# Apprentissage/didactique des langues étrangères et TALN : analyse de corpus écrits à l'aide d'outils d'extraction automatique du langage

Isabelle Audras, Jean-Gabriel Ganascia

# Apprentissage/didactique des langues étrangères et TALN : analyse de corpus écrits à l'aide d'outils d'extraction automatique du langage.

Isabelle Audras[1], Jean-Gabriel Ganascia[1]

[1] LIP6 – 8, rue du Capitaine Scott – 75015 Paris – France

## Abstract

New text analysis softwares issued from fields of research such as Machine Learning and Natural Languages Processing prove to be relevant tools for the language sciences. *Littératron* is a new data-processing tool for the automatic extraction of syntactic patterns, designed at LIP6 by Jean-Gabriel Ganascia. Associated with a linear text analyser, it reveals the stylistic peculiarities of a text.

We will see that *Littératron* carries out a linguistic diagnosis of learners if used in language sciences, especially in the field of acquisition of written French as a foreign language. The learner can be from a heterogeneous group (various language levels and various mother tongues) or from a homogeneous group (only one language level and one mother tongue, here, Arabic). The interest of this approach is related to three fields: first, language didactics, on a purely educational basis; next, computational linguistics; finally, computer-assisted learning.

## Résumé

De nouveaux logiciels d'analyse textuelle tirent partie des progrès récents effectués en apprentissage symbolique et dans le traitement automatique des langues naturelles. Conçu au LIP6 par Jean-Gabriel Ganascia, le *Littératron* est l'un d'entre eux ; il extrait automatiquement des motifs syntaxiques[i] à partir de textes écrits en langage naturel. Plus exactement, le *Littératron* prend comme entré un arbre d'analyse syntaxique et donne en sortie un certain nombre de motifs syntaxiques récurrents. Associé à un analyseur de textes, qui engendre l'arbre d'analyse syntaxique à partir de textes écrits en langage naturel, il révèle les singularités stylistiques de ces textes.

Nous allons voir qu'utilisé en sciences du langage, dans le domaine de l'acquisition du français écrit, le *Littératron* permet d'effectuer un diagnostic linguistique de l'apprenant, que celui-ci provienne d'une classe de langue hétérogène (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 :** acquisition d'une langue étrangère écrit, didactique de l'écrit en langue étrangère, TALN, extraction de motifs récurrents, stylistique, diagnostic linguistique

**Keywords** : foreign-language acquisition, foreign-language written didactic, NLP, stylistics, extraction of recurrent patterns, linguistic diagnosis

[i] Un motif syntaxique est une association d'unités linguistiques cohérentes

# 1. Introduction

Cette communication croise les domaines de la didactique des langues étrangères et de la stylistique. Les recherches sur l'apprentissage d'une langue étrangère à partir de supports informatisés, suscitent aujourd'hui de plus en plus d'engouement. Néanmoins, il apparaît qu'elles n'ont pas donné lieu aux prolongements didactiques auxquels l'on pouvait logiquement s'attendre, et notamment dans l'acquisition du français langue étrangère écrit.

Ce type d'apprentissage suscite un intérêt scientifique autour de la notion de style produit par l'apprenant. A partir d'un besoin de communication, le style se définit par l'ensemble des expressions linguistiques choisies parmi celles connues, pour y répondre. Travailler le style en classe de langue contribue à l'enrichissement de la grammaire et du vocabulaire en apportant les nuances nécessaires à une expression juste et variée.

Or, tout apprenant d'une langue étrangère commet des erreurs lors de son apprentissage, qui sont constructives et permettent une progression par leur compréhension et leur correction. De plus, les erreurs de style jouent un rôle d'indicateur du niveau de langue atteint pour le formateur, qui sur la base de ces erreurs (types d'erreurs, fréquences de l'erreur) établissent un programme de remédiation[ii]. C'est pourquoi nous nous proposons ici de repérer, grâce aux techniques actuelles du traitement automatique des langues, les erreurs usuelles dans le passage à l'écrit 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 étude dans un cadre narratif précis 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. Nous avons utilisé dans nos expériences le Littératron, mis au point au LIP6 par Jean-Gabriel Ganascia : il détecte des motifs syntaxiques récurrents, présents dans un texte et absents d'un autre. Un motif syntaxique est une unité syntaxique cohérente (groupe sujet, groupe verbal, groupe complément). Les analyses linguistiques proposées renseignent l'utilisateur du Littératron (formateur ou apprenant) sur la présence ou l'absence de tel motif syntaxique. De plus, lorsqu'elles sont confrontées avec une grille de points morpho-syntaxiques attendue pour un certain niveau d'apprentissage - comme c'est le cas pour chaque épreuve écrite de DELF - elles peuvent contribuer à évaluer le niveau de l'apprenant.

Des études empiriques conduites autour de trois populations d'apprenants, l'une à Paris, à l'Alliance Française, l'autre à l'université de Naplouse, auprès d'un public arabophone et la troisième à l'Ecole Normale de Port-au-Prince (Haïti) valident l'approche proposée.

# 2. Présentation des 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

[ii] Cf. Veltcheff & Hilton

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 [iii]. Les sorties sont ensuite transformées en arbres stratifiés ordonnés (ASO) pour servir d'entrée au Littératron.

### *2.1. Première étape : de l'analyse syntaxique du texte au graphe de similarité des sous-arbres*

Plus précisément, l'analyseur textuel associe une étiquette à chaque mot (nom, verbe, etc.) ou groupe de mots (syntagme nominal, syntagme verbal, syntagme prépositionnel, etc.) et les transforme en arbre. Un ASO est une partition d'étiquettes dont les classes dépendent de la profondeur du nœud dans l'arbre d'analyse. Par exemple, un niveau correspond à la phrase (analyse logique), un second à des syntagmes non récursifs, et un dernier à des lemmes[iv].

Cette étape d'analyse est importante car l'algorithme d'extraction du Littératron repose en grande partie sur ces structures ordonnées. En effet, étant donnée une structure d'ASO, le Littératron calcule une mesure de similarité entre plusieurs ASO, fondée sur la notion de distance d'édition. Le concept d'édition consiste en une opération qui transforme un caractère ou un nœud d'une chaîne par un autre. Il peut s'agir d'une opération d'insertion, de substitution ou de destruction d'un caractère ou nœud dans une chaîne. Une distance d'édition entre deux chaînes est le nombre d'opérations minimales nécessaires pour remplacer une chaîne par une autre. Pour étendre cette notion aux arbres, il est nécessaire d'avoir recours à des ASO (Ganascia, 2001). L'algorithme d'extraction de motifs, construit sur la base de distance d'édition, génère un graphe de similarité enregistrant les sous-arbres les plus proches de l'ASO en entrée.

### *2.2. Deuxième étape, l'algorithme "centre-étoiles"*

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.
Une étoile centrée sur un nœud N est un graphe dont toutes les arêtes contiennent le nœud N. 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 ce score associé à chaque étoile, l'algorithme "centre-étoiles" prend celle qui a la plus forte évaluation, c'est-à-dire celle qui correspond au motif le plus récurrent dans les textes étudiés. 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 n'avoir que des nœuds marqués (Ganascia, 2004).
En résumé, toute étoile est un sous-graphe du graphe de similarité qui est lui-même centré sur un nœud. Le centre d'une étoile correspond à un des motifs parmi ceux qui sont les plus récurrents dans les textes étudiés.

### *2.3. Etape de description*

L'étape finale de l'algorithme de classification consiste à décrire chaque classe induite. Une étoile induit une classe de nœuds. Le centre de l'étoile est représenté par un motif syntaxique

[iii] Un syntagme non récursif est un segment intermédiaire, un groupe d'unités syntaxiques intermédiaire entre le mot et la phrase

[iv] Un lemme est une unité constituante du lexique ou du mot

récurrent. Pour chaque classe construite, l'algorithme choisit les motifs les plus similaires au centre de l'étoile, pour illustrer la signification de l'étoile. Autrement dit, l'algorithme choisit le motif qui maximise la similarité avec les autres membres de la classe et qui minimise la similarité avec les membres des autres classes. Il donne également la partie extraite des textes sources représentée par chacun des motifs.

Voici l'exemple d'un centre d'étoile, illustré par la figure 1 :
[PREP ["de"]] + [GN [ART ["la"]] + [NOM ["forêt"]]] (texte : "de la forêt"), auquel sont associés les 5 motifs syntaxiques suivants :
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')
Ceci signifie que la mesure de similarité entre le premier motif ('de la forêt') et l'un des arbres dérivés des arbres syntaxiques de chacun de ces cinq groupes nominaux est supérieure à un certain seuil. Ces cinq motifs font partie de la même étoile dont le centre est de la forme : [PREP ["de"]] + [GN [ART ["la"]] + [NOM ["forêt"]]].

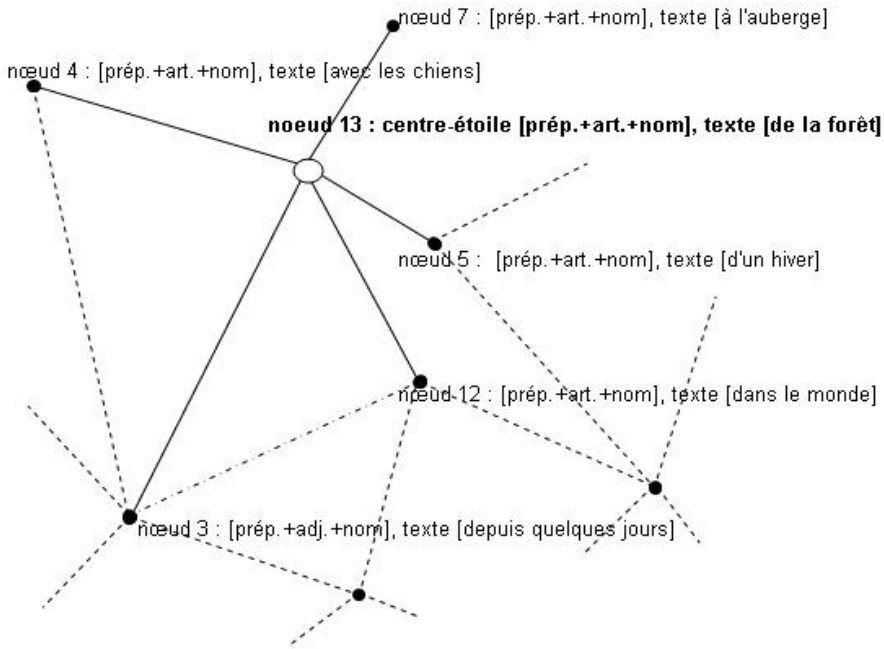


*Figure 1 – Graphe du centre-étoile présenté en exemple.*

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.

Par exemple, voici l'une des sorties du Littératron analysant trois groupes de scripteurs distincts à partir de lettres de motivation (LM app : lettres de motivation d'apprenants de langues maternelles diverses, LM appa : lettre de motivation d'apprenants arabophones, LM frcph : lettres de motivation de francophones).

Les sorties du Littératron se présentent sous l'intitulé de 'patron'. Le patron décrit la structure syntaxique du motif extrait. Ici il s'agit d'une proposition indépendante de forme sujet + verbe + COD. Il y est précisé que le sujet est un pronom personnel à la première personne du singulier et que le COD est un nom commun féminin singulier. Suivent trois exemples de cette structure, chacun extrait d'un des trois groupes de scripteurs présentés ci-dessus (apprenants de langue maternelle diverse, apprenants arabophones, francophones).

Patron N°46
[Indépendante] = [(sujet-pronom personnel 1ère personne singulier) + (verbe) + (COD-nom commun féminin singulier)]
Exemples:
Fichier LMapp: Je parle la langue anglaise et française
Fichier LMfrcph: Je maîtrise la mise en place de l'organisation de l'archivage
Fichier LMappra: J'apprends la presse à l'université de Naplouse

Les exemples extraits de chaque groupe de scripteurs donnent un aperçu des différents textes en langage naturel que le Littératron détecte comme étant proches de cette structure centrale.

## 3. Problématique : l'écrit en classe de langue

Toute production écrite laisse une trace du fonctionnement cognitif du scripteur apprenant, même dans des productions scolaires comme la rédaction ou la dictée (Besse, 2003). En effet, la production écrite en classe de langue est le reflet des compétences de l'apprenant lors du passage à l'écrit. Ses compétences se révèlent à la fois dans la fréquence des expressions observées, dans ses prises de risques et dans l'originalité de ses idées (Carroll, M. & Stutterheim Ch., 1997). Par ailleurs, selon Tuffs (Tuffs, 1993), travailler sur des genres textuels différents facilite l'acquisition des langues étrangères. De façon générale, l'écrit en classe de langue est associé à une consigne qui prévoit l'intention de communication, même à l'extérieur d'un genre. 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 réemplois, et ceci est d'autant plus vrai si celui-ci se trouve dans 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 (Tagliante, 1994).

Par ailleurs, et nous y reviendrons plus loin, l'apprentissage du FLE est sanctionné par une certification appelée DELF (Diplôme d'Etudes en Langue Française) aligné sur le cadre européen commun de référence dans l'apprentissage des langues. Les épreuves écrites A1, A2 et A3 ont pour cadre narratif, respectivement : la carte postale, la lettre amicale, la lettre de motivation. Les erreurs linguistiques et stylistiques détectées dans ces productions, au cadre narratif contraignant, sont autant de traces cognitives laissées par l'apprenant. Ainsi, le niveau de l'apprenant est validé par rapport à sa capacité à exprimer un message à travers un modèle appris et reconnu et non simplement par rapport à ses compétences grammaticales.

C'est pourquoi 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

effectuée entre productions d'apprenants et de francophones. Les outils informatiques se révèlent un outil précieux, en sciences cognitives, pour révéler un 'style' en langue seconde. Rappelons ici que la portée des analyses linguistiques qui découlent des sorties du Littératron s'arrête à renseigner l'utilisateur (formateur ou apprenant) sur la présence ou l'absence de tel motif, sur la construction syntaxique d'une unité linguistique. Le Littératron n'a donc aucune prétention à vérifier si telle production vérifie les objectifs communicationnels de la consigne. Il est vrai qu'il y a un lien entre style et motifs syntaxiques tels que les extrait le Littératron.

C'est pourquoi, voici d'abord quelques exemples d'expérimentations pertinentes qui se sont montrées efficaces sur des apprenants scripteurs :

- utiliser régulièrement les sorties du Littératron auprès d'apprenants en difficulté sur diverses productions permet de vérifier la qualité stylistique d'un même apprenant à plusieurs moments.
- pour connaître l'activité d'un apprenant à un moment donné (avant une certification DELF par exemple) : l'utilisation du Littératron est pertinente à condition de dresser une liste des points de morpho-syntaxe à vérifier (d'où la nécessité de productions aux consignes contraignantes c'est à dire ouvertement dirigées sur des compétences communicatives et morpho-syntaxiques). Ainsi, en analysant les sorties, on peut voir nettement quel scripteur apprenant a acquis tel point de morpho-syntaxe, lequel au contraire a des difficultés etc.

Ces expérimentations du quotidien sont autant d'exemple d'utilisabilité du Littératron dans le quotidien en classe de langue mais ne font pas l'objet des expériences présentées ici.

## 4. 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.

### *4.1. Présentation des productions écrites et méthodologie expérimentale*

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

Toutes les productions d'apprenants ont été faites en classe, entre le mois d'avril et le mois de juin 2002. La plupart se sont déroulées à l'Alliance Française de Paris. Certaines descriptions ont été réalisées dans une formation en FLE et en alphabétisation dans le Foyer de travailleurs Pinel, à Saint Denis.

Le tableau 1 a une double fonction. Premièrement, il récapitule les expérimentations réalisées par genre textuel. Par exemple : en ce qui concerne la 'carte postale' (CP), vont être introduits simultanément dans les analyseurs les productions d'apprenants débutants et de francophones. Deuxièmement, il détaille le nombre total de production de chaque type.

Concernant la description, les productions des 4 groupes de scripteurs sont introduites en même temps dans les analyseurs.

| | apprenants | | | |
|---|---|---|---|---|
| | débutants (niveau A1 du CECR[v]) | intermédiaires (A2 niveau du CECR) | avancés (A3 niveau du CECR) | francophones |
| carte postale (CP) | 6 | | | 6 |
| lettre amicale (LA) | | 4 | | 4 |
| lettre de motiv. (LM) | | | 6 | 6 |
| Description (Des) | 5 | 5 | 5 | 5 |

*Tableau 1 : Tableau récapitulatif des productions et leur nombre.*

### *4.2.Résultats et commentaires*

Les résultats obtenus sont de nature statistique, auxquels nous ajoutons des commentaires linguistiques sur les motifs extraits.

| | CP déb. | CP frcph. | LA inter. | LA frcph. | LM av. | LM frcph. | Des deb. | Des inter. | Des. av. | Des frcph |
|---|---|---|---|---|---|---|---|---|---|---|
| nb étoiles | 6 | 10 | 2 | 5 | 6 | 6 | 2 | 3 | 3 | 13 |
| % texte | 50 | 50 | 60 | 30 | 25 | 17 | 33 | 33 | 35 | 14 |

*Tableau 2 : Nombre d'étoiles et pourcentage de texte représenté par celles-ci.*

Le tableau 2 ci-dessus donne les résultats numériques des calculs statistiques effectués par l'analyseur. Il indique, pour chaque classe de scripteurs (francophones : frcph ; apprenants débutants : deb ; apprenants intermédiaires : inter ; apprenants avancés : av) et pour chaque type de production, le nombre d'étoiles détectées par le Littératron ainsi que le pourcentage de texte représenté par ces étoiles. Les paramètres d'entraînement du Littératron sont identiques sur tous ces ensembles de productions, en particulier les seuillages de l'algorithme centre étoile et du graphe de similarité. Autrement dit, Le nombre d'étoiles détectées est donc un bon indicateur de la richesse stylistique : plus il y a d'étoiles, plus le style est riche, c'est-à-dire moins les automatismes prévalent. Il en va de même pour le pourcentage de texte couvert par les étoiles détectées : plus celui-ci est faible, plus les patrons varient, ce qui signifie que le style est plus riche.

Notons que cette notion de richesse stylistique doit être relativisée ; en effet, un grand écrivain pourrait se caractériser par la singularité d'un style qui déclinerait une palette restreinte de patrons, tandis qu'un écrivain sans style les déploierait tous. En dépit de ces quelques réserves, dans le cas particulier de la didactique qui nous intéresse, nous assimilons la richesse d'un texte (ou d'un ensemble de textes) au nombre de figures syntaxiques employées.

[v] CECR : Cadre Européen Commun de Référence

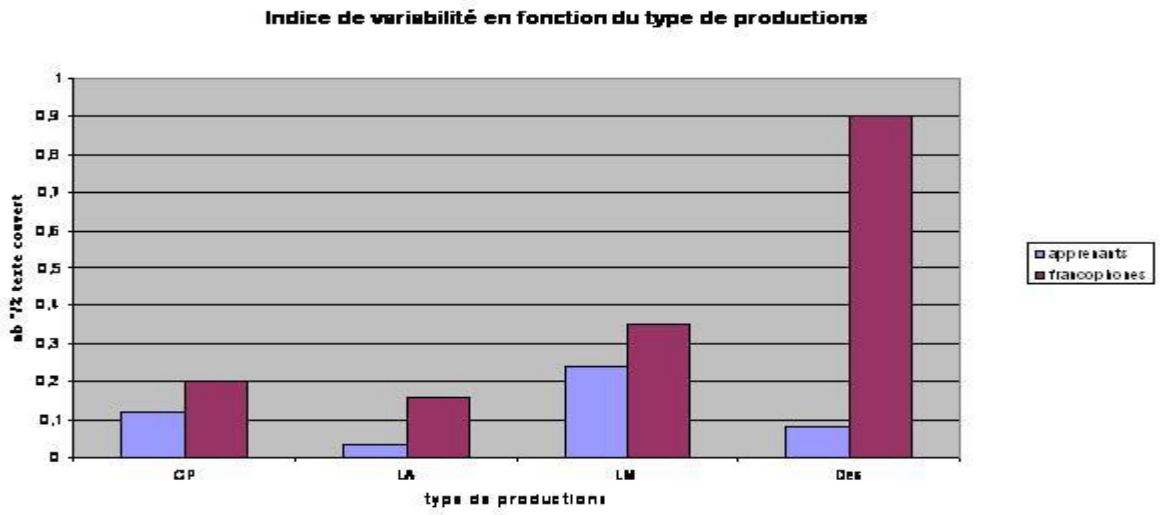


*Figure 2 – Indice de variabilité en fonction du type de production*

Sur le graphe de la figure 2, nous définissons un indice de variabilité qui est, pour chaque type de texte, le rapport du nombre d'étoiles détectées sur le pourcentage de texte, parmi l'ensemble des textes de même consigne soumis à comparaison, utilisé par l'application.
Que conclure du nombre de motifs syntaxiques récurrents et du pourcentage de texte recouvert par ces motifs ? D'une part, les résultats statistiques représentés par l'indice de variabilité nous montre 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 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 partie de texte, où le Littératron n'a pas détecté de motifs récurrents, pourrait être utilisée pour définir l'originalité du scripteur.

Cette analyse a révélé des automatismes de l'écrit à l'intérieur de certains types de production. Ces automatismes concernent aussi bien des textes d'apprenants du français que ceux des francophones. Il y a donc des matrices d'écriture de cartes postales, de lettres d'invitation ou de 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 compétence morpho-syntaxique s'acquiert.

Enfin, concernant la description, nous sommes en mesure de rajouter quelques commentaires sur la structure syntaxique des motifs extraits. Les motifs de base extraits en qualité de syntagme nominal et en qualité de syntagme verbal ont, respectivement, la composition suivante : préposition + substantif + adjectif qualificatif et pronom sujet + verbe + adverbe. Ces motifs de base s'enrichissent progressivement en fonction de la maîtrise du français écrit.

Par exemple, le syntagme nominal "des paysages variés" issu d'une production d'un scripteur débutant évolue en "un paysage bien vert" chez un scripteur natif francophone.

## 5. Deuxième type d'expérience : les apprenants sont de même langue maternelle

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

### *5.1. Présentation des productions écrites et méthodologie expérimentale*

L'ensemble des productions analysées correspond aux examens d'histoire et de civilisation du 1er semestre 2004 d'étudiants en 3ème année du département de français de l'Université An-Najah de Naplouse (Territoires Palestiniens). Les étudiants de l'université sont tous de langue maternelle arabe, l'anglais est leur première langue étrangère, le français leur deuxième. La plupart des étudiants ont commencé à apprendre le français en arrivant au département, sauf pour un petit nombre d'entre eux issus de collèges et lycées où l'enseignement du français en primaire et secondaire est obligatoire. Les mêmes étudiants ont écrit les deux examens. Les productions de chaque type sont au nombre de dix.
La méthodologie expérimentale est la même que pour la première expérience : chaque type de productions est introduit successivement dans l'analyseur morpho-syntaxique puis dans le Littératron, en même temps que des productions de même consigne écrites par des francophones. Les scripteurs francophones sont des natifs français de niveau d'étude au moins équivalent à bac+4.

### *5.2.Résultats et commentaires*

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 ou DE+nom+adjectif, comme dans les exemples : 'de choses magnifiques' et 'd'autres villages'.
Ce motif syntaxique révèle une utilisation massive de groupes nominaux compléments du verbe de forme adjectif + nom commençant par DE, au détriment d'autres articles et d'autres prépositions. 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.
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.
Prenons l'un des deux motifs extraits souvent répétés : « de choses magnifiques ». Ce motif apparaît dans plusieurs phrases et chez différents apprenants. En voici deux exemples, pris chez deux apprenants :
- apprenant 1 : « J'ai visité de choses magnifiques »
- apprenant 2 : « J'ai vu de choses magnifiques »
Ces deux occurrences montrent une utilisation erronée du déterminant DE qui introduit le complément d'objet du verbe 'visiter' dans le premier exemple et 'voir' dans le deuxième.

### *5.3.Population créolophone*

Une expérimentation en cours, sur des productions de même consigne, auprès d'étudiants de l'Ecole Normale de Port-au-Prince (Haïti) semble converger vers la même conclusion, avec une flagrante récurrence de motifs extraits de la forme : DE + adjectif + nom.

## 6. Conclusion

L'analyse syntaxique d'un scripteur par le Littératron est un outil qui vérifie l'acquisition ou l'absence d'acquisition d'un point de morpho-syntaxe. Autrement dit, 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 linguistique 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

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. Quant au type de productions analysées, une étude similaire sur des productions de DELF est envisagée.

Par ailleurs, une autre recherche, en caractérisation lexicale cette fois-ci, serait riche d'enseignement pour l'accompagnement à l'acquisition de l'écrit.

## Références

Adam, J.-M. (1992). Les textes : types et prototypes. Récit, description, argumentation, explication et dialogue. Paris : Nathan.

Besse, J.-M. (dir.) (2003). Qui est illettré ? Décrire et évaluer les difficultés à se servir de l'Ecrit. Paris : Retz.

Carroll, M. & Stutterheim Ch. (1997). "Relations entre grammaticalisation et conceptualisation et implications sur l'acquisition d'une langue étrangère". *Acquisition et Interaction en Langue Etrangère (AILE)*, vol. 9, pp. 14-19.

Dupoux, E. (ed.) (2001). *Language, brain and cognitive development : Essays in Honor of Jacques Mehler,* Cambridge, Mass. : MIT Pr. Morais J. & Kolinsky R. (2001). "The literate mind and the universal human mind". In Dupoux. pp. 463-480.

Ganascia, J.-G. (2001). "Extraction automatique de motifs syntaxiques". *In actes du colloque Traitement Automatique du Langage Naturel 2001* (*TALN 2001).* Consulté en juillet 2005. http://www.li.univ-tours.fr/taln-recital-2001/index1.html

Ganascia, J.-G. (2001). "Extraction of Recurrent Patterns from Stratified Ordered Trees". *In actes de Machine Learning : 12th European Conference of Machine Learning 2001 (ECML 2001)*. Freiburg : Springer-Verlag, pp. 167-179.

Ganascia, J.-G. (2004). "Detection of Statistically Abnormal Patterns from Stratified Ordered Trees". In Milutinovic & Vujovic Milutinovic, (dir.) *Advances in the Internet Technology, Concepts and Systems.* .

Garcia-Debanc, C. (1993). "Enseignement de la langue et production d'écrits". In *Ecriture et langue*, Garcia-Debanc (dir.). *Pratiques*, 77. pp 3-23.

Giguet, E. (1998). "Méthode pour l'analyse automatique de structures formelles sur documents multilingues". Thèse de doctorat en informatique, Université de Caen. Consulté en juillet 2005. http://users.info.unicaen.fr/~giguet/these/

Kahn, G. (dir.) (1993). *Des pratiques de l'écrit*. numéro spécial *Le Français dans le Monde*. Paris : Hachette.

Mangenot, F. (1998). "Outils textuels pour l'apprentissage de l'écriture en L1 et en L2"**.** In *Pratiques discursives et acquisition des langues étrangères*, Souchon, M. (dir.). *Actes du Xe colloque international "Acquisition d'une langue étrangère : perspectives et recherches"*. Besançon : Université de Franche-Comté. pp 515-525.

Moirand, S. (1990). *Une grammaire des textes et des dialogues*. Paris : Hachette FLE.

Pery-Woodley, M.-P. (1993). *Les écrits dans l'apprentissage*. *Clés pour analyser les productions des apprenants*. Paris: Hachette FLE.

Scardamalia, M. & Bereiter, C. (1986). "Research on written composition". In *Handbook of research on teaching*, Wittrock, M.C. (dir.). New York : McMillan. pp 778-803.

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, pp. 691-721.

Vergne, J. (1999). "Etude et modélisation de la syntaxe des langues à l'aide de l'ordinateur. Analyse syntaxique non combinatoire Synthèse et résultats". Habilitation à Diriger des Recherches, Université de Caen. Consulté en juillet 2005. http://users.info.unicaen.fr/~jvergne/#HDR.