Méthodes TAL pour la prédiction automatisée de la difficulté lexicale

Thomas François

Séminaire à l'Université Aix-Marseille

June 27th, 2016

Plan

Plan

Problématique

Situation-problème : un professeur sélectionne un texte pour un exercice de lecture.

Une phrase exemple

La présidente nouvellement élue demande l'abolition de la taxe sur le capital.

Intuition du professeur concernant les mots complexes (pour A2)

La présidente nouvellement élue demande l'abolition de la taxe sur le capital.

Problématique

Mots réellement difficiles pour l'apprenant A

La présidente nouvellement élue demande l'abolition de la taxe sur le capital.

- L'apprenant A a déduit que *nouvellement* est un adverbe basé sur la forme *nouvelle*.
- Il a aussi associé *élue* avec la forme infinitive *élire*.
- Cependant, il n'a jamais rencontré le mot *taxe*.

Problématique

Variations en fonction de la L1 :

Mots en réalité difficiles pour un apprenant anglophone

La présidente nouvellement élue demande l'abolition de la taxe sur le capital.

- *taxe* et *capital* sont des congénères en anglais.

Mots réellement difficiles pour un apprenant japonais

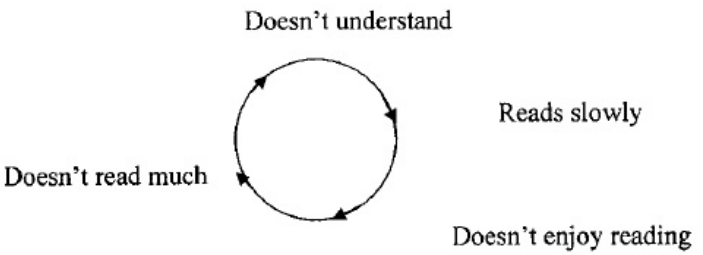
La présidente nouvellement élue demande l'abolition de la taxe sur le capital.

- Ce n'est par contre pas le cas en japonais (il y a *tax-free*)

Lien entre acquisition et difficulté lexicale

- Il semblerait que la lecture joue un rôle important dans l'acquisition 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, 1997],

  qui peut entraîner un cercle vicieux.



On peut briser ce cercle vicieux en contrôlant la difficuté

Recherches présentées dans cette présentation

Nous présentons trois approches de la prédiction de la difficulté lexicale :

FLELex : un lexique gradué en fonction du CECR pour le FLE

- Contient les distributions de +/- 15 000 mots

ReSyf : une liste graduée de synonymes pour la L1 et la L2

- ReSyf est destiné à des enfants, mais il existe une version pour le FLE.
- Les synonymes y sont ordonnés automatiquement à l'aide d'un modèle statistique.

Expériences sur la prédiction au niveau de l'individu

- Expériences préliminaires d'une thèse de master (Anais Tack).

Plan

Le projet CEFRLex

- Objectif : offrir des ressources lexicales décrivant la distribution du lexique de diverses langues dans des manuels L2.
  $\rightarrow$ Cette distribution est établie en fonction des 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).

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)

Méthodologie commune

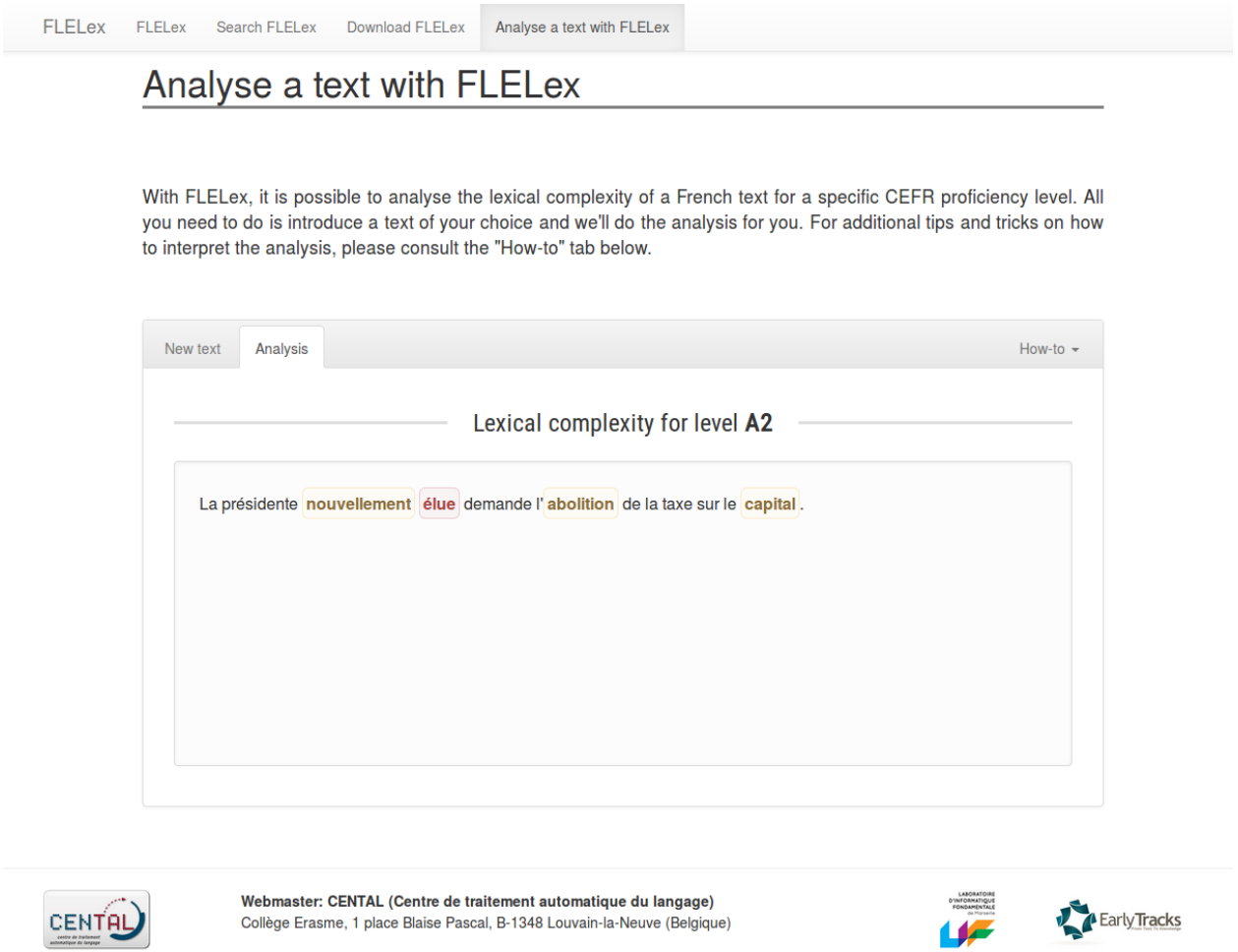
1. Collecter un corpus de textes de manuels L2 ou livres simplifiées dans la langue donnée
2. Lemmatiser et POS-tagger le corpus
3. Estimer la distribution de fréquence de chaque lemme, à l'aide d'un estimateur robuste
4. Processus itératif : nettoyage manuel pour éliminer les erreurs de TAL, avant ré-estimation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

Illustration de cette méthodologie avec FLELex

FLELex : exemples 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

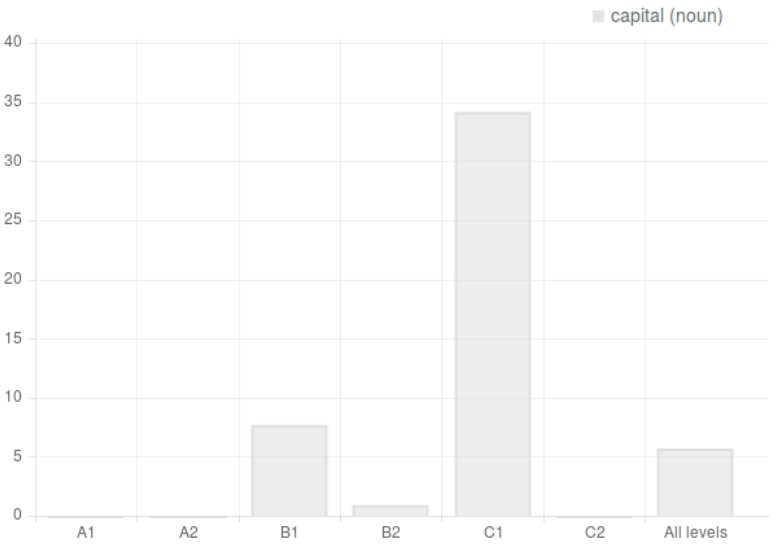
La prédiction des mots inconnus avec FLELex



FLELex : prédire les mots inconnus à un niveau CECR

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

Par exemple : la distribution de *capital*



* Frequencies of capital in FLELex-TT.

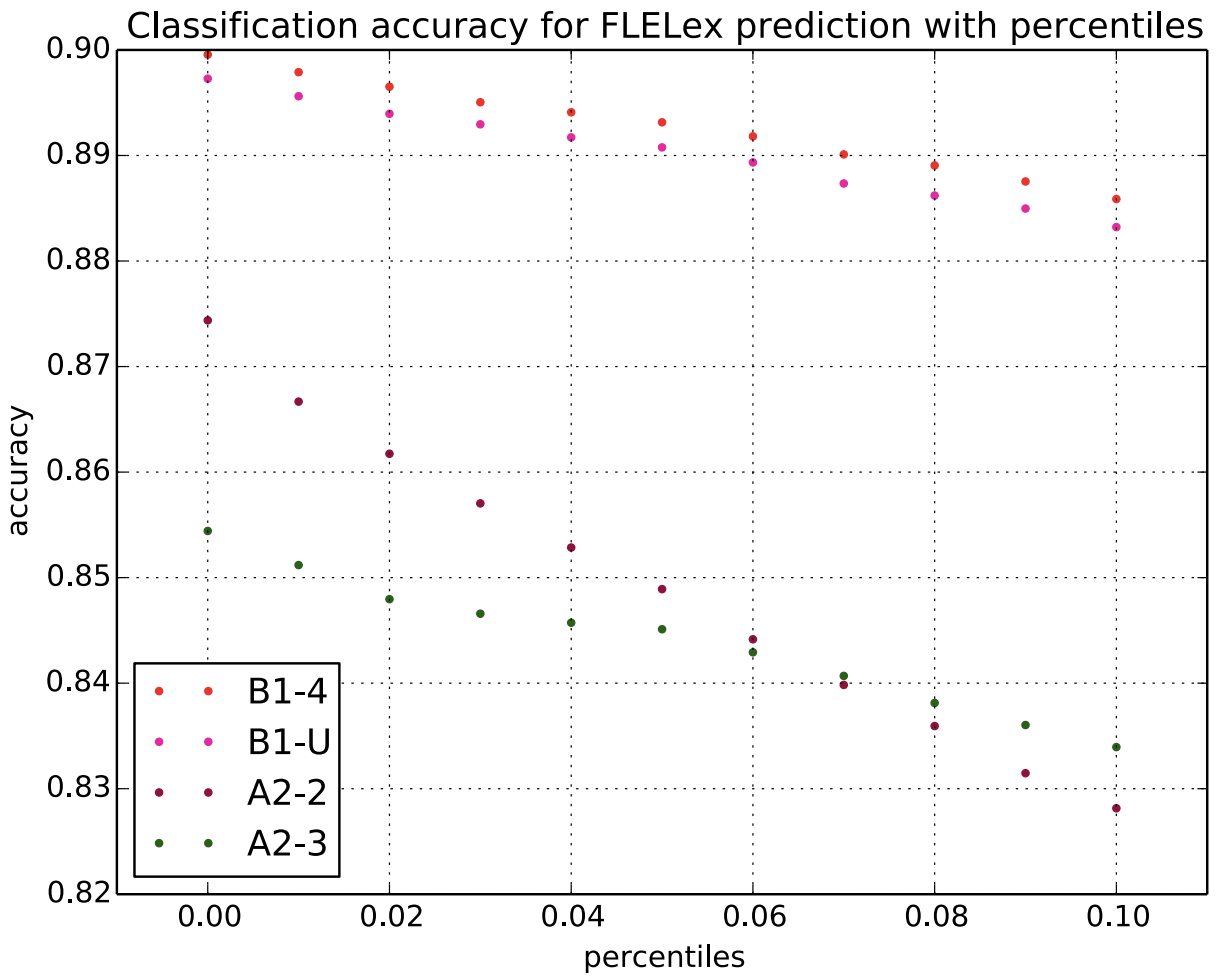
... est transformée en B1

FLELex : prédire les mots inconnus à un niveau CECR

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 : prédire les mots inconnus à un niveau CECR



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 "première occurrence", 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.

Perspectives

- Collecter des données plus représentatives d'apprenants (plus test de positionnement) pour reproduire l'expérience
- Extraire le vocabulaire de base en se basant sur la répartition des mots dans les manuels d'un niveau.
- Etendre le projet à d'autres langues (espagnol et anglais en cours)
- Développer un outil similaire, mais directement basé sur les référentiels du CECR [Beacco and Porquier, 2007]

Plan

Objectifs du projet ReSyf

- **Contexte :** Investiguer la difficulté lexicale d'un point de vue TAL
  $\rightarrow$ Est-il possible de prédire la complexité des mots de façon intrinsèque (sur la base de leurs caractéristiques) ?
- **Objectifs :**
  - Identifier les variables (predictors) qui caractérisent les mots "simples"
  - Développer un modèle de la difficulté lexicale afin de proposer une ressource graduée de synonymes (ReSyf)
- **Concrètement :** Il s'agit de généraliser les échelles de difficulté de ressources comme Manulex ou FLELex à un vocabulaire plus large Ce modèle est croisé avec des ressources de synonymes.

Team : Núria Gala, Delphine Bernhard, Mokhtar Billami, Cédrick Fairon

Défi 1 : grader les synonymes par difficulté

Remplacer les termes complexes par des synonymes plus simples

La présidente nouvellement élue

1. justement
2. récemment
3. dernièrement
4. fraîchement
5. nouvellement

demande l'abolition de la taxe sur le capital.

1. annulation
2. suppression
3. abolition
4. abrogation
5. abolissement

1. capital
2. capitalisation

Défi 2 : opérer le remplacement

Le simple remplacement est souvent inapproprié !

La présidente nouvellement élue

élue depuis peu

inversion syntaxique

demande l'abolition de la taxe sur le capital.

la suppression

modification du contexte

demande l'abolition de la taxe sur le capital.

la destruction

"synonymes" inadaptés

Focus sur l'algorithme de ranking

- **Technique** : algorithme d'ordonnancement de type *pairwise*, à savoir SVMRank [Herbrich et al., 2000]
- **Données d'entraînement** : ressource similaire à FLELex, mais pour des enfants L1 : Manulex [Lété et al., 2004]
  $\longrightarrow$ Problème : définit une distribution de fréquence pour chaque mot, pas un niveau unique !
- 2 fonctions $\phi(D)$ testées :
  - $\phi(D) = L$, où L est le premier niveau où la fréquence du mot n'est pas nulle.
  - $\phi(D)$ qui renvoie une valeur continue ($1 \leq \phi(D) \leq 3$ en fonction de la distribution :

$$\phi(D) = L + e^{-r} \quad \text{où} \quad r = \frac{\sum_{l=1}^{L} U_l}{\sum_{l=L+1}^{3} U_l}$$

Les variables pour la difficulté lexicale

Les critères basés sur la forme orthographique :

- Nombre de lettres, phonèmes, syllabes
- Densité et fréquence du voisinage orthographique du mot cible
- Structure syllabique (structures les plus fréquentes V, CVC, CV, CYV)
- Consistance graphème-phonème :
  - 0 = transparence : 'abruti' `[abRyti]`
  - $<$ 2 caractères : 'abrite*r*' `[abRite]`
  - $>$ 2 caractères : 'l*ent*em*ent*' `[l@tm@]`
- Patrons orthographiques : doubles voyelles (ex. ée `[e]`), doubles consonnes (ex. pp `[p]`), digraphes (ex. ch `[S]`)

Variables pour la difficulté lexicale (2)

Les critères morphologiques :

- Nb. morphèmes, préfixation (oui/non), suffixation (oui/non), est composé (oui/non), fréq. minimale préf/suf, fréq. moyenne préf/suf, taille famille morphologique
- Nouvelles variables : fréq. mot le plus fréquent dans la famille, fréq. moyenne (ou cumulée) des mots de la famille.

L'analyse morphologique, automatique non supervisée

- Découpage en segments morphémiques étiquetés (base, préfixe, suffixe, élé. liaison)
- Ensuite, identification de familles morphologiques [Bernhard, 2010]
- Exemples :

  rouille – antirouille ; rouilleux
  dérouiller – dérouillage ; dérouillement ;
  débrouille – brouilleur ; brouilleuse ; débrouilleur ; débrouilleuse

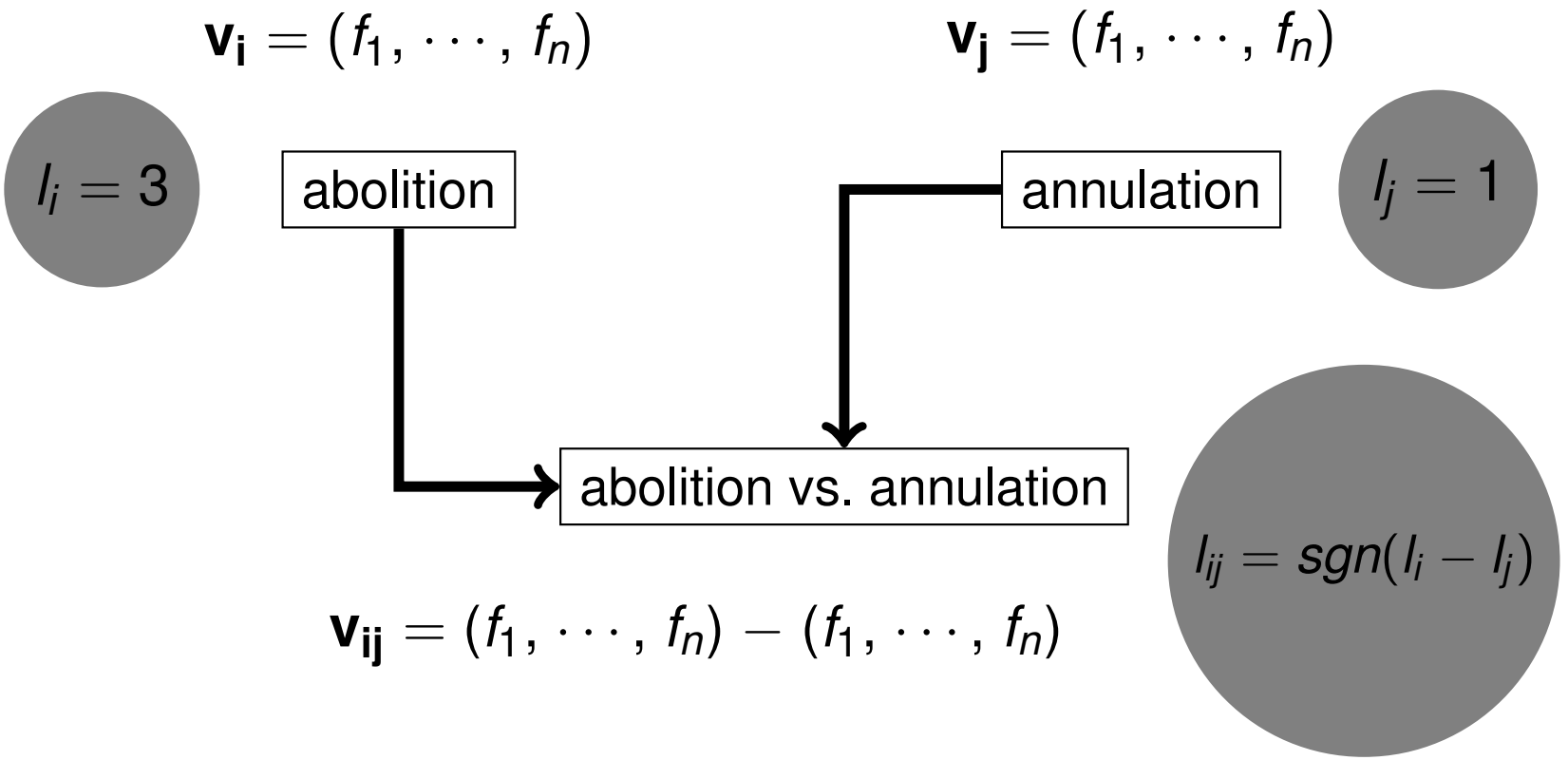
  brouille – brouillerie ; brouilleux

Variables pour la difficulté lexicale (3)

Autres critères :

- Polysémie :
  - Variable binaire indiquant si le mot est polysémique dans *JeuxDeMots* [Lafourcade, 2007]
  - Nombre de synsets répertoriés dans *BabelNet* [Navigli and Ponzetto, 2010]
- Informations de fréquence :
  - Logarithme de la fréquence reprise dans Lexique3 [New et al., 2007]
  - Présence ou absence de la liste de Gougenheim [Gougenheim et al., 1964]

Algorithme de création des paires



Si abolition est plus difficile que annulation, l_{ij} vaut 1.

Résultats (1/2)

L'efficacité de chaque variable est évaluée à l'aide d'une corrélation de Spearman (sur Manulex et sur les paires) :

Variables	**Manulex (ρ)**	**Paires (ρ)**
17 Freq. Lex3	-0,51	-0.57
18 AbsGoug (6000)	-0,41	-0.46
02 Nb. phon	0,30	0,35
15 Polysémie	-0,29	-0.33
01 Nb. lettres	0,27	0,32
03 Nb. syllables	0,27	0,32
4a Nb. voisins	-0,25	-0,23
15 Fréq. moyenne de la famille morpho.	-0,24	-0,27
15 Fréq. cumulée de la famille morpho.	-0,24	-0,27
15 Fréq. maximum de la famille morpho.	-0,24	-0,27
4b Voisin freqcum	-0,25	-0,23
6 Nombre de sens dans BabelNet	-0,20	-0,19

Résultats (2/2)

- La meilleure variable de chaque famille est retenue, si significative (total = 21).
- Evaluation via exactitude, estimée en validation croisée à 10 plis

Modèle SVM	**C**	**21 var.**	**C**	**69 var.**
Manulex-3N	0,01	77,4%	0,01	77,8%
Manulex-Cont	0,01	72,4%	0,01	71,4%

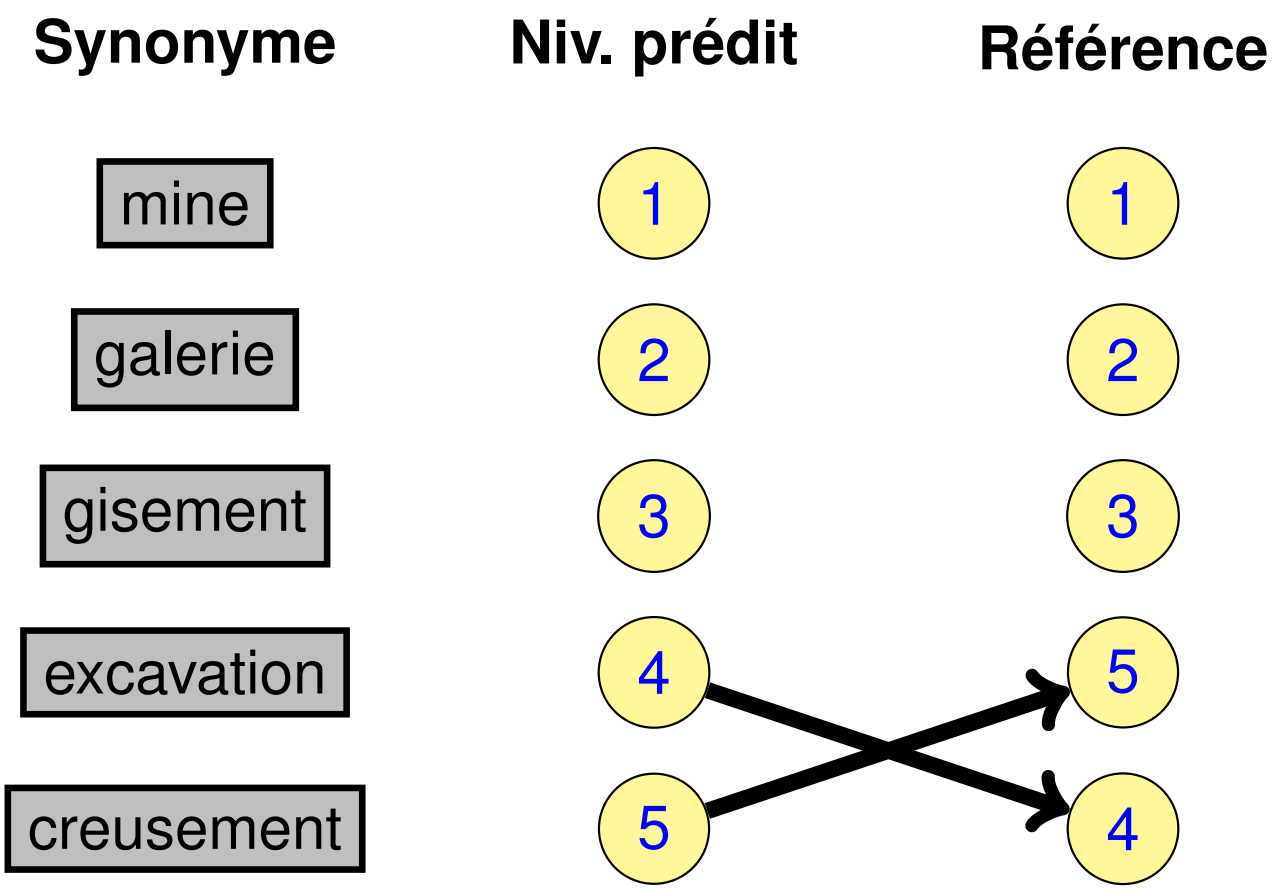
Evaluation avec des juges humains

- Évaluation externe : comparaison entre les prédictions du modèle et l'avis de juges humains.
- Données : 40 vecteurs de synonymes (de 3,5 mots en moyenne) soumis à l'avis de 40 juges.
- Accord moyen entre les juges : α de Krippendorff = $0,4$.
- Résultats de notre modèle : *k* de Cohen quadratique de $0,63$ (accord fort).
- MRR (rang réciproque moyen) = 0,84.

Evaluation extrinsèque : exemple

Synonyme	Niv. prédit	Référence
associer	1	1
combiner	2	2
assimiler	3	3
entremêler	4	4
amalgamer	5	5

Evaluation extrinsèque : exemple



Plan

Vers la prédiction de la difficulté lexicale individualisée

- Mémoire de Mlle Anaïs Tack [Tack et al., 2016]
- Jeu de données : annotations de 4 apprenants (A2 et B1) sur 51 courts textes (via une interface web).
- Variables simples (w_{-1}, w, w_{+1}) :
  $\longrightarrow$ POS, informations de FLELex, nb. lettres, nb. voisins orthographiques, patrons orthographiques, nb. synsets dans BabelNet
- Modèle : réseau de neurones

Evaluation des modèles personnalisés

	observations	A2-2	A2-3	B1-4	B1-U
M_E	toutes	89,9%	87,2%	92,2%	92%
	lexicales	86,4%	80,7%	91,2%	90,5%
M_P	toutes	95,8%	87,8%	97%	96,7%
	lexicales	92%	80,1%	94%	93,6%

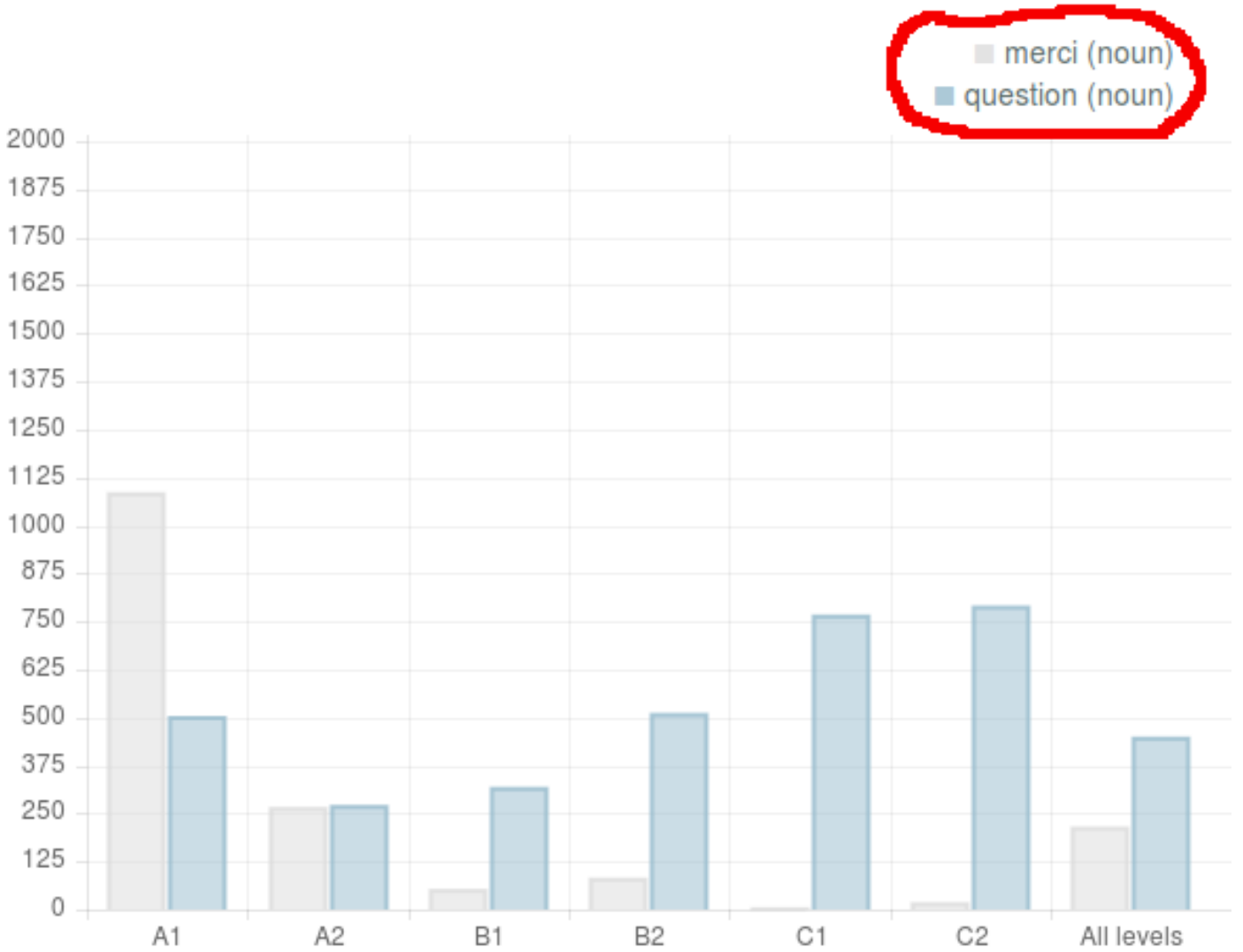
TABLE : Comparaison entre le modèle FLELex (M_E) et le modèle individualisé (M_P)

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 ReSyF propose une ressource de synonymes gradués, utile pour la substitution lexicale automatique.
- Les ressources sont disponibles sur le web.

- Trouver d'autres fonctions de discrétisation pour transformer les distributions en un niveau (ex. distribution dans les manuels).
- Affiner la prédiction personnalisée en combinant le travail de [Tack, 2015] et [Gala et al., 2014] (mémoire ou stage ?)
- Évaluer (et substituer) en contexte !

Merci pour votre attention



References I

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

Bernhard, D. (2010).
Apprentissage non supervisé de familles morphologisques : Comparaison de méthodes et aspects multilingues.
Traitement Automatique des Langues, 51(2) :11–39.

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

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 II


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.


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


Herbrich, R., Graepel, T., and Obermayer, K. (2000).
Large margin rank boundaries for ordinal regression.
chapter 7, pages 115–132. MIT Press, Cambridge.


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


Lafourcade, M. (2007).
Making people play for lexical acquisition with the jeuxdemots prototype.
In *SNLP'07 : 7th international symposium on natural language processing*.

References III


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.


Navigli, R. and Ponzetto, S. P. (2010).
Babelnet : Building a very large multilingual semantic network.
In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225.


New, B., Brysbaert, M., Veronis, J., and Pallier, C. (2007).
The use of film subtitles to estimate word frequencies.
Applied Psycholinguistics, 28(04) :661–677.


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 IV

Tack, A., François, T., Ligozat, A.-L., and Fairon, C. (2016).
Modèles adaptatifs pour prédire automatiquement la compétence lexicale d'un apprenant de français langue étrangère.
In *Actes de la 23e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2016)*, pages 1–14.