. *Recherches en rédaction professionnelle*, vol. 1, n° 1, p. 1-19.
- Brown J., Eskenazi M. (2004). Retrieval of authentic documents for reader-specific lexical practice. *Proceedings of InSTIL/ICALL Symposium 2004*. Venice, Italy.
- Brown J.C., Frishkoff G.A., Eskenazi M. (2005). Automatic question generation for vocabulary assessment. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Association for Computational Linguistics Stroudsburg, PA, USA, p. 819-826.
- Carrell P.L. (1982). Cohesion Is Not Coherence. *TESOL Quarterly*, vol. 16, n° 4, dec, p. 479-488.
- DuBay W.H. (2004). *The Principles of Readability*. Impact Information, Costa Mesa, California.
- Foltz P.W., Kintsch W., Landauer T.K. (1998). The Measurement of Textual Coherence with Latent Semantic Analysis. *Discourse Processes*, vol. 25, n° 2-3, p. 285-307.
- Foltz P.W., Laham D., Landauer T.K. (1999). The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, vol. 1, n° 2.
- Foulonneau M., Grouès V. (2012). Common vs. Expert knowledge: making the Semantic Web an educational model. *Proceedings of the 2nd International Workshop on Learning and Education with the Web of Data (LiLe-2012 at the WWW-2012 conference)*, CEUR workshop proceedings, vol. 840.
- Foulonneau M., Ras E., Latour T. (2011). Reusing multimedia resources in assessment items – practices and impact. *Computer Assisted Assessment conference Southampton*.
- François T., Fairon C. (2012, in press). An "AI readability" formula for French as a foreign language. *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

- Fry E. (2002). Readability versus Leveling. *The Reading Teacher*, vol. 56, n° 3, nov., p. 286-291, International Reading Association.
- Grunwald T., Corsbie-Massay C. (2006). Guidelines for Cognitively Efficient Multimedia Learning Tools: Educational Strategies, Cognitive Load, and Interface Design. *Academic Medicine*, vol. 81, n° 3, p. 213-223.
- Halliday M.A.K., Hasan R. (1976). *Cohesion in English*, Harlow, Longman.
- Hearst M.A. (1997, mars). TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Comput. Linguist.*, vol. 23, n° 1, p. 33-64.
- Heilman M., Eskenazi M. (2008). Self-Assessment in Vocabulary Tutoring. *Ninth International Conference on Intelligent Tutoring Systems*, Lecture Notes in Computer Science, vol. 5091, Springer, p. 656-658.
- Heilman M., Collins-Thompson K., Eskenazi M. (2008). An Analysis of Statistical Models and Features for Reading Difficulty Prediction. *Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications*, Columbus, Ohio, USA, June, Association for Computational Linguistics, p. 71-79.
- Heilman M., Zhao L., Pino J., Eskenazi M. (2008). Retrieval of Reading Materials for Vocabulary and Reading Practice. *Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications*, Columbus, Ohio, USA, June, Association for Computational Linguistics, p. 80-88.
- Huff L. (2008). Review of *netTrekker d.i.*. *Language Learning & Technology*, vol. 12, n° 2, June, p. 17-25.
- Innajih A.A. (2007). *The impact of textual cohesive conjunctions on the reading comprehensive of 4th year English major students in Libyan universities*. Thèse de doctorat, University of Newcastle upon Tyne.
- Le Deuff O. (2010). Réseaux de loisirs créatifs et nouveaux modes d'apprentissage. *Distances et savoirs*, vol. 8, n° 4, p. 601-621.
- Mayer R.E. (2005). Cognitive Theory of Multimedia Learning. *The Cambridge Handbook of Multimedia Learning*. Richard mayer (Ed.), Cambridge University Press, New York.
- McLaughlin G.H. (1969). SMOG grading: A new readability formula. *Journal of Reading*, vol. 12, n° 8, p. 639-646.
- McNamara D.S., Louwerse M.M., McCarthy P.M., Graesser A.C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, vol. 47, n° 4, p. 292-330.
- Miltsakaki E., Troutt A. (2007). Read-X: Automatic Evaluation of Reading Difficulty of Web Text. *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2007*, n° 1, p. 7280-7286.
- Miltsakaki E., Troutt A. (2008). Real-Time Web Text Classification and Analysis of Reading Difficulty. *Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications*, Columbus, Ohio, USA, June, Association for Computational Linguistics, p. 89-97.
- Morris J., Hirst G. (1991). Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, vol. 17, n° 1.

- Naudet Y., Groues V., Foulonneau M. (2010). Introduction to Fuzzy-Ontological Context-Aware Recommendations in Mobile Environments. *1st International Workshop on Adaptation, Personalization and REcommendation in the Social-semantic Web*, 7th Extended Semantic Web Conference (ESWC 2010), 31 May Heraklion, Greece.
- Nelson J., Perfetti C., Liben D., Liben M. (2011). *Measures of Text Difficulty: Testing their Predictive Value for Grade Levels and Student Performance*. Rapport technique, Student Achievement Partners.
- Ras E., Plichart P., Latour T. (2010). Addressing the needs of today's e-learning environments via an open and versatile architecture for assessment. *ICL 2010 Conference*, September 15-17, 2Hasselt, Belgium.
- Sauermann L., van Elst L., Dengel A. (2007). PIMO – a framework for representing personal information models. *I-SEMANTICS'07*, Graz, Austria, p. 270-277.
- Schmidt K., Schmitz H.-C., Wolpers M. (2011). Developing a Network of Cultural Heritage Objects Repositories for Educational Purposes. *Proceedings Fifth International Conference on Metadata and Semantic Research (MTSR'11)*. Yasar University 12-14 October 14, 2011, Izmir, Turkey, Communications in Computer and Information Science, Springer, vol. 240, part 3, p. 337-348.
- Séré A., Bassy A.M. (2010). *Le manuel scolaire à l'heure du numérique - Une « nouvelle donne » de la politique des ressources pour l'enseignement*. Rapport n° 2010-087 au ministre de l'Éducation nationale.
- Sheehan K.M., Kostin I., Futagi Y., Flor M. (2010). *Generating automated text complexity classifications that are aligned with targeted text complexity standards*. Rapport de recherche ETS RR-10-28, Educational Testing Service.
- Smith D., Stenner A.J., Horabin I., Smith M. (1989). *The Lexile scale in theory and practice*. Washington, DC, MetaMetrics.
- Suvorov R. (2009). Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format. *Developing and evaluating language learning materials*, C.A. Chapelle, H.G. Jun, & I. Katz (Eds.). Ames, IA: Iowa State University, p. 53-68.
- Van de Vijver F., Tanzer N.K. (2004). Bias and equivalence in cross-cultural assessment: an overview. *Revue européenne de psychologie appliquée*, vol. 54, n° 2, p. 119-135.
- Youmans G. (1991). A New Tool for Discourse Analysis: The Vocabulary-Management Profile. *Language* vol. 67, n° 4, December 1, p. 763-789.