Analyses comparatives de motifs syntaxiques de francophones et d'apprenants du français arabophones, à l'aide d'outils d'extraction automatique du langage.

Isabelle Audras, Jean-Gabriel Ganascia

Analyses comparatives de motifs syntaxiques de francophones et d'apprenants du français arabophones, à l'aide d'outils d'extraction automatique du langage.

Isabelle Audras
Jean-Gabriel Ganascia

LIP 6 - Université Pierre et Marie Curie

Isabelle.Audras@lip6.fr
Jean-Gabriel.Ganascia@lip6.fr

Résumé – Abstract

De nouveaux logiciels d'analyse textuelle s'avèrent être des outils pertinents dans les domaines de recherches tels que ceux de l'apprentissage symbolique et du Traitement Automatique des Langues Naturelles. Le Littératron[1] est un nouvel outil informatique d'extraction automatique de motifs syntaxiques, réalisé au LIP 6, par Jean-Gabriel Ganascia. Associé à l'analyseur de textes linéaire de Jacques Vergne[2], il révèle les singularités stylistiques d'un texte.

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.

New software of textual analysis prove to be relevant tools in the fields of research such as those of the training symbolic system and the Automatic Treatment of the Natural Languages. Littératron is a new data-processing tool for automatic extraction of syntactic patterns, produced in the LIP 6, by Jean-Gabriel Ganascia[1]. Associated the linear analyzer of texts of Jacques Vergne[2], it reveals the singularities stylistics of a text.

We will see that used in sciences of the language, in the field of acquisition in foreign language of written French, Littératron carries out a cognitive diagnosis of learning, that it acts of a class of heterogeneous language (with learners of various language level and various mother tongues) or homogeneous (only one language level and one mother tongue, in fact here Arabic) ; the interest of this approach relates to three fields: on the one hand the didactic one of the languages, on a purely educational basis; in addition, computational linguistics, and finally computer-assisted learning.

[1] GANASCIA, J-G, 2001
[2] VERGNE, J., 1999
[1] GANASCIA, J-G, 2001
[2] VERGNE, J., 1999

Keywords – Mots-clefs :

Mots-clés : acquisition d'une langue étrangère, TALN, stylistique, extraction de motifs récurrents

Keywords : foreign-language acquisition, TALN, stylistic, extraction of recurring patterns

1 Introduction

L'approche de la didactique des langues étrangères pourrait être considérablement transformée par l'emploi des techniques du traitement automatique des langues. L'assertion peut paraître étrange à première vue, puisque le traitement automatique des langues vise, entre autres, à supprimer les barrières linguistiques grâce à l'emploi des ordinateurs, et donc à rendre moins nécessaire l'apprentissage des langues étrangères. Dans ce contexte, la didactique des langues étrangères ne serait pas transformée ; elle disparaîtrait, tout simplement… On peut toutefois voir les choses d'une autre façon et c'est ce que nous faisons ici. Ainsi, dans la perspective qui est la nôtre, il ne s'agit pas de supprimer l'enseignement 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 peut se faire de deux façons, soit dans l'absolu, par détection des fautes syntaxiques, soit relativement aux usages, par une études des tournures propres à une catégorie d'apprenants, et qui se trouvent absentes ou peu usitées chez les locuteurs natifs. C'est cette seconde approche que nous adopterons ici, sachant que le rôle des enseignants de langue n'est pas d'apprendre une langue abstraite parfaite, mais de transmettre les usages d'une langue.

Plus exactement, le travail présenté ici recourt à l'emploi d'outils d'analyse stylistique pour dégager les caractéristiques des apprenants, selon leur niveau, et les distinguer des locuteurs natifs. Des études empiriques conduites autour de deux populations d'apprenants, l'une à Paris, à l'Alliance Française, l'autre à l'université de Naplouse, auprès d'un public arabophone, valident l'approche proposée.

2 Les outils informatiques

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. Nous avons eu recourt à l'analyseur linéaire avec dictionnaire partiel Vergne qui a été élaboré par Jacques Vergne de l'Université de Caen, en 1998. Cet analyseur découpe un texte en langage naturel en syntagmes non récursifs. Les sorties sont ensuite transformées en arbres stratifiés ordonnés (ASO) pour servir d'entrée à un outil d'analyse stylistique, le Littératron (Ganascia 2001), qui dégage les motifs récurrents présents dans des arbres.

Plus exactement, à chaque mot ou groupe de mots l'analyseur de Vergnes associe une étiquette ; un arbre stratifié est donc une partition d'étiquettes dont les classes dépendent de la

profondeur du nœud dans l'arbre d'analyse. Etant 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. En effet, l'algorithme centre-étoile évalue d'abord l'ensemble des étoiles centrées sur les différents nœuds puis il prend, pour chacune, la somme des valeurs de similarité des nœuds de chaque étoile au centre. Une fois calculée la valeur de chaque étoile, l'algorithme 'centre-étoiles' prend celle qui à la plus forte évaluation. On marque ensuite, les nœuds qui appartiennent à cette première étoile, avant d'appliquer récursivement le même algorithme sur les nœuds non marqués, jusqu'à épuisement des nœuds non marqués.

En résumé, toute étoile est un sous-graphe du graphe de similarité centré sur un nœud. Pour chaque classe ainsi construite, l'algorithme choisit les motifs les plus similaires au centre de l'étoile, pour illustrer la signification de l'étoile. Il indique aussi le texte source couvert par chacun des motifs.

Voici cinq exemples de motifs syntaxiques extraits 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')
4. [PREP ['avec']] + [GN [ART ['les']] + [NOM ['chiens']]] (texte : 'avec les chiens')
5. [PREP ['depuis']] + [GN [ADJ ['quelques']] + [NOM ['jours']]] (texte : 'depuis quelques jours')

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.

3 Premier type d'expérience : analyses de productions écrites issues de classes de langue hétérogènes (apprenants de différents niveaux d'apprentissage et de diverse langue maternelle).

L'idée de cette recherche est de recueillir des productions écrites en classe de langue d'apprenants du français de différents niveaux et d'étudier les sorties des analyseurs textuels présentés ci-dessus, en les comparant avec celles de textes de francophones (natifs français bac+4), répondant aux mêmes consignes.

3.1 Problématique : l'écrit en classe de langue

La façon d'écrire en langue seconde est pragmatique. Elle est le reflet des compétences en présence lors du passage à l'écrit de l'apprenant, qui se révèlent à la fois dans les fréquences d'expression observées et mais aussi dans ses 'prises de risques' et l'originalité

des idées. Dans un article[3], R. Tuffs appuie sur le fait que travailler sur des genres textuels facilite l'acquisition de la langue étrangère. De même, à l'extérieur d'un genre, l'écrit en classe de langue est toujours associé à une consigne qui prévoit l'intention de communication. En effet, le cadre narratif choisi, par le genre ou la consigne, définit un objectif de communication précis, celui-ci appelle des objectifs fonctionnels dont l'expression morphosyntaxique et lexicale est vue en classe. Ce contenu linguistique, découvert à l'intérieur d'une situation de communication, est automatisé lors de réemploi d'autant plus si celui-ci utilise un contexte similaire. Enfin, l'analyse des besoins communicatifs du cadre narratif aide l'apprenant à s'adapter face à une nouvelle situation de communication dans laquelle il doit réagir.

L'acquisition du français langue étrangère est observable, à l'écrit, par la comparaison de la nature des motifs syntaxiques extraits et de leur fréquence, comparaison entre productions d'apprenants et de francophones. Les outils informatiques aident, en sciences cognitives, à révéler un 'style' en langue seconde.

3.2 Présentation des productions écrites :

Quatre types de production ont été choisis : la carte postale (CP), la lettre amicale (LA), la lettre de motivation (LM), la description (Des). Chaque production correspond à un niveau d'apprentissage du français langue étrangère.

Derrière le terme général de description, se regroupent dans le corpus ici rassemblé:
- pour le degré 1 :
des descriptions d'objets (consigne : jeu de 'qu'est-ce que c'est ?', faire deviner un objet en en disant le moins possible),
des présentations de lieux (exemple de consigne : votre ville préférée en France : la décrire et dire pourquoi)
- pour le degré 2 :
des critiques de film dont toute une partie consiste à le décrire,
d'autres présentations de lieux (exemple de consigne : décrire votre pays en l'an 2020)
- pour le degré 3 :
des descriptions de tableaux ou photos en y ajoutant un jugement de goût
encore d'autres encore descriptions de lieux (consigne : présenter un lieu historique de votre pays)

Toutes les productions d'apprenants ont été faites en classe, entre le mois d'avril et le mois de juin 2002, pour la plupart à l'Alliance Française, pour certaines descriptions dans une classe FLE et alphabétisation d'un foyer de travailleurs, le Foyer Pinel, à Saint Denis.

[3] TUFFS R., 1993

	Apprenants			
	Débutants	intermédiaires (125 h de français)	avancés (250 h de français)	Francophones
carte postale	X			X
lettre amicale		X		X
lettre de motivation			X	X
Description	X	X	X	X

Figure 1 : Tableau récapitulatif des productions

	Apprenants			
	débutants	intermédiaires (125 h de français)	avancés (250 h de français)	Francophones
carte postale	6			6
lettre amicale		4		4
lettre de motivation			6	6
Description	5	5	5	5

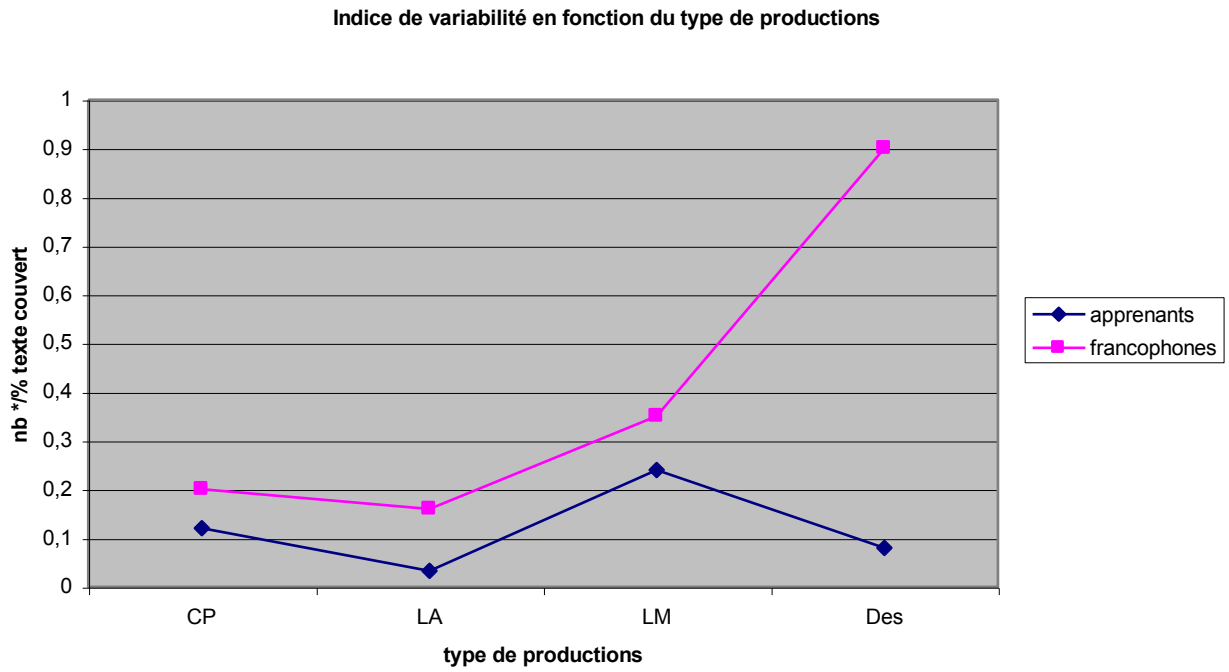
Figure 2 : Nombre des productions

3.3 Présentation des résultats

	6 CP déb.	6 CP frcph	4 LA interm.	4 LA frcph	6 LM av.	6 LM frcph.
nb *	6	10	2	5	6	6
% texte recouvert	50	50	60	30	25	17

	5 Des deb.	5 Des inter.	5 Des. Av.	5 Des francph
nb *	2	3	3	13
% texte recouvert	33	33	35	14

Figure 3 : Nombre d'étoiles et pourcentage de texte couvert par celles-ci



3.4 Commentaires :

Sur le graphe, l'indice de variabilité est, pour chaque type de texte, le rapport du nombre d'étoiles détectés sur le pourcentage de texte couvert par l'application.

Les nombre de motifs syntaxiques récurrents et le pourcentage de texte recouvert par ces motifs, ne laissent aucun doute : 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 bac+. Il y a donc une façon d'écrire les cartes postales, les lettres d'invitation ou les lettres de motivation. 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. Il est à noter que le motif de base semble être pour le syntagme nominal : préposition+substantif+adjectif qualificatif et le syntagme verbal : pronom sujet+verbe+adverbe, et ce motif de base d'enrichit progressivement en fonction de la maîtrise du français.

4 Deuxième type d'expérience : analyses de productions écrites d'apprenants issus d'une classe homogène (de même niveau d'apprentissage et de même langue maternelle, en l'occurrence ici l'arabe)

4.1 Présentation de l'expérience

L'idée est de montrer une spécificité de la syntaxe française au sein d'une classe homogène.

L'ensemble des productions analysées correspond aux examens d'histoire et de civilisation du 1er semestre 2003 de 3ème année du département de français de l'Université An-Najah de Naplouse (Cisjordanie). Les étudiants de l'université sont tous de langue maternelle arabe, l'anglais est leur première langue étrangère, le français leur deuxième. L'apprentissage du français a commencé pour la plupart à l'université, sauf pour un petit nombre d'étudiants issus du collège et du lycée privé, l'enseignement du français en primaire et secondaire dans les écoles publiques étant une initiative récente.

Les mêmes étudiants ont écrit les deux examens.
Il s'agit de texte de type argumentatif et descriptif. Les questions sont : « Vous gagnez un voyage de 15 jours en France, choisissez un itinéraire et expliquez votre choix .», quant au corpus de civilisation et : « A partir de ce qui a été vu en cours, et de ce que vous lu et appris dans vos recherches personnelles, écrivez un petit texte sur la première guerre mondiale » pour celui d'histoire. Il s'agit à chaque fois de productions d'une quinzaine de lignes.
Les deux textes sont introduits dans le Littératron, pour être comparer avec des textes de francophones.

4.2 Présentation des résultats

Trois mêmes motifs syntaxiques ressortent systématiquement des productions arabophones. Ces motifs recouvrent ¼ du texte analysé. Il s'agit de deux motifs nominaux et d'un verbal. Les deux motifs nominaux sont de construction : DE + adjectif + nom, comme dans les exemples : 'de choses magnifiques' et 'd'autres villages'.

4.3 Commentaires

Ce motif syntaxique révèle une utilisation massive de groupes nominaux de forme adjectif + nom commençant par DE, au détriment d'autres articles et d'autres prépositions. L'hypothèse qui est que l'étudiant, à défaut de connaître la bonne rection d'un verbe, directe ou indirecte, ou le bon emploi de l'article partitif sur le défini ou l'indéfini, va utiliser systématiquement la préposition DE pour introduire ses compléments d'objet, s'avère vérifiée en relisant les phrases correspondantes aux motifs récurrents extraits.

L'apprenant ne semble pas maîtriser une bonne utilisation des prépositions et des articles. Cette expérience révèle pour une classe homogène d'apprenants arabophones, une spécificité de la syntaxe française dont l'acquisition nécessite un accompagnement particulier. La rection

des verbes en arabe est conditionnée par un placement spécifié de ces mots. Plus précisément, le "recteur" est un trait qui, par définition, fait partie du lexique des items de la langue[4].

5 Conclusion :

Utilisé en sciences du langage, dans le domaine de l'acquisition en langue étrangère du français écrit, le Littératron est en mesure de déterminer le diagnostic cognitif de l'apprenant ; en effet, les invariants et diversités syntaxiques extraits témoignent des compétences en présence lors du passage à l'écrit de l'apprenant.

Ainsi, ces applications révèlent clairement :

- dans la première expérience :

- des automatismes morphosyntaxiques propres à un genre textuel, que l'écrivant soit francophone ou natif.
- des invariants dans la composition des syntagmes nominaux et verbaux chez les apprenants et les francophones
- une évolution vers la complexification et la diversification dans la composition de ces syntagmes, de l'apprenant débutant au francophone.

- dans la deuxième expérience :

- des compétences en manque lors du passage à l'écrit, démasquées
- des reflets de l'apprentissage de la langue-cible et des effets imputables à la langue-source ; donc l'utilisation du Littératron est d'autant plus intéressante que la distance linguistique entre la langue-cible et la langue-source est grande, comme c'est notamment le cas entre le français et l'arabe.

A terme, ce travail doit faire l'objet de deux types de développements complémentaires, aux plans technique et expérimental.

D'un côté, nous nous sommes limités ici à une décomposition en syntagmes, et à une étude de la structure de la phrase relativement à cette décomposition. Cela restreint assez fortement le type de motifs détectés. Nous allons faire appel à une décomposition plus riche qui prendra en compte la structure propositionnelle. L'algorithme d'extraction de motifs est identique, mais l'analyse syntaxique diffère et, surtout, l'arbre résultant de cette analyse doit être considérablement enrichi.

D'un autre côté, les résultats obtenus auprès d'étudiants arabophones nous encourage à poursuivre plus loin l'étude des différences spécifiques, auprès d'apprenants venant de différentes régions du monde, et dont la langue première varie. Le but, à terme, serait de caractériser les modes d'influence d'une langue maternelle, sur l'apprentissage d'autres langues.

Références

Carroll M., Stutterheim Ch. Von, (1997), Relations entre grammaticalisation et conceptualisation et implications sur l'acquisition d'une langue étrangère*, AILE*, Vol. 9, pp. 14-19.

Ganascia, J-G, (2001), Extraction automatique de motifs syntaxiques*,* Actes de *TALN 2001.*

[4] Mejri S., 2000

Gaonac'h D., (1987), *Théories de l'apprentissage et acquisition d'une langue étrangère,* Paris, Hatier.

Giguet, E. (1998), *Méthode pour l'analyse automatique de structures formelles sur documents multilingues*, Thèse de doctorat en informatique, Université de Caen.

Labbe C., Labbe D., Hubert P. (2002) 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*.

Morais J. et Kolinsky R., (2000), The literate mind and the universal human mind, *Langage, Brain and cognitive development*.

Tagliante C., (1994), *La classe de langue,* Paris, CLE International.

Tuffs R., (1993), 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, p. 691-721

Vergne, J. (1999), *Analyseur linéaire avec dictionnaire partiel,* convention d'utilisation de l'analyseur J. Vergne.