Analyses comparatives de productions écrites d'apprenants de français et de locuteurs francophones, à l'aide d'outils d'extraction automatique du langage

Isabelle Audras, Jean-Gabriel Ganascia

Analyses comparatives de productions écrites d'apprenants de français et de locuteurs francophones, à l'aide d'outils d'extraction automatique du langage.

Isabelle Audras, Jean-Gabriel Ganascia, LIP6, Université Pierre et Marie Curie

1. Introduction

La didactique des langues étrangères pourrait utilement intégrer 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 disparaîtrait, tout simplement… On peut toutefois envisager d'autres perspectives : ainsi, il ne s'agirait 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 écrites usuelles 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 soit dans l'absolu, par détection des erreurs syntaxiques, soit par rapport aux usages, par une étude des tournures propres à une catégorie d'apprenants dans un cadre narratif précis, celles-ci se trouvant absentes ou peu usitées chez les locuteurs natifs. C'est cette seconde approche que nous avons adoptée, sachant que le rôle des enseignants de langue n'est pas d'enseigner 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 (Territoires Palestiniens), 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 des différentes populations. Un motif syntaxique se définit comme une association d'unités linguistiques cohérentes. Voici un exemple de motifs extraits des analyseurs ayant la structure syntaxique [préposition + pronom personnel réfléchi + verbe à l'infinitif] : "de vous adresser", "afin de vous donner", "de m'investir", "de vous donner". Ces quatre motifs ont été extraits ensemble d'un même groupe de scripteurs de lettres de motivation.

Le premier outil informatique requis est un analyseur morphosyntaxique du français qui construit des arbres syntaxiques à partir de productions écrites. Le deuxième est l'analyseur stylistique Littératron, mis au point au LIP6 par Jean-Gabriel Ganascia (Ganascia, 2002) qui dégage les motifs syntaxiques récurrents présents dans ces arbres.

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

Dans les premières expériences, 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 (Giguet, 1998 et Vergne, 1999). Dernièrement, nous avons utilisé l'*analyseur Cordial* (© Synapse Développement). Ces analyseurs découpent un texte en langage naturel en syntagmes non récursifs[1]. Les sorties sont ensuite transformées en arbres stratifiés ordonnés (ASO) pour servir d'entrée au *Littératron*.

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[2].

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,

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

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

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 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 :

- [PREP ["à"]] + [GN [ART ["l"]] + [NOM ["auberge"]]] (texte : "à l\'auberge ") ;
- [PREP ["d\"]] + [GN [ART ["un"]] + [NOM ["hiver"]]] (texte : "d\'un hiver") ;
- [PREP ["dans"]] + [GN [ART ["le"]] + [NOM ["monde"]]] (texte : "dans le monde") ;
- [PREP ["avec"]] + [GN [ART ["les"]] + [NOM ["chiens"]]] (texte : "avec les chiens") ;
- [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"]]].

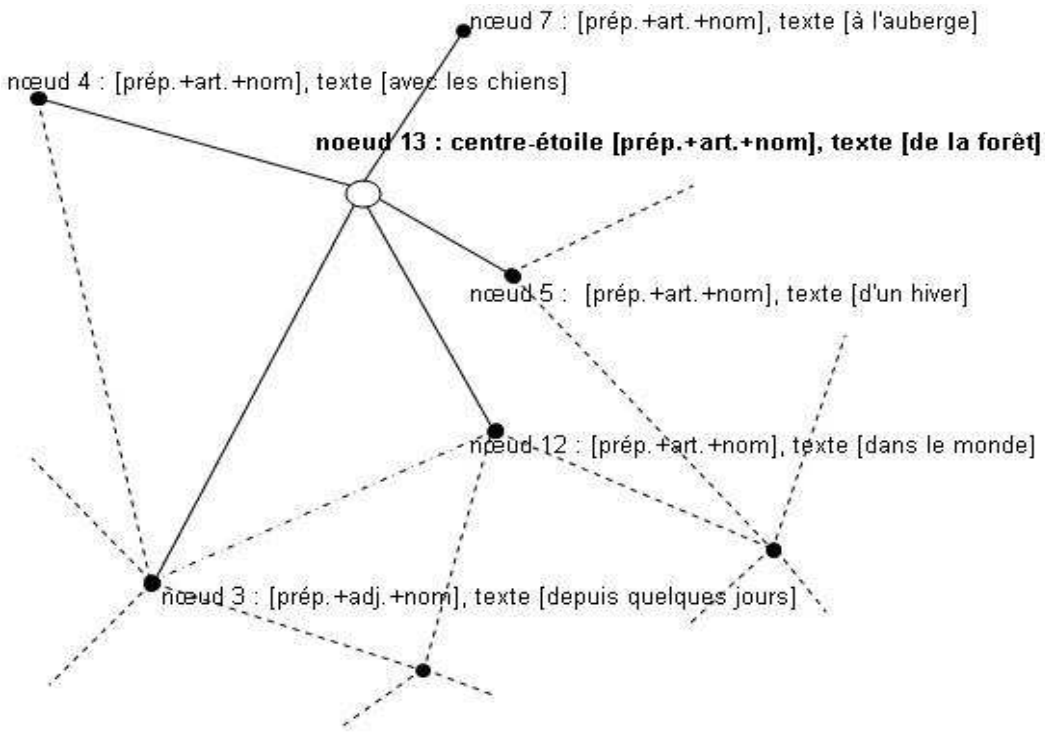

Figure 1 – Graphe du centre-étoile présenté en exemple.

2.4. Comparaison de motifs entre deux textes distincts

Outre la construction d'étoiles et l'extraction de motifs, le *Littératron* procède à un second type d'opérations qui consiste à comparer les étoiles issues de plusieurs textes afin de repérer les étoiles présentes dans l'un et absentes dans 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).

Patron N°46

- Indépendante

  S

  P p 1 s

  V

  COD

  N c f s

- 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

Le patron décrit la structure syntaxique du motif extrait. Ici il s'agit d'une proposition indépendante de forme sujet S + verbe V + COD. Il y est précisé que le sujet est un pronom personnel à la première personne du singulier (Pp 1 s) et que le COD est un nom commun féminin singulier (Nc f s).

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.

4. Premier type d'expérience : les apprenants sont de langues maternelles diverses.

L'idée de cette recherche est de comparer des productions écrites en classe de F.L.E. de différents niveaux avec des productions de francophones répondant aux mêmes consignes. Les scripteurs francophones sont des natifs français de niveau d'étude au moins équivalent à bac+4. Une autre étude, qui pourrait se révéler intéressante, travaillerait avec des francophones d'un niveau d'études moins élevé, voire des débutants scripteurs adultes (Morais J. & Kolinsky R., 2001). L'idée de cette perspective est de montrer que le critère du niveau d'éducation n'est pas négligé dans cette approche.

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[3])	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

[3] CECR : Cadre Européen Commun de Référence

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.

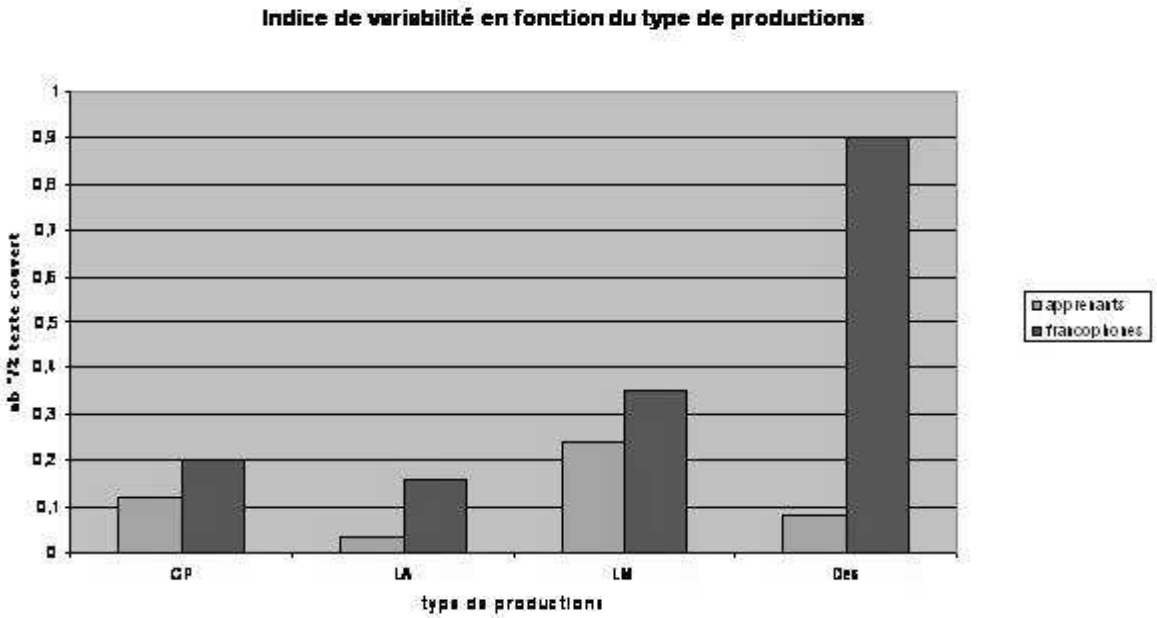

Figure 2 – Indice de variabilité en fonction du type de production

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 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'arabe.

L'objectif de cette seconde expérience est de détecter, à partir de la liste des compétences requises pour chaque unité du DELF, celles qui sont présentes et celles qui sont absentes chez l'apprenant lorsqu'il écrit, que ce soit pour une préparation à l'épreuve ou une épreuve réelle. Suite au repérage éventuel de ces lacunes, un programme de remédiation personnalisé peut être proposé à l'apprenant.

5.1. Présentation des productions

Deux types de productions ont été choisies : la lettre de motivation (LM) et la lettre amicale (LA). Ces deux productions ont été tirées d'épreuves du DELF session 2004 ;

le DELF solaire - public adolescent - pour les lettres amicales, et l'unité A3 - public adulte - 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, quinze productions d'apprenants et quinze de francophones sont introduites dans l'analyseur textuel *Cordial* dont la sortie est ensuite traitée par le *Littératron*.

5.2. Points de morpho-syntaxe à vérifier

Nous décidons de vérifier successivement la présence ou l'absence des compétences morpho-syntaxiques suivantes :

- la ponctuation et la coordination, dans une structure de phrase du type "…, … et …" ;
- la subordination, avec la complétive de structure : sujet + verbe + que ;
- les expansions du nom, avec la relative introduite par "qui".

Ces compétences morpho-syntaxiques ont été choisies parce qu'elles nous semblaient correspondre à des compétences langagières minimales.

A cette fin, nous avons calculé pour chaque point morpho-syntaxique une figure qui rassemble l'ensemble des patrons correspondants. Par exemple, la figure de ponctuation et de coordination recouvre l'ensemble des patrons qui contiennent une ponctuation ou une coordination. De même, la figure de subordination recouvre l'ensemble des patrons qui introduisent une subordination. Enfin, la figure des expansions du nom recouvre les patrons qui font intervenir un pronom relatif. Nous avons ensuite calculé la fréquence de ces figures au sein des textes, puis nous avons centré le tableau de fréquence sur l'ensemble des productions de façon à faire apparaître les excès relatifs et, à l'opposé, les défauts.

5.3. Résultats et commentaires

Le tableau 3 donne l'inventaire des trois figures présentées plus haut relativement aux sorties du *Littératron*.

Points de morpho-syntaxe et motifs observés	Lettres de motivation d'apprenants arabophones (LM appra)	Lettres de motivation de francophones (LM frcph)	Lettres amicales d'apprenants arabophones (LA appra)	Lettres amicales de francophones (LA frcph)
Ponctuation/coordination : …, … et …	- 0.046	**0.041**	- 0.075	- 0.0082
Subordination : sujet + verbe + que	- 0.0034	**0.018**	- 0.025	**0.0063**
Relative : qui	- 0.026	**0.079**	- 0.071	- 0.085

Tableau 3 : Sorties du *Littératron*.

Les chiffres qui apparaissent dans le tableau n'ont pas d'unité : ce sont des valeurs de fréquences relatives qui n'ont d'intérêt que si elles sont comparées deux à deux, d'où la coloration pour les valeurs les plus pertinentes qui correspondent à un excès de fréquences. Enfin, ces valeurs permettent de construire des graphes comme celui-ci :

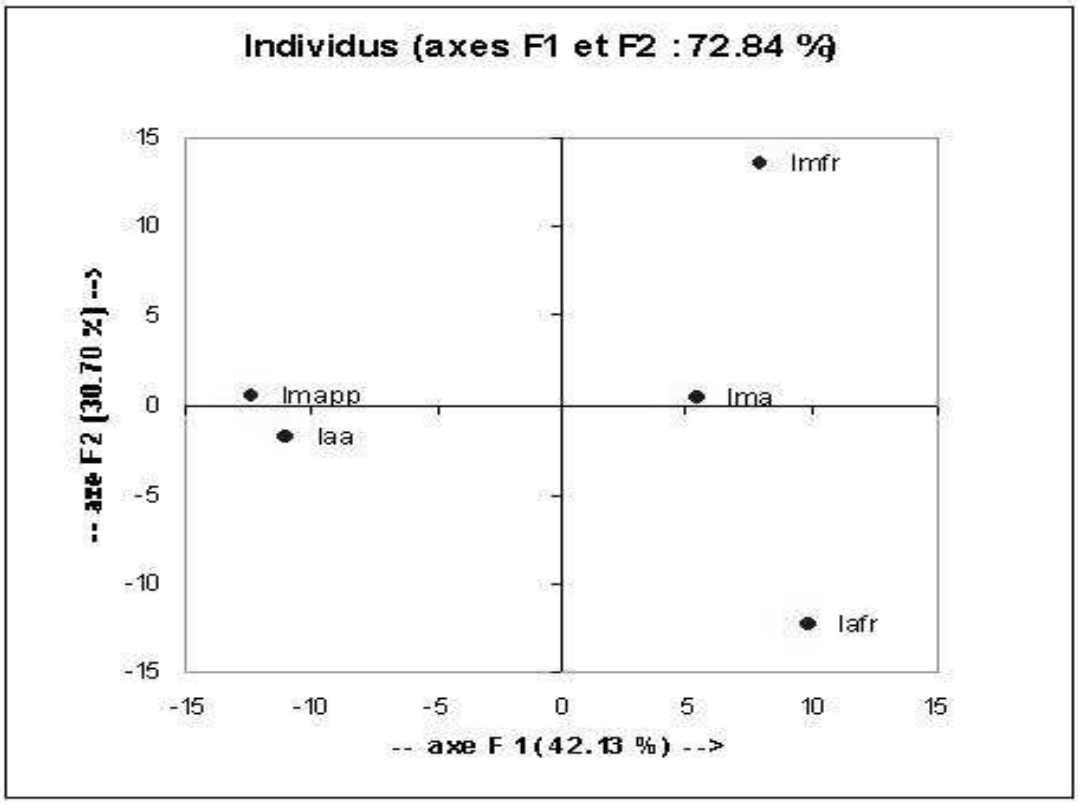

Figure 3 – Analyse en composante principale des sorties du *Littératron*

Les valeurs colorées du tableau 3 font apparaître clairement les motifs syntaxiques choisis dans les productions de francophones. La lettre de motivation s'avère encore

plus riche que la lettre amicale : est-ce parce qu'il s'agit d'adolescents dans la lettre amicale ?

Ces contrastes sont encore plus visibles sur le graphe. L'analyse en composante principale permet une meilleure visibilité du grand nombre de sorties à analyser. Chaque zone délimitée correspond à l'ensemble des motifs extraits pour chaque production (LM app, LM frcph, etc.). Nous avons ajouté aux productions du tableau 3 les lettres de motivation d'apprenants de la première expérience (LM appr).

Les axes F1 et F2 du graphe sont tracés en proportion avec les plus grandes fréquences de motifs extraits. Ils ont un sens linguistique. Ici, étant donné la pauvreté dans la diversité et la complexité des motifs extraits, ces axes représentent probablement, l'un les motifs représentant des structures de phrase simples et l'autre les syntagmes nominaux. LM appra et LM appr ne couvrent pas la même zone du graphe, même si 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. Par contre, chaque type de production francophone a l'air bien éloigné de son pendant arabophone ou autre apprenant (LM frcph et LM appra ; LA appra et LA frcph).

A partir des motifs extraits caractéristiques de chaque zone, il est ensuite aisé de construire un tableau récapitulatif des compétences présentes / absentes et de les spécifier selon le niveau attendu pour chaque unité de DELF.

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).

6. Conclusion

Utilisé en sciences du langage dans le domaine de l'acquisition en langue étrangère du français écrit, le *Littératron* est en mesure d'effectuer un diagnostic linguistique de l'apprenant sur des productions au cadre narratif contraignant. En effet, les invariants et diversités syntaxiques extraits témoignent des compétences présentes lors du passage à l'écrit de l'apprenant.

Ainsi, ces applications révèlent clairement :

a) 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.

b) dans la deuxième expérience :

- des compétences non acquises ou en cours d'acquisition lors du passage à l'écrit, démasquées ;

- des reflets de l'apprentissage de la langue-cible et des effets imputables à la langue-source ; ainsi 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, sur les 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 ; 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 de langues maternelles diverses.

Références.

Références bibliographiques

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. (2002). "Extraction of Syntactical Patterns from Parsing Trees". *In actes de Internationale Conference on Textual Data Statistical Analysis*.

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/

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

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.

Logiciels

Littératron (2001). Le Littératron a été conçu par Jean-Gabriel Ganascia. Université Pierre et Marie Curie : Paris.

Analyseur Vergne (1998). L'analyseur Vergne a été conçu par Jacques Vergne. Université de Caen Jacques Vergnes (voir http://www.info.unicaen.fr/~jvergnes et (Vergnes 1999)) au GREYC (Groupe de Recherche en Informatique, Image, Instrumentation de Caen).

Cordial Analyseur (version 8). Synapse développement : Toulouse. Consulté en juillet 2005 : http://www.synapse-fr.com/

Remerciements

Je tiens à remercier particulièrement, pour leurs conseils et leurs remarques en tant que relecteurs : Nathalie Hirschsprung (Centre Culturel Français Romain Gary), Julien Velcin (LIP6) et Philippe Boula de Mareuil (LIMSI).