# La prédiction automatisée de la difficulté lexicale en FLE


Thomas François


Séminaire à l'Université de Fukuoka

January 13th, 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 *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*)

# Objectifs

1. Mieux appréhender les différentes caractéristiques des mots qui les rendent difficiles (“difficulté intrinsèque”)
2. Relier ces caractéristiques avec celles des apprenants (“difficulté extrinsèque”)
   $\longrightarrow$ ex. la L1, la culture, le niveau d’éducation, la motivation, la connaissance du sujet traité, etc.
3. Proposer des modèles linguistico-statistiques capables de prédire la difficulté des mots...
   - Pour une large population
     $\longrightarrow$ apprenants de FLE, enfants, dyslexiques, etc.
   - Pour une classe
     $\longrightarrow$ vue comme un groupe cohérent avec une histoire d’enseignement commune
   - Au niveau individuel
     $\longrightarrow$ L1, niveau d’éducation, culture, intertextes (reading history), connaissance du sujet, etc.

# Pourquoi détecter les mots difficiles ?

- Sélectionner des matériaux de lecture adaptés à une classe ou à un individu
  $\longrightarrow$ pour une lecture guidée ou une lecture indépendante
- Rendre les systèmes d'ALAO actuels plus adaptatifs au nivau du lexique
  $\longrightarrow$ sélection d'activités de niveau adaptée, focus sur les mots réellement inconnus, etc.
- Détecter les mots difficiles en vue d'une simplification (manuelle ou automatique) de textes
- ...

# 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

- Présentation des dernières avancées.

## 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 niveaux de difficulté sont prédits 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

# 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

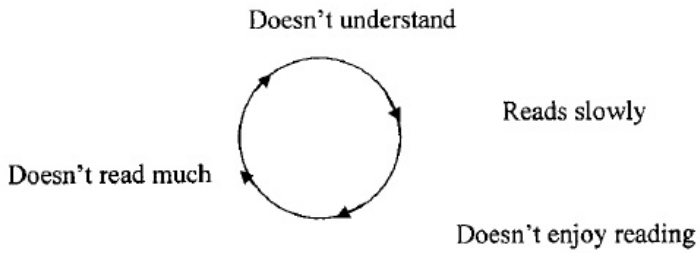
## Question 2 : comment apprend-on de nouveaux mots ?

- 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],

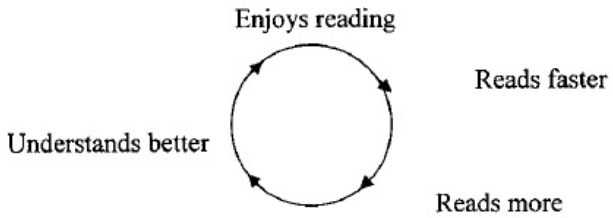
  qui peut entraîner un cercle vicieux.




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.



# 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 et référentiels

- 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) |

# 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 !

# 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 |

# Perspectives présentées l'année dernière

- Nettoyer manuellement la version CRF

# Perspectives présentées l'année dernière

- Nettoyer manuellement la version CRF
- Utiliser FLELex pour prédire le vocabulaire connu et inconnu d'un lecteur donné

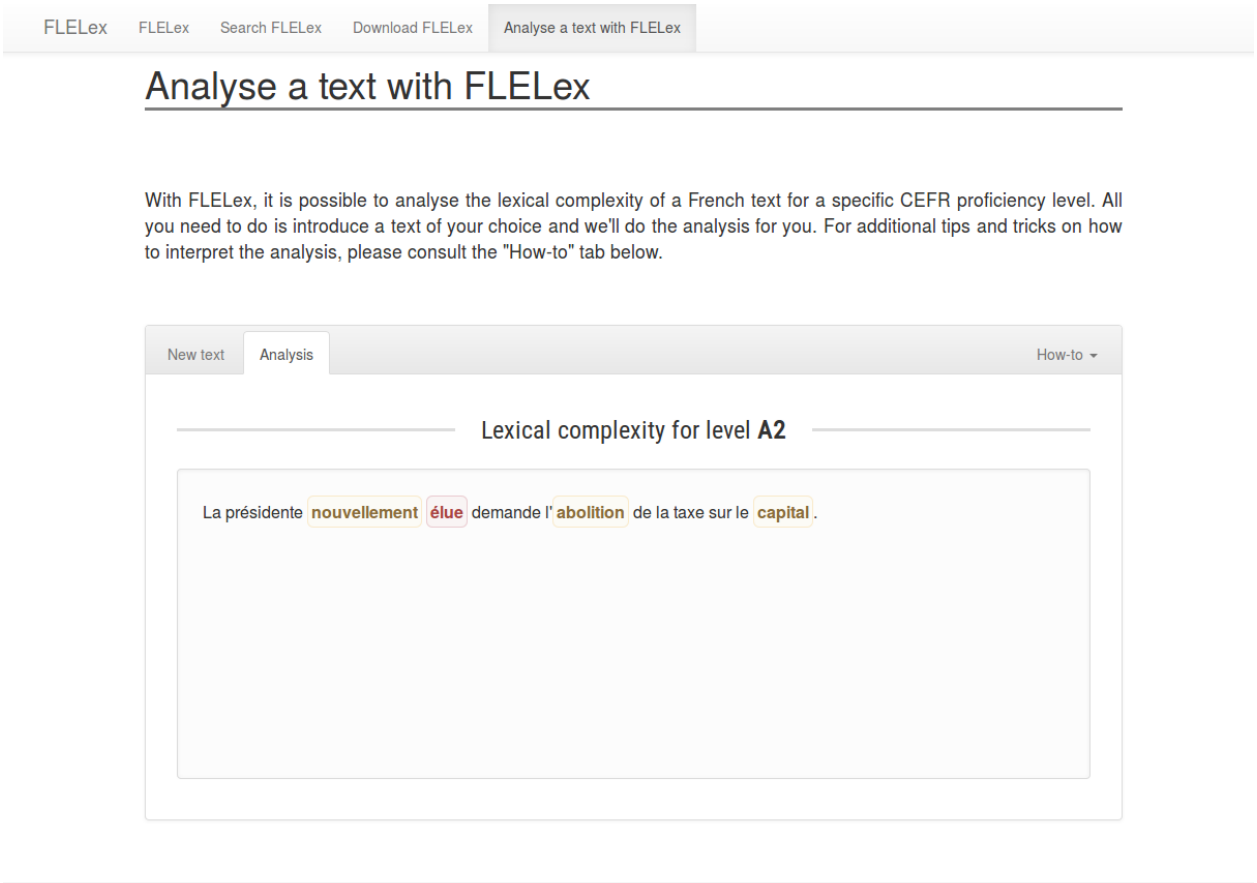
# Perspectives présentées l'année dernière

- Nettoyer manuellement la version CRF
- Utiliser FLELex pour prédire le vocabulaire connu et inconnu d'un lecteur donné
- Offrir des versions de "FLELex" pour d'autres langues

# Perspectives présentées l'année dernière

- Nettoyer manuellement la version CRF
- Utiliser FLELex pour prédire le vocabulaire connu et inconnu d'un lecteur donné
- Offrir des versions de "FLELex" pour d'autres langues
- Ajouter un onglet sur le site, permettant d'analyser directement un texte

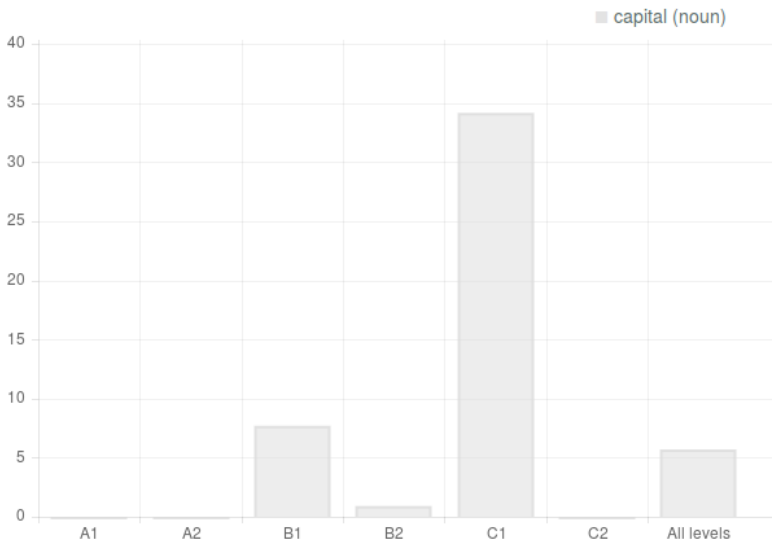
# Démonstration



# 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.

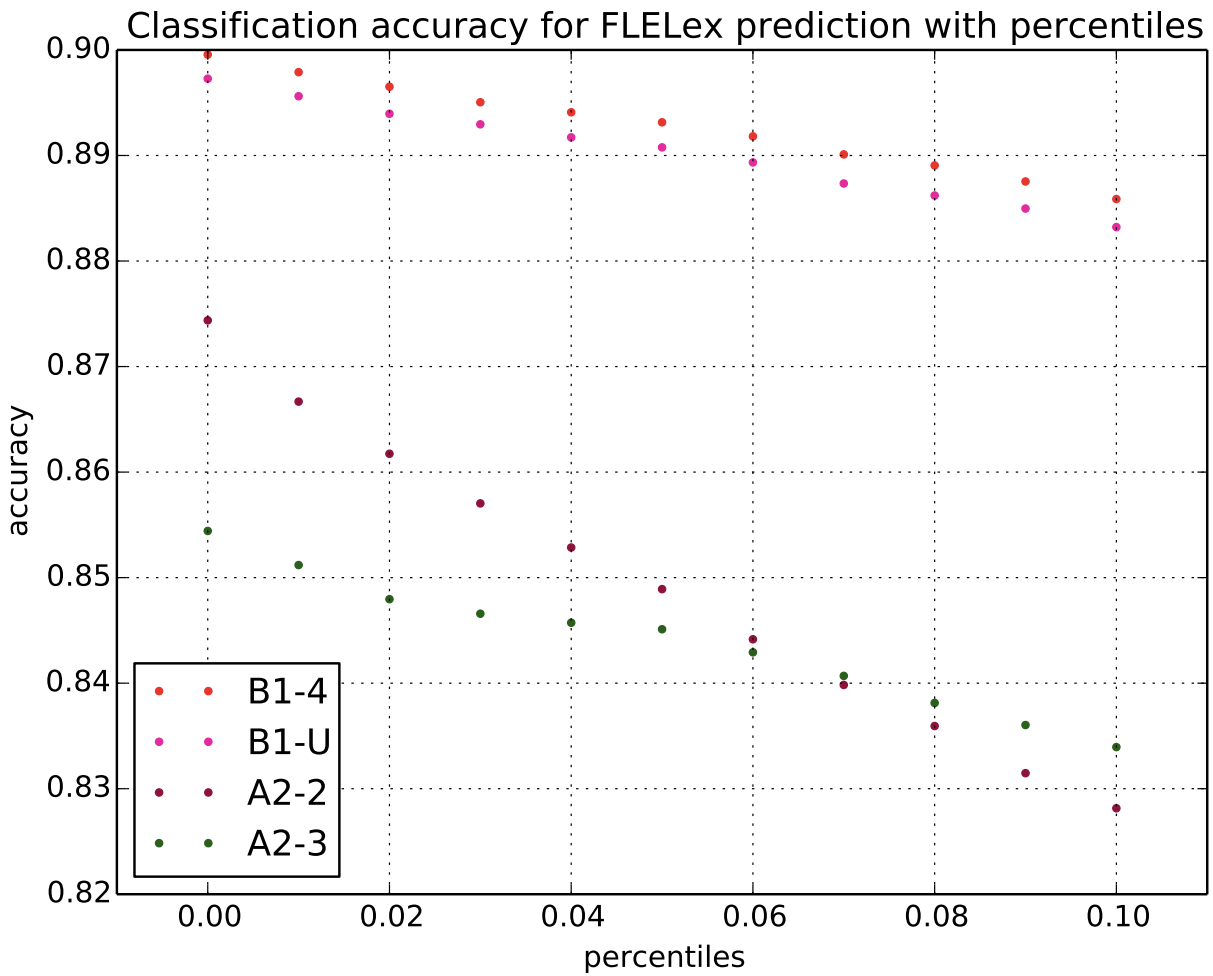
... est transformée en B1

# 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



# 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

# Problématique

Il arrive souvent qu'un texte contienne des termes trop complexes pour un apprenant !

**Intuition du professeur concernant les mots complexes (pour A2)**

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

2 possibilités :

- Utiliser un autre texte (pas toujours possible en fonction du sujet)
- Simplifier les termes les plus problématiques
  $\longrightarrow$ travail habituellement manuel ... et long !

# Objectifs du projet ReSyf

Remplacer les termes complexes par des synonymes plus simples

La présidente nouvellement élue

depuis peu
récemment
...

demande l'abolition de la taxe sur le capital.

abrogation
suppression
annulation
destruction
effacement

# Défi 1

Il est nécessaire de graduer les synonymes pour choisir le plus adapté.

B1
La présidente nouvellement élue

| | |
|---|---|
| A2 | depuis peu |
| A2 | récemment |
| | ... |

B2
demande l'abolition de la taxe sur le capital.

| | |
|---|---|
| C1 | abrogation |
| A2 | suppression |
| A2 | annulation |
| B1 | destruction |
| B1 | effacement |

# Défi 2

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

# 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, Mokthar Billami, Cédrick Fairon

# Ressources d'apprentissage

- **Besoin** : une liste de mots graduée (chaque mot est associé à un niveau)
  → Par ex. Manulex (L1) ou FLELex (L2)
- **Problème :** Ces deux ressources définissent une distribution de fréquence pour chaque mot, pas un niveau unique !
  → 3 approches testées :
  - Niveau de la première occurrence
  - Considérer chaque distribution comme une série statistique et prendre le premier quartile
  - Idem, mais prendre la moyenne
- La 1re semble être la meilleure (validation intrinsèque !)

# Variables pour la difficulté lexicale

- Nombre de lettres, phonèmes, syllabes
- Structure syllabique (structures 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)

- Morphèmes :
  - analyse morphologique automatique non supervisée, découpage en segments morphémiques étiquetés (base, préfixe, suffixe, élt. liaison) et identification de familles morphologiques [Bernhard, 2010]
  - nb morphèmes, préfixation (oui/non), suffixation (oui/non), est composé (oui/non), fréq. minimale préf/suf, fréq. moyenne préf/suf, taille famille morphologique

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

# Variables pour la difficulté lexicale (3)

- Polysémie :
  - utilisation de lexiques sémantiques (réseaux lexicaux)
  - plusieurs sens dans *JeuxDeMots* (oui/non) (`http://www.jeuxdemots.org`) [Lafourcade, 2007]
  - nombre de synsets (groupes de synonymes) dans *BabelNet* (`http://babelnet.org/`) [Navigli and Ponzetto, 2010]

# Résultats (1/2)

- L'efficacité de chaque variable est d'abord évaluée isolément à l'aide d'une corrélation de Spearman (sur Manulex et FleLex) :

| **Variables** | **Manulex ($\rho$)** | **FLELex ($\rho$)** |
|---|---|---|
| 17 Fréquences dans Lexique3 | -0,51 | -0.53 |
| 18 % d'absents de Goug. (5000) | -0,41 | -0.46 |
| 18 % d'absents de Goug. (4000) | -0,41 | -0.47 |
| 02 Nombre de phonèmes | 0,30 | 0,27 |
| 15 Polysémie | -0,29 | -0.38 |
| 01 Nombre de lettres | 0,27 | 0,25 |
| 03 Nombre de syllables | 0,27 | 0,26 |
| 4a Nombre de voisins | -0,25 | -0,23 |
| 4b Voisin freq. cumulée | -0,25 | -0,23 |
| 16 Synset BabelNet | -0,20 | -0,19 |
| 6b Voy. Nasale | 0,08 | 0,07 |
| 14 Taille famille (morphoclust_10) | -0,08 | -0,05 |
| 08 Nombre de morphèmes (seg_10) | 0,06 | 0,08 |
| 06 Patrons orthographiques (a-d) | 0,05 | 0,06 |

# Résultats (2/2)

- Entraînement d'un algorithme d'apprentissage SVM sur la base des meilleurs prédicteurs par famille.
- Evaluation des performances du modèle sur les données de Manulex (26 var.) et de FLELex (24 var.)
- Baseline 1 : prédiction de la classe majoritaire dans les données
- Baseline 2 : modèle basé uniquement sur la fréquence des mots

| Liste | Modèle | Coût | Exac. | Ecart-type |
|---|---|---|---|---|
| Manulex | Classe majoritaire | / | 48% | / |
| | Baseline Fréq. | 0,1 | 61% | 0,4% |
| | Modèle | 0,5 | 63% | 0,7% |
| FLELex | Classe majoritaire | / | 28,8% | / |
| | Baseline Fréq. | 0,5 | 39% | 0,8% |
| | Modèle | 0,001 | 43% | 0,5% |

# Retour à l'exemple

Exemple de résultats avec ReSyf v1.0

| | lvl 3 | lvl 1 |
|---|---|---|
| La présidente | nouvellement | élue |

| | |
|---|---|
| lvl 1 + lvl 1 | depuis peu |
| lvl 3 | récemment |

| | lvl 3 | | lvl 3 | |
|---|---|---|---|---|
| demande l' | abolition | de la taxe sur le | capital | . |

| | |
|---|---|
| lvl 1 | abrogation |
| lvl 3 | suppression |
| lvl 3 | annulation |
| lvl 2 | (destruction) |
| absent | (effacement) |

# Retour à l'exemple

Même analyse, mais en utilisant FLELex

| | B1 | absent |
|---|---|---|
| La présidente | nouvellement | élue |

| | |
|---|---|
| B1 | depuis peu |
| A2 | récemment |

| | B1 | | B1 | |
|---|---|---|---|---|
| demande l' | abolition | de la taxe sur le | capital | . |

| | |
|---|---|
| absent | abrogation |
| B1 | suppression |
| A2 | annulation |
| B1 | (destruction) |
| absent | (effacement) |

# Conclusions et perspectives

- Complexité lexicale : modèle capable de prédire le niveau de difficulté en fonction de paramètres intralexicaux est-il possible ?
- Ressources lexicales : développement de lexiques gradués intégrant la notion de difficulté
- Applications de CALL : pour l'aide à la lecture, l'accès facilité à des documents importants, etc.
- Difficulté en contexte : nécessité d'évaluer la complexité des mots en contexte, pour un apprenant donné !

# Plan

# Vers la prédiction de la difficulté lexicale individualisée

- Mémoire de Mlle Anaïs Tack sur la question.
- Jeu de données : annotations de 4 apprenants (A2 et B1) sur 51 courts textes (via une interface web).
- 2 expériences :
  1. Utilisation de FLElex pour prédire les mots inconnus
  2. Entraînement d'un modèle spécifique à chaque apprenant

# 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.

# Modèle individuel de prédiction de la difficulté lexicale

- Données d'entraînement : annotations d'un apprenant
- Variables simples :
  - Partie du discours
  - Informations sur le mot issues de FLELex
  - nb. lettres, nb. voisins orthographiques, patrons orthographiques
  - nb. synsets dans BabelNet
  - Mêmes variables pour le mot suivant et le mot précédent
- Modèle : réseau de neurones
- Evaluation : validation croisée à 10 échantillons

# 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 et le modèle individualisé

# Discussion

- Apparemment, 3 modèles individualisés sur 4 améliorent les prédictions...
- Problème : les modèles manquent de mots inconnus !

**A2-2**

| $M_P$ | | *inconnu* | *connu* |
|---|---|---|---|
| | *inconnu* | **0** | 0 |
| | *connu* | 431 | **4 986** |
| | | R : 0,0% | S : 100,0% |

**A2-3**

| $M_P$ | | *inconnu* | *connu* |
|---|---|---|---|
| | *inconnu* | **410** | 380 |
| | *connu* | 737 | **3 977** |
| | | R : 35,7% | S : 93,1% |

**B1-4**

| $M_P$ | | *inconnu* | *connu* |
|---|---|---|---|
| | *inconnu* | **28** | 39 |
| | *connu* | 288 | **5 062** |
| | | R : 8,9% | S : 99,2% |

**B1-U**

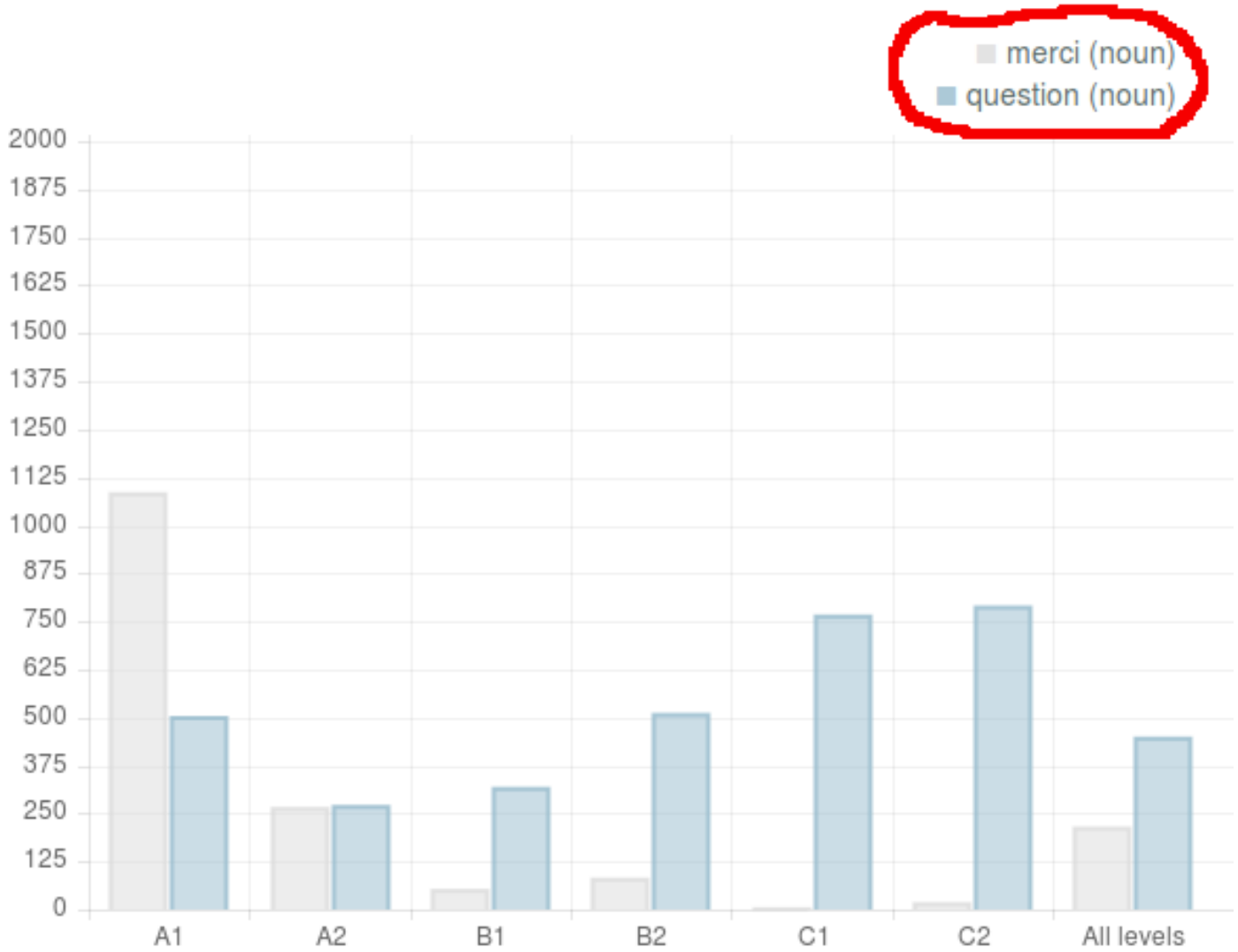
| $M_P$ | | *inconnu* | *connu* |
|---|---|---|---|
| | *inconnu* | **12** | 20 |
| | *connu* | 327 | **5 058** |
| | | R : 3,5% | S : 99,6% |

# 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.
- 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).
- 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



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.


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. (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 II



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.


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)*.


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.

# References III



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 IV


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

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

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

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 V



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 VI



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.


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

# References VII


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


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.


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.