Cahiers de l'ILSL 56, 2018, 311-331

Sur le rôle de la diversité lexicale dans la mesure de la diversité flexionnelle

Aris XANTHOS
Université de Lausanne

Cette contribution se propose de revisiter une partie des résultats de la recherche de Xanthos et al. *(2011) sur la relation entre diversité flexionnelle dans le langage adressé aux enfants et développement de la flexion dans le langage enfantin. À la lumière des progrès méthodologiques récents en matière de mesure de la diversité flexionnelle, je chercherai à mettre en évidence la nécessité de contrôler non seulement les fluctuations de taille d'échantillon mais aussi celles de diversité lexicale dans les données. Les résultats empiriques suggèrent que l'adoption d'une mesure robuste vis-à-vis de ce type de variations revêt une importance cruciale dans le contexte de l'étude de l'acquisition*[1].

[1] Je remercie mes collègues du *Crosslinguistic Project on Pre- and Protomorphology in Language Acquisition* pour la mise à disposition des corpus d'acquisition qu'ils ont récoltés, transcrits et enrichis au fil des années : Ayhan Aksu-Koç (Université du Bosphore), Wolfgang Dressler (Académie autrichienne des sciences, Vienne), Natalia Gagarina (Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin), Steven Gillis (Université d'Anvers), Gordana Hrzica, Melita Kovačević et Marijan Palmović (Université de Zagreb), Katharina Korecky-Kröll et Sabine Laaha (Université de Vienne), Klaus Laalo (Université de Tampere), F. Nihan Ketrez (Université Bilgi d'Istanbul) et Maria D. Voeikova (Académie des sciences de Russie). Ils sont heureux de contribuer au moins par leurs données à cet hommage à Marianne Kilani-Schoch – qui de son côté me pardonnera, j'espère, d'avoir utilisé les siennes sans demander son autorisation. Merci enfin à François Bavaud et Christian Surcouf pour leur relecture attentive.

1. INTRODUCTION

La mesure de la diversité lexicale est l'un des sujets les plus abondamment étudiés en linguistique quantitative, en particulier du point de vue de la relation existant entre la *variété* lexicale, soit le nombre *V* d'unités lexicales distinctes (ou *types*) dans un échantillon textuel, et la *taille* de cet échantillon, soit le nombre *N* de *tokens* qui le composent. En effet, la variété *V* ne peut que croître avec la taille *N* de l'échantillon, si bien que cet indice ne permet pas de comparer directement des échantillons de tailles variables – pas plus que ne le permettent le *ratio types–tokens RTT = V/N* et ses diverses transformations (Tweedie & Baayen 1998).

De façon très générale, les propositions méthodologiques récentes en matière de mesure de diversité lexicale sont des variations plus ou moins sophistiquées d'une idée introduite par Johnson (1944). Pour remédier au problème de la dépendance à la taille de l'échantillon, Johnson propose de découper l'échantillon considéré en *B* segments de *S* tokens chacun et d'évaluer la diversité lexicale de l'échantillon par le RTT moyen dans ces segments. Dans ce contexte, le nombre *B* de séquences est le résultat de la division entière de la taille *N* de l'échantillon par la taille *S* choisie pour les segments. Le seul paramètre de la méthode est donc *S*, qui peut en principe être compris entre 1 et la taille du plus petit des échantillons à comparer.

L'une des principales évolutions de cette approche est celle proposée par Dubrocard (1988) et consistant à remplacer les segments de *S* tokens contigus de Johnson (1944) par des sous-échantillons de *S* tokens obtenus par tirage aléatoire sans remise à partir de l'échantillon entier[2]. Cette pratique a l'avantage de faire du nombre *B* de sous-échantillons un paramètre que l'on peut choisir

[2] Ici et dans la suite, j'utilise systématiquement le terme *échantillon* pour désigner l'objet dont la diversité est mesurée et celui de *sous-échantillon* pour référer à une sélection de tokens tirés d'un échantillon.

indépendamment de *S* ainsi que de permettre d'exploiter toute l'information présente dans l'échantillon original[3]. C'est notamment l'approche retenue par la méthode *VOCD* (pour *vocabulary diversity*, cf. Malvern & Richards 1997), qui constitue actuellement un standard *de facto* en matière de diversité lexicale. La mesure repose sur le calcul du RTT moyen dans des sous-échantillons de taille *S* croissante (35, 36, …, 50 tokens) pour construire une courbe de RTT dite *empirique*; une technique d'ajustement de courbe permet ensuite de trouver, au sein d'une famille de courbes *théoriques* spécifiées par la variation d'un paramètre unique dans un modèle formel de la relation entre RTT et nombre *N* de tokens, celle qui est la plus similaire à la courbe empirique. La valeur du paramètre générant cette courbe constitue alors la mesure de diversité rapportée.

En fait, comme le montrait déjà le développement de Serant (1988), il est possible de calculer de façon analytique la moyenne de la variété *V* (ou du RTT) attendue sur *tous* les sous-échantillons possibles de taille *S* donnée[4]. Ce calcul basé sur la loi hypergéométrique n'a même pas le désavantage d'être particulièrement coûteux en temps d'exécution, si bien que l'échantillonnage aléatoire mis en œuvre dans VOCD notamment lui est strictement inférieur. McCarthy & Jarvis (2007) arguent à ce propos que le seul intérêt de la procédure d'ajustement de courbe qui sous-tend la méthode est de réduire le bruit engendré par l'échantillonnage aléatoire, et qu'en ce sens VOCD est essentiellement une façon d'approximer $V^{theo}(S)$.

---

[3] La méthode de Johnson (1944) implique en effet de négliger une partie des tokens de l'échantillon dans le cas général où sa taille *N* n'est pas un multiple de la longueur *S* des séquences.

[4] Dans ce qui suit, je parlerai de *variété ré-échantillonnée observée*, notée $V^{obs}(S)$, pour désigner la moyenne de la variété calculée à partir d'un ensemble de sous-échantillons aléatoires, telle qu'utilisée dans VOCD par exemple; le terme *variété ré-échantillonnée théorique* et la notation $V^{theo}(S)$ feront référence à la quantité calculée selon les indications de Serant (1988).

McCarthy & Jarvis (2010) proposent une alternative intéressante à la méthodologie de Johnson (1944) et ses dérivés : plutôt que d'évaluer le RTT moyen dans des segments de longueur *S* donnée, leur méthode nommée *MTLD* (*measure of textual lexical diversity*) vise à estimer la longueur moyenne de segments dont le *RTT* est fixé. Mes propres expériences visant à comparer une variante de cette mesure[5] et $V^{obs}(S)$ suggèrent que les deux approches sont largement interchangeables : « grosso modo, elles présentent la même réaction (ou absence de réaction) aux variations de taille et de diversité des données » (Xanthos 2013, 250). Considérant par ailleurs que $V^{obs}(S)$ est une approximation de $V^{theo}(S)$, si l'on admet avec McCarthy & Jarvis (2007) que c'est aussi le cas de VOCD, il semble en définitive qu'aussi bien $V^{theo}(S)$ que MTLD puissent être tenues pour représentatives de l'état de l'art en matière de mesure de la diversité lexicale.

La mesure de la diversité *flexionnelle* est loin d'avoir engendré à ce jour un volume de recherche comparable à la mesure de la diversité lexicale[6]. Les méthodes proposées dans ce domaine reposent généralement sur un modèle simple de la flexion, selon lequel chaque token de l'échantillon textuel considéré (par exemple les formes fléchies *va*, *vais*, *aller* ou *fera*) appartient à un et un seul lexème (respectivement les verbes ALLER ou FAIRE). Jusque dans les années 2000, l'indice le plus utilisé pour quantifier la diversité

---

5 La particularité de cette variante est de reposer sur un échantillonnage aléatoire plutôt que sur un découpage du texte en segments (voir Xanthos 2013 pour plus de détails).

6 Il ne sera fait état ici que des indices portant spécifiquement sur la dimension paradigmatique de la flexion, à l'exclusion notamment des propositions visant à caractériser sa dimension syntagmatique (p. ex. Greenberg 1954; Nichols 1992), ainsi que des travaux abordant l'évaluation de la complexité morphologique sous l'angle de la théorie de l'information (p. ex. Juola 1998; Bane 2008; Moscoso del Prado Martín 2011), dans la mesure où ils ne font pas explicitement le départ entre ces deux dimensions.

flexionnelle d'un échantillon est ce que Xanthos & Gillis (2010) nomment *MSP* (pour *mean size of paradigm*, la taille moyenne du paradigme), soit le rapport $V/\tilde{V}$ entre la variété lexicale V et la variété des lexèmes, notée $\tilde{V}$. Cette quantité, qui s'interprète comme le nombre moyen de formes distinctes par lexème, est toutefois dépendante de la taille N de l'échantillon et s'avère donc aussi inadéquate que la variété lexicale ou le *RTT* pour comparer des échantillons de tailles variables.

C'est à Malvern, Richards, Chipere & Durán (2004) qu'est due la première tentative de développer une mesure de diversité flexionnelle prenant en compte la question de la dépendance à la taille d'échantillon. Ces auteurs proposent d'interpréter la différence entre le résultat de l'application de la méthode VOCD aux formes fléchies et celui, systématiquement inférieur, de son application aux lexèmes comme une mesure de diversité flexionnelle, mesure qu'ils nomment *ID* pour *inflectional diversity*. Cette proposition présente toutefois certains inconvénients, en particulier le fait que l'échelle dans laquelle ID est exprimée n'a pas d'interprétation intuitive. C'est ce qui conduit Xanthos & Gillis (2010) à défendre une autre mesure, qu'ils désignent par le terme de MSP *normalisé* et qu'on peut concevoir comme une transposition de l'idée générale de Johnson (1944) au domaine de la flexion : il s'agit de fixer une taille S de sous-échantillon, construire B sous-échantillons de S tokens (par tirage aléatoire sans remise), calculer le MSP de chaque sous-échantillon, et finalement rapporter le MSP moyen dans les B sous-échantillons, noté MSP(S) car il dépend directement de la valeur choisie pour le paramètre S [7].

La spécificité des approches les plus récentes de la mesure de la diversité flexionnelle est de tenir compte non seulement de variations de taille d'échantillon mais aussi de diversité lexicale – ou plus

[7] L'indice *MCI* proposé par Pallotti (2015) met en œuvre un raisonnement similaire pour calculer non pas le nombre moyen de formes distinctes par lexème mais le nombre moyen d'*exposants*, soit de procédés tels qu'affixation d'un morphe particulier, apophonie, etc.

précisément lexématique. En effet, en fixant le nombre de tokens par sous-échantillons, la méthode de calcul du MSP normalisé s'expose à un nouveau biais : toutes choses étant égales par ailleurs, les lexèmes apparaissant dans les sous-échantillons auront en moyenne une fréquence d'autant plus élevée que la diversité lexématique de l'échantillon de base est faible. Ainsi, même si le nombre de tokens dans les sous-échantillons est constant, le nombre moyen de tokens *par lexème* dépend de la diversité lexématique et les variations potentielles de celle-ci contrecarrent l'objectif de normalisation qui sous-tend la méthode.

Conscients de cet écueil, Krajewski, Lieven & Theakston (2012) proposent de ne prendre en compte dans le calcul du MSP normalisé que les lexèmes dont l'occurrence est attestée dans tous les échantillons qu'il s'agit de comparer. Cette option méthodologique implique toutefois de négliger une part potentiellement importante des données, notamment dans le contexte des corpus d'acquisition, où l'inventaire des lexèmes présents dans un échantillon mensuel de la production d'un jeune enfant, par exemple, peut être très restreint. En outre, opérer sur l'intersection des lexiques revient à uniformiser la variété lexématique des échantillons, mais dans le cas général où la distribution des lexèmes varie d'un échantillon à l'autre, rien ne garantit que la variété lexématique des sous-échantillons, ni par conséquent le nombre moyen de tokens par lexème, soient uniformes.

La méthodologie *RMSP* (pour *robust* MSP) introduite par Xanthos & Guex (2015) vise à remédier à ces inconvénients. Son principe est de calculer le MSP normalisé en ajustant séparément la taille *S* des sous-échantillons pour chacun des échantillons à comparer, de sorte à affecter une taille plus réduite aux échantillons dont la diversité lexématique est moindre. En particulier, il s'agit de déterminer, pour chaque échantillon, la taille de sous-échantillon aboutissant à uniformiser autant que possible le nombre moyen de tokens par lexème (dans tous les sous-échantillons de tous les échantillons), ou de façon équivalente, le nombre moyen de lexèmes par token, soit le TTR lexématique. En pratique, on fixe d'abord la taille maximale S^{max} des sous-échantillons, puis on identifie l'échantillon dont le TTR

lexématique (ré-échantillonné sur S^{max} tokens) est maximal et l'on enregistre sa valeur, notée ici TTR^{max}. On calcule $MSP(S^{max})$ pour cet échantillon, puis pour chaque autre échantillon *i*, on recherche la taille de sous-échantillon $S^i \geq 2$ telle que le TTR lexématique ré-échantillonné sur S^i tokens de l'échantillon *i* soit aussi proche que possible de TTR^{max}, et l'on calcule $MSP(S^i)$ pour cet échantillon (les détails de cet algorithme sont donnés dans Xanthos & Guex 2015).

Les simulations conduites par Xanthos & Guex (2015) montrent que le RMSP est significativement moins affecté par les variations de diversité lexématique que le MSP normalisé. L'objectif de la présente contribution est d'évaluer l'impact de cette nouvelle méthodologie sur des données « réelles » (plutôt que générées artificiellement). À cet effet, je me propose de revisiter une partie des résultats de la recherche conduite au sein du *Crosslinguistic Project on Pre- and Protomorphology in Language Acquisition* sur la relation entre diversité flexionnelle dans le langage adressé aux enfants et taux d'acquisition de la flexion dans le langage enfantin (Laaha & Gillis 2007; Xanthos *et al.* 2011). Le choix de ce cas d'étude n'est pas seulement motivé par le fait qu'il s'agit d'une des applications les plus systématiques du MSP normalisé, portant sur des données issues de plusieurs langues typologiquement variées. C'est aussi, d'un point de vue personnel, un point d'orgue de ma collaboration avec le groupe de chercheuses et de chercheurs réunis à Vienne autour de Wolfgang Dressler, dans lequel j'ai eu l'honneur d'être invité par ma collègue et amie Marianne Kilani-Schoch, et qui a fourni le cadre de bon nombre de mes échanges et collaborations avec elle depuis près de 20 ans.

Dans la section suivante, je présente les données utilisées dans cette étude ainsi que les traitements différenciés appliqués au langage adressé à l'enfant et au langage enfantin, dont l'exposé des résultats fait l'objet de la section 3. La section 4 propose une discussion plus générale de ces observations en relation avec les objectifs de cette contribution. Je conclus par une synthèse de ce travail et une réflexion plus générale sur sa pertinence pour l'étude de l'acquisition.

2. MÉTHODE

2.1. Données

Les données utilisées pour cette contribution forment un sous-ensemble de celles utilisées par Xanthos *et al.* (2011) et décrites en détail dans cet article. Il s'agit de corpus longitudinaux de 7 enfants acquérant 7 langues : allemand, croate, finnois, français, néerlandais, russe et turc[8]. Les enfants ont été enregistrés à leur domicile plusieurs fois par mois sur une période de temps variable (entre l'émergence de la parole et l'âge de 3 ans environ), en interaction avec leur mère ou les autres personnes en charge de leur garde dans le cadre d'activités quotidiennes telles que jeu, bain, repas, etc. Les données ont été transcrites en format CHAT et codées morphologiquement selon les normes du projet CHILDES (MacWhinney, 2000).

Les analyses présentées ci-dessous portent exclusivement sur les formes verbales présentes dans ces données. Le codage morphologique spécifie le lexème auquel chaque forme est associée. Les verbes à particule séparable allemands et néerlandais sont traités comme des occurrences d'un seul et même lexème (p. ex. all. *machen* 'faire', *aufmachen* 'ouvrir' et *zumachen* 'fermer'). Les flexions homophones sont traitées comme des occurrences d'une seule et même forme (p. ex. *parle* et *parles*). Notons enfin qu'une construction périphrastique comme le passé composé *est allé* est comptée comme une occurrence de la forme *est* suivie d'une autre, distincte, de la forme *allé*.

L'analyse du langage adressé à l'enfant (*LAE*) porte sur l'ensemble des 7 langues, tandis que celle du langage enfantin (*LE*) se concentre sur les corpus néerlandais et russe, en vertu de leur

[8] Seules les langues pour lesquelles un échantillon d'au moins 1000 tokens de langage adressé à l'enfant était disponible ont été retenues, à l'exclusion donc du grec et du yucatec maya, présents dans l'article original.

caractère particulier tel que révélé par l'étude du LAE (cf. section 3.1). Les données du LE sont subdivisées en 12 échantillons mensuels successifs, de 1 an et 5 mois à 2 ans et 4 mois (1;5–2;4). Le tableau 1 ci-dessous résume les tailles (en nombre de tokens verbaux) des divers sous-ensembles de ces données.

Langue	**Nb. tokens verbaux**	
	LAE	**LE**
Allemand (ALL)	13975	–
Croate (CRO)	10795	–
Finnois (FIN)	4329	–
Français (FRA)	13768	–
Néerlandais (NEE)	4362	1067
Russe (RUS)	7394	1434
Turc (TUR)	1193	–
Total	55816	2501

Tableau 1 : Nombre de tokens verbaux dans le LAE et le LE des corpus analysés dans cette étude.

2.2. Analyse du langage adressé à l'enfant

Cette étude réplique à quelques détails près l'analyse du LAE effectuée par Xanthos *et al.* (2011) et compare les résultats obtenus avec le MSP normalisé et le RMSP. Pour chaque langue, on considère un échantillon unique composé de l'ensemble des tokens verbaux présents dans le corpus. Les deux mesures de diversité flexionnelle sont calculées sur $B = 1000$ sous-échantillons de chaque langue. Pour le MSP normalisé, les sous-échantillons comptent $S = 1000$ tokens; pour le RMSP, ils contiennent 1000 tokens au maximum (c'est-à-dire qu'on fixe $S^{max} = 1000$). Afin de mieux comprendre les différences observées entre MSP normalisé et RMSP, on rapporte également la variété lexématique théorique (sur 1000 tokens) dans les échantillons de chaque langue.

Il est à noter que dans l'article original, l'algorithme de ré-échantillonnage aléatoire est implémenté de façon telle que le nombre

moyen de tokens par sous-échantillon vaut environ 1000, mais le nombre de tokens dans les sous-échantillons individuels reste variable. Pour la présente étude, les calculs ont été effectués avec le logiciel *Textable* (Xanthos 2014), dans lequel l'implémentation de l'algorithme de ré-échantillonnage garantit que chaque sous-échantillon contient le nombre exact de tokens désiré.

2.3. Analyse du langage enfantin

Le traitement appliqué au LE par Xanthos *et al.* (2011) est plus complexe que celui appliqué au LAE. D'une part, il prend en compte la dimension longitudinale des données et vise à construire une courbe d'évolution de la diversité flexionnelle en fonction de l'âge dans chaque corpus plutôt qu'à évaluer globalement sa diversité flexionnelle. D'autre part, chaque point successif de la courbe de diversité est calculé non pas sur la base de l'échantillon mensuel correspondant mais du cumul de l'ensemble des échantillons mensuels jusqu'à ce point dans le temps, ceci afin de capturer des relations flexionnelles entre formes pouvant apparaître à des périodes distinctes.

L'implémentation du mécanisme de ré-échantillonnage du LE de Xanthos *et al.* (2011) a pour objectif de contrôler non pas le nombre de tokens par mois mais l'accroissement mensuel moyen ΔS de ce nombre. Ainsi, pour un corpus de N tokens répartis sur M mois, chaque sous-échantillon est construit en donnant aux tokens du corpus une probabilité $P = \Delta S \cdot M/N$ d'être sélectionnés, quel que soit le mois de leur apparition. Ce sous-échantillon (qui compte en moyenne $\Delta S \cdot M$ tokens) fait alors l'objet du cumul mensuel indiqué précédemment et M valeurs de MSP sont calculées pour les données cumulées des M mois successifs. Enfin, on calcule la moyenne du MSP pour chaque mois donné dans tous les sous-échantillons et la séquence de ces MSP mensuels moyens forme la courbe de diversité flexionnelle du corpus considéré.

Une adaptation directe de la méthode RMSP à cette procédure consisterait à ajuster la probabilité P de sélection des tokens dans chaque corpus en fonction de la diversité lexématique du

corpus. Toutefois, comme cette probabilité est fixée pour l'ensemble du corpus, un tel ajustement ne permettrait de contrôler la diversité lexématique qu'au niveau des corpus entiers. Or on peut s'attendre à observer des variations non négligeables au niveau plus fin des échantillons mensuels d'un corpus donné – en particulier dans le contexte de données d'acquisition. Pour pouvoir contrôler la diversité lexématique à ce niveau de granularité, il est nécessaire que l'échantillon *mensuel*, qu'il soit cumulé ou non, constitue l'unité de base de l'analyse.

Pour ces raisons, la présente étude ne réplique pas directement l'analyse du LE faite par Xanthos *et al.* (2011) mais en propose plutôt une variante mieux adaptée à la comparaison avec le RMSP. La courbe de diversité flexionnelle est obtenue dans ce cas en cumulant d'abord les échantillons mensuels, puis en calculant le MSP normalisé sur 50 tokens et le RMSP (avec $S^{max} = 50$) dans chaque échantillon mensuel cumulé. Comme pour l'analyse du LAE, on rapporte également la variété lexématique théorique (sur 50 tokens) $\tilde{V}^{theo}(50)$, cette fois-ci dans chaque échantillon mensuel cumulé.

3. Résultats

3.1. Analyse du langage adressé à l'enfant

La figure 1 (p. 322) représente les valeurs obtenues pour les deux mesures de diversité flexionnelle dans le LAE. On observe que le RMSP est systématiquement inférieur au MSP normalisé, à l'exception du corpus dont la diversité lexématique est maximale, soit le russe. Dans l'ensemble, les variations du RMSP sont réduites relativement à celles du MSP normalisé (l'écart-type sur les 7 valeurs de RMSP mesurées vaut 0,33, contre 0,74 pour les valeurs de MSP normalisé). Les valeurs sont pareillement ordonnées, à l'exception du corpus RUS, 6ème sur 7 du point de vue du MSP normalisé mais 4ème selon le RMSP (passant ainsi devant ALL et NEE).

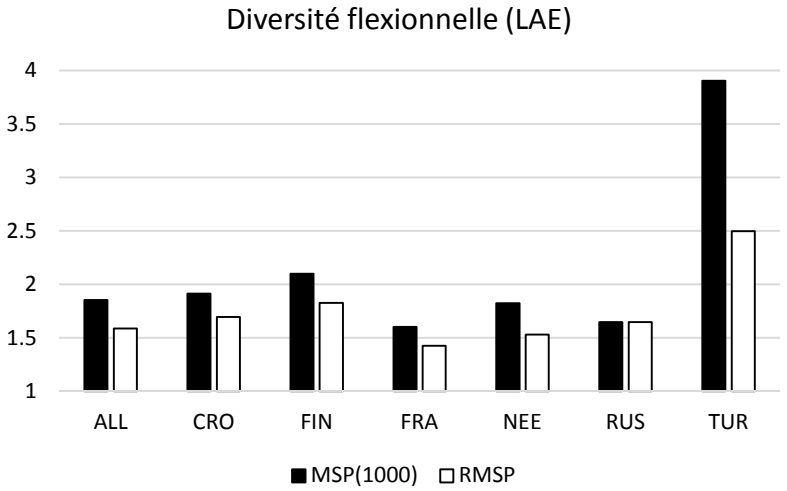

Figure 1 : MSP normalisé (sur S = 1000 tokens) et RMSP (sur S^{max} = 1000 tokens au maximum) dans le langage adressé à l'enfant[9].

Les figures 2 et 3 représentent respectivement la variété lexématique ré-échantillonnée théorique dans le LAE et la taille de sous-échantillon déterminée par l'algorithme RMSP pour chacun des corpus. Leur comparaison met en évidence la façon dont la méthode ajuste la taille de sous-échantillon en fonction de la variété lexématique (la corrélation entre les deux valant ici 0,97, $p < 0{,}001$). La diversité lexématique considérablement plus élevée du corpus RUS, qui pénalise d'autant l'évaluation de la diversité flexionnelle par le MSP normalisé, est ainsi contrebalancée par une réduction drastique de la taille de sous-échantillon des autres corpus – jusqu'à un minimum de 142 tokens (au lieu de 1000) pour le corpus NEE.

9 Les intervalles de confiance à 95 % sont trop restreints pour être utilement représentés ici (ils n'excèdent pas ±0,02).

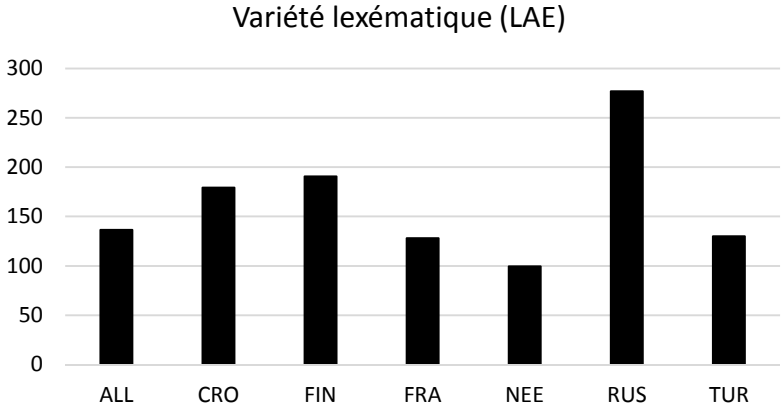

Figure 2 : Variété lexématique ré-échantillonnée théorique (sur S = 1000 tokens) dans le langage adressé à l'enfant.

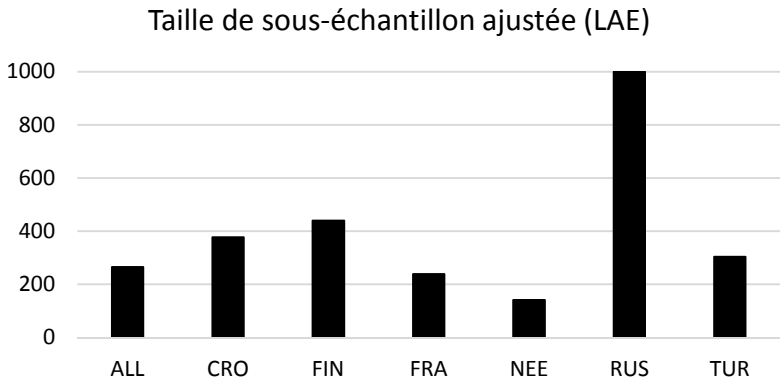

Figure 3 : Taille de sous-échantillon après ajustement selon la méthode RMSP dans le langage adressé à l'enfant.

3.2. Analyse du langage enfantin

Les figures 4 et 5 (p. 324) représentent respectivement l'évolution du MSP normalisé et celle du RMSP dans le LE des corpus NEE et RUS. La différence entre les deux figures est remarquable.

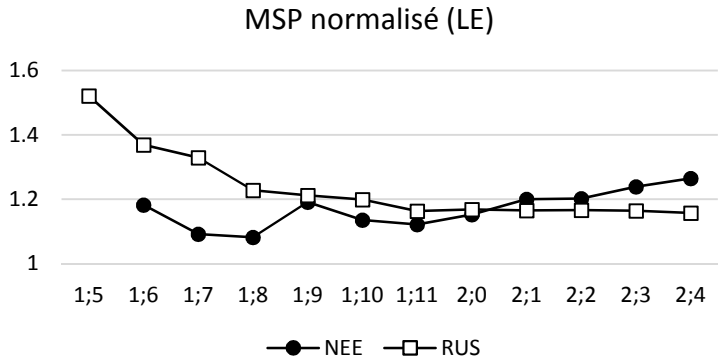

Figure 4 : MSP normalisé (sur S = 50 tokens) dans le langage enfantin[10]*.*

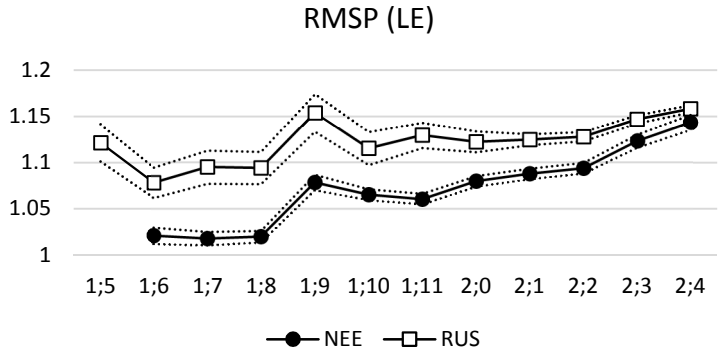

Figure 5 : RMSP (sur $S^{max} = 50$ *tokens au maximum) dans le langage enfantin*[11]*.*

[10] Le premier mois (1;5) du corpus NEE ne compte que 14 tokens, si bien qu'il n'est pas possible de calculer la variété lexématique ré-échantillonnée sur 50 tokens pour cet échantillon (ni le MSP normalisé ou le RMSP correspondants).

[11] Les lignes pointillées délimitent les intervalles de confiance à 95 % (qui n'ont pas été représentés sur la figure 4 car ils n'excèdent pas ±0,005).

À l'aune du MSP normalisé, si la croissance attendue de la diversité flexionnelle est perceptible dans le corpus NEE (la corrélation de Spearman avec l'âge valant 0,76, $p < 0{,}01$), c'est une nette *dé*croissance qu'on observe dans le corpus RUS (la corrélation de Spearman avec l'âge valant ici $-0{,}92$, $p < 0{,}001$). Par contraste, le RMSP détecte dans les deux cas une croissance significative : la corrélation de Spearman avec l'âge vaut 0,94 pour le corpus NEE ($p < 0{,}001$) et 0,73 pour le corpus RUS ($p < 0{,}01$).

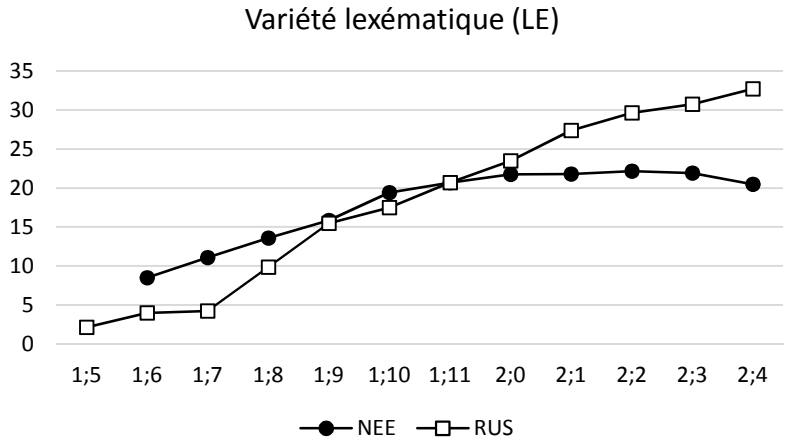

Figure 6 : Variété lexématique ré-échantillonnée théorique (sur S = 50 tokens) dans le langage enfantin.

La figure 6 représente l'évolution de la variété lexématique ré-échantillonnée théorique en fonction de l'âge dans les deux corpus. Elle est globalement croissante au fil des échantillons mensuels cumulés sur cette année de vie des deux enfants considérés (la corrélation de Spearman entre âge et variété lexématique vaut 0,85 pour NEE et 1 pour RUS, $p < 0{,}001$). Dans le corpus NEE, la croissance est plus rapide entre 1;6 et 2;0, puis atteint un plateau qui perdure jusqu'au terme de la période observée. Dans le corpus RUS, la croissance ne démarre véritablement qu'après 1;7, mais elle est plus rapide et se poursuit ensuite sans interruption. Il s'agit dans les deux cas d'importantes variations, largement attendues dans le

contexte de l'acquisition et permettant de mieux comprendre l'ampleur de la différence observée précédemment entre MSP normalisé et RMSP.

La taille de sous-échantillon déterminée par l'algorithme RMSP pour chacun des échantillons mensuels est donnée sur la figure 7. À l'exception des 4 derniers mois du corpus RUS, dont la diversité lexématique est la plus élevée, tous les échantillons voient leur taille de sous-échantillon réduite à un nombre de tokens compris entre 2 et 15 (sur un maximum de 50). Si la relation entre diversité lexématique et taille de sous-échantillon est moins évidente que dans le cas du LAE (cf. section 3.1), les deux variables n'en restent pas moins fortement liées (leur corrélation vaut 0,8, $p < 0{,}001$).

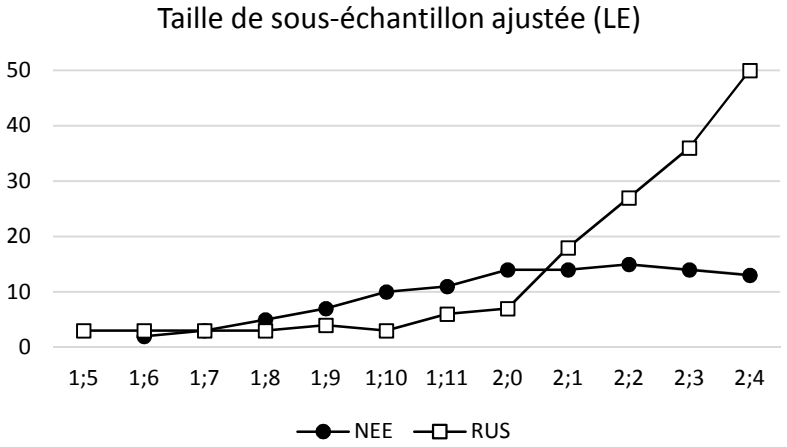

Figure 7 : Taille de sous-échantillon après ajustement selon la méthode RMSP dans le langage enfantin.

4. Discussion

L'analyse du LAE montre que des corpus de langues différentes peuvent présenter des variations de diversité lexématique considérables en dépit de la relative homogénéité de leurs conditions de production. En pareilles circonstances, uniformiser la taille de sous-

échantillon (comme c'est le cas dans le calcul du MSP normalisé) aboutit à sous-estimer la diversité flexionnelle des échantillons dont la diversité lexématique est plus élevée. C'est ce qui explique que, dans nos données, le RMSP dans le LAE de tous les autres corpus que RUS soit systématiquement réduit relativement au MSP normalisé correspondant. Si dans cet exemple, la correction apportée par le RMSP ne bouleverse pas les relations d'ordre entre les corpus, elle est suffisamment conséquente pour conduire à des conclusions différentes concernant une partie d'entre eux (en particulier RUS, ALL et NEE).

L'analyse du LE a porté plus spécifiquement sur les corpus RUS et NEE, dont la diversité lexématique dans le LAE s'est avérée la plus contrastée. Dans la période couverte par cette étude (1;5–2;4), il est bien documenté que le développement lexical est en plein essor (cf. p. ex. Caselli, Casadio & Bates 1999), ce que confirme l'examen de la diversité lexématique dans les échantillons mensuels cumulés du LE. Dans le corpus RUS, où ce phénomène est le plus marqué, cette mesure varie entre un peu plus de 2 lexèmes distincts par sous-échantillon de 50 tokens (à 1;5) et plus de 30 lexèmes (à 2;4). En d'autres termes, les lexèmes ont une fréquence moyenne d'environ 25 occurrences à 1;5 contre moins de 2 occurrences à 2;4. La conséquence de ce état de fait est la sous-estimation croissante de la diversité flexionnelle telle que mesurée par le MSP normalisé au fil des échantillons mensuels, à tel point qu'on en vient à conclure à une baisse graduelle de cette diversité dans le temps. Dans ce contexte, l'effet du RMSP est drastique : il aboutit à renverser littéralement la tendance et rétablir la progression développementale attendue pour le corpus RUS, ainsi qu'à renforcer celle déjà présente dans le corpus NEE. Comme l'algorithme RMSP fonctionne sur le principe d'une réduction par rapport à une taille de sous-échantillon maximale, il est vraisemblable que son impact soit encore plus marqué avec une taille de sous-échantillon plus élevée que 50 tokens, mais les effectifs mensuels de nos données sont trop restreints pour tester cette conjecture.

5. Conclusion

L'objectif de cette contribution était de mettre la mesure de diversité flexionnelle RMSP à l'épreuve de données réelles. Après avoir retracé le cheminement historique et conceptuel qui conduit des propositions initiales de Johnson (1944) aux plus récents développements dans ce domaine, j'ai présenté les données analysées dans cette étude (qui forment un sous-ensemble de celles exploitées dans Laaha & Gillis 2007 et Xanthos *et al*. 2011), ainsi que la méthodologie conçue pour tenter d'évaluer l'impact de l'approche RMSP relativement au MSP normalisé. La comparaison des deux mesures sur les données du LAE et, surtout, sur celles du LE, montre que cet impact est considérable. Dans les deux cas, il aboutit à des conclusions différentes concernant le degré de diversité flexionnelle relatif de certains échantillons. L'effet du RMSP est le plus remarquable dans le cas des données longitudinales du LE, dans la mesure où le développement lexical qui s'y reflète induit un biais systématique de nature à occulter le développement flexionnel tel que mesuré par le MSP normalisé.

Pour le traitement de données d'acquisition, en particulier dans une perspective translinguistique, il semble en définitive que la validité d'une mesure de diversité flexionnelle soit largement tributaire de la mise en place d'un dispositif de contrôle des variations de diversité lexématique tel que la méthode RMSP cherche à l'effectuer. Toutefois, ce gain de robustesse a un coût : il implique d'opérer sur des sous-échantillons encore plus réduits que le MSP normalisé, et donc de négliger une partie encore plus grande de l'information initiale. La condition pour diminuer cette perte serait de disposer d'échantillons plus grands; il faut y voir une raison supplémentaire, pour la recherche dans ce domaine, de collecter des corpus toujours plus denses.

RÉFÉRENCES

Bane Max (2008), Quantifying and Measuring Morphological Complexity, in *Proceedings of the 26th West Coast Conference on Formal Linguistics,* 67-76.

Caselli Cristina, Casadio Paola & Bates Elizabeth (1999), A comparison of the transition from first words to grammar in English and Italian, *Journal of Child Language 26,* 69-111.

Dubrocard Michel (1988), Évaluation de l'étendue du lexique, quelques essais de simulation, in Labbé Dominique, Thoiron Philippe & Serant Daniel (éd.), *Études sur la richesse et la structure lexicale*, Paris-Genève, Slatkine–Champion, 43-66.

Greenberg Joseph (1954), A quantitative approach to morphological typology of language, in Spencer Robert F. (éd.), *Method and perspective in anthropology*, Minneapolis, University of Minnesota Press, 192-195.

Johnson Wendell (1944), Studies in language behavior : I. A program approach, *Psychological Monographs 56*, 1-15.

Juola Patrick (1998), Measuring linguistic complexity : The morphological tier, *Journal of Quantitative Linguistics 5(3)*, 206-213.

Krajewski Grzegorz, Lieven Elena V. M. & Theakston Anna L. (2012), Productivity of a Polish child's inflectional noun morphology : a naturalistic study, *Morphology 22*, 9-34.

Laaha Sabine & Gillis Steven (éd.) (2007), Typological perspectives on the acquisition of noun and verb morphology, *Antwerp Papers in Linguistics 112*.

MacWhinney Brian (2000), *The CHILDES project : tools for analyzing talk*, Mahwah, NJ, Lawrence Erlbaum Associates.

Malvern David D. & Richards Brian J. (1997), A new measure of lexical diversity, in Ryan Ann & Wray Alison (éd.), *Evolving models of language*, Clevedon, UK, Multilingual Matters, 58-71.

Malvern David D., Richards Brian J., Chipere Ngoni & Durán, Pilar (2004), *Lexical diversity and language development : quantification and assessment,* Basingstoke, Palgrave Macmillan.

McCarthy Philip M. & Jarvis Scott (2007), vocd : A theoretical and empirical evaluation, *Language Testing 24(4)*, 459-488.

McCarthy Philip M. & Jarvis Scott (2010), MTLD, vocd-D, and HD-D : A validation study of sophisticated approaches to lexical diversity measurement, *Behavior Research Methods 42(2)*, 381-392.

Moscoso del Prado Martín Fermin (2011), The mirage of morphological complexity, In *Proceedings of Quantitative Measures in Morphology and Morphological Development*, Center for Human Development, UC San Diego, 3524-3529.

Nichols Johanna (1992), *Language diversity in space and time*, Chicago, University of Chicago Press.

Pallotti Gabriele (2015), A simple view of linguistic complexity, *Second Language Research 31(1)*, 117-134.

Serant Daniel (1988), À propos des modèles de raccourcissement de textes, in Labbé Dominique, Thoiron Philippe & Serant Daniel (éd.), *Études sur la richesse et la structure lexicale*, Paris-Genève, Slatkine–Champion, 77-91

Tweedie Fiona J. & Baayen R. Harald (1998), How variable may a constant be? Measures of lexical richness in perspective, *Computers and the Humanities 32*, 323-352.

Xanthos Aris (2013), L'évaluation (de l'évaluation)+ de la diversité lexicale, in Prikhodkine Alexei & Xanthos Aris (éd.), *Mélanges offerts en hommage à Remi Jolivet, Cahiers de l'ILSL 36*, 231-252.

Xanthos Aris (2014), Textable : programmation visuelle pour l'analyse de données textuelles, in *Actes des 12es Journées internationales d'analyse statistique des données textuelles (JADT 2014)*, 691-703.

Xanthos Aris & Gillis Steven (2010), Quantifying the development of inflectional diversity, *First Language 30(2),* 175-198.

Xanthos Aris & Guex Guillaume (2015), On the robust measurement of inflectional diversity, in Tuzzi Arjuna, Benesova Martina & Macutek Ján (éd.) *Recent Contributions to Quantitative Linguistics*, De Gruyter, 241-254.

Xanthos Aris, Laaha, Sabine, Gillis Steven, Stephany Ursula, Aksu-Koç Ayhan, Christofidou Anastasia, Gagarina Natalia, Hrzica Gordana, Ketrez Fatma Nihan, Kilani-Schoch Marianne, Korecky-Kröll Katharina, Kovačević Melita, Laalo Klaus, Palmović Marijan, Pfeiler Barbara, Voeikova Maria D. & Dressler Wolfgang U. (2011), On the role of morphological richness in the early development of noun and verb inflection, *First Language 31(4)*, 461-479.