Des outils informatiques au service du passage à l'écrit d'apprenants.

Isabelle Audras*, Jean-Gabriel Ganascia**

** LIP 6*
8 rue du Capitaine Scott
75015 Paris
Isabelle.Audras@lip6.fr
*** LIP 6*
8 rue du Capitaine Scott
75015 Paris
Jean-Gabriel.Ganascia@lip6.fr

RÉSUMÉ :

L'analyse textuelle et le Traitement Automatique des Langues peuvent tirer parti de nouvelles techniques d'apprentissage symbolique et de fouille de données. De nombreux travaux sur les analyseurs robustes (« Robust parsing ») en attestent. Réalisé au LIP 6, par Jean-Gabriel Ganascia, le Littératron extrait automatiquement des motifs syntaxiques récurrents dans un texte ; associé à un analyseur de textes en arbres, il révèle les singularités stylistiques d'un auteur ou d'un genre.

Nous allons voir qu'utilisé en sciences du langage, dans le domaine de l'acquisition en langue étrangère du français écrit, le Littératron effectue un diagnostic cognitif de l'apprenant, qu'il s'agisse d'une classe de langue hétérogène (avec différentes langues maternelles) ou homogène (une seule langue maternelle, en l'occurrence ici l'arabe) ; l'intérêt de cette approche concerne trois domaines : d'une part la didactique des langues, à titre éducatif ; d'autre part, la linguistique computationnelle, et enfin l'enseignement assisté par ordinateur.

MOTS-CLÉS : Apprentissage d'une langue étrangère écrit, TALN, stylistique, extraction de motifs récurrents, diagnostic cognitif.
KEY-WORDS : foreign-language acquisition, TALN, stylistic, extraction of recurring patterns, cognitive diagnosis

1. Introduction

L'approche de la didactique des langues étrangères pourrait être considérablement transformée par l'emploi des techniques du traitement automatique des langues. Ainsi, ce que nous proposons n'est pas de supprimer l'enseignement 'académique' des langues étrangères, mais au contraire de le faciliter en tirant parti des connaissances acquises grâce aux outils de traitement automatique des langues.

En d'autres termes, nous souhaitons repérer, grâce aux techniques actuelles du traitement automatique des langues, les erreurs usuelles, caractéristiques d'une population d'apprenants, ce qui permettra de mettre l'accent, au cours de l'enseignement, sur la correction de ces erreurs. Ce repérage des erreurs se fait ici relativement aux usages, par une étude des tournures propres à une catégorie d'apprenants, et qui se trouvent absentes ou peu usitées chez les locuteurs natifs.

Des études empiriques conduites autour de deux populations d'apprenants du français langue étrangère, l'une à Paris, à l'Alliance Française, l'autre au département de français de l'université de Naplouse, auprès d'un public arabophone, valident l'approche proposée.

2. Présentation des outils informatiques utilisés

Deux outils informatiques sont nécessaires pour extraire les motifs syntaxiques caractéristiques de différentes populations. Le premier est un analyseur morphosyntaxique du français qui construit un arbre syntaxique à partir de productions écrites[1]. A chaque mot ou groupe de mots l'analyseur textuel associe une étiquette. L'arbre résultant est alors un arbre stratifié ordonné (ASO). Plus exactement, un arbre est dit stratifié si les étiquettes des nœuds sont partitionnées en classes de telle sorte que l'attribution d'une classe à un nœud dépende de la profondeur de ce nœud dans l'arbre d'analyse. Et, un arbre est dit stratifié ordonné (ASO en abrégé) si c'est un arbre stratifié dans lequel l'ordre des fils est pris en considération. Étant donnée une structure d'ASO, le Littératron calcule une mesure de similarité entre plusieurs ASO, fondée sur la notion de distance d'édition, et génère un graphe de similarité enregistrant les sous-arbres les plus proches de l'ASO en entrée.

C'est ce graphe de similarité qui sert ensuite d'entrée à l'algorithme de classification du Littératron, appelé 'centre-étoiles', qui construit des classes de motifs similaires et leur attribue un nom significatif.

---

[1] Dans les premières expériences, nous avons eu recours à l'analyseur linéaire avec dictionnaire partiel Vergne qui a été élaboré par Jacques Vergne de l'Université de Caen, en 1998 ; dernièrement, nous avons utilisé l'analyseur Cordial

Voici trois exemples de motifs syntaxiques extraits sur un texte et associés à l'étoile dont le centre est : [PREP ['de']] + [GN [ART ['la']] + [NOM ['forêt']]] (texte : 'de la forêt')

1. [PREP ['à']] + [GN [ART ['l'']] + [NOM ['auberge']]] (texte : 'à l\'auberge')

2. [PREP ['d\'']] + [GN [ART ['un']] + [NOM ['hiver']]] (texte : 'd\'un hiver')

3. [PREP ['dans']] + [GN [ART ['le']] + [NOM ['monde']]] (texte : 'dans le monde')

Outre la construction d'étoiles et l'extraction de motifs, le Littératron procède à un second type d'opérations qui consistent à comparer les étoiles issues de plusieurs textes afin de repérer les étoiles présentes dans l'un et absentes de l'autre. Ceci permet de discriminer, parmi les motifs présents dans une production, ceux qui le distinguent d'autres productions. C'est à partir de ce type de discrimination que l'on construira les tournures caractéristiques de populations d'apprenants.

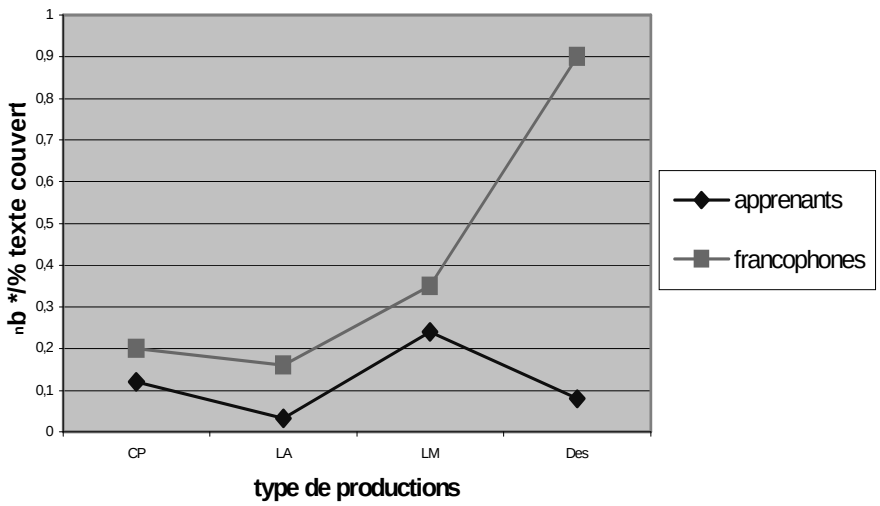


Figure 4. *Indice de variabilité en fonction du type de productions*

3. Premier type d'expérience : les apprenants sont de langue maternelle diverse

3.1. Présentation de l'expérience

Cette première expérience propose de comparer des productions d'apprenants et de francophones natifs (de niveau d'étude bac+4), à consigne égale : carte postale

(CP) pour le niveau débutant, lettre amicale (LA) pour le niveau intermédiaire, lettre de motivation (LM) pour le niveau avancé et description (Des)[2], pour tous.

3.2. Résultat

Pour apprécier la richesse stylistique, nous avons eu recours à un indicateur qui fait état de la diversité des motifs récurrents. Plus exactement, l'indice de variabilité correspond, pour chaque type de texte, au rapport du nombre d'étoiles détectées sur le pourcentage de texte couvert par l'application. C'est ce qui a conduit au tableau ci-dessus.

3.3. Commentaires

Nous remarquons que, pour un même genre de production écrite, les motifs syntaxiques retenus par l'application sont plus nombreux, divers et dans une proportion de texte plus petite chez les francophones que chez les apprenants ; de plus, la partie de texte non recouvert par les motifs syntaxiques récurrents – qui définit l'originalité de l'écrivant - varie dans un rapport 2 (pour les CP, LM et LA) à 9 (pour la Des) fois plus important chez les francophones que chez les apprenants, même les plus avancés.

Cette analyse a mis au jour des automatismes de l'écrit à l'intérieur de certains types de production. Ces automatismes concernent aussi bien des textes d'apprenants que des textes de natifs français. Pour ce qui concerne les descriptions, la comparaison entre les différents niveaux fait apparaître des fréquences de motifs qui évoluent vers une complexification dans la composition et les liens de dépendance, donc une aisance d'écriture qui s'installe au fur et à mesure que la grammaire s'acquiert.

4. Deuxième type d'expérience : les apprenants sont de même langue maternelle, l'arabe.

4.1. Présentation de l'expérience

L'apprentissage du FLE est sanctionné par une certification appelée DELF (Diplôme d'Etudes en Langue Française), dont les épreuves écrites A1, A2, A3 ont pour consigne, respectivement : la carte postale, la lettre amicale, la lettre de motivation. L'objectif de cette expérience est de détecter les compétences en manque lors du passage à l'écrit pour chaque unité de DELF ; suite au repérage de ces lacunes, un programme de re-médiation personnalisée, qui pourrait être contenu dans un EIAH[3], peut être proposé à l'apprenant..

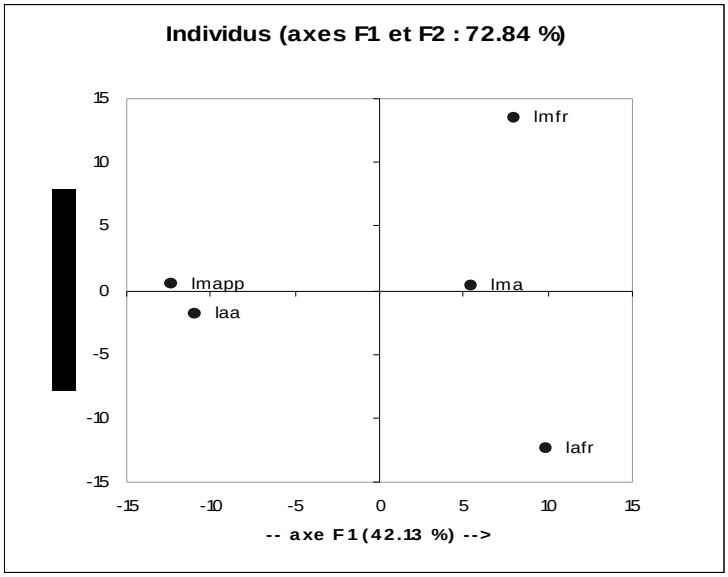
---

[2] Chaque apprenant, tout niveau confondu, est soumis à l'observation puis la description d'un même dessin en couleurs de format A3 (place de village, art naïf).

[3] En effet, le Littératron pourrait modéliser l'apprenant grâce à l'étude des figures

Deux types de productions ont été choisis : la lettre de motivation et la lettre amicale ; ces deux productions ont été tirées d'épreuves de DELF session 2004, le DELF scolaire pour les lettres amicales et l'unité A3 pour les lettres de motivation. De l'autre côté, des productions de même consigne ont été recueillies auprès de francophones (natifs, de niveau d'étude équivalent au moins à bac + 2). Suivant le même procédé que précédemment, les productions d'apprenants et de francophones de même consigne sont introduitse dans l'analyseur textuel puis le Littératron.

4.2. Résultat



Chaque zone délimitée en bleu nous donne les motifs syntaxiques spécifiques pour chaque type de production. A partir des motifs extraits caractéristiques de chaque zone, il est ensuite aisé de construire un tableau récapitulatif des compétences en absence / en présence et de les spécifier selon le niveau attendu pour chaque unité de DELF.

4.5. Commentaires

---

qu'il recense au moyen des patrons récurrents. Le logiciel relève dans les productions écrites un ensemble de figures caractéristiques relatives par exemple à l'expansion du nom, à l'emploi des adjectifs, des adverbes, des ponctuations etc. Puis, à partir de ces figures, il analyse les erreurs ou les emplois excessifs de certaines tournures. Ensuite, ces données sont transmises au module d'inférence qui s'appuie sur une base de règles pour déterminer le profil de l'apprenant.

Sur le graphe issu de la décomposition en composantes principales, nous avons ajouté à notre corpus les lettres de motivation d'apprenants de la première expérience (lm app). Lm a et lm app ne couvrent pas la même zone du graphe, même elles suivent le même axe F1 : il s'agit ici sans doute d'une spécificité des lettres arabophones dont nous pouvons retrouver les motifs syntaxiques. De même les productions francophones semblent suivre une ordonnée commune, mais chaque type de production francophone a l'air bien éloigné de son pendant arabophone ou autre apprenant (lm fr et lm a ; la a et la fr).

Les productions arabophones présentées dans cette expérience s'avèrent nettement trop faibles pour un niveau Delf scolaire 1 (lettre amicale) ou A3 (lettre de motivation).

5. Conclusion

Utilisé en sciences du langage, dans le domaine de l'acquisition en langue étrangère du français, le Littératron est en mesure de déterminer le diagnostic cognitif au moment du passage à l'écrit de l'apprenant.

A terme, ce travail doit faire l'objet de deux types de développements complémentaires, aux plans technique et expérimental. D'un côté, nous allons faire appel à une décomposition plus riche que la décomposition en syntagmes qui prendra en compte la structure propositionnelle. L'arbre résultant de cette analyse doit être considérablement enrichi. D'un autre côté, les résultats obtenus auprès d'étudiants arabophones nous encouragent à poursuivre plus loin l'étude des différences spécifiques, auprès d'apprenants venant de différentes régions du monde, et dont la langue première varie.

6. Bibliographie

[GANASCIA02] Ganascia J-G, "Extraction of Syntactical Patterns from Parsing Trees", *Internationale Conference on Textual Data Statistical Analysis, 13-15 mars 2002*, 2002.

[GANASCIA04] Ganascia J-G, "Detection of Statistically Abnormal Patterns from Stratified Ordered Trees" *Advances in the Internet Technology, Concepts and Systems*, ouvrage publié sous la direction de Veljko Milutinovic et Ivana Vujovic, 2004.

[LABBE02] Labbe C., Labbe D., Hubert P., " Segmentation automatique des corpus : Voyage de l'autre côté, de JMG Le Clézio", *Actes des JADT 2002, Communication aux VIe Journées d'Analyse de Données Textuelles*, 2002.

[TUFFS93] Tuffs R., "A genre approach to writing in the second language classroom : the use of direct mail letters", *Revue belge de philology et d'histoire,* Vol. 71, n°3, 1993.

[VERGNE99] Vergne, J., Analyseur linéaire avec dictionnaire partiel, convention d'utilisation de l'analyseur J. Vergne, 1999.