Formation de futurs enseignants à l'exploitation de corpus d'apprenants pour l'évaluation qualitative et quantitative de la production orale en FLE

Minerva Rojas

Formation de futurs enseignants à l'exploitation de corpus d'apprenants pour l'évaluation qualitative et quantitative de la production orale en FLE

Training of future teachers in the use of learner corpora for quantitative and qualitative assessment of oral production in French as a foreign language

Minerva Rojas

1. Introduction

1 La contribution de la linguistique de corpus à la didactique des langues est notable (Granger 2009), pourtant plusieurs chercheurs avancent que l'exploitation pédagogique et didactique de corpus d'apprenants est peu explorée (Meunier 2010, Gilquin *et al.* 2007, Mas & Gil 2018).

2 Un des domaines pour lesquels les corpus d'apprenants pourraient avoir une application pertinente est celui de l'évaluation des langues (Cushin 2017) car ils permettent de recueillir une quantité importante d'informations sur l'évolution des apprenants (Xi 2017). De plus, l'analyse quantitative de corpus d'apprenants peut compléter l'évaluation qualitative de la production en L2. Selon Weir (2005) des éventuels biais, que nous aborderons plus tard, surviennent lors de l'évaluation qualitative basée sur l'application d'échelles, comme celles du *Cadre Européen Commun de Référence pour les Langues* (CECRL) (Conseil de l'Europe, 2001, 2018).

3 En accord avec O'Keeffe & Farr 2003, l'utilisation de corpus d'apprenants pour l'évaluation des langues passe par la formation des futurs enseignants. Pour ce faire,

nous avons mis en place une expérience pilote avec des étudiantes de Master FLE de l'Université Côte d'Azur.

2. Cadre théorique

2.1 Corpus d'apprenants et applications didactiques

4 L'enseignement des langues secondes et l'étude de corpus de référence ont un lien étroit depuis la seconde moitié du XX^{e} siècle (Kennedy 1992). À titre d'exemple, les contenus lexicaux et grammaticaux du *Français fondamental* (Gougenheim *et al*. 1964) ont été définis à partir de l'analyse de fréquence d'un corpus de référence de plus de 800 000 mots. À l'heure actuelle, la place des corpus dans l'enseignement des langues est largement soutenue et justifiée car ils constituent des données authentiques (McEnery & Xiao 2011) permettant à l'apprenant de se confronter à différents usages, régularités et exceptions de la langue cible (Osborne 2002).

5 Dans le cas de l'enseignement du FLE, Boulton signalait dès 2008 que l'exploitation de corpus de référence et l'enseignement basé sur les données (*data-driven learning*) étaient peu répandus. Cependant on constate un intérêt croissant depuis les dix dernières années tant dans l'exploitation de corpus de référence pour l'enseignement du FLE (Auzéau & Abiad 2018, André 2020, Bilger & Cappeau 2013, Cavalla 2019, Di Vito 2013, Giuliani & Hannachi 2010, Sockett 2014, Tran *et al.* 2018) que dans la création des corpus de données authentiques à visée didactique (corpus *FLEURON* (André 2017) ; projet *ClapiFLE* (Ravazzolo & Etienne 2019)).

6 L'exploitation pédagogique des corpus d'apprenants ne semble pas avoir eu le même rayonnement que l'utilisation de corpus de référence ou des corpus de documents authentiques, même s'il existe des travaux en enseignement de l'anglais (Gilquin *et al*. 2009) ou de l'espagnol (Mas & Gil 2018). Plusieurs raisons pourraient en être la cause, telles qu'une réticence à utiliser les ordinateurs (Boulton 2008), un sentiment de distance entre les enseignants de L2 et les chercheurs, ou simplement le fait que les enseignants ignorent leur existence (Meunier 2012).

7 Pourtant les corpus d'apprenants contribuent principalement à la construction des théories sur l'acquisition car leur analyse aide à tracer le développement de la L2 (Larsen-Freeman & Cameron 2008, De Cock & Tyne 2014). Or, l'exploitation de ce type de corpus en enseignement des langues reste faisable. Granger et Lefer (2017) identifient la remédiation d'erreurs grammaticales et pragmatiques comme application. En outre, la consultation par des apprenants des corpus comparables de locuteurs experts peut les aider à contraster leurs productions afin de trouver des formes lexicales imprécises ou des mauvaises usages collocationnels ou encore les expressions figées (Cushing 2017).

8 Xi (2017) suggère que l'évaluation de la production en L2 à l'aide de corpus d'apprenants est un domaine fertile, ouvrant la voie aux enseignants pour fournir un retour détaillé dans le cadre de l'évaluation formative. En effet, l'étude de corpus constitue un bon moyen d'évaluation de la L2 car elle permet d'analyser de manière précise la production. De plus, Erard et Schneuwly (2005), soulignent que l'exploitation des corpus d'apprenants permet aux apprenants mêmes d'analyser leurs réussites et leurs difficultés. En outre, l'analyse de corpus d'apprenants peut être utile comme

complément des jugements qualitatifs effectués lors de l'évaluation de la production orale basée sur les échelles du CECRL (Cushing 2017, Mas & Gil 2018, Palacios *et al.* 2022).

2.2 L'évaluation de la production orale et les échelles qualitatives

9 L'évaluation de la production orale en L2 est devenue un objet de recherche à part entière depuis l'apparition de l'approche communicative (Bachman & Palmer 1996). Encore aujourd'hui, il existe des divergences sur les tâches communicatives déclenchant la production orale ou écrite objet de l'évaluation, sur les dimensions langagières à évaluer et sur la création de différents types d'échelle de mesure de la production (Bordón 2015). La création d'échelles découle de la nécessité de donner une note aussi « objective » que possible lors de l'évaluation (Alderson 1991) et leur application en évaluation des langues se répand après la deuxième guerre mondiale (voir Fulcher 2003, Hamps-Lyons 2016). D'après Figueras *et al.* (2008), depuis les années 2000 une des échelles des plus influentes est celle du CECRL (Conseil de l'Europe 2001, 2018), élément unificateur de l'évaluation proposant 47 échelles analytiques dans sa version de 2018, retenue pour cette étude.

10 Il existe deux types d'échelles d'évaluation : celles analytiques et celles holistiques. L'évaluation à l'aide d'une échelle holistique évalue la production orale de manière globale sans distinguer différents aspects (Luoma 1994). En revanche, l'évaluation à l'aide d'une échelle analytique identifie différentes caractéristiques de la production orale et les analyse séparément, procurant ainsi une gamme plus riche d'informations pour l'évaluation (Taylor & Galaczi 2011). Il y a un accord pour affirmer que les échelles analytiques sont plus adéquates pour l'évaluation en L2 par rapport aux échelles holistiques (Hamp-Lyons 2016, Taylor & Galaczi 2011). Cela dit, quand on utilise des échelles holistiques il peut y avoir des instabilités quant à leur fiabilité[1] car il est possible que certains biais apparaissent du fait de facteurs liés à la fiabilité inter-juges (Alderson & Banerjee 2001, 2002). En effet, il a été constaté que les juges non-natifs sont plus inflexibles que les juges natifs (Hill 1997, Hyland & Anan 2006) ainsi que les moins expérimentés par rapport aux juges plus expérimentés (Weigle 1998). Il a été aussi démontré que les évaluateurs qui ne connaissent pas préalablement les apprenants sont plus sévères et que la formation en évaluation basée sur des échelles rend plus uniformes les jugements (McNamara 1996, Weigle 1998).

11 En outre, l'échelle doit comporter un nombre limité de niveaux pour ne pas distraire les évaluateurs ; plus de huit niveaux peut perturber le processus d'évaluation (Bachman & Palmer 1996). Enfin, le nombre de catégories à évaluer doit être limité car il peut affecter la concentration (Diez Bedmar 2012), « au-delà de quatre ou cinq catégories on est cognitivement saturé et [...] sept catégories constituent un seuil psychologique à ne pas dépasser. » (Conseil de l'Europe 2001 : 145).

12 En somme, bien que les échelles soient conçues pour uniformiser les évaluations, il existe certains aspects pouvant conditionner les résultats. Ainsi l'évaluation de la production orale fondée sur l'analyse quantitative de corpus peut être un complément d'une évaluation qualitative faite à partir du CECRL. Des études sur la production écrite en anglais L2 montrent comment des mesures textuelles aident à prédire les niveaux des apprenants (Crossley & McNamara 2012, Gaillat *et al.* 2021, Vajjala 2018), mais il existe peu d'études pour l'oral (Zechner & Evanini 2019), d'autant moins se penchant sur le FLE et sur des données longitudinales.

13 Pour explorer l'application des échelles du CECRL et des mesures textométriques à l'évaluation d'un corpus d'apprenants de FLE, nous avons développé une démarche auprès de futurs enseignants pour les former à l'évaluation de la production orale. Cette démarche avait pour objectif de répondre aux questions suivantes :

- L'analyse qualitative de la production orale basée sur des échelles du CECRL est-elle fiable ?
- L'analyse qualitative permet-elle de saisir l'évolution de la production orale en FLE ?
- L'analyse quantitative permet-elle de mesurer l'évolution de la production orale en FLE ?

3. Méthodologie

14 La méthodologie de cette étude comprend deux démarches. D'une part, l'analyse qualitative d'un échantillon de productions du dit corpus à l'aide des échelles du CECRL (Conseil de l'Europe, 2018) effectuée par un groupe d'étudiantes de Master 2 FLE. D'autre part, l'analyse quantitative d'un corpus de production orale en FLE (corpus SCFLE).

3.1 Le corpus SCFLE

15 Le corpus SCFLE[2] a été créé dans le cadre d'une recherche doctorale à l'Université Savoie Mont Blanc (Chambéry, France) visant l'étude longitudinale de la production orale en FLE (Rojas Madrazo 2020). Ce corpus compile les productions orales monologiques et dialogiques de 12 locuteurs de FLE et de 14 locuteurs francophones. Il est composé de 47 025 mots distribués en deux sous-corpus : le sous-corpus de productions de FLE contenant 22 397 mots (dont 6 785 mots en monologue et 15 612 mots en interaction) et le sous-corpus de natifs qui comprend 24 628 mots (dont 8 767 mots en monologue et 15 861 en interaction).

16 Tous les participants étaient inscrits dans la même licence, Langues Étrangères Appliquées à l'Université Savoie Mont Blanc. Les locuteurs de FLE étudiaient la susdite licence (2015-2018), spécialité Anglais-FLE. Le recueil longitudinal des données a été effectué en quatre vagues successives au cours des trois années de la licence : la première (t1) a été menée en octobre 2015 ; la deuxième (t2) en mars 2016 ; la troisième (t3) en mars 2017 et la dernière (t4) en novembre 2017. Lors de chacune des vagues, les participants ont effectué une tâche en interaction et une tâche en monologue (*story retelling*) déclenchée par des séquences vidéo. Ce sont les tâches monologiques de t1 et de t4 qui ont été retenues pour cette étude.

17 Les données orales ont été transcrites et analysées avec la suite de logiciels EXMARaLDA (Schmidt 2004) et CLAN (MacWhinney 2000). La transcription des productions a été effectuée avec *Partitur Editor* d'EXMARaLDA qui permet une transcription détaillée des phénomènes reliés à la fluidité énonciative (Segalowitz 2010) (pauses, hésitations, allongements phonémiques, répétitions et autres redémarrages).

3.2 Productions orales retenues pour l'analyse et présentation des locuteurs FLE

18 Nous avons sélectionné[3] 12 productions orales monologiques du corpus SCFLE dont 6 de la première collecte de données (t1) et 6 de la quatrième (t4). À chaque locuteur

correspondent deux productions, ce qui signifie un écart de deux ans entre chaque production. Celles-ci ont été divisées en deux lots et randomisées pour éviter un éventuel biais : le premier lot est composé de 3 productions de t1 et de 3 productions de t4, et le deuxième lot est constitué de 3 productions de t1 et de trois productions de t4.

3.3 Les évaluatrices

19 Le groupe d'évaluatrices est composé de sept étudiantes de Master 2 FLE à l'Université Côte d'Azur. Les étudiantes ont accepté que leurs évaluations fassent l'objet d'une recherche (fiche de consentement signée). Un questionnaire a permis le recueil de métadonnées relatives à leurs profils linguistiques (L1, L2 parlées), ainsi que leurs années d'expérience en enseignement et évaluation du FLE. Les évaluatrices ont des profils différents du fait de leurs origines : trois évaluatrices francophones, trois sinophones et une hispanophone. Aucune des évaluatrices n'a d'expérience en évaluation de l'oral en FLE.

3.4 Les variables en production orale

20 Nous avons sélectionné quatre variables à analyser qualitativement avec quatre échelles du CECRL (2018) : la production orale générale, l'étendue linguistique générale, l'étendue de vocabulaire et l'aisance, dont les deux dernières seront aussi analysées quantitativement.

3.4.1 Les échelles qualitatives du CECRL retenues

21 Les quatre échelles du CECRL (Conseil de l'Europe, 2018) utilisées portent sur i) la production orale générale évaluant globalement la production de l'utilisateur de la L2 (*Ib.*: 72) ; ii) l'étendue linguistique générale qui sert à juger globalement la compétence linguistique (*Ib.*: 137) ; iii) l'étendue de vocabulaire fait référence à « l'ampleur et la variété des mots et des expressions utilisés » (*Ib.*: 138) ; iv) et l'aisance, qui correspond à la « capacité à poursuivre une longue production » (*Ib.*: 151) (Annexe 1).

3.4.2 Les variables quantitatives

22 Les unités d'analyse quantitatives que nous avons analysées sont l'indice D de diversité lexicale (ou VocD) (Jarvis 2002) et la longueur moyenne des segments (LMS) (Kormos 2006).

23 L'indice D se calcule automatiquement dans le logiciel CLAN donnant comme résultat un coefficient allant de 0 à 125 ; plus le score est haut, plus la diversité lexicale est élevée. Ce coefficient s'obtient à partir d'une formule corrigée du TTR (*type token ratio*[4]) (McCarthy & Jarvis 2007) et est souvent utilisé dans la recherche en acquisition des langues pour illustrer le développement lexical dans la langue cible (David 2008, Jarvis 2002).

24 Pour l'analyse quantitative de l'aisance, nous avons utilisé la LMS (MLR en anglais : *mean length of runs*). Cette mesure exprime le nombre de mots qu'un locuteur peut produire sans hésiter[5] et est souvent utilisée pour mesurer l'évolution de la production orale en L2 (Kormos 2006, Segalowitz 2010). Elle est calculée en divisant le nombre total

de mots d'une production par le nombre de segments. Un segment correspond aux mots produits entre deux pauses de plus ou égal à 250 millisecondes.

$$LMS = \frac{nombre\ de\ mots}{nombre\ de\ segments}$$

25 Il faut signaler que dans le cas des deux variables, VocD et LMS, on peut s'attendre à ce que les résultats augmentent de t1 à t4 de manière générale et individuellement.

3.5 Le protocole d'évaluation qualitative et la formation des évaluatrices

26 Trois séances de deux heures ont été consacrées à la formation des évaluatrices. L'objectif était de porter l'attention sur l'utilisation des descripteurs du CECRL retenus, et sur les concepts véhiculés par ces descripteurs, et notamment la notion d'aisance. Nous avons aussi présenté les déclencheurs des tâches et les différents outils de transcription et d'analyse quantitative.

27 Le premier lot de productions a été envoyé en octobre 2021 et rendu en novembre 2021. Le deuxième lot a été adressé aux évaluatrices en novembre 2021 et rendu en décembre 2021. Les évaluatrices disposaient d'un *livret d'évaluateur* (Annexe 2), contenant une grille pour attribuer le niveau CECRL des quatre variables et le justifier. Dans le livret, une grille d'annotation de temps d'évaluation a été ajoutée pour que les étudiantes précisent le temps consacré à l'évaluation. Le même livret a été envoyé une deuxième fois, accompagnant le deuxième lot d'enregistrements, auquel nous avons ajouté une grille pour annoter les difficultés rencontrées.

3.6 Analyses

28 L'objectif des analyses qualitatives est de mesurer le degré d'accord inter-juges utilisant les échelles du CECRL comme instrument d'évaluation. Pour cela, nous avons calculé la distribution des jugements pour chaque échelle évaluée, et le coefficient kappa de Fleiss (Fleiss 1971). Celui-ci détermine la proportion dans laquelle plus de deux évaluateurs sont d'accord entre eux. Le coefficient affiche des valeurs allant de 0 à 1, 0 étant la valeur qui exprime le désaccord entre les juges et 1 la valeur exprimant l'accord total inter-juges[6].

29 Dans le cas des échelles portant sur la Production Orale Générale (POG) et l'Étendue Linguistique Générale (ELG), nous avons soumis aux analyses la totalité des jugements (7 évaluatrices). Pour l'analyse qualitative des deux échelles portant sur le vocabulaire et l'aisance, nous avons seulement retenu les jugements des 3 évaluatrices francophones afin d'éviter des éventuels biais dus à leurs origines.

30 L'objectif des analyses des mesures quantitatives (diversité lexicale, LMS) est de décrire l'évolution de la production orale chez chaque individu et dans l'ensemble du groupe. Les résultats individuels ont été soumis à des analyses statistiques descriptives (moyenne, médiane et écart-type) pour les mesures de tendance centrale et de dispersion. Les résultats ont ensuite été soumis au test de Wilcoxon afin de mesurer si

les changements survenus entre t1 et t4 dans l'ensemble de l'échantillon était statistiquement significatifs.

4. Résultats

4.1 Production orale générale et étendue linguistique générale

Tableau 1. Évaluation qualitative Production Orale Générale (POG) et Étendue Linguistique Générale (ELG)

Image 1001DB5C00003A23000032B2A5BAA534F0CF62C0.emf

		A1	A2	B1	B2	C1	C2	Niveaux attribués	Distance
Meg	POG t1	0	3	4	0	0	0	2	1
	POG t4	0	0	3	4	0	0	2	1
	ELG t1	0	5	2	0	0	0	2	1
	ELG t4	0	2	3	2	0	0	3	2
Ser	POG t1	0	1	4	1	1	0	4	3
	POG t4	0	0	4	3	0	0	2	1
	ELG t1	0	1	5	1	0	0	3	2
	ELG t4	0	0	5	2	0	0	2	1
Mat	POG t1	0	2	4	1	0	0	3	2
	POG t4	0	0	5	1	1	0	3	2
	ELG t1	0	2	5	0	0	0	2	1
	ELG t4	0	0	6	0	1	0	2	2
Pet	POG t1	0	1	4	2	0	0	3	2
	POG t4	0	2	5	0	0	0	2	1
	ELG t1	0	2	5	0	0	0	2	1
	ELG t4	0	2	5	0	0	0	2	1
Yoa	POG t1	0	1	4	2	0	0	3	2
	POG t4	0	2	5	0	0	0	2	1
	ELG t1	0	2	5	0	0	0	2	1
	ELG t4	0	2	5	0	0	0	2	1
Uya	POG t1	0	3	3	1	0	0	3	2
	POG t4	0	0	4	3	0	0	2	1
	ELG t1	0	2	4	1	0	0	3	2
	ELG t4	0	0	5	2	0	0	2	1

31 Concernant les deux descripteurs POG et ELG, nous constatons qu'il n'existe pas d'accord qui fasse l'unanimité (tableau 1), ni pour les évaluations des productions correspondant à 2015 (t1), ni pour les évaluations des productions correspondant à 2017 (t4). Cela est observé dans la colonne *Niveaux attribués*, recueillant le nombre de niveaux cumulés pour un seul cas ; par exemple, la production de l'apprenante Meg a été classée A2 et B1 pour la POGt1 et B1 et B2 pour la POG t4.

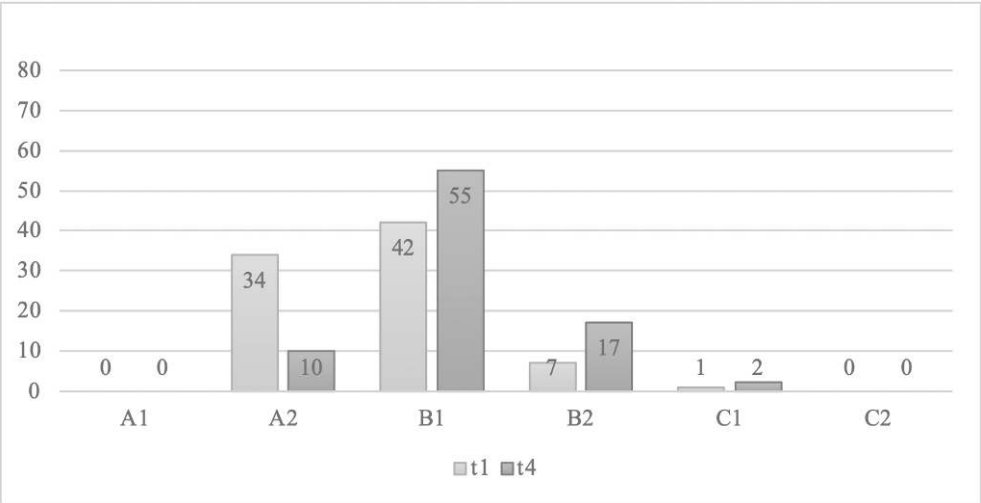
32 Les désaccords des jugements de POG en t1, POG en t4, ELG en t1 et ELG en t4 sont corroborés par les résultats du taux kappa (tableau 2). Pour POG en t1, l'accord inter-juges est presque nul (k = .029) ce qui survient aussi en POG t4 (k = .042). Pour ELG en t1, le degré d'accord entre les juges est très faible (k = 0,103) et diminue en t4 car le degré diminue étant les évaluatrices en désaccord total (k = -0.002).

Tableau 2. Concordance globale. Évaluations Production Orale Général (POG) et Étendue Linguistique Générale (ELG)

Concordance globale						
		Asymptotique			Intervalle de confiance asymptotique à 95 %	
	Kappa	Erreur standard	z	Sig.	Limite inférieure	Limite supérieure
POG t1	.029	.064	.448	.654	-.097	.155
POG	.042	.067	.632	.527	-.088	.173
ELG t1	.103	.079	1.309	.191	-.051	.257
ELG t4	-.002	.063	-.035	.972	-.126	.121

33 En rassemblant les jugements des deux échelles (POG et ELG, 168 jugements dont 84 pour t1 et 84 pour t4), nous constatons que l'ensemble du groupe d'apprenants cumule des niveaux plus élevés dans les évaluations des productions en t4 par rapport aux productions en t1 (figure 1). Le niveau A2 représente 40 % des jugements des productions t1 contre 12 % des productions t4 ; le niveau B1 passe de 51 % à 66 % de t1 à t4 ; enfin, le niveau B2 augmente de 12 points, de 8 % en t1 à 20 % en t4. Il faut remarquer que le niveau B1 reste le niveau qui cumule le pourcentage le plus élevé, 58 % au total (t1 et t4), (97 des 168 jugements effectués). Malgré les désaccords inter-juges constatés et de manière générale, les évaluations qualitatives reflètent l'évolution longitudinale de la production orale chez les locuteurs de FLE ; de manière très claire dans le cas de Meg (POG), de Yoa (POG et ELG) et de Mat (ELG) (voir tableau 1).

Figure 1. Distribution des niveaux du CECRL attribués à la Production Orale Générale et à l'Étendue Linguistique Générale



4.2 Étendue de vocabulaire et aisance

Tableau 3. Évaluation qualitative de l'étendue de vocabulaire

		A1	A2	B1	B2	C1	C2	niveaux attribués
Meg	ét. de vocabulaire t1		2	1				2
	ét. de vocabulaire t4				3			1
Ser	ét. de vocabulaire t1			2	1			2
	ét. de vocabulaire t4				3			1
Mat	ét. de vocabulaire t1		1	2				2
	ét. de vocabulaire t4			3				1
Pet	ét. de vocabulaire t1		3					1
	ét. de vocabulaire t4			3				1
Yoa	ét. de vocabulaire t1		2	1				2
	ét. de vocabulaire t4			1	2			2
Uya	ét. de vocabulaire t1		1	2				2
	ét. de vocabulaire t4			3				1

34 Pour l'évaluation qualitative portant sur le vocabulaire et l'aisance, seuls les jugements des trois évaluatrices francophones ont été retenus. Les jugements de l'étendue de vocabulaire (tableau 3) présentent des désaccords en t1 dans tous les cas, sauf dans le cas de Pet où elle obtient l'accord total des trois juges. Les jugements des productions t4 ne coïncident pas pour cinq apprenants. Chez tous les apprenants l'étendue de vocabulaire est jugée plus élevée dans les productions t4.

35 Ces observations sont corroborées par les résultats de kappa (tableau 4) ; en t1 le degré d'accord entre les évaluatrices est presque nul (0.062), et augmente en t4 à un accord faible ($k = 0.364$) qui est significatif statistiquement ($p < 0.05$).

Tableau 4. Concordance globale. Évaluation de l'étendue de vocabulaire

Concordance globale						
		Asymptotique			Intervalle de confiance asymptotique à 95 %	
	Kappa	Erreur standard	z	Sig.	Limite inférieure	Limite supérieure
EVoc t1	.062	.180	.347	.729	-.291	.416
EVoc t4	.364	.175	2.073	.038	.020	.707

36 Quant aux analyses quantitatives de la diversité lexicale (tableau 5), de manière générale (chez 4 apprenants sur 6) l'indice D augmente de t1 à t4. La moyenne de l'échantillon augmente de 16,63 points en t4, et la médiane augmente de 6,59 points. Or, il faut signaler qu'il existe une variabilité visible entre les apprenants, vu les valeurs de l'écart type (SD) qui augmentent aussi en t4. Bien que les valeurs de D augmentent, le test de Wilcoxon ($z = -1.15$) révèle que les différences entre t1 et t4 ne sont pas significatives.

Tableau 5. Résultats VocD, diversité lexicale. Résultats individuels, tendances centrales de l'échantillon et test Wilcoxon

							Statistiques descriptives				Wilcoxon	
	Meg	Ser	Mat	Pet	Yoa	Uya	N	M	Mdn	SD	Z	Sig. Asint.
VocD_t1	58.52	31.32	34.01	61.14	49.10	58.78	6	48.811	53.810	13.196		
VocD_t4	53.36	77.84	57.25	63.55	94.11	46.54	6	65.441	60.400	17.604	-1.15	.249

37 Quant à l'évaluation qualitative de l'aisance (tableau 6), on constate un meilleur accord dans les jugements des productions en t4 que t1. En t4, 4 apprenants sont classés au même niveau d'aisance par les trois évaluatrices.

Tableau 6. Évaluation qualitative de l'aisance

		A1	A2	B1	B2	C1	C2	niveaux attribués
Meg	aisance t1		1	1	1			3
	aisance t4				3			1
Ser	aisance t1			2	1			2
	aisance t4				3			1
Mat	aisance t1		1	2				2
	aisance t4			3				1
Pet	aisance t1			3				1
	aisance t4			3				1
Yoa	aisance t1		2	1				2
	aisance t4		1	2				2
Uya	aisance t1			3				1
	aisance t4			2		1		2

38 Concernant le degré d'accord inter-juges dans l'évaluation de l'aisance (tableau 7), en t1 il y a un désaccord total inter-juges (k = -0.012), en revanche en t4, les jugements présentent un accord fort (k = 0.613) et significatif statistiquement.

Tableau 7. Concordance globale. Évaluation de l'aisance

Concordance globale						
		Asymptotique			Intervalle de confiance asymptotique à 95 %	
	Kappa	Erreur standard	z	Sig.	Limite inférieure	Limite supérieure
Aisance t1	-.012	.182	-.069	.945	-.369	.344
Aisance t4	.613	.179	3.428	.001	.262	.963

39 Pour ce qui concerne les analyses quantitatives (tableau 8), de manière générale, la LMS augmente chez les apprenants, mais l'augmentation est moins importante que celle de l'indice de diversité lexicale. La moyenne augmente de 0,714 points, alors que la médiane diminue de 0,520 points. La variabilité entre les apprenants s'observe grâce aux valeurs de l'écart-type (SD). Les valeurs de LMS baissent entre t1 et t4 mais ne sont pas significatives avec le test de Wilcoxon (z = -0.73).

Tableau 8. Longueur Moyenne des Segments. Résultats individuels, tendances centrales de l'échantillon et test Wilcoxon

							Statistiques descriptives				Wilcoxon	
	Meg	Ser	Mat	Pet	Yoa	Uya	N	M	Mdn	SD	Z	Sig. Asint.
LMS_t1	6.41	6.00	6.09	6.37	3.28	2.37	6	5.086	6.045	1.782		
LMS_t4	7.10	4.89	5.36	6.40	5.42	5.63	6	5.800	5.525	.805	-.73	.463

40 En somme, le degré d'accord est moins élevé dans l'évaluation qualitative du vocabulaire. Les résultats quantitatifs montrent que la diversité lexicale est un indicateur de l'apprentissage du FLE car les valeurs augmentent de manière générale. En revanche, l'évaluation de l'aisance présente des accords plus élevés. Les analyses quantitatives montrent que la moyenne des valeurs de LMS augmente légèrement alors que la médiane diminue. Ce résultat suggère que la LMS n'est pas un indicateur de l'évolution de la production en FLE dans l'ensemble du groupe d'apprenants. Ces conclusions doivent être prises avec prudence étant donné la taille de l'échantillon et la variabilité entre les résultats individuels des apprenants.

5. Conclusions

41 Dans notre étude, nous avons essayé de répondre à trois questions portant sur l'évaluation qualitative et la mesure quantitative de la production orale en L2. Tout d'abord, afin de savoir si l'analyse qualitative basée sur les échelles du CECRL est fiable, nous avons constaté qu'il n'existe pas d'accord inter-juges ni sur l'analyse qualitative de la production orale générale (POG) ni sur l'étendue linguistique générale (ELG). L'utilisation des échelles s'avère peu solide comme seul instrument d'évaluation de la production orale en FLE. Il se peut que les divergences et l'absence d'accord sont dues aux origines des évaluatrices, pourtant, dans les livrets d'évaluation saisis par elles-mêmes, toutes signalent la difficulté de l'interprétation des descripteurs des niveaux au moment de placer les apprenants.

42 Malgré les désaccords entre les évaluatrices, la distribution des résultats des jugements de la POG et de l'ELG montre que les apprenants sont placés dans des niveaux supérieurs après deux ans d'immersion. Donc, concernant la deuxième question, l'analyse qualitative permettrait de saisir l'évolution de la production orale de manière générale dans l'ensemble du groupe d'apprenants.

43 Concernant l'analyse qualitative de l'étendue du vocabulaire et de l'aisance, il existe un désaccord généralisé inter-juges en t1 mais des accords significatifs dans les évaluations de t4. Ces observations suggèrent que ces paramètres de la production orale seraient perçus par les évaluatrices de manière plus variable, peut-être dû à des différences de niveau entre les apprenants en début de formation dans les premiers mois d'une immersion à long terme. On peut noter que l'origine des évaluatrices n'est pas la cause du désaccord en t1 étant donné que toutes sont francophones et ont suivi le même nombre d'heures de formation.

44 Pour ce qui concerne la troisième question de recherche, d'abord, l'analyse quantitative de la diversité lexicale permet de montrer l'évolution de l'utilisation du vocabulaire en production orale en FLE. En revanche, l'analyse quantitative de la longueur moyenne de segments n'est pas un indicateur solide pour mesurer l'évolution de l'aisance dans

l'ensemble du groupe. Il faut également évoquer que l'analyse de la fluidité est normalement effectuée à travers une large liste d'indicateurs (Segalowitz *et al.* 2017), mais nos résultats suggèrent que la fluidité présente une dimension subjective sujette à la perception de ce qu'est un locuteur fluent (Segalowitz 2010).

45 En somme, cette étude suggère que les échelles du CECRL (Conseil de l'Europe, 2018) sont des instruments qui peuvent donner lieu à des désaccords quand l'évaluation de la production orale générale et de l'étendue linguistique générale est effectuée par plusieurs évaluateurs. En revanche, l'utilisation d'autres échelles, telles que celles de vocabulaire ou d'aisance peuvent présenter un degré d'accord plus élevé, ce qui les rendraient plus fiables. L'analyse quantitative apparaît comme un complément d'évaluation utile même si les résultats en fin de séjour à long terme ne sont pas significatifs statistiquement.

46 En conclusion, nous voudrions ajouter que cette expérience a été très positive. Les étudiantes, qui ont suivi cette formation, se considèrent mieux préparées à évaluer la production orale et sont plus conscientes des enjeux en termes d'interprétation des données quantitatives et qualitatives. Il est vrai que la préparation de cette formation est laborieuse, mais nous pensons que cette démarche peut favoriser l'utilisation des corpus d'apprenants en enseignement et en évaluation des langues et contribuer à l'enseignement supérieur pour et par la recherche.

BIBLIOGRAPHIE

André V. (2020). « Faire de la linguistique de corpus avec des apprenants de français langue étrangère », *in* P. Larrivée & F. Lefeuvre, *La didactisation du français vernaculaire*. Caen : Presses Universitaires de Caen, 37-66.

André V. (2017). « Un corpus multimédia pour apprendre à interagir en situations universitaires en France », *Actes du troisième colloque international de l'ATPF. Enseigner le français : s'engager et innover*. Bangkok, 292-315.

Alderson J. (1991). « Bands and Scores », *in* J. C. Alderson & B. North (dir.) *Language testing in the 1990s : The communicative legacy*. Londres : Modern English Publications, 71-86.

Alderson J. & Banerjee J. (2001). « Language testing and assessment (Part 1) », *Language Teaching* 34 : 213-236.

Alderson J. & Banerjee J. (2002). « Language testing and assessment (Part 2) », *Language Teaching* 35 : 79-113.

Auzéau F. & Abiad L. (2018). « Le corpus : un outil inductif pour l'enseignement-apprentissage de la grammaire », *Synergies France* 12 : 175-187.

Bachman L.F. (1990). *Fundamental considerations in language testing*. Oxford : Oxford University Press.

Bachman L.F. & Palmer A.S. (1996). *Language testing in practice*. Oxford : Oxford University Press.

Biber D., Conrad S. & Reppen R. (1998). *Corpus linguistics : Investigating language structure and use*. Cambridge : Cambridge University Press.

Bilger M. & Cappeau P. (2013). « Comment les données de corpus pourraient renouveler les manuels de grammaire ? », *Linx* 68-69, en ligne : http://journals.openedition.org/linx/1526.

Boulton A. (2008). « Esprit de corpus : Promouvoir l'exploitation de corpus en apprentissage des langues », *Texte et corpus* 3 : 37-46.

Bordón T. (2015). « La evaluación de segundas lenguas (L2) », *Balance y perspectivas. Revista Internacional.*

Cavalla C. (2019). « Comment former les étudiants de Master FLE à l'utilisation pédagogique des corpus numériques ? », *in* J. Goes, L. Meneses-Lerin, J.-M. Mangiante, F. Olmo & C. Pineira-Tresmontant, *Apports et limites des corpus numériques en analyse de discours et didactique des langues de spécialité*. Editura Universitaria, 79-92.

Conseil de l'Europe (2001). *Un cadre européen commun de référence pour les langues : Apprendre, enseigner, évaluer.* Strasbourg : Conseil de l'Europe.

Conseil de l'Europe (2018). *Un cadre européen commun de référence pour les langues : Apprendre, enseigner, évaluer. Un volume complémentaire avec des nouveaux descripteurs.* Strasbourg : Conseil de l'Europe.

Crossley S. A. & McNamara D. S. (2012). « Predicting second language writing proficiency : The roles of cohesion and linguistic sophistication », *Journal of Research in Reading* 35(2) : 115-135. https://doi.org/10.1111/j.1467-9817.2010.01449.

Cushing S.T. (2017). « Corpus linguistics in language testing research », *Language Testing* 34(4) : 441-449.

David A. (2008). « A developmental perspective on productive lexical knowledge in L2 oral interlanguage », *Journal of French Language Studies* 18 : 315-331.

De Cock S. & Tyne H. (2014). « Corpus d'apprenants et acquisition des langues », *Recherches en didactique des langues et des cultures. Les cahiers de l'Acedle* 11(1). En ligne : 10.4000/rdlc.1716.

Delahaie J. (2013). « Constitution et exploitation de corpus d'interactions verbales pour le FLE : problèmes et programme », *Linx. Revue des linguistes de l'université Paris X Nanterre* 68-69 : 95-114.

Diez Bedmar (2012). « El uso del MCERL para evaluar las redacciones en el examen de inglés en las Pruebas de Acceso a la Universidad », *Revista de Educación* 357 : 55-80.

Di Vito S. (2013). « L'utilisation des corpus dans l'analyse linguistique et dans l'apprentissage du FLE », *Linx* 68-69. http:// journals.openedition.org/linx/1519.

Erard S. & Schneuwly B. (2005). « La didactique de l'oral : savoirs ou compétences », *in* J.-P. Bronckart, E. Bulea & M. Pouliot, *Repenser l'enseignement des langues : comment identifier et exploiter les compétences*. Presses universitaires du Septentrion, 69-97.

Figueras N., North B., Takala S., Verhelst N. & Van Avermaet P. (2005). « Relating examinations to the Common European Framework : A manual », *Language Testing* 22(3) : 261-279.

Fleiss J. L. (1971). « Measuring nominal scale agreement among many raters », *Psychological Bulletin* 76 : 378-382.

Fulcher G. (2003). *Testing Second Language Speaking*. Londres : Pearson/Longman.

Gaillat T., Simpkin A., Ballier N., Stearns B., Sousa A., Bouyé M. & Zarrouk M. (2021). « Predicting CEFR levels in learners of English : The use of microsystem criterial features in a machine learning approach », *ReCALL* 34(2). https://doi.org/10.1017/S095834402100029X.

Giuliani D. & Hannachi R. (2013). « Linguistique de corpus et didactique du F.L.E. Une exploitation du corpus IntUne », *Cahiers de praxématique* 54-55. http://journals.openedition.org/praxematique/1136.

Granger S. (2008). « Learner corpora », *in* A. Lüdeling et M. Kytö (dir.) *Corpus Linguistics. An International 5 Handbook. Volume 1*. Berlin/New York : Walter de Gruyter, 259-275.

Granger S. (2009). « The contribution of learner corpora to second language acquisition and foreign language teaching : A critical evaluation », in K. Aijmer (dir.) *Corpora and Language Teaching*. Amsterdam/Philadelphia : John Benjamins, 13-32.

Granger S. & Lefer M.A. (octobre 2017). « Bridging the gap between learner corpus research and translation studies : The Multilingual Student Translation corpus », *4th Learner Corpus Conference.* Bolzano.

Gilquin G., Granger S. & Paquot M. (2007). « Learner corpora : The missing link in EAP pedagogy », *Journal of English for Academic Purposes* 6(4) : 319-335.

Gougenheim G., Rivenc P., Michéa R., Sauvageot A. (1964). *L'élaboration du français fondamental (1er degré) : étude sur l'établissement d'un vocabulaire et d'une grammaire de base (Vol. 2)*. Didier.

Hamp-Lyons L. (2016). « Purposes of assessment », *in* D. Banerjee & J. Tsagari (dir.) *Handbook of Second Language Assessment*. De Gruyter Mouton, 13-27.

Hill K. (1997). « Who should be the judge ? The use of non-native speakers as raters on a test of English as an international language », *in* A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (dir.) *Current developments and alternatives in language assessment*. Jyväskyla : University of Jyväskyla, 275-290.

Hilton H., Osborne J., Derive M. J., Suco N., O'Donnell J., Rutigliano S. & Billard S. (2008). *Corpus PAROLE. Architecture du corpus et conventions de transcription*. Laboratoire LLS. Université de Savoie. http://archive.sfl.cnrs.fr/sites/sfl/IMG/pdf/PAROLE_manual.pd.

Hyland K. & Anan E. (2006). « Teachers' perceptions of errors : The effects of first language and experience », *System* 34 : 509-519.

Jarvis S. (2002). « Short texts, best-fitting curves and new measures of lexical diversity », *Language Testing* 19(1) : 57-84.

Kennedy G. (1992). « Preferred ways of putting things with implications for language teaching », *in* J. Svartvik (dir.) *Trends in linguistics. Studies and Monographs 65. Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82*. Berlin : Mouton De Gruyter, 335-378.

Larsen-Freeman D. L. & Cameron L. (2008). « Research methodology on language development from a complex systems perspective », *The Modern Language Journal* 92(2) : 200-213.

Luoma S. (2004). *Assessing speaking*. Cambridge : Cambridge University Press.

MacWhinney B. (2000). *The CHILDES Project : Tools for Analyzing Talk*. Mahwah, NJ : Lawrence Erlbaum.

Mas Álvarez I. & Gil Martínez A. (2018). « Los corpus de aprendices : un terreno en expansión para la enseñanza de español », *in* M. Ellison, M. Anido, P. Nicolás & S. Valente-Rodrigues, *As linguas estrangeiras no ensino superior : propostas didáticas e casos em estudo*. Porto : Universidade do Porto. Faculdade de Letras, 35-55.

McCarthy P. M. & Jarvis S. (2007). « Vocd : A theoretical and empirical evaluation », *Language Testing* 24(4) : 459-488.

McEnery T. & Xiao R. (2011). « What corpora can offer in language teaching and learning », *in* E. Hinkel (dir.) *Handbook of research in second language teaching and learning*. New York : Routledge, 382-398.

McNamara T. (1996). *Measuring Second Language performance*. Londres : Longman.

Meunier F. (2010). « Learner corpora and English language teaching : Checkup time », *Anglistik : International Journal of English Studies* 21(1) : 209-220.

Meunier (2012). « Learner corpora in the classroom: a useful and sustainable didactic resource », *in* L. Pedrazzini & A. Nava (dir.) *Learning and Teaching English : Insights from Research.* Milano : Poliletrica, 211-228.

O'Keeffe A. & Farr F. (2003). « Using Language Corpora in Initial Teacher Education : Pedagogic Issues and Practical Applications », *TESOL Quarterly* 37(3) : 389-418.

Osborne J. (2002). « Integrating corpora into a language-learning syllabus », *in* B. Lewandowska-Tomaszczyk (dir.) *PALC 2001 : Practical applications in language corpora.* Francfort : Peter Lang, 479-492.

Palacios Martínez I., Barcala F.M. & Rojo G. (2022). « El Corpus de Aprendices de Español (CAES) y sus aplicaciones para la enseñanza/aprendizaje del español como lengua extranjera », *in* M. Blanco, H. Olbertz & V. Vázquez Rozas (dir.) *Corpus y construcciones. Perspectivas hispánicas. Anejo 79 de Verba 2022*, 273-303.

Ravazzolo E. & Etienne C. (2019). « Nouvelles ressources pour le FLE à partir des études en interaction », *LINX* 79 : 177-199. doi.org/10.4000/linx.1526.

Rojas Madrazo M. (2020). *Stratégies de communication, fluidité et lexique en production orale. Étude longitudinale d'apprenants de FLE en immersion*. Thèse de doctorat. Université Savoie Mont Blanc. Chambéry.

Schmidt T. (2004). « Transcribing and annotating spoken language with EXMARaLDA », *in Proceedings of the LREC-Workshop on XML based richly annotated corpora, Lisbon 2004*. Paris : ELRA, 69-74.

Sockett G. (2014). « Corpus et perspectives pour l'enseignant : Compétences, formation, outils, besoins, activités, objectifs », *Recherches en didactique des langues et des cultures* 11(1). http://journals.openedition.org/rdlc/1691.

Taylor L. & Galaczi E. (2011). « Scoring Validity of speaking tests », *in* L. Taylor, *Studies in Language Testing : Examining Speaking*. Cambridge : Cambridge University Press, 171-233.

Tran T. T. H., Tutin A. & Cavalla C. (2016). « Pour un enseignement systématique des marqueurs discursifs à l'aide de corpus en classe de FLE : l'exemple des marqueurs de reformulation », *Linguistik online* 78(4) : 113-128.

Vajjala S. (2018). « Automated assessment of non-native learner essays : Investigating the role of linguistic features », *International Journal of Artificial Intelligence in Education* 28 : 79-105. https://doi.org/10.1007/s40593-017-0142-3.

Warrens M.J. & Pratiwi B.C. (2016). « Kappa Coefficients for Circular Classifications », *Journal of Classification* 33 : 507-522. https://doi.org/10.1007/s00357-016-9217-3.

Weigle S.C. (1998). « Using FACETS to model rater training effects », *Language Testing* 15 : 263-287.

Weir C. J. (2005). « Limitations of the Common European Framework for developing comparable examinations and tests », *Language testing* 22(3) : 281-300.

Xi X. (2017). « What does corpus linguistics have to offer to language assessment ? », *Language Testing* 34(4) : 565-577.

Zechner K. & Evanini K. (éd.). (2019). *Automated Speaking Assessment : Using Language Technologies to Score Spontaneous Speech*. New York : Routledge. https://doi.org/10.4324/9781315165103.

ANNEXES

Annexe 1.

Descripteurs et échelles utilisés pour l'évaluation des productions orales : production orale générale ; étendue linguistique générale ; étendue de vocabulaire et aisance à l'oral.

PRODUCTION ORALE	
PRODUCTION ORALE GÉNÉRALE	PROSIGN
C2	Peut produire un discours, clair, limpide et fluide, avec une structure logique efficace qui aide le destinataire à remarquer les points importants et à s'en souvenir.
C1	Peut faire une présentation ou une description claire d'un sujet complexe en intégrant des thèmes secondaires et en développant des points particuliers pour parvenir à une conclusion appropriée.
B2	Peut méthodiquement développer une présentation ou une description soulignant les points importants et les détails pertinents.
	Peut faire une description et une présentation détaillées sur une gamme étendue de sujets relatifs à son domaine d'intérêt en développant et justifiant les idées par des points secondaires et des exemples pertinents.
B1	Peut assez aisément mener à bien une description directe et non compliquée de sujets variés dans son domaine en la présentant comme une succession linéaire de points.
A2	Peut décrire ou présenter simplement des gens, des conditions de vie, des activités quotidiennes, ce qu'on aime ou pas, par de courtes séries d'expressions ou de phrases non articulées.
A1	Peut produire des expressions simples isolées sur les gens et les choses.
Pré-A1	Peut produire des phrases courtes pour parler de soi, donner des renseignements simples personnels (par exemple, le nom, l'adresse, la situation familiale, la nationalité).

ÉTENDUE LINGUISTIQUE GÉNÉRALE	PROSIGN
C2	Peut tirer profit d'une maîtrise globale et fiable d'une gamme très étendue de langue pour formuler précisément sa pensée, insister, discriminer et lever l'ambiguïté. Ne montre aucun signe de devoir réduire ce qu'il/elle veut dire.
C1	Peut utiliser une gamme étendue de structures grammaticales complexes de façon appropriée et avec beaucoup de souplesse. Peut choisir la formulation appropriée dans un large répertoire de langue pour exprimer sans restriction ce qu'il/elle veut dire.
B2	Peut s'exprimer clairement et sans donner l'impression d'avoir à restreindre ce qu'il/elle souhaite dire. Possède une gamme assez étendue de langue pour pouvoir faire des descriptions claires, exprimer des points de vue et développer des arguments sans chercher ses mots de manière évidente et en utilisant des phrases complexes.
B1	Possède une gamme assez étendue de langue pour décrire des situations imprévisibles, expliquer les points principaux d'un problème ou d'une idée avec assez de précision et exprimer sa pensée sur des sujets abstraits ou culturels tels que la musique ou le cinéma. Possède suffisamment de moyens linguistiques pour s'en sortir, et suffisamment de vocabulaire pour s'exprimer avec quelques hésitations et périphrases sur des sujets tels que la famille, les loisirs et centres d'intérêt, le travail, les voyages et l'actualité mais le vocabulaire limité conduit à des répétitions et parfois même à des difficultés de formulation.
A2	Possède un répertoire de langue élémentaire qui lui permet de se débrouiller dans des situations courantes au contenu prévisible, bien qu'il lui faille généralement chercher ses mots et trouver un compromis par rapport à ses intentions de communication.
A2	Peut produire de brèves expressions courantes afin de répondre à des besoins simples de type concret : détails personnels, routines quotidiennes, désirs et besoins, demandes d'information. Peut utiliser des modèles de phrases élémentaires et communiquer à l'aide de d'expressions mémorisées, de groupes de quelques mots et d'expressions toutes faites, sur soi, les gens, ce qu'ils font, les lieux, les biens, etc. Possède un répertoire limité de courtes expressions mémorisées couvrant les premières nécessités vitales des situations prévisibles ; des pannes fréquentes et des malentendus surviennent dans les situations imprévues.
A1	Possède un choix élémentaire d'expressions simples pour les informations sur soi et les besoins de type courant. Peut utiliser quelques structures simples dans des phrases simples en supprimant ou en simplifiant des éléments
Pré-A1	Peut utiliser des mots isolés et des expressions simples pour donner des informations simples sur soi.

ÉTENDUE DU VOCABULAIRE	PROSIGN
C2	Possède une bonne maîtrise d'un vaste répertoire lexical incluant des expressions idiomatiques et des termes familiers ; est conscient des niveaux de connotation sémantique.
C1	Possède une bonne maîtrise d'un vaste répertoire lexical lui permettant de surmonter facilement les lacunes par des périphrases avec une recherche peu apparente d'expressions et de stratégies d'évitement. Peut choisir entre plusieurs possibilités lexicales dans pratiquement toutes les situations en utilisant des synonymes même pour des mots non familiers. Maîtrise bien les expressions idiomatiques familières et fait des jeux de mots avec facilité. Peut comprendre et utiliser de façon appropriée la gamme de vocabulaire technique et d'expressions idiomatiques propres à son domaine de spécialité.
B2	Peut comprendre et utiliser les termes techniques généraux de son domaine, quand il/elle en discute avec d'autres spécialistes. Possède une bonne gamme de vocabulaire pour les sujets relatifs à son domaine et les sujets plus généraux. Peut varier sa formulation pour éviter des répétitions fréquentes, mais des lacunes lexicales peuvent encore provoquer des hésitations et l'usage de périphrases. Peut produire assez systématiquement de nombreux mots adéquats dans la plupart des contextes. Peut comprendre et utiliser une grande partie du vocabulaire spécialisé de son domaine mais a des difficultés avec la terminologie d'une spécialité différente de la sienne.
B1	A une bonne gamme de vocabulaire en rapport avec des sujets familiers et des situations quotidiennes. Possède un vocabulaire suffisant pour s'exprimer à l'aide de périphrases sur la plupart des sujets relatifs à sa vie quotidienne tels que la famille, les loisirs et les centres d'intérêt, le travail, les voyages et l'actualité.
A2	Possède un vocabulaire suffisant pour mener des transactions quotidiennes courantes dans des situations et sur des sujets familiers. Possède un vocabulaire suffisant pour satisfaire les besoins communicatifs élémentaires. Possède un vocabulaire suffisant pour satisfaire les besoins primordiaux.
A1	Possède un répertoire élémentaire de mots isolés et d'expressions relatifs à des situations concrètes précises.
Pré-A1	*Pas de descripteur disponible*

AISANCE À L'ORAL	PROSIGN
C2	Peut s'exprimer longuement dans un discours naturel et sans effort. Ne s'arrête que pour réfléchir au mot juste qui exprimera précisément sa pensée ou pour trouver un exemple approprié qui illustre l'explication.
C1	Peut s'exprimer avec aisance et spontanéité presque sans effort ; seul un sujet conceptuellement difficile est susceptible de gêner le flot naturel et fluide du discours.
B2	Peut communiquer avec spontanéité, montrant souvent une remarquable aisance et une facilité d'expression même dans des énoncés complexes assez longs.
	Peut parler relativement longtemps avec un débit assez régulier bien qu'il/elle puisse hésiter en cherchant tournures et expressions ; l'on remarque peu de longues pauses.
	Peut communiquer avec un degré d'aisance et de spontanéité qui rend tout à fait possible une interaction régulière avec des locuteurs de la langue cible sans imposer d'effort de part et d'autre.
B1	Peut s'exprimer avec une certaine aisance. Malgré quelques problèmes de formulation ayant pour conséquence pauses et impasses, est capable de continuer effectivement à parler sans aide.
	Peut discourir de manière compréhensible même si les pauses pour chercher ses mots et ses phrases et pour faire ses corrections sont très évidentes, particulièrement dans les séquences plus longues de production libre.
A2	Peut se faire comprendre dans une brève intervention, même si la reformulation, les pauses et les faux démarrages sont très évidents.
	Peut construire des phrases sur des sujets familiers avec une aisance suffisante pour gérer des échanges courts et malgré des hésitations et des faux démarrages évidents.
A1	Peut se débrouiller avec des énoncés très courts, isolés, généralement stéréotypés, avec de nombreuses pauses pour chercher ses mots, pour prononcer les moins familiers et pour remédier à la communication.
Pré-A1	Peut se débrouiller en prononçant des paroles très courtes, isolées et répétées à l'aide de gestes et en demandant de l'aide quand cela est nécessaire.

Annexe 2.

Échantillon des grilles pour les évaluations des productions orales (grille 1) et pour l'annotation du temps consacré à l'évaluation (grille 2).

Grille 1

PRODUCTION Nº 1 _	
Niveau :	**PRODUCTION ORALE GÉNÉRALE**
OBSERVATIONS. JUSTIFICATION	
Niveau :	**COMPÉTENCE LINGUISTIQUE. ÉTENDUE LINGUISTIQUE GÉNÉRALE**
OBSERVATIONS. JUSTIFICATION	
Niveau :	**COMPÉTENCE LINGUISTIQUE. ÉTENDUE DE VOCABULAIRE**
OBSERVATIONS. JUSTIFICATION	

Niveau :	**COMPÉTENCE PRAGMATIQUE. AISANCE**
OBSERVATIONS. JUSTIFICATION	

Grille 2

DATE	**TEMPS CONSACRÉ À L'ÉVALUATION**	**DATE**	**TEMPS CONSACRÉ À L'ÉVALUATION**	**DATE**	**TEMPS CONSACRÉ À L'ÉVALUATION**

NOTES

1. La fiabilité est une qualité des résultats d'un test qui est en relation avec le degré auquel les scores sont le résultat d'une évaluation libre d'erreur de mesure (Fulcher 2003), indépendamment du moment où l'évaluation s'est effectuée, de ses caractéristiques formelles, des évaluateurs qui la notent, etc. (Bachman 1990).

2. SCFLE : Stratégies de Communication en Français Langue Étrangère.

3. La sélection a été effectuée à l'aide d'un échantillonnage aléatoire basé sur un tirage au sort avec l'application Padlet.

4. Le type/token ratio (TTR) est un taux calculé en mettant en relation les différents mots utilisés (*type*) et le nombre total des mots utilisés (*token*).

5. L'utilisation des pauses de 250 millisecondes pour délimiter les segments est la plus utilisée dans la recherche en fluidité (Kormos 2006, Segalowitz 2010) ; une pause de plus de 200 millisecondes est considérée comme une hésitation (Hilton *et al*. 2008).

6. Pour interpréter les résultats Altman (1991, cité par Warrens & Pratiwi 2016) propose une classification où les Kappas peuvent être très faibles (0 à 0,20), faibles (0,21 à 0,40), modérés (0,41 à 0,60), forts (0,61 à 0,80) et presque parfaits (0,81 à 1,00).

RÉSUMÉS

Cet article présente une expérience de formation à l'évaluation de la production orale en français langue étrangère (FLE) auprès de futurs enseignants de FLE. La formation s'est déroulée en deux étapes, l'une pour les former à l'évaluation de la production orale en FLE avec quatre échelles du CECRL, et l'autre pour montrer comment évaluer la production orale avec des mesures textométriques et quantitatives. Les résultats montrent que les deux méthodologies présentent des inconsistances ; les résultats qualitatifs montrent des désaccords inter-juges dans la plupart des échelles, et les résultats quantitatifs ne montrent pas de changements significatifs à long terme dans l'ensemble du groupe.

This article deals with an experiment in training future teachers of French as a foreign language (FFL) in the assessment of oral production. The training took place in two stages, one to train them to assess oral production in FFL with four CEFR scales and the other to show how to assess oral production with textometric and quantitative measures. The results show inconsistencies in the two methodologies; the qualitative results show inter-rater disagreement in most scales, and the quantitative results do not show significant long-term changes in the group.

INDEX

Mots-clés : corpus d'apprenants, production orale, Français Langue Étrangère, formation des enseignants, évaluation des langues, CECRL
Keywords : learner corpus, oral production, French Foreign Language, teacher training, language assessment, CEFR

AUTEUR

MINERVA ROJAS

Université Côte d'Azur. UMR 7320 Bases Corpus Langage Université Côte d'Azur | CNRS