Cahiers de l'ILSL-45 (2016 : 95-122)

L'usage des corpus oraux pour la recherche sur l'acquisition[1]

Steven GILLIS
Computational linguistics & psycholinguistics (Clips) research center
Universiteit Antwerpen (B)
steven.gillis@uantwerpen.be

1. Introduction

Mon exposé portera sur l'usage des corpus oraux dans la recherche sur l'acquisition du langage. Je parlerai d'abord des corpus utilisés dans ce domaine et en particulier du problème de la rareté des données. Ensuite je présenterai trois solutions à ce problème : la base de données CHILDES[2], la technologie LENA™ [3] et finalement Deb ROY. Enfin, j'aimerais revenir à la question de la rareté des données et proposer quelques pistes pour l'aborder.

Plusieurs options méthodologiques sont envisageables dans le contexte de l'étude du langage enfantin. L'une d'entre elles consiste à mener des expériences comportementales en demandant à l'enfant de jouer certaines actions : par exemple, l'expérimentateur dit « Je pose ma tasse sur la table » et l'enfant doit effectuer cette action. On peut aussi utiliser des tâches de dénomination, des techniques de suivi oculaire ou d'imagerie cérébrale, ou encore l'inventaire parental du développement communicatif, entre autres options.

La méthode que je privilégie repose sur l'utilisation de données observationnelles recueillies en milieu naturel,

[1] *The use of speech corpora in language acquisition research*. Transcription, traduction et adaptation par Guillaume Feigenwinter et Aris Xanthos.
[2] http://childes.psy.cmu.edu/
[3] http://www.lenafoundation.org/

autrement dit des corpus. Je me sers d'enregistrements audio et vidéo de conversations, provenant d'études longitudinales. Mon implication dans ce type de recherche ne date pas d'hier, puisque j'ai commencé à constituer mon premier corpus en 1981. J'ai également participé à la création du Corpus Oral du Néerlandais[4], un corpus riche de dix millions de mots, qu'évoque plus amplement Mirjam ERNESTUS (voir page 65). Plus récemment, j'ai constitué un corpus oral assez étendu, à partir de l'observation de trente enfants suivis de l'âge de six mois jusqu'à deux ans. J'ai donc une certaine expérience dans ce domaine et je peux témoigner de son évolution spectaculaire.

Figure 1 – Magnétophone à bandes, à cassette et enregistreur de minidisques

La figure 1 montre le genre d'équipement que nous utilisions quand j'ai commencé à travailler dans ce champ. Un peu plus récemment, l'invention des magnétophones à cassette a permis de limiter un peu l'encombrement. Pour récolter les données du Corpus Oral de Néerlandais, au début des années 2000, ce sont des enregistreurs de minidisques que nous avons employés.

[4] http://lands.let.ru.nl/cgn/

Figure 2 – Deux modèles de caméra, l'un du début des années 1980, l'autre actuel

Par ailleurs, je me suis plongé dans mes archives à la recherche des caméras que j'avais à disposition pour mon tout premier corpus. Comme on peut l'imaginer, il fallait une grande voiture et un entrainement spécial pour transporter cet équipement. À titre de comparaison, le type de caméra que nous utilisons aujourd'hui est sensiblement plus réduit – l'évolution est effectivement spectaculaire (voir figure 2).

La constitution d'un corpus demande beaucoup de temps et d'énergie, ce qui contribue à expliquer pourquoi l'on dispose de relativement peu de données. Pour illustrer ce point, considérons ce qu'implique une heure d'enregistrement. Premièrement, on peut compter au moins 5 heures pour se rendre au domicile de l'enfant, préparer et effectuer l'enregistrement, puis revenir. Il faut alors importer l'enregistrement sur un ordinateur, le transcrire, et synchroniser la vidéo et la transcription. La transcription orthographique est le cas le plus favorable : à peu près 10,5 heures pour une heure d'enregistrement ; une transcription phonétique simple demandera plutôt de l'ordre de 24 heures. Ensuite, en utilisant un système de transcription phonétique automatique pour les productions des adultes, comme nous le faisons, il faut compter 3 heures supplémentaires. En somme, une heure d'enregistrement demande environ 43

heures de travail, dont la transcription occupe la plus grande partie, soit près de 86% (voir tableau 1).

ACTIVITÉ	TEMPS REQUIS
Visite au domicile de l'enfant, enregistrement	5 h
Importation de la vidéo sur ordinateur	0,5 h
Transcription orthographique et synchronisation avec la vidéo (x8 – x40, moy. 10,5)	10,5 h
Transcription phonétique simple (x14 – x62, moy. 24)	24 h
Vérification de la transcription phonétique automatique des productions des adultes	3 h
Comptabilité et administration	0,5 h
Total	43,5 h

Tableau 1 – Estimation du temps requis pour la production d'un corpus transcrit à partir d'une heure d'enregistrement (sans tests de fiabilité, annotation, etc.)

Supposons maintenant qu'on enregistre ainsi dix enfants pendant une heure mensuellement. Il faut donc grosso modo 435 heures de travail par mois, soit 5220 heures par année. Si l'on imagine que la semaine compte soixante heures de travail, une année d'enregistrement peut alors être traitée en 87 semaines environ. En Belgique, nous ne travaillons officiellement que 38 heures par semaine, donc il nous faudrait plutôt 137 semaines, soit deux ans et demi. En confiant ce travail à un assistant de recherche, qui coute 1500€ par mois, on arrive à un cout total d'environ 51 000€ pour une heure d'enregistrement mensuelle de dix enfants sur une année. Ça n'est pas le plus grand corpus dont on puisse rêver, mais il aura néanmoins demandé une quantité incroyable d'argent, de temps, d'énergie et de frustration. Supposons toutefois qu'un enfant soit éveillé et parle pendant dix heures par jour. Douze heures enregistrées sur une année ne représenteront alors que 0,33% de sa production. Avec un régime d'une heure d'enregistrement par semaine, soit quatre fois plus que dans notre scénario, on n'échantillonnera toujours que 1,33% de la production. Cela reste bien maigre en regard des ressources nécessaires à l'obtention de ces données.

2. La base de données CHILDES

Il existe cependant des solutions à ce problème de rareté des données. L'une d'entre elles consiste à réunir tous les petits corpus disponibles dans le monde entier au sein d'une archive centralisée que tout le monde puisse consulter et exploiter. C'est précisément la vocation de la base de données CHILDES, dans le cas particulier de l'étude du langage enfantin. CHILDES permet d'accéder à des transcriptions (parfois synchronisées avec des fichiers audio et vidéo) ainsi qu'à des logiciels spécifiques pour l'analyse de ces transcriptions, spécifiant par ailleurs des méthodes pour leur annotation linguistique. CHILDES contient des corpus monolingues et bilingues, des récits, ainsi que des données provenant de populations cliniques, par exemple des enfants trisomiques, aphasiques, etc. L'archive complète compte quelque treize millions d'énoncés et cinquante millions de mots (voir tableau 2).

	ENFANTS		ADULTES		TOTAL	
	Énoncés	**Mots**	**Énoncés**	**Mots**	**Énoncés**	**Mots**
Monolingue	4 233 036	12 577 726	7 001 529	30 333 810	11 234 565	42 911 536
Bilingue	391 415	1 410 953	581 080	2 422 625	972 495	3 833 578
Récit	77 160	528 398	40 376	257 211	117 536	785 609
Clinique	224 308	671 129	530 295	2 168 969	754 603	2 840 098
Total	4 925 919	15 188 206	8 153 280	35 182 615	13 079 199	50 370 821

Tableau 2 – Répartition des données par type de corpus dans CHILDES

Ces nombres sont impressionnants, mais il faut prendre en compte leur répartition dans plus de 150 corpus en 39 langues différentes (voir tableau 3 page suivante). Pour cette conférence, j'ai examiné ce qu'on peut trouver dans les corpus de français. Essentiellement, il y a des corpus longitudinaux (deux ou plusieurs observations par enfant) et des corpus transversaux (une seule observation par enfant).

	ENFANTS		ADULTES		TOTAL	
	Énoncés	**Mots**	**Énoncés**	**Mots**	**Énoncés**	**Mots**
Anglais	1 460 992	4 320 926	2 917 501	13 772 058	4 378 493	18 092 984
Allemand	474 258	1 430 003	703 980	3 334 142	1 178 238	4 764 145
Français	227 006	720 222	423 832	2 019 858	650 838	2 740 080
Indonésien	270 930	739 721	540 750	1 606 552	811 680	2 346 273
Espagnol	244 559	856 765	330 918	1 406 854	575 477	2 263 619
Japonais	288 528	791 590	329 372	1 000 938	617 900	1 792 528
Néerlandais	180 670	444 933	271 278	1 130 194	451 948	1 575 127
Polonais	133 420	654 053	96 200	460 285	229 620	1 114 338
Serbe	95 143	219 260	226 853	807 587	321 996	1 026 847
Cantonais	81 038	207 184	147 882	657 602	228 920	864 786

Tableau 3 – Répartition des données de dix langues les plus représentées de CHILDES

On dispose en tout de plus de 2000 enregistrements (au moins une session par enfant et parfois plus de vingt), à des âges variant entre 7 mois et 6 ans et 9 mois. La longueur médiane de ces enregistrements est de 44 énoncés ou 171 mots. C'est donc relativement peu, avec un total de 200 000 énoncés et moins d'un million de mots (voir tableau 4).

	ÉNONCÉS	MOTS
Médiane	44	171
Empan	5 – 709	0 – 4165

Tableau 4 – Longueur des enregistrements dans les corpus en français de CHILDES

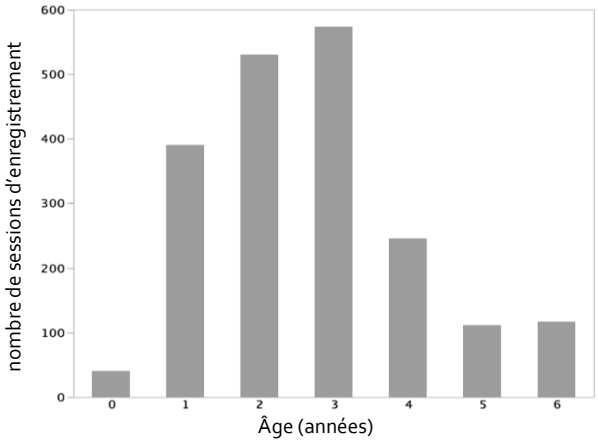

Figure 3 – Corpus français de CHILDES : nombre de sessions d'enregistrement par âge

Comme l'illustre la figure 3 (page précédente), l'examen du nombre de sessions d'enregistrement en fonction de l'âge montre que c'est entre 2 et 3 ans que les données sont les plus fournies. On voit par ailleurs que les enregistrements diffèrent drastiquement du point de vue de leur longueur (voir figure 4 ci-dessous, et figure 5 page suivante). Certains enregistrements comptent jusqu'à 700 énoncés, mais les longueurs médianes sont beaucoup plus faibles, et l'on observe à peu près le même phénomène concernant le nombre de mots : à deux ou trois ans, certains enregistrements contiennent jusqu'à 4500 mots, mais la plupart sont nettement plus courts. En somme, si l'on dispose effectivement de données en français, leur distribution est inégale.

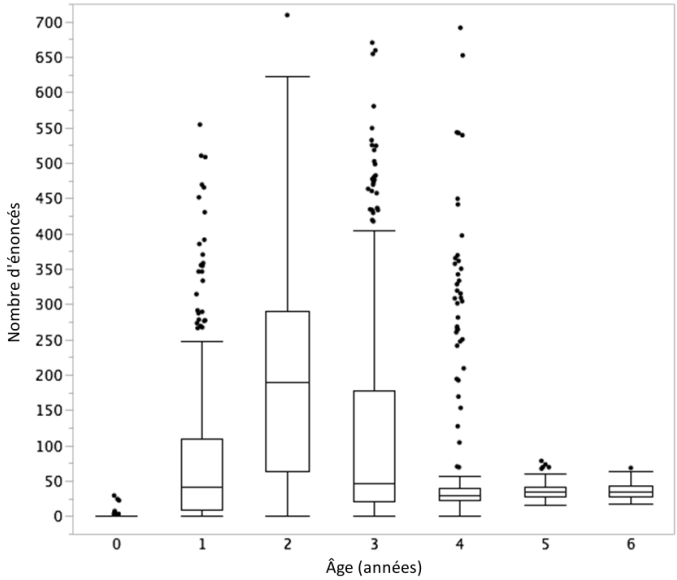

Figure 4 – Corpus français de CHILDES : distribution du nombre d'énoncés selon l'âge (langage enfantin)

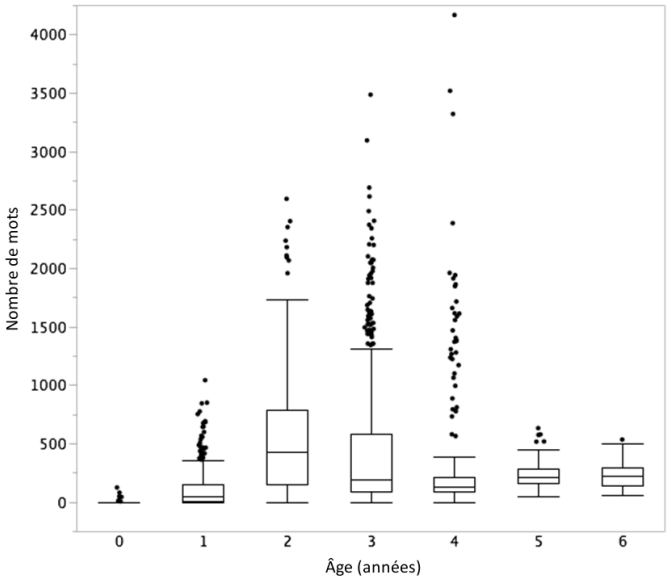

Figure 5 – Corpus français de CHILDES : distribution du nombre de mots selon l'âge (langage enfantin)

Afin de suivre le développement des enfants de manière plus précise, il est parfois souhaitable d'analyser des enregistrements sur une base mensuelle plutôt qu'annuelle. Or, comme il apparait dans la figure 6 (page suivante), on ne dispose d'un nombre raisonnable d'enregistrements que pour des enfants de 2 à 3 ans ; en particulier il y a peu de données d'enfants de 5 ou 6 ans. La même conclusion s'impose en ce qui concerne le nombre de mots dans les productions enfantines : il y a jusqu'à 40 000 mots pour certains mois entre deux et trois ans, mais c'est exceptionnel (voir figure 7 page suivante). Enfin, on dispose de données pour près de 90 enfants de 3 ans, mais pour les autres âges, les effectifs sont bien moindres (voir figure 8 page 104).

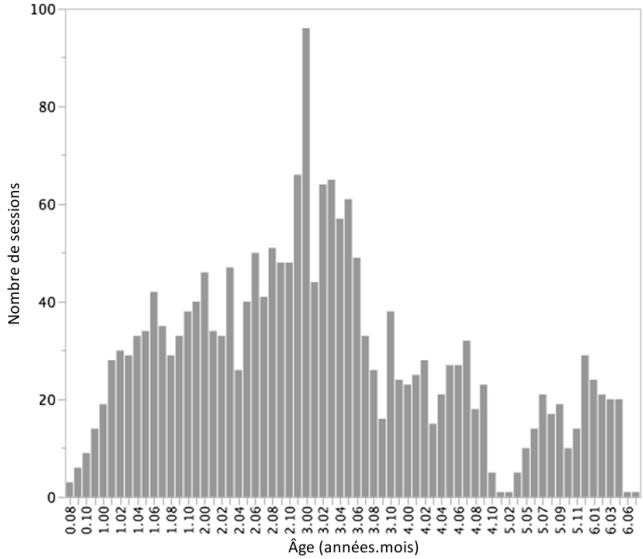

Figure 6 – Corpus français de CHILDES : nombre de sessions d'enregistrement par âge

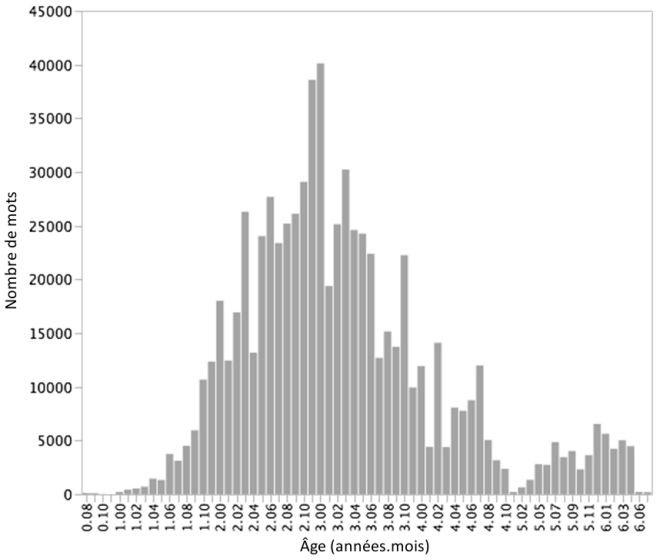

Figure 7 – Corpus français de CHILDES : nombre de mots par âge (langage enfantin)

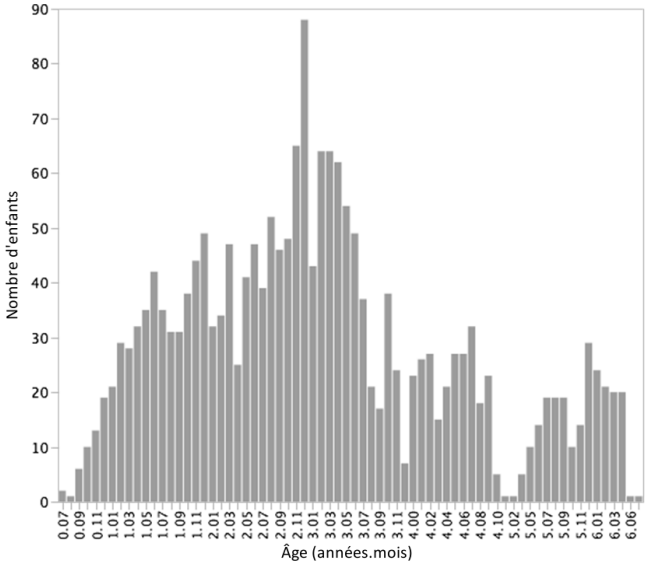

Figure 8 – Corpus français de CHILDES : nombre d'enfants par âge

On voit ainsi que de rassembler toutes sortes de corpus dans une base de données unique telle que CHILDES contribue à résoudre une partie du problème de la rareté des données : on peut en obtenir ainsi plus qu'on pourrait jamais en récolter par soi-même. Toutefois elles sont distribuées de façon inégale et, en définitive, il n'y en a pas tant que cela.

On me demande souvent : « est-ce que CHILDES peut suffire pour ma recherche ? ». Ma réponse est toujours la même : tout dépend de la question de recherche, de l'objet et des objectifs de l'étude. Par exemple, la question de recherche vous permet-elle d'agréger des données ? Si vous travaillez sur le français et que vous pouvez vous permettre d'agréger les données de tous les enfants de 2 et 3 ans, alors CHILDES fournira une bonne base. L'étude implique-t-elle une dimension longitudinale ? et ainsi de suite. Toutes ces questions jouent un rôle, et voici une statistique qui résume bien la situation : pour le français, entre 18 mois et 4 ans, on dispose pour chaque mois de données concernant plus de 10

enfants (médiane 42,5, empan 17 – 88), de plus de 2000 énoncés (médiane 6303, empan 2443 – 11 565), enfin de plus de 3000 mots (médiane 20 776, empan 3120 – 39 093). Donc, pour cette tranche d'âge, si cette quantité de données est suffisante, on peut se satisfaire de CHILDES.

Pour ma part, je ne peux pas savoir quelle quantité est nécessaire pour une étude donnée, parce que cela dépend aussi d'éléments tels que le régime d'échantillonnage : à quelle fréquence les données sont-elles échantillonnées ? Pour illustrer l'importance de ce paramètre, supposons qu'on étudie le lexique cumulatif, soit le nombre cumulé de mots distincts qu'un enfant acquiert : par exemple, la relation entre la richesse lexicale et le développement phonologique, la spécificité lexicale, ou encore les constructions lexicales spécifiques. Quelle est alors l'influence du régime d'échantillonnage ? Considérons le cas du lexique cumulatif observé dans un corpus que j'ai constitué à partir d'enregistrements d'un enfant nommé Maarten (voir figure 9 page suivante) : en échantillonnant toutes les deux semaines, le lexique croît de 20 à 250 mots environ. Toutefois lors de la constitution de ce corpus, j'ai effectué des enregistrements tous les 4 jours, et en calculant le lexique cumulatif sur cette base, on obtient un maximum de 739 mots au terme de la période. L'évaluation du lexique cumulatif est crucialement dépendante du régime d'échantillonnage. Je suggère donc – et c'est assez intuitif – d'utiliser le même régime d'échantillonnage pour comparer des enfants dans une perspective longitudinale.

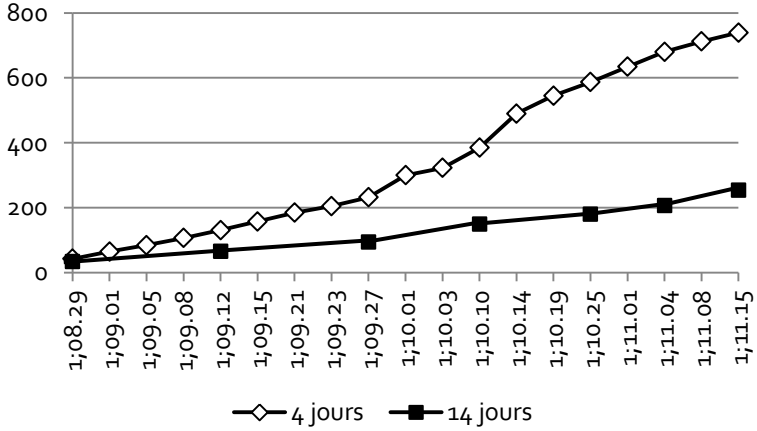

Figure 9 – Développement du lexique cumulatif selon le régime d'échantillonnage (enfant Maarten)

De ce point de vue, l'usage de CHILDES pose problème : certains enfants sont enregistrés toutes les semaines, d'autres toutes les deux semaines, d'autres encore tous les mois, voire tous les trois mois, par exemple. À cela s'ajoutent d'autres difficultés, dans la mesure où ces enregistrements sont de durées variables : certains ne durent qu'une demi-heure, d'autres une heure, d'autres encore sont plus longs. Il y a des différences en termes de nombre de mots et d'énoncés, ainsi que du point de vue de la volubilité des enfants. Même avec des enregistrements de durée égale, par exemple une heure, peut-être qu'un enfant parlera beaucoup et produira 500 énoncés, alors qu'un autre parlera beaucoup moins et n'en produira qu'une centaine. Cela aura aussi des implications selon le type de mesures utilisées. Par exemple, lorsqu'on s'intéresse au nombre de mots distincts (types) dans les productions d'un enfant, on peut constater qu'il est fonction du nombre d'occurrences (tokens) dans le corpus (voir figure 10 page suivante).

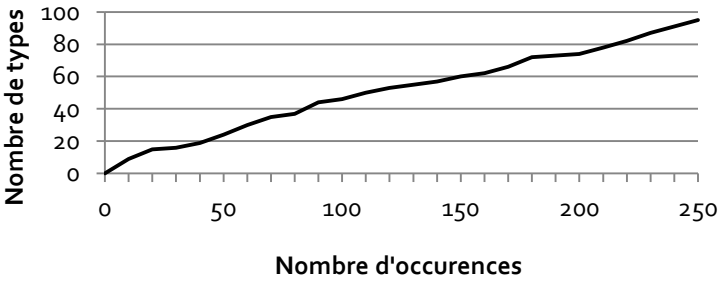

Figure 10 – Dépendance entre nombre de types et nombre d'occurrences

De la même façon, le rapport types-tokens, un indice très courant dans l'étude de l'acquisition, dépend dans une certaine mesure de la taille de l'échantillon (voir figure 11).

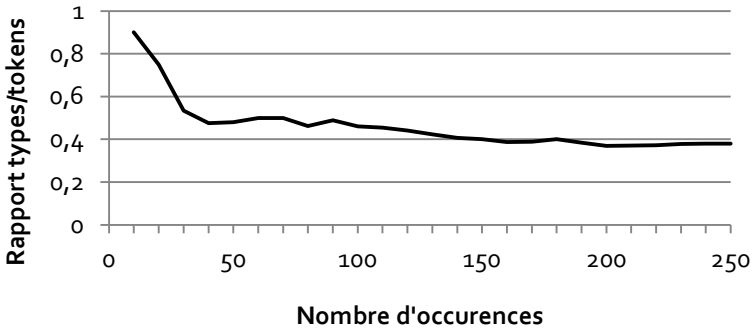

Figure 11 – Dépendance entre rapport types/tokens et nombre d'occurrences

Lorsqu'on compare des enfants du point de vue de certaines mesures, il est donc essentiel d'utiliser le même régime d'échantillonnage : nombre d'échantillons par période de temps, nombre d'unités linguistiques (mots, énoncés) par échantillon, etc. TOMASELLO a formulé des remarques très précieuses à ce sujet, notamment dans un article au titre fort bien trouvé « How much is enough ? » (TOMASELLO & STAHL 2004). Les contributions de HUTCHINS, BRANNICK, BRYANT & SILLIMAN (2005) et ROWLAND, FLETCHER & FREUDENTHAL

(2008) sont également pertinentes dans ce contexte. Le propos commun à ces publications est qu'on n'a pas prêté assez attention aux aspects quantitatifs de l'étude des données de parole spontanée. Il faut aller beaucoup plus loin dans l'examen des fondements quantitatifs de la collecte et l'usage de ces données :

> There has been relatively little discussion in the field of child language acquisition about how best to sample from children's spontaneous speech, particularly with regard to quantitative issues. (TOMASELLO & STAHL 2004, 101).

Pour en revenir au régime d'échantillonnage, l'une des conclusions de TOMASELLO & STAHL est qu'il faut être attentif à la granularité des données. Pour répondre à la question de la quantité de données nécessaire pour l'étude d'un phénomène particulier (un type d'erreur, de construction, de mot, etc.), il est possible d'estimer le nombre de fois qu'on peut s'attendre à observer ce phénomène en fonction de la proportion de données échantillonnées. C'est ce que TOMASELLO appelle le taux d'observation (« hit rate ») : la probabilité d'observer au moins une fois le phénomène-cible dans un échantillon. Sans entrer dans les détails mathématiques, on peut comprendre le raisonnement sous-jacent à partir de la figure 12, donnant sur l'axe horizontal, le nombre d'heures d'enregistrement par semaine, et sur l'axe vertical, le taux d'observation ; la ligne pointillée horizontale indique le taux minimal pour être raisonnablement sûr d'observer au moins une occurrence du phénomène en question, soit 95%.

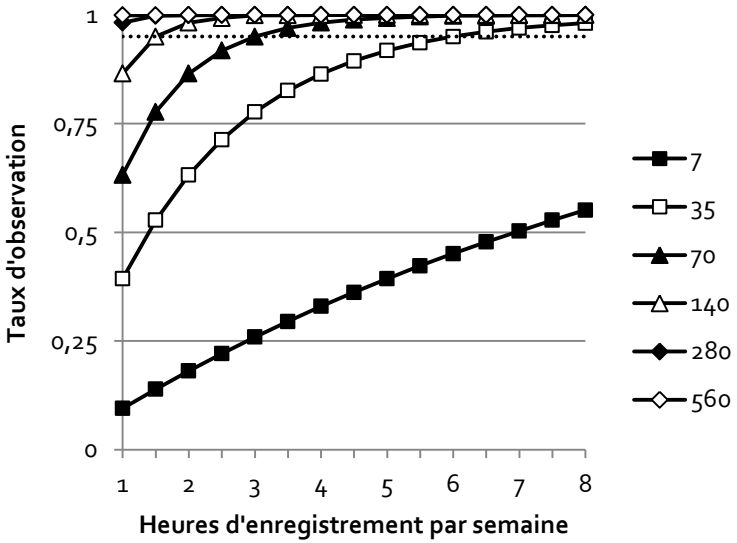

Figure 12 – Taux d'observation d'un phénomène en fonction du nombre d'heures d'enregistrement par semaine et du nombre d'occurrences attendues par heure

Prenons l'exemple d'un phénomène dont la fréquence attendue est de 7 occurrences par heure. Dans ce cas, même en enregistrant 8 heures par semaine, la probabilité de détecter le phénomène est à peine en-dessus de 50%. Si l'on s'attend à ce que le phénomène se produise 35 fois par heure, il faudra tout de même enregistrer 5 à 6 heures par semaine pour atteindre la barre des 95%. Avec 70 occurrences par heure, il faut enregistrer entre 2,5 et 3 heures par semaine, et si on s'attend à ce que le phénomène se produise 560 fois par heure, alors il n'y a plus besoin que d'une demi-heure d'enregistrement hebdomadaire. La leçon à tirer de tout cela est triviale : il est vraisemblable d'observer des phénomènes fréquents même si la fréquence d'échantillonnage est faible ; en revanche la combinaison d'un phénomène rare et d'un échantillonnage parcimonieux implique que les estimations de fréquence seront sans doute très peu fiables. Pour cette

raison, le moins que l'on puisse dire en examinant le détail de certaines études dans la littérature est qu'il n'est pas certain que les mesures utilisées soient fiables. Et il y a lieu de s'inquiéter lorsque des chercheurs affirment que « les enfants ne font pas ceci à l'âge de… ». La suggestion, dès lors, est que les chercheurs donnent une indication quant au degré de confiance qu'on peut avoir dans les fréquences qu'ils rapportent – c'est à mon sens une question d'éthique scientifique.

3. La technologie LENA™

Je n'ai découvert la seconde solution au problème de la rareté des données que récemment : il s'agit du système LENA™. Il s'agit d'un appareil de taille assez réduite (il ne pèse que 70 grammes) qui peut être placé dans les vêtements d'un enfant pour faire des enregistrements de 12 à 16 heures (voir figure 13).

Figure 13 – L'enregistreur LENA™ [5]

J'ai indiqué précédemment que la transcription était un problème de taille : il faut près de 40 heures pour transcrire une heure d'enregistrement. Dès lors, que faire avec 12 à 16 heures d'enregistrement quotidien ? Dans un scénario de science-fiction, un logiciel pourrait se charger de la

[5] Source : http://shop.lenafoundation.org/products/97-lena-digital-language-processor-dlp.aspx

transcription. Et il se trouve que LENA propose ce type de logiciel, dans une certaine mesure. La question de la transcription automatique est abordée plus longuement dans la contribution de Mirjam ERNESTUS (voir page 65), aussi je n'en parlerai que brièvement ici. Le logiciel LENA Pro[6] est capable de segmenter le signal de parole et d'identifier les sources les plus probables. Il peut ainsi déterminer quelles parties de l'enregistrement correspondent à la parole de l'enfant ou d'un adulte, au bruit de la télévision ou de la radio, ou encore à du silence, et il segmente automatiquement l'enregistrement en fonction de ces catégories (figure 14).

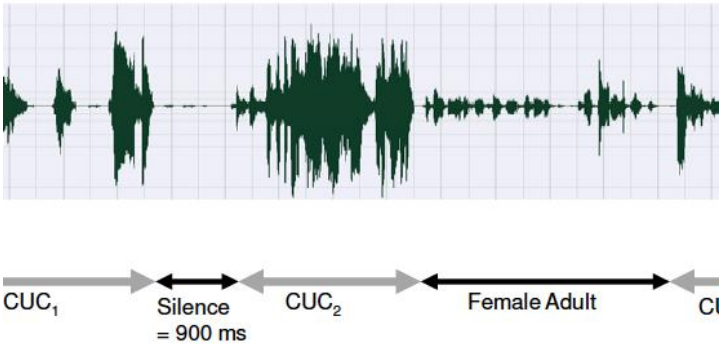

Figure 14 – Segmentation et catégorisation du signal audio par LENA Pro (OLLER *et al.* 2010, Supporting Information Appendix: p. 16)

CUC est l'abréviation de « child utterance cluster » (*groupe d'énoncés de l'enfant*)

		Système LENA (en %)			
		Adulte	Enfant	TV	Autre
Transcription humaine (en %)	Adulte	**82**	2	4	12
	Enfant	7	**76**	0	17
	TV	8	0	**71**	21
	Autre	14	4	6	**76**

Tableau 5 – Concordance en % de la segmentation produite par LENA Pro et par des experts humains (XU, YAPANEL & GRAY 2009, 5)

[6] http://www.lenafoundation.org/lena-pro/

Cette analyse automatique est assez fiable : dans trois cas sur quatre, la décision du système concorde avec l'avis des transcripteurs humains, ce qui est une proportion raisonnable (voir tableau 5 page précédente). De plus, le logiciel peut segmenter le signal plus finement (voir figure 15 ci-dessous). Il peut ainsi découper les groupes d'énoncés de l'enfant en groupes de souffle (« child utterances », CU) sur la base des silences. Puis il identifie les ilots de vocalisation (*child vocal islands*, *CVI*), c'est-à-dire les portions de segments où l'énergie est la plus élevée, par exemple les voyelles prononcées par l'enfant. Ensuite il distingue parmi les ilots de vocalisation ceux qui sont liés à la parole (« speech-related vocal islands », SVI) et les autres (pleurs et sons végétatifs), de manière assez fiable (voir tableau 6 page suivante). Enfin les ilots de vocalisation liés à la parole sont regroupés pour former les énoncés linguistiques (« speech-related child utterances », SCU) à proprement parler.

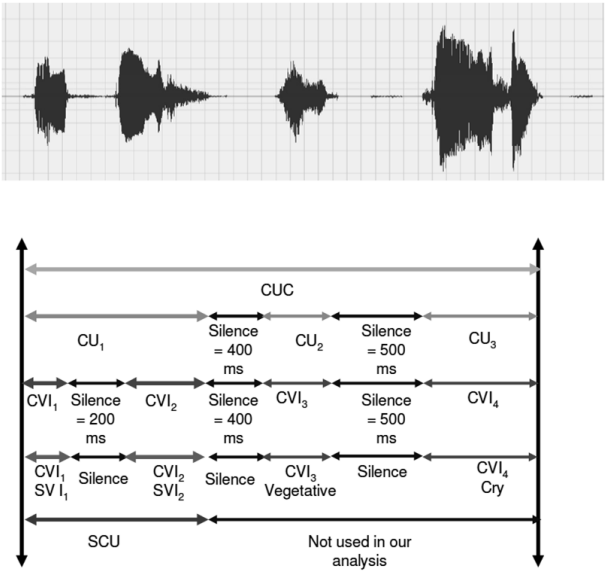

Figure 15 – Segmentation et catégorisation hiérarchique du signal audio par LENA Pro (OLLER *et al.* 2010, Supporting Information Appendix: p 18)

		Système LENA (en %)	
		SVI	Pleurs / sons végétatifs
Classification humaine (en %)	SVI	**75**	25
	Pleurs / sons végétatifs	16	**84**

Tableau 6 – Concordance entre classification des ilots de vocalisation par LENA Pro et par des experts humains (OLLER *et al.* 2010, Supporting Information Appendix: p 26)

Par ailleurs, le système LENA est capable d'identifier les mots en détail, non seulement dans la parole des enfants, mais aussi dans celle des adultes. Les deux décomptes coïncident à peu près dans les cas favorables, comme celui d'une mère qui joue tranquillement avec son enfant à la maison. Dans un environnement plus bruyant, la machine identifie moins d'éléments, et de manière moins précise (XU *et al.* 2009, 9-11)

En dépit de ces problèmes, les résultats obtenus avec le système LENA sont intéressants : le logiciel permet de visualiser le nombre de mots produits par l'enfant ou par l'adulte à chaque heure de la journée, et toutes sortes d'autres vues d'ensemble (voir figure 16 ci-dessous, et figure 17 page suivante).

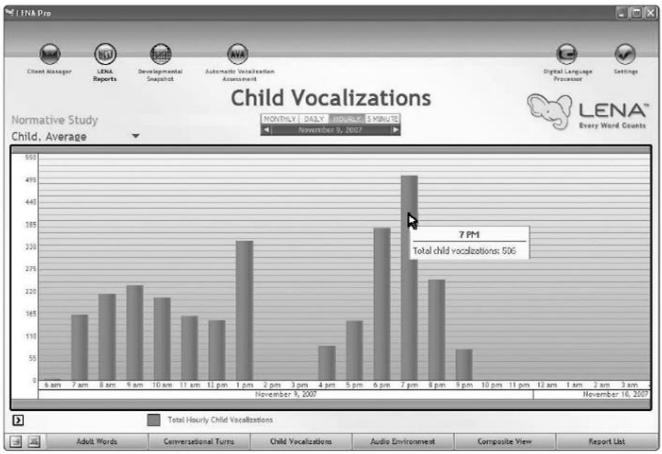

Figure 16 – Nombre moyen de vocalisations de l'enfant par heure de la journée, calculé par LENA Pro (LENA Pro Brochure 2012 : 5)

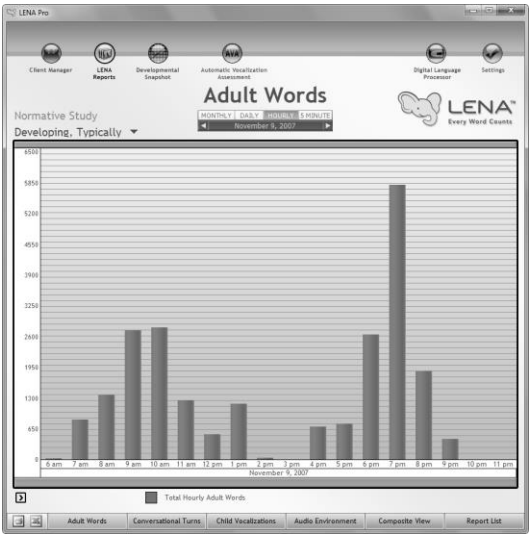

Figure 17 – Nombre moyen de mots dans les productions adultes par heure de la journée, calculé par LENA Pro (LENA Pro Brochure 2012, 3)[7]

Pour prendre la mesure de l'utilité de ces résultats, on peut rappeler l'étude de HART & RISLEY (1995), qui visait à détecter des différences significatives dans la parole adressée par les adultes aux enfants et dans le développement du langage en fonction de la classe sociale. Il s'agissait d'une très vaste étude longitudinale, avec 42 familles et des observations mensuelles d'une heure, de 7 mois à 3 ans, pour un total de 1318 heures d'enregistrement. HART & RISLEY évoquent dans un passage de leur livre qu'après avoir assemblé ce corpus, ils ont passé les six années suivantes à le transcrire. Le système LENA a été utilisé pour réaliser une étude comparable, mais bien plus vaste encore (GILKERSON & RICHARDS 2009) : plus de 300 familles, des observations mensuelles de 12 heures, sur une période totale plus longue

[7] Illustrations disponibles dans la documentation de la fondation : http://www.lenafoundation.org/wp-content/uploads/2014/10/LTR-11-1_LENA-Pro-Brochure.pdf

d'une année : au total, 32 000 heures d'observation au lieu de 1300 – la différence est considérable (voir tableau 7).

	HART & RISLEY	GILKERSON & RICHARDS
Nombre de familles	42	329
Fréquence des sessions	mensuelle	mensuelle
Tranche d'âge	0;7 – 3;0	0;2 – 4;0
Durée des sessions (heures)	1	12
Durée totale d'observation (heures)	1318	32 000+

Tableau 7 – Comparaison de HART & RISLEY (1995) et GILKERSON & RICHARDS (2009)

Le tableau 8 montre la répartition cumulée du nombre quotidien de mots et de tours de paroles dans ces données à l'âge de 24 mois et l'on peut voir que les parents ne parlent pas tous autant à leurs enfants. Par exemple, un enfant de cet âge entend presque 30 000 mots par jour au 99ème centile, contre 6000 au 10ème centile, soit presque 5 fois moins ; la proportion est comparable en ce qui concerne le nombre de tours de parole. Une différence significative peut également être observée concernant le nombre de mots produits par l'enfant, soit 4500 mots au 99ème centile mais seulement le quart au 10ème centile. Ces décomptes ont été effectués avec la technologie LENA.

	ADULTES		ENFANTS
Centile	Mots	Tours de parole	Mots
99	29 428	1163	4406
90	20 824	816	3184
80	17 645	688	2728
70	13 338	603	2422
60	13 805	535	2174
50	12 297	474	1955
40	10 875	418	1747
30	9451	361	1538
20	7911	300	1310
10	6003	225	1024

Tableau 8 – Répartition cumulée du nombre quotidien de mots et de tours de parole dans la production des adultes et des enfants à l'âge de 24 mois (GILKERSON & RICHARDS 2009 : 10)

Dans la figure 18, qui présente une comparaison entre deux niveaux de formation, on peut voir qu'en moyenne, le nombre quotidien de mots est plus faible dans les productions d'adultes dont le niveau de formation est plus bas que chez ceux dont le niveau de formation est plus élevé. Ce résultat confirme la conclusion de HART & RISLEY (1995) de l'existence de différences significatives en fonction de la classe socioéconomique.

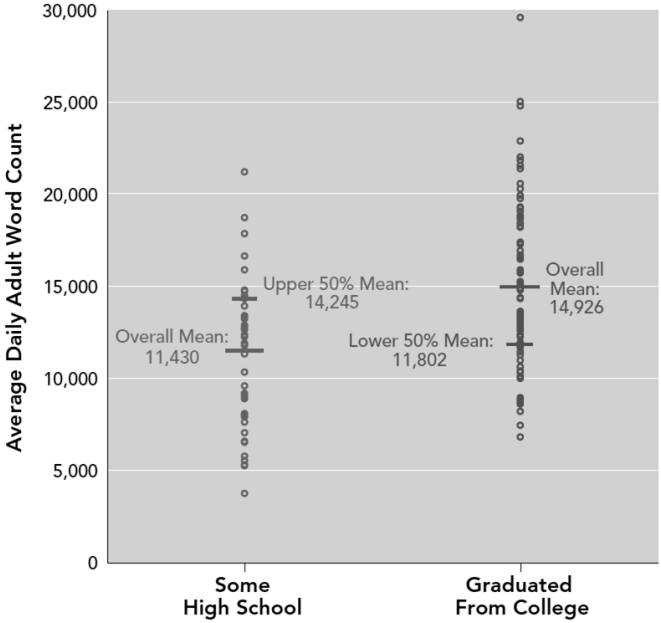

Figure 18 – Distribution du nombre moyen de mots par jour dans la production des adultes en fonction du niveau socioéconomique (GILKERSON & RICHARDS 2009 : 20)

En tant que solution au problème de la rareté des données, la technologie LENA est très prometteuse et constitue un pas concret en direction de la transcription automatique. Toutefois des améliorations sont nécessaires et c'est un problème de ne disposer que de l'audio et non de la vidéo. Quiconque a déjà transcrit des paroles d'enfant sait à

quel point la vidéo est essentielle pour permettre de comprendre ce que l'enfant dit précisément, grâce aux informations contextuelles qu'elle fournit. En conclusion, la solution LENA fonctionne jusqu'à un certain point et peut d'ores et déjà être utilisée pour effectuer des décomptes approximatifs.

4. L'approche de Deb ROY

J'aimerais encore évoquer une troisième solution. C'est une alternative qu'a trouvée Deb ROY (2011) : avec suffisamment d'argent, il est possible d'acheter un terrain et la maison qui va avec, équiper celle-ci de nombreux micros et caméras, et ensuite tout enregistrer. C'est naturellement le fantasme de tout chercheur en acquisition du langage que de disposer d'enregistrements continus de tout ce que les enfants entendent et disent. Mais considérons les chiffres : trois ans d'enregistrement, près de 10 000 heures de vidéo. ROY était affilié au MIT et avait à sa disposition beaucoup d'argent et une armée d'assistants de recherche. Or, en trois ans, un enfant de parents volubiles entend en moyenne un peu plus de 26 millions de mots, et en produit en moyenne près de 3,5 millions s'il est lui aussi volubile. ROY et ses nombreux assistants de recherche ont transcrit 7 millions de mots, soit à peine le quart des données qu'il a récoltées, ce qui est peu. Cela souligne l'importance, dans un futur proche, d'être capable de transcrire automatiquement, afin de pouvoir laisser de côté ces tout petits corpus, qui coutent beaucoup d'argent et d'énergie – je le sais pour en avoir constitué moi-même – mais qui restent des corpus miniatures. Il faut des données massives de ce type pour faire avancer la recherche dans ce domaine.

Questions

Légende : « Q » pour « question », « SG » pour Steven GILLIS.

Q : Vous parlez beaucoup d'échantillonnage, et vous avez montré de manière très convaincante que les échantillons dont nous disposons sont trop petits pour être utilisés de manière légitime dans des études détaillées. Mais que dire du contenu des données ? Quand vous agrégez des données, est-ce que vous êtes d'accord avec ce que vous voyez dans les données transcrites ?

SG : Regrouper des données est toujours une entreprise risquée. Par exemple en discutant des transcriptions phonétiques dans le Corpus Oral du Néerlandais, nous avions observé des pratiques différentes alors même que les transcripteurs travaillaient tous au sein du même projet. Donc si vous collectez et rassemblez des données provenant de divers projets, et que ces derniers ne sont pas bien documentés, c'est-à-dire que vous ignorez ce que les chercheurs ont fait exactement, avec quels protocoles, comment ils ont géré certains types de difficultés, alors cela peut engendrer de sérieux problèmes. En définitive, tout dépend de la question de recherche.

Q : Vous avez posé la question « Quelle quantité est suffisante pour que l'échantillon soit représentatif ? » et j'aimerais vous demander : pour être représentatif de quoi ?

SG : Ultimement, de la population.

Q : Mais dans la base de données CHILDES, il y a beaucoup de langues différentes, donc votre point de départ est de dire que toutes les langues sont égales.

SG : Non, je n'ai jamais dit ça. Ce que vous pourriez assembler, si vous étudiez le français par exemple, ce sont tous les corpus en français. Mais vous pourriez aussi traiter de questions translinguistiques telles que « Quelle est la progression du nombre de mots que les enfants apprennent et produisent en français, comparé à l'arabe ou au néerlandais, etc. ». Mais je n'ai jamais dit que cela devrait être représentatif du langage avec un grand « L ». J'ai moi-même conduit des recherches sur plusieurs langues en comparant l'acquisition du néerlandais et de l'hébreu, et je serais

bien le dernier à vous dire que le néerlandais et l'hébreu sont exactement les mêmes langues, évidemment.

Q : J'aimerais revenir au problème de la rareté des données, qu'on rencontre partout mais pour lequel il existe bien sûr diverses solutions. Dans le cas du rapport types/tokens, par exemple. Il y a des stratégies de normalisation et encore bien d'autres mesures. Je pense donc qu'il est possible de trouver des réponses dans l'usage de la statistique, jusqu'à un certain point.

SG : C'est ce que j'aurais présenté si j'avais eu plus de temps. L'exemple du rapport types/tokens est bien connu. Ce que nous avons publié avec Aris XANTHOS, par exemple, sont des méthodes de normalisation que vous pouvez employer dans le domaine de la richesse morphologique (XANTHOS & GILLIS 2010), et il y a toutes sortes de solutions pour cela. Mais vous ne pouvez jamais compenser le problème du manque de données. Vous pouvez essayer de le normaliser, et vous arriverez à certains résultats, mais au final, vous ne pouvez pas le résoudre.

Q : Je pense que les mesures basées sur l'entropie peuvent offrir une réponse partielle. Et les modèles de langage basés sur la fréquence d'usage sont aussi prometteurs, parce que les phénomènes les plus fréquents sont aussi les plus importants.

SG : Je ne peux qu'être d'accord avec vous.

Q : J'ai une autre question à propos du temps nécessaire à la transcription d'une heure de parole d'enfants. Vous avez dit passer dix heures sur la transcription orthographique, ce qui m'a surpris, parce que dans mon expérience avec des conversations entre adultes, nous arrivons souvent à près de quarante heures de transcription pour une heure d'enregistrement. Savez-vous d'où cette différence peut venir ? Est-ce que c'est à cause des rires et des pleurs, ou bien les enfants parlent-ils moins ?

SG : C'était une moyenne, comparable au Corpus Oral du Néerlandais, par exemple. Certaines parties de ce corpus étaient facile à transcrire, parce qu'elles avaient été enregistrées dans des conditions tranquilles, ou alors ce sont des textes lus à voix haute, etc. À l'autre extrême, vous arriverez à quarante voire soixante heures, parce que ce ne sont plus des monologues mais des conversations avec trois, quatre ou cinq locuteurs qui parlent

ensemble. Cela rend la tâche très difficile. Dans notre cas, ce sont des petits enfants, à un stade pré-linguistique, donc la transcription orthographique est relativement limitée. Dans un environnement calme, une mère parle beaucoup plus lentement lorsqu'elle s'adresse à son enfant que si elle parlait de football. Mais c'est assez comparable, en termes de variation, avec ce qu'on a dans le Corpus Oral du Néerlandais. C'est juste que nous avons bien plus de données « faciles ».

Q : Durant votre présentation, plusieurs questions me sont venues, dont la suivante : avez-vous tenté de mélanger des transcriptions automatiques et humaines ? Pas pour les enfants, mais peut-être que vous pourriez utiliser la transcription automatique pour les adultes, puisqu'ils parlent très lentement, quand ils s'adressent aux enfants.

Mirjam ERNESTUS : Ça n'est pas le cas. Les énoncés adressés aux enfants par les adultes sont également très difficiles à transcrire. Tout ce qui est spontané est difficile à transcrire.

Q : Alors des tentatives ont déjà été faites pour tester ce que je propose, mais n'ont pas obtenu de bons résultats ?

Mirjam ERNESTUS : Oui, en fait le pourcentage de mots transcrits correctement est très élevé pour les textes dictés : on arrive presque à 100%. Mais dès que vous essayez de faire la même chose avec de la parole spontanée, vous tombez à 20 ou 30%.

Q : D'accord. Ma deuxième question concernait les types et les occurrences. Vous avez dit que le nombre de types dépend du nombre d'occurrences, ce qui est plutôt intuitif. Est-ce que cela vaut aussi pour les adultes, quand ils parlent à des enfants ?

SG : Je ne pense pas qu'il y aurait de grande différence.

Q : Est-ce que cela a été analysé, ou est-ce une supposition de votre part ?

SG : C'est une supposition de ma part.

Q : Je pense qu'il serait intéressant d'étudier ce qui se passe à ce niveau.

Elena TRIBUSHININA : Il y a eu des études au sujet de la diversité de la parole adressée aux enfants, révélant d'énormes différences entre des parents éduqués qui utilisent un vocabulaire très varié

pour parler à leurs enfants, et d'autres qui emploient toujours les mêmes mots. Donc ça dépend.

SG : Il a été montré que les parents qui ont un niveau socioéconomique faible emploient un lexique plutôt limité. Ceux qui ont un niveau socioéconomique élevé utilisent beaucoup, beaucoup plus de mots différents, ce qui fait que leur fréquence est plus basse, mais leur diversité bien supérieure.

Q : Donc les conditions sociales et l'éducation, ce genre de choses, ont une influence importante ?

SG : Oui. Un point que j'aurais dû soulever est que je vous ai parlé de *l'*enfant, alors que *cet* enfant n'existe pas. On peut le voir dans les données socioéconomiques : les enfants pauvres, nés dans un environnement de faible niveau socioéconomique, reçoivent beaucoup moins d'input, apprennent beaucoup moins. Les résultats de scanners cérébraux à six ou sept mois montrent qu'ils ont moins de tissus et un niveau d'activation moindre. Donc si un enfant est né dans un environnement pauvre, les conséquences sont immédiatement perceptibles. Je n'aurais donc pas dû parler de *l'*enfant, tout comme je n'aurais pas dû intituler un livre « *The* acquisition of Dutch » (GILLIS & DE HOUWER 1998).

Références

GILKERSON Jill & RICHARDS Jeffrey A. (2009), The Power of Talk: Impact of Adult Talk, Conversational Turns, and TV during the Critical 0-4 Years of Child Development, Boulder, LENA Research Foundation, *LENA Technical Report*: *ITR-01-2*.

GILLIS Steven & DE HOUWER Annick (1998), *The Acquisition of Dutch*, Amsterdam/Philadelphia: John Benjamins.

HART Betty & RISLEY Todd R. (1995), *Meaningful Differences in the Everyday Experience of Young American Children*, Baltimore: Paul H. Brookes.

HUTCHINS Tiffany L., BRANNICK Michael, BRYANT Judith B. & SILLIMAN Elaine R. (2005), Methods for Controlling Amount of Talk: Difficulties, Considerations and Recommendations, *First Language 25-3*, 347-363.

OLLER D. Kimbrough, NIYOGI Partha, GRAY Sharmistha, RICHARDS Jeffrey A., GILKERSON Jill, XU Dongxin, YAPANEL Umit & WARREN Steven F. (2010), Automated Vocal Analysis of Naturalistic Recordings from Children with Autism, Language Delay, and Typical Development, *Proceedings of the National Academy of Sciences 107-30*, 13354-13359.

ROWLAND Caroline F., FLETCHER Sarah L. & FREUDENTHAL Daniel (2008), How Big is Big Enough? Assessing the Reliability of Data from Naturalistic Samples, in BEHRENS Heike (Ed.), *Corpora in Language Acquisition Research. History, methods, perspectives*, Amsterdam/Philadelphia: John Benjamins, 1-24.

ROY Deb (2011), The Birth of a Word, http://www.ted.com/talks/deb_roy_the_birth_of_a_word, [22/08/2015].

TOMASELLO Michael & STAHL Daniel (2004), Sampling Children's Spontaneous Speech: How Much is Enough?, *Journal of Child Language 31-01*, 101-121.

XANTHOS Aris & GILLIS Steven (2010), Quantifying the Development of Inflectional Diversity, *First Language 30-2*, 175-198.

XU Dongxin, YAPANEL Umit & GRAY Sharmi (2009), Reliability of the LENA™ Language Environment Analysis System in Young Children's Natural Home Environment, Boulder, LENA Reseach Foundation, *LENA Technical Report*: *LTR 05-2*.