Classification d'apprenants francophones de l'anglais sur la base des métriques de complexité lexicale et syntaxique

Nicolas Ballier[1] Thomas Gaillat[1, 2]
(1) CLILLAC-ARP (EA3967), rue Thomas Mann, 75013 PARIS, FRANCE
(2) SCELVA, Campus Beaulieu 35042 RENNES cedex, FRANCE
nicolas.ballier@univ-paris-diderot.fr, thomas.gaillat@univ-rennes1.fr

RESUME

Cette contribution examine les monologues en oral spontané du corpus ANGLISH (Tortel 2009). Les productions orales de vingt locuteurs natifs sont comparées aux monologues produits par les quarante locuteurs francophones de niveau intermédiaire et avancé. Les métriques de complexité syntaxique et lexicale implémentées dans des analyseurs (Lu 2014) sont utilisées pour essayer de classer les locuteurs. Enfin, à partir des courbes de croissance du vocabulaire et des modèles LNRE (Baayen 2008), on cherche à évaluer la pertinence de ces métriques de l'écrit pour la classification des locuteurs en fonction de leur production orale.

ABSTRACT

Classifying French learners of English with written-based lexical and complexity metrics.
This paper assesses spontaneous oral monologues in the ANGLISH corpus (Tortel 2009). Twenty oral transcriptions of NS English are compared with forty French-L1 transcriptions of NNS English of intermediate and advanced levels. Syntactic and complexity metrics (Lu 2014) and Vocabulary Growth Curves (Evert & Baroni 2008, Baayen 2008) are used to classify speakers. We analyse how significant these written-based metrics are in the classification of speakers for their oral production.

MOTS-CLES : complexité lexicale, complexité syntaxique, métriques d'apprenants, modèles LNRE
KEYWORDS: syntactic complexity, lexical complexity, vocabulary growth curves, LNRE models.

1 Introduction

Les corpus d'apprenants oraux sont encore rares (Ballier & Martin 2015), alors que bien des corpus d'apprenants écrits ont fait l'objet d'analyses de traitements automatiques (Cf., entre autres, Díaz et al. 2013; Tono 2013). Pour bien des corpus d'apprenants (ANGLISH, NOCE, LeAP), c'est le niveau d'étude à l'université des apprenants qui sert de référence à l'évaluation de leur niveau initial. Il y a donc une sorte de présomption d'homogénéité de la promotion d'étudiants, l'appartenance à un niveau universitaire devant garantir le niveau des sujets enregistrés et étudiés. Dans le cas du corpus oral ANGLISH (Tortel 2009), le choix des sujets enregistrés pour le corpus s'est de plus un peu fait sur la base des situations professionnelles, puisque les locuteurs de niveau intermédiaire (FR1) ont été en partie recrutés sur le campus de l'université (et non sur la base de tests spécifiques, mais aucun n'avait étudié l'anglais en études supérieures), les étudiants

spécialistes d'anglais en troisième année à l'université ont constitué le groupe de locuteurs francophones avancés en anglais (FR2). Le corpus a échantillonné trois populations de locuteurs (GB : natifs anglophones, FR1 et FR2) et permet des comparaisons éclairantes sur des tâches identiques. Pour l'analyse du texte lu (enjeu de la thèse), l'analyse révèle une certaine forme d'homogénéité, au moins en termes de variance, au sein des trois groupes ; le pari global de niveaux d'étudiants conçus à partir de promotion n'est donc pas complètement invalidé. Pour autant, les compétences en anglais des sujets moins avancés (FR1) sont sans doute extrêmement variables, qu'il s'agisse d'un expert lisant régulièrement en anglais dans son domaine ou de sujets ayant conservé une certaine proximité professionnelle en raison de leur activité professionnelle avec les textes écrits en anglais. Dans quelle mesure peut-on utiliser les métriques de complexité lexicale et syntaxique comme des procédures *a posteriori* établissant le niveau des apprenants ?

La conversion en niveaux de référence du CECRL est encore programmatique (mais voir Ovtcharov et al. 2006, ainsi que Leclercq et al. 2014) et cela constitue en quelque sorte l'horizon de ce travail, qui s'inscrit dans la recherche des traits critériés (Hawkins & Filipovic 2012). Sur la base des performances lexicales et syntaxiques, on cherche à distinguer des productions des natifs et des non-natifs, et surtout des stades d'interlangue chez les non-natifs (Gaillat 2013), voire des profils d'apprenants (Chitez 2014). La deuxième partie présente le corpus analysé et la méthode suivie. La troisième section expose les résultats, que la quatrième section discute et met en perspective.

2 Méthode

Cette expérimentation va appliquer à des données orales des métriques mises au point pour l'écrit. Nous présentons successivement les données étudiées et les différentes méthodes d'analyse suivies, par ordre décroissant d'intervention sur les données.

2.1 Description du corpus

Le corpus ANGLISH compte 60 locuteurs : 20 anglophones natifs, 20 francophones natifs ayant arrêté l'anglais au bac (FR01) et 20 étudiants en troisième année de Licence d'anglais à l'Université de Provence (FR02). Une partie en texte lu a fait l'objet d'une thèse qui a cherché à classer les locuteurs sur la base des différences de rythme (Tortel 2009). Le corpus a été transcrit dans les annexes de cette thèse, en suivant des conventions de transcription assez standard pour un corpus d'oral (transcription de pauses pleines, absence de majuscule). Etablies par une spécialiste de l'oral, ces données ne sont pas directement interprétables par les logiciels de l'écrit (nombreux problèmes de casse, de répétitions pour le *parsing* des phrases, étapes indispensables pour les analyses automatiques de la complexité syntaxique). Nous avons dû modifier le format des données pour l'analyseur de complexité : nous avons ponctué le texte et supprimé les répétitions. L'accord inter-annotateur entre les deux experts a été de 85%, les différences portant essentiellement sur les virgules après les adverbiales. Pour l'analyseur de complexité lexicale, il y a une procédure de lemmatisation ; pour les courbes de croissance du vocabulaire, nous avons utilisé les données brutes.

Pour une même consigne ("vous allez parler librement en anglais pendant deux minutes sur le sujet de votre choix. Vous entendrez un signal de "top départ" et je vous ferai signe lorsque les deux minutes seront atteintes. Je vous laisse réfléchir pendant une minute afin de choisir votre thème, si vous n'avez pas d'idées je vous propose de raconter vos dernières vacances"), les trois groupes de locuteurs s'exécutent différemment : 5 027 tokens pour les FR1, 5 683 pour les FR2 et 8 293 pour les natifs. On a prélevé un échantillon comparable des 5 027 tokens dans les trois corpus. Nous avons suivi des exemples de

Baayen 2008 et en quelque sorte «redécoupé » les corpus à des dimensions identiques, une randomisation des occurrences retenues pourrait donner des résultats différents.

2.2 Les métriques de complexité syntaxique

Comme expliqué dans (Lu 2010, Lu 2016) et dans la documentation, l'analyseur de complexité syntaxique L2SCA repose sur le parser de Stanford (Klein & Manning, 2003), génère les fréquences des principales unités textuelles et syntaxiques et calcule les indices de complexité syntaxique en L2 proposés dans la littérature et notamment ceux compilés dans (Wolfe-Quintero et al. 1998). Il produit donc pour chaque fichier texte en entrée une série de 23 traits :

- neuf mesures syntaxiques : le nombre de mots (W), de phrases (S), de syntagmes verbaux (VP), de propositions (C), de T-units (T), de propositions subordonnées (DC), de T-units complexes (CT), de syntagmes coordonnés (CP), et de SN complexes (CN). L'unité centrale à ces mesures est la T-unit, qui est ainsi définie : ''one main clause plus any subordinate clause or nonclausal structure that is attached to or embedded in it'' Hunt (1970:4).

- quatorze indices de complexité syntaxique : la longueur moyenne de la phrase (MLS), la longueur moyenne de la T-unit (MLT), la longueur moyenne de la proposition (MLC), le nombre de propositions par phrase (C/S), le nombre de syntagme verbal par T-unit (VP/T), le nombre de proposition par T-unit (C/T), le nombre de propositions subordonnées par propositions (DC/C), le nombre de propositions dépendantes par T-unit (DC/T), le nombre de T-units par phrase (T/S), le ratio nombre de T-units complexes /nombre de T-Unit (CT/T), le nombre de syntagmes coordonnés par T-unit (CP/T), le nombre de syntagmes coordonnés par proposition (CP/C), le nombre de groupes nominaux complexes par T-unit (CN/T), et le nombre de groupes nominaux complexes par proposition (CP/C). On voit qu'il s'agit d'un ensemble de ratios calculés par rapport à la proposition ou à la T-unit. Cette *T(erminable)- unit* met en avant un concept de clôture du vouloir-dire. Il existe aussi une caractérisation en un sens encore plus lâche, qui comptabilise de la sorte toute proposition indépendante et ses modificateurs, toute proposition qui ne serait pas une indépendante mais serait ponctuée comme une phrase ou bien même des formes impératives (Schneider & Connor, 1990, cité dans Paris 2015 :203-4).

2.3 Traits de la complexité lexicale

Nous avons utilisé l'analyseur de complexité lexicale développé par Xiaofei Lu (Lu 2012). Ce *Lexical Complexity Analyzer* (LCA) est décrit dans plusieurs travaux de ses travaux (Lu 2013, Lu 2014) et à été mis au point pour l'anglais. Il suppose un choix de variété de référence, britannique ou américaine, les fréquences lexicales de référence ayant été calculées (respectivement) à partir du British National Corpus ou de l'American National Corpus.

L'analyseur LCA suppose en entrée un format de données lemmatisées (au format TreeTagger) et produit 24 mesures. Nous reportons ici par commodité les principaux indices et renvoyons à la lecture de (Lu 2012) pour une discussion critique de ces différents indices. Les mesures portent sur la densité lexicale, la variation lexicale et l'élaboration lexicale (*sophistication*). Les mesures de densité lexicale portent sur les types, les tokens et sur le ratio de certaines catégories. La variation lexicale est essentiellement comptabilisée à partir du *type-to-token ratio* et de ses variantes. Les mesures d'élaboration lexicale reposent sur des proportions des unités lexicales plus élaborées.

Métriques	Signification
Sentence	Le nombre de segments séparés par un point dans chaque transcription
Wordtypes, swordtypes	Trois indicateurs dénombrant le nombre de mots en fonction de leur catégorie grammaticale
Lextype, slextypes	Indicateurs du nombre de lemmes lexicaux
Wordtokens, swordtokens	Indicateurs du nombre de tokens liés aux mots différents
lextokens, slextokens	Indicateurs du nombre de tokens lexicaux liés aux lemmes différents
Ld	Indice de densité lexicale (= nombre de mots lexicaux et non grammaticaux/nombre de tokens)
ls1, ls2	Proportions de lexèmes parmi les 2000 mots les plus fréquents (Laufer & Nation 1994)
vs1, vs2, cvs1	Ratio de verbes ne figurant pas parmi les 20 ou 200 verbes les plus fréquents en français / nombre de verbes utilisés (et ses variantes, cf. Wolfe-Quintero et al. 1998)
ndw, ndwz, ndwerz, ndwesz	Indices fondés sur le nombre de mots différents pris dans des échantillons de 50 items.
ttr, msttr, cttr, rttr, logttr, Uber	Indices fondés sur le ratio entre le nombre de mots différents et le nombre total de mots (*type-to-token* ratio et ses avatars normalisés ou transformés :TTR, MSTTR, CTTR, RTTR, et LogTTR). L'index Uber est le Log2(types)/log(types/tokens).
lv, vv1, svv1, cvv1, vv2	Indices de variation concernant le lexique et les formes verbales
nv, adjv, advv, modv	Indices de variation concernant les noms, adjectifs, adverbes et modificateurs (englobant adjectifs et adverbes)

TABLE 1 : Synthèse des principales métriques implémentées dans LCA (Lu 2012, Lu2014).

La réflexion sur la complexité lexicale est dans ce cadre indissociable de la lemmatisation, contrairement aux techniques suivantes.

2.4 L'étude des courbes de croissance du vocabulaire

Bien que pratiquées sur des échelles de données singulièrement différentes (la partie étudiée ici du corpus ANGLISH compte 19 003 tokens, à comparer avec le million du corpus Brown), l'intérêt de la problématique demeure : dans quelle mesure les données d'apprenants, sujettes à l'appauvrissement du lexique au fur et à mesure de leur expression (Bentz et al. 2013) sont-elles justiciables de la loi de Zipf (Zipf 1949) ? L'examen individuel des courbes d'accroissement du vocabulaire de chaque apprenant insisterait sur les spécificités. L'objectif général est de modéliser, à partir de spectres de fréquences, les courbes de l'accroissement du vocabulaire (nombre de types et d'hapax) en fonction de la taille de l'échantillon (nombre de *tokens*, voir Baayen 2008 et, pour une description rapide en français, Turenne 2016). Le spectre fréquentiel est ainsi défini " A frequency spectrum summarizes a frequency distribution in terms of number of types (Vm) per frequency class (m), i.e., it reports how many distinct types occur once, how many types occur twice, and so on. " (Baroni & Evert 2014:2). On a donc établi pour les trois groupes de locuteurs la courbe de croissance du vocabulaire, qui est en partie fondée sur une mesure des hapax (Evert & Baroni 2004, Baroni 2008).

Ensuite, nous avons utilisé la fonction de comparaison de richesse lexicale "compare.richness.fnc" implémentée dans le package {languageR} (Baayen 2008), qui compare les documents deux à deux à partir de leurs spectres fréquentiels. Dans leurs analyses, (Evert & Baroni 2004 et Baayen 2008) ne procèdent pas à une lemmatisation, contrairement aux méthodes décrites dans (Lu 2014).

Enfin, nous avons essayé d'exploiter les courbes de croissance du vocabulaire. L'objectif est la caractérisation de traits grammaticaux et lexicaux permettant pour chaque enregistrement d'établir une différence de niveau, nous avons donc aussi utilisé les cinq premières valeurs des courbes de croissance du vocabulaire pour chaque transcription, ce qui correspond aux effectifs cumulés des hapax par tranches de 100 occurrences. Nous avons ici restreint la fenêtre de la mesure des hapax, en raison de la petitesse de la taille des donnés. Dans les travaux précédents (Baayen 2001, 2008), le nombre d'hapax est mesuré par incrémentations successives de 500 ou 1000 tokens. Un jeu de données fondé sur l'augmentation des hapax au fil du monologue est donc construit avec cinq effectifs d'hapax pour chaque sujet. On a ensuite procédé à une classification automatique à partir de TiMBL (Daelmans et al. 1994).

2.5 Classifieur retenu

Le classifieur TiMBL, déjà mis à contribution dans plus d'une centaine d'études, permet l'estimation des traits les plus déterminants dans la classification, le calcul du GainRatio (Daelemans et al. 2005) rend possibles une hiérarchisation des traits (*features*) et une optimisation de la classification fondée sur la pondération des traits les plus pertinents dans la classification. C'est ce qui le distingue de la famille des algorithmes k-nn dont il relève.

Pour rappel, le classifieur procède dans un premier temps à un apprentissage en mémorisant les différents vecteurs et leur classe, et en calculant une pondération des variables basée sur la notion d'entropie. Dans un second temps, de nouveaux vecteurs (sans classe) sont présentés au module de classification pour se voir attribuer une classe. Il y a donc un échantillon pour chaque phase. Dans le cas de notre expérience, l'échantillon que nous utilisons ne comprend que 60 observations (converties en 60 vecteurs de variables), ce qui est relativement faible. Afin de pallier ce manque, nous utilisons l'option 'leave-one-out' de TiMBL qui permet d'utiliser le même échantillon pour l'apprentissage et le test. Le programme effectue 60 passages de classification. A chaque passage, 59

vecteurs sont utilisés pour l'apprentissage, le dernier étant utilisé pour le test. Une fois la classification effectuée, des statistiques de classification sont renvoyées, ce qui permet de mesurer la précision de la classification. En outre, la pondération des variables composant les vecteurs est affichée. Cela permet d'avoir une vue précise sur la pondération attribuée à chacune d'entre elles, et ainsi de voir les variables les plus significative du point de vue de leur potentiel de réorganisation des informations. Nous avons donc utilisé l'option *leave-one-out* dans nos classifications.

3. Résultats

A titre de comparaison, rappelons d'abord que la classification obtenue par Anne Tortel sur la base des métriques du rythme était de 69% pour les trois groupes de locuteurs, mais les données portaient sur les productions lues des sujets, non sur leur production spontanée.

3.1 Classification fondée sur la complexité syntaxique

La précision de la classification sur la base des 23 traits retenus pour l'analyse syntaxique est très moyenne (48%) et n'a de sens que pour la classification des locuteurs les moins avancés (FR1)

	FR1	FR2	GB
FR1	**14**	3	3
FR2	6	**8**	6
GB	4	9	**7**

TABLE 2 : Matrice de confusion des trois groupes de locuteurs (toutes métriques confondues)

3.2 Complexité lexicale

La précision est de 48,33% (29/60) et donne la matrice de confusion suivante :

	FR1	FR2	GB
FR1	**12**	5	3
FR2	8	**9**	3
GB	5	7	**8**

TABLE 3 : Matrice de confusion des trois groupes de locuteurs (toutes métriques confondue

TiMBL fait apparaître la pertinence des métriques suivantes : adjv (le rôle des adjectifs) et le type-token-ratio (TTR) brut, ce qui s'explique en partie par la brièveté des monologues. Comme le

rappelle (entre autres) Baayen 2001, la mesure du TTR est problématique en ce qu'elle est sensible à la taille des échantillons. La taille recommandée par (Biber 2012) dans une étude sur le TTR est de 450 tokens. Le rapport quantitatif entre le nombre de types et le nombre de tokens est au cœur des modélisations LNRE (*Large Number of Rare Events*).

3.3 Courbes de l'accroissement du vocabulaire et modèles LNRE

La Figure 1 présente la croissance des types (en ordonnées) en fonction de l'accroissement de l'échantillon (nombre de *tokens* en abscisse). La première série de courbes montre la croissance des hapax (en bas), et la deuxième représente la croissance des types en fonction de l'accroissement de l'échantillon (N). On voit que, sur l'échantillon analysé (tous les locuteurs de chaque groupe confondus), la courbe de croissance ne permet pas de discriminer entre les différents groupes. Les micro-variations (les courbes se croisent) sont également imputables à la variation individuelle des locuteurs analysés, car chaque courbe est constituée à partir des effectifs cumulés des 20 locuteurs.

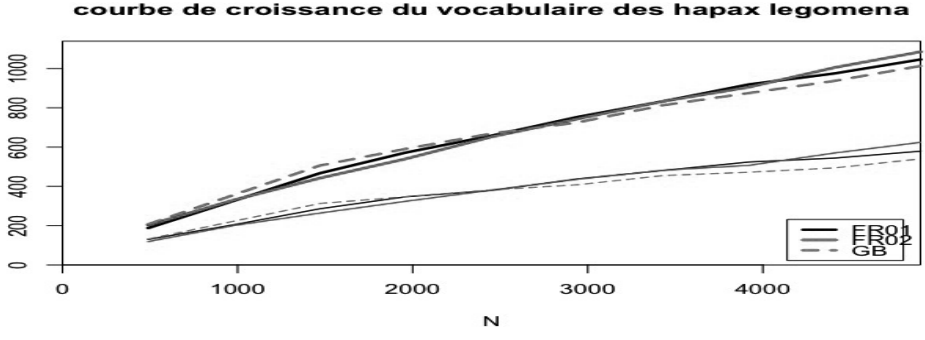

FIGURE 1 : Courbes des *hapax legomena* et du nombre de types en fonction du nombre de tokens

Nous avons également calculé la richesse lexicale, implémentée par Baayen dans le package {languageR} par la fonction compare.richness.fnc(), qui compare les courbes de croissance du vocabulaire (Baayen 2008, Turenne 2016). La comparaison des corpus deux à deux (ramenée, pour les trois sous-corpus, aux 5 027 premiers tokens) ne donne pas de différence significative (p= 0,1969) pour des taux d'accroissement du vocabulaire de 0,11717 pour FR01, de 0,10961 pour GB et de 0.12751 pour FR02. La traduction intuitive que Baayen 2008 donne à cette métrique est que la probabilité que le mot suivant le 5 027^{e} soit une occurrence d'un type inédit jusque là est d'une chance 1 sur 9. Cette fonction est sensible à la taille de l'échantillon, ce qui rend problématique son utilisation avec les corpus d'apprenants, de taille plus restreinte.

4. Discussion et conclusion

Plusieurs ordres de problématiques sont mobilisables, au-delà des problèmes de taille d'échantillon et du traitement statistique choisi. Même avec la méthode du *leave one out*, le nombre de documents pour entraîner le classifieur est très limité. On pourrait affiner les classifications en ne retenant que les traits sélectionnés par TiMBL comme étant davantage pertinents (par exemple le TTR brut), mais on ne saurait rien du caractère *ad hoc* de cette restriction du nombre de métriques pertinentes et on multiplierait les risques d'*overtfitting* sur les données. D'autres méthodes d'analyses

statistiques que la classification étaient possibles pour les données recueillies, en particulier des modèles *stepwise* de régression (cf, entre autres, Johnson 2008).

4.1 Métriques de l'écrit et métriques de l'oral

L'apprenant parle comme il écrit, c'est à ce type de caractéristique qu'on reconnaît qu'il ne parle pas comme un natif. Derrière cette boutade, se dissimulent l'importance de la parataxe en AND chez les natifs et l'importance de l'hypotaxe en SO chez les non-natifs. Les métriques ne rendent pas immédiatement compte des différences, flagrante à l'oral, des constructions prosodiques et de la structuration du discours. Nous y reviendrons et détaillons peu les métriques de complexité syntaxique, le statut de la T-unit pour la caractérisation de l'oral n'est pas idoine, pas plus que notre ponctuation comme pis-aller de la structuration prosodique.

Dans cette approche, les propositions verbales non-finies ne sont pas comptées comme des propositions. Il existe par ailleurs d'autres métriques de complexité, comme le nombre moyen de mots devant le verbe de la principale, le nombre de modificateurs (adjectif ou adverbe) par syntagme nominal, ainsi que des mesures plus techniques fondées sur le corpus arboré (nombre de nœuds dans les phrases). Pour les propositions non-finies, les infinitives et les formes en *-ing* gagneraient à être distinguées, particulièrement pour l'analyse des constructions des francophones.

Sur la base de l'ensemble des métriques mobilisées dans nos analyses (52), la précision obtenue est assez faible (43%). La tâche est plus complexe avec l'ensemble des traits.

	FR1	FR2	GB
FR1	**12**	8	0
FR2	6	**9**	5
GB	5	10	**5**

TABLE 4 : Matrice de confusion des trois groupes de locuteurs (toutes métriques confondues)

4.2 Prolongements LNRE

La distribution de la loi de Zipf vaut également pour les corpus oraux, mais, chez les apprenants, à condition de disposer de suffisamment de corpus, les événements rares (hapax, mots rares) ne seront pas si fréquents que cela et d'autres éléments sont susceptibles d'être répétés.

Nous avons signalé que l'analyse a été conduite par courbe représentant un groupe et n'avons pas randomisé les occurrences attribuables aux différents locuteurs (nous n'avons donc pas neutralisé la variation inter-locuteurs). Une approche des courbes de vocabulaire sujet par sujet, dans une version plus subtile faisant intervenir la randomisation, pourrait donner des résultats différents. Nous n'avons pris en compte la variation inter-locuteur que dans les cinq relevés du nombre d'hapax par tranche de 100 occurrences. Le vocabulaire des apprenants étant plus limité, il nous a paru intéressant d'en mesurer les effets sur la raréfaction des mots nouveaux au fur et à mesure de l'évolution du discours. A titre expérimental, nous avons comptabilisé l'accroissement des hapax

par tranches de 100 tokens. Sur les cent premiers mots, on mesure le nombre d'hapax, sur les tranches suivantes, le différentiel d'hapax entre deux tranches.

La classification sur ces valeurs initiales des courbes d'accroissement du vocabulaire (*option leave one out*) ne donne que 40% de précision (Cf Table 5). La différence de longueur des productions reste l'indice le plus clair de distinction entre groupes de locuteurs, c'est ce qui explique la bonne classification du groupe FR1 : les valeurs cumulées sont plus souvent plus faibles au-delà de 400 mots pour le nombre d'hapax de la dernière tranche d'occurrences, car il y a parfois moins de 400 mots. Contrairement à ce que l'on pourrait penser, les locuteurs natifs n'ont pas une expression plus variée que les non-natifs du groupe FR2, ce qui explique leur très mauvaise classification. La capacité à produire régulièrement des mots nouveaux ne permet pas à elle seule une identification automatique de locuteurs avancé ou natifs. Les locuteurs avancés sont (mal) classés comme des locuteurs moins avancés car ils ont des scores avoisinants. En clair, c'est la différence de *fluency* qui est la plus manifeste entre les groupes de locuteurs, ce que reflète le nombre de mots prononcés dans les deux minutes que dure l'exécution de la consigne.

	FR1	FR2	GB
FR1	**16**	3	1
FR2	17	**2**	1
GB	6	8	**6**

TABLE 5 : Matrice de confusion des trois groupes de locuteurs (méthode des hapax)

Les limites en expression des apprenants pourraient avoir une incidence sur l'occurrence des événements rares au profit de formules récurrentes rassurantes type 'doudous' ('*teddy bear', cf.* Hasselgren 1994 ; Ellis 2012) ou de séquences préfabriquées, « reliability islands » (Dechert 1983). Ce type d'approfondissement gagnerait à envisager l'analyse en n-grams et, surtout, l'analyse de corpus longitudinaux de taille plus conséquente pourrait révéler des courbes de croissance du vocabulaire plus rapidement asymptotiques que pour les natifs. Reste à voir si le discours des apprenants (Gries & Ellis 2015) se laisse simplement régler par les paramètres alpha, A et C (Baayen 1991, Evert & Baroni 2004, Baayen 2008) des modèles LNRE. Le package {zipfR} permettrait d'essayer d'extrapoler les courbes de croissance du vocabulaire pour tester la robustesse des modèles LNRE et sans doute affiner les paramètres alpha, A et C au sein des modèles LNRE pour la modélisation des courbes de croissance selon les groupes de locuteur, voire selon les locuteurs et enfin évaluer la pertinence des trois modélisations possibles à partir de leurs trois implémentations dans le package {zipfr} (lnre.gigp; décrite dans Baayen 2001) Generalized Inverse Gauss Poisson Zipf-Mandelbrot (lnre.zm;Evert, 2004) et finite Zipf-Mandelbrot (lnre.fzm; Evert, 2004). Voici les spectres fréquentiels des trois groupes (toujours en effectifs cumulés, représentés en ordonnée). La première série d'histogrammes à gauche correspond aux hapax, puis aux mots ayant deux occurrences dans le corpus, et ainsi de suite.

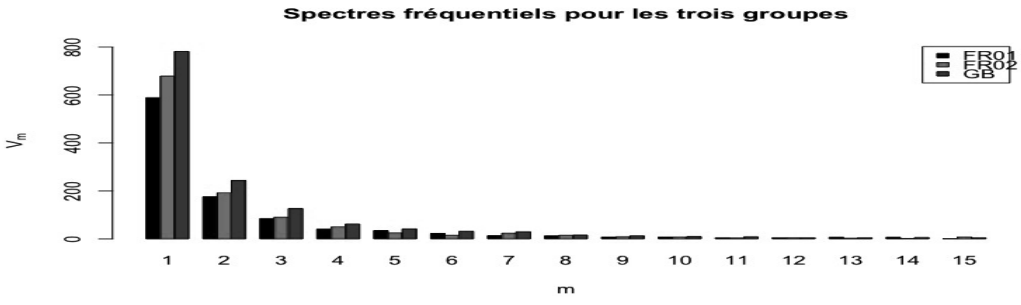

FIGURE 2 : Spectres fréquentiels des trois groupes (effectifs des 20 locuteurs cumulés)

4.3 Approche davantage qualitative : les profils lexicaux (Laufer & Nation 1995)

Il existe une version plus élaborée que les deux mesures ls1 et ls2 pour juger de la rareté des mots employés. Le *Lexical Frequency profile* de lextutor.ca (inspiré de Laufer & Nation 1995) procède par échantillonnage des productions selon quatre catégories de fréquences de référence (inventaire des 1 000 premiers mots les plus fréquents, inventaire des 2 000 mots suivants les plus fréquents, familles lexicales du vocabulaire universitaire et domaine « hors liste »). La composition de chaque production est alors analysée comme l'addition des quatre pourcentages de ces quatre « inventaires » lexicaux. Cette approche davantage qualitative donnerait peut-être de meilleurs résultats, étant donné la nature du corpus. Bien des locuteurs étant des personnels universitaires, le pourcentage de mots « académiques » est conforme au personnel recruté pour les locuteurs du groupe FR1. La fréquence des mots employés semble assez judicieuse en première approximation. En raison de nombreux problèmes de lemmatisation et de mots non-reconnus (*Aix, Aix-en-Provence*), nous n'avons pas pris en compte des résultats obtenus sur l'interface en ligne http://www.lextutor.ca/freq/train/. Les critiques adressées à cette méthodologie dont (Lu 2014) se fait l'écho laissent entendre que la discrimination n'est vraiment opératoire que pour un certain volume de données. On voit la difficulté posée par l'absence de base de données de référence.

4.4 Productions des apprenants à l'oral et fréquences de référence

De manière plus générale, il convient d'établir des inventaires fréquentiels de référence pour les non-natifs. Ce type de recherche devient pressant, dans la mesure où les travaux fondés sur l'usage sont obligés de se rabattre sur les bases de données des fréquences lexicales établies sur des corpus natifs (pour l'anglais, c'est souvent le CELEX, cf. Baayen et al. 1994). Pour l'analyse des productions d'apprenants, le paramètre de la tâche devra être prise en compte : le nombre de locuteurs qui se sont appuyés dans leur monologue sur les éléments du décor est impressionnant, comparés aux natifs, qui ont abordé des sujets plus divers.

Pour les corpus d'apprenants oraux, une difficulté supplémentaire réside dans le statut non-lexical sans doute trop facilement accordé aux pauses pleines. Parmi les points qui nous apparaissent les plus déterminants, les choix de transcrire les pauses par des formes distinctes interrogent à la fois la structuration de l'énoncé, la réalisation des voyelles et le niveau de l'apprenant. Il se trouve que les pauses pleines sont plutôt codées « heu » ou « euh » pour les apprenants FR1, «hum » pour les FR2 et « erm » pour les natifs. Sans même nous livrer à une analyse précise des réalisations phonétiques de leurs traits acoustiques (voir Chlébowski 2015 pour une analyse sur l'anglais de Newcastle), la

qualité de la réalisation de la voyelle diffère. Une observation un peu fine des pauses pleines et de leur transcription textuelle dans le corpus ANGLISH révèle la complexité de ce phénomène. Il y a véritablement toute une réflexion à conduire, tant sur le statut de tokenisation des réalisations phonétiques dans les corpus oraux (Ballier 2016) que sur leur signification (Käger 2016). Ceci dépasse le cadre de cet article, mais on peut songer que les niveaux des apprenants et les strates d'interlangue pourraient également s'analyser en termes de maîtrise de la position des articulateurs, et notamment de la langue (les *articulatory settings* de Catford 1984). On assisterait à une réalisation davantage labialisée par défaut chez les francophones moins avancés (le « *heu* » des FR1), à des réalisations plus proches des natifs en termes de timbre vocalique ou de recours au trait nasal chez les FR2 (le « *hum »),* et à une position de la langue plus en avant et une labialisation moins marquée chez les anglophones (« *erm* »).

L'examen de l'ordre dans la fréquence des occurrences (loi de Zipf) des pauses pleines souligne les problèmes de lemmatisation des pauses pleines dans une perspective interlangue (le statut et la fréquence de *heu, euh, hum*), sans compter les cas de réalisations clitiques (telles *but ahem* réalisé en deux syllabes). *Erm* est au 6e rang des occurrences chez les natifs (après les classiques *the, and, I, to, a*), *hum* est au sixième rang chez les locuteurs avancés FR2 et les transcriptions retenues pour les pauses pleines des FR1 se répartissent entre *hum* (8^{e} rang), *heu* (15^{e} rang) et *euh* (18^{e} rang). Les effectifs cumulés situant les occurrences de pause pleine au troisième rang des tokens des locuteurs FR1. Plutôt que de voir dans cette distribution quasi-complémentaire des pauses (*erm/hum/heu*) des incohérences transcriptionnelles, on peut y voir des variantes sémantiquement significatives des pauses pleines dont le rôle éventuel doit être analysé, des marqueurs du discours dont la contribution à la structuration de la hiérarchie des constituants prosodiques ne sont sans doute pas réductibles à la T-unit. (Paris 2015 :203-5) montre les limites de la notion de T-unit pour l'analyse de l'oral et notamment pour la prise en compte des faux-départs, des incises, du discours rapportés ou de ce qu'elle nomme des « propositions concaténées » dans l'analyse de ses enregistrements. Pour des données orales, le constituant prosodique inter-pauses pleines serait une unité plus pertinente que la T-unit. Manque cruellement une métrique de complexité à l'oral davantage fondée sur la prosodie, quelle que soit son empan dans la hiérarchie prosodique (Nespor & Vogel 2007).

4.5 Effets du genre

Le corpus ANGLISH, étant contrôlé, il y a 10 hommes et 10 femmes dans chaque groupe de locuteurs. Nous avons essayé de voir si les métriques rendaient compte de ces différences entre hommes et femmes. La tâche est plus complexe, mais la matrice de confusion pour les niveaux et les genres donne les résultats suivants, avec l'ensemble des métriques considérées, pour une précision de 23% seulement :

	FFR1	FFR2	FGB	HFR1	HFR2	HGB
FFR1	5	2	0	2	1	0
FFR2	2	4	1	3	0	0
FGB	1	2	0	2	4	1
HFR1	5	4	0	0	1	0

HFR2	0	2	3	1	3	1
HGB	1	2	2	2	1	2

TABLE 6 : Matrice de confusion par niveau et genre des groupes de locuteurs

4.5 Conclusions

Le projet d'ensemble est celui d'une analyse multidimensionnelle des productions orales des apprenants. On cherche, entre autres propriétés, à établir des corrélations ou des distinctions strictes entre les compétences lexicales, syntaxiques et phonologiques des apprenants. La comparaison n'est ici que partielle et a concerné la partie spontanée du corpus, là où la thèse d'Anne Tortel a porté sur les parties lues du corpus ANGLISH. Pour le moment, ce corpus dans sa dimension phonétique n'a été étudié qu'à partir de sa composante lue (Tortel 2008, Ballier et al. 2016), afin de favoriser les comparaisons inter-sujets. À l'inverse, le travail sur la production libre en spontané permet le type de raffinement que nous venons de proposer. Soulignons donc une fois encore l'intérêt de ce corpus ANGLISH et la diversité des réalisations qui s'y donne à voir et à entendre. En particulier, les productions des hommes et des femmes ne sont sans doute pas équivalentes au plan de la prosodie, alors que la classification ne semble pas l'établir aussi nettement.

Le deuxième point à faire valoir porte sur la diversité des technologies maintenant abordables pour l'étude des corpus d'apprenants (Díaz et al. 2013). Comme l'écrit (Lu 2014), il existe maintenant une panoplie d'outils qui constituent un terrain d'entente entre linguistes et programmeurs : "something in the middle ground, something that enables novice language and linguistics researchers to use more sophisticated and powerful corpus annotation and analysis tools than concordancing programs and yet still does not require programming".

En première analyse, sur des échantillons de données assez restreints, les métriques de complexité lexicale et syntaxique ne se substituent pas à des tests initiaux d'évaluations du niveau des apprenants. La taille restreinte du corpus est évidemment un obstacle, ce qui nous a conduits à multiplier les approches quantitatives par métriques et les différentes métriques possibles dans ce qui reste une expérimentation à très petite échelle. La taille restreinte pourrait être un avantage pour conduire également des approches plus qualitatives fondées sur les récurrences des n-grams. Reste que la piste automatique n'est pas complètement satisfaisante, même s'il conviendrait pour la réfuter complètement sur le jeu de données étudié de refaire une classification du niveau relatif des différents locuteurs francophones sur la base d'évaluations d'experts. C'est notre prochaine étape, en plus de la réplication de la méthodologie sur un corpus longitudinal de 135 locuteurs.

Remerciements

Nous remercions chaleureusement Anne Tortel pour la mise à disposition de son corpus, déposé sur le SLDR. Nous remercions Xiaofei Lu pour ses explications et ses démonstrations de ses solutions logicielles et Stefan Evert et Marco Baroni pour la clarté de leur documentation du package {ZipfR} et de son site dédié http://zipfr.r-forge.r-project.org ainsi que, pour leurs commentaires, les participants du workshop du 30 mars (Paris Diderot) et les relecteurs anonymes de TALN.

Références

BALLIER N. MARTIN, PH. (2015). Speech annotation of learner corpora. In: Granger, S., Gilquin, G., Meunier, F., (eds), *The Cambridge Handbook of Learner Corpus Research*, Cambridge: Cambridge University Press.

BALLIER N. (2016). Du dictionnaire lexico-phonétisé aux corpus oraux, quelques problèmes épistémologiques pour l'école de Guierre. *Histoire Epistémologie Langage*, à paraître.

BALLIER N., MARTIN, PH. AMAND, M. (2016). Variabilité des syllabes réalisées par des apprenants de l'anglais, *JEP 2016*, Paris, 8 pages.

BAAYEN R.H. (2001). *Word Frequency Distributions*. Dordrecht, Boston & London: Kluwer.

BAAYEN R.H. (2008). *Analysing Linguistic Data with R*. Cambridge : CUP.

BARONI M. (2009). Distributions in Text. In A. Lüdeling, M. Kytö (eds), *Corpus Linguistics. An International Handbook*, 803–821. Berlin, New York: de Gruyter Mouton.

BARONI M. EVERT, S. (2014). The zipfR package for lexical statistics: A tutorial introduction , zipfR version 0.6-7.

BENTZ C., BUTTERY P. (2014). Towards a Computational Model of Grammaticalization and Lexical Diversity. Proceedings of 5th *Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)*, 38-42.

CHITEZ M. (2014). *Learner corpus profiles: The case of Romanian Learner English*. Oxford : Peter Lang.

CHLEBOWSKI A. (2015*). Nasal Grunts » in the NECTE corpus An experimental investigation*. Mémoire de M2, Université Paris Diderot.

DAELEMANS W., ZAVREL, J., VAN DER SLOOT, K., & VAN DEN BOSCH A. (2004). Timbl: Tilburg memory-based learner. *Tilburg University*.

DAELEMANS W., VAN DEN BOSCH A. (2005). *Memory-based Language Processing*. Cambridge: CUP.

DECHERT, H.W., 1983, How a story is done in a second language, in: Farch, C. & Kasper, G. 1983d, Strategies in interlanguage communication, London, Longman., 175-196.

ELLIS, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*, 17-44.

DIAZ-NEGRILLO A., BALLIER N., THOMPSON P. (eds.). (2013). *Automatic treatment and analysis of learner corpus data*. (Studies in Corpus Linguistics 59). Amsterdam: Benjamins.

DRAGER, K. (2016). Constructing style: phonetic variation in quotative and discourse particle *like* In Heike Pichler (ed). *Discourse-Pragmatic Variation and Change in English:New Methods and Insights*, Cambridge : CUP, 232-251.

EVERT, S., & BARONI, M. (2007). zipfR: Word frequency distributions in R. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions,* 29-32. Association for Computational Linguistics.

EVERT, S., BARONI, M., (2006). The zipfR package.*http://cran.r-project.org/doc/packages/zipfR.pdf*

GAILLAT, T. (2013). Annotation automatique d'un corpus d'apprenants d'anglais avec un jeu d'étiquettes modifié du Penn Treebank. Actes de *la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, 271–284.

GRIES, S. T., & ELLIS, N. C. (2015). Statistical Measures for Usage-Based Linguistics. *Language Learning*, *65* (S1), 228-255.

HAWKINS J., BUTTER P. (2010). Criterial Features in Learner Corpora: Theory and Illustrations. *English Profile Journal* 1(01), 1-23.

HAWKINS J. A., FILIPOVIĆ L. (2012). *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework*. United Kingdom: Cambridge University Press.

HASSELGREN, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, *4*(2), 237-258.

HUNT K.W. (1965). *Grammatical Structures Written at Three Grade Levels*. Champaign IL: National Council of Teachers of English.

HUNT K.W. (1970). Do Sentences in the Second Language Grow Like Those in the First? *TESOL Quarterly* 4(3), 195–202.

JOHNSON, K. (2008). *Quantitative Methods in Linguistics*. Londres :Blackwell.

KAMEEN P.T. (1979). Syntactic Skill and ESL Writing Quality. In C. Yorio, K. Perkins, J. Schachter (eds), *On TESOL '79: The Learner in Focus,* 343–364. Washington DC: TESOL.

KLEIN, D., & MANNING, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1* (pp. 423-430). Association for Computational Linguistics.

LAUFER, B., et NATION, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16 (3), 307-322

LECLERCQ P., EDMONDS A., HILTON H. (Eds.). (2014). *Measuring L2 Proficiency: Perspectives from SLA*. Bristol: Multilingual Matters.

LU X. (2010). Automatic Analysis of Syntactic Complexity in Second Language Writing. *International Journal of Corpus Linguistics* 15, 474–496.

LU X. (2011). A Corpus-based Evaluation of Syntactic Complexity Measures as Indices of College-level ESL Writers' Language Development. *TESOL Quarterly* 45, 36–62.

LU X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal* 96, 190–208

LU X. (2014). *Computational Methods for Corpus Annotation and Analysis*, Dordrecht: Springer.

LU, X. (2016). L2 Syntactic Complexity Analyzer, http://www.personal.psu.edu/xxl13/download.html, consulté le 20/04/2016

NESPOR, M., & VOGEL, I. (2007). *Prosodic Phonology,* Berlin : Mouton de Gruyter.

OVTCHAROV V., COBB T., HALTER R. (2006). La richesse lexicale des productions orales: mesure fiable du niveau de compétence langagière. *Canadian modern language review* 63(1), 107-125.

PARIS, J. (2015). *Lumière sur le développement de la production de langage non-littéral en L2. Pour une comparaison avec l'acquisition des langues maternelles*, thèse non publiée, Paris 3.

TONO Y. (2013). Automatic Extraction of L2 Criterial Lexicogrammatical Features across Pseudo-longitudinal Learner Corpora: Using Edit Distance and Variability-based Neighbour Clustering. In C. Bardel, C. Lindqvist & B. Laufer (eds.), *L2 Vocabulary Acquisition, Knowledge and Use: New Perspectives on Assessment and Corpus Analysis*, 149–176. (Eurosla Monographs Series 2). The European Second Language Association.

TORTEL A. (2008). ANGLISH: Base de données comparatives de l'anglais lu, répété et parlé en L1 & L2, *Travaux Interdisciplinaires du Laboratoire Parole et Langage (TIPA)* 27, 111-122.

TORTEL, A. (2009). *Evaluation qualitative de la prosodie d'apprenants français: apport de paramétrisations prosodiques*. Thèse de doctorat non publiée. Aix-Marseille University.

TURENNE N. (2006). *Analyse de données textuelles sous R*. Paris : ISTE.

WOLFE-QUINTERO, K., INAGAKI, S. & KIM, H.-Y. 1998. *Second language deveopment in writing: Measures of fluency, accuracy and complexity* [Technical Report 17]. Honolulu: University of Hawaii, Second Language Teaching and Curriculum Center.

ZIPF G.K. (1949) *Human Behavior and the Principle of Least Effort*. Cambridge (Massachusetts): Addison-Wesley.