Modèles statistiques pour l'estimation automatique de la difficulté lexicale et syntaxique en FLE.

Thomas François

Aspirant FNRS, CENTAL

Université Catholique de Louvain

Plan de la présentation

١. Introduction : contexte et problématique
٢. Modèles de la difficulté d'un texte
٣. Les étapes méthodologiques :
   A) Collecte du corpus
   B) Les facteurs de la difficulté
   C) Les modèles statistiques utilisés
٤. L'évaluation
٥. Perspectives et conclusion
٦. Bibliographie

Le Contexte

- **Un constat :** Le secteur de l'enseignement des langues étrangères est en pleine croissance et évolution...
  - **Le nombre de professionnels ne suffit plus à satisfaire à la demande.**
    - Il faut décharger au maximum les professeurs des tâches répétitives (ex. de drill, recherche de matériaux pour des tâches de lecture...)
  - **Les apprenants désirent plus de souplesse dans les méthodes d'enseignement.**
    - Développement de logiciels d'auto-apprentissage, mais qui pêchent encore au niveau de leur capacité d'adaptation.

Le Contexte

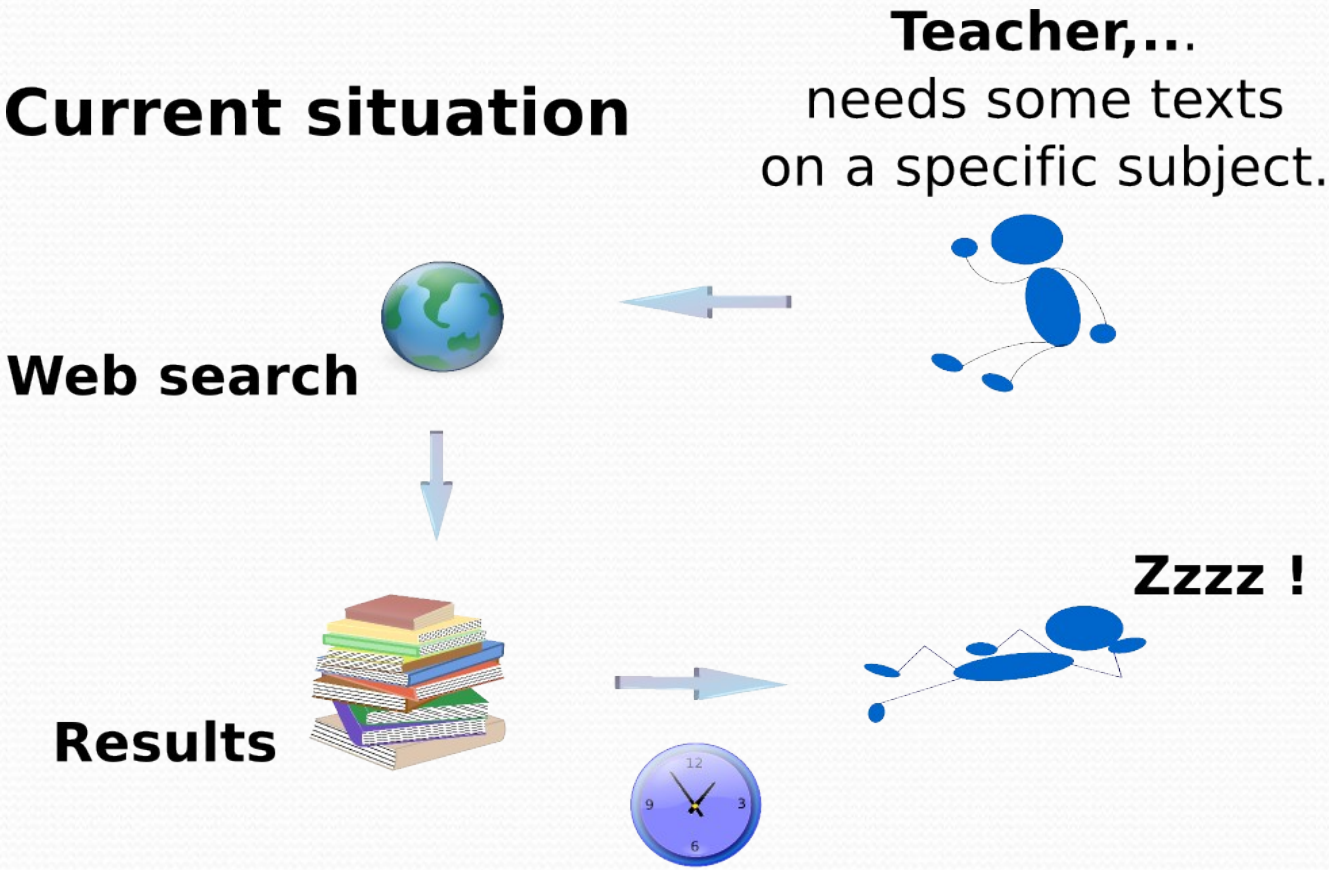
- **Deux applications possibles** :
  - moteur de recherche pour le repérage automatique de textes authentiques portant sur un sujet particulier.

    Collins-Thompson and Callan (2004), Miltsakaki (2009)
  - conception automatique d'exercices sur base d'un corpus de textes.

    Chanier & Selva (2000), Selva (2002)

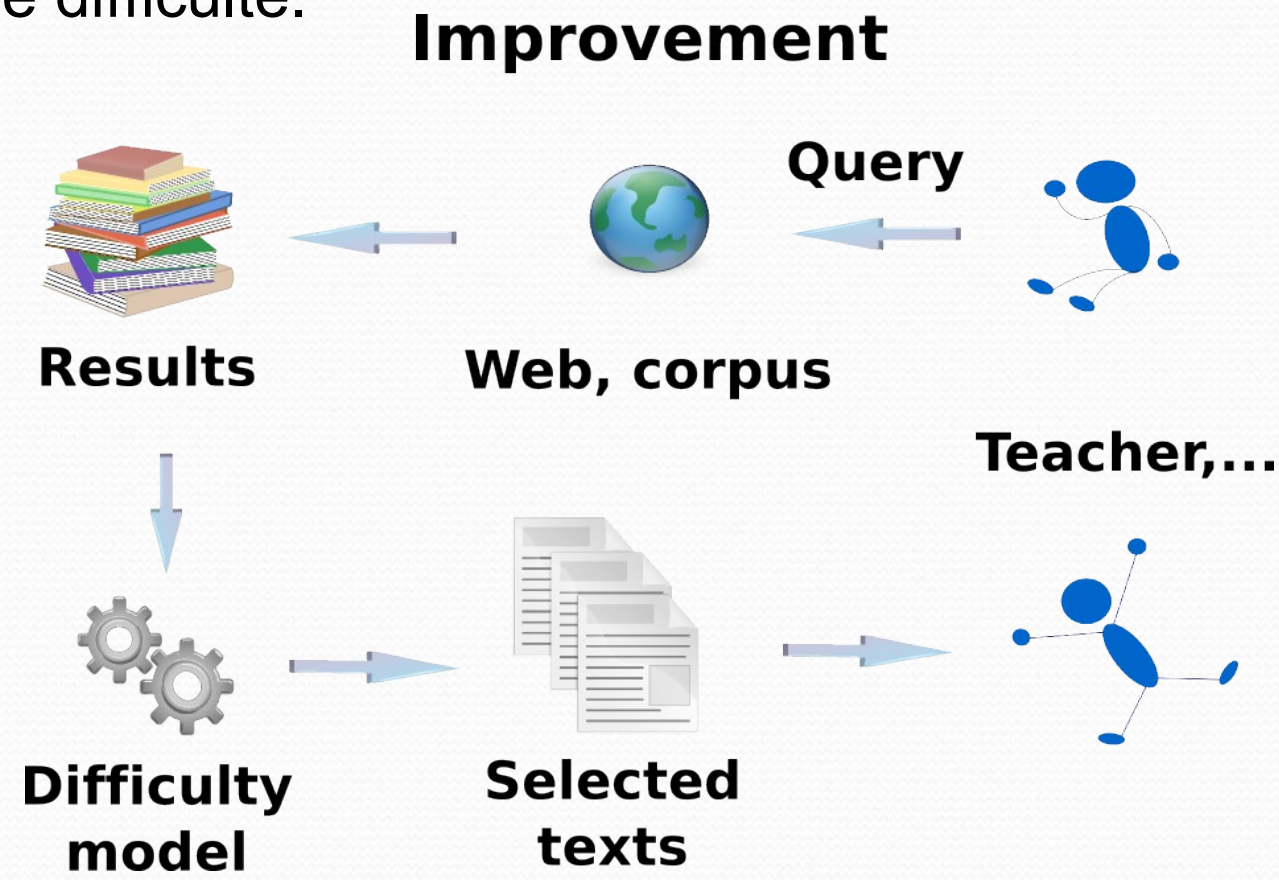
Problématique :
estimer la difficulté des textes

- Pour ces deux tâches, il est nécessaire de contrôler le niveau des documents authentiques utilisés comme matériaux.



Problématique :
estimer la difficulté des textes

- **Possibilité d'amélioration:** sélection automatique de textes d'un niveau de difficulté précis à l'aide d'un filtre, basé sur un modèle de difficulté.



Plan de la présentation

1. Introduction : contexte et problématique
2. Modèles de la difficulté d'un texte
3. Les étapes méthodologiques :
   A) Collecte du corpus
   B) Les facteurs de la difficulté
   C) Les modèles statistiques utilisés
4. L'évaluation
5. Perspectives et conclusion
6. Bibliographie

Modèles de la difficulté d'un texte

- Le développement d'un modèle apte à déterminer la complexité d'un texte à la lecture ressort à la **lisibilité**.
- Domaine essentiellement anglo-saxon :
  - Premières formules (USA) : ne considèrent que le lexique
    Lively and Pressey (1923), Vogel and Washburne (1928)
  - Formules classiques : basées sur la régression linéaire et 2 prédicteurs (un lexical, un syntaxique)
    Flesch (1948), Dale and Chall (1948)
  - Approches cognitivistes et critiques des formules « classiques »
    Kintsch and Vipond (1977, 1979), Kemper (1983)

Peu de travaux en français et FLE

- Français L1: on compte peu de travaux
  - « Adaptation » des formules anglophones
    Kandel and Moles (1958) ; de Landsheere (1963)
  - Création de formules particulières au français
    Henry (1975) ; Richaudeau (1979) ; Mesnager (1989)
- Très peu de travaux pour le FLE
  Cornaire (1989) ; Uitdenbogerd (2005)
  François (2009a), François (2009b)

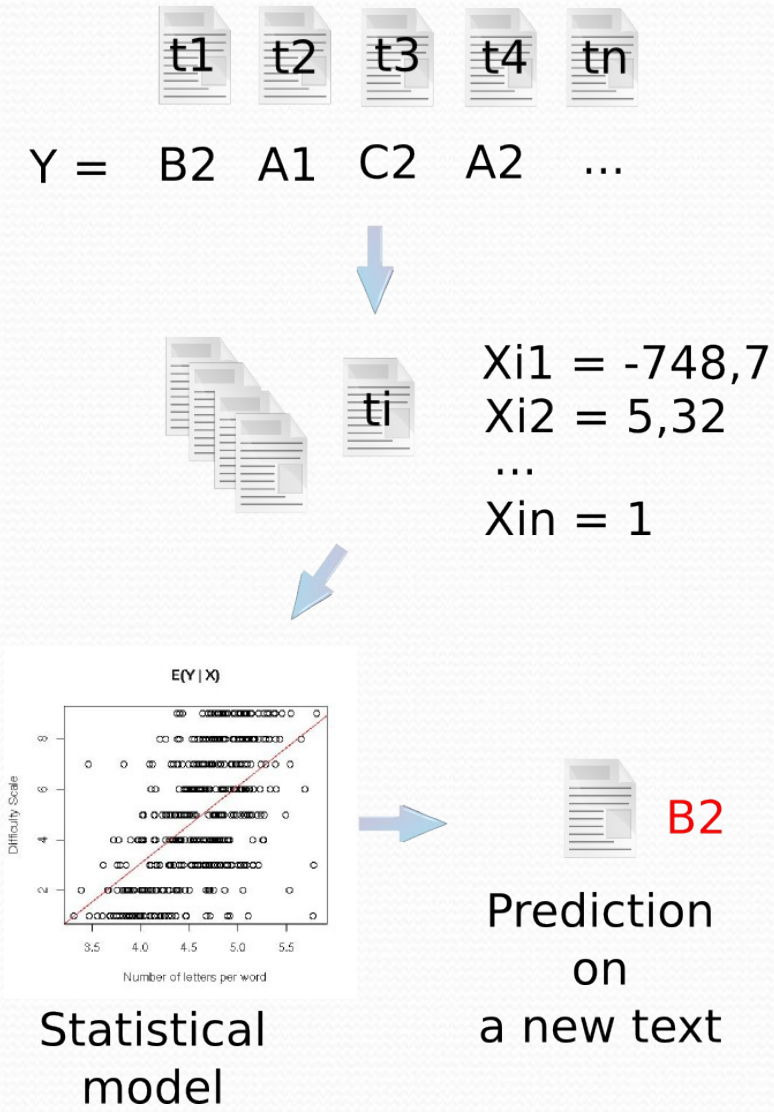
→ **Il existe bien un réel besoin pour un tel modèle !**

Plan de la présentation

1. Introduction : contexte et problématique
2. Modèles de la difficulté d'un texte
3. Les étapes méthodologiques :
   A) Collecte du corpus
   B) Les facteurs de la difficulté
   C) Les modèles statistiques utilisés
4. L'évaluation
5. Perspectives et conclusion
6. Bibliographie

Méthodologie pour un modèle de difficulté

1. Rassembler un corpus de textes annotés en fonction d'une échelle de difficulté.
   - Partie 3.A
2. Définir une liste d'indices linguistiques de la difficulté d'un texte.
   - Partie 3.B
3. Entraîner un modèle statistique prédictif (de classification) sur la base de ces variables.
   - Partie 3.C
4. Valider le modèle.



Plan de la présentation

1. Introduction : contexte et problématique
2. Modèles de la difficulté d'un texte
3. Les étapes méthodologiques :
   - A) Collecte du corpus
   - B) Les facteurs de la difficulté
   - C) Les modèles statistiques utilisés
4. L'évaluation
5. Perspectives et conclusion
6. Bibliographie

Collecte du corpus

- **Problématique :** obtenir un corpus annoté en fonction d'une échelle de difficulté et d'un critère.
- **L'approche classique :**
  - Critère : tests (de compréhension ou test de closure) sur une population.
    + prise en compte de la variable lecteur.
    - peu de textes : peu pratique pour le TAL.
  - L'échelle de mesure retenue est quantitative : pourcentages, correspondant à la réussite moyenne aux tests.
    → Induit comme modèle statistique une régression linéaire.

Collecte du corpus

- **Notre approche** :
  - Les besoins :
    - L'approche TAL requiert de nombreux textes
    - L'échelle de mesure doit avoir une dimension pratique
  - La solution :
    - Depuis 2001, le niveau de difficulté des manuels de FLE doit être exprimé selon l'échelle du « *Cadre européen commun de référence pour les langues* » (CECR).
    - Il est dès lors possible d'utiliser les manuels comme corpus.

L'échelle du CECR

- **L'échelle du CECR :**
  - Elle comprend 6 niveaux de base :
    - A1 (+ facile), A2, B1, B2, C1, C2 (+complexe)
  - Certains auteurs/professeurs préconisent de raffiner l'échelle en divisant certains niveaux :
    - J'obtiens 9 niveaux : A1, A1+, A2, A2+, B1, B1+, B2, C1, C2.
  - ➔ Cette division permet de mieux prendre en compte les différences de compétence pour les apprenants de niveau plus faible, où elles sont plus notables que dans les niveaux supérieurs.

Critères de sélection des textes

- Il n'est pas possible d'utiliser tous les manuels comme corpus, ni l'ensemble des textes.

- **Critères de sélection :**
  1. Manuels postérieurs à 2001 (publication du CECR).
  2. Manuels destinés à un public d'adultes ou de jeunes gens.
  3. Manuels généralistes (pas de langage de spécialité)
  4. À l'intérieur des manuels, seuls les textes associés à une tâche de compréhension à la lecture ont été considérés.

- **Taille actuelle du corpus** :
  - Environ 2,000 documents et 500,000 tokens.

Le bruit dans le corpus

- **Un problème** : le bruit dans le corpus.
- **Causes possibles :**
  - Dans les manuels par tâches, il arrive de rencontrer un texte complexe associé à une tâche plus simple.
    - Ex. : RP LM = -731 au niveau A1 (moyenne = -700)
  - Parfois, au contraire, un texte simple est associé à une tâche complexe (ex. une chanson au niveau B2)
- **Solutions :**
  - Actuellement, retrait brutal des « outliers »
  - Voir Perspectives...

Hétérogénéité des manuels

A1	A2	B1	B2	C1	C2
/	/	-746	-763	-766	-787
-705	-723	/	/	/	/
/	-749	-757	/	/	/
-690	/	/	/	/	/
/	/	/	-758	-766	-777
-694	/	-746	/	/	/
-725	/	/	/	/	/
-696	-730	-753	/	/	/
-731	-742	-733	-766	/	/
/	/	/	/	-787	-778
-664	-712	-756	/	/	/
-711	-740	-752	/	/	/
-683	-740	/	/	/	/
-700.09	-732.9	-750.75	-763.52	-771	-779

Plan de la présentation

1. Introduction : contexte et problématique
2. Modèles de la difficulté d'un texte
3. Les étapes méthodologiques :
   - A) Collecte du corpus
   - B) Les facteurs de la difficulté
   - C) Les modèles statistiques utilisés
4. L'évaluation
5. Perspectives et conclusion
6. Bibliographie

Les indices de la difficulté

- **Les indices de la complexité lexicale (4)**
  - Modèle de langue (LM1, LM2, LM3)
  - Nombre moyen de lettres par mot (NLM)
- **Les indices de la complexité syntaxique (12)**
  - Nombre moyen de mots par phrases (NMP)
  - 11 variables binaires : temps et modes verbaux
- **Les variables de dialogue (5)**
  - 5 variables qui visent à déterminer le genre de texte (dialogue >< narration), basées sur l'étude de Henry (1975).

Pourquoi un modèle de langue ?

- Actuellement, le modèle de langue est le seul indice TAL.
  - **Hypothèse** : divers travaux en psycholinguistique suggèrent une association entre la difficulté des mots et leur fréquence.

    Howes and Salomon (1951) ; Brysbaert et al. (2000)
  - **L'approche classique :** on emploie un pourcentage de mots absents d'une liste reprenant les mots les plus fréquents de la langue (ex. Français Fondamental, Gougenheim, 1964).

    Dale and Chall (1948) ; Henry (1975)
  - **Amélioration** : Un modèle unigramme lissé peut être utilisé avantageusement à la place de ces listes.

    Collins-Thompson and Callan (2005)

Principes du modèle de langue

- **Difficulté d'un texte = probabilité du texte :**
  - Sur base de cette hypothèse, on peut donc estimer la difficulté d'un texte T (avec N tokens) comme la probabilité de ce texte dans la langue :

$$P(T)=P(t_1\cap\ldots\cap t_n)=\prod_{i=1}^{n} p(t_i|t_{i-1},\cdots,t_1)$$

- Ce modèle, afin d'être entraînable, est approximé en un unigramme (ce qui revient à postuler l'hypothèse des tokens) :

$$P(T)=\prod_{i=1}^{n} p(t_i)$$ où $p(t_i)$ représente la probabilité de rencontrer le token t_i en français.

La meilleure unité ?

- **Une question** : étant donné la nature flexionnelle du français, la forme fléchie (token) est-elle bien la meilleure unité pour l'unigramme ?
- **Unités possibles** :
  - Le lemme (LM1)
  - La forme fléchie (= token) (LM2)
  - La forme fléchie désambiguïsée (LM3)

    (using TreeTagger, Schmid, 1994)
- Les coefficients de corrélations avec la difficulté sont similaires :

Unit	LM1	LM2	LM3
Correlation (r)	-0.58	-0.58	-0.59

- Nous avons choisi le lemme.

Le modèle de langue : détails

- La liste des lemmes et de leur probabilité vient de *Lexique3*, développé par New et al. (2001).
  - Les fréquences ont été estimées sur un corpus de sous-titres de films comprenant environ 50 millions de mots.

```
1_lemme 2_cgram  9_freqlemfilms

aïe     ONO          18.25
cuire   VER          21.65
cuisant ADJ          0.49
île     NOM          58.35
```

  - Les probabilités ont été lissées à l'aide de l'algorithme « Simple Good-Turing ».

    Gale and Sampson (1995)

Les indices classiques

- **Deux indices classiques de la difficulté**
  - Au niveau lexical : nombre moyen de lettres par mots (NLM)
  - Au niveau syntaxique : nombre moyen de mots par phrases (NMP)
  - Leur efficacité est bien connue. Ils ont été utilisés comme indices (avec une variante : nombre moyen de syllabes par mots) dans de nombreuses formules.

    Mc Clusky (1934) ; Flesh (1948) ; Smith (1961) ; Flesch-Kincaid (1975)...

Les indices syntaxiques

- **Emploi des temps et modes verbaux** :
  - Liste des 11 variables binaires (choix lié au TreeTagger) :

Conditionnel	Future	Impératif
Imparfait	Infinitif	Participe passé
Participe présent	Indicatif présent	Passé simple
Subjonctif présent	Subjonctif imparfait	

  - **Objectif** : modéliser le rythme de l'enseignement de la grammaire dans un contexte de FLE.
  - **Approche optimale** : reconnaissance automatique des structures.
    Heilman et al. (2007)
  - **Problème** : les parseurs syntaxiques pour le français manquant encore de précision, nous avons opté pour cette paramétrisation particulière.

Les variables de dialogue

- **Hypothèse** : les dialogues sont plus simples que les narrations, les textes didactiques ou scientifiques :
  - Vocabulaire familier ;
  - Structures plus simples ;
  - Sujets plus quotidien.

  Henry (1975)

- **Variables de dialogue :**
  - Ratio des pronoms personnels de dialogue (1re, 2e pers.)
  - Ratio du nombre d'interjection sur le nombre de mots
  - Ratio de ! et ? par rapport à ! ? . ;
  - Ratio de ! et ? par rapport à ! ? . ; :
  - Présence de guillemets de dialogue (variable binaire)

Les données...

- À la fin de cette opération de paramétrisation, on obtient le tableau de données suivant :

Diff	ML1	ML2	ML3	NLM	NMP	PPD	PI	PPEI_1	...
1	-646.86	-604.59	-647.63	3.59184	5.15	0.33	0.019	0.412	...
1	-669.90	-709.11	-731.24	3.54	4.875	0.7	0.026	0.375	...
1	-686.26	-684.10	-725.34	4.54	14.5	0.0	0.0	0.11	...
1	-656.56	-649.68	-695.13	3.90	5.90	0.47	0.021	0.81	...
1	-633.65	-685.45	-703.32	3.86	6.90	0.41	0.037	0.57	...
1	-677.87	-686.38	-721.16	4.03	3.81	0.65	0.04	0.65	...
1	-673.63	-646.13	-687.27	4.675	2.11	1.0	0.1	0.63	...
1	-698.23	-690.75	-740.27	4.23	12.44	0.04	0.01	0.5	...
...									
4	-758.3	-782.57	-788.77	4.32	21.66	0.0	0.0	0.33	...
...									
6	-823.42	-859.91	-893.27	5.42	26.76	0.0	0.0	0.037	...

Plan de la présentation

1. Introduction : contexte et problématique
2. Modèles de la difficulté d'un texte
3. Les étapes méthodologiques :
   - A) Collecte du corpus
   - B) Les facteurs de la difficulté
   - C) Les modèles statistiques utilisés
     - a) La modélisation statistique
     - b) La régression logistique
     - c) Les méthodes d'agrégation de modèles
4. L'évaluation
5. Perspectives et conclusion
6. Bibliographie

Pourquoi un modèle... statistique ?

- De nombreux phénomènes réels varient d'une manière qui peut sembler aléatoire.
  - Ex. : pourquoi certains noms prennent-ils un -s final et d'autres pas ?
  - Le phénomène à expliquer est appelé **variable dépendante**.

Exemples de noms :
- amis
- chien
- tours
- soirs
- famille
- modèle
- statistiques
- années

Pourquoi un modèle... statistique ?

- De nombreux phénomènes réels varient d'une manière qui peut sembler aléatoire.
  - Ex. : pourquoi certains noms prennent-ils un -s final et d'autres pas ?
  - Le phénomène à expliquer est appelé **variable dépendante**.
- En réalité, il existe des régularités cachées au sein de ces phénomènes.
  - On peut chercher à les découvrir à l'aide d'informations supplémentaires (**variables explicatives ou indépendantes**)
  - Ici, c'est le mot qui précède.

Exemples de noms :
- Les amis
- Un chien
- Quelques tours
- Les soirs
- La famille
- Un modèle
- Des statistiques
- Vingt années ??
- ...

- Hypothèse :
  Le nom prend un s quand le mot qui le précède a aussi un s.

Pourquoi un modèle... statistique ?

- Les modèles statistiques cherchent les régularités à partir de données observables :
  - Dans notre exemple simpliste, un modèle ne pourra pas postuler un niveau caché de description linguistique tel que le « nombre ».
  - Mais, il va tenter de modéliser ce concept sur la base des informations observables qu'il possède.
  - Cette manière de faire peut-être pratique en TAL où il faut expliciter des concepts à la machine.
- Pour des phénomènes simples, un modèle statistique n'apporte rien à une démarche par règles.
- Par contre, il aide à déterminer des régularités dans des phénomènes complexes tels que la lisibilité d'un texte.

Conception d'un modèle statistique

- **Un modèle statistique** vise soit à :
  - **expliquer** une variable dépendante Y par les variables explicatives X_1, ..., X_n, soit à
  - **prédire** la valeur de Y pour une observation particulière en fonction des valeurs prises par X_1, ..., X_n.

- **Le type de modèle** dépend de la nature de la variable Y :
  - **Nominale** : les résultats sont des « étiquettes », des « classes ».
    ex. : le genre d'un texte, le sexe...
  - **Ordinale** : les résultats sont des classes ordonnées.
    ex. : le niveau social, l'année scolaire, une classe d'âge...
  - **Quantitative** (discrète ou réelle) : les résultats sont des nombres.

Les types de modèles statistiques

- **Les modèles de régression** :
  - Le choix dépend de la nature de la variable dépendante :

Continue	→	Régression linéaire
Ordinale	→	Régression logistique ordinale (PO)
Nominale	→	Régression logistique multinomiale

- **Les modèles par arbres de décision :**
  - Arbre de classification (baseline) *(Breiman et al., 1984)*
  - Boosting *(Freund & Schapire, 1996)*
  - Bagging *(Breiman, 1996)*

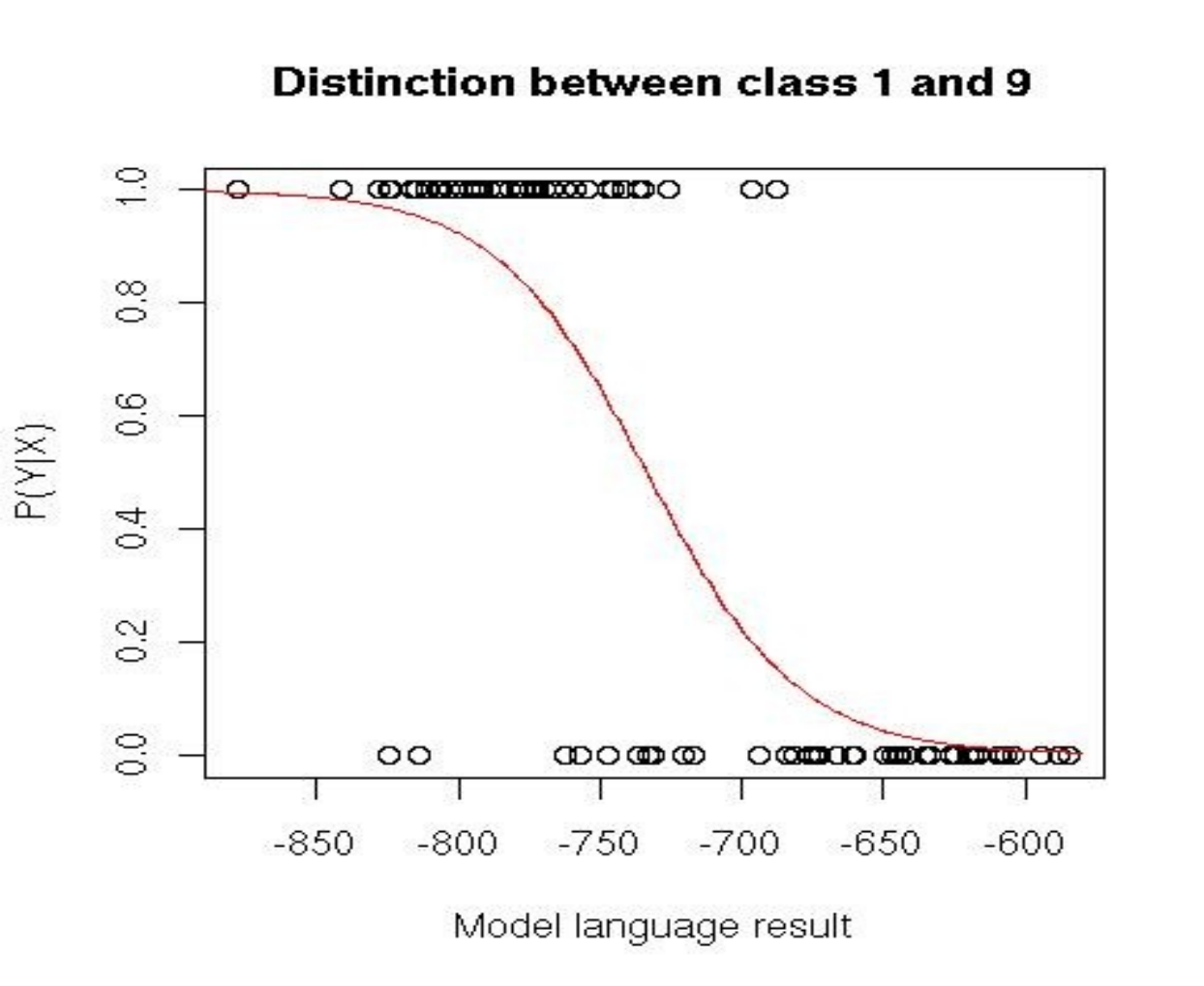
- La régression linéaire multiple a été rejetée suite à de mauvais résultats.

Plan de la présentation

1. Introduction : contexte et problématique
2. Modèles de la difficulté d'un texte
3. Les étapes méthodologiques :
   - A) Collecte du corpus
   - B) Les facteurs de la difficulté
   - C) Les modèles statistiques utilisés
     - a) La modélisation statistique
     - b) La régression logistique
     - c) Les méthodes d'agrégation de modèles
4. L'évaluation
5. Perspectives et conclusion
6. Bibliographie

Introduction aux modèles logistiques

La régression logistique vise à modéliser E(Y | X), non pas comme une droite, mais à l'aide d'une sigmoïde.



Introduction aux modèles logistiques

- **Notre problème** :
  - N (21) variables indépendantes : X_1, X_2, ..., X_n
  - K (6 ou 9) niveaux pour la variable dépendante Y.
- **Solution** :
  - Réduction à K-1 modèles binaires → K-1 courbes de réponse.
- **2 approches :**
  - **PO modèle** : toutes les courbes ont la même forme (les coefficients sont les mêmes) ; seule l'ordonnée à l'origine change.
  - **RLM** : chaque fonction est déterminée par un ensemble différent de paramètres → modèle est plus complexe.

Plan de la présentation

1. Introduction : contexte et problématique
2. Modèles de la difficulté d'un texte
3. Les étapes méthodologiques :
   - A) Collecte du corpus
   - B) Les facteurs de la difficulté
   - C) Les modèles statistiques utilisés
     - a) La modélisation statistique
     - b) La régression logistique
     - c) Les méthodes d'agrégation de modèles
4. L'évaluation
5. Perspectives et conclusion
6. Bibliographie

Bagging et Boosting : principe

- **Principe** :
  - Une hypothèse simple (linéaire) n'est jamais parfaite.
  - La combinaison de plusieurs hypothèses simples peut se révéler fort efficace.
- **Méthode** :
  - Entraîner une série de classifieurs simples (arbres de classification).
  - Les combiner pour former un classifieur performant.
- **Problèmes** :
  - Comment entraîner T classifieurs, sachant qu'il faut alors T corpus différents ?
  - Comment combiner ces T classifieurs en un seul ?

Bagging (Bootstrap AGGregatING)

- **Technique basée sur le rééchantillonnage.**
  - Génération de T échantillons « bootstrap ».
  - Sur la base de ces T échantillons, on entraîne t classifieurs.
  - On agrège ces T modèles :
    - Par vote (classe majoritaire) pour une variable qualitative.
    - Par moyenne pour une variable quantitative.
- Le bagging réduit fortement la variance d'un modèle par rapport à un seul arbre de classification, lequel ne peut trouver que des optimum locaux.
- Par contre, réduit aussi la lisibilité du modèle...

Boosting

- **Technique adaptative :**
  1) Le boosting utilise un seul échantillon, dans son entièreté.
  2) Il se concentre sur les données difficiles à modéliser.
  3) Chaque modèle s'adapte en fonction du précédent : les individus ayant été mal classés voient leur poids augmenter.
  4) Adapté à des données binaires.

- **Agrégation** :
  - On fait une moyenne pondérée des différents modèles, où les poids sont définis en fonction de la qualité d'ajustement de chaque modèle :

$$f_{Ens}(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t h_t(\boldsymbol{x})$$

Boosting : algorithme

- **Input** : N observations { ($\mathbf{x}_1$, y_1),..., ($\mathbf{x}_N$, y_N) }
- **Initialisation** : $w_n^{(1)} = \frac{1}{N} \quad pour\ tout\ n = 1, \cdots, N$
- **Boucle** (pour t = 1, ..., T) :

  Entraînement du classifieur h_t en fonction de **x** et des poids.

  2. Calcul de l'erreur pondérée : $\epsilon_t = \sum_{n=1}^{N} w_n^{(t)} \boldsymbol{I}\left(y_n \neq h_t(\boldsymbol{x_n})\right)$

  3. Calcul du poids de ce modèle : $\alpha_t = \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t}$

  4. Adaptation des poids des observations :

$$w_n^{t+1} = w_n^t + \exp\left(-\alpha_t\, y_n\, h_t(\boldsymbol{x_n})\right) / Z_t$$

- **Output** : $f_{Ens}(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t\, h_t(\boldsymbol{x})$

Plan de la présentation

1. Introduction : contexte et problématique
2. Modèles de la difficulté d'un texte
3. Les étapes méthodologiques :
   A) Collecte du corpus
   B) Les facteurs de la difficulté
   C) Les modèles statistiques utilisés
4. L'évaluation
5. Perspectives et conclusion
6. Bibliographie

Les données d'évaluation

- **2 échelles de difficulté = 2 jeux de données :**
  - Modèle à 6 niveaux : 299 textes formant un échantillon i.i.d. (Corp6) du corpus entier.
  - Modèle à 9 niveaux : 449 textes formant un échantillon i.i.d. (Corp9) du corpus entier .
- **Traitement des données aberrantes :**
  - Un « outlier » = donnée située à plus de 3 écart-types de la moyenne de sa classe (pour une variable).
  - Corp6 : 299 textes – 11 outliers = 288 textes
  - Corp9 : 449 textes – 12 outliers = 437 textes

La procédure d'évaluation

- **Deux procédures d'évaluation** :
  - La sélection « pas à pas » des variables (modèle logistique).
    - Entraînement d'un modèle de base avec la variable prédictive la plus efficace.
    - Second modèle : ajout d'une seconde variable et comparaison avec le premier modèle.
    - Critère : AIC (Akaike's Information Criterion)

      AIC = - 2* log-vraisemblance + 2 k (où k = nb. paramètres)

    - L'algorithme s'arrête lorsque l'AIC du modèle ne diminue plus en ajoutant de nouvelles varibables.
  - Une validation croisée à 10 échantillons.

Mesures d'évaluation

- **Pour déterminer les variables importantes :**
  - AIC (Akaike's Information Criterion)

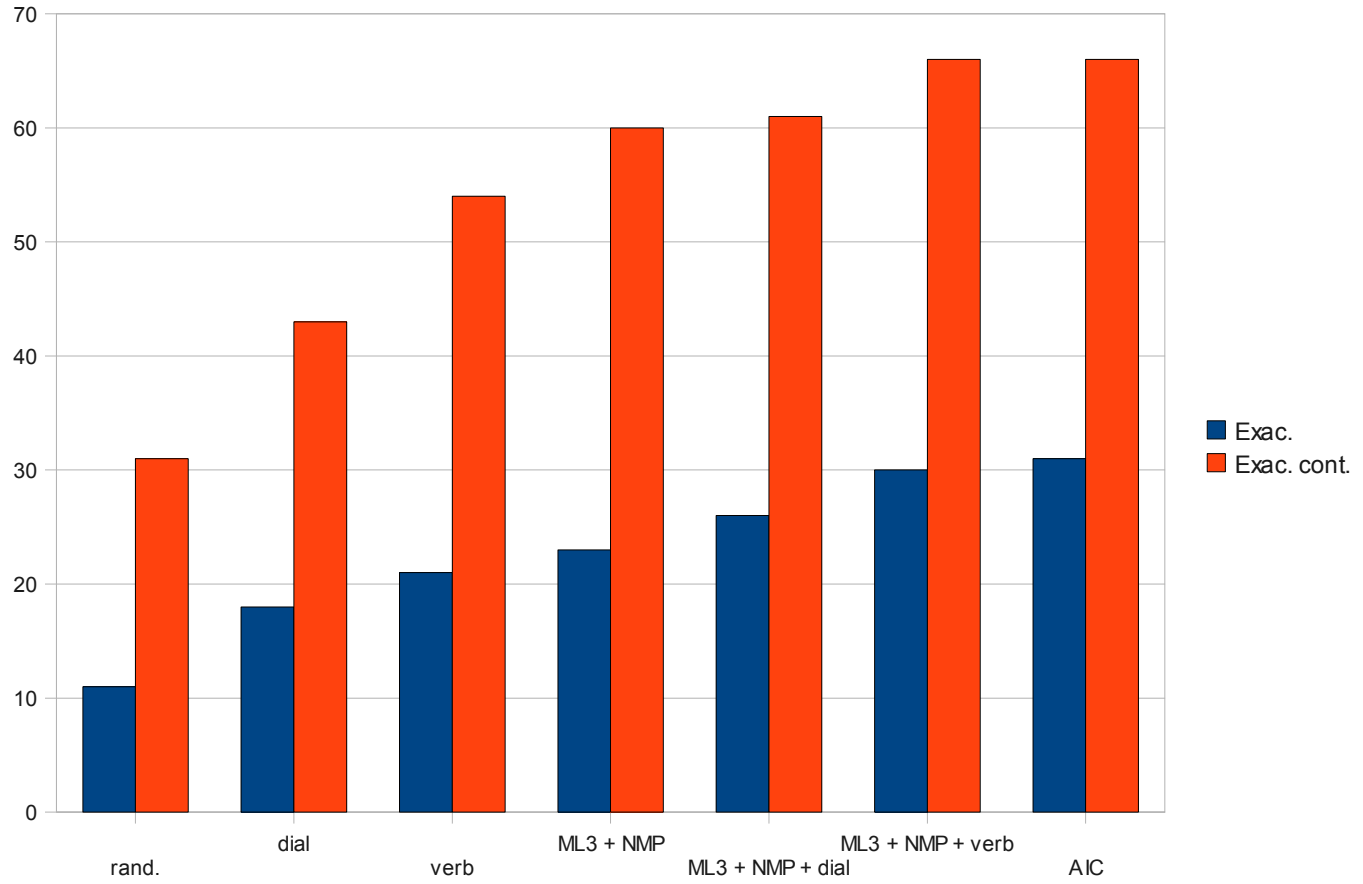
- **Pour l'exactitude (= « accuracy ») des modèles :**
  - Exactitude
  - Exactitude contiguë *(Heilman et al., 2008)*
    - La proportion de prédictions correctes ou erronées d'un seul niveau.
    - Elle vise à prendre en compte la difficulté des experts humains à s'entendre sur une même classification.

Sélection des variables

- La sélection « pas à pas » est sensible aux variations du modèle et des données d'entraînement.
- Liste des variables conservées par modèle et par échantillon :
  - **Régression logistique ordinale :**
    - **Corp6** : ML1 + ML3 + NMP + PPD + PI + PPEI_1 + BINGUI + Futur + Impf + Infi + PPasse + Subp
    - **Corp9** : ML3 + NMP + PPD + PPEI_2 + BINGUI + Cond + Futur + Impf + Infi + PPasse + Pres + Subp
  - **Régression logistique multinomiale :**
    - **Corp6 :** ML1 + NLM + NMP + BINGUI + Futur + Impf + Infi + PasseSim
    - **Corp9** : ML2 + ML3 + NMP + PPD + PPEI_1 + Cond + Futur + Impf + PPasse + Subi + Subp

Importance des variables

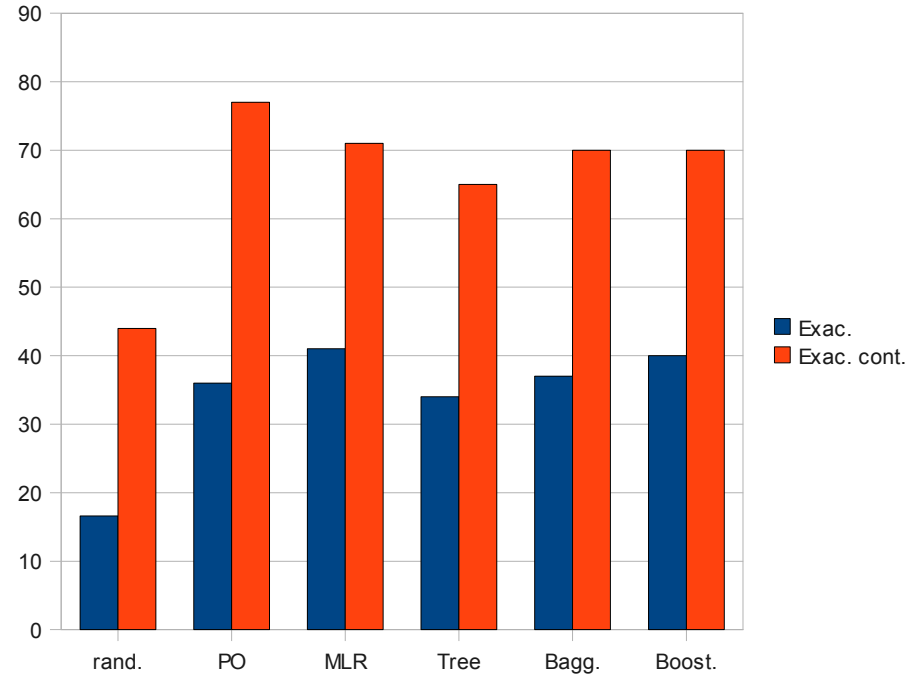
- On retrouve souvent deux variables de type lexical (ML... ou NLM) et NMP : c'est la base de la formule.
- Exemple de décomposition de l'exactitude (modèle PO, Corp9)



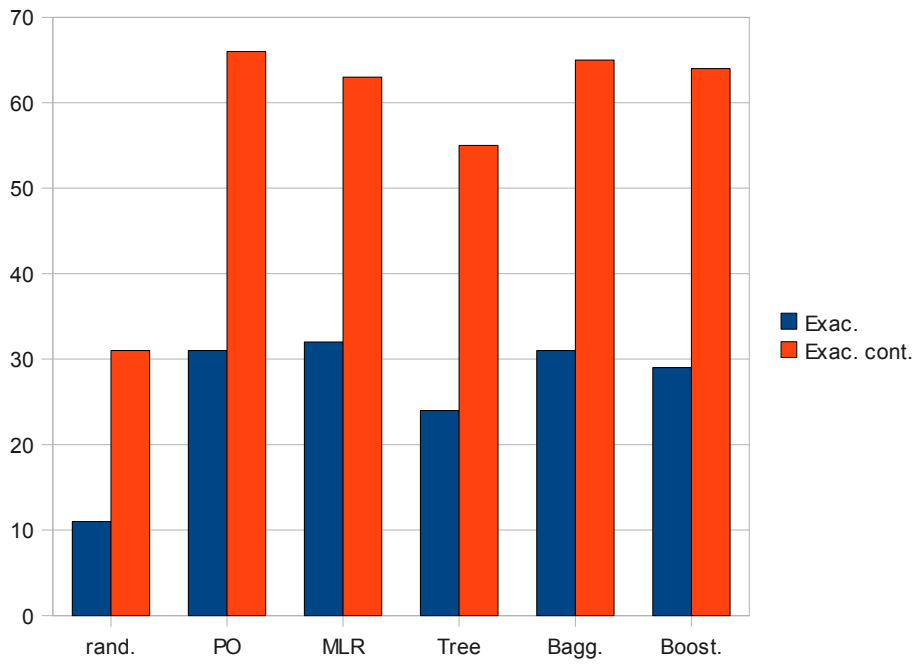
Efficacité des modèles

- Résultats de la validation croisée à 10 échantillons :

Corp6



Corp9



- Exactitude : peu satisfaisante, mais l'exactitude contiguë est meilleure !

Discussion

- Exemples de matrice de confusion :

MLR

real \ prediction	A1	A1+	A2	A2+	B1	B1+	B2	C1	C2
A1	1	1	0	0	1	0	0	0	0
A1+	0	2	1	0	0	0	0	0	0
A2	0	1	0	0	0	0	0	0	0
A2+	0	1	1	1	1	0	1	1	0
B1	0	1	0	1	0	0	0	1	0
B1+	0	0	0	0	2	0	2	0	0
B2	0	1	2	0	1	2	3	0	0
C1	0	0	0	0	1	1	0	3	3
C2	0	0	0	1	0	0	0	0	5

PO

real \ prediction	A1	A1+	A2	A2+	B1	B1+	B2	C1	C2
A1	1	3	0	1	0	0	0	0	0
A1+	0	0	1	0	0	0	0	0	0
A2	1	2	0	0	0	0	1	1	0
A2+	0	0	1	3	2	0	1	0	1
B1	0	0	0	1	1	0	0	0	0
B1+	0	0	0	0	0	0	2	2	1
B2	0	0	0	0	1	0	2	0	2
C1	0	0	0	0	1	0	0	4	1
C2	0	0	0	0	1	0	1	0	4

- Prédictions ne sont pas trop éloignées de la diagonale !
- MLR est meilleure en termes d'exactitude ;
- Le modèle PO est meilleure en termes d'exactitude contiguë.

Les autres études

Sur le français L1:

- Pour un problème à 5 classes : **R = 0,64**; Exac. et exac. cont. ne sont pas spécifiées.

  Collins-Thompson and Callan (2005)

Sur l'anglais L1:

- Pour un problème à 12 classes : R = 0,64 (grades 1-6) and **0,79** (grades 7-12); Exac. et exac. cont. ne sont pas spécifiées.

  Collins-Thompson and Callan (2005)

Sur l'anglais L2:

- Pour un problème à 12 classes : **R = 0,773** (PO) et 0,582 (MLR)

  exac. cont. = **52%** (PO) et 45% (MLR).

  Heilman et al. (2008)

Perspectives

- **3 voies principales de recherches :**
  - **Autres modèles statistiques**
    - Les machines à vecteurs de support (SVM)
    - Réseaux de neurones multicouches
  - **Recherche de nouvelles variables**
    - Variables syntaxiques : recherche de structures particulières, fréquence des structures...
    - Variables sémantiques : mots abstraits/concrets
    - Variables du lecteur : pourcentage de « cognates ».
  - **Réduction du bruit dans le corpus.**
    - Régression logistique régularisée ??
    - Entraînement sur base d'un corpus validé par des professeurs ou des apprenants.

Conclusions

- Il s'agit de la première formule spécifique au FLE qui utilise une approche TAL.
  - Le corpus comprend une grande variété de types de textes, ce qui garantit une couverture plus large à la formule.
- Nos expériences semblent montrer la supériorité de la régression logistique.
  - La RLM est supérieure pour l'exactitude ; le modèle PO pour l'exactitude contiguë.

  (divergence avec Heilman et al., 2008).
- Nos expérimentations ne permettent pas de décider quelle est la meilleure unité pour les modèles de langue appliqués à la lisibilité.

Bibliographie (1)

Breiman, L. and Friedman, J. H. and Olsen, R. and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman and Hall : New York.

Breiman, L. (1996). « Bagging predictors » in *Machine Learning*, 24(2), 123-140.

Brysbaert, M. and Lange, M. and Wijnendaele, I. V. (2000). « The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition: Further evidence from the Dutch language » in *European Journal of Cognitive Psychology*, 12, 65-85.

Chanier T., Selva T. (2000). « Génération automatique d'activités lexicales dans le système ALEXIA » in *Sciences et Techniques Educatives (STE)*. Editions Hermès, Paris : vol 7, 2, 385-412.

Collins-Thompson, K. and Callan, J. (2005.) « Predicting reading difficulty with statistical reading models » in *Journal of the American Society for Information Science and Technology*, 56(13), 1448-1462.

Collins-Thompson, K. and J. Callan. (2004). « Information retrieval for language tutoring: an overview of the REAP project (poster description) » in *Proceedings of SIGIR 2004*, Sheffield, UK.

Bibliographie (2)

Dale, E. and J. S. Chall. (1948). « A formula for predicting readability. » in *Educational research bulletin* Jan. 21 and Feb. 17, 27:1-20, 37-54.

De Landsheere, G. (1963). « Pour une application des tests de lisibilité de Flesch à la langue française » in *Le travail humain*, vol. XXVI, nos 1-2, 141-154.

Flesch, R. (1948). « A new readability yardstick » in *Journal of Applied Psychology*, Vol. 32, 221-233.

François T. (2009b, to be published). « Modèles statistiques pour l'estimation automatique de la difficulté de textes de FLE » in *Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2009)*, Senlis, 24-26/06/2009.

Francois, T. (2009a). « Combining a Statistical Language Model with Logistic Regression to Predict the Lexical and Syntactic Difficulty of Texts for FFL » in *Proceedings of the EACL 2009 Student Research Workshop*, Athens, 19-27.

Freund, Y. and Schapire, R. (1996) « Experiments with a New Boosting Algorithm. » in *International Conference on Machine Learning*, 148-156.

Bibliographie (3)

Gale, W. and Sampson, G. (1995). « Good-Turing frequency estimation without tears » in *J. of Quant. Linguistics*, v. 2, 217-237.

Gougenheim, G. and Michéa, R. and Rivenc, P. and Sauvageot, A. (1964). *L'élaboration du français fondamental (1er degré): étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris: Didier.

Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008). « An Analysis of Statistical Models and Features for Reading Difficulty Prediction » in *Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications*. Columbus, Ohio, 71-79.

Howes, Davis H. & Solomon, R. L. (1951). « Visual duration threshold as a function of word-probability » in *Journal of Experimental Psychology*. Jun. Vol 41(6), 401-410.

Lively, B. A. and Pressey, S. L. (1923). « A method of measuring vocabulary burden of textbooks » in *Educational Administration and Supervision*, 9, 389–398.

Kandel, L. et A. Moles (1958). « Application de l'indice de Flesch à la langue française » in *Cahiers Études de Radio-Télévision*, vol. 19, 253-274.

Bibliographie (4)

Kincaid, J.P. and Fishburne, R.P. and Rodgers, R.L. and Chisson, B.S. (1975). « Derivation of new readability formulas for Navy enlisted personnel » in *Research Branch Report 8-75*, U.S. Naval Air Station, Memphis.

Kintsch, W. and Vipond, D. (1979). « Reading comprehension and readability in educational practice and psychological theory » in L. G. Nilsson (Ed.), *Perspectives on memory research*. Hillsdale, NJ: Erlbaum. 329-365.

Kintsch, W., & Vipond, D. (1977). « Reading comprehension and readability in educational practice ». Paper presented at the Conference on Memory, June 1977, University of Uppsala (see the published version, 1979).

Miltsakaki, Eleni (2009). « Matching Readers' Preferences and Reading Skills with Appropriate Web Texts » in *Proceedings of the EACL 2009 Demonstrations Session*, Athens, 49-52.

New, B. and Pallier, C. and Brysbaert, M. and Ferrand, L. (2004). « Lexique 2: A new French lexical database » in *Behavior Research Methods, Instruments, & Computers*, 36, 516–524.

Bibliographie (5)

Selva T. (2002). « Génération Automatique d'Exercices Contextuels de Vocabulaire » in *Proceedings of TALN 2002*, 185-194.

Vogel M. and Washburne, C., (1928). « An odjective Method of Determining Grade Placement of Children's Reading Material » in *Elementary School Journal*, 28 : 373-81.