La prédiction automatisée de la difficulté lexicale par la combinaison de ressources et de méthodes d'apprentissage automatisé

Thomas François

Chargé de recherche FNRS

Séminaire du Cental, UCL

6 Novembre 2015

Plan

Plan

La problématique de la connaissance lexicale

Importance de la composante lexicale en L2

- Le lien entre la connaissance lexicale et la compréhension à la lecture est largement reconnu, tant pour les lecteurs L1 que L2. [Alderson, 1984, Carrell, 1998, Koda, 2005]
- En lisibilité, les variables prédisant la difficulté du lexique expliquent souvent le mieux la difficulté des textes [Chall and Dale, 1995]

Deux corrolaires :

- Développer la connaissance lexicale est donc intéressant chez un apprenant d'une L2.
- De même, évaluer sa connaissance lexicale à un moment donné permet de le situer par rapport à un parcours d'apprentissage/textes.

La connaissance lexicale et la lecture

- La connaissance lexicale est un facteur plus corrélé avec la compréhension que d'autres facteurs tels :
  - la conscience morphologique
    [Ulijn and Strother, 1990, Koda, 1989]
  - les stratégies de lecture [Haynes and Baker, 1993]
- [Hu and Nation, 2000] postulent l'existence d'un seuil lexical : pour qu'un texte soit bien compris par un lecteur, celui doit connaître X% des mots du texte.

La connaissance lexicale et la lecture

Etudes	seuil	taille du vocabulaire
[Hu and Nation, 2000]	> 95%	/
[Hirsh and Nation, 1992]	98%	5000 familles de mots
[Nation, 2006]	98%	8000-9000 familles de mots
[Laufer and Ravenhorst-Kalovski, 2010]	98%	6000-8000 familles de mot
	95%	4000-5000 familles de mot

TABLE : Taille du vocabulaire nécessaire pour atteindre le seuil lexical de la compréhension.

seuil lexical	couverture lexicale	type de compréhension
optimal	98%	lecture indépendante
minimal	95%	lecture assistée

TABLE : Deux seuils lexicaux définis par Laufer et Ravenhorst-Kaloski (2010)

L'acquisition de la connaissance lexicale

Question 1 : que signifie connaître un mot ?

- Plusieurs catégorisations [Anderson and Nagy, 1991, Nation, 2001]
- [Nation, 2001] distingue 3 grandes dimensions :
  - Forme (prononciation, orthographe, morphologie)
  - Sens (lien entre forme et ses différents sens, associations sémantiques, etc.)
  - Usage (collocations, fonctions grammaticales, contraintes de registre, de fréquence, etc.)
- Ces différentes caractéristiques sont acquises indépendamment [Schmitt, 1998]

L'acquisition de la connaissance lexicale

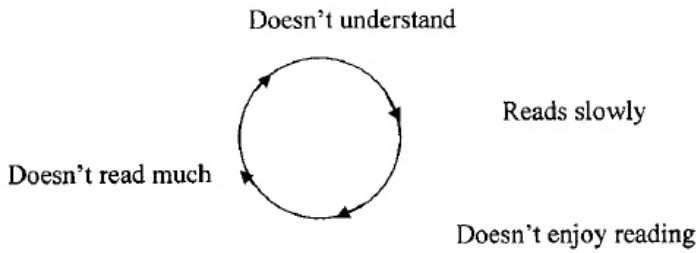
Question 2 : comment apprend-on de nouveaux mots ?

- En L1, les enfants apprennent environ 2000 à 3000 nouveaux par an, en les rencontrant dans divers contextes [Nagy and Herman, 1987]
- "incidental learning from context accounts for a substantial proportion of the vocabulary growth that occurs during the school years" [Nagy et al., 1985, 233].
- En L2, on retrouve plusieurs positions [Coady, 1997a] :
  - Acquisition à partir du contexte seul, en particulier la lecture, quand il y compréhension (Input Hypothesis de [Krashen, 1989])
  - Acquisition à partir du contexte, mais l'usage de stratégies est parfois nécessaire (mnémotechnique, structure, etc.).
  - Instruction + Contexte, l'instruction explicite étant surtout nécessaire pour le vocabulaire de base.
  - Apprentissage basé uniquement sur des activités de classe.

Lien entre la connaissance lexicale et l'acquisition

- Il semblerait que la lecture joue un rôle important dans l'acquisiton de nouveaux mots, pour autant qu'il y ait compréhension.
- Or, [Hu and Nation, 2000], parmi d'autres, montrent que l'existence d'une couverture lexicale est nécessaire pour cette compréhension.
- **Paradoxe du débutant** [Coady, 1997b],

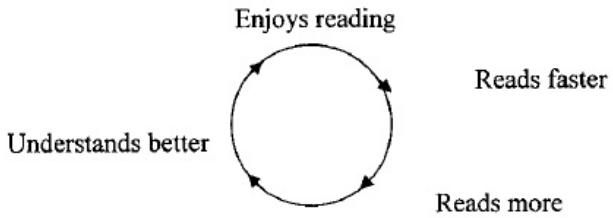
  qui peut entraîner un cercle vicieux.



Comment briser ce cercle vicieux ?

Vers un cercle vertueux

- Un moyen de sortir du cercle vicieux est de proposer des textes adaptés à l'apprenant, cad. avec 1-5% de mots inconnus.
- Cela soulève deux problèmes concrets :
  - Estimer le niveau de connaissance lexicale de l'apprenant par rapport à des textes données.
    $\rightarrow$ [Nation, 2006] indiquent une couverture lexicale idéale, mais pas pour un texte donné.
  - Déterminer quels sont les nouveaux mots à proposer à l'apprenant.



Objectifs de notre présentation

1. Proposer une cartographie de l'usage des mots dans un corpus de L2 (FLELex, SVALex, etc.)
   - Peut servir à identifier des priorités dans l'acquisition du vocabulaire
   - Pourrait être utilisé pour prédire, pour une texte, les mots inconnus d'un apprenant d'un niveau donné
2. Essayer de prédire la complexité des mots à partir de leurs caractéristiques lexicales (ReSyf)
3. Approche personnalisée de la prédiction de la connaissance lexicale [Tack, 2015].

Plan

L'existant : listes de fréquence

- De nombreuses listes de fréquences existent, à vocation linguistique, didactique, psycholinguistique, etc.
- [Thorndike, 1921] est une des premières : liste de 10 000 mots avec leurs fréquences collectées sur un corpus de 4 500 000 mots.
- Pour le français : [Gougenheim et al., 1964, New et al., 2004, Lonsdale and Le Bras, 2009]
  $\rightarrow$ Ces listes sont définies à partir de textes authentiques (pour L1).
- Plusieurs défauts à cette approche :
  - L'état de la langue L1 ne correspond pas nécessaire à l'interlangue d'un apprenant
  - [Michéa, 1953] souligne que les "mots disponibles" ne sont pas correctement estimés (ex. plafond, dentifrice, etc.).
  - Problème : comment transformer des fréquences en niveaux scolaires ?

Les listes de fréquences ne sont pas réellement des ressources gradués en fonction de niveaux scolaires !

Listes graduées

- Il existe une liste graduée pour le français L1 : Manulex [Lété et al., 2004] :
  - Il contient environ 23 900 lemmes dont la distribution par niveau a été estimée sur des manuels de primaire.
  - Leur corpus inclut 54 manuels de la CP (6 ans) à la CM2 (11 ans).
  - La ressource comprend 3 niveaux : CP = 1 ; CE1 = 2, et le 3 s'étend de la CE2 à la CM2.

Mot	Pos	Niveau 1	Niveau 2	Niveau 3
pomme	N	724	306	224
vieillard	N	-	13	68
patriarche	N	-	-	1
cambrioleur	N	2	-	33
Total		31%	21%	48%

Le CECR et les référentiels

- Le Cadre Européen commun de référence pour les langues (CECR) définit les niveaux de maîtrise d'une langue étrangère en fonction de savoir-faire, selon 6 niveaux (A1 à C2).
  $\rightarrow$ Il reste très vague en ce qui concerne l'acquisition du lexique et de la grammaire.
- Plus récemment, des référentiels ont été publiés pour chaque langue, précisant les apprentissages [Beacco and Porquier, 2007]
- Il reste cependant des limites :
  - Pas de distinction plus fine au sein d'un même niveau
  - Le format n'est pas adéquat pour le TAL
  - Diverses critiques sur le mode de conception des référentiels [Hulstijn, 2007]

Une approche alternative : le projet CEFRLex

- Objectif : offrir des ressources lexicales décrivant la distribution du lexique de diverses langues dans des manuels L2.
  $\rightarrow$ Cette distribution est faite sur les six niveaux du CECR.
- La distribution est estimée à partir d'un corpus de textes issus de manuels de langue et les fréquences sont adaptées (*cf.* ci-après).
- Usage possible :
  - Définition du parcours d'apprentissage (quels mots à quel niveau/sous-niveau)
  - Comparer la fréquence d'utilisation de synonymes (substitution lexicale en SAT)
  - Intégration comme modèle de langue dans diverses tâches d'ALAO (ex. lisibilité)

Une approche alternative : le projet CEFRLex

FLELex (Français L2)

- Disponible à l'adresse http ://cental.uclouvain.be/flelex/
- Publication : [François et al., 2014]
- Equipe : Núria Gala, Patrick Watrin, Cédrick Fairon, Anaïs Tack, Thomas François

SVALex (Suédois L2)

- Disponible à l'adresse http ://cental.uclouvain.be/svalex/
- Publication : en cours
- Equipe : Elena Volodina, Ildikó Pilán, Anaïs Tack, Thomas François

En cours : Espagnol (avec Barbara Decock) et Anglais

Méthodologie commune

1. Collecter un corpus de textes de manuels L2 ou livres simplifiées dans la langue donnée
2. Lemmatiser et POS-tagger le corpus
3. Estimer la distribution de fréquence de chaque lemme, à l'aide d'un estimateur robuste
4. Processus itératif : nettoyage manuel pour éliminer les erreurs de TAL, avant ré-estimation des fréquences.
5. Analyse de la ressource et mise à disposition sur un site.

Illustration de cette méthodologie avec FLELex

FLELex : le corpus

Collecte de 28 manuels de FLE et de 29 livres simplifiés, pour un total de 2 071 textes et 777 000 mots

Genre	A1	A2	B1	B2
Dialogue	153 (23,276)	72 (17,990)	39 (11,140)	5 (1,698)
E-mail, mail	41 (4,547)	24 (2,868)	44 (11,193)	18 (4,193)
Phrases	56 (7,072)	21 (4,130)	12 (1,913)	5 (928)
Varias	31 (3,990)	36 (4,439)	23 (5,124)	14 (1,868)
Textes	171 (23,707)	325 (65,690)	563 (147,603)	156 (63,014)
Livres simplifiés	8 (41,018)	9 (71,563)	7 (73,011)	5 (59,051)
Total	460 (103,610)	487 (166,680)	688 (249,984)	203 (130,752)

Genre	C1	C2	Total
Dialogue	/	/	269 (54,104)
E-mail, mail	8 (2,144)	1 (398)	136 (25,343)
Phrases	/	/	94 (14,043)
Varias	1 (272)	/	105 (15,693)
Textes	175 (89,911)	48 (34,084)	1,438 (424,009)
Livres simplifiés	/	/	29 (244,643)
Total	184 (92,327)	49 (34,482)	2,071 (777,835)

L'analyse morpho-syntaxique

- **Objectif :** obtenir le lemme de chaque forme observée dans le corpus et désambiguïser les formes homographiques qui se différencient par leur catégorie de discours
  - Utiliser des formes fléchies impliquerait de disperser la masse fréquentielle parmi plusieurs formes fléchies.
  - On considérerait aussi que les apprenants sont incapables d'analyser les formes fléchies.

- **Problème :** La précision du taggeur utilisé est cruciale, sans quoi on obtiendra :
  - entrées avec une mauvaise catégorie (ex. *adoptez* PREP ou *tu* ADV) ;
  - entrées avec un lemme non attesté (ex. *faire partir* plutôt que *faire partie*) ;
  - catégorie possible pour un mot, mais incorrect dans un contexte donné.

Les taggeurs et les multi-mots

- Une limitation bien connue des taggeurs est leur inaptitude à extraire les expressions polylexicales (EPs) !
- Les expressions polylexicales incluent un ensemble hétérogène d'objets linguistiques (collocations, mots composés, idiomes, etc.)
- Des études [Bahns and Eldaw, 1993] montrent que la connaissance qu'ont les apprenants des EPs est à la traîne par rapport à leur connaissance lexicale en générale
  $\rightarrow$ En conséquence, inclure les EPs dans un lexique gradué pour le FLE apparaît essentiel !

Les taggeurs sélectionnés

Nous avons sélectionné 2 taggeurs :

TreeTagger

- Treetagger [Schmid, 1994] : utilisé largement et bien connu
- Facile à employer (il existe des wrappers pour différents langages de programmation)
- Performance plus faible que l'état de l'art et ne détecte pas les EPs

un taggeur CRF

- Les taggeurs CRF sont l'état de l'art et peuvet être entraînés pour détecter les EPs.
- Nous avons utilisé le taggeur CRF développé par EarlyTracks et inspiré de [Constant and Sigogne, 2011].

Phase d'évaluation des taggeurs

Performances des deux taggeurs ne sont pas connues sur le FLE :

Méthodologie de comparaison des deux taggeurs

- Données de test = échantillon de 100 phrases tirées du corpus et divisé en deux *batchs*
- Chaque *batch* a été évalué par deux experts, pour chaque taggeur.
- Le schéma d'annotation des erreurs :

0. pas d'erreur ;
1. lemme est correct, mais pas le POS ;
2. le POS est correct, mais pas le lemme ;
3. erreur au niveau du POS et du lemme ;
4. erreur de segmentation (uniquement pour le taggeur CRF)

Résultats

	TreeTagger	**Taggeur CRF**
correct	94.2%	95.8%
erreur de POS	2.6%	1%
erreur de lemma	1.3%	0.5%
POS + lemme	1.9%	1.1%
segmentation	/	1.6%

- Accord inter-anotateur généralement bon : *kappa* varie entre 0.66 et 0.90
- Le taggeur CRF se comporte mieux que le TreeTagger !
- Les deux commettent quelques erreurs → une validation manuelle sera nécessaire !

Le calcul des distributions de fréquence

Pourquoi ne pas utiliser les fréquences brutes ?

- Étape 1 = compter les fréquences des lemmes par niveau
- Cependant, on ne peut se satisfaire des fréquences brutes...
  - [Francis and Kucera, 1982] ont mis en évidence que les mots à faible fréquence tendent à être spécifiques à certains contextes.
  - On les retrouve dans peu de textes, mais parfois de façon (trop) fréquente dasn un texte donné
    $\rightarrow$ Ex. : livre simplifié sur les chevaliers de la table ronde (chevalier, épée, heaume, etc.)
    $\rightarrow$ *Heaume* apparaît dès A1 dans FLELex !

Le calcul des distributions de fréquence

- Pour réduire cet effect, utilisation de l'indice de dispersion [Carroll et al., 1971]

$$D_{w,K} = [\log(\sum p_i) - \frac{\sum p_i \log(p_i)}{\sum p_i}] / \log(I) \quad (1)$$

K = niveau CECR ; I = nombre de manuels dans le niveau K ;
p_i = probability du mot dans le manuel i.

- Ensuite, les fréquences brutes (RFL) sont normalisés comme suit :

$$U = (\frac{1\ 000\ 000}{N_k})[RFL * D + (1 - D) * f_{min}] \quad (2)$$

où N_k = nombre de tokens au niveau k ;
$f_{min} = \frac{1}{N} \sum f_i s_i$ avec f_i = probability du mot dans le manuel i et s_i = nombre de mots dans le manuel i

Les deux versions de FLELex

FLELex-TT

- Inclut 14 236 entrées, mais pas d'expression polylexicales !
- Il est basé sur le Treetagger et est donc simple à utiliser dans des applications de TAL
- La ressource a été vérifiée manuellement (sans processus itératif, jusqu'à présent).

FLELex-CRF

- Inclut 17 871 entries, parmi lesquelles plusieurs milliers d'EPs
- Les meilleurs performances de ce taggeur signifie une meilleure estimation des distributions de fréquence
- Par contre, les erreurs de segmentations engendrent l'apparition de séquences erronnées (ex. *académisme et avant-garde*)
- Pas encore vérifié manuellement (mais cela serait bien nécessaire)

Exemple d'entrées

lemma	tag	A1	A2	B1	B2	C1	C2	total
voiture (1)	NOM	633.3	598.5	482.7	202.7	271.9	25.9	461.5
abandonner (2)	VER	35.5	62.3	104.8	79.8	73.6	28.5	78.2
justice (3)	NOM	3.9	17.3	79.1	13.2	106.3	72.9	48.1
kilo (4)	NOM	40.3	29.9	10.2	0	1.6	0	19.8
logique (5)	NOM	0	0	6.8	18.6	36.3	9.6	9.9
en bas (6)	ADV	34.9	28.5	13	32.8	1.6	0	24
en clair (7)	ADV	0	0	0	0	8.2	19.5	1.2
sous réserve de (8)	PREP	0	0	0.361	0	0	0	0.03

Quelques chiffres à propos de FLELex

- Une majorité des entrées dans les deux listes sont des noms (respectivement 51% et 55%).
- La version TreeTagger contient 33% d'hapaxes, alors que seules 26% des entrées ont 10 occurrences ou plus.
- La version CRF contient 20% d'hapaxes pour 31% d'entrées avec 10 occurrences ou plus.
- Comparaison de FLELex-TT avec un autre lexique : Lexique 3 [New et al., 2004]
  $\rightarrow$ Seules 622 entrées de FLELex-TT manquent dans Lexique 3
- Corrélation entre les fréquences globales de FLELex-TT et de Lexique3 est élevée : $0,84$

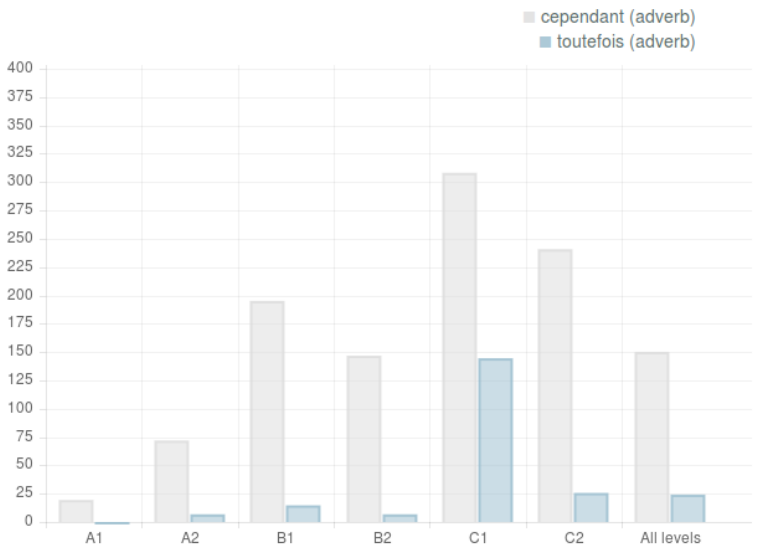
Démonstration

Make a query in FLELex

Enter a word

cependant | toutefois | – | Search

Frequencies by CEFR levels for the words cependant* and toutefois**.



* Frequencies of cependant in FLELex-TT.

** Frequencies of toutefois in FLELex-TT.

Perspectives

- Nettoyage manuel de la version CFR.
- Utiliser FLELex pour prédire les mots connus/inconnus d'un lecteur donné.
- Développer le projet "CEFRLex" pour d'autres langues (espagnol et anglais en cours)
  $\rightarrow$ Pourquoi pas le néérlandais ou l'allemand.
- Développer un filtre entre le tagset du TreeTagger et celui du DELAF (permettrait d'utiliser TT avec la version CRF)
- Estimer la perte de la masse fréquentielle due aux erreurs de TAL.

Plan

La prédiction de la connaissance lexicale

La prédiction lexicale en général

- [Crossley et al., 2010] évalue la compétence lexicale globale à partir de diverses variables (diversité, fréquence, polysémie, etc.)
- [Gala et al., 2013] ont développé un modèle de difficulté lexical basé sur des caractéristiques intrinsèques des mots + carac. psycholinguistiques.
  $\rightarrow$ Ils obtiennent 62% de prédictions exactes sur trois classes (Manulex), pour la L1.
- [Shardlow, 2013] utilise aussi un SVM avec des variables variées
  $\rightarrow$ [Shardlow, 2014] montre que ce type d'approche peine à dépasser une baseline basée sur la fréquence
- [Gala et al., 2014] confirme ce résultat, obtenant un gain de 2 à 4% par rapport à une baseline fréquentielle.

Prédire la difficulté des mots semble une tâche fort complexe !

La prédiction de la connaissance lexicale

Alternative : prédire les mots inconnus d'un lecteur dans un texte.

- **Méthode** : pour un apprenant d'un niveau donné, identifier les mots qui sont d'un niveau supérieur.
- FLELex pourrait être utilisé pour prédire les mots inconnus d'un apprenant.
- [Gala et al., 2014] testent aussi leur méthode de prédiction sur FLELex
  - Les distributions de FLELex sont transformées en un niveau unique, qui sert de données d'entraînement pour un algorithme statistique
  - Classe majoritaire = 28,8% ; Baseline fréquentielle = 39% ; modèle = 43% d'exactitude.
- Approche de [Tack, 2015] : utiliser directement les niveaux de FLELex.

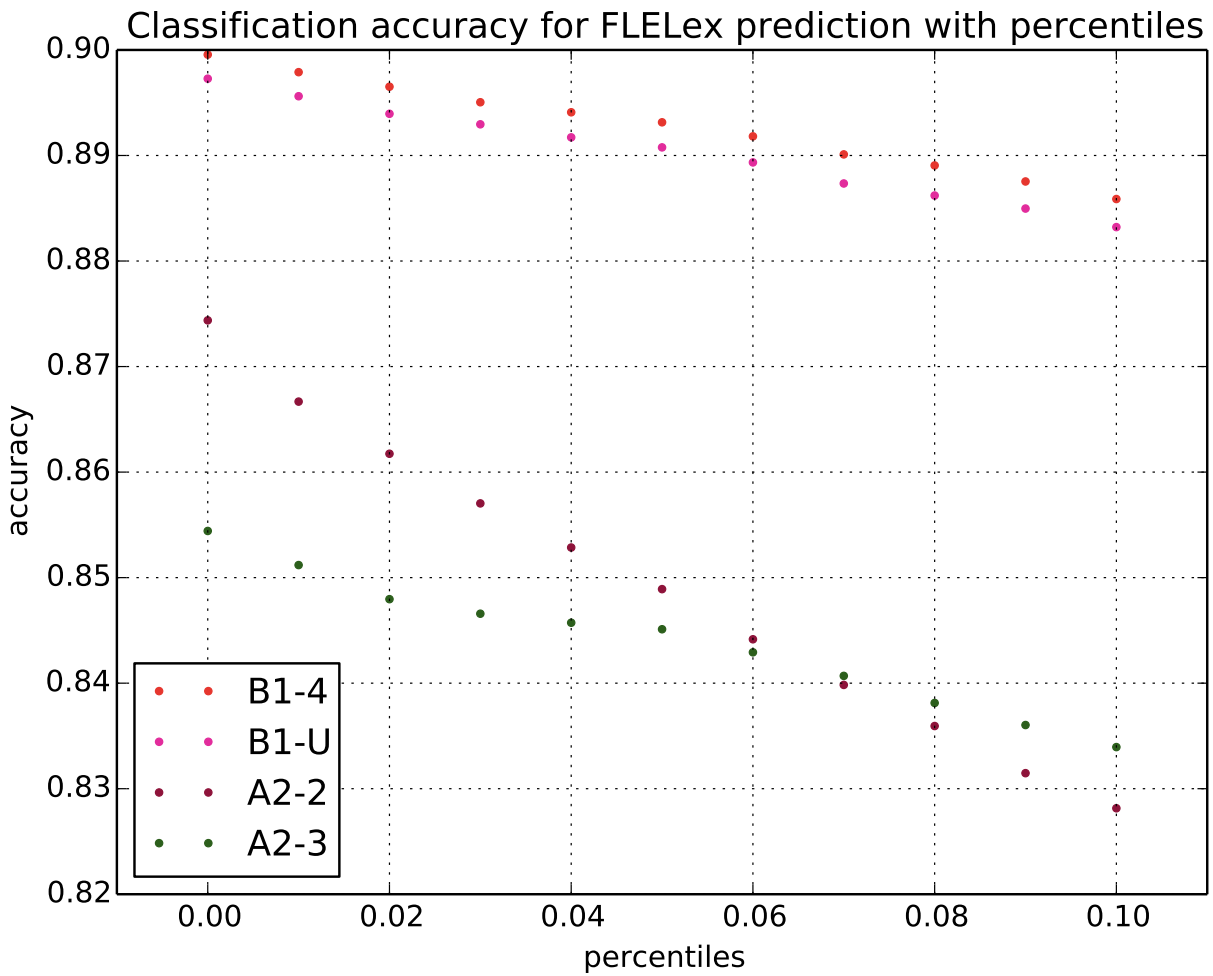
FLELex pour prédire les mots inconnus

Problème : Comment transformer au mieux les distributions en un niveau unique ?

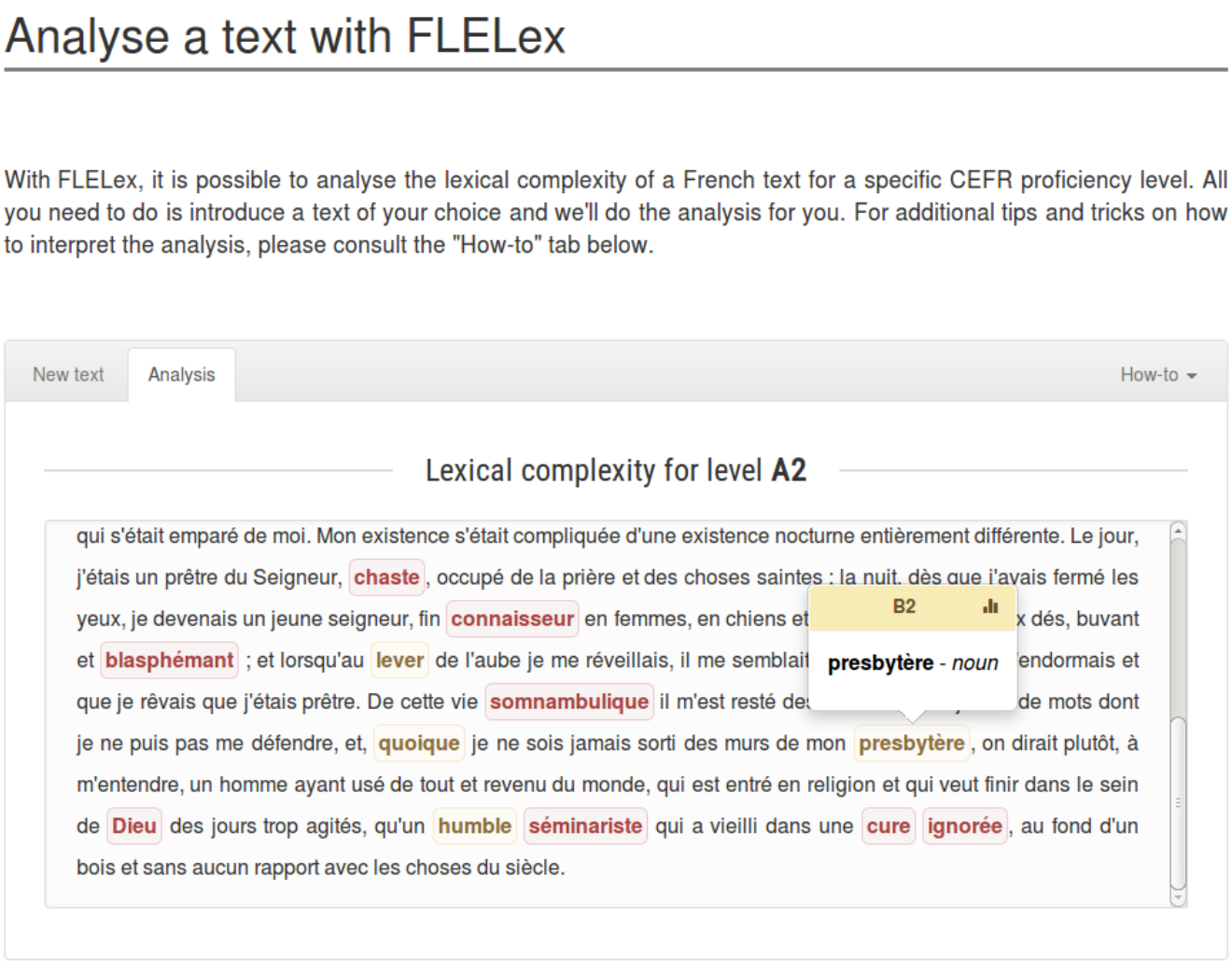
Expérience de [Tack, 2015]

- Collecte les annotations de 4 apprenants (A2 et B1) sur 51 courts textes $\rightarrow$ apprenants identifient les mots inconnus via une interface web.
- Ensuite, expérimentations de différents critères (seuil de fréquence, quantile) dans le but de prédire au mieux les mots inconnus des 4 apprenants.
- Etonnament, la meilleure fonction de discrétisation est la première occurrence !

FLELex pour prédire les mots inconnus



Démonstration



Evaluation de FLELex comme prédicteur

	Mots lexicaux	mots grammaticaux	Total
apprenant A2-2	86.6%	99.2%	89.7%
apprenant A2-3	81.1%	99.2%	87.4%
apprenant B1-4	91.3%	99.7%	92.3%
apprenant B1-U	90.8%	99.8%	92.0%

TABLE : Exactitude des prédictions de la connaissance lexicale des 4 apprenants via FLELex.

Discussion

- D'après les résultats de l'interface, les prédictions sont trop optimistes (trop de mots A1)
- D'après l'évaluation, les prédictions globales sont bonnes, mais...
  $\longrightarrow$ Le modèle se comporte mieux sur les mots connus que les mots inconnus (bcp moins fréquents).
- Conséquence de l'heuristique ci-dessus, qui est trop optimiste !

	Connu		Inconnu	
apprenant A2-2	95.7%	(0.92)	4.3%	(0.42)
apprenant A2-3	88.1%	(0.94)	11.9%	(0.38)
apprenant B1-4	97.0%	(0.94)	3.0%	(0.40)
apprenant B1-U	96.7%	(0.94)	3.3%	(0.37)

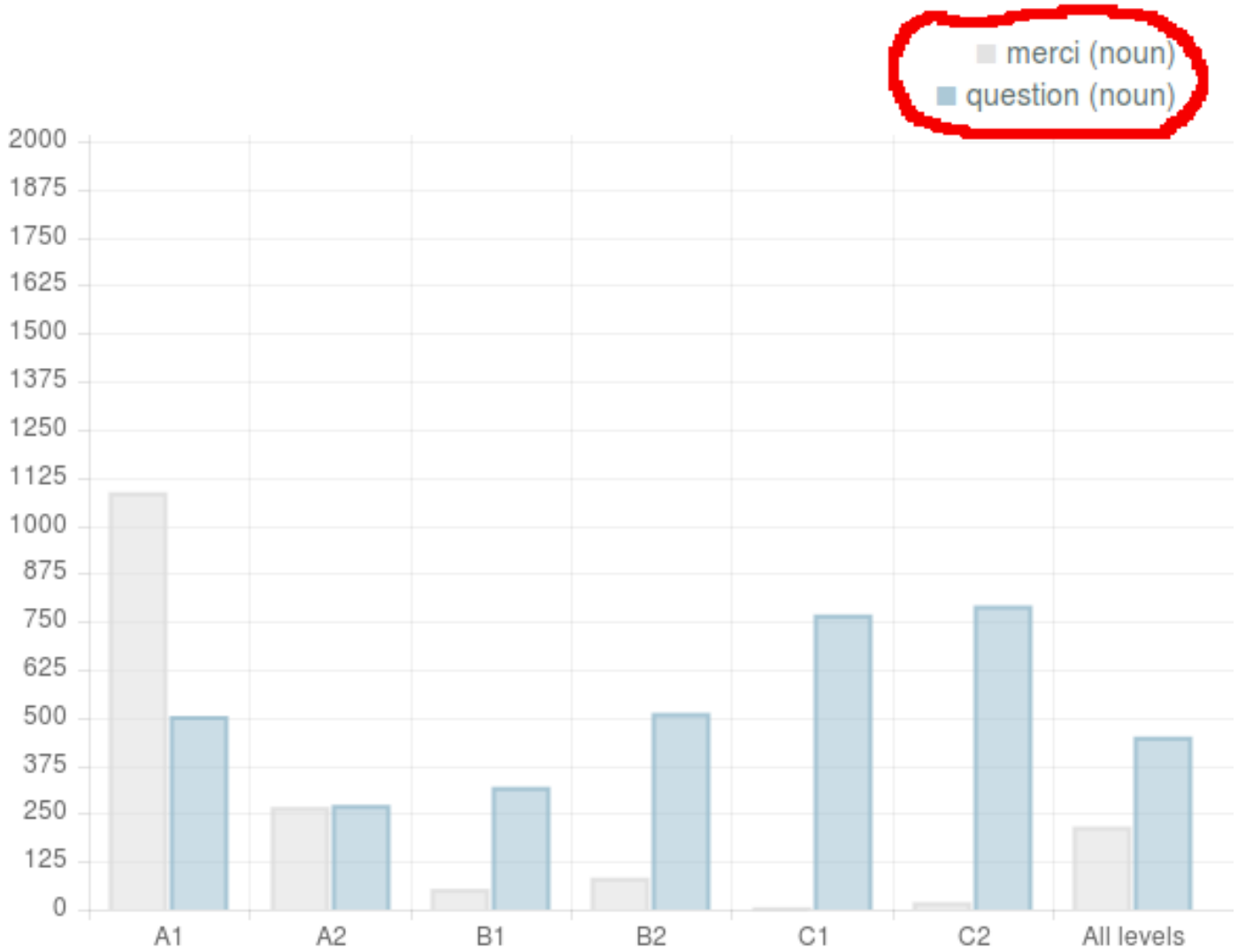
TABLE : Pourcentage de mots connus et inconnus des apprenants + rappel des prédictions de FLELex.

Conclusions et perspectives

- Le projet CEFRLex (et FLELex) propose une cartographie de l'usage des lemmes de L2 à destination des professeurs, des appprenants et des chercheurs.
- Le projet vise à couvrir plusieurs langues européennes (anglais, espagnol, allemand, italien, néérlandais) à terme.
- Les ressources sont disponibles via un site web auquel d'autres fonctions viendront s'ajouter (ex. évaluation d'un texte en fonction des référentiels).

- Trouver d'autres fonctions de discrétisation pour transformer les distributions en un niveau (ex. distribution dans les manuels).
  $\rightarrow$ Objectif = distinguer le vocabulaire de base du vocabulaire périphérique.
- Affiner la prédiction personnalisée en combinant le travail de [Tack, 2015] et [Gala et al., 2014] (mémoire ou stage ?)

Merci pour votre attention



References I


Alderson, J. (1984).
Reading in a foreign language : a reading problem or a language problem ?
In Alderson, J. and Urquhart, A., editors, *Reading in a Foreign Language*, pages 1–24. Longman, New York.


Anderson, R. and Nagy, W. (1991).
Word meanings.
In Barr, R., Kamil, M. L., Mosenthal, P., and Pearson, P., editors, *Handbook of Reading Research*, pages 512–538. Longman, New York.


Bahns, J. and Eldaw, M. (1993).
Should We Teach EFL Students Collocations ?
System, 21(1) :101–14.


Beacco, J.-C. and Porquier, R. (2007).
Niveau A1 pour le français : utilisateur-apprenant élémentaire.
Didier.

References II


Carrell, P. (1998).
Introduction : Interactive approaches to second language reading.
In Carrell, P., Devine, J., and Eskey, D., editors, *Interactive approaches to second language reading*, pages 1–7. Cambridge Univ. Press, Cambridge.


Carroll, J., Davies, P., and Richman, B. (1971).
The American Heritage word frequency book.
Houghton Mifflin Boston.


Chall, J. and Dale, E. (1995).
Readability Revisited : The New Dale-Chall Readability Formula.
Brookline Books, Cambridge.


Coady, J. (1997a).
L2 vocabulary acquisition : A synthesis of the research.
In Coady, J. and Huckin, T., editors, *Second language vocabulary acquisition*, pages 273–290. Cambridge University Press, Cambridge.

References III


Coady, J. (1997b).
L2 vocabulary acquisition through extensive reading.
In Coady, J. and Huckin, T., editors, *Second language vocabulary acquisition*, pages 225–237. Cambridge University Press, Cambridge.


Constant, M. and Sigogne, A. (2011).
Mwu-aware part-of-speech tagging with a crf model and lexical resources.
In *Proceedings of the Workshop on Multiword Expressions : from Parsing and Generation to the Real World*, pages 49–56.


Crossley, S., Salsbury, T., McNamara, D., and Jarvis, S. (2010).
Predicting lexical proficiency in language learner texts using computational indices.
Language Testing, pages 1–20.


François, T., Gala, N., Watrin, P., and Fairon, C. (2014).
FLELex : a graded lexical resource for French foreign learners.
In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*.

References IV


Francis, W. and Kucera, H. (1982).
Frequency analysis of english usage.


Gala, N., François, T., Bernhard, D., and Fairon, C. (2014).
Un modèle pour prédire la complexité lexicale et graduer les mots.
In *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2014)*, pages 91–102.


Gala, N., François, T., and Fairon, C. (2013).
Towards a french lexicon with difficulty measures : Nlp helping to bridge the gap between traditional dictionaries and specialized lexicons.
In *Electronic lexicography in the 21st century : thinking outside the paper (eLex2013)*, pages 132–151.


Gougenheim, G., Michéa, R., Rivenc, P., and Sauvageot, A. (1964).
L'élaboration du français fondamental (1er degré).
Didier, Paris.

References V


Haynes, M. and Baker, I. (1993).
American and chinese readers learning from lexical familiarization in english texts.
Second language reading and vocabulary learning, pages 130–152.


Hirsh, D. and Nation, P. (1992).
What vocabulary size is needed to read unsimplified texts for pleasure ?
Reading in a foreign language, 8(2) :689–689.


Hu, M. and Nation, P. (2000).
Unknown vocabulary density and reading comprehension.
Reading in a foreign language, 13(1) :403–30.


Hulstijn, J. (2007).
The shaky ground beneath the cefr : Quantitative and qualitative dimensions of language proficiency.
The Modern Language Journal, 91(4) :663–667.

References VI

Koda, K. (1989).
The effects of transferred vocabulary knowledge on the development of l2 reading proficiency.
Foreign language annals, 22(6) :529–540.

Koda, K. (2005).
Insights into second language reading : A cross-linguistic approach.
Cambridge University Press, Cambridge.

Krashen, S. (1989).
We acquire vocabulary and spelling by reading : Additional evidence for the input hypothesis.
The Modern Language Journal, 73(4) :440–464.

Laufer, B. and Ravenhorst-Kalovski, G. (2010).
Lexical threshold revisited : Lexical text coverage, learners' vocabulary size and reading comprehension.
Reading in a foreign language, 22(1) :15–30.

References VII

Lété, B., Sprenger-Charolles, L., and Colé, P. (2004).
Manulex : A grade-level lexical database from French elementary-school readers.
Behavior Research Methods, Instruments and Computers, 36 :156–166.

Lonsdale, D. and Le Bras, Y. (2009).
A frequency dictionary of French : core vocabulary for learners.
Routledge, London, UK.

Michéa, R. (1953).
Mots fréquents et mots disponibles. un aspect nouveau de la statistique du langage.
Les langues modernes, 47(4) :338–344.

Nagy, W. and Herman, P. (1987).
Breadth and depth of vocabulary knowledge : Implications for acquisition and instruction.
The nature of vocabulary acquisition, 19 :35.

References VIII


Nagy, W., Herman, P., and Anderson, R. (1985).
Learning words from context.
Reading research quarterly, 20(2) :233–253.


Nation, I. (2001).
Learning vocabulary in another language.
Cambridge University Press.


Nation, I. (2006).
How large a vocabulary is needed for reading and listening ?
Canadian Modern Language Review, 63(1) :59–82.


New, B., Pallier, C., Brysbaert, M., and Ferrand, L. (2004).
Lexique 2 : A new French lexical database.
Behavior Research Methods, Instruments, & Computers, 36(3) :516.


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.

References IX


Schmitt, N. (1998).
Tracking the incremental acquisition of second language vocabulary : A longitudinal study.
Language learning, 48(2) :281–317.


Shardlow, M. (2013).
A comparison of techniques to automatically identify complex words.
In *ACL Student Research Workshop*, pages 103–109.


Shardlow, M. (2014).
Out in the open : Finding and categorising errors in the lexical simplification pipeline.
LREC 2014, pages 1583–1590.


Tack, A. (2015).
Modèles adaptatifs pour évaluer automatiquement la connaissance lexicale d'un apprenant de FLE.
Master's thesis, Université catholique de Louvain.
Thesis Supervisors : C. Fairon and T. François.

References X


Thorndike, E. (1921).
Word knowledge in the elementary school.
The Teachers College Record, 22(4) :334–370.


Ulijn, J. and Strother, J. (1990).
The effect of syntactic simplification on reading est texts as l1 and l2.
Journal of research in reading, 13(1) :38–54.