Nouvelles technologies et standards méthodologiques en linguistique

Cahiers de l'ILSL n° 45, 2016

Ont déjà paru dans cette série :

Lecture de l'image (1992, n°1)
Langue, Littérature et altérité (1992, n°2)
Relations inter- et intraprédicatives (1993, n°3)
Travaux d'étudiants (1993, n°4)
L'Ecole de Prague : l'apport épistémologique (1994, n°5)
Fondements de la recherche linguistique : perspectives épistémologiques (1996, n°6)
Formes linguistiques et dynamiques interactionnelles (1995, n°7)
Langues et nations en Europe centrale et orientale (1996, n°8)
Jakobson entre l'Est et l'Ouest, 1915-1939 (1997, n°9)
Le travail du chercheur sur le terrain (1998, n°10)
Mélanges en hommage à M. Mahmoudian (1999, n°11)
Le paradoxe du sujet : les propositions impersonnelles dans les langues slaves et romanes (2000, n°12)
Descriptions grammaticales et enseignement de la grammaire en français langue étrangère (2002, n°13)
Le discours sur la langue en URSS à l'époque stalinienne (2003, n°14)
Pratiques et représentations linguistiques au Niger (2004, n°15)
Le discours sur la langue sous les pouvoirs autoritaires (2004, n°17)
Le slipping dans les langues médiévales (2005, n°18)
Travaux de linguistique (2005, n°19)
Un paradigme perdu : la linguistique marriste (2005, n°20)
La belle et la bête: jugements esthétiques en Suisse romande et alémanique sur les langues (2006, n°21)
Etudes linguistiques kabyles (2007, n°22)
Langues en contexte et en contact (2007, n°23)
Langage et pensée : Union Soviétique, années 1920-30 (2008, n°24)
Structure de la proposition (histoire d'un métalangage) (2008, n°25)
Discours sur les langues et rêves identitaires (2009, n°26)
Langue et littératures pour l'enseignement du français en Suisse romande: problèmes et perspectives (2010, n°27)
Barrières linguistiques en contexte médical (2010, n°28)
Russie, linguistique et philosophie (2011, n°29)
Plurilinguismes et construction des savoirs (2011, n°30)
Langue(s). Langage(s). Histoire(s). (2011, n°31)
Identités en confrontation dans les médias (2012, n°32)
Humboldt en Russie (2013, n°33)
L'analyse des discours de communication publique (2013, n°34)
L'édification linguistique en URSS : thèmes et mythes (2013, n°35)
Mélanges offerts en hommage à Remi Jolivet (2013, n°36)
Histoire de la linguistique générale et slave : "sciences et traditions (2013, n°37)
Ireland and its Contacts/L'Irlande et ses contacts (2013, n°38)
La linguistique urbaine en Union Soviétique (2014, n°39)
La linguistique soviétique à la recherches de nouveaux paradigmes (2014, n°40)
Le niveau méso-interactionnel : lieu d'articulation entre langage et activité (2014, n°41)
L'expertise dans les discours de la santé : du cabinet médical aux arènes publiques (2015, n°42)
L'école phonologique de Leningrad : histoire et modernités (2016, n°43)

Les Cahiers de l'ILSL peuvent être commandés à l'adresse suivante :
CLSL, Faculté des Lettres, Anthropole
CH-1015 LAUSANNE
Renseignements : http://www.unil.ch/clsl

Nouvelles technologies et standards méthodologiques en linguistique

Édité par
Marianne KILANI-SCHOCH, Christian SURCOUF et Aris XANTHOS

Cahiers de l'ILSL, n° 45, 2016

UNIL | Université de Lausanne

Illustration de couverture basée sur un design de stALLio!
https://www.flickr.com/photos/stallio/3149911976

Les Cahiers de l'ILSL
(ISSN 1019-9446)
sont une publication du Centre de Linguistique et
des Sciences du Langage de l'Université de
Lausanne (Suisse)

Centre de Linguistique et des Sciences du Langage
Quartier UNIL-Dorigny, Bâtiment Anthropole
CH-1015 Lausanne

Table des matières

Cahiers de l'ILSL-45 (2016 : 1-18)

Nouvelles technologies et standards méthodologiques en linguistique

Marianne KILANI-SCHOCH[1]
Christian SURCOUF[2]
Aris XANTHOS[3]
École de français langue étrangère (EFLE) (1, 2) & Section des Sciences du Langage et de l'Information (SLI) (1, 3), Université de Lausanne (CH)
Correspondance : marianne.kilanischoch@unil.ch

> L'informatique permet de renouer en morphologie avec la tradition philologique qui fondait la linguistique sur la lecture critique des textes. (HATHOUT, NAMER, PLÉNAT & TANGUY 2009)

1. Introduction

Dans le contexte toujours plus riche des Humanités numériques, ce numéro des Cahiers de l'ILSL présente les contributions à une journée d'étude consacrée aux nouvelles technologies et standards méthodologiques en linguistique qui s'est tenue en octobre 2014 à l'Université de Lausanne.

L'objectif de cette journée était de concentrer l'attention sur le volet linguistique des Humanités numériques pour en interroger la méthodologie. Les six contributions réunies ici explorent ainsi la question de la méthodologie dans des (sous)-disciplines linguistiques où l'empirie et les méthodes d'analyse ont été transformées par les nouvelles technologies. Cette thématique très générale abordée dans différents domaines linguistiques vise à évaluer les questions communes, et à travers elles à esquisser les directions vers lesquelles la linguistique tend à s'orienter aujourd'hui.

Un des premiers thèmes de réflexion est de toute évidence celui de l'extension des données empiriques. En rendant possible l'étude de collections toujours plus

importantes de données, dans des proportions que l'on n'avait pas soupçonnées auparavant, les nouvelles technologies étendent le champ et la portée de la recherche linguistique. Mais cet effet comporte le risque d'être surévalué et doit donc être étudié notamment pour éviter l'illusion que des données extensives se suffiraient à elles-mêmes. Car leur apport est naturellement fonction des objectifs linguistiques : sans hypothèses théoriques spécifiques, la linguistique de corpus et la linguistique quantitative demeurent purement descriptives.

La question est encore de savoir dans quelle mesure l'accès à un grand nombre de données tel qu'il est rendu possible par les nouvelles technologies influence et favorise certains choix théoriques (cf. VINCENT-DURROUX & CARR 2013). Par exemple, un objectif aussi ambitieux que celui de rendre compte de l'ensemble des comportements langagiers semble devenu concevable et caractérise même certaines versions radicales des grammaires basées sur l'usage et sur les exemplaires (voir PORT 2010, par exemple, et LADD 2011 pour une synthèse).

Cela revient à dire qu'aujourd'hui la question méthodologique des corpus et plus généralement celle de la linguistique de corpus ou basée sur des corpus (SCHEER 2013 : 19), en d'autres termes, la question de l'empirisme (VINCENT-DURROUX & CARR 2013 ; CHATER, CLARK, PERFORS & GOLDSMITH 2015) demeure une question épistémologique centrale. Avec elle, les conditions de validation des modèles linguistiques et la nature des généralisations constituent tout l'enjeu.

Dans la linguistique de corpus, on peut distinguer les problèmes méthodologiques relatifs aux corpus eux-mêmes, qu'il s'agisse des limites inhérentes à l'échantillonnage ou de celles qui tiennent à leur utilisation et aux outils de description, et les problèmes méthodologiques spécifiques aux instruments de traitement et de modélisation. Ces problèmes sont à envisager dans leurs conséquences sur la

nature des généralisations linguistiques (ERNESTUS & BAAYEN 2011 : 388). Nous verrons que les contributions des conférenciers s'inscrivent toutes dans l'une ou l'autre de ces problématiques.

1.1. Recueil des ressources linguistiques, construction des corpus

Les problèmes classiques de la constitution d'un corpus, à savoir la pertinence quantitative (quel seuil d'occurrence ? DAL & NAMER 2012 : 1265), la représentativité et l'équilibre (entre des données de différents genres, registres, par exemple), l'homogénéité et l'exhaustivité (TOMASELLO & STAHL 2004) se posent-ils différemment aujourd'hui ? La question de savoir quel volume de données est suffisant (TOMASELLO & STAHL 2004) et quelle stratégie d'échantillonnage (*sampling*) est la plus appropriée n'a évidemment pas disparu avec l'accès à des données toujours plus nombreuses. Elle a pris un nouveau relief avec l'avènement des mégadonnées (*big data*). Mais la réflexion sur les limites des corpus ainsi que sur les méthodes pour en amoindrir l'impact s'est également développée (BAAYEN 2008 ; MALVERN, RICHARDS, CHIPERE, & DURÁN 2004 ; XANTHOS & GILLIS 2010).

Du côté d'une perspective critique, SCHEER (2013) par exemple, dénonce l'illusion selon laquelle le corpus produirait la science par lui-même ou même garantirait le statut scientifique de l'analyse menée (VINCENT-DURROUX & CARR 2013). SCHEER considère que le corpus n'est pas plus pertinent aujourd'hui qu'il ne l'était hier même si « meilleures » sont les données, plus grandes sont les possibilités de réfuter les théories. Il rappelle le rôle déterminant de la conception et de la construction du corpus (*corpus design*) et des finalités théoriques qu'il sert.

TOMASELLO & STAHL (2004 : 119) avaient déjà montré, pour le domaine de l'acquisition, que les procédures

d'échantillonnage, par exemple, dépendent des questions de recherche et ne peuvent être universelles. On observe aujourd'hui des approches diverses et complémentaires dans le recours aux données linguistiques empiriques et la construction des corpus. D'une part la recherche et l'exigence de données multiformes et de mégadonnées – renforçant par exemple le potentiel statistique d'un corpus mais posant toutes sortes de problèmes techniques – connaissent un grand essor (notamment en psycholinguistique, voir KEULEERS & BALOTA 2015 ; TSUJI 2014 ; ROY, FRANK & ROY 2014 ; NEWMAN, RATNER & ROWE 2014). Il ne semble plus y avoir de limite supérieure à ce que peut être un corpus (VAUGHAN & CLANCY 2013), « toujours augmentable et jamais fini » selon la perspective de SINCLAIR (1991) (ARBACH & ALI 2013 : 23).

D'autre part, la nécessité de corpus plus petits, facilement annotables, accessibles et adaptés à des questions spécifiques, c'est-à-dire des corpus sur mesure, refait surface (KOESTER 2010 ; VAUGHAN & CLANCY 2013 ; ROMERO-TRILLO 2013). Elle soulève la question de la pertinence de la recherche empirique basée sur des corpus informatisés selon les domaines ou sous-disciplines linguistiques : les problèmes techniques rencontrés dans certains d'entre eux peuvent constituer une restriction sévère (SCHEER 2013).

Par exemple, les problématiques phonologiques ou morphologiques sont généralement plus faciles à étudier empiriquement que les problématiques pragmatiques, syntaxiques ou prosodiques. En pragmatique les recherches informatisées sont rendues difficiles, par exemple, par le fait qu'il est particulièrement laborieux d'annoter/coder les actes de langage, et parce que les phénomènes sont par définition dépendants du contexte (VAUGHAN & CLANCY 2013). Plusieurs publications récentes plaident en faveur de petits corpus pragmatiques contextualisés, relevant que la pertinence de ceux-ci est un effet secondaire des développements

technologiques qui facilitent l'accessibilité des grands corpus et rendent le choix possible (VAUGHAN & CLANCY 2013, ROMERO-TRILLO 2013). Ces petits corpus permettent l'accès à des métadonnées complètes (voir 1.2).

Et la voie moyenne consistant à adopter des corpus de référence, c'est-à-dire de larges ensembles de données sans qu'il s'agisse de mégadonnées[1], rigoureusement construits et répondant aux exigences de différents descripteurs, est une autre option, nécessaire pour un « usage responsable des données de corpus », dans la recherche sur la variation en phonologie, par exemple (ERNESTUS & BAAYEN 2011 : 384).

1.2. Utilisation des ressources : métadonnées

La méthodologie appliquée dans l'utilisation des ressources électroniques satisfait-elle aux exigences de rigueur établies par une longue tradition en sciences du langage, qu'il s'agisse de la tradition philologique ou de la tradition sociolinguistique, avec l'objet desquelles les données et corpus d'internet peuvent être comparés ? On constate le plus souvent que les métadonnées, c'est-à-dire l'ensemble standard de descripteurs de données permettant de caractériser les ressources digitales (BURNARD 2005 : 40), à savoir les conditions de constitution des corpus ou du recueil des données, l'information sur les locuteurs/scripteurs, les éléments de contextualisation, le statut discursif, diatopique, diastratique, diaphasique, diamésique et sociolinguistique en général, sont inexistantes. Or, privé de ces métadonnées un corpus n'est qu'une collection décontextualisée de mots (BURNARD 2005 : 41 ; ADOLPHS & KNIGHT 2010) sans représentativité possible (ARBACH & ALI 2013 : 17).

Il faut relever par ailleurs qu'on ne dispose pas d'un inventaire standard des outils descriptifs et analytiques (par

[1] La question étant bien entendu celle des critères de distinction entre les deux types de données.

ex. les différents types de mesures) disponibles pour interroger les ressources digitales alors que des initiatives ont été prises il y a plus de 30 ans pour standardiser la création et la gestion de données linguistiques et textuelles digitales (par exemple TEI = Text Encoding Initiative, BURNARD 2014). Le projet de *General Ontology for Linguistic Description* (*GOLD*) discuté par FARRAR 2013, à savoir une base partagée de connaissances linguistiques destinées au traitement informatique (*machine processing*) relève probablement de la même préoccupation du partage des ressources et des instruments.

Un autre problème méthodologique tient au fait que la diversification des sources, leur comparabilité relativement à la sélection et à la taille des corpus/échantillons ou à la fréquence (relative ou absolue) des unités, par exemple, le codage des corpus, leur annotation (souvent manquante, même au niveau morphosyntaxique) et le contrôle de la qualité de ces traitements, sont peu problématisés. De même, la différenciation entre divers types de données informatisées, par exemple les données de la Toile et les données de corpus, ne connait pas encore de critères établis (DAL & NAMER 2012). Pourtant, la question de la quantité et de la variété des sources informatisées devant produire le même résultat pour qu'une généralisation soit considérée comme fiable fait partie des problèmes élémentaires de validation.

L'accessibilité de sources informatisées variées soulève ensuite la question des conséquences du caractère de plus en plus interdisciplinaire de la recherche linguistique sur la nature des indices ou éléments de preuve. Dans quelle mesure et jusqu'à quel point les résultats des autres disciplines et sous-disciplines doivent-ils être pris en considération dans le système de validation ? (ERNESTUS & BAAYEN 2011 : 389 ; DRESSLER 2013).

2. Outils de traitement et de modélisation

Le problème de la méthodologie, et celui des conditions de validation à l'aide de ces données, concerne aussi bien les ressources fournies par les nouvelles technologies que les outils de traitement utilisés.

Certaines sous-disciplines de la linguistique, telle la psycholinguistique de l'acquisition dont l'approche a été entièrement renouvelée par les technologies numériques, mettent la méthodologie quantitative au centre de la recherche et de l'argumentation (voir par exemple SAFFRAN, ASLIN & NEWPORT 1996 ; SAFFRAN, NEWPORT & ASLIN 1996 ; REDINGTON, CHATER & FINCH 1998 ; ASLIN, SAFFRAN & NEWPORT 1999 ; LEWIS & ELMAN 2001 ; BOD 2009 ; LIEVEN 2014).

On peut se demander si ces sous-disciplines préfigurent le devenir de la linguistique, voire des sciences humaines dans leur ensemble et si un nouveau standard méthodologique est en train d'émerger rendant potentiellement caduques les recherches conduites sans outillage informatique.

On notera dans ce sens que le tournant quantitatif à la fin du siècle passé et le développement des modèles basés sur l'usage et les exemplaires semblent avoir réorienté les problèmes de modélisation linguistique du côté de la nature de la modélisation informatique. Car, comme le rappellent DAL & NAMER :

> depuis une dizaine d'années, [...] en quelque sorte, l'usage est à la portée du linguiste, grâce d'une part à la démocratisation spectaculaire des capacités de stockage des ordinateurs de plus en plus performantes, et d'autre part à l'évolution des techniques informatiques de recherche d'information, qui simplifient et accélèrent la fouille de ces grandes quantités de données informatisées. (DAL & NAMER 2012 : 1264)

Dans une certaine mesure, ce développement a remis sur le devant de la scène le débat ancien concernant l'opposition langue-parole/compétence-performance et interroge l'objet

ultime de la linguistique. Les outils technologiques fournissent potentiellement aux théories linguistiques les moyens de rendre compte non seulement des systèmes linguistiques dans leur complexité mais encore des actes de parole individuels (LADD 2011) dans la double perspective du locuteur et de l'interlocuteur, c'est-à-dire en intégrant à la fois la production et la compréhension (ERNESTUS 2014). Il convient de se demander dans quelle mesure les modèles linguistiques devront désormais chercher à intégrer toute la diversité des usages et donc d'élaborer des outils informatiques à même de rendre compte de la nature variée des dimensions qui y sont impliquées (ERNESTUS & BAAYEN 2011 ; ERNESTUS 2014). Ceux-ci auront à traiter à la fois les probabilités spécifiques à certains items, l'analogie dynamique (*dynamic analogy-driven computation*) et la compression des données (BAAYEN 2007 : 98 ; ERNESTUS & BAAYEN 2011).

L'intérêt croissant, cette dernière décennie, pour la complexité des systèmes linguistiques notamment en morphologie (ALBRIGHT & HAYES 2002 ; BAAYEN 2007 ; BAAYEN, MILIN, ĐURKĐEVIĆ, HENDRIX & MARELLI 2011 ; BAERMAN, BROWN & CORBETT 2015) est un autre exemple de recherches récentes soutenues par les nouvelles technologies et représentant un défi informatique aussi bien que linguistique.

En somme, la question fondamentale suscitée par les nouvelles technologies au niveau des données, de leur traitement ou de la modélisation est celle de savoir si elles représentent un défi pour les modèles linguistiques antérieurs sur le plan de l'objet comme sur celui des résultats acquis. Dans quelle mesure remettent-elles en question la pertinence de ces modèles et jettent-elles le discrédit sur les généralisations qui ont été énoncées ?

Ce sont les approches connexionnistes, il y a quelques décennies, qui ont inauguré le débat sur la validité des

catégories et unités linguistiques fondamentales telles que le phonème, le morphème, le mot-forme, etc. et sur la plus grande pertinence des données statistiques et des architectures subsymboliques (cf. BAAYEN *et al.* 2011 ; PORT 2010, etc.).

Aujourd'hui, en morphologie, pour ne prendre que cet exemple, des chercheurs comme Harald BAAYEN, Gilles BOYÉ, Olivier BONAMI, parmi d'autres, font l'impasse sur le concept de règle, lui préférant celui de généralisations probabilistes qui peuvent être obtenues en recourant à des techniques statistiques et d'apprentissage automatique (*machine learning*). Certaines versions de la théorie par exemplaires vont jusqu'à rejeter l'existence de sous-disciplines linguistiques comme la phonologie (voir LADD 2011 : 368).

Est-il ainsi plausible que, pour suivre SCHEER (2013 : 1, 4), toute proposition scientifique doive dorénavant être statistiquement pertinente ? Une manière plus spécifique et constructive de formuler la question consiste à se demander, à la suite d'ERNESTUS & BAAYEN (2011 : 387), « comment des analyses statistiques basées sur des corpus sont articulées avec la théorie de la grammaire ».

Parallèlement à ces débats, il y a à s'interroger sur les conséquences de ces évolutions pour le reste de la discipline. Les prochaines années diront la place réservée aux recherches qui ne recourent pas aux techniques informatiques. Elles montreront si celles-ci deviennent simplement obsolètes.

Les auteurs qui ont contribué à ce numéro ne prétendent ni ne peuvent évidemment répondre à l'ensemble de ces questions. Mais chacun apporte une réflexion particulière sur les possibilités et/ou les limites de l'apport des nouvelles technologies. Leurs travaux portent sur les divers domaines de l'acquisition du langage, du traitement du langage, de la phonologie, de la morphologie flexionnelle, dérivationnelle et computationnelle. Ils partagent une vaste expérience en

matière de bases de données et de corpus de grandes dimensions, une connaissance approfondie des nouvelles technologies et l'exigence d'une méthodologie solide.

Dans ce numéro nous proposons la traduction française des conférences de cette journée tenues en anglais ainsi que celle d'une contribution allemande de DRESSLER, KORECKY-KRÖLL & MÖRTH dont la participation à la journée n'a pas été possible. À l'exception de l'article de Nabil HATHOUT, qui a rédigé en français la version écrite de sa conférence, et de celui de DRESSLER, KORECKY-KRÖLL & MÖRTH, traduit par Marianne Kilani-Schoch, les textes de ce numéro ne sont donc pas des articles préparés par les auteurs eux-mêmes mais des traductions de leur conférence du 17 octobre 2014, élaborées par nos soins à partir de la transcription des enregistrements de la Journée effectuée par Guillaume Feigenwinter et supervisée par Marianne Kilani-Schoch, et d'une première traduction de Guillaume Feigenwinter que Marianne Kilani-Schoch, Christian Surcouf et Aris Xanthos ont adaptée et reformulée. Nous avons pris le parti de conserver certaines caractéristiques de la présentation orale, qu'il s'agisse du style, de l'adresse à l'audience, des renvois aux autres conférences ou du jeu de questions-réponses en fin de texte.

3. Présentation des contributions

La contribution de BOYÉ aborde une série de problèmes méthodologiques liés à l'étude de la morphologie flexionnelle et en particulier la question du remplissage des paradigmes (ACKERMAN *et al.* 2009) dans la conjugaison des verbes en français. Dans une première partie, différents aspects problématiques sont définis : (i) au niveau des formes, les représentations phonologiques, la variation des formes et l'influence des fréquences sur le lexique, (ii) au niveau des cases, la surabondance et la défectivité. Un tour d'horizon des données disponibles tend à montrer que la prise en

compte de ces obstacles n'est pas chose aisée pour l'instant. La deuxième partie présente l'ébauche d'un modèle à même de contourner ces difficultés sur la base d'un lexique d'entrainement prenant en compte des fréquences, des représentations variées, la surabondance (sans inclure la défectivité). L'analyse est basée sur la collecte des analogies entre formes dans l'échantillon lexical et des classes de compétition entre analogies pour produire en masse des formes-candidates. Le paradigme flexionnel de chaque lexème est ensuite extrait en choisissant une clique à couverture maximale parmi l'ensemble des formes produites.

La contribution de DRESSLER, KORECKY-KRÖLL & MÖRTH s'attache à la question de la prévisibilité et de la prédictibilité probabiliste en linguistique, ainsi qu'au rôle des corpus électroniques dans l'élaboration des prévisions et de la vérification empirique, en se concentrant plus particulièrement sur deux domaines : l'acquisition de la langue première et le développement diachronique.

Les auteurs détaillent la variété des facteurs à prendre en considération dans le premier domaine pour rendre compte de sa complexité et effectuer des prédictions relatives à l'acquisition des pluriels allemands, turcs et anglais, notamment, qu'il s'agisse des facteurs typologiques de richesse, transparence et univocité morphologique de la flexion de la langue-cible, de la relation entre intégration et production chez l'enfant ou du niveau socioéconomique des familles. Les prédictions probabilistes sont limitées à la richesse de l'input et de l'output et à la relation entre les deux dimensions pour lesquelles on dispose de mesures quantifiables.

Les rétrodictions de la diachronie sont également discutées à l'aune d'un exemple de développement morphologique dans les dialectes italo-romans qui a pu faire l'objet de rétrodictions relativement précises. Les auteurs

montrent pourquoi une telle prévisibilité demeure exceptionnelle en diachronie.

Dans sa contribution, ERNESTUS évoque l'intérêt des grands corpus oraux dans la recherche en linguistique et en psycholinguistique, par le fait même qu'ils permettent de dépasser les limites de l'expérimentation en laboratoire ou du recours à l'intuition par le linguiste. Ainsi dans un premier temps, sur la base d'une analyse quantitative du Corpus Oral du Néerlandais, l'auteure démontre que l'assimilation régressive dans cette langue ne fonctionne pas toujours conformément à la description proposée par les linguistes. Si le recours à des corpus oraux de grande taille permet de déterminer la structure phonétique des mots dans le flux effectif de la parole, ERNESTUS rappelle néanmoins que leur utilisation doit s'accompagner de certaines précautions. Elle souligne plus particulièrement les difficultés de la transcription phonétique, qu'elle soit réalisée manuellement par des transcripteurs humains ou automatiquement à l'aide d'un dispositif informatique de reconnaissance vocale, qu'il faudra de toute façon alimenter correctement pour l'obtention de résultats exploitables. L'auteure aborde par la suite les écueils du traitement statistique des données, notamment celui de la manière d'intégrer les prédicteurs en fonction de leur nombre et de leur degré de corrélation. À la suite de ces divers rappels sur les précautions nécessaires à la manipulation des corpus oraux, ERNESTUS conclut que la recherche en linguistique et en psycholinguistique ne peut se dispenser de leur apport dans la mesure où les corpus oraux, contrairement aux expérimentations en laboratoire, donnent accès à ce que font *réellement* les locuteurs dans leur pratique quotidienne de l'oral.

La contribution de GILLIS dresse un panorama critique de l'usage des technologies pour l'étude de l'acquisition – en particulier dans son volet observationnel, centré sur les

corpus oraux. Partant du constat que la manière dont de tels corpus sont traditionnellement constitués est extrêmement couteuse en temps de travail, alors même qu'une proportion très restreinte des productions enfantines est échantillonnée, l'auteur examine successivement plusieurs façons de remédier au problème de la rareté des données : la base de données CHILDES, le système LENA™ et l'approche « big data » mise en place par Deb ROY (2011). Ce passage en revue suggère que le problème ne se limite pas à la difficulté d'enregistrer une portion représentative de l'input et de l'output enfantin ; il faut encore et surtout pouvoir transcrire et annoter les données récoltées, ce que GILLIS conçoit comme un défi majeur dans la perspective d'une avancée substantielle de ce domaine de recherche.

L'article de HATHOUT présente en détail les avantages et les couts de l'évolution consécutive au développement des nouvelles technologies pour la recherche en morphologie, et soulève la question de la nature et de la place des données dans cette recherche.

Ce qu'on appelle désormais la morphologie extensive, fondée sur de vastes quantités de données, a connu déjà différentes périodes dans une histoire que HATHOUT retrace en quelques pages et qui va de l'accès très large à la Toile dans les années 90 aux restrictions que les moteurs de recherche ont imposées aujourd'hui par la protection des index.

Si le chercheur montre, à l'exemple des travaux qu'il a menés en équipe sur les adjectifs en *-esque* et en *-able*, que la quantité de données prises en compte détermine directement la qualité des résultats, il expose aussi la longue liste des problèmes méthodologiques posés par les données de la Toile. Il évoque en outre les transformations que l'approche extensive engendre dans la recherche en morphologie devenue expérimentale, pour appeler à une

revalorisation du travail de constitution de ressources et de collections de données morphologiques en même temps qu'à une politique de partage dont il esquisse les contours.

TRIBUSHININA & MAK abordent la question des limites de l'étude de corpus, donc, de la production verbale dans l'analyse du développement linguistique chez des enfants bilingues et des enfants atteints d'un trouble du langage.

Ils montrent à travers leur recherche sur les connecteurs et les pronoms sujets en russe et en néerlandais que la différenciation linguistique des deux populations, souvent confondues au niveau de la production, nécessite le recours au paradigme méthodologique du monde visuel et à la technique de l'oculométrie (*eye-tracking*) : dans les deux domaines de la langue étudiés, l'écart significatif de compétence linguistique entre enfants bilingues et enfants atteints d'un trouble spécifique du langage (TSL/SLI) n'apparait qu'avec les mesures précises de l'activité de traitement rendues possibles par l'oculométrie.

Nos remerciements vont à François Rosset (doyen de la Faculté des lettres en 2014) et Thérèse Jeanneret (directrice de l'EFLE) pour leur soutien financier dans l'organisation de la Journée ainsi qu'au Centre de linguistique et des sciences du langage (CLSL) et à son directeur, Marcel Burger, pour les fonds accordés en vue de cette publication.

Références

ACKERMAN Farrell, BLEVINS James P. & MALOUF Robert (2009). Parts and Wholes: Implicative Patterns in Inflectional Paradigms. In BLEVINS James P. & BLEVINS Juliette (Eds), *Analogy in Grammar: Form and Acquisition*. Oxford: Oxford University Press, 54-82.

ADOLPHS Svenja & KNIGHT Dawn (2010). Building a Spoken Corpus: what are the Basics? In O'KEEFFE Anne & MCCARTHY Michael (Eds), *The Routledge Handbook of Corpus Linguistics.* Abingdom: Routledge, 38-52.

ALBRIGHT Adam & HAYES Bruce (2002), Modeling English Past Tense Intuitions with Minimal Generalization. In MAXWELL Michael (Ed.), *Proceedings of the sixth Meeting of the ACL Special Interest Group in Computational Phonology*. Philadelphia: ACL, 58-69.

ARBACH Najib & SAANDIA Ali. (2013). Aspects Théoriques et Méthodologiques de la Représentativité des Corpus. *Corela-HS-13.* Publié en ligne le 10.12.2013.

ASLIN Richard N., SAFFRAN Jenny R. & NEWPORT Elissa L. (1999). Statistical Learning in Linguistic and Nonlinguistic Domains. In MACWHINNEY Brian (Ed.), *The Emergence of Language.* Mahwah, NJ: Lawrence Erlbaum, 359-380.

BAAYEN R. Harald. (2007). Storage and Computation in the Mental Lexicon. In JAREMA G. & LIBBEN G. (Eds), *The Mental Lexicon*. Amsterdam: Elsevier, 81-104.

BAAYEN R. Harald. (2008). *Analyzing Linguistic Data*. Cambridge: Cambridge University Press.

BAAYEN R. Harald, MILIN Petar, ĐURĐEVIĆ Dusica Filipović, HENDRIX Peter & MARELLI Marco (2011), An Amorphous Model for Morphological Processing in Visual Comprehension based on Naive Discriminative Learning, *Review 118-3*, 438–482.

BAERMAN Matthew, BROWN Dunstan & CORBETT Greville G. (Eds) (2015). *Understanding and Measuring Complexity*. Oxford: Oxford University Press.

BOD Rens (2009). From Exemplar to Grammar: a Probabilistic Analogy-Based Model of Language Learning. *Cognitive Science* 33, 752-793.

BURNARD Lou (2005), Metadata for Corpus Work. In WYNNE Martin (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice*. Chapter 3. Produced by ahds literature, language and linguistics. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm

BURNARD Lou (2014). *What is the Text Encoding Initiative?* OpenEdition Press.

CHATER Nick, CLARK Alexander, PERFORS Amy & GOLDSMITH John A. (2015). *Empiricism and Language Learnability*. Oxford: Oxford University Press.

DAL Georgette & NAMER Fiametta (2012). Faut-il Bruler les Dictionnaires ? Ou comment les Ressources Numériques ont Révolutionné les Recherches en Morphologie. *Congrès Mondial de Linguistique Française*, 1261-1276. EDP Sciences. DOI10.1051/shsconf/20120100217.

DRESSLER Wolfgang U. (2013). Quo vadis linguistica ? Conférence plénière. Salzburg. 40. *Oesterreichische Linguistiktagung*. Ms.

ERNESTUS Mirjam (2014). Acoustic Reduction and the Roles of Abstractions and Exemplars in Speech Processing. *Lingua 142*, 27-41.

ERNESTUS Mirjam & BAAYEN, R. Harald (2011). Corpora and Exemplars in Phonology. In GOLDSMITH John, RIGGLE Jason & YU Alan C.L., *The Handbook of Phonological Theory.* Malden, Oxford: Wiley-Blackwell, 374-400.

FARRAR Scott (2013). An Ontological Approach to Canonical Typology: Laying the foundations for e-Linguistics. In BROWN Dunstan, CHUMAKINA Marina & CORBETT Greville G., *Canonical Morphology and Syntax.* Oxford: Oxford University Press, 239-261.

HATHOUT Nabil, NAMER Fiametta, PLÉNAT Marc & TANGUY Ludovic (2009). La Collecte et l'Utilisation des Données en Morphologie. In FRADIN Bernard, KERLEROUX Françoise & PLÉNAT Marc (ss la dir.), *Aperçus de Morphologie du Français*. Paris : Presses Universitaires de Vincennes, 267-289.

KEULEERS Emmanuel & BALOTA David A. (2015). Megastudies, Crowdsourcing, and Large Datasets in Psycholinguistics: An Overview of Recent Developments. *Quarterly Journal of Experimental Psychology 68-8*, 1457-1468.

KOESTER Almut (2010). Building Small Specialized Corpora. In O'KEEFFE Anne & MCCARTHY Michael (Eds), *The Routledge Handbook of Corpus Linguistics.* Abingdom: Routledge, 66-79.

LADD, D. Robert (2011) Phonetics in Phonology. In GOLDSMITH John, RIGGLE Jason & YU Alan C.L., *The Handbook of Phonological Theory.* Wiley-Blackwell, 348-373.

LEWIS John D. & ELMAN Jeffrey L. (2001). Learnability and the Statistical Structure of Language: Poverty of Stimulus Arguments Revisited. *Proceedings of the Twenty-Sixth Annual Boston University Conference on Language Development.* Somerville, MA: Cascadilla Press, 359-370.

LIEVEN Elena (2014) First Language Development: a Usage-Based Perspective on Past and Current Research. *Journal of Child Lang*uage *41-Supplement S1*, 48-63.

MALVERN David D., RICHARDS Brian J., CHIPERE Ngoni & DURÁN Pilar. (2004). *Lexical Diversity and Language Development.* Basingstoke: Palgrave Macmillan.

NEWMAN Rochelle, RATNER Nan & ROWE Meredith (2014). Big Data: Challenges of Conducting Longitudinal Studies. Amsterdam. *IASCL Symposium*. Abstracts, 324.

PORT Robert F. (2010). Rich Memory and Distributed Phonology. *Language Sciences* 32, 43-55.

REDINGTON Martin, CHATER Nick & FINCH Steven. (1998). Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. *Cognitive Science 22-4*, 425-469.

ROMERO-TRILLO Jesús. (2013). *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*.

ROY Brandon C., FRANK Michael C. & ROY Deb (2014). Harnessing Big Data in a Naturalistic Study of one Child's Early Word Learning. Amsterdam. *IASCL Symposium*. Abstracts, 323.

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

SAFFRAN Jenny R., ASLIN Richard N. & NEWPORT Elissa L. (1996), Statistical Learning by 8-Month-Old Infants, *Science 274-5294*, 1926-1928.

SAFFRAN Jenny R., NEWPORT Elissa L. & ASLIN Richard N. (1996). Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language 35-4*, 606-621.

SCHEER Tobias (2013). The Corpus: a Tool among Others. *Corela-HS-13.* Publié en ligne le 25.11.2013.

SINCLAIR, John (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

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

TSUJI Sho. (2014). Big Data in Infant Language Acquisition. Chances and Challenges. Amsterdam. *IASCL Symposium*, Absracts, 322.

VAUGHAN Elaine & CLANCY Brian (2013). Small Corpora and Pragmatics. *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies.* Dordrecht: Springer, 53-76.

VINCENT-DURROUX Laurence & CARR Philippe (2013). Statut et Utilisation des Corpus en Linguistique. *Corela-HS-13.* Publié en ligne le 11.12.2013.

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

Cahiers de l'ILSL-45 (2016 : 19-44)

Une analyse de la conjugaison française basée sur les données[1]

Gilles BOYÉ
Université Bordeaux-Montaigne & CNRS (F)
gilles.boye@u-bordeaux-montaigne.fr

1. Introduction

À quoi devrait ressembler une analyse de la conjugaison française basée sur les données ? D'abord, une telle analyse devrait probablement prendre en compte de larges quantités de données. Or, je ne dispose que de lexiques, c'est-à-dire d'une petite quantité de données, compilées à partir de données plus importantes collectées par d'autres personnes. Parlons donc en premier lieu de ce problème.

Il me manque un corpus, un vrai corpus duquel je pourrais tirer les formes conjuguées et les rentrer dans mon moteur pour effectuer un traitement informatique. Il me faudrait plus de matière, mais je n'en ai pas. L'autre élément nécessaire est de réduire au maximum le nombre des hypothèses. Je n'en utilise que deux. J'ai besoin de descriptions phonémiques des segments, parce que je travaille au niveau phonologique. Je n'utilise que des analogies simples entre les formes. Par exemple, si on a une représentation phonologique donnée dans une case du paradigme, et une autre représentation phonologique dans une autre case, comment passe-t-on de l'une à l'autre par analogie ? Ensuite, à partir des larges quantités de données – que je n'ai pas… –, je voudrais extraire des généralisations et générer des formes candidates, en partant d'une case par exemple, et tenter de

[1] *A data-driven analysis of French conjugation: driving around the curbs*. Transcription, traduction et adaptation par Guillaume Feigenwinter et Christian Surcouf.

remplir la totalité du paradigme d'un verbe, voire de tous les verbes. Ensuite, parce qu'il y a beaucoup de bruit, en raison du grand nombre d'analogies, j'essaie de faire un tri et peu à peu de trouver quels paradigmes peuvent être retenus, et lesquels devraient être éliminés. J'ai donc des problèmes avec les données et des problèmes avec les analyses.

2. Le problème avec les données

Commençons par les données. Dans une langue, il y a beaucoup de variation, ce qui s'avère difficile à gérer, et pas seulement au niveau de la prononciation, comme, par exemple, les réductions de mot (voir ERNESTUS, page 65). Les locuteurs ont par ailleurs tendance à avoir plusieurs formes conjuguées différentes pour une même case, et je ne sais pas quoi faire avec ce genre de problème. Quel que soit le niveau d'expertise des transcripteurs, des désaccords subsistent sur la phonologie. Alors que faire ? Encoder moins de choses, ne garder que les points sur lesquels tout le monde s'accorde, ou bien encoder plus et trouver un moyen de gérer les dissensions ? Il nous faut prendre une décision.

Quant au lexique, que dois-je mettre dedans ? Quels mots existent ? Quels mots n'existent pas ? Pour chaque mot, chacune de ses formes potentielles existe-t-elle, c'est-à-dire *concrètement* dans la langue orale ? En français, combien de fois avez-vous entendu le subjonctif imparfait à la deuxième personne du singulier d'un verbe ? De toute votre vie ? Probablement jamais ! Nous faisons l'hypothèse que la case existe, mais en réalité personne n'a jamais entendu ce qu'elle contient. Alors qu'est-ce qu'on fait ? On l'enlève ou on la laisse ? Est-ce que c'est une vraie case ? Voilà le problème avec les données en général, ensuite viennent ceux de l'analyse.

3. Les difficultés de l'analyse

Dès qu'on envisage la morphologie, se pose la question des phénomènes à analyser. C'est compliqué parce qu'en morphologie, bien des choses ne se déroulent pas comme prévu, comme par exemple la prononciation. Pour cette raison, des cadres théoriques très variés ont été proposés pour décrire des aspects très différents des données. Que devrait-on garder ou éliminer ? Toutes ces questions sont en fait élémentaires. Prenons l'exemple de la variation.

3.1. La question de la variation

Dois-je intégrer une forme comme « ils croivent », censée être socialement marquée ? Je ne dirais pas qu'elle est inacceptable parce qu'elle est marquée, d'ailleurs beaucoup de gens l'emploient. Dois-je faire comme certains et prétendre qu'une telle forme n'existe pas ou dois-je au contraire la prendre en compte puisque, effectivement, des locuteurs l'utilisent ? Et que faire de *frire* ? *Frire* a cette capacité intrigante d'être un verbe différent selon la zone de la francophonie dans laquelle on se trouve. Pour certains, c'est un verbe défectif, qui n'a que quelques cases remplies. D'autres pensent que c'est un verbe tout à fait normal, et que toutes ses cases sont remplies. Beaucoup de locuteurs pensent un peu des deux, et décrivent une défectivité plus ou moins importante. Devrais-je alors considérer *frire* comme complètement défectif, ou l'inverse puisque que certaines personnes en connaissent toutes les formes ? Et qu'en est-il de *chourave* ? Je ne sais pas ce qu'il en est en Suisse, mais en France, *chourave* est un mot étrange. Il ne varie jamais, il n'a qu'une seule forme pour moi, mais pour beaucoup de gens, c'est devenu *chouraver*, c'est-à-dire un verbe tout à fait régulier du premier groupe. Mais le verbe français que, moi, je connais se conjugue : « je chourave, tu chouraves, il chourave, nous… volons, vous… volez, ils chouravent » et « je l'ai chourave ». Vous voyez que le participe est aussi

« chourave », tout comme l'infinitif. C'est à peu près tout ce que nous avons : le participe passé, l'infinitif et quatre formes du présent de l'indicatif et du subjonctif, c'est tout. On a environ une demi-douzaine de verbes de la sorte, qui se terminent tous en *-ave* parce qu'ils proviennent de la même langue, et qu'ils ont gardé les mêmes propriétés. Cependant, avec le temps, ce verbe a évolué en un verbe normal. Devrais-je prendre le premier paradigme ou le nouveau paradigme ? Et de quelle population ? En fait, je ne sais pas.

Ces exemples illustrent des points de données spécifiques à un lexème, mais plus généralement se pose la question de savoir comment traiter l'intégralité du contenu d'un corpus. En effet, si j'ai un corpus global, alors il ne sera pas homogène. Est-ce représentatif de prendre des points de données d'individus, sachant que ces individus ont été choisis de manière à être représentatifs de leur population ? Ou devrais-je au contraire mélanger les données de plusieurs corpus et essayer de représenter l'entièreté des données, d'une manière ou d'une autre ? Là encore, je ne sais pas. Ma conférence n'est pas très informative…

Et jusqu'à quel point devrais-je manipuler les données ? Est-ce légitime ? Tout le monde le fait. On a parfois des données aberrantes, mais signifient-elles quelque chose ? Peut-on démontrer qu'elles indiquent quelque chose du signal, mais trop insignifiant pour qu'on en tienne compte ? Quelles sont exactement les données aberrantes ? La forme *croivent* est-elle une erreur ou un dialecte ? Comment décider ? Ce sont là des questions linguistiques générales. Abordons maintenant la question de la fréquence.

3.2. La question de la fréquence

À supposer que j'aie une grande quantité de données, j'ai l'avantage de pouvoir en tirer des fréquences. Je peux tenter d'avoir une couverture étendue, au lieu d'une couverture standard. Je peux essayer d'obtenir des fréquences orales, au

lieu de fréquences écrites, les seules dont je dispose pour le moment. Je n'ai aucune idée des fréquences orales : aucune donnée, rien, pourtant il est probable que la fréquence ait une influence sur le problème que je tente de résoudre. Maintenant, imaginons que j'aie toutes ces fréquences : comment savoir, quand on voit des formes homophones, laquelle est vraiment comptabilisée quand je compte les fréquences ? La forme *mange* serait-elle un subjonctif ? Une première, une troisième personne ? Est-ce important ? Si oui, quand ? Est-ce important au sein du même lexème, ou seulement entre différents lexèmes ? Est-ce important entre différentes catégories ? Je ne sais pas. Et comment traiter la défectivité ? Comment prendre en compte la fréquence d'une entité qui n'est jamais là ? Je m'intéresse plus particulièrement à la défectivité, parce que j'ai pendant longtemps travaillé sur ce problème. Au début, on ne me laissait même pas en avoir un modèle, on me disait : « Ça n'existe pas. », parce que ce n'est pas un point de donnée positif. À priori, on pourrait effectivement parvenir à une telle conclusion, mais l'observation des données invite au contraire à reconnaitre la défectivité. Par exemple, pour le verbe *clore*, bien qu'ils sachent que la forme est censée être *closons*, les locuteurs s'efforcent de l'éviter comme si leur vie en dépendait. Qu'en est-il maintenant des problèmes phonologiques.

3.3. La question de la phonologie

Imaginons que nous nous mettions d'accord sur un système phonémique. Il subsistera malgré tout le problème des voyelles moyennes en français. Alors, que fait-on des voyelles moyennes ? Dans bien des contextes en français, elles sont neutralisées, mais pas chez tous les locuteurs dans toutes les régions des pays francophones. À certains endroits, on distingue presque toutes les voyelles moyennes, alors qu'ailleurs on les confond presque toutes en une seule. Le

même type de problèmes se pose avec le schwa. Mais il y a plus encore, en français, en général, il n'y a pas de contraste entre un /ʁ/ et deux /ʁʁ/. Donc ce n'est pas comme si c'était un problème phonémique, ça ne vient pas des phonèmes eux-mêmes, mais de leur contraste. Il n'existe que dans trois lexèmes : *courir, acquérir, mourir*. Et on ne peut pas dire [ilkuʁa] (« Il *coura »), ce n'est pas du français. [ilkuʁa] n'est pas reconnaissable en tant que forme conjuguée de *courir*. On a la même chose avec *vingt-deux*, qui ne peut pas être prononcé [vɛ̃dø], qui correspondrait en l'occurrence à la séquence de *vingt* et *deux*, mais pas à *vingt-deux*. Quand je dis qu'on ne peut pas le dire, je parle depuis mon fauteuil, je ne suis pas allé vérifier dans un corpus. Je ne fais qu'écouter les gens. Pour *courir*, c'est évident. D'autres, comme [tʁije], soit « faire un tri », soit « émettre des trilles », ont exactement la même forme mais pas le même paradigme, parce qu'il y a neutralisation en français dans ce contexte. Il est impossible de distinguer la présence ou non d'une voyelle avant cette semi-consonne.

Il y a encore d'autres choses, comme quand vous faites un futur en français : le futur de *batter*[2] se prononce en général [batʁa], mais peut aussi être [batəʁa], alors que le futur de *battre* ne peut qu'être prononcé [batʁa] sans le schwa. Comment peut-on en rendre compte ? Devrait-on inclure le schwa dans tous les futurs de *batter*, en perdant le fait important que dans 99% des cas il n'y est pas ? Est-ce une séparation numérique entre sa présence et son absence ou a-t-on affaire à un phénomène graduel ? Et comment le représenter ? Au final, la seule chose dont je suis certain, c'est que j'aurai besoin d'un ensemble de traits. Il s'agit là de mon hypothèse de base, celle qui relève de ma responsabilité. Pour toutes les autres questions, je n'ai aucune idée.

[2] NdE : Frapper une balle à l'aide d'une batte.

3.4. Les problèmes avec le lexique

Venons-en maintenant à la question du lexique, qui soulève au moins un problème. Tous les lexiques ont été constitués automatiquement, et s'avèrent pour cette raison très cohérents. Comme on peut s'y attendre de tout procédé automatique, la moindre erreur est répercutée de manière systématique et constante à tout le corpus. Il est alors impossible de se fier à ce genre de lexique. Quant aux fréquences, elles proviennent de textes. Il n'y a pas de traitement spécifique de l'homophonie, ou de l'homographie, juste du texte, ce qui d'une manière générale n'a rien à voir avec de l'oral véritable.

J'ai deux sortes de lexiques. En premier lieu : les dictionnaires de paradigmes, très bien organisés, et s'intégrant parfaitement dans mon système parce qu'une forme donnée apparait dans une case spécifique, et qu'il n'y a qu'une forme par case. Mais les variantes et la surabondance ne sont en l'occurrence jamais prises en compte, ce qui, pour moi, constitue un problème, puisque je veux justement intégrer ces deux phénomènes. En second lieu, j'ai des dictionnaires de formes, qui ignorent l'existence des paradigmes. Ainsi perd-on le fait qu'*assoir* – verbe surabondant en français – a deux séries : *assois, assois, assoit, assoyons, assoyez, assoient,* et *asseye, asseyes, asseye, asseyons, asseyez, asseyent*. Avec un simple lexique de formes, il est impossible de savoir que ces deux séries vont ensemble, ce qui constitue une perte d'information.

Prenons le cas du verbe *courir*. Quelle est la forme qui apparait dans la case 'deuxième personne du pluriel du conditionnel présent' ? C'est un verbe très courant, que tout le monde connait et que tout le monde est convaincu de pouvoir conjuguer facilement. On hésiterait à priori entre

/kuʁje/ et /kuʁʁje/[3]. En français, on ne peut pas avoir deux /ʁ/ suivis d'un /j/, donc il faut changer quelque chose, par exemple enlever un /ʁ/. On peut également insérer un schwa, brisant ainsi le groupe consonantique. Certains locuteurs proposent même /kuʁiʁje/. Quoi qu'il en soit, tout le monde ressent la même chose et je suppose que tout le monde a un trou à cet endroit du paradigme. Ainsi évite-t-on d'utiliser cette forme. Ce n'est pas que nous sachions que les locuteurs évitent effectivement une telle forme, mais nous savons qu'ils devraient l'éviter parce qu'elle leur semble étrange. Les formes de ce genre ne sont pas à proprement parler problématiques, mais on a l'impression qu'elles ne devraient pas être dans cette case-là, qu'elles devraient en être expulsées, parce qu'elles ne se comportent pas comme les autres.

Si l'on évoque la défectivité, qui est mon domaine de recherche, on voit des choses comme *distraire*, dont les locuteurs ne savent pas former le passé simple. N'ayant jamais appris à le faire, ils n'en ont pas la moindre idée. Pour le pluriel de l'adjectif *nasal*, là on connait la réponse : c'est soit *nasals* /nazal/, soit *nasaux* /nazo/. Mais si vous enseignez la morphologie ou la phonologie comme moi, vous évitez ce mot au masculin pluriel, parce que soit vous dites /nazal/ et tout le monde rigole, soit vous dites /nazo/ et… tout le monde rigole ! C'est une autre sorte de défectivité, une défectivité apprise. Quand vous êtes au milieu d'une foule – comme je suis Français, on va dire que c'est une manifestation ou une grève, évènements courants en France – vous suivez les autres. Et si pour une raison ou une autre, les manifestants devant vous changent de route et se séparent, que faites-vous ? Continuez-vous tout droit ou suivez-vous tout simplement les gens ? Nous suivons normalement ceux qui

[3] Durant la conférence, Gilles BOYÉ a sollicité le public, et obtenu ces deux versions.

nous précèdent et c'est ce que nous avons tendance à faire avec *clore*. Nous remarquons que les gens ont tendance à l'éviter, nous ignorons pourquoi, mais c'est comme dans une grève ou une manifestation : si au lieu d'aller tout droit comme on s'y attend, la foule évite quelque chose, alors on choisit de faire pareil.

Comme je travaille sur la prédictibilité, je rencontre un autre problème. Par exemple, il est possible de prédire deux choses à partir de /ilɛ/. En fait, dès qu'on le dit à voix haute, le problème disparait : /il'ɛ/ (*il hait*) n'est effectivement pas la même chose que /ilɛ/ (*il est*). En français *oral*, il n'y pas d'ambigüité dans ce cas, et c'est seulement parce que les données dont je dispose n'incluent pas le fait que *hait* de *haïr* ne commence pas par une voyelle, au sens français du terme.

Voyons un autre problème courant : /ʒəsɥi/ *je suis* et /ʒəsɥi/ *je suis*. On sait que dans la réalité, les locuteurs ont différentes représentations pour ces deux-là, parce que l'énoncé « je suis une fille » n'est pas problématique. Si je dis [ʒəsɥiynfij], c'est évidement dans le sens de « je poursuis une fille », en revanche, si je dis [ʃɥiynfij], personne ne me croira, parce que [ʃɥi] est seulement une contraction possible pour le verbe *être*. Il y a donc, dans notre représentation, une différence concrète entre ces deux verbes dans cette case, invisible cependant dans les points de données à ma disposition.

Ensuite, il y a des choses identiques qui devraient effectivement l'être, parce que *assoir* reste le même mot, peu importe que nous ayons /aswa/ ou /asje/ ou /asɛj/, c'est d'une certaine manière toujours la même identité. Quant à *ficher*, c'est un verbe qui signifie « noter sur une fiche », et il n'a pas le même sens que *fiche*, employé dans la phrase *j'ai fichu mon vélo par terre*. Ils ont une grande partie de leur paradigme en commun, mais diffèrent en certains points. La plupart des Français les confondent et ne connaissent que la différence entre leurs participes passés.

3.5. La question des paradigmes

Venons-en aux paradigmes. Je pourrais faire des paradigmes syntactiques, comme tout le monde : une case par contexte syntactique avec des syncrétismes entre cases homophoniques. Ou alors je pourrais faire des paradigmes de formes. Et ensuite, il suffirait d'avoir un identifiant pour chaque forme, et des correspondances entre ces identifiants et les contextes syntactiques. Ainsi pour chaque forme, on aurait la liste des contextes syntaxiques où elle apparait. Ceci résoudrait le problème des homophonies et expliciterait la présence des syncrétismes, mais je perdrais alors l'uniformité de mes paradigmes, car les verbes auraient des paradigmes différents en fonction du nombre de formes qu'ils possèdent et des contextes syntactiques qu'elles occupent. En fait, mon problème est généralement de remplir le paradigme. Cela porte un nom : « le problème du remplissage des cases du paradigme »[4] (ACKERMAN, BLEVINS & MALOUF 2009).

Ces auteurs proposent une solution, en calculant l'entropie, définie comme la mesure du manque d'information à combler pour résoudre un problème. Selon eux, la morphologie flexionnelle est un problème à basse entropie, parce qu'en général il manque peu d'information pour remplir les cases des paradigmes. La flexion est effectivement un système très bien organisé à très faible entropie. Toutefois, leur solution présente un problème, que je ne discuterai pas ici. En effet, les auteurs ne proposent aucun modèle de flexion. Ils parlent juste d'entropie, mais jamais de flexion *réelle* : quelles sont les formes ? comment elles sont calculées ? Pour ACKERMAN, BLEVINS & MALOUF, il s'agit de remplir une case du paradigme à partir de l'information dont on dispose, mais ils ne précisent pas quelle

[4] NdE: « Paradigm Cell Filling Problem: What licences reliable inferences about the inflected (and derived) surface forms of lexical items » (voir ACKERMAN, BLEVINS & MALOUF 2009, 54).

est cette information. Il n'y a aucune mesure de la connaissance initiale. Pourtant dès qu'on travaille sur l'entropie, définie comme le manque de connaissance à combler pour résoudre un problème, une telle mesure devrait être obligatoire. Les auteurs prétendent inférer le contenu d'*une* case à partir d'une forme, mais en définitive, ils essaient d'inférer le contenu de *toutes* les cases à partir d'une seule forme. Même le fait qu'ils partent vraiment d'une forme n'est pas certain, parce qu'on n'a aucun contrôle sur la connaissance dont ils disposent au départ. En fait, il me semble qu'ils savent déjà presque tout du paradigme, et ils infèrent quel est le paradigme. Mon collègue, Olivier BONAMI a quant à lui mené de véritables recherches sur la manière dont on peut inférer le paradigme à partir d'une, deux, ou trois formes, en mesurant ce qu'il connait des formes initiales (BONAMI & BENIAMINE 2015). Bref, je vois là beaucoup de problèmes avec ces deux types de réponses à la question du remplissage des paradigmes et j'aimerais tenter d'en résoudre une partie.

3.5.1. La résolution des problèmes

Alors qu'en est-il de la variation ? Bien que je n'aie aucune idée de comment collecter les données, je sais en revanche comment les rentrer dans mon système. J'essaie donc de résoudre ce premier problème en intégrant non seulement des zéros et des uns, mais aussi des proportions, et des estimations. Pour la phonologie, j'essaie différents modèles, différents encodages, et j'analyse les résultats. À partir de là, je formule différentes hypothèses et tente de voir quelles sont les conséquences, et laquelle semble être la plus appropriée et adaptée aux données. Pour le lexique, je commence avec les lexiques que j'ai à ma disposition, bien qu'ils soient imparfaits, mais c'est tout ce que j'ai pour le moment. En ce qui concerne la morphologie, j'essaie de couvrir la supplétion, la défectivité et la surabondance. Quant

au cadre théorique, étant donné que je ne trouvais rien d'adapté, j'en ai proposé un nouveau, que j'ai nommé « distributions de paradigme ». Pour la morphologie des distributions de paradigme, je propose d'avoir des distributions pour les formes : plusieurs formes dans une case, mais pas n'importe quelle forme (par exemple, on aurait 10% d'une forme donnée, 20% d'une autre : c'est une distribution). Je procède de même pour les paradigmes tout en leur attribuant un indice de confiance. J'utilise également les réseaux « petit monde » (*small-world networks*), comme sur Facebook, où les amis de vos amis sont généralement vos amis. En d'autres termes, si vous êtes dans un paradigme, et que quelqu'un d'autre y est aussi, ses connaissances sont probablement dans le même paradigme que vous.

Je n'ai pas encore évoqué les problèmes de l'appariement un à un, généralement sous-jacent dans les modèles de la flexion. Étant donné l'absence d'analyse syntagmatique, je n'ai affaire qu'à des formes pas à des radicaux et des affixes donc le principe « un lexème, un radical » n'a pas de place ici. En ce qui concerne le principe « une case, une forme », je dois prendre en compte la surabondance et les variantes, fournies par mon échantillon lexical défini pour l'entrainement, et comme je n'ai pas encore de variante dans mon lexique, la variation ne peut être incluse. Quant au principe « un lexème, un paradigme », je sortirai plusieurs paradigmes si je constate qu'un lexème a plusieurs paradigmes possibles. En définitive, ce n'est pas un problème de remplissage de case, mais un problème de remplissage de paradigmes.

Venons-en aux implications des formes candidates. Depuis longtemps, nous avons des généralisations sur les implications. À ma connaissance, WURZEL (1984) est le premier à avoir proposé les conditions de structure paradigmatique du type : « Si ce mot a telle terminaison et est masculin dans cette case, alors il aura telle terminaison au féminin dans une autre case ». On parle alors d'implications

au sein du contenu. Pour les espaces thématiques, on aurait : « Si on a ce type de thème dans cette zone, alors toutes les racines dans cette zone sont identiques ». Par exemple, en français, le futur et le conditionnel ont toujours les mêmes thèmes. Même le verbe le plus irrégulier n'y échappe pas.

Abordons maintenant le modèle de l'» apprentissage par généralisation minimale » [5] que j'expliquerai plus loin. Présenté succinctement, il s'agit d'une manière de calculer les analogies entre des formes. La façon dont ALBRIGHT & HAYES procèdent est très spécifique : ils font la généralisation la plus petite possible pour chaque paire d'analogies qui subissent les mêmes transformations. Si un segment change d'une manière donnée dans deux contextes différents, ALBRIGHT & HAYES essaient de rassembler ces deux contextes et d'en proposer la plus petite généralisation possible. De telles implications ont été employées pour créer des modèles non pas de flexion, mais uniquement de prédictibilité.

3.5.2. Les « petits mondes »

Venons-en maintenant aux « petits mondes ». Les petits mondes sont utilisés en sociologie, parce que, apparemment, les réseaux sociologiques, comme les « réseaux sociaux », fonctionnent de cette manière : des individus connaissent d'autres individus qui eux-mêmes en connaissent d'autres, et les communautés sont en général des ensembles d'individus qui se connaissent, formant ce qu'on appelle une « clique ». Une clique est une partie d'un graphique, contenant des individus tous en relation mutuelle. Nous avons déjà utilisé les petits mondes dans notre laboratoire, pour des analyses de synonymie et de morphologie dérivationnelle, donc ça fait déjà partie de notre culture. Et en l'occurrence, c'est un très petit monde, puisque les paradigmes verbaux n'ont que quarante-huit cases en français, ce qui est très peu.

[5] NdE: « minimal generalization learner » (voir ALBRIGHT & HAYES 2002).

Cependant dès qu'on considère le lexique dans son entier, l'ensemble constitue un réseau immense. Chaque forme appartient à un petit monde, mais ils sont très nombreux dans le lexique.

L'avantage avec les petits mondes, c'est que les propositions de devenir amis proviennent d'individus probablement déjà connus. Si vous êtes relié à quatre personnes qui sont en relation avec un individu X, qui ne fait pas partie de votre réseau et si ces quatre personnes sont toutes reliées entre elles, alors la probabilité que vous connaissiez en fait cet individu X s'avère très élevée. Wiktionnaire utilise ce principe pour les synonymes. Si vous proposez un nouveau mot au Wiktionnaire, il vous dira en général : « pensez-vous que ce mot soit synonyme avec cette liste de mots ? » parce qu'il possède déjà un graphique des synonymes. Si vous suggérez un synonyme, il le prend et propose en retour tous les synonymes plausibles. Donc il y a des manières d'étendre des petits mondes et de créer plus de connexions par l'intermédiaire des voisins. Les petits mondes sont par ailleurs très stables. Si l'on enlève 10% des liens, et qu'on essaie d'étendre à nouveau les petits mondes, on verra qu'on retrouve probablement 99% de la configuration initiale.

Pour la flexion, un petit monde serait une clique. Dans un paradigme, les individus devraient tous être voisins. Je me sers de ce principe comme solution pour extraire des paradigmes : créer des petits mondes et ensuite en extraire des cliques. Voici ma procédure : j'analyse le lexique et j'essaie de créer toutes les analogies possibles entre chaque forme d'un même lexème en cherchant à faire des généralisations de manière minimale sur tout le lexique. Pas minimale, dans le sens d'ALBRIGHT, qui s'efforce d'avoir une généralisation minimale très fine. Moi, je garde seulement la généralisation la plus importante, celle qui a le contexte le plus vaste. Un contexte aussi généralisé que possible, pas

quelque chose de trop fin. À partir de ces analogies, je produis chaque forme, en partant de toutes les représentations lexicales d'un verbe. En définitive, pour chaque case du paradigme, je teste toutes les analogies qui fonctionnent. J'obtiens alors un réseau de grande taille, dont j'essaie d'extraire des cliques pour voir si je peux y trouver un paradigme utile. J'emploie le modèle d'apprentissage par généralisation minimale – ou plutôt une ré-implémentation, mais c'est presque pareil – et j'extrais toutes les analogies. Un tel modèle ne peut extraire que des affixes, ce qui s'avère suffisant pour le français. Je ne conserve que le contexte général, sans indice de confiance, parce que ceux fournis par le modèle d'apprentissage par généralisation minimale sont inutiles pour ma recherche.

Par exemple un tel modèle permet d'extraire la règle : « si on part d'un imparfait et qu'on passe au présent, la règle la plus générale est d'enlever le [ɛ] à la fin du mot ». Cette règle fonctionne pour 5237 verbes sur 6440, ce qui constitue une bonne généralisation, mais j'ai peut-être une vingtaine d'analogies différentes pour ça. Sur tout le lexique, en fusionnant l'ensemble, j'obtiens environ 1600 analogies de case à case, de mot à mot. Avec ces analogies, on n'utilise pas directement le modèle d'apprentissage par généralisation minimale, mais on essaie d'appliquer à nouveau les analogies aux lexèmes, et cela nous donne une distribution des possibilités qui existent réellement dans le lexique. Donc pour chaque forme, on regarde toutes les analogies qu'elle pourrait engendrer et on compte chaque fois que l'analogie est correcte. On parvient alors à une distribution qui couvre le lexique entier, de la manière dont cette ambigüité est effectivement réalisée.

Prenons un exemple. Disons que vous regardez les classes de la figure 1 (page suivante). Si vous regardez la classe 29, c'est censé être ambigu parce que les individus dans cette classe pourraient passer par ces deux règles. Mais en réalité,

aucun ne passe par cette règle-ci, la totalité des cas passent par cette règle-là. Ce n'est pas toujours le cas, comme on peut le voir avec la classe 27, on a plutôt une distribution, et l'output qu'on obtient suit la distribution du lexique d'entrainement (82% de /ɛ/ → [] et 18% de /tɛ/ → [] dans les contextes correspondants à ces deux transformations).

```
class 27 ( syportE ~ syport ):  231 members  points
E --> [] / X[p,t,k,b,d,g,f,s,S,v,z,Z,m,n,J,j,l,r,w,H,i,y,E,6,u,o,ê,û,ô] ___ # : 189—81.82% (supporter, etc.)
tE --> [] / X[p,t,b,d,f,s,v,z,m,n,r,E,6,a,o,ê,û,â,ô][r,E,a,ê,â] ___ # : 42—18.18% (sortir, etc.)
class 28 ( fryskE ~ frysk ):  4027 members  points
E --> [] / X[p,t,k,b,d,g,f,s,S,v,z,Z,m,n,J,j,l,r,w,H,i,y,E,6,u,o,ê,û,ô] ___ # : 4027—100.00% (frusquer, etc.)
local conditional entropy: -0.0
class 29 ( r6d6vE ~ r6dwa ):  2 members  points
E --> [] / X[p,t,k,b,d,g,f,s,S,v,z,Z,m,n,J,j,l,r,w,H,i,y,E,6,u,o,ê,û,ô] ___ # : 0—0.00%
6vE --> wa / X[t,d,s,z] ___ # : 2—100.00% (redevoir, etc.)
```

Figure 1 – Distribution entre deux analogies

On commence avec une représentation du paradigme, qui pourrait être un paradigme entier, une seule forme ou encore plusieurs formes dans une case, qui suivent une distribution : toutes les configurations sont imaginables. On a soit des variantes dans une case, avec des pondérations différentes, ou diverses formes du paradigme et on établit des inférences à partir de l'ensemble. On n'a par conséquent aucune des restrictions qu'on aurait eues si on partait du principe qu'il n'y a qu'une forme par case. Avec des formes cohérentes dans différentes cases, on peut commencer avec ce qu'on veut et voir où ça mène. À titre d'exemple simple, prenons le cas où l'on a une forme par case. Il nous faut développer une paire de paradigmes, parce que la première fois, on a seulement ce que le premier individu indique comme étant ses amis, mais on veut également savoir qui sont les amis de ses amis. Il faut donc développer deux paradigmes. La première et la deuxième fois suivent le même procédé. On prend toutes les formes de toutes les cases et on calcule les analogies. Dans la

figure 2 ci-dessous, les analogies n'ont pas toutes été calculées.

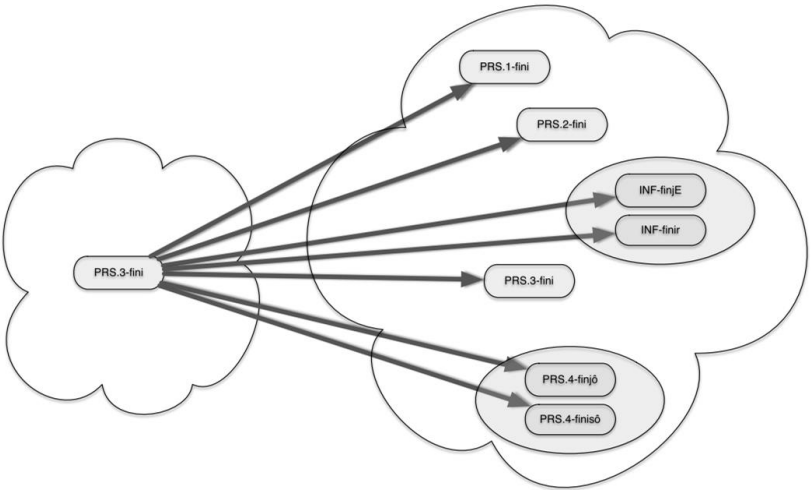

Figure 2 – L'expansion d'un paradigme : première étape

On obtient l'expansion des paradigmes en procédant de la même manière une seconde fois.

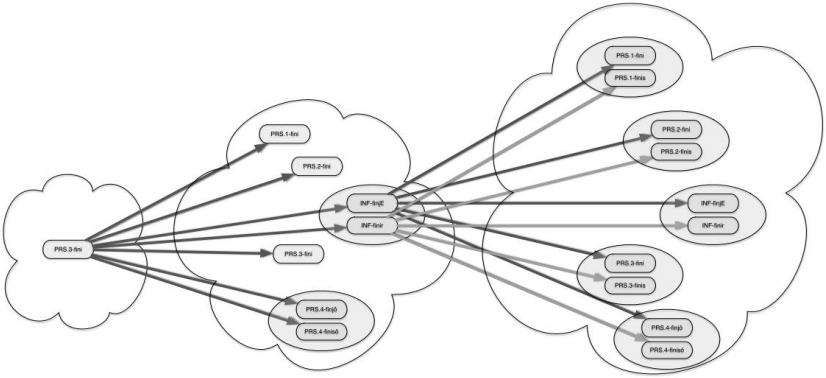

Figure 3 – L'expansion d'un paradigme

Dans la figure 3, la partie générée à la seconde étape et ses liens avec le point précédent sont importants pour collecter les paradigmes, et trouver les cliques. Pour ce faire, au sein

des deux paradigmes obtenus, on examine tous les points connectés, qui forment donc des cliques, qui sont ensuite extraites et devraient constituer le paradigme.

Prenons un exemple (voir figure 4), en démarrant avec la forme /kas/ pour l'impératif 2SG, au terme des deux étapes, on obtient un graphe qui ne contient qu'une seule clique, et donc un seul paradigme correspondant directement à celui de *casser*, rien de compliqué. Il n'existe aucun risque de confusion avec un autre verbe. Cette clique ne pose aucun problème.

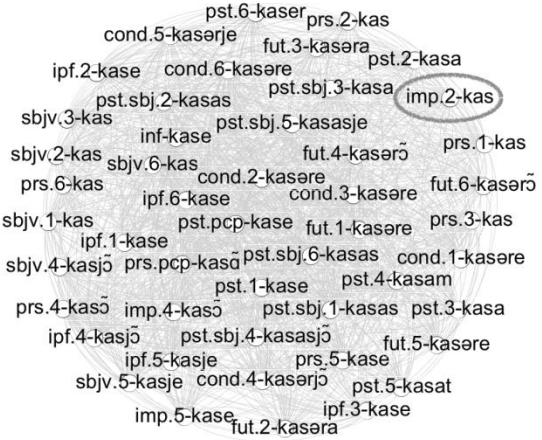

Figure 4 – Graphe à partir de 'imp.2-kas' : une seule clique

Maintenant, si on part de /kas/ non plus à l'impératif 2SG mais au présent 3PL, c'est une autre histoire, parce que 'prs.6-kas' produit plusieurs cliques (voir figure 5, page suivante). Il existe certes une clique complète de 48 formes qui constitue le paradigme du verbe *casser*, mais d'autres formes apparaissent qui n'appartiennent pas au paradigme de *casser*, et qui ont cependant été produites parce que ces analogies pourraient théoriquement convenir. Apparaissent notamment les participes passés *cassi et cassu*.

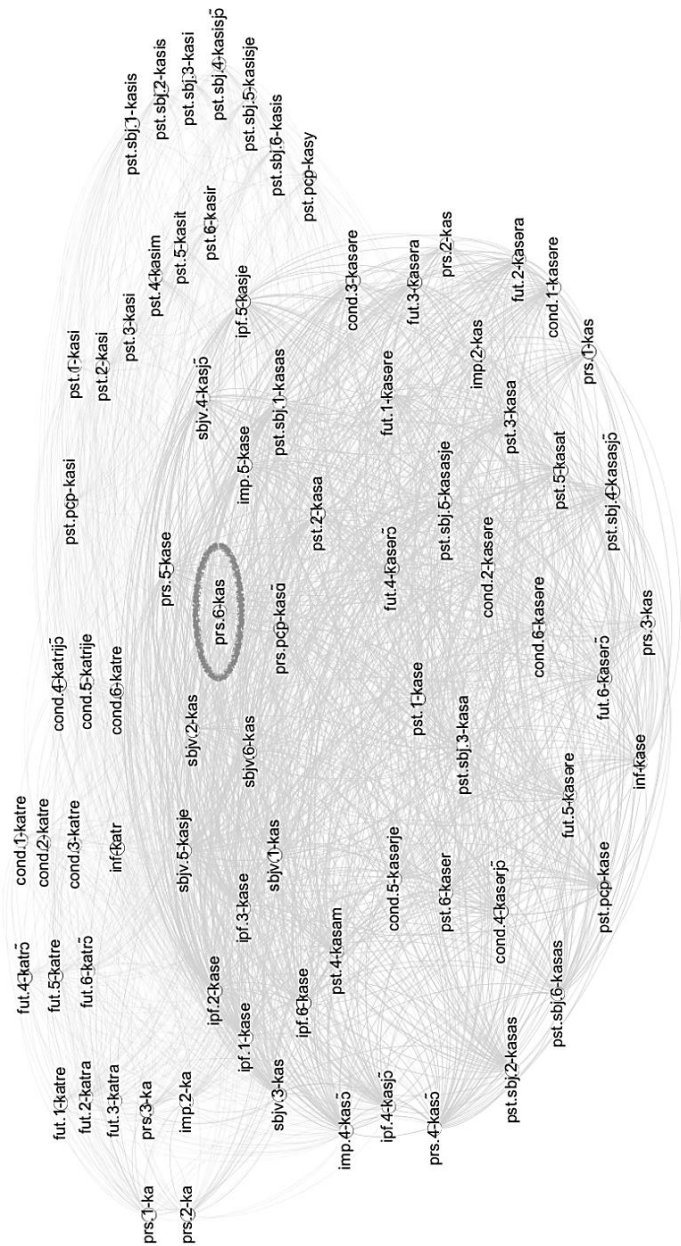

Figure 5 – Graphe à partir de 'prs.6-kas' : plusieurs cliques

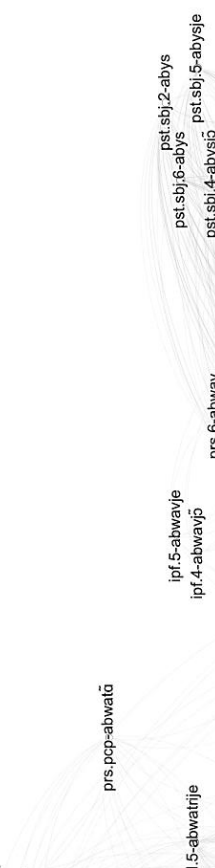

Figure 6 – Graphe à partir de 'prs.3-abwa'

Si nous observons maintenant le graphe de la figure 6 (page précédente), en partant de /abwa/ (*il aboie*), on obtient effectivement le verbe *aboyer* (en bas à droite sur la figure). Cependant on a également d'autres éléments étranges. Certains relèvent de la conjugaison normale du verbe *aboyer*, d'autres pas du tout. Ainsi en est-il de la partie en haut à droite qui donne l'impression qu'on a appliqué le paradigme de *boire* à *aboyer*. Mais si on extrait les cliques, on a bien le paradigme d'*aboyer* avec ses quarante-huit formes. Toutefois, la seconde clique sur le graphique est la suivante (tableau 1). Elle ne contient que trente-cinq membres, et s'avère par conséquent incomplète.

aboyer-PRS3 35 0.541714285714

temps	1SG	2SG	3SG	1PL	2PL	3PL
present	abwa	abwa	abwa			abwav
imperfective						
past	aby	aby	aby	abym	abyt	abyr
future	abware	abwara	abwara	abwarɔ̃	abware	abwarɔ̃
present subj.	abwav	abwav	abwav			abwav
imperfective subj.	abys	abys	aby	abysjɔ̃	abysje	abys
conditional	abware	abware	abware	abwarjɔ̃	abwarje	abware
imperative		abwa				

inf	abwar
pst.pcp	aby

Tableau 1 – La seconde clique de /abwaje/

Le nombre indiqué en haut est l'indice de confiance pour cette clique, qui en l'occurrence s'avère plutôt bonne, avec des liens très forts, puisque /abwa/ (*aboie*) et /bwa/ *boit* sont très proches. Cependant, le réseau ne nous donne pas les quarante-huit formes attendues.

Pour conclure, le procédé proposé ici permettrait de modéliser la flexion sur la base d'analogies entre formes, la surgénération serait tempérée par la recherche de petits

mondes entièrement connectés qui formeraient les paradigmes flexionnels, le meilleur paradigme flexionnel correspondant à une clique couvrant toutes les cases du paradigme flexionnel.

Questions

Légende : « Q » pour « Question », « GB » pour « Gilles BOYÉ »

Q : J'essaie de comprendre quelle est la nature des données que vous entrez dans votre modèle, et je comprends – ou du moins je crois comprendre – que vous avez des formes qui sont déjà complètement analysées, dans le sens où vous savez que c'est par exemple une troisième personne du singulier…

GB : Oui, elles sont complètement analysées, puisqu'elles sont classées dans telle ou telle case.

Q : Mais dans ce cas, d'où provient /aby/ (voir tableau 1, page précédente), je ne comprends pas.

GB : En fait, la première chose que nous faisons est d'analyser le lexique pour avoir un point de départ et pouvoir dire : « Ceux-ci, on les connait. » Nous avons la base, c'est tout ce que nous savons à propos des formes, mais nous n'avons pas les formes, tout ce que nous gardons, ce sont les analogies. Vous pouvez alimenter le programme avec une ou plusieurs formes à la fois, et c'est tout ce qu'il connait. Seul le programme connait toutes les paires entre les 6048 cases – mais il ne sait rien à propos d'une forme en particulier, il ne connait pas le lexique, seulement les analogies qu'il en a extraites. Donc quand il tombe sur quelque chose comme *avoir*, une analogie simple est de dire que ça ressemble à *voir*. L'analogie commence en appliquant le paradigme de *voir* à *avoir*. Mais à la deuxième étape, les analogies vont étendre cet embryon de paradigme sans pour autant parvenir à en étendre la portée sur l'ensemble des cases. De ce fait, cette famille d'analogies ne forme pas une clique complète. Voilà comment nous avons extrait une clique de seulement trente-cinq éléments, et une autre de quarante-huit, parce que des liens manquent. Seules les règles sont reconstruites, dans le cas de la clique complète. Le programme ne connait pas le lexique, mais il sait des choses à propos de ce lexique.

Q : Il sait des choses telles que « si un verbe se termine en /vwaʁ/, alors il pourrait bien donner /vy/ au participe passé » ?

GB : Il a des informations sur les cases. Donc si dans une case, on a une terminaison en /vwaʁ/, il se peut qu'une autre case contienne une terminaison en /vy/. Ça, il le sait grâce à la généralisation de cette alternance entre /wa/ et /y/, à partir de verbes concrets.

Q : Vous n'avez qu'une forme en entrée ou bien…

GB : Des paires, nous n'avons que des paires !

Q : Vous avez besoin de deux formes, de deux formes conjuguées pour générer le paradigme ?

GB : J'ai essayé ça, ce n'est pas un problème. Il est possible de partir d'un paradigme rempli, tout comme il est possible de partir d'une forme seule, ou de n'importe quelle configuration entre ces deux extrêmes.

Q : Qu'est-ce qui se passe si vous avez deux formes divergentes ? Par exemple si vous prenez une forme d'un de vos paradigmes et l'autre forme à partir de…

GB : J'ai essayé avec *assoir*, et vous arrivez à deux cliques.

Q : Donc par exemple, une pour *assieds* et une autre pour *assois* ?

GB : Oui, et elles sont distinctes. C'est parce que pour commencer, mon lexique n'a pas de surabondance. J'imagine que si j'incluais la surabondance dans mon lexique au départ, j'obtiendrais de la surabondance dans mes résultats, avec probablement des cliques plus grandes que quarante-huit formes.

Q : Quelle pourrait-être la réalité psycholinguistique de ce modèle ? Y en a-t-il une, est-ce que vous la cherchez ?

GB : Deux choses à ce propos : mes précédents travaux ont été critiqués sur ce point, et ma défense, auparavant, aurait été de dire « comment est-ce que vous rendez compte de la grammaire des locuteurs adultes ? » et je n'ai pas la moindre idée de comment on peut prendre une grammaire d'enfant et l'amener linéairement à une grammaire d'adulte. Alors tout le monde me disait que ça n'était pas possible, parce que linéairement, on ne peut pas aller d'un point donné à un autre point. À mon avis, en tout cas, il n'y a

pas de raison que ce soit linéaire. Des réorganisations se produisent en permanence, et la progression en U a quelque chose à voir là-dedans. Deuxièmement : cette fois-ci, j'ai pris quelques précautions, et en fait je pense que si vous entrez les données que vous voulez dans le système, vous n'avez plus qu'à savoir qu'il faut associer tel point de données avec tel autre. Vous devez savoir que c'est le même lexème, à un moment, mais vous pourriez trouver les analogies de toute façon, comme vous le voulez, et contrôler le réseau à chaque étape. Donc cela pourrait être linéaire, de cette manière.

Q : Et les fréquences ?

GB : J'ai fait une longue liste de problèmes, et la fréquence en fait partie, je ne veux pas m'y attaquer, parce que je n'ai pas les données nécessaires. Si c'était le cas, je proposerais d'utiliser des distributions, dans les cases et dans les paradigmes, mais je n'ai rien que je puisse employer.

Q : Quand même, dans le *Français fondamental*[6], vous savez que *bois* apparait là comme dans *je bois*, *tu bois*.

GB : Non, c'est une vraie question : de quelle fréquence parle-t-on ? Parce que j'en ai besoin, il me faut une idée de la distribution. À défaut, je n'ai rien. Je sais juste que *bois*, en tant que lexème, est fréquent. Mais il pourrait être pertinent de savoir qu'une forme donnée est fréquente, qu'une autre ne l'est pas, mais je ne peux pas faire de généralisation sans données concrètes.

Q : J'ai pensé à NEW, BRYSBAERT, VERONIS & PALLIER (2007) qui s'occupent d'analyser des sous-titres de films en français. Est-ce que ça aiderait ?

GB : Non.

Q : Et pourquoi pas ? C'est au moins un reflet écrit de l'oral, donc vous pouvez en savoir d'avantage sur les différentes formes.

GB : Bon, j'ai utilisé les fréquences des sous-titres. Je les ai regardées, mais il n'y a pas de futurs. Pourtant, j'entends tout le temps des futurs ! Si je veux intégrer les fréquences, il me faudrait

[6] NdE : le locuteur fait allusion à l'ouvrage de GOUGENHEIM, MICHÉA, RIVENC & SAUVAGEOT (1964).

des données concrètes, provenant d'expériences concrètes de locuteurs concrets. Les sous-titres, c'est loin, loin, très loin de remplir ces conditions. J'adorerais regarder d'autres corpus. Je sais juste que les sous-titres, ça ne va pas suffire.

Q : D'accord.

GB : Mais c'est effectivement une des choses que je dois dire : il me faut des données concernant les fréquences. Est-ce que les lexèmes les plus fréquents ont plus d'influence, ou au contraire moins d'influence sur le système ?

Q : Mais c'est intéressant, non ? Le fait qu'on ne puisse pas faire de prédiction est une bonne raison de s'y intéresser.

GB : En fait, depuis le départ, ma méthodologie a été de travailler avec des psycholinguistes, parce que utiliser les formes du *Bescherelle*, tout le monde peut le faire, de toutes les manières que vous voulez : il n'y a pas de restriction dans l'analyse. La seule contrainte intéressante pour l'analyse, c'est « Comment ça fonctionne dans le cerveau ? », « Comment est-ce que les gens produisent une conjugaison ? », « Comment ça fonctionne ? ». Ce n'est pas intéressant de connaitre *je crois, tu crois, il croit*, etc. Vous pouvez apprendre le *Bescherelle* par cœur, mais d'après ma propre expérience, j'en connais un rayon sur la conjugaison espagnole et italienne, et pourtant je ne peux pas utiliser des verbes correctement ni en espagnol, ni en italien. Je sais tout ce qu'il y a à savoir sur eux, et comment les former, si on me laissait prononcer un mot par minute… mais ça n'a rien à voir avec la connaissance des vrais locuteurs. Donc il me faut cette information psycholinguistique, tout comme il me faut les fréquences réelles, mais comme je l'ai déjà dit : je ne sais pas comment faire pour les obtenir.

Références

ACKERMAN Farrell, BLEVINS James P. & MALOUF Robert (2009). Parts and Wholes: Implicative Patterns in Inflectional Paradigms. In BLEVINS James P. & BLEVINS Juliette (Eds), *Analogy in Grammar: Form and Acquisition*. Oxford: Oxford University Press, 54-82.

ALBRIGHT Adam & HAYES Bruce (2002). Modeling English Past Tense Intuitions with Minimal Generalization. In *Proceedings of the Sixth Meeting of the ACL Special Interest Group in Computational Phonology*, MAXWELL Michael (Ed.), Philadelphia: ACL, 58-69.

BONAMI Olivier & BENIAMINE Sarah (2015). Implicative Structure and Joint Predictiveness. In PIRRELLI Vito, MARZI Claudia, FERRO Marcello (Eds), *Word Structure and Word Usage. Proceedings of the NetWordS Final Conference*, Pisa, March 30-April 1, 2015: http://ceur-ws.org.

GOUGENHEIM Georges, MICHÉA René, RIVENC Paul & SAUVAGEOT Aurélien (1964), *L'Élaboration du Français Fondamental : Étude sur l'Établissement d'un Vocabulaire et d'une Grammaire de Base*, Paris : Didier.

NEW Boris, BRYSBAERT Marc, VERONIS Jean & PALLIER Christophe (2007). The Use of film Subtitles to Estimate Word Frequencies, *Applied Psycholinguistics 28*, 661-677.

WURZEL Wolfgang U. (1984). Flexionsmorphologie und Natürlichkeit. Ein Beitrag zur morphologischen Theoriebildung, *Studia Grammatica 21*, Berlin: Akademie-Verlag.

Cahiers de l'ILSL-45 (2016 : 45-64)

Prévisibilité dans le développement linguistique et importance des corpus électroniques[1]

Wolfgang U. DRESSLER[1]
Katharina KORECKY-KRÖLL[2]
Karlheinz MÖRTH[3]
Institut für Corpuslinguistik und Texttechnologie der Österreichischen Akademie der Wissenschaften (A) (1, 2)
Austrian Centre for Digital Humanities, Österreichische Akademie der Wissenschaften (A) (3)
Correspondance : wolfgang.dressler@univie.ac.at

1. Prévisibilité et prédictibilité

Les notions de prévisibilité et de prédictibilité, la seconde plus puissante que la première (cf. BARRETT & STANFORD 2006), sont discutées dans de nombreux champs de la linguistique, mais souvent dans une perspective différente. Par exemple, en phonétique expérimentale, les résultats des expérimentations sont prédits sur la base d'hypothèses que la méthodologie a opérationnalisées, parallèlement à ce qui se pratique dans les autres sciences expérimentales. En sociolinguistique, la prévisibilité est comprise de la même manière que dans les autres sciences humaines. Dans les grammaires formelles, selon le modèle de CHOMSKY (1957) qui cherche à traiter la théorie de la grammaire comme une science exacte telle la physique ou l'astronomie, une prédiction signifie habituellement que la grammaire d'une langue prédit quelles sont les phrases à classer comme correctes et incorrectes parmi les phrases réalisées ou possibles.

Les prédictions probabilistes sont plus faibles. Prenons par exemple la distribution des composés et de leurs

[1] Traduction de *Vorhersehbarkeit in der Sprachentwicklung und die Bedeutung elektronischer Corpora*, effectuée par Marianne Kilani-Schoch.

constituants respectifs dans les textes. Nous avons soutenu que ceux-ci se répartissent de façon inverse suivant qu'il s'agit du titre ou du corps du texte : les composés apparaissent plutôt dans les titres (ex. *Ausfuhrzölle* 'droits à l'exportation') et les constituants ou membres des composés plutôt dans le corps des textes (*Ausfuhr* 'exportation', *Zölle* 'droits') (DRESSLER & MÖRTH 2012a; b). Une telle prédiction est rendue possible par la convergence de deux facteurs : la pression actuelle en faveur de la brièveté des titres et le fait que les composés sont plus courts que la combinaison syntaxique de leurs parties, comparez *Meeresfreiheit* 'liberté des mers' et *Freiheit der Meere* 'liberté des mers' (sur la convergence des motivations en linguistique, voir DRESSLER *et al.* 2014).

La vérification empirique de cette prédiction probabiliste est difficilement réalisable à partir de la lecture de textes imprimés, à la différence d'une analyse linguistique de corpus basée sur des textes électroniques annotés. C'est ainsi que dans l'étude de DRESSLER & MÖRTH (2012b), le nombre des textes à examiner a considérablement diminué et est passé de 13 000 à 2087 grâce la segmentation automatique des composés en leurs constituants (au moyen du programme *Noun Splitter* de l'*Institut für Corpuslinguistik und Texttechnologie* ou ICLTT 'Institut de linguistique de corpus et de technologie du texte' de l'Académie des sciences autrichienne). Le temps de lecture en a été réduit de plus de 80%. Cette diminution a été d'autant plus marquée que les composés identifiés dans les textes et leurs constituants potentiels ont pu être extraits automatiquement. Les résultats ont dû néanmoins être vérifiés manuellement dans chaque texte.

Notons que les composés représentent un défi considérable pour tout ce qui est traitement automatique des langues naturelles (au sens de *Natural Language Processing*), en particulier pour les nombreuses procédures de traduction

automatique ou de récupération de l'information. On compte de nombreux essais d'automatisation de la segmentation des composés – des techniques d'apprentissage automatique sont notamment utilisées – mais nos travaux avec le *Noun Splitter* ont montré que dans bien des cas le problème ne peut être résolu que par le recours à de grands dictionnaires numériques, malheureusement encore en nombre insuffisant aujourd'hui.

Dans la présente contribution, nous allons concentrer notre attention sur la prévisibilité et la prédictibilité probabiliste dans deux domaines du développement linguistique, d'une part l'acquisition par l'enfant dans sa phase précoce (particulièrement l'acquisition de la grammaire), d'autre part le changement diachronique, au cours de l'histoire, où nous tenterons aussi une comparaison avec les sciences historiques classiques.

2. Acquisition de la langue première

Considérons d'abord l'acquisition de la langue première par les enfants ainsi que l'acquisition successive d'une langue seconde. Ici nous pouvons sans grands risques prévoir des différences interindividuelles dans le rythme d'acquisition des enfants en fonction de la typologie linguistique, c'est-à-dire par exemple prévoir que la vitesse d'acquisition de la morphologie, en l'occurrence celle de la flexion, dépend de de la richesse relative, de la transparence et de l'univocité de la morphologie de la langue-cible (DRESSLER 2010 ; SLOBIN 1985a ; 1985b ; 1992 ; 1997a ; 1997b ; XANTHOS *et al.* 2011). Ainsi, dans les langues qui font l'objet des recherches que nous venons de mentionner, le pluriel est-il d'abord acquis en turc, parce que la formation biunivoque du pluriel turc peut être prévue, c'est-à-dire que la terminaison du pluriel est toujours *-ler* ou *-lar*, selon la voyelle qui précède, par ex. *ev-ler* 'maisons', *oda-lar* 'chambres'. En allemand, par contre, le pluriel, *-(e)n, -s, -e, -er* est largement imprévisible – mais

presque toujours conventionnel, c'est-à-dire fixé pour chaque mot (sur les tendances en matière de prévisibilité partielle des formes de pluriel allemand, voir KÖPCKE 1993 ; WEGENER 2004).

L'étude longitudinale de la langue spontanée requiert une préparation informatique précise des données, par exemple à l'aide de l'ensemble des programmes CLAN (MACWHINNEY 2000) qui permettent une annotation automatique et l'exploitation de données massives (voir également KORECKY-KRÖLL soumis ; LAAHA & KORECKY-KRÖLL à paraitre). Dans un premier temps les enregistrements sont saisis dans un format défini, le format CHAT (MACWHINNEY 2000), puis ils sont vérifiés selon trois procédures différentes : d'abord la transcription est contrôlée par un expert pour minimiser les erreurs d'écoute, ensuite l'alignement des niveaux de CHAT est vérifié à l'aide du programme CHECK, enfin les lacunes dans la standardisation de mots (par ex. *gehma* pour *gehen wir* 'allons') sont repérées. Une fois ces étapes franchies avec succès, le programme MOR ainsi qu'un lexique électronique recherchent quelles entrées de la nouvelle transcription à coder ne sont pas encore enregistrées dans le lexique. L'annotation de ces entrées doit être effectuée manuellement. Selon les objectifs de la recherche, il peut être décidé d'ajouter des annotations supplémentaires et de réaliser une annotation plus détaillée (par exemple distinguant entre verbes faibles et forts et les divisant en sous-classes).

Dans un deuxième temps, on procède à l'annotation morphologique automatique des mots de chaque ligne de la transcription. Comme toutes les formes ambigües ont été relevées, un travail supplémentaire de désambigüisation est nécessaire. Il consiste à décider laquelle des annotations possibles s'impose en contexte, par exemple si le mot *der* dans *da ist der Hund* 'le chien est là' est seulement un article et non pas un pronom démonstratif ou un pronom relatif.

Grâce aux divers programmes de contrôle, ces données codées peuvent rapidement se prêter à une analyse quantitative.

L'acquisition du pluriel en turc est facilitée par le fait que la formation du pluriel est transparente : dans *ev-ler* 'maison-pluriel', *oda-lar* 'salle-pluriel', etc. il est très facile de séparer le marqueur du pluriel du radical du mot. L'allemand en revanche a deux marqueurs, la désinence et, dans beaucoup de mots, l'umlaut, dont l'occurrence n'est généralement pas prévisible. Dans les pluriels allemands *Häus-er* 'maisons', *Mütter* 'mères', le marqueur de pluriel ne peut être isolé aussi aisément des bases *Haus* 'maison', *Mutter* 'mère' qu'il ne l'est en turc. À ces pluriels s'ajoutent des pluriels sans marqueur au niveau du radical, comme *der – die Lehrer* 'le, les professeurs'. Ils constituent un troisième exemple de pluriel qui n'est pas non plus transparent et univoque en général. Car dans le pluriel *die* du singulier *der/das*, il n'y a pas de désinence plurielle transparente segmentable ; en outre *die* est aussi le féminin singulier, et l'indéfini se caractérise au pluriel par l'absence d'article. Ceci montre que la formation du pluriel en allemand n'est ni biunivoque ni univoque, mais ambigüe.

Enfin, intervient le fait que le turc est morphologiquement bien plus riche que l'allemand (spécialement dans la flexion), c'est-à-dire qu'il exprime un plus grand nombre de catégories grammaticales, et par exemple recourt souvent à des désinences au lieu de subordonnées. Un exemple de langue très pauvre en flexion est l'anglais. Or on constate que les enfants turcs prêtent beaucoup plus d'attention à l'acquisition de la flexion (par ex. le pluriel ou les formes verbales) que les enfants qui acquièrent l'anglais comme langue première. La flexion anglaise n'est en outre ni biunivoque (cf. *cow-s* 'vaches' et *ox-en* 'bœufs'), ni toujours transparente (par ex. dans la formation du pluriel *wife* 'femme' – *wiv-es*, *mouse* 'souris' – *mice*). La prédiction,

avérée, est donc que les petits anglophones acquièrent la morphologie de l'anglais plus tard et plus lentement que les enfants turcs. Et, comme prédit, les enfants germanophones acquièrent la flexion, par ex. le pluriel, plus tôt et plus vite que les enfants anglophones, mais plus lentement et plus tard que les enfants turcs. Des langues slaves comme le russe ont aussi une flexion peu transparente et le plus souvent ambigüe (par ex. dans la formation du pluriel), mais plus riche que l'allemand et un peu moins que le turc. À partir de l'interaction des trois facteurs de richesse morphologique, biunivocité et transparence, l'échelle suivante d'acquisition de la morphologie, basée sur des données empiriques, est prévisible : turc – russe – allemand – anglais.

Cette échelle est valable non seulement pour l'acquisition typique, sans problème spécifique, mais aussi pour l'acquisition non-typique perturbée ou retardée. Néanmoins, dans le cas de l'acquisition non-typique, la nature et le degré de sévérité du trouble sont des facteurs plus importants que ceux que nous venons de mentionner (cf. BARTKE & SIEGMÜLLER 2004 ; BAVIN 2009). La prévisibilité du développement langagier y est comparable à la prévisibilité de l'évolution de la maladie dans le pronostic médical, à ceci près que dans la recherche sur l'acquisition du langage, le degré d'incertitude du pronostic ne peut pas (pas encore ?) être mesuré quantitativement.

Dans les prévisions typologiques que nous venons de présenter, nous avons procédé à une simplification importante. Car ce n'est naturellement pas la structure de la langue à acquérir qui agit directement sur le processus de développement mais la réalisation de ce système linguistique dans le parler adressé à l'enfant, ce qu'on appelle l'*input*, dont dépend la production de l'enfant telle que nous l'analysons. Et cet *input* que la personne responsable (le plus souvent la mère) adresse au jeune enfant peut se différencier fortement de la langue des adultes (CAMERON-FAULKNER *et al.* 2003 ;

RAVID *et al.* 2008 ; XANTHOS *et al.* 2011). Par exemple, la morphologie turque comporte un certain nombre de difficultés, tout particulièrement les longues formes fléchies, c'est-à-dire les chaines de désinences successives, auxquelles s'ajoutent les éventuels changements de position de ces suffixes, sans que la signification morphologique n'en soit modifiée. Aucune mère turque, cependant, ne parle à des enfants en bas âge en utilisant des formes aussi complexes. C'est la raison pour laquelle, dans notre « Crosslinguistic Project on Pre- and Protomorphology in Language Acquisition » (DRESSLER 2010 ; XANTHOS *et al.* 2011) qui porte sur la production d'enfants provenant de 18 pays différents, nous enregistrons et étudions aussi bien le développement longitudinal de l'input que celui de l'output.

3. Input, saisie, intégration, production

Il n'y a pas non plus de relation directe entre l'input et l'output de l'enfant, dans la mesure où les êtres humains ne sont pas des perroquets. La chaine causale entre production de l'enfant (*output*) et input est la suivante :

> input → saisie (*intake*) → intégration (*uptake*) → production (*output*).

La saisie (*intake*) correspond à cette partie de l'input que l'enfant capte, notamment pour des raisons de perception. Les syllabes accentuées sont ainsi mieux perçues que les syllabes atones. Les enfants (en tout cas les plus jeunes) enregistrent plus facilement la partie finale d'unités structurales comme les syntagmes nominaux *des klein-en Kind-es* 'du petit enfant', *die klein-en Kind-er* 'les petits enfants' que la partie initiale. Bien que l'occurrence de l'article soit beaucoup plus fréquente dans l'input que celle des désinences flexionnelles (comme dans l'exemple *-en, -es, -er* ci-dessus), les désinences sont acquises plus tôt par les jeunes enfants que l'article (KORECKY-KRÖLL 2011: 190).

Les différences interindividuelles mentionnées plus haut dépendent d'autres facteurs encore, par exemple de l'étape intermédiaire d'intégration (*uptake*) qui consiste dans la forme que les enfants donnent à leur grammaire en la construisant à partir des éléments saisis (*intake*). Dans la relation entre intégration et production enfantine interviennent des facteurs relatifs à l'extraction des données : ainsi les enfants produisent-ils moins "d'erreurs" (dans la perspective de la langue adulte) dans leur usage spontané de la langue que dans des tests formels, comme nous l'avons montré pour l'acquisition des pluriels et des cas en allemand (KORECKY-KRÖLL 2011 ; KORECKY-KRÖLL & DRESSLER 2015). C'est la raison pour laquelle la recherche ne peut se limiter à des tests formels, comme les tests psychologiques et pédagogiques réalisés à des fins de diagnostic et de thérapie (qui ne prennent ni l'input, ni le développement langagier individuel en considération).

En d'autres termes, la prévisibilité en ce qui concerne l'acquisition de la langue première d'un enfant est tributaire d'un nombre important de facteurs. Le principal problème réside donc dans la manière de les évaluer et de les hiérarchiser. L'ordre de présentation suivi ici reflète notre proposition de hiérarchisation.

4. Variable sociolinguistique

En ce qui concerne le recueil des données, une variable supplémentaire est pertinente, à savoir une variable sociolinguistique. Les recherches sur le langage des enfants, en particulier sur la langue spontanée, sont dans la majeure partie des cas conduites auprès de familles dont le niveau de formation est élevé, parce que l'accès est à bien des égards beaucoup plus facile (c'est également le cas dans notre « Crosslinguistic Project on Pre- and Protomorphology in Language Acquisition »). Mais la question se pose de savoir

comment le développement du langage s'effectue dans des familles dont le niveau de formation est peu élevé.

Cet aspect est étudié de manière systématique dans notre programme de recherche INPUT[2] (KORECKY-KRÖLL *et al.* 2015), dans le cadre d'une approche linguistique basée sur corpus, s'appuyant sur des travaux américains et israéliens (HART & RISLEY 1995 ; WEISLEDER & FERNALD 2013). La variable socioculturelle du niveau socioéconomique[3] et du niveau de formation déterminent le degré de richesse de l'input enfantin, parce qu'en règle générale, les mères (plus exactement les personnes en charge des enfants) dont le niveau de formation est plus élevé développent plus d'interactions verbales avec leurs enfants que celles dont le niveau de formation est moins élevé. La chaine causale est donc la suivante :

> niveau socioéconomique → richesse de l'input → saisie → intégration → production

Elle permet de formuler deux prévisions : premièrement l'output des enfants provenant de familles dont le niveau de formation est limité est moins riche que celui d'enfants provenant de familles dont le niveau de formation est plus élevé ; deuxièmement leur développement langagier est plus tardif que celui d'enfants de familles dont le niveau de formation est plus élevé. Les observations dans ce sens se sont multipliées à partir du 20e siècle (cf. OEVERMANN 1972). Mais il n'y a pas pour autant de rapport direct entre niveau socioéconomique et output enfantin, car entre les deux intervient de manière déterminante le style communicatif des parents avec les enfants.

La richesse de l'input et de l'output peut être évaluée par diverses mesures. Dans notre projet nous recourons à la

---

[2] Le projet est financé par l'Académie des sciences de Vienne et le fonds Technologie.

[3] En anglais, *SES* pour *socioeconomic status*.

longueur moyenne des phrases, la diversité du vocabulaire, la quantité de mots utilisés, des mesures de complexité des différentes catégories grammaticales, l'élaboration textuelle d'un récit rapporté, ainsi que la manière dont les adultes s'adressent à leurs enfants (par exemple comment ils réagissent aux erreurs des enfants, cf. KILANI-SCHOCH *et al.* 2009). Sur tous ces points on observe des différences dans l'input et dans l'output en fonction du niveau socioéconomique. À ce stade de la recherche, ces différences correspondent au minimum à des tendances, mais nous pensons parvenir, au terme de l'analyse, à des résultats statistiquement significatifs. À ces différences s'ajoute, comme prévu, un retard du développement linguistique des enfants dont les familles ont un niveau de formation limité. On constate chez ces enfants un retard dans l'acquisition des groupes consonantiques, par exemple à la finale des mots *Obst* 'fruit', *du lobst* 'tu loues' (KORECKY-KRÖLL & DRESSLER à paraitre). La question de savoir si ces enfants, devenus adultes, n'atteindront pas non plus le niveau linguistique des enfants de familles dont le niveau de formation est plus élevé se pose. On s'attend à la persistance d'une différence en ce qui concerne le degré de complexité des phrases et des mots (voir déjà OEVERMANN 1972).

Notre recherche porte aussi sur l'effet de l'input adressé aux enfants dans les garderies. Cependant il est encore plus difficile de prévoir si les éducatrices compensent dans leurs interactions avec les enfants les déficits linguistiques familiaux. En effet, des facteurs relatifs à la formation des éducatrices à côté de variables pédagogiques et personnelles, de la taille et de la composition du groupe d'enfants, ainsi que des facteurs spécifiques aux garderies interviennent.

Considérons maintenant le deuxième domaine de prévisibilité du développement linguistique, le changement des langues au cours de l'histoire, c'est-à-dire le changement

diachronique. Ce domaine est en relation étroite avec celui que nous venons d'examiner puisque le produit de l'acquisition du langage ne coïncide pas avec tous les aspects de la langue des parents, c'est-à-dire avec la langue adulte, et qu'il implique donc le changement linguistique. Le rôle précis de l'acquisition du langage par les enfants dans le changement diachronique est toutefois fortement débattu.

Commençons par un type de prévision qui implique une langue dans sa totalité. Il s'agit de la prédiction très commune selon laquelle une langue minoritaire menacée dans un État où une langue est dominante disparaitra rapidement. Cette prédiction souvent n'est pas réalisée. Elle s'apparente le plus souvent à une projection démographique très vague concernant la diminution du nombre des locuteurs qui ont cette langue minoritaire comme langue maternelle. Car la réduction du système linguistique, au sens d'érosion ou de déclin linguistique, est bien plus importante pour le devenir d'une langue que la réduction du nombre de locuteurs. Dans ce domaine du déclin linguistique, DRESSLER (2011) a proposé de considérer le tarissement de la créativité linguistique comme un facteur prédictif prometteur. Il se manifeste dans l'incapacité à exprimer de nouveaux concepts à travers de nouveaux mots (néologismes) acceptés par la communauté. Par exemple en breton, au 19^{e} siècle, non seulement les néologismes français comme *moissonneuse-batteuse* étaient traduits (breton *dorn-erezh*), mais de nouvelles formations comme *marc'h-houarn* 'bicyclette' (littéralement 'cheval de fer', cf. allemand *Drahtesel*) étaient créées. Après la première guerre mondiale, cependant, ce mouvement a cessé et les nouvelles formations bretonnes proposées n'ont plus été acceptées. Il faut souligner que le déclin linguistique tel qu'il est illustré par le breton ne peut être empêché que par des tentatives énergiques et réussies de revitalisation linguistique comme dans le cas de l'hébreu moderne.

5. Changement diachronique

En ce qui concerne la prévisibilité du changement linguistique dans une langue dont la vitalité est entière et n'est pas menacée de disparition ou de déclin, nous prendrons comme exemple de prévisibilité partielle le pluriel allemand des mots étrangers. WEGENER (2004) a constaté que la terminaison du pluriel des mots étrangers en allemand dans un premier temps est *-s*, puis est remplacée par des désinences plus courantes, par exemple *Ballon-s* 'ballons' devient *Ballon-e*. Cette observation permet de prévoir que les pluriels en *-s* des mots étrangers actuels seront dans le futur substitués par d'autres désinences. Mais, naturellement, elle ne prédit pas quand le changement se produira. Il est aussi possible d'appliquer cette prédiction aux mots étrangers qui ont été empruntés antérieurement. Ce faisant on ne réalise pas une prédiction au sens propre mais une rétrodiction (BARRETT & STANFORD 2006). Une telle rétrodiction s'applique à des cas comme en allemand *General* dont le pluriel a d'abord été *General-s*, puis est devenu *General-e* et enfin *Generäl-e*. Toutefois, dans notre corpus électronique (DRESSLER & MÖRTH 2012a ; MÖRTH & DRESSLER 2014) nous avons trouvé beaucoup d'exemples de pluriels en *-s* utilisés concurremment à d'autres pluriels dans un même mot dès la première attestation, par exemple au début du 19^{e} siècle *Pizza-s* et *Pizz-en*, *Scheich-s* 'cheikhs' et *Scheich-e.* On ne peut donc, en l'occurrence, prévoir qu'une tendance dont la probabilité dépend aussi d'autres facteurs.

De telles rétrodictions ne sont habituellement que probabilistes, et souvent aussi très faibles. Une rare exception est constituée par notre étude sur le développement de la 1ère personne du pluriel du présent des dialectes italoromans depuis leur formation jusqu'à aujourd'hui (SPINA & DRESSLER 2011), c'est-à-dire ceux des dialectes romans (le standard y compris) qui n'appartiennent pas à une autre langue romane d'Italie. En proto-italien, c'est-

à-dire les étapes intermédiaires entre le latin vulgaire et l'attestation la plus ancienne de l'italien, étapes reconstruites avec un haut degré de certitude, on peut partir des formes de 1ère personne du pluriel du présent des trois classes flexionnelles telles qu'elles sont illustrées par les verbes 'aimer', 'craindre', 'terminer' :

> indicatif : *-amo* (par ex. *amamo* 'nous aimons'), *-emo* (par ex. *tememo* 'nous craignons'), *-imo* (par ex. *finimo* 'nous terminons') ;
> subjonctif : *-emo* (*amemo*), *-iamo* (*temiamo*), *-iamo* (*finiamo*).

Dans plusieurs dialectes la distribution de ces six formes est restée la même jusqu'à aujourd'hui. Nous ne cherchons évidemment pas à prédire rétrodictivement quel changement s'est produit dans quel dialecte, et quand (le "problème d'actualisation" de LABOV 2001 : 466 ; 2014). Nous ne traitons pas non plus la question de savoir si la désinence flexionnelle *-mo* a changé et comment. Seul nous intéresse ici le destin des voyelles *-a-*, *-e-*, *-i-* et *-ia-* du radical. Nous ne pouvons donc pas prédire que dans certains dialectes *-e-* accentué s'est développé en *-i-* selon les lois phonétiques générales. Nous limitons le champ de recherche de notre rétrodiction à la distribution morphologique des voyelles du radical dans les classes flexionnelles, distribution qui est largement indépendante des changements des autres formes personnelles (même de la 2e personne du pluriel). Nous prédisons deux résultats pour le domaine de la 1ère personne du pluriel du présent : premièrement, les systèmes susceptibles de succéder au système proto-italien décrit ci-dessus, tels qu'ils sont possibles déductivement et les systèmes impossibles déductivement ; deuxièmement, la probabilité d'occurrence de ces systèmes possibles. Les deux prédictions ont été comparées au développement des dialectes italoromans actuels – lorsque les données disponibles le permettaient. Le champ de recherche est donc

suffisamment vaste pour que la réfutation des rétrodictions soit facilitée. En ce qui concerne l'examen de textes italiens plus anciens, le *Corpus testuale del Tesoro della lingua italiana delle origini* de l'Accademia della Crusca (Florence) a constitué un instrument très utile.

D'abord nous pouvons déduire du modèle de la morphologie naturelle, et spécialement de son application à l'histoire de la langue (DRESSLER 1997 ; 2002 ; KILANI-SCHOCH & DRESSLER 2005), quels changements concevables dans la distribution des formes du radical du proto-italien sont permis par la théorie et quels changements sont exclus. Ce champ de prédictions est ensuite restreint aux changements internes à la morphologie ; en d'autres termes, les changements qui trouvent leur origine dans la phonologie ou la syntaxe ainsi que les dialectes dans lesquels un changement syntaxique analogue à la substitution actuelle en français de *nous parlons* par *on parle* s'est produit, sont exclus de la recherche. Enfin, une prémisse supplémentaire pose que dans le développement typologique du latin à l'italien ainsi qu'à la plupart des langues romanes, la morphologie flexionnelle n'a connu qu'une réduction ou des échanges de formes, et aucun développement (voir, dans les langues romanes, premièrement la disparition des cas latins, du participe futur, du supin, du gérondif, du passif, deuxièmement de l'impératif, de l'infinitif passé, de l'imparfait et du parfait du subjonctif ainsi que la réduction d'autres catégories). Autrement dit il s'est produit une perte de complexité morphologique. Il s'agit donc seulement d'établir quelles formes parmi les six formes flexionnelles mentionnées plus haut se sont substituées aux autres et lesquelles ont disparu. La disparition la plus marquée est survenue en italien standard et dans les dialectes toscans de sa base qui ont remplacé toutes les autres formes par *-iamo*.

À partir du modèle de développement morphologique diachronique de la morphologie naturelle et de son

application aux dialectes italiens (sur une période de plus de mille ans), on peut prédire que 64 changements sont concevables. Parmi ceux-ci la théorie en exclut 52 comme étant impossibles et en retient 12. À notre connaissance les changements exclus par la théorie n'ont effectivement pas eu lieu. Parmi les 12 changements admissibles, seuls deux ne se sont pas produits ; ce sont des changements consécutifs à d'autres changements très rares, de nature vraisemblablement accidentelle parce que des aires dialectales de petite dimension tendent à ne pas se subdiviser en dialectes encore plus limités.

Comme la morphologie naturelle est une théorie des préférences, on peut prévoir quels changements parmi les changements admissibles sont les plus vraisemblables. Ces rétrodictions doivent être compatibles avec le nombre relatif de dialectes différents qui ont connu le même changement. C'est-à-dire que plus un changement déterminé est préféré (dans la dérivation déductive à partir de la théorie), plus son occurrence dans les dialectes italoromans doit être fréquente. De fait, six des changements documentés correspondent à des aires plus étendues et souvent discontinues, tandis que quatre d'entre eux sont limités à des aires très restreintes. Cette contradiction apparente est compatible avec les hypothèses théoriques, dans la mesure où le nombre exact de dialectes qui manifestent un changement particulier/déterminé ne peut être prédit.

Un tel exemple de forte prévisibilité du changement linguistique historique est exceptionnel. Il est le résultat des contraintes très restrictives imposées aux prémisses et aux conditions des rétrodictions. Généralement la prévisibilité des changements diachroniques est plus faible et seulement partielle.

Néanmoins, la prévisibilité en linguistique reste supérieure à la prévisibilité dans les sciences historiques, car ni le développement de l'enfant dans le cours de l'acquisition du

langage, ni le changement diachronique des langues ne connaissent des latitudes de variation comparables à celles des changements historiques. Et dans les deux domaines linguistiques, les intentions des personnes et des groupes jouent un rôle beaucoup plus limité.

De même que les autres sciences cognitives, sociales et culturelles, la linguistique a donc affaire à des phénomènes complexes, comme on l'a vu également avec le développement du langage. Ces phénomènes ne sont que partiellement saisissables du point de vue quantitatif et, en raison du nombre des facteurs, leur prévisibilité ne peut correspondre qu'à des degrés variables de probabilité.

L'ample travail de vérification des phénomènes langagiers qui ont fait l'objet de prévisions est opéré au moyen de l'analyse informatique de grands corpus électroniques. À nos yeux cette entreprise constitue le fondement de la linguistique de corpus. En ce qui concerne les méthodes, celles qui se basent sur des modèles probabilistes se sont largement imposées en linguistique informatique au cours des années. Elles s'appliquent aussi bien aux exigences de base de la linguistique de corpus, telle la lemmatisation automatique (l'attribution/assignation de formes de mots comme *Haus* 'maison', *Hauses, Häuser* au même lemme *Haus*) qu'à l'attribution automatique de classes de mots à l'ensemble des lemmes (par ex. nom à *Haus*, adjectif à *häuslich*, verbe à *hausen*). Dans toutes les procédures d'analyse qui concernent les textes, celles-ci représentent le niveau de base. Leur large diffusion tient principalement au fait que l'application de méthodes statistiques à de nouvelles données apparait comme plus robuste et plus efficace que d'autres méthodes. Ainsi, par exemple l'adaptation d'outils de la linguistique de corpus à de nouvelles langues au moyen d'une approche stochastique est-elle dans de nombreux cas réalisée beaucoup plus rapidement et avec des résultats aussi bons sinon meilleurs que ceux qui étaient obtenus avec une

approche basée sur des règles. Les modèles de Markov cachés font partie des algorithmes le plus souvent utilisés en linguistique informatique aujourd'hui (CARSTENSEN *et al.* 2009).

Références

BARRETT Jeff & STANFORD P. Kyle. (2006). Prediction. In PFEIFER Jessica & SARKAR Sahotra (Eds), *The Philosophy of Science: An Encyclopedia*, New York: Routledge, 585-599.

BARTKE Susanne & SIEGMÜLLER Julia (Eds), (2004). *Williams Syndrome across Languages*. Amsterdam/Philadelphia: John Benjamins.

BAVIN EDITH LAURA (Ed.) (2009). *The Cambridge Handbook of Child Language*. New York: Cambridge University Press.

CAMERON-FAULKNER Thea, LIEVEN Elena & TOMASELLO Michael. (2003). A Construction Based Analysis of Child Directed Speech. *Cognitive Science 27-6*, 843-873.

CARSTENSEN Kai-Uwe, EBERT Christian, EBERT Cornelia, JEKAT Susanne, KLABUNDE Ralf & LANGER Hagen (Eds), (2009). *Computerlinguistik und Sprachtechnologie – Eine Einführung*. Dordrecht: Springer Verlag.

CHOMSKY Noam. (1957). *Syntactic Structures*. The Hague: Mouton.

DRESSLER Wolfgang U. (1997). "Scenario" as a Concept for the Functional Explanation of Language Change. In GVOZDANOVIĆ Jadranka (Ed.), *Language Change and Functional Explanations*, Berlin: Mouton de Gruyter, 109-142.

DRESSLER Wolfgang U. (2002). Naturalness and Morphological Change. In JOSEPH Brian D. & JANDA Richard D. (Eds), *The Handbook of Historical Linguistics*, Oxford: Blackwell, 461-471.

DRESSLER Wolfgang U. (2010). A Typological Approach to First Language Acquisition. In KAIL Michèle & HICKMANN Maya (Eds), *Language Acquisition Across Linguistic and Cognitive Systems*, Amsterdam: Benjamins, 109-124.

DRESSLER Wolfgang U. & MÖRTH Karlheinz. (2012a). Vom Einfluss der Pragmatik auf die Grammatik, insbesondere in der Entwicklung der Pluralbildung. Eine corpusbasierte Untersuchung. *Historische Pragmatik. Jahrbuch für Germanistische Sprachgeschichte 3-1*, 75-93.

DRESSLER Wolfgang U. & MÖRTH Karlheinz. (2012b). Produktive und weniger produktive Komposition in ihrer Rolle im Text an Hand der Beziehungen zwischen Titel und Text. In GAETA Livio & SCHLÜCKER Barbara. (Eds), *Das Deutsche als kompositionsfreudige Sprache*, Berlin: de Gruyter, 219-233.

DRESSLER Wolfgang U., LIBBEN Gary & KORECKY-KRÖLL Katharina. (2014). Conflicting vs. Convergent vs. Interdependent Motivations in Morphology. In MACWHINNEY Brian, MALCHUKOV Andrej & MORAVCSIK Edith (Eds), *Competing Motivations in Grammar and Usage*, Oxford: Oxford University Press, 181-196.

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

KILANI-SCHOCH Marianne & DRESSLER Wolfgang U. (2005). *Morphologie Naturelle et Flexion du Verbe Français*. Tübingen: Gunter Narr.

KILANI-SCHOCH Marianne, BALČIUNIENE Ingrida, KORECKY-KRÖLL Katharina, LAAHA Sabine & DRESSLER Wolfgang U. (2009). On the Role of Pragmatics in Child-Directed Speech for the Acquisition of Verb Morphology. *Journal of Pragmatics 41*, 219-239.

KÖPCKE Klaus-Michael. (1993). *Schemata bei der Pluralbildung im Deutschen: Versuch einer kognitiven Morphologie*. Tübingen: Narr.

KORECKY-KRÖLL Katharina (2011). *Der Erwerb der Nominalmorphologie bei zwei Wiener Kindern: Eine Untersuchung im Rahmen der Natürlichkeitstheorie*. Wien: Universität Wien. Dissertation.

KORECKY-KRÖLL Katharina & DRESSLER Wolfgang U. (2015). The Acquisition of Case in German: A Longitudinal Study of Two Viennese Children. *Studi e Saggi Linguistici 53-1*, 9-36.

KORECKY-KRÖLL Katharina, UZUNKAYA-SHARMA Kumru, CZINGLAR Christine & DRESSLER Wolfgang U. (2015). Das INPUT-Projekt: Herausforderungen auf dem Weg zum Bildungserfolg von ein- und zweisprachigen Wiener Kindergartenkindern. In ANREITER Peter, MAIRHOFER Elisabeth & POSCH Claudia (Eds), *ARGUMENTA. Festschrift für Manfred Kienpointner zum 60. Geburtstag*, Wien: Praesens, 201-213.

KORECKY-KRÖLL Katharina & DRESSLER Wolfgang U. (à paraitre). (Mor)phonotactics in High vs. Low SES Children. In DZIUBALSKA-KOŁACZYK Katarzyna & WECKWERTH Jaroslaw (Eds), *Volume in Memoriam of Rajendra Singh*, Poznań: Adam Mickiewicz University Press.

KORECKY-KRÖLL Katharina. (soumis). Kodierung und Analyse mit CHILDES: Erfahrungen mit kindersprachlichen Spontansprachkorpora und erste Arbeiten zu einem rein erwachsenensprachlichen Spontansprachkorpus. In RESCH Claudia & DRESSLER Wolfgang U. (Eds), *Korpusbasierte Linguistik*, Wien: Verlag der Österreichischen Akademie der Wissenschaften.

LAAHA Sabine & KORECKY-KRÖLL Katharina. (à paraitre). Verschriftung, Kodierung und Analyse von Kindersprache mit CHILDES. In WANDL-VOGT Eveline & KORECKY-KRÖLL Katharina (Eds), *Transkriptionssysteme im Vergleich: Sprache - Ton - Bild. Kodierung gesprochener Sprache*, Wien: Praesens.

LABOV William. (2001). *Principles of Linguistic Change: Social Factors*. Oxford: Blackwell.

LABOV William. (2014). The Sociophonetic Orientation of the Language Learner. In CELATA Chiara & CALAMAI Silvia. (Eds), *Advances in Sociophonetics*, Amsterdam: John Benjamins, 17-29.

MACWHINNEY Brian. (2000). *The CHILDES Project: Tools for Analyzing Talk. Volume 1: Transcription Format and Programs*. Mahwah: Lawrence Erlbaum.

MÖRTH Karlheinz & DRESSLER Wolfgang U. (2014). German Plural Doublets with and without Meaning Differentiation. In RAINER Franz, DRESSLER Wolfgang U., GARDANI Francesco & LUSCHÜTZKY, Hans C. (Eds), *Morphology and Meaning*, Amsterdam: John Benjamins, 249-258.

OEVERMANN Ulrich. (1972). *Sprache und soziale Herkunft. Ein Beitrag zur Analyse schichtenspezifischer Sozialisationsprozesse und ihrer Bedeutung für den Schulerfolg*. Frankfurt am Main: Suhrkamp.

RAVID Dorit, DRESSLER Wolfgang U., NIR-SAGIV Bracha, KORECKY-KRÖLL Katharina, SOUMAN Agnita, REHFELDT Katja, LAAHA Sabine, BERTL Johannes, BASBØLL Hans & GILLIS Steven. (2008). Core morphology in child directed speech: Crosslinguistic corpus analyses of noun plurals. In BEHRENS Heike (Ed.), *Corpora in Language Acquisition Research: History, Methods, Perspectives*, Amsterdam/Philadelphia: John Benjamins, 25-60.

SLOBIN Dan Isaac (Ed.) (1985a). *The Crosslinguistic Study of Language Acquisition. Volume 1.* Hillsdale: Lawrence Erlbaum.

SLOBIN Dan Isaac (Ed.) (1985b). *The Crosslinguistic Study of Language Acquisition. Volume 2.* Hillsdale: Lawrence Erlbaum.

SLOBIN Dan Isaac (Ed.) (1992). *The Crosslinguistic Study of Language Acquisition. Volume 3*. Hillsdale: Lawrence Erlbaum.

SLOBIN Dan Isaac (Ed.) (1997a). *The Crosslinguistic Study of Language Acquisition. Volume 4*. Mahwah: Lawrence Erlbaum.

SLOBIN Dan Isaac (Ed.) (1997b). *The Crosslinguistic Study of Language Acquisition. Volume 5*. Mahwah: Lawrence Erlbaum.

SPINA Rossella & DRESSLER Wolfgang U. (2011). How far Can Diachronic Change be Predicted: the Case of Italo-Romance First Person Plural Present Indicative. *Diachronica 28*, 499-544.

WEGENER Heide. (2004). Pizzas und Pizzen, die Pluralformen (un)assimilierter Fremdwörter im Deutschen. *Zeitschrift für Sprachwissenschaft 23 - 1*, 47-112.

WEISLEDER Adriana & FERNALD Anne. (2013). Talking to Children Matters: Early Language Experience Strengthens Processing and Builds Vocabulary. *Psychological Science 24-11*, 2143 –2152.

XANTHOS Aris, LAAHA Sabine, GILLIS Steven, STEPHANY Ursula, AKSU-KOÇ Ayhan, CHRISTOFIDOU Anastasia, GAGARINA Natalia, HRZICA Gordana, KETREZ F. Nihan, KILANI-SCHOCH Marianne, KORECKY-KRÖLL Katharina, KOVAČEVIĆ Melita, LAALO Klaus, PALMOVIC 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.

Cahiers de l'ILSL-45 (2016 : 65-93)

L'utilisation des corpus oraux pour la recherche en (psycho)linguistique[1]

Mirjam ERNESTUS
Radboud University Nijmegen & Max Planck Institute for Psycholinguistics (NL)
m.ernestus@let.ru.nl

1. Introduction

Ma présentation abordera l'usage des corpus de langue dans la recherche en (psycho)linguistique. Elle se divise en deux parties. La première concerne la question : quelle est la structure phonétique des mots dans le flux de parole ? En général, ce sont des « linguistes de bureau », qui, en songeant à leur propre manière de produire des énoncés, tentent d'y répondre. Mon second questionnement portera sur la manière dont les locuteurs produisent de l'oral, point habituellement traité à l'aide d'expériences en laboratoire. Cependant, j'évoquerai pour ma part l'apport des corpus oraux à ces deux questions. Le problème avec « la linguistique de bureau » ou les expériences de production en laboratoire, c'est qu'elles ne montrent pas réellement comment les individus se comportent lors d'une conversation naturelle. Or, je pense que les corpus de conversations spontanées peuvent contribuer à répondre à ces deux questions, et j'espère pouvoir vous en convaincre. Je commencerai par discuter de l'utilité de ces corpus en évoquant succinctement les types de questions qu'ils permettent de résoudre. J'évoquerai également les problèmes auxquels nous avons été confrontés lorsque que nous avons utilisé des corpus pour répondre à nos questions.

[1] *The use of speech corpora for (psycho)linguistic research*. Transcription, traduction et adaptation par Guillaume Feigenwinter et Christian Surcouf.

2. Quelques questions autour de la structure phonétique des mots dans le flux de parole

Commençons par la première question : quelle est la structure phonétique des mots dans le flux de parole ? On trouve évidemment des réponses dans les écrits spécialisés, et j'aimerais poser à ce sujet deux questions, que je vais accompagner chaque fois d'un exemple : 1) « les faits qu'on trouve dans la littérature spécialisée sont-ils exacts ? », 2) : « la littérature spécialisée passe-t-elle à côté de faits intéressants ? ». Mon attention se portera sur une étude qu'on a réalisée il y a déjà plusieurs années, au sujet de l'assimilation régressive de voisement en néerlandais (ERNESTUS *et al.* 2006).

2.1. L'assimilation régressive de voisement en néerlandais

L'assimilation régressive de voisement en néerlandais est importante d'un point de vue théorique parce que cette langue se comporte apparemment de manière étrange, et que ce phénomène a un impact dans toutes sortes de théories concernant le voisement. Dès lors, nos données doivent être correctes. Je vais d'abord présenter ce que dit la littérature spécialisée sur le néerlandais, pour en venir à ce que nous avons trouvé en nous basant sur le Corpus Oral du Néerlandais[2], également évoqué par Steven GILLIS (page 95).

On a les explications traditionnelles, qu'on trouve partout dans la littérature spécialisée, y compris dans celle que j'estime, et donc pas seulement dans la « mauvaise » littérature, mais bien dans des textes plutôt sérieux, écrits par des linguistes reconnus (voir par exemple BOOIJ 1999). Il y est rapporté qu'en néerlandais, si une syllabe se termine par une constrictive sourde, ce qui est systématique en raison du dévoisement en position finale – comme en allemand –, et qu'elle est suivie d'un /d/ ou d'un /b/, alors on a une

[2] Corpus Gesproken Nederlands (voir OOSTDIJK 2002).

assimilation régressive de voisement, c'est-à-dire que l'obstruante sourde devient sonore. Par exemple, si on a un /p/ suivi d'un /d/, alors ce /p/ devient /b/. En revanche, le contraire ne pourrait jamais se produire en néerlandais, c'est-à-dire que, dans une séquence /p/ – /d/, la seconde consonne soit dévoisée. Donc on n'aurait jamais d'assimilation progressive en néerlandais. C'est très important pour de nombreuses théories sur le voisement.

Nous avons procédé à l'étude de ce phénomène en recourant au Corpus Oral du Néerlandais. Étant donné la complexité de la tâche, nous avons simplifié la recherche en nous concentrant sur la partie du corpus la plus simple à transcrire : les histoires lues aux aveugles. Ces textes sont racontés de manière vivante, mais restent de l'écrit oralisé, ce qui toutefois vaut mieux que d'essayer d'imaginer dans l'abstrait comment les mots sont prononcés. Tous les mots dans lesquels une obstruante sourde précédait un /b/ ou un /d/ ont été sélectionnés et retranscrits par trois personnes. Le tableau 1 présente le nombre d'occurrences dans lesquelles nous avons trouvé une assimilation régressive de voisement, censée être obligatoire dans ces groupes consonantiques. Or on n'en trouve pas dans plus de la moitié des cas ! D'où notre surprise, puisque ce phénomène est prétendument obligatoire.

CATÉGORISATION	ASSIMILATION	N	POURCENTAGE
+sonore +sonore	Régressive	261	42,9%
-sonore +sonore	Non	121	19,9%
-sonore -sonore	Progressive	151	24,8%
absent présent		57	9,4%
autre		19	3,1%

Tableau 1 – L'assimilation dans les histoires lues aux aveugles issues du Corpus Oral du Néerlandais (ERNESTUS *et al.* 2006 : 1042)

Nous avons ensuite étudié le nombre de groupes consonantiques dans lesquels on ne trouvait aucun voisement, soit presque 20% des cas, ce à quoi on pouvait

encore s'attendre. Mais ce qui nous a particulièrement étonnés, c'est que 25% des cas relèvent d'une assimilation *progressive* de voisement, c'est-à-dire un phénomène censé *ne pas exister* en néerlandais, mais se produisant en réalité dans un cas sur quatre (voir tableau 1 page précédente). En somme, il ne s'agit pas d'un ou deux cas isolés, mais bien d'un nombre élevé d'occurrences. Nous avons également trouvé d'autres cas, sur lesquels je ne m'attarderai pas, où parfois la première obstruante était absente, et d'autres phénomènes encore. Mais comme le démontre cette analyse, il est indispensable d'étudier ce que les gens produisent *effectivement* dans leur quotidien pour savoir quelles configurations se trouvent dans la langue. En l'absence de données issues de corpus oraux, le simple fait d'affirmer que l'assimilation régressive de voisement est obligatoire en néerlandais ne suffit pas à décrire ce que font les locuteurs dans la réalité. Cela montre peut-être ce qu'ils « devraient » faire, mais pas ce qu'ils font. En définitive, les faits décrits dans la littérature phonologique ne sont pas toujours exacts.

2.2. Le recours aux corpus oraux et ses difficultés

À présent, j'aimerais parler des difficultés rencontrées lors de notre analyse. Notre étude portait sur tous ces groupes d'obstruantes, et nous avions trois transcripteurs – en l'occurrence des personnes qualifiées pour ce travail. Cependant – c'est un fait reconnu –, même les transcripteurs hautement qualifiés font beaucoup d'erreurs. Si on leur demande une première fois de transcrire un certain nombre d'obstruantes, et qu'on leur demande d'effectuer la même tâche une seconde fois, on observe souvent des différences. C'est qu'il est très difficile de faire ces transcriptions, en particulier quand la langue n'est pas articulée très clairement. Comme ce travail demande beaucoup de concentration, les transcriptions contiennent des erreurs si elles sont réalisées par des transcripteurs humains. Un des problèmes provient

notamment des attentes des transcripteurs face au phénomène analysé : s'ils croient qu'un schwa sera présent, il est probable qu'ils en entendent effectivement un. S'ils pensent que l'assimilation régressive de voisement se produit en néerlandais, ils auront tendance à transcrire ces groupes consonantiques avec des assimilations régressives de voisement. Le transcripteur humain, quel que soit son niveau de compétence, tend à se faire influencer par ses attentes. Voilà pourquoi on trouve des différences entre les participants.

J'ai moi-même effectué une expérimentation assez ingrate, il y a longtemps, dans laquelle j'étudiais la prononciation du mot néerlandais « natuurlijk » (« bien entendu », « naturellement ») (voir ERNESTUS 2000). Pour l'occasion j'ai demandé à deux excellents transcripteurs, de transcrire 274 occurrences de ce mot. Dans 58% des cas, il y avait désaccord sur la présence ou l'absence d'un schwa. J'étais quelque peu désemparée et j'ai décidé de ne pas étudier ces données… D'autres fois, ça s'est mieux déroulé. Par exemple, lorsqu'il a fallu transcrire le voisement des occlusives intervocaliques parmi deux mille occurrences, nous avons obtenu seulement 15% de désaccord, ce qui est beaucoup mieux. Dans cette étude-ci, nous avions trois transcripteurs pour travailler sur le voisement des groupes consonantiques d'obstruantes, et dans un cas sur trois, ils n'ont pas pu se mettre d'accord. C'est donc un véritable problème dans ce genre d'études. Alors que faire ? Comment gérer ces nombreux points de vue conflictuels ? En cas de désaccords entre transcriptions, une des solutions consisterait à demander à l'un ou aux deux transcripteurs s'ils seraient d'accord de changer leur transcription. C'est un peu risqué. Ce qui arrive en général, c'est que le plus jeune des transcripteurs aura davantage tendance à modifier son travail. Donc vous obtiendrez les transcriptions des personnes les plus sures d'elles ; ce qui ne garantit pas pour

autant la qualité de la transcription. J'écarterais donc cette option. L'autre possibilité consisterait à éliminer toutes les transcriptions conflictuelles, ce qui par exemple dans mon étude, où 58% des jugements étaient problématiques – reviendrait à éliminer de grandes quantités de données, voire la plus grande partie. Et on connait les problèmes soulevés par les travaux sur des données trop restreintes. Ceci constituait un premier problème.

Un autre problème vient de ce qu'en mettant de côté les résultats problématiques, on risque d'écarter des occurrences présentant en fait un grand intérêt. En effet, si les transcripteurs ne sont pas d'accord, c'est peut-être pour une bonne raison. Par exemple, si nous revenons à notre étude sur le voisement des obstruantes, il se pourrait que les transcripteurs n'aient pas été d'accord parce qu'il est relativement difficile de déterminer avec certitude si une occurrence est sourde ou sonore. Il est possible que le locuteur ait produit un son intermédiaire, ne permettant pas aux transcripteurs de trancher pour l'une ou l'autre des options, conduisant dès lors à des jugements divergents. Voilà pourquoi ces occurrences pourraient en réalité s'avérer vraiment intéressantes. Aussi me parait-il préférable de les conserver pour les analyser, surtout si j'observe que la plupart des conflits surviennent dans un contexte particulier. En somme, si on écarte tous les jugements conflictuels, on élimine non seulement une des conditions que l'on entendait tester à l'origine, mais aussi un phénomène potentiellement digne d'intérêt pour la recherche. Exclure de telles données n'est donc, à mon avis, pas un bon choix. Alors que peut-on faire d'autre ?

Il reste en somme deux solutions. La première est de prendre la transcription de la majorité. Nous aurions pu choisir cette option et opter pour la version qui mettait d'accord au moins deux des trois transcripteurs. C'est envisageable, mais on perd cependant de l'information

puisque pour certaines occurrences, les trois transcripteurs étaient d'accord alors que pour d'autres ils ne l'étaient pas, ce qui en soi peut justement s'avérer intéressant. Au final, pour cette étude, nous avons décidé de garder toutes les transcriptions et de les analyser statistiquement. De cette manière, nous savions quelles occurrences posaient problème. Je n'entrerai pas ici dans les détails statistiques, mais je tiens à insister sur l'intérêt de conserver toutes les versions, même si elles sont conflictuelles, parce que, en définitive, ces informations sont pertinentes.

On pourrait se demander pourquoi pour cette étude sur le voisement des obstruantes, nous n'avons pas tout simplement recouru à des mesures acoustiques, d'autant plus qu'un tel traitement peut en grande partie s'effectuer de manière automatique et par exemple fournir le pourcentage d'obstruantes produites avec vibration des cordes vocales, permettant ainsi de discriminer les groupes consonantiques sourds et sonores. En fait, si nous n'avons pas recouru à cette méthode, c'est qu'elle est en réalité impossible. En effet, un trait tel que le voisement n'est pas seulement déterminé par la vibration des cordes vocales, mais aussi par la durée de la voyelle qui précède, la durée de l'obstruante, l'intensité de la friction, etc. Plusieurs dimensions seraient donc à prendre en compte, et mesurer n'est qu'une première étape. Il faudrait ensuite entrer les résultats dans des catégories spécifiques. Or nous n'avons pas la moindre idée de comment procéder à une telle catégorisation. En somme, les mesures acoustiques n'étaient pas possibles non plus, nous contraignant donc à recourir à des transcripteurs humains.

Pour conclure cette première partie, comme le montrent les résultats de nos recherches (ERNESTUS, LAHEY, VERHEES & BAAYEN 2006), il faut prendre avec beaucoup de prudence les descriptions fournies par la littérature spécialisée.

3. L'intérêt des corpus oraux

J'aimerais maintenant évoquer l'un des intérêts des corpus oraux, qui permettent aux chercheurs de découvrir des phénomènes dont ils ignoraient complètement l'existence auparavant. Je montrerai quelques-uns de ces aspects, à propos de la réduction, sur laquelle je travaille depuis plusieurs années (ERNESTUS 2000 ; ERNESTUS & WARNER 2011) (ERNESTUS 2014). Prenons l'exemple en français du syntagme « le ministre de la culture », extrait d'un des corpus, que nous avons compilés. Chacun sait comment prononcer le mot « ministre ». Or dans notre corpus, sa prononciation la plus fréquente est [mis], c'est tout ! (voir BRAND & ERNESTUS 2015). Un tel phénomène pourrait laisser perplexe, pourtant si on entend cette même séquence dans son contexte d'origine, on la trouve alors naturelle, même si une fois averti, on entend bien que le mot est réduit. Maintenant, ce qui est intrigant et à mon sens très intéressant, c'est que les francophones natifs n'ont pas conscience de tout le temps parler de cette manière, et je pense vraiment que les natifs le font en permanence sans en être conscients, ni en tant que locuteur ni en tant qu'auditeur. La seule façon de trouver la prononciation de mots comme celui-ci est d'étudier les corpus d'oral spontané. Prenons en anglais cette fois-ci l'exemple de *probably*. Une des prononciations possibles, et même la plus courante ressemble à [pɹɒbli]. On a deux syllabes au lieu de trois, et c'est la prononciation la plus fréquente. Ce que l'on constate, c'est que dans la conversation, les mots ont tendance à subir des réductions. Des phonèmes, voire des syllabes complètes peuvent disparaitre. Ce n'est pas limité au français ou à l'anglais. On a trouvé des phénomènes identiques dans de nombreuses langues, comme en néerlandais, où *wedstrijd* (« jeu ») ressemble parfois à /ʊɛs/ au lieu de /ʊɛtstrɛit/. En allemand, on peut dire [vaːŋ] au lieu de [vaːgən] *Wagen* (« véhicule »). En français, *fenêtre* peut se prononcer [fnɛːtʁ].

En espagnol, où on est censé dire [entonθes] *entonces* (*alors*), ça ressemble souvent plus à [entons] (« alors »). En tchèque, on remplace souvent la prononciation formelle /jɛstli/ (*si*) par [ɛs] (ERNESTUS & WARNER 2011). J'insiste sur le fait qu'il ne s'agit pas d'exemples isolés.

De temps à autre, je remarque que des chercheurs utilisent les corpus oraux pour y trouver à tout prix une confirmation de leur théorie. Ainsi, partant de l'hypothèse qu'une prononciation donnée devrait exister, leur théorie se verrait confirmée s'ils en trouvent ne serait-ce qu'une seule occurrence. Je pense qu'une telle démarche est inacceptable. En effet, une occurrence unique peut tout à fait résulter d'une erreur ou d'une idiosyncrasie du locuteur. Bien qu'il s'agisse là d'une évidence, je tiens à insister sur ce point : observer une seule occurrence ne suffit pas, il en faut au contraire une grande quantité avant de pouvoir se prononcer sur les phénomènes se produisant *effectivement* dans une langue. Telle est en définitive la finalité de notre recherche. Aux États-Unis, JOHNSON (2004) a cherché à savoir avec quelle fréquence les mots se trouvaient réduits en anglais, et a établi qu'il manquait au moins un son dans 25% des mots, et au moins une syllabe dans 6% des mots. En somme, un tel phénomène s'avère relativement fréquent. Nous avons mené le même type de recherche sur le français et le néerlandais, langues pour lesquelles nous obtenons des résultats similaires (ADDA-DECKER *et al.* 2005 ; SCHUPPLER *et al.* 2011). Il y a donc toutes ces séries de données, auxquelles on peut s'intéresser pour étudier la réduction des mots, et on pense immédiatement à toutes les transcriptions déjà faites, et au temps considérable qu'elles ont demandé (voir GILLIS, page 95).

4. Les défis de la transcription

Si on veut connaitre la fréquence des réductions dans une langue, il faut de grandes quantités de données orales, et on

se retrouve confronté au cout élevé de la transcription, aux erreurs de transcription des transcripteurs humains, etc. Pour ces raisons, la transcription humaine nous a paru impossible, il fallait procéder autrement. Nous avons donc essayé de la faire de manière automatique à l'aide d'un dispositif de reconnaissance vocale (« Automatic Speech Recognizer » en anglais). À cet effet, nous avons utilisé le logiciel HTK (« Hidden Markov Model Toolkit »), téléchargeable gratuitement et facile à manipuler en raison de la qualité de son manuel d'utilisation. Il se base sur les modèles cachés de Markov. On l'alimente avec un fichier .wav constituant le signal de parole. Mais il faut là encore une transcription orthographique, que l'on doit faire faire à des humains s'il s'agit d'oral conversationnel. Le dispositif nécessite deux autres apports :

1) un dictionnaire de prononciation, indiquant quels mots peuvent se trouver dans la transcription orthographique et comment ils peuvent être prononcés. Ce qui signifie que le chercheur doit d'abord réfléchir à toutes les prononciations possibles. Par exemple, on pourrait imaginer les variantes [fənɛtʁə], [fənɛtʁ], [fnɛtʁ], [fnɛt] pour *fenêtre*, ou [ministʁə], [ministʁ], [minisʁə], [minisʁ], [minis], [mis] pour *ministre* ;
2) des modèles phoniques, c'est-à-dire des symboles, des lettres, qui doivent être associés au signal. En d'autres termes, les modèles phoniques sont une sorte de moyenne, une manière de montrer ce à quoi ressemblerait normalement un segment donné du signal acoustique. Donc il faut traduire ces symboles en format .wav.

Une fois toutes ces données rentrées dans le dispositif, on obtient un résultat comme celui de la figure 1 ci-après :

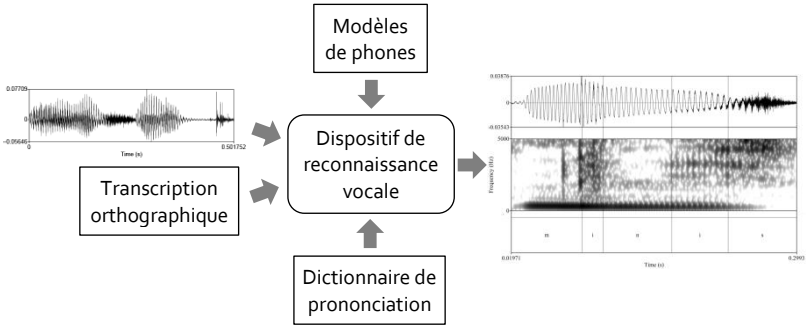

Figure 1 – Le dispositif de reconnaissance vocale automatique

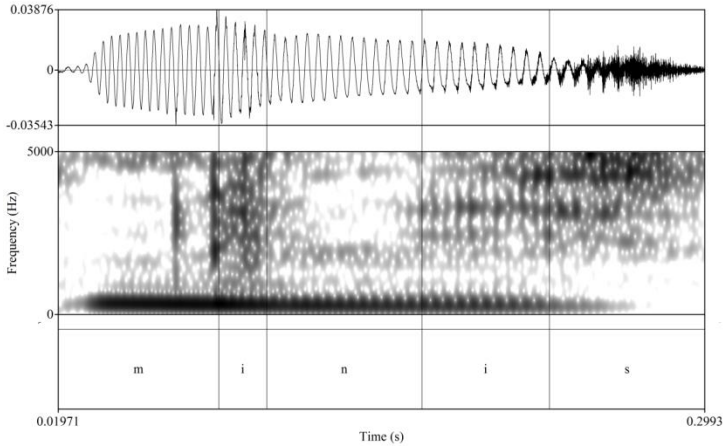

Figure 2 – Le spectrogramme d'une prononciation de *ministre*

La question est évidemment de savoir si la reconnaissance vocale automatique offre de bons résultats, ce qui dépend de plusieurs facteurs, dont deux très importants. Premièrement, la qualité des modèles phoniques, que l'on crée à partir de données orales déjà alignées au niveau des phones, par un transcripteur humain qui a procédé à ce travail d'alignement en décidant que tel son était un /s/, tel autre un /b/, etc. Plus on a de données, meilleurs seront les modèles de phones. Bien entendu, les erreurs dans la transcription phonétique –

et il y en a parce qu'elle est effectuée par des humains – se répercutent dans les modèles phoniques. Prenons par exemple le cas du néerlandais, où, pour certaines fricatives, nous confondons souvent la sourde et la sonore, par exemple /ɣ/ et /x/ ou /v/ et /f/. Donc quand on transcrit ces phonèmes, on voit apparaitre beaucoup d'erreurs au niveau du voisement rendant les modèles phoniques et la transcription phonétique finale moins fiables à cet égard. Il faut par conséquent de bons modèles phoniques, dont on peut par ailleurs évaluer les limites.

Le second point important est évidemment la qualité du lexique. Le problème avec cette procédure, c'est qu'il est impossible de trouver des variantes de prononciation qui n'ont pas été introduites au préalable. Par exemple, si on n'a pas envisagé que les Français prononcent [miz] au lieu de [ministʁ] (*ministre*), alors la transcription phonétique finale ne montrera pas les occurrences de cette prononciation. En somme, on est vraiment limité par ce qu'on entre soi-même dans le lexique. Il parait alors judicieux de mettre le plus de variantes possible, et c'est ce que nous avons fait. Mais il faut être prudent et ne pas entrer absolument tout ce qui nous passe par la tête non plus, parce que plus le dispositif de reconnaissance vocale a de choix, plus il risque de se tromper. Par exemple, si l'on rentre dans le lexique une prononciation qui ne se produit jamais, le programme va quand même la considérer comme une option valable et, de temps en temps, choisir cette option au lieu de la prononciation réelle. On voit donc toute la difficulté de la tâche. Mais au final, si l'on dispose d'une bonne reconnaissance vocale, je suis convaincue qu'il est possible d'obtenir une très bonne transcription, suffisante en tout cas pour travailler.

Nous avons comparé les transcriptions effectuées automatiquement par le dispositif de reconnaissance vocale avec celles des humains. En fait, il ne faudrait pas qu'il y ait 100% de correspondance, puisque, comme on le sait, les

humains font des erreurs. Cependant une certaine correspondance s'avère évidemment souhaitable. Dans une comparaison où nous avons regardé toutes les séquences, nous avons trouvé des désaccords avec les transcripteurs humains dans environ 14% des cas (SCHUPPLER, ERNESTUS, SCHARENBORG & BOVES 2011) ; ce qui parait acceptable. Dans une autre étude, nous avons étudié la présence ou l'absence de schwa (HANIQUE *et al.* 2013), qui, comme je l'ai évoqué plus haut, est une voyelle particulièrement difficile à transcrire. Nos résultats montrent que le dispositif de reconnaissance vocale est en désaccord dans 26% et 23% des cas, suivant le transcripteur. Les transcripteurs humains eux-mêmes étaient en désaccord dans 18% des cas. La différence entre ces trois résultats n'est pas significative. Donc il semble que le transcripteur automatique se comporte comme un humain.

On s'est ensuite interrogé sur la portée de ces constats en ce qui concerne la durée des séquences. Les durées trouvées par le transcripteur automatique sont-elles comparables à celles établies par les humains ? Nous pensons que oui. Dans au moins 95% des cas, nous avons une différence inférieure à vingt millisecondes, un résultat assez similaire à ce qui est décrit dans la littérature sur les différences entre transcripteurs humains et automatiques.

Comme il est très difficile de faire travailler pendant de nombreuses heures des transcripteurs humains sur de l'oral spontané, il nous reste cette option et je pense que la qualité est suffisante s'il s'agit d'effectuer des recherches quantitatives et d'étudier, par exemple, la fréquence à laquelle un segment de mot disparait. Toutefois pour des recherches phonétiques plus précises, les transcriptions humaines s'avèrent nécessaires, mais il faut alors limiter la quantité de données.

Pour clore cette partie sur les transcriptions humaine et automatique, j'aimerais, en rapport avec l'exposé de Steven

GILLIS (voir page 95), faire une remarque sur la fusion des corpus. Steven GILLIS suggère qu'en réunissant nos corpus – pas tous bien sûr –, on bénéficierait d'un potentiel accru, avec des données livrant davantage d'informations. Il faut toutefois rester prudent à cet égard. Des différences existent entre les corpus parce que les situations de production ne sont pas exactement identiques même s'ils ont été enregistrés et transcrits de manière similaire.

Le Corpus Oral du Néerlandais propose divers contextes de production : conversations spontanées en face-à-face, entretiens avec des enseignants de néerlandais, conversations téléphoniques, etc., bref, toutes sortes de contextes qui peuvent paraitre similaires. Dans une de nos études, nous avons choisi dix mots néerlandais ayant une forte tendance à se réduire, et nous avons cherché à établir à quelle fréquence, en fonction des situations de production, ils se voyaient réduits de manière extrême, c'est-à-dire d'au moins une, voire deux syllabes. Comme l'illustre la figure 3, le pourcentage d'occurrences de mots fortement réduits varie de 5% à 40%.

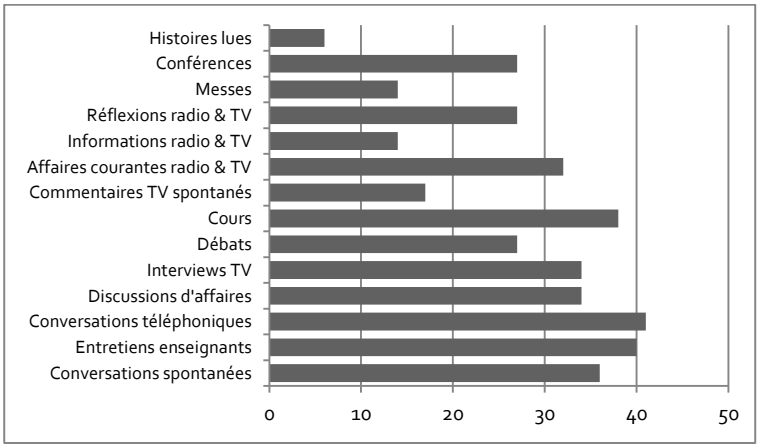

Figure 3 – Variation d'apparition de dix mots néerlandais très fréquents et extrêmement réduits selon la situation de production (ERNESTUS *et al.* 2015)

Les histoires lues aux aveugles font apparaitre un taux de réduction de 5,5%, ce qui parait conforme à l'intuition. Dans le cas des entretiens avec les enseignants de néerlandais et des conversations téléphoniques, ce taux se situe aux alentours de 40%. On pourrait alors dire que ce sont des registres de langue très différents, expliquant ainsi la variation. Cependant on observe également des différences parmi les conversations spontanées. Le pourcentage est plus élevé dans le cas des entretiens avec les enseignants (40%) que pour les conversations spontanées (36%). J'ignore la raison pour laquelle un tel écart existe. Quoi qu'il en soit, même de légères différences entre situations d'enregistrement peuvent déboucher sur des résultats différents. En somme, il faut être prudent si on compare ou rassemble des corpus, car il suffit d'une simple différence de situation pour faire varier les résultats. Si on veut comparer plusieurs langues, le choix de corpus similaires s'impose.

Pour ce faire, nous avons créé une série de trois corpus d'oral familier, enregistrés de la même manière : un en français (TORREIRA *et al.* 2010), un en espagnol (TORREIRA & ERNESTUS 2010) et un en tchèque (ERNESTUS *et al.* 2014). Dès lors, quand on utilise les données de ces corpus, on peut vraiment comparer les langues, comme par exemple le débit de parole. Et comme on sait qu'il existe des différences entre les corpus, il est préférable d'en utiliser beaucoup afin d'éviter que les pourcentages calculés soient uniquement basés sur la nature du corpus retenu. En somme, d'un côté, il faut être prudent et n'utiliser que des corpus similaires, et de l'autre, ces résultats montrent qu'on doit utiliser différents types de corpus.

J'ai consacré la première partie de mon exposé à insister sur la manière dont les mots sont réellement prononcés, ce qui constitue une question plutôt linguistique. Maintenant, j'aimerais passer à des considérations plus psycholinguistiques en essayant de répondre à la question de

savoir comment les locuteurs réduisent leur discours oral, ce qui conduira également à interroger les raisons pour lesquelles les locuteurs réduisent certains segments. Il s'agira ensuite de déterminer si l'absence d'un segment est catégorielle ou graduelle.

5. L'apport des corpus oraux en psycholinguistique

Dans notre recherche, nous avons utilisé divers corpus pour voir quand les réductions apparaissent (ERNESTUS *et al.* 2014 ; TORREIRA *et al.* 2010 ; TORREIRA & ERNESTUS 2010). Nous avons trouvé, sans surprise, que les réductions sont plus fréquentes quand le débit de parole est plus rapide, quand le mot se trouve au milieu d'un syntagme et qu'il n'est pas accentué (HANIQUE *et al.* 2013 ; PLUYMAEKERS *et al.* 2005a ; b). Dans les trois cas, les mots sont prononcés à une vitesse relativement élevée. Mais le débit n'est pas le seul facteur déterminant. La nature du mot suivant est un autre bon prédicteur. Il y a plus de réduction si le locuteur n'a pas de problème à articuler le mot suivant. Si un mot est suivi d'une hésitation, il n'y aura pas de réduction, alors qu'il y en aura plus si le mot est suivi d'un mot fortement probable. Par exemple, pour les phrases « children swim » (*Les enfants nagent*) et « children smoke » (*Les enfants fument*), le mot « children » tend à être plus réduit dans la première que dans la seconde, simplement parce que « children smoke » est une phrase qu'on ne produit pas souvent ; elle est peu probable, ce qui la rendrait plus difficile à articuler pour les locuteurs. En somme, le locuteur réduirait quand il n'est pas obligé de ralentir, et qu'il n'a pas besoin de temps pour préparer le mot suivant. En définitive, il semblerait que les mots réduits constituent la norme, les mots non-réduits survenant uniquement lorsque le locuteur rencontre un problème et doit gagner du temps. Que nous apprennent ces résultats sur les raisons de la réduction ? D'après les données, bien qu'il s'agisse là de spéculation, nous pensons que les locuteurs

réduisent afin de diminuer leur effort articulatoire. Quant à la question de savoir à quel point le locuteur prend en compte son interlocuteur, il nous est impossible d'y répondre sur la base de nos études sur corpus.

Si nous savons désormais *quand* les mots sont réduits, comment avons-nous déterminé les facteurs susceptibles d'influencer le degré de réduction ? Nos corpus sont constitués d'oral spontané. Aussi, conformément à notre volonté de recueillir de l'oral aussi spontané que possible, aucune indication n'était fournie quant au sujet de conversation durant les séances d'enregistrement. On ne savait jamais ce qu'on allait obtenir, et la variété des sujets de conversation s'avère très importante. Dès lors, les mots recueillis dans le corpus se révèlent très différents les uns des autres. Ils sont d'ailleurs produits par des locuteurs eux aussi différents. Chaque mot présente sa propre identité. Certains mots sont en première position dans la phrase, d'autres à la fin et d'autres encore au milieu. Les uns sont accentués, les autres non. Leur fréquence peut varier ainsi que la fréquence du mot qui précède et du mot qui suit. Ils diffèrent donc sur de nombreux points. Il n'en reste pas moins que nous aimerions identifier les facteurs responsables de leur réduction, et pour y parvenir, il faut accumuler beaucoup de données. Il ne s'agit pas de recueillir seulement cent ou deux cents occurrences, mais plusieurs milliers. C'est la raison pour laquelle un transcripteur automatique s'avère indispensable. À défaut, la recherche devient irréalisable.

6. L'utilisation des corpus oraux et le modèle statistique

Maintenant évoquons brièvement la dimension statistique, et plus particulièrement les modèles linéaires à effets mixtes. On utilise un modèle de régression, et on prédit la réduction ou non d'un mot ou bien la durée du segment, et on entre chaque prédicteur pour voir s'il améliore les résultats. À titre d'illustration, imaginons que

l'effet du débit de parole nous intéresse. On a alors le « débit de parole » comme prédicteur. On doit cependant gérer beaucoup d'autres variables, parce que si on n'entre que le prédicteur de débit de parole, on ne va probablement rien trouver. Il faut en effet prendre en compte tous les autres facteurs susceptibles de jouer un rôle. On ajoute donc la « fréquence du mot », sa « position prosodique », etc. Toutes ces variables de contrôle sont indispensables, sinon on n'obtient aucun résultat, et il faut par ailleurs recourir à des outils statistiques complexes. Notre expérience avec ce modèle n'est pas récente, et nous avons essayé beaucoup d'approches différentes. Dans la littérature scientifique, parait tous les six mois un nouvel article sur la bonne manière d'utiliser ce modèle, mais les opinions continuent de diverger à cet égard. Nous avons étudié ces propositions, et nous sommes désormais assez conscients des écueils. Deux points fondamentaux ressortent.

Premièrement, il faut être attentif aux prédicteurs hautement corrélés. Un article de WURM & FISICARO (2014) évoquait la possibilité de rentrer pratiquement tous les prédicteurs, même s'ils sont corrélés, mais pas s'ils sont *hautement* corrélés. Par exemple, si on s'intéresse à l'impact de la fréquence d'occurrences d'un mot, et qu'on rentre parallèlement comme prédicteur le nombre de phonèmes – ce qui représente un exemple de deux prédicteurs hautement corrélés –, le modèle ne peut pas vraiment déterminer lequel des deux a le plus d'impact. Il est donc impossible de rentrer plusieurs variables hautement corrélées. On pourrait imaginer que pour résoudre le problème, il suffirait d'éliminer par exemple le prédicteur du 'nombre de phonèmes' et de conserver celui de la 'fréquence d'occurrence'. Toutefois, si une telle solution nous permet d'observer un impact du prédicteur retenu, on ne peut pas garantir qu'il résulte uniquement de la fréquence d'occurrence et non pas de

l'*absence* du prédicteur 'nombre de phonèmes', préalablement écarté.

Deuxièmement, il ne faut pas entrer tous les prédicteurs auxquels on pourrait à priori penser. En effet, il est parfois tentant d'essayer d'en intégrer le maximum – sans justification théorique – juste « pour voir » et observer les interactions possibles. Cependant, une telle stratégie est insatisfaisante. On obtient parfois un modèle à priori très performant, qui prédit bien les données, mais dès qu'on le teste sur un autre ensemble de données les prédicteurs spécifiques ne fonctionnent plus. En somme, on se retrouve avec différents ensembles de prédicteurs pour différents ensembles de données, les premiers résultats obtenus se révélant non-généralisables à de nouveaux ensembles de données. Évidemment, ce n'est pas ce qu'on veut. Voilà pourquoi j'insiste : il faut faire attention au nombre de prédicteurs qu'on utilise et aux interactions. Certains soutiennent que pour un prédicteur, il faut au moins trente points de données. Je pense que c'est insuffisant et qu'il en faut beaucoup plus[3].

7. L'apport des corpus oraux dans la détermination de la nature de la réduction des mots

Abordons désormais ma dernière question – typiquement psycholinguistique – sur la nature de la réduction des mots : est-elle graduelle ou catégorique ? Dans le premier cas, on pourrait alors la considérer comme le résultat d'un processus phonétique ou articulatoire. Imaginons un mot comportant un schwa au départ dont on réduirait de plus en plus la durée. Ce schwa finira par complètement disparaitre, et uniquement en raison de cette réduction progressive, soit un processus phonétique ou articulatoire, qu'on peut qualifier de

[3] Il existe d'autres manières de résoudre ce problème que je n'évoquerai pas ici.

« graduel ». À l'inverse, il se pourrait que le schwa soit absent en raison d'une règle phonologique, qui, par exemple dans un mot comme *fenêtre*, impliquerait la chute du /ə/ et aboutirait à /fnɛtʁ/. On pourrait également envisager que le locuteur possède deux variantes de prononciation dans son lexique mental – /fənɛtʁ/ et /fnɛtʁ/ – entre lesquelles il peut choisir. Il est en fait impossible d'établir une distinction entre règle phonologique et variantes intégrées dans le lexique mental, mais une différence claire existe entre réduction graduelle et catégorique. La question est alors de savoir si les corpus peuvent nous apporter des éléments de réponse.

On peut bien entendu étudier la réduction en laboratoire en faisant des expérimentations, mais on n'aura plus affaire à de l'oral spontané. Il serait donc intéressant de tirer parti des corpus oraux pour obtenir des informations supplémentaires. À cet effet, nous avons étudié à quel moment surviennent les réductions, en partant de l'hypothèse que si, dans les mêmes conditions, un segment (par exemple un schwa, un /t/) est absent et plus réduit, l'absence et la réduction sont soumises aux mêmes prédicteurs. En d'autres termes, celles-ci résultent d'un processus identique, en l'occurrence nécessairement phonétique, puisque la réduction est un processus phonétique. Nous constatons que c'est effectivement le cas pour la chute du /t/ en position finale en néerlandais (HANIQUE *et al.* 2013).

PRÉDICTEUR DE PRÉSENCE/ABSENCE	PRÉDICTEUR DE DURÉE
débit	débit
registre	registre
segment précédent	segment précédent
	dernier mot du segment
	dernier mot du segment & débit

Tableau 2 – Prédicteurs de l'absence du /t/ final ou de sa réduction en néerlandais

Comme l'indique la colonne de gauche du tableau 2, le /t/ est plus souvent absent à débit rapide, dans un registre de langue plus familier, et lorsque le segment précédent est un

/s/ ou un /m/ plutôt qu'une autre consonne. La durée du /t/ obéit également à ces mêmes prédicteurs. En d'autres termes, les prédicteurs pertinents pour l'absence du /t/ le sont aussi pour sa durée, ce qui laisse à penser que l'absence et la durée du /t/ sont déterminées par le même processus, qui doit donc être *graduel*, *phonétique*, *articulatoire*. Toutefois, la colonne de droite fait apparaitre deux autres prédicteurs, en l'occurrence la position du mot en fin de segment, ainsi qu'une interaction avec le débit de parole (signalée par '&'). On pourrait à priori en conclure que les prédicteurs de l'absence de /t/ diffèrent de ceux de sa durée. Certes, mais ces ensembles de données montrent malgré tout que les mêmes prédicteurs ont une influence et sur l'absence et sur la durée du /t/, et ce en raison du fait que dans le cas de la durée, on a juste une puissance statistique plus grande. Dès lors, il est bien plus probable de trouver des prédicteurs additionnels si la durée constitue la variable dépendante. En définitive ceci prouverait que l'absence du /t/ final en néerlandais résulte d'un processus phonétique.

L'autre possibilité concerne les cas dans lesquels les prédicteurs déterminant l'absence ou la durée d'un segment sont très différents. Ainsi en est-il du schwa en position médiane en français (par exemple [fənɛtʁ], *fenêtre*), comme l'illustre le tableau 3 (BÜRKI *et al.* 2011).

PRÉDICTEUR DE PRÉSENCE/ABSENCE	PRÉDICTEUR DE DURÉE
débit	débit
position du schwa dans le mot	position du schwa dans le mot
position du mot dans la phrase	voisement de la consonne suivante
nombre de consonnes de la séquence	voisement de la consonne précédente
respect du principe de sonorité[4]	probabilité d'apparition du mot avec le mot précédent

Tableau 3 – Prédicteurs de l'absence du schwa médian ou de sa réduction en français

[4] NdE : Correspondant à l'échelle : « occlusive sourde < occlusive sonore < fricative sourde < fricative sonore < nasale < liquide » (BÜRKI *et al.* 2011 : 3983).

À débit rapide, le schwa disparait plus souvent. Déterminante s'avère également la position du mot dans l'énoncé, et, dans le cas où la disparition du schwa donne lieu à un groupe consonantique, il faut regarder combien de consonnes le composent et si le principe de sonorité y est respecté. Pour la durée du schwa, on observe que le débit et la position du segment dans le mot sont importants. Trois nouveaux prédicteurs interviennent également : le voisement de la consonne suivante, celui de la consonne précédente, et la probabilité d'apparition du mot avec le mot précédent. Nous sommes donc en présence de deux ensembles de prédicteurs différents, ce qui prouverait qu'en français, l'absence du schwa peut résulter d'un processus catégorique, comme une règle phonologique ou encore d'un stockage du mot sans schwa dans le lexique mental. Ces résultats confirment les données en production que nous avons collectées en laboratoire (BÜRKI *et al.* 2010).

À ce stade, il faudrait encore répondre à la question de savoir à partir de quel moment on peut dire que deux ensembles de prédicteurs diffèrent. Au cours de cet exposé, j'ai montré deux exemples, et il me semble que mes arguments suffisent pour affirmer que, dans un cas, il y avait un ensemble commun de prédicteurs, et dans l'autre, deux ensembles distincts. On peut cependant imaginer qu'il est parfois impossible de trancher, d'autant plus que la méthodologie n'est pas toujours évidente.

Pour conclure, j'insisterai sur le fait que les corpus de langue, et plus particulièrement les corpus d'oral spontané ont une réelle importance pour notre recherche. Ils nous révèlent en effet les processus de production de la parole en conversation naturelle, et la véritable prononciation des mots. J'ai montré qu'ils pouvaient s'avérer utiles tant pour des questions linguistiques, que psycholinguistiques, même si de nombreuses interrogations subsistent, auxquelles les corpus ne permettent pas de répondre.

8. Conclusion

Dans cet exposé, je n'ai pas du tout abordé la question de la compréhension, pour laquelle, on peut en partie recourir aux corpus. Par ailleurs, de nombreux points précis de psycholinguistique restent difficiles à traiter à l'aide d'un tel outil. Aussi parait-il souhaitable, dans la mesure du possible, de recourir de manière complémentaire aux corpus et aux expérimentations en laboratoire pour voir si l'on obtient des résultats convergents.

Toutefois, comme je l'ai montré, il faut garder à l'esprit que l'utilisation des corpus n'est pas exempte de problèmes. Un grand nombre de conversations s'avèrent nécessaires, si possible transcrites automatiquement, entrainant dès lors un traitement statistique complexe, parsemé de nombreux écueils, dont il faut savoir se préserver. De plus, la comparaison des corpus est souvent malaisée, en raison de la différence des niveaux de langue. Enfin, il peut s'avérer difficile de déterminer si deux ensembles de prédicteurs sont identiques ou pas.

En définitive, la recherche psycholinguistique ne peut pas se passer des corpus d'oral spontané, car sans eux, on aurait seulement accès à ce que les locuteurs produisent lors d'expérimentations en laboratoire, ce qui manifestement ne correspond pas à ce qu'ils font réellement dans leur quotidien.

Questions

Légende : « Q » pour « Question », « ME » pour « Mirjam ERNESTUS »

Q : Vous n'avez pas parlé des dysfluences, alors que je suppose que c'est un problème constant quand on utilise des corpus oraux. Comment les gérez-vous, par exemple, quand vous effectuez une transcription automatique ?

ME : Avant d'effectuer la transcription automatique, on filtre en écartant tous les éléments perturbateurs, comme les rires et les silences. En fait, on ne donne pas l'entièreté du signal au

transcripteur automatique. Au lieu d'un seul bloc d'une heure et demie, on découpe nos fichiers en segments d'environ deux secondes chacun, ce qui nous permet d'isoler les dysfluences. Donc il n'y en a pas vraiment, seulement dans les parties non-filtrées de la transcription.

Q : Mais… ici, une question méthodologique se pose : vous ignorez les dysfluences, alors que certaines personnes les considèrent comme une part essentielle du langage.

ME : Absolument. Mais on a aussi montré – pas seulement nous – que les dysfluences ont une influence sur le degré des réductions. Donc on ne les ignore pas ; ce n'est que pour la transcription automatique qu'on les filtre, parce qu'il est impossible de les transcrire. Sinon, le lexique des prononciations contiendrait un nombre gigantesque de transcriptions pour chaque mot. C'est pourquoi on les ignore dans nos transcriptions automatiques. Mais de toute évidence, elles sont très importantes pour la recherche.

Q : Puisqu'on parle de dysfluences, que se passe-t-il quand on hésite ? En français par exemple, si on dit « euh » avant « fenêtre » : est-ce que vous savez si les gens ont tendance à prononcer plus souvent le schwa dans cette condition ?

ME : À ma connaissance, ça n'a jamais été démontré. Mais le contraire, oui : les locuteurs prononceront « fenêtre » en entier si une dysfluence vient juste après, mais pas dans l'autre sens.

Q : Quand vous parliez de transcription humaine, dans un sens, c'est une manière de « discrétiser » un phénomène continu. A-t-on des moyens pour coder cette continuité ? Parce que, en cas de conflit entre transcripteurs, on peut penser que le phénomène en soi n'est en réalité pas catégorique. Est-ce qu'on arrive à coder, à rendre compte du flou des données ?

ME : On y a pensé, mais on ne l'a pas fait. Une des choses qu'on a songé à faire est qu'au lieu d'avoir des prédicteurs binaires en 0 ou 1 – sourd ou sonore, par exemple – on pourrait utiliser des valeurs intermédiaires, comme ⅔ ou ⅓ pour refléter le désaccord des transcripteurs.

Q : Avez-vous une idée de comment on pourrait le visualiser ? Parce qu'on aimerait pouvoir lire ces transcriptions.

ME : Je n'ai pas de solution qui me vienne à l'esprit. C'est un problème qui nous a préoccupés, mais on n'a pas encore essayé de trouver une solution. Normalement, on fait toutes les mesures acoustiques, car je pense qu'elles représentent ce continuum. Après, bien sûr, il y a le problème de la multiplicité des propriétés acoustiques d'un phénomène à prendre en compte, particulièrement si vous vous intéressez au voisement, par exemple. Mais les mesures acoustiques rendent compte du continuum. Voilà ce qu'on a fait pour le moment.

Q : Une idée pourrait être de superposer à la transcription une sorte d'indication visuelle de la confiance que vous accordez à la transcription, ou au désaccord des transcripteurs.

Q : Votre exemple des phrases « children swim » et « children smoke » me rappellent un des travaux de PIANTADOSI[5], qui parlait d'efficacité dans la communication. Sa théorie était qu'on investit plus d'effort dans l'articulation lorsqu'on s'attend à ce que l'auditeur ait des difficultés à comprendre. Il a une théorie complète et un modèle statistique là-dessus.

ME : Oui, c'est une grande question en psycholinguistique : à quel point est-ce que vous ajustez votre production verbale en fonction de l'auditeur, à quel point est-ce que vous adaptez votre discours par rapport à lui ? Il y a des gens qui pensent que ces réductions résulteraient de la prise en compte de l'auditeur dans la mesure où, en tant que locuteur, vous utiliseriez ce qui est préférable pour votre auditeur, et que vous réduiriez, parce que vous savez que l'auditeur a entendu ce mot il y a peu de temps, et qu'il n'est pas important, alors que ne pas le réduire reviendrait au contraire à signaler à votre auditeur que ce mot est important. À vrai dire, je ne crois pas du tout à cette théorie, et ce pour deux raisons : premièrement, je ne peux tout simplement pas croire qu'en tant que locuteur, on soit capable de savoir ce qui est préférable pour l'auditeur. Deuxièmement, de nombreuses études montrent que ça ne fonctionne pas comme ça. Il y a une étude – je suis d'ailleurs jalouse de son auteur, j'aurais vraiment adoré la faire moi-même – faite par Ellen BARD de l'Université d'Édimbourg (BARD *et al.* 2000). On sait qu'en ce qui concerne la réduction, plus on

[5] NdE : voir PIANTADOSI *et al.* (2011).

répète un mot dans une conversation, et plus il est réduit. On pourrait dire que vous signalez à votre auditeur qu'il a déjà entendu ce mot avant. BARD a conçu une expérience dans laquelle deux personnes avaient devant elles deux cartes semblables mais pas identiques : dans l'une, un itinéraire était indiqué, et le participant qui l'avait devant les yeux devait l'expliquer à l'autre personne. Certains éléments revenaient périodiquement dans la conversation, comme par exemple le mot « cathédrale », dans des phrases comme « Tourne à droite après la cathédrale ! » auxquelles l'interlocuteur répondrait par « Je ne vois pas de cathédrale », etc. Donc vous imaginez bien que ces mots sont de plus en plus réduits, au fur et à mesure qu'ils sont employés. Ensuite, les chercheurs ont gardé le « locuteur-guide » mais lui ont présenté un autre auditeur, en donnant à chacun deux nouvelles cartes, la grande question étant « Qu'est-ce qui va se passer pour le mot "cathédrale" ? ». Le locuteur-guide va-t-il continuer à réduire ou au contraire repartir à la case zéro et réduire à partir de là ? On constate que le locuteur-guide continue tout simplement à réduire, sans se préoccuper du changement d'auditeur, ce qui indique clairement qu'on ne tient pas compte à ce point de l'auditeur.

Q : Comment était organisée l'expérience ? Ils avaient un retour ?

ME : Oui, tout le temps.

Q : Un retour visuel ?

ME : Non, mais ils pouvaient se parler.

Q : Mais pas de retour visuel.

ME : Il y avait plusieurs situations différentes pour cette expérience : parfois, il y avait un contact visuel, parfois pas.

Q : Parce que, quand vous dites qu'on ne s'adapte pas à ce niveau de précision, je suis tout à fait d'accord. Mais on s'adapte quand même plus précisément lorsqu'on a un retour visuel.

ME : Certes, mais ça ne signifie pas que si vous réduisez, le mot va devenir incompréhensible. Ça ne signifie pas que si vous réduisez plus qu'on ne s'y attend, l'auditeur va tout de suite signaler qu'il ne comprend pas le mot.

Q : J'ai une question au sujet de la reconnaissance vocale. Je suppose que la réponse est la même que pour les dysfluences, mais que faites-vous du bruit, des chevauchements dans une conversation ?

ME : Pour les chevauchements, c'est la même réponse que dans le cas des dysfluences : on découpe le signal en segments d'environ deux secondes et on vérifie que les limites entre ces segments se situent à des endroits logiques, de préférence entre des silences, de façon à perdre le moins possible de données. Mais on laisse les chevauchements de côté, parce que c'est impossible à gérer pour la reconnaissance vocale. Si on a trop de bruit, c'est pareil. Là encore, ça veut dire qu'on n'a pas de transcription automatique de ces segments, mais on s'y intéresse quand même. Évidemment que ce qui se passe pendant les chevauchements nous intéresse, mais on ne peut pas traiter cette question d'après des transcriptions automatiques.

Références

ADDA-DECKER Martine, BOULA DE MAREÜIL Philippe, ADDA Gilles & LAMEL Lori. (2005). Investigating Syllabic Structures and their Variation in Spontaneous French. *Speech Communication 46-2*, 119-139.

BARD Ellen Gurman, ANDERSON Anne H., SOTILLO Catherine, AYLETT Matthew, DOHERTY-SNEDDON Gwyneth & NEWLANDS Alison. (2000). Controlling the Intelligibility of Referring Expressions in Dialogue. *Journal of Memory and Language 42-1*, 1-22.

BOOIJ Geert. (1999). *The phonology of Dutch*. Clarendon: Oxford University Press.

BRAND Sophie & ERNESTUS Mirjam. (2015). Reduction of Obstruent-Liquid-Schwa Clusters in Casual French. *Proceeding of the 18th International Congress of Phonetic Sciences*, ICPHS0251.1-5.

BÜRKI Audrey, ERNESTUS Mirjam & FRAUENFELDER Ulrich H. (2010). Is There Only one "Fenêtre" in the Production Lexicon? On-Line Evidence on the Nature of Phonological Representations of

Pronunciation Variants for French Schwa Words. *Journal of Memory and Language 62-4*, 421-437.

BÜRKI Audrey, ERNESTUS Mirjam, GENDROT Cédric, FOUGERON Cécile & FRAUENFELDER Ulrich Hans. (2011). What Affects the Presence versus Absence of Schwa and its Duration: A Corpus Analysis of French Connected Speech. *Journal of the Acoustical Society of America 130-6*, 3980-3991.

ERNESTUS Mirjam. (2000). *Voice Assimilation and Segment Reduction in Casual Dutch*. Utrecht: LOT (Landelijke Onderzoekschool Taalwetenschap).

ERNESTUS Mirjam, LAHEY Mybeth, VERHEES Femke & BAAYEN R Harald. (2006). Lexical Frequency and Voice Assimilation. *The Journal of the Acoustical Society of America 120-2*, 1040-1051.

ERNESTUS Mirjam & WARNER Natasha. (2011). An Introduction to Reduced Pronunciation Variants. *Journal of phonetics 39-3*, 253-260.

ERNESTUS Mirjam. (2014). Acoustic Reduction and the Roles of Abstractions and Exemplars in Speech Processing. *Lingua 142*, 27-41.

ERNESTUS Mirjam, KOČKOVÁ-AMORTOVÁ Lucie & POLLAK Petr. (2014). The Nijmegen Corpus of Casual Czech. *LREC 2014: 9th International Conference on Language Resources and Evaluation*, 365-370.

ERNESTUS Mirjam, HANIQUE Iris & VERBOOM Erik. (2015). The Effect of Speech Situation on the Occurrence of Reduced Word Pronunciation Variants. *Journal of phonetics 48*, 60-75.

HANIQUE Iris, ERNESTUS Mirjam & SCHUPPLER Barbara. (2013). Informal Speech Processes can be Categorical in Nature, even if they Affect many Different Words. *The Journal of the Acoustical Society of America 133-3*, 1644-1655.

JOHNSON Keith. (2004). Massive Reduction in Conversational American English. In YONEYAMA Kiyoko & MAEKAWA Kikuo (Eds), *Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium*, Tokyo: The National International Institute for Japanese Language, 29-54.

OOSTDIJK Nelleke. (2002). The Design of the Spoken Dutch Corpus. In SMITH Adam, COLLINS Peter & PETERS Pam (Eds), *New Frontiers of Corpus Research*, Amsterdam: Rodopi, 105-112.

PIANTADOSI Steven T., TILY Harry & GIBSON Edward. (2011). Word Lengths are Optimized for Efficient Communication. *Proceedings of the National Academy of Sciences 108-9*, 3526-3529.

PLUYMAEKERS Mark, ERNESTUS Mirjam & BAAYEN R. Harald. (2005a). Articulatory Planning is Continuous and Sensitive to Informational Redundancy. *Phonetica 62-2/4*, 146-159.

PLUYMAEKERS Mark, ERNESTUS Mirjam & BAAYEN R. Harald. (2005b). Lexical Frequency and Acoustic Reduction in Spoken Dutch. *The Journal of the Acoustical Society of America 118-4*, 2561-2569.

SCHUPPLER Barbara, ERNESTUS Mirjam, SCHARENBORG Odette & BOVES Lou. (2011). Acoustic Reduction in Conversational Dutch: A Quantitative Analysis Based on automatically Generated Segmental Transcriptions. *Journal of phonetics 39-1*, 96-109.

TORREIRA Francisco, ADDA-DECKER Martine & ERNESTUS Mirjam. (2010). The Nijmegen Corpus of Casual French. *Speech Communication 52-3*, 201-212.

TORREIRA Francisco & ERNESTUS Mirjam. (2010). The Nijmegen Corpus of Casual Spanish. *Seventh Conference on International Language Resources and Evaluation (LREC'10)*, 2981-2985.

WURM Lee H. & FISICARO Sebastiano A. (2014). What Residualizing Predictors in Regression Analyses does (and what it does not do). *Journal of Memory and Language 72*, 37-48.

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

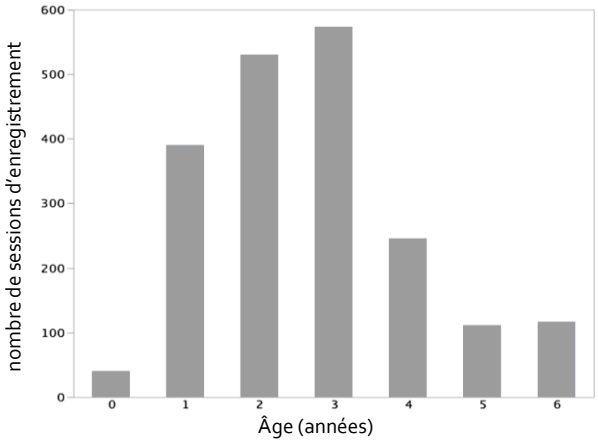

Figure 3 – Corpus français de CHILDES : nombre de sessions d'enregistrement par âge

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.

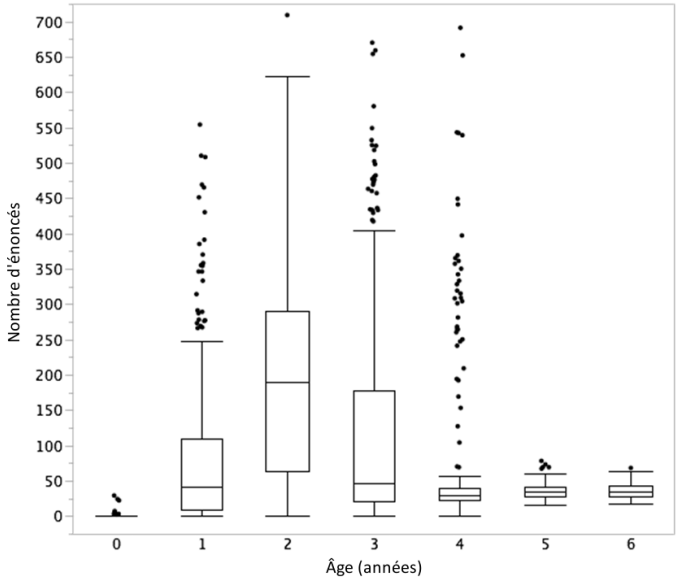

Figure 4 – Corpus français de CHILDES : distribution du nombre d'énoncés selon l'âge (langage enfantin)

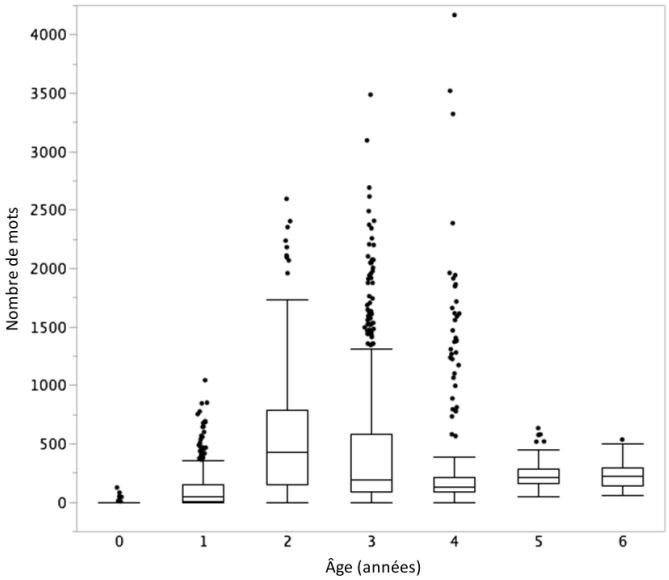

Figure 5 – Corpus français de CHILDES : distribution du nombre de mots selon l'âge (langage enfantin)

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).

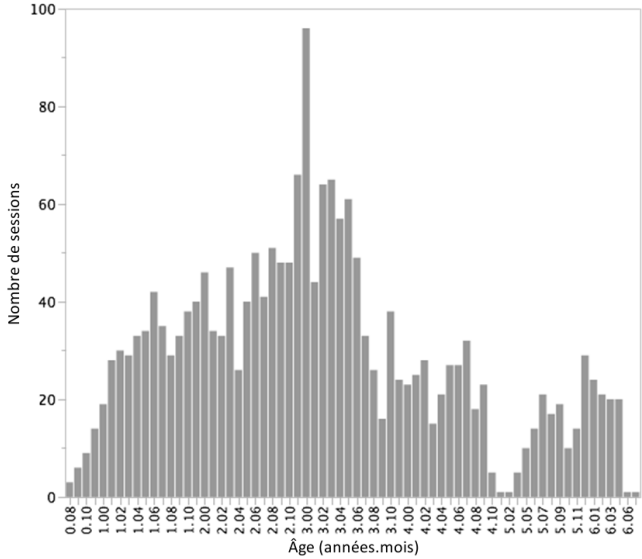

Figure 6 – Corpus français de CHILDES : nombre de sessions d'enregistrement par âge

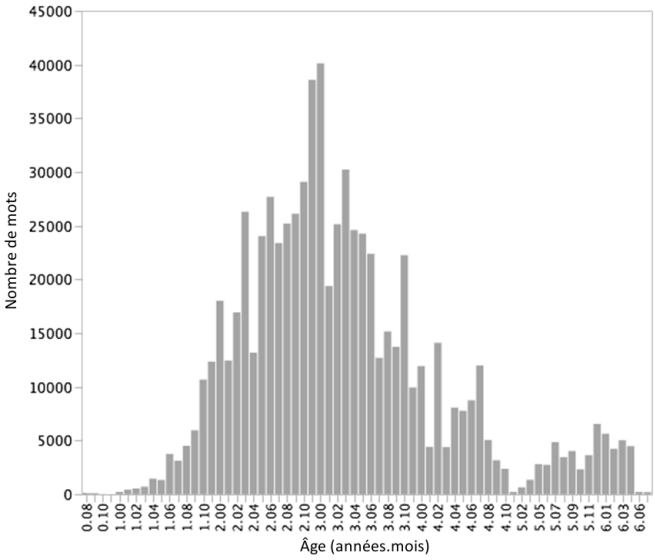

Figure 7 – Corpus français de CHILDES : nombre de mots par âge (langage enfantin)

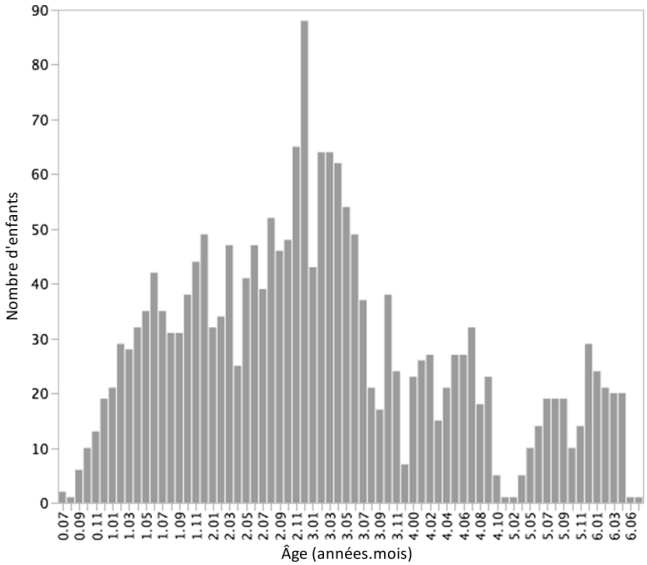

Figure 8 – Corpus français de CHILDES : nombre d'enfants par âge

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.

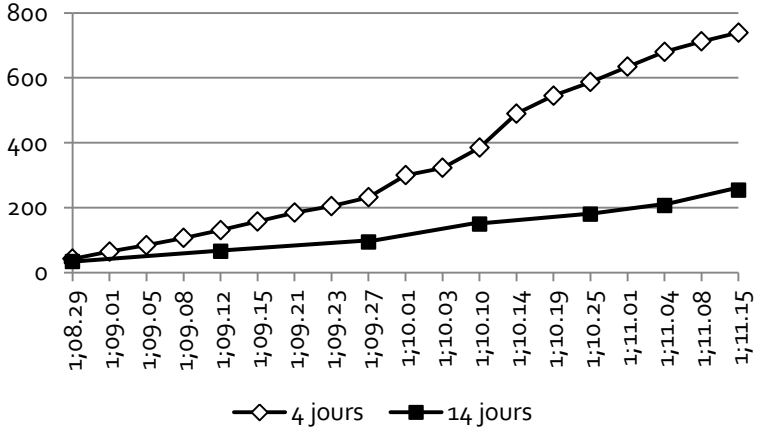

Figure 9 – Développement du lexique cumulatif selon le régime d'échantillonnage (enfant Maarten)

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).

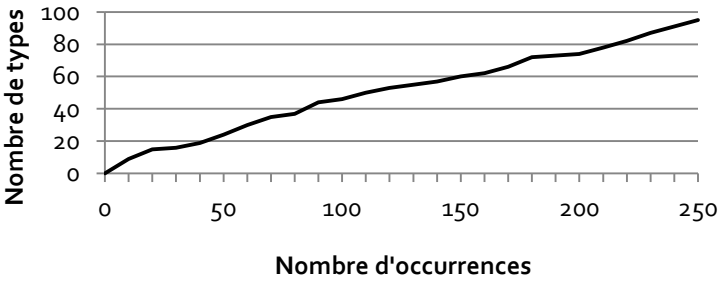

Figure 10 – Dépendance entre nombre de types et nombre d'occurrences

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).

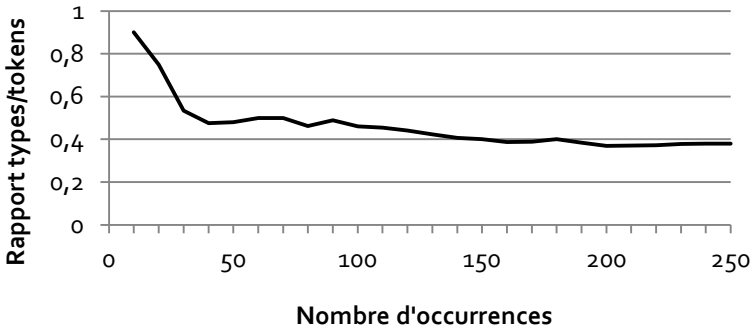

Figure 11 – Dépendance entre rapport types/tokens et nombre d'occurrences

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%.

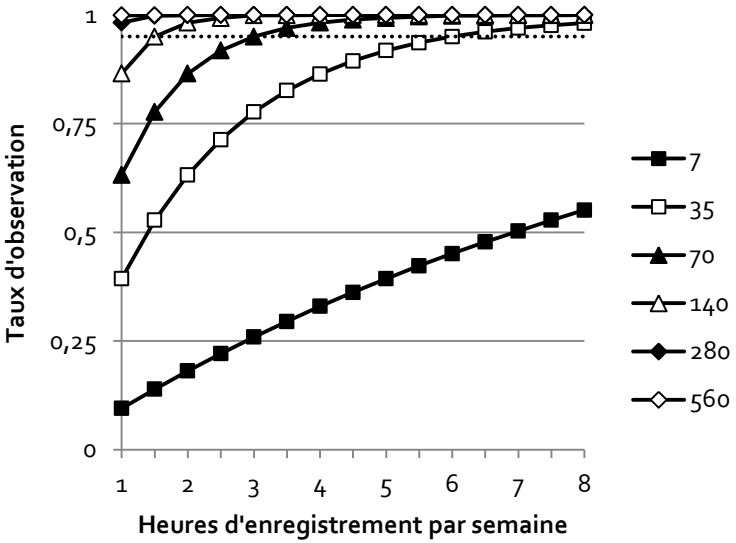

Figure 12 – Taux d'observation d'un phénomène en fonction du nombre d'heures d'enregistrement par semaine et du nombre d'occurrences attendues par heure

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).

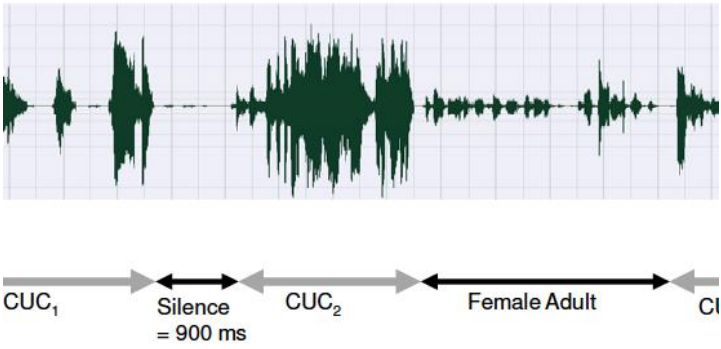

Figure 14 – Segmentation et catégorisation du signal audio par LENA Pro (OLLER *et al.* 2010, Supporting Information Appendix: p. 16)

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.

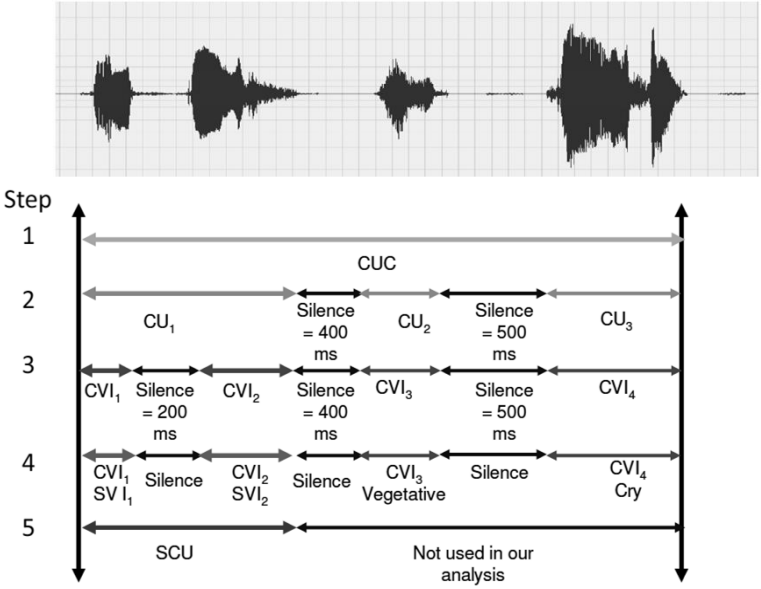

Figure 15 – Segmentation et catégorisation hiérarchique du signal audio par LENA Pro (OLLER *et al.* 2010, Supporting Information Appendix: p 18)

		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).

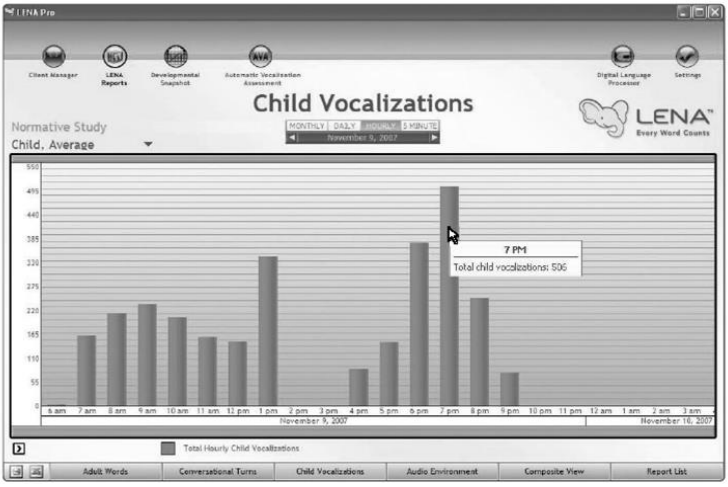

Figure 16 – Nombre moyen de vocalisations de l'enfant par heure de la journée, calculé par LENA Pro (LENA Pro Brochure 2012 : 5)

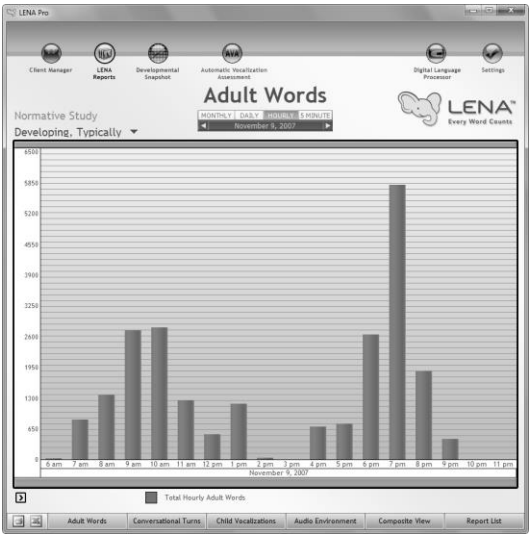

Figure 17 – Nombre moyen de mots dans les productions adultes par heure de la journée, calculé par LENA Pro (LENA Pro Brochure 2012, 3)[7]

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.

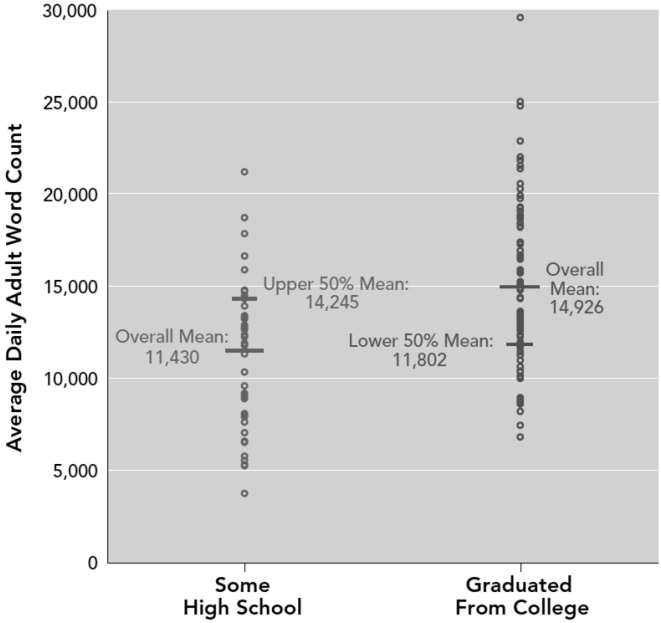

Figure 18 – Distribution du nombre moyen de mots par jour dans la production des adultes en fonction du niveau socioéconomique (GILKERSON & RICHARDS 2009 : 20)

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*.

Cahiers de l'ILSL-45 (2016 : 123-160)

La question des données en morphologie

Nabil HATHOUT
Université de Toulouse & CNRS (F)
nabil.hathout@univ-tlse2.fr

1. Introduction

Je présente dans cet article un ensemble de réflexions qui ont été menées sous l'impulsion de Marc PLÉNAT, au sein du laboratoire ERSS puis de la composante CLLE/ERSS, sur les données qu'il convient d'utiliser en morphologie[1]. Cette question générale peut être déclinée selon différents points de vue relatifs à l'objet d'étude, les jeux de données, leur création, leur origine ou leur devenir :

- Quelle est la nature des données en morphologie ? Quels sont les objets auxquels s'intéresse la morphologie ?
- De quelles données a-t-on besoin pour réaliser des recherches en morphologie, pour l'analyse des phénomènes morphologiques ?
- Comment peut-on obtenir ces données ? Où est-il possible de les collecter ?
- Comment rentabiliser les efforts nécessaires à la constitution des jeux de données morphologiques[2] ?

[1] Les travaux et les réflexions présentés dans cet article ont été réalisés et menés en collaboration avec Marc PLÉNAT, Ludovic TANGUY, Fiammetta NAMER et Fabio MONTERMINI. Ils ont fait l'objet de plusieurs publications, notamment (HATHOUT, PLÉNAT & TANGUY 2003 ; HATHOUT, NAMER, PLÉNAT & TANGUY 2009 ; HATHOUT, MONTERMINI & TANGUY 2008 ; HATHOUT & NAMER, 2014a, 2014b). Mon texte reprend en grande partie les trois dernières références et la présentation de Démonette dans la quatrième. Le lecteur intéressé pourra également se reporter à PLÉNAT (2000), PLÉNAT *et al.* (2002) ou TANGUY (2012, 2013).

[2] Dans ce texte, j'utilise indifféremment les termes de « jeux de données morphologiques » et « collections d'exemples » pour désigner les données prêtes à l'emploi, directement exploitables dans le cadre de recherches en morphologie.

Les réponses que l'on peut apporter à ces questions dépendent du type d'analyse morphologique que l'on souhaite réaliser. En morphologie descriptive, les données jouent un rôle central dans les analyses, ce qui oblige les morphologues à consacrer à leur constitution une part importante de leur travail. La qualité des analyses descriptives dépend en effet directement de celle de données. En morphologie, comme dans les autres sciences, l'analyse vise à organiser les données en catégories et à identifier les propriétés qui caractérisent leurs éléments. Il faut donc que les données utilisées soient suffisamment variées pour contenir des représentants de chacune des catégories pertinentes pour le phénomène étudié. Par ailleurs, la **quantité** de ces données doit être suffisante pour permettre l'identification des régularités qui s'établissent dans ces catégories et qui interviennent dans la compréhension et l'explication du phénomène.

Les nouvelles technologies et le Web ont amélioré très significativement la quantité et la qualité des données rassemblées et utilisées par les morphologues du fait d'un accès plus facile à des ressources volumineuses et au développement d'outils permettant de les exploiter. La variété des données est également améliorée par la possibilité d'observer des productions langagières peu normées présentes dans les forums de discussion, les tweets, etc. Mais ces évolutions ont un cout qui n'est pas négligeable : les morphologues se trouvent en effet aujourd'hui dans l'obligation de prendre en compte des quantités énormes d'exemples dont le traitement est à la fois long et fastidieux ; les exemples collectés sont fortement bruités et nécessitent un nettoyage soigneux suivi d'une préparation préalable à toute utilisation (formatage et annotation). Une autre conséquence de l'augmentation de la quantité des données disponibles est un changement dans la nature même du travail des morphologues : la recherche

devient plus expérimentale ; une part du travail de plus en plus grande est consacrée à la collecte et au traitement des données ; les exemples décrits dans la littérature perdent de leur importance. Ce surcout soulève plusieurs questions :

- A-t-on réellement besoin d'autant de données ? Est-il rentable d'utiliser plus de données ? Les analyses morphologiques sont-elles significativement améliorées par la masse des exemples collectés ?
- Comment préserver les données collectées, nettoyées, préparées et analysées ? Cette question concerne notamment les formats, la complétude des jeux de données et les dépôts dans lesquels il faut les placer. Se pose en outre la question du cout de la mise au format et de la complétion des données.
- Quelles sont les utilisations possibles des données morphologiques une fois terminée l'analyse du phénomène étudié ? Une collection d'exemples peut-elle servir à d'autres études que celles pour laquelle elle a été créée ?
- Comment favoriser la constitution de jeux de données réutilisables ? Comment les diffuser ? Dans quelles conditions ? Sous quelle licence ?
- Peut-on mieux rentabiliser le temps et les efforts consacrés à la création des jeux de données ? Comment faire en sorte que la constitution d'une collection d'exemples ne soit pas seulement une sous-tâche préalable dans l'analyse d'un phénomène particulier ? Les jeux de données peuvent-ils avoir une valeur en eux-mêmes ? Comment mieux faire reconnaitre ce travail, notamment par les tutelles et les instances d'évaluation ?

Les réponses à cette dernière question passent par une meilleure considération du travail de constitution de ressources et de collections de données morphologiques. Actuellement, elles sont souvent considérées comme un

sous-produit de l'analyse morphologique sans réelle valeur. Le changement de conception passera probablement d'abord par l'usage : la mise à disposition systématique des données sur lesquelles reposent les analyses favorisera leur utilisation par d'autres chercheurs pour de nouvelles études. Un changement dans la politique éditoriale des journaux de morphologie et plus généralement de linguistique sera aussi nécessaire : il est essentiel de publier davantage d'articles consacrés à la description des jeux de données disponibles.

2. La nature des données utilisées en morphologie

Le point d'entrée pour toute recherche en morphologie – qu'elle soit morphématique ou lexématique – est le mot. Dans le premier cas, l'analyse morphologique vise à décomposer le mot en morphèmes et à organiser ces derniers dans une structure généralement arborescente comme en (1). Dans le second, l'analyse morphologie cherche à identifier les relations de forme et de sens qui s'établissent entre les mots. Ces relations décrivent notamment l'histoire dérivationnelle des lexèmes comme en (2).

(1) *décomposition* : [[dé- [composer]$_V$]$_V$ -ion]$_N$
(2) *décomposition* : décomposition$_N$ → décomposer$_V$ → composer$_V$

2.1 Les mots

Avoir les mots comme objet d'étude apporte à la morphologie un avantage considérable sur d'autres sous-disciplines de la linguistique car ces unités peuvent être facilement collectées dans les textes. Les mots sont notamment très faciles à identifier et à traiter. Leur identification repose sur une approximation de la réalisation du lexème par le mot graphique que l'on définit très simplement comme une chaine de caractères délimitée par des espaces ou des signes de ponctuation. Les mots graphiques sont par ailleurs faciles à manipuler. Une simple

substitution permet par exemple de transformer la chaine `décomposition` en `décomposer`. Un autre avantage des mots est la correspondance très régulière qui existe dans des langues comme le français entre les graphies et les formes phonémiques avec un corolaire intéressant : les mots construits par une règle de construction de lexèmes particulière peuvent être identifiés avec une grande précision à partir de leurs graphies. Par exemple, un mot qui se termine par la sous-chaine `ion` a de fortes chances d'être un nom déverbal construit par suffixation en *-ion*, comme c'est le cas de *décomposition*. Par ailleurs, la graphie d'une forme fléchie ne varie pas en fonction du contexte dans lequel elle apparait, contrairement par exemple aux structures syntaxiques. Au singulier, *décomposition* s'écrit toujours `décomposition`. À l'inverse, la construction *dire que* suivie d'une proposition peut se réaliser de façon contigüe ou non-contigüe comme en (3)

(3) a. Elle **dit qu'**il reviendra.
b. Elle ne **dit** pas **qu'**il reviendra.
c. Elle m'a **dit** à plusieurs reprises **qu'**il reviendra.

Les principales conclusions que l'on peut tirer des remarques précédentes sont d'abord que les documents écrits sont particulièrement bien adaptés à la collecte de données pour la morphologie, y compris dans des textes qui ne peuvent pas être utilisés pour la recherche en phonologie ou en syntaxe. Dans le premier cas, les limitations sont principalement dues à la rareté des enregistrements oraux retranscrits phonologiquement. Dans le second, c'est l'impossibilité d'exploiter les très nombreux textes en français peu voire pas normé, disponibles dans les forums de discussion ou sur les plateformes de microblogage comme Twitter. Si la syntaxe de ces productions est souvent difficile à analyser du fait des interférences avec l'organisation discursive de ces textes, la graphie des mots est en général suffisamment standard pour permettre de les reconnaitre et

de les interpréter comme dans l'exemple (4a) où *gentillable* est clairement construit par suffixation en *-able* sur l'adjectif *gentil* dont il semble être ici un synonyme. Le second extrait (4b) met en évidence les connotations qui lui sont associées.

(4) a. c vré vré vré vrement tres **gentillable** de ta par...merci ma cherie..serieu ta du metre trop longtemps pour faire cet article... `http ://f-ceriiise-f.skyrock.com/...`

b. elle c une fille incroyable formidable **gentillable** et inoubliable (sa existe gentillable??)... `http ://latitmarionette.skyrock.com/1049941156-Lelex.html`

Sans prendre ici position sur la qualité (ou la recevabilité) de ces exemples (voir aussi section 3.2), il est indéniable que les extraits en (4), difficilement exploitables par la syntaxe et par la phonologie, sont susceptibles d'intéresser un morphologue qui envisage de réaliser une analyse sur les adjectifs en *-able*. Cet adjectif n'est pas présent dans les données utilisées par HATHOUT *et al.* (2003) ; aucune étude de cette suffixation ne prévoit ce type de dérivé ; le statut marginal de cet exemple peut être relativisé en exhibant d'autres adjectifs en *-able* construits sur des bases adjectivales comme *difficilable* (5a), *facilable* (5b), *seulable* (5c) ou *tristable* (5d) même s'ils sont relativement difficiles à interpréter.

(5) a. Xerath n'est pas non plus très **difficilable** à prendre en main `ffrlol.forumgratuit.org`

b. Je nsuis po... zune fille facilement **facilable**... `gniuuu.skyrock.com`

c. Dire qu'à une époque, les garçons faisaient la queue pour me bécoter l'oreille et que je me retrouve aujourd'hui plus **seulable** pour le restant des mes jours. `indra-nimportkoi.blogspot.fr`

d. ..Mdrrrrrrrrrr. ..vraiment c'est **tristable** hein.... En plus il te largue salement comme ça et tu oses encore le pleuré. ...otula oooh nini (tu galère niveau chopage de ... `/fr-fr.facebook.com/LaGossiPeuZe`

L'accumulation de tels exemples complique la tâche du morphologue, qui ne peut les ignorer en bloc en les déclarant simplement ininterprétables ou agrammaticaux (voir BAUER 2014 pour une discussion sur les conséquences de l'utilisation des grands corpus sur les notions de grammaticalité et de productivité).

2.2 L'évolution des ressources et du nombre des exemples

Les exemples précédents sont des exemplaires caractéristiques du type de données que l'on peut actuellement collecter pour l'étude d'un phénomène morphologique. Auparavant, c'est-à-dire avant la création du Web, les recherches en morphologie s'appuyaient sur des relevés effectués dans des ouvrages imprimés et des dictionnaires, à un cout prohibitif en temps puisqu'il fallait lire dans leur intégralité la totalité des livres du corpus pour compiler les listes de mots sur lesquels portaient les études. Ce cout était d'autant plus élevé que le nombre des exemples intéressants contenus dans les livres est généralement très faible. En contrepartie, ces exemples étaient irréprochables car provenant de textes édités qui ont subi des révisions nombreuses. À titre d'exemple, on trouve **183** adjectifs ayant la finale `able` dans un corpus d'environ 800 000 mots composé des huit romans de la base Frantext-démonstration parus entre 1803 et 1908. La collecte d'exemples dans les dictionnaires exige un effort moindre car la lecture se limite à la seule nomenclature. Ainsi, la nomenclature de la 8e édition du dictionnaire de l'Académie (31 934 entrées) contient **444** adjectifs se terminant par `able`. Un premier changement d'échelle a eu lieu à partir des années 1990, lorsque les morphologues ont eu accès à des corpus électroniques comme Frantext, mis en ligne en 1992, et à des dictionnaires informatisés comme le *Trésor de la Langue Française informatisé* (TLFi), à partir de 2000. La première évolution concerne le temps nécessaire à la collecte des exemples, qui

est réduite à quelques minutes, voire quelques secondes. La seconde est l'augmentation du nombre d'exemples et leur plus grande diversité. Les données extraites des dictionnaires électroniques sont les mêmes que celles qui se trouvent dans les versions imprimées. On peut par exemple extraire du TLFi **1034** adjectifs en `able` que l'on peut utiliser sans révision ni correction. Ce n'est pas le cas des données provenant d'autres types de corpus, notamment journalistiques qui s'avèrent fortement bruités et qui imposent une vérification et un nettoyage de chacun des exemples qui en sont extraits. Signalons que le rendement de ces corpus n'est pas toujours très élevé : 5 années d'archives du quotidien *Libération* couvrant la période 1995-1999 (87 millions de mots) contiennent **767** adjectifs finissant en `able` tandis que 4610 documents (227 millions de mots) de la base Frantext-intégral (consultée en mars 2015) en fournissent près de **2900**[3]. Ces quelques exemples de ressources imprimées et électroniques donnent une idée de l'évolution de leur taille et de leur capacité à fournir aux morphologues un nombre croissant d'exemples.

À partir de la fin des années 1990, les morphologues ont eu accès à ce que l'on peut considérer comme la ressource ultime : le Web. Cette source d'exemples innombrables présente des caractéristiques uniques qui la rendent incontournable, facilement accessible mais relativement difficile à utiliser et à exploiter (GREFENSTETTE 1999 ; KILGARRIFF & GREFENSTETTE 2003 ; HATHOUT *et al.* 2009a ; TANGUY 2013).

[3] Le nombre exact de ces adjectifs ne peut pas être calculé informatiquement car l'information catégorielle n'est pas disponible pour les listes de mots.

Ses principaux avantages sont :

> 1/ sa taille exceptionnelle. Le Web est sans nul doute la plus grande collection de documents disponibles même si l'on ne peut accéder qu'à une fraction des pages existantes. En effet, l'accès à cette ressource ne peut se faire que par l'intermédiaire des moteurs de recherche qui n'enregistrent dans leurs index qu'un petit sous-ensemble du Web (TANGUY 2012). Cette fraction reste néanmoins d'une taille qui dépasse celle de tout autre corpus. Or la taille d'une ressource détermine sa richesse, tant en nombre de lexèmes différents, d'emplois différents pour ces lexèmes et par suite de sens différents.
>
> 2/ la diversité des documents sur les plans diatopique (influence des substrats régionaux), diastratique (sociolectes) et diaphasique (styles et registres de langue). On trouve également sur le Web énormément de documents techniques qui relèvent d'un grand nombre de domaines de spécialité, même si certains comme l'informatique sont mieux représentés que d'autres. Le Web enregistre notamment un nombre exceptionnel de discussions informelles sur les forums, les blogs, les plateformes de microblogage, etc. qui donnent accès à une langue très spontanée, souvent peu normée, dans laquelle la créativité lexicale est forte. Ces types de textes ne sont (et n'ont jamais été) disponibles dans aucune autre ressource.
>
> 3/ le Web fournit un accès rapide voire immédiat aux évolutions des langues, par exemple à l'explosion des constructions en *-itude* en 2007 et à l'extension des conditions d'emploi de cette suffixation qui l'a accompagnée (KOEHL 2012a ; KOEHL & LIGNON, 2014).

Il faut néanmoins garder à l'esprit que le Web n'est pas un corpus. Nul ne connait sa composition, ne dispose d'un inventaire des documents qu'il contient, ni de leurs caractéristiques fondamentales comme leur date et lieu de publication, le nom, l'âge, la nationalité ou le sexe de leurs auteurs, la langue dans laquelle ils sont rédigés, le niveau de maitrise de cette langue par le ou les auteurs, l'utilisation

éventuelle d'outils de traduction automatique, etc. Il n'est pas équilibré comme l'est par exemple le British National Corpus. Il n'est représentatif de rien, sinon de lui-même. La taille du Web n'est pas connue et ne peut être mesurée. Du fait de sa constante évolution, les expériences réalisées sur le Web ne sont généralement pas reproductibles. Il est impossible d'obtenir la liste des mots utilisés dans le Web. Les moteurs de recherche ne donnent accès qu'à une sélection des documents qu'ils ont indexés et l'interrogation ne peut s'effectuer qu'au moyen de formes (mots simples ou de séquences de mots). On ne peut pas obtenir l'ensemble des pages indexées par un moteur de recherche qui contiennent un mot donné. Google annonce par exemple que son index contient plus de 4,3 millions de documents comprenant le mot *décomposition* mais limite à 406 la liste des résultats affichables. Malgré ces limitations, le Web reste une mine d'exemples inégalable qui nous a notamment permis, en 2002, de collecter près de **4000** adjectifs en *-able* ne figurant pas dans le TLFi.

3. « More data is better data »

Il apparait de façon implicite dans les chiffres présentés ci-dessus que la taille des corpus et le nombre des exemples sont des facteurs importants pour les études en morphologie et que l'on pourrait reprendre un slogan bien connu de la linguistique de corpus : *More data is better data* (CHURCH & MERCER 1993) ou sa version français « gros c'est beau » (PÉRY-WOODLEY 1995). Plus précisément, l'approche extensive en morphologie (PLÉNAT 2000 ; HATHOUT *et al.* 2003) consiste à fonder les analyses morphologiques sur le plus grand nombre possible d'exemples. Elle considère en effet que la quantité d'exemples pris en compte détermine directement la qualité des analyses et que ces derniers doivent être collectés de manière systématique. De leur nombre dépend la bonne compréhension des procédés et des phénomènes étudiés.

Les analyses des phénomènes qui sont fondées sur de grands nombres d'exemples sont plus fines et rendent mieux compte des données moins centrales, plus « exceptionnelles ». Les recherches menées à l'ERSS sur les adjectifs en *-esque* et en *-able* illustrent parfaitement les progrès qui ont été rendus possibles par la morphologie extensive.

3.1 Les voyelles moyennes devant *-esque*

La suffixation en *-esque* construit des adjectifs dont les bases peuvent être des noms communs (6a) et des noms propres (6b). Elle a fait l'objet de plusieurs études menées à l'ERSS pendant une dizaine d'années par PLÉNAT et ses collaborateurs. Ces dérivés constituent en effet un matériau adapté à l'étude des contraintes dissimilatives, qui constitue l'objet réel des recherches de PLÉNAT (2011). Ces contraintes pénalisent l'apparition à faible distance de phonèmes identiques ou similaires.

(6) a. sultan → sultanesque
b. Molière → moliéresque

Dans ces adjectifs, PLÉNAT s'intéresse principalement au comportement des voyelles moyennes antérieures (/e, ɛ, ø, œ/) qui se trouvent à la fin des radicaux des dérivés en *-esque* et qui sont suivies d'une consonne fixe (i.e. non-latente) comme en (7) :

(7) a. Cervantes → cervantesque
b. enchanteur → enchanteuresque

Ces exemples montrent que dans certains dérivés, la rime tombe mais pas dans d'autres. Quels sont les conditions du maintien ou de la chute de la rime ?

Une consultation du TLFi permet de collecter 104 adjectifs dérivés en *-esque* qui ne posent aucun problème car ils sont formés par simple concaténation du thème de la base et de l'exposant du suffixe, comme en (8).

(8) a. Molière → moliéresque
b. Raphaël → raphaélesque

PLÉNAT a entrepris avec l'aide de SERNA une collecte systématique de dérivés en *-esque*, notamment dans les romans de San Antonio. En 1997, 800 dérivés ont ainsi été rassemblés, qui font apparaitre que les rimes en /ɛ/ suivies d'une consonne fixe peuvent tomber lorsque la base comporte au moins quatre syllabes (9). Les bases de trois syllabes qui finissent en /s/ sont normalement raccourcies (10). PLÉNAT est cependant intrigué par l'exemple (11) dans lequel la finale *-eur* est supprimée alors que la base ne comporte que trois syllabes.

(9) a. Pantagruel → pantagruesque
b. consommateur → consommatesque
(10) a. Cervantes → cervantesque
b. cosinus → cosinesque
(11) tirailleur → tiraillesque

La collecte s'est poursuivie pendant encore quelques années. En 2001, PLÉNAT dispose de plus de 3000 dérivés qui confirment que les rimes composées d'une voyelle moyenne antérieure et d'une consonne fixe tombent dans les radicaux de quatre syllabes ou plus. De même, les rimes dont la consonne finale est identique à l'une des consonnes du suffixe (/s/ ou /k/) tombent dans les radicaux de deux et trois syllabes (12). Mais ces données ont également permis à PLÉNAT de mettre au jour une régularité inédite et bien plus surprenante : la rime tombe aussi lorsque la consonne finale du radical est répétée (PLÉNAT & ROCHÉ 2003), c'est-à-dire lorsqu'elle y apparait une seconde fois comme en (13).

(12) (Louis de) Funès → funesque

(13)

		consonne répétée
a.	co**l**one**l** → colonesque	/l/
b.	Be**n** Lade**n** → benladesque	/n/
c.	ti**r**ailleu**r** → tiraillesque	/r/
d.	In**t**erne**t** → internesque	/t/

Cette étude montre qu'il est possible d'aborder la morphologie comme une science d'observation (HATHOUT *et al.* 2009a). L'augmentation du nombre des exemples pris en compte par les analyses morphologiques a un effet similaire à l'introduction du microscope dans les sciences naturelles. Quand l'observation d'une centaine de dérivés ne permet pas de dégager de généralité intéressante, celle de 3000 exemples fait apparaitre des régularités inédites qui conduisent à de nouvelles conclusions.

Ces régularités concernent notamment des configurations qui, dans la collection réduite, paraissent exceptionnelles. Mais dès lors qu'elles sont mieux représentées dans les collections étendues il devient possible d'identifier les facteurs qui expliquent leur fonctionnement et de proposer une analyse du phénomène capable de les intégrer au cas général.

3.2 La sémantique des dérivés en *-able*

Les avancées rendues possibles par l'approche extensive en morphologie ne concernent pas seulement la dimension morphophonologique. Elles peuvent également être significatives sur le plan sémantique comme l'illustre l'étude de la suffixation en *-able* de HATHOUT *et al.* (2003). Plusieurs études de cette suffixation (DUBOIS 1969 ; PLÉNAT 1988 ; LEEMAN & MELEUC 1990 ; LEEMAN 1992 ; ANSCOMBRE & LEEMAN 1994 ; FRADIN 2003) avaient été réalisées antérieurement à partir de collections dont la taille n'excède pas les 1400 adjectifs ; ce qui correspond approximativement au nombre des dérivés en *-able* des grands dictionnaires de langue comme le TLF ou le GRLF (*Grand Robert de la Langue Française*). Sur le plan sémantique, la suffixation en *-able* était analysée comme ayant un « sens passif ». Les dérivés en *-able* sont en effet principalement construits sur des verbes et sont utilisés pour modifier des noms qui, dans l'évènement dénoté par la base verbale, ont un rôle de patient (14). Dans

une étude plus ancienne, GAWELKO (1977) a identifié trois petites séries de dérivés construits sur des bases nominales, en l'occurrence des noms de taxes (corvée, taille, gabelle, etc.), de véhicules et de titres (15).

(14) réparer → réparable
'on peut réparer le téléphone'
= 'le téléphone peut être réparé'
= 'le téléphone est réparable'

(15) a. corvée → corvéable
b. cycle → cyclable
c. président → présidentiable

Les données disponibles au début des années 1990 comportaient cependant des dérivés dont l'analyse est problématique. On trouve d'une part des dérivés qui se comportent différemment des passifs de leurs verbes de base (16). Plus gênants sont les dérivés pour lesquels le nom recteur ne peut pas être analysé comme correspondant à un argument du verbe de base (17).

(16) a. Marie répare le téléphone
Le téléphone est réparable
b. Cette robe coute 100 euros
* 100 euros sont coutables
c. Un terrain atterrissable
* L'avion atterrit le terrain

(17) Une robe à un prix abordable

En 2003, HATHOUT, PLÉNAT & TANGUY ont collecté et analysé 5286 dérivés. Ce nombre est plus de trois fois supérieur à celui des exemples considérés dans les études antérieures. Les auteurs constatent tout d'abord que le sens de la grande majorité des adjectifs en *-able* peut être décrit comme « passif ». Ils observent cependant que les noms recteurs des adjectifs en *-able* peuvent représenter une grande part des participants à l'évènement dénoté par le

verbe de base. On trouve par exemple des sujets comme en (18) et des compléments indirects comme en (19).

(18) D'une manière générale, la sensibilité au gel d'une pâte de ciment est étroitement liée à la quantité d'eau "**gelable**".

(19) Le premier PC «**parlable**». On pourra maintenant causer à son ordinateur.

Ces deux exemples soulèvent une question essentielle, relative à la nature et à l'acceptabilité des données prises en compte dans les analyses (voir section 2.1). Les dérivés *gelable* en (18) et *parlable* en (19) sont-ils des lexèmes du français ? Sont-ils des lexèmes acceptables ? Peut-on (ou doit-on) les utiliser comme des instances de dérivés en *-able* ? La réponse de HATHOUT *et al.* (2003) est clairement affirmative. Ils considèrent en effet qu'en l'absence d'indices clairs permettant d'affirmer qu'un énoncé *n'*a *pas* été produit par un locuteur ayant une bonne maitrise du français, et si cet énoncé ne comporte ni erreur ni dysfluence, alors il doit être intégré à la collection des exemples à analyser. On ne peut en effet en aucun cas se limiter au français fortement normé que l'on trouve dans les grands dictionnaires de langue. Le rôle du linguiste est d'expliquer le fonctionnement de la langue parlée par les locuteurs et non de celle qui est idéalisée par les instances de normalisation institutionnelles. La lecture des documents dont sont extraits (18) et (19) montrent clairement que les locuteurs qui les ont produits ont une maitrise parfaite de la langue, qu'ils connaissent les normes comme le montre l'emploi des guillemets, qu'ils ont considéré que la création et l'utilisation de ces dérivés est légitime et qu'elles ne posent pas de problème de compréhension ou d'interprétation aux lecteurs. Dans ces conditions, rien ne justifie l'exclusion de ces exemples.

La question de l'acceptabilité des exemples collectés est permanente. Elle se pose pour chaque dérivé présent dans les

corpus ou sur le Web et doit faire l'objet d'une réponse au cas par cas dont l'une des conséquences est le cout prohibitif de la constitution des collections d'exemples. Signalons que la collecte intensive d'exemples a elle-même une influence sur les jugements d'acceptabilité dans la mesure où il est parfois difficile d'estimer la qualité d'un dérivé non-standard isolé, mais son interprétation peut devenir plus facile lorsqu'il est rapproché d'autres mots similaires. Remarquons par ailleurs que les exemples illustrant les emplois des dérivés ne sont pas édités : les fautes d'orthographe et de typographie sont conservées.

La diversité des participants à l'évènement dénoté par un verbe de base que son dérivé en *-able* permet de modifier peut être illustrée par les exemples suivants de l'adjectif *pêchable* (20-29). La plupart proviennent de blogs ou de groupes de discussion. Certains sont plus marqués que d'autres.

(20) Poisson Avec ce concept révolutionnaire, enfin les gros **poissons** sont **pêchables** au coup !

(21) Taille des poissons La sur-pêche et le non respect de la **taille pêchable** en Guadeloupe a entraîné une forte régression de la population.

(22) Lieu de pêche [...] 3 km de **rives pêchables,** bien aménagées pour le lancer [...]

(23) Longueur du lieu de pêche La **longueur pêchable** sur les 2 berges est de 2 025 mètres.

(24) Étendue d'eau La **rivière** reste **pêchable** en été , [...]

(25) Jour 31 Aout Eau très haute (9,7 m3/s) et froide (9°C), premier **jour pêchable** depuis le 15 Aout. Quelques gobages, surtout des petits poissons, ...

(26) Saison de pêche C'est vrai, la carte de pêche complète à 75€, rapportée aux nombres de **jours pêchables**, et même si ça augmente chaque année, ce n'est pas hors de prix. ...

(27) Vent Jusqu'à 14 ça va, au delà je sors pas car le **vent** devient trop gênant voir **impêchable**. ...

(28) Conditions météorologiques Si le vent monte trop et que les **conditions** ne deviennent plus **pêchables,** plusieurs solutions s'offrent à vous :. - tout plier et attendre une accalmie ...

(29) Fil de pêche je remarque aprés quelques lancers (je peche generalement a 40 metres en etang) que **mon nylon** se met a vriller et devient **impechable**. ...

La plasticité sémantique de ces adjectifs n'était pas signalées dans les études de la suffixation en *-able* publiées avant 2003. Elle permet notamment de modifier les objets, les lieux où se trouvent les proies et ceux où se postent les pêcheurs, les instruments, etc. mais aussi les propriétés de ces participants, notamment leur dimension ou leur force (pour le vent). On peut rattacher à ce dernier cas et plus précisément à (21), l'exemple en (16) où le prix est une propriété de l'un des participants, en l'occurrence la robe. Dans le cas de *pêchable*, il semble en effet que tout participant à l'évènement dénoté par le verbe *pêcher* ainsi que toutes ses propriétés puissent être modifiés par l'adjectif, à l'exception de l'agent, le pêcheur. Cette conclusion peut être reformulée comme suit :

> X peut être dit *pêchable* si
> - X a une propriété qui favorise l'évènement dénoté par le verbe *pêcher* ;
> - X intervient dans l'évènement mais ne peut pas être l'agent.

Dans HATHOUT (2009), j'ai proposé une analyse plus générale en termes de dynamique des forces (TALMY 2000) qui ne nécessite pas d'exclure explicitement les agents :

> X peut être dit *pêchable* si X est susceptible d'exercer une force antagoniste qui s'oppose à la réussite de l'évènement dénoté par le verbe *pêcher*, mais qui n'est pas suffisante pour l'empêcher.

Par exemple, une berge est *pêchable* si les éventuelles difficultés liées à son accès et à son utilisation n'empêchent pas que l'on puisse y pêcher du poisson avec succès. Notons que cette analyse peut aussi rendre compte de certains dérivés désadjectivaux comme *seulable* : l'auteure a du succès auprès des garçons ; ce succès exerce une force qui s'oppose à la réalisation (l'avenance) de l'état de solitude ; cette force n'est pas suffisante.

Ces avancées dans la description du sens des adjectifs en *-able* dépendent directement de la taille de la collection d'exemples réunis pour cette étude. Les analyses qui en découlent permettent de réintégrer des dérivés jusque-là considérés comme exceptionnels comme *abordable* et d'expliquer que *coutable* ne soit pas attesté : *couter* décrit une propriété qui n'implique pas de succès ni d'échec.

L'étude de HATHOUT *et al.* (2003) a aussi permis de compléter celle de GAWELKO (1977) en mettant au jour quatre nouvelles séries d'adjectifs dénominaux dont les bases dénotent des noms de construction (30), de lieu (31), de peine (32) et de finalité (33).

(30) Terrain 1200m M2 arboré et **piscinable**.
(31) L'objet **muséable** est à votre image [...]
(32) le simple fait de prier un dieu, ou même de prêcher le Juge, était un fait grave, **peinable de mort**.
(33) L'évolution du prix de la commission pour les Bintje **fritables** est présentée à la figure 4.

Par ailleurs, HATHOUT *et al.* (2003) ont montré que les dérivés dénominaux font partie des mêmes séries que les dérivés déverbaux et que leur création s'explique d'abord par l'absence de verbe permettant de dénoter l'évènement évoqué par le nom de base.

4. Les hauts et les bas dans la collecte de données extensives pour la morphologie

L'utilisation de données réelles en grande quantité se généralise. En témoignent les collections réunies dans le cadre des thèses en morphologie soutenues ces dernières années comme celles de TRIBOUT (2010) ou KOEHL (2012b). Les données utilisées dans KOEHL (2012b) proviennent en grande partie du Web, qui tend à devenir la source quasi-universelle de tous les exemples utilisés dans les recherches morphologiques en synchronie. Je rappelle qu'en réalité, nous n'avons accès qu'à une petite partie des pages Web : seules les occurrences de mots qui figurent dans les pages indexées par les moteurs de recherches peuvent être retrouvées. Idéalement, les morphologues pourraient constituer leurs collections directement à partir de ces index qui normalement contiennent l'ensemble des mots qui apparaissent dans les pages référencées. Le résultat serait quasi-parfait : le nombre des exemples maximal, la précision très élevée, et la collection peu biaisée car ne reflétant pas les intuitions des linguistes qui les ont constituées, etc. Cet idéal est malheureusement totalement inaccessible car la qualité des moteurs de recherche dépend crucialement de celle de leurs index ; leur valeur est colossale et leur protection maximale.

4.1 Interrogation automatisée des moteurs de recherche

La protection des index n'a pas toujours été aussi forte qu'elle l'est aujourd'hui et les possibilités d'interrogation offertes par les premiers moteurs étaient plus variées que celles que nous connaissons actuellement. Au début des années 2000, les morphologues avaient la possibilité de soumettre aux moteurs de recherche comme AltaVista ou Yahoo! des quantités importantes de requêtes automatiques. Ils pouvaient ainsi récupérer en peu de temps des nombres élevés de candidats dérivés. Le moteur AltaVista acceptait en

outre des requêtes par patron dans lesquelles une partie des mots pouvait être remplacée par un joker. Par exemple, une requête `pro*able` permettaient de récupérer des pages contenant des mots comme (34).

(34) probabilisable, professorable, promenable, promotable, promouvable, prononciable, protégeable

Ces requêtes permettaient en théorie de récupérer dans l'index du moteur tous les mots susceptibles d'avoir été construits par une règle spécifique de construction de lexèmes. Leur intérêt le plus important résidait dans le fait qu'elles fournissaient des mots que le morphologue n'avait pas prédits, parce qu'ils n'entraient pas dans sa conception du phénomène. Un tel apport pouvait se révéler décisif pour l'analyse.

La possibilité de soumettre des requêtes automatiques aux moteurs de recherche a été exploitée par plusieurs outils comme Webaffix (TANGUY & HATHOUT 2002 ; HATHOUT & TANGUY 2003) ou WaliM (NAMER 2003, 2009). Webaffix est une boite à outils d'acquisition lexicale à partir du Web qui dispose de plusieurs modules. Le premier permet de construire un ensemble de requêtes incluant des jokers permettant de récupérer les mots connus du moteur et qui contiennent un affixe donné. Le second crée des requêtes en prédisant des formes possibles à partir de schémas d'affixation appris sur un lexique flexionnel comme TLFnome[4]. Le troisième permet de soumettre ces deux types de requêtes à un moteur de recherche et de réaliser diverses opérations de nettoyage des résultats en éliminant les pages qui ne contiennent pas le ou les mots recherchés, celles qui ne sont pas rédigées dans la langue souhaitée, celles où le

[4] TLFnome est un lexique flexionnel du français construit à l'INaLF/ATILF à partir de la nomenclature du Trésor de la Langue Française. Morphalou, la version XML de cette ressource est distribuée par le CNRTL à l'adresse suivante : www.cnrtl.fr/lexiques/morphalou/

mot se trouve dans une liste, etc. Webaffix a été utilisé pour constituer les collections d'exemples de plusieurs études en morphologie extensive menées à l'ERSS et a fortement contribué à la démonstration de la supériorité de l'approche extensive en morphologie.

Plusieurs outils similaires ont été développés à la même époque, notamment le méta-moteur WaliM réalisé par NAMER pour vérifier si des mots prédits sont attestés sur le Web. Les formes des mots dérivées sont générées à partir de TLFnome au moyen de GéDériF (NAMER & DAL 2000), un générateur morphologique qui implémente des règles de construction de lexèmes conçues et mises au point par des linguistes. Ces formes font ensuite l'objet de requêtes soumises au moteur Yahoo! dont les résultats sont filtrés pour ne conserver que les mots ayant au moins une attestation.

4.2 Les gros corpus de page Web

À partir de 2003, les possibilités d'interroger les moteurs de recherche au moyen de robots sont petit à petit devenues plus limitées jusqu'à disparaitre complètement. Aujourd'hui, plus aucun moteur ne les accepte. Seule l'interrogation manuelle au moyen d'un navigateur est autorisée. Ces restrictions constituent un retour en arrière dont les conséquences sur la morphologie extensive sont importantes. Les linguistes doivent se contenter de vérifier l'attestation des dérivés les plus probables qui sont généralement les moins intéressants, dans la mesure où ce sont les moins susceptibles de faire progresser les descriptions. La taille des collections d'exemples sur lesquelles ils fondent leurs études se trouve réduite de façon significative, conduisant à un moins grand nombre de généralisations et à des généralisations plus grossières. Enfin, ces collections sont plus biaisées que par le passé car elles dépendent de l'intuition de celui qui prédit les formes

dérivées, qui, involontairement tend à favoriser les formes compatibles avec sa théorie et à pénaliser celles dont elles seraient des contre-exemples.

Ces restrictions ont également conduit les morphologues à se tourner vers un « succédané » du Web, à savoir les gros corpus de pages Web comme frWaC (BARONI *et al.* 2009). Ce corpus français de 1,6 milliard de mots, librement disponible à des fins de recherche, a été constitué par BARONI et son équipe pour réaliser des études de sémantiques distributionnelles. D'autres corpus, plus gros encore, ont été compilés dans le cadre de programmes de recherche comme Quaero. Exalead, le maitre d'œuvre, a ainsi crée un corpus de plus 2,5 millions de pages contenant plus de 3,3 milliards de mots. SAJOUS, TANGUY et moi avons réalisé une étude d'acquisition morphologique sur le corpus Exalead, afin notamment de comparer les exemples collectés avec ceux que l'on pouvait obtenir en utilisant Webaffix (HATHOUT *et al.* 2009b). L'étude portait sur les suffixations déverbales en *-age*, *-ment* et *-ion* pour lesquelles nous disposons de collections réunies dans le cadre du projet WesConVa (DAL *et al.* 2004). L'un des enseignements de cette étude est qu'on trouve en moyenne un nouveau déverbal (i.e. absent du TLF) dans 2000 pages Web, soit environ 1200 dérivés dans les 2,5 millions de pages. À titre de comparaison, le lexique Verbaction, créé à partir du TLF contient 3800 déverbaux en *-age*, *-ment* et *-ion*. Le corpus Exalead permet donc d'augmenter la collection des exemples disponibles de l'ordre de 30%, là où Webaffix permettait des progressions allant de 300% à 3000% ! Malgré sa taille exceptionnelle, ce corpus est de fait tout petit. Il impose de plus au morphologue d'avoir des compétences minimales en traitement automatique des langues, du type de celles présentées dans TANGUY & HATHOUT (2007) car il ne dispose pas d'une interface d'interrogation.

Cette baisse considérable dans les capacités de découverte de nouveaux dérivés a de nombreuses conséquences : le cout de création des jeux de données augmente dans la mesure où elle nécessite une plus grande intervention des morphologues, qui doivent notamment soumettre leurs requêtes manuellement ; les jeux de données sont plus petits ; il faut plus de temps pour trouver des exemples intéressants, non-prévus par les théories et les analyses actuelles.

Je voudrais enfin signaler que la constitution d'une collection d'exemples pour une étude morphologique comporte aussi des « frais cachés » souvent assez lourds. Les attestations récupérées sur le Web sont fortement bruitées : chaque exemple doit être soigneusement examiné. Or une campagne peut ramener plusieurs dizaines de milliers de candidats dont il faut vérifier l'acceptabilité. Les faux positifs sont en effet très nombreux. Une forme peut être le résultat d'une faute de frappe, d'une faute d'orthographe, d'une erreur de découpage dues à une césure ou à l'omission d'un espace ; elle peut appartenir à une partie du discours autre que celle qui est visée, être un nom propre (par exemple, un identifiant dans un blog) ou un mot d'une autre langue ; elle peut apparaitre dans du code informatique, dans une adresse mail ou une URL ; elle peut être produite par un traducteur automatique ou par une personne qui n'a clairement pas une maitrise suffisante de la langue ; etc. À cela s'ajoutent les questions d'ambigüité, comme par exemple les noms en *-eur* en français qui peuvent être des noms d'agent (*tailleur*) et des noms de propriétés (*longueur*), même si cette seconde dérivation tend à être aujourd'hui remplacée par la suffixation en *-ité* (KOEHL 2012b). J'ajoute enfin que ce travail philologique doit être répété pour chaque nouvelle collecte.

5. La préservation des collections de données morphologiques

Les données, devenues plus couteuses à obtenir que par le passé, n'en demeurent pas moins indispensables à toute recherche en morphologie. L'histoire esquissée dans les sections précédentes est celle d'une évolution, dont le point de départ a été l'utilisation de collections limitées car difficiles à constituer. À ces époques, les philologues produisaient des ouvrages dans lesquels ils compilaient à la main le vocabulaire. Les ressources sont ensuite devenues plus accessibles, puis de plus en plus conséquentes jusqu'à l'abondance amenée par le Web et les premiers moteurs de recherche. Cette mutation a permis de problématiser la place des données réelles dans le dispositif de recherche en morphologie et d'établir définitivement (1) qu'elles sont indispensables et (2) que la qualité des analyses dépend directement de la quantité des données utilisées. Aujourd'hui la situation redevient plus complexe, et il faut inventer de nouvelles manières de travailler afin de s'adapter aux restrictions imposées par les moteurs de recherche. L'une d'elles est d'utiliser des ressources où l'on sait que la créativité morphologique est peu contrainte par les normes institutionnelles comme les tweets de la plateforme Twitter où les enfants se déclarent in-dormables, où les restaurants sont étoilables, etc. Si le travail de nettoyage et de vérification des exemples constitue une part importante dans le cout de création d'une collection d'exemples, les morphologues ont généralement peu conscience de sa valeur et rares sont ceux qui finalisent leurs jeux de données et les mettent à la disposition de la communauté. C'est à cet aspect qu'est consacrée cette dernière section : la préservation des collections qui va de pair avec une meilleure reconnaissance du travail investi dans leur constitution.

5.1 La conservation et la dissémination des données en morphologie

Dans de nombreuses sous-disciplines de la linguistique, notamment en socio- ou en psycholinguistique, le partage et la réutilisation de jeux de données existants est une pratique bien établie ; ce n'est pas encore le cas en morphologie. En psycholinguistique, par exemple, des ressources comme CELEX (pour l'anglais, l'allemand ou le néerlandais ; BAAYEN *et al.* 1995) ou Lexique (pour le français ; NEW 2006) servent à constituer du matériel expérimental pour les études sur le lexique mental et sur les traitements morphologiques. D'autres ressources comme CHILDES (MACWHINNEY 2000) sont utilisées dans de très nombreuses recherches sur l'acquisition du langage. Il n'existe, en revanche, rien de comparable pour la description morphologique. Chaque étude débute par la constitution d'une nouvelle collection d'exemples dont le point de départ est généralement un dictionnaire électronique comme le TLFi. Le recours au Wiktionnaire est plus rare car il n'existe pas pour l'instant d'outil d'interrogation adapté à la recherche morphologique[5]. Suit une pêche aux exemples sur le Web plus ou moins fastidieuse, plus ou moins fructueuse. Les dérivés sont ensuite décrits dans des tables ou une base de données dont le contenu dépend essentiellement des facteurs qui interviennent dans les analyses prévues. Il arrive souvent par ailleurs, que certaines parties de la collection soient traitées de manière plus approfondie que d'autres. Le format de ces jeux de données est également *ad hoc* et parfois hétérogène, notamment lorsqu'ils se composent de plusieurs fichiers. Ce « manque de considération » envers les données s'explique

[5] Cette situation devrait néanmoins évoluer grâce à GLÀFFOLI (http://redac.univ-tlse2.fr/glaffoli/), l'interface d'interrogation du lexique GLÀFF (SAJOUS *et al.*, 2013 ; HATHOUT *et al.*, 2014), qui permet à l'utilisateur d'extraire les entrées du Wiktionnaire en combinant différents critères catégoriels et de forme.

d'abord par le fait que ces collections sont créées spécifiquement pour une étude particulière et que leur dissémination ne fait pas partie des objectifs du morphologue qui les constitue : l'utilisation de données déjà analysées ne fait pas (encore) partie des pratiques de la communauté ; la finalisation et la dissémination des données comportent par ailleurs un surcout qui ne saurait se justifier que si ce travail était considéré comme ayant une valeur en soi, s'il pouvait être valorisé par des publications et obtenir une reconnaissance suffisante des instances d'évaluation.

Notons qu'il existe de très nombreux travaux sur les formats de données lexicales, conçus et développés essentiellement en traitement automatique des langues, comme LMF (Lexical Markup Format ; FRANCOPOULOS 2006). Ces formats se caractérisent par une grande généricité mais ils ne sont pas utilisés pour la description des données en morphologie parce que la plupart des morphologues ne les connaissent pas et qu'ils ne sont pas suffisamment adaptés aux besoins de ces derniers.

Pour amener les morphologues à partager et réutiliser plus systématiquement les collections et les analyses associées, il faudrait concevoir une ressource à large couverture, disposant d'une interface d'interrogation et d'outils de gestion intuitifs. Une telle ressource doit disposer d'une architecture qui accepte une grande variété de descriptions morphologiques. Elle doit être aussi œcuménique que possible sur le plan théorique et ne doit pas être fondée sur des hypothèses ou des présupposés sur le contenu ou la forme des analyses morphologiques. Afin de permettre l'alignement des différentes descriptions, les informations doivent être suffisamment décomposées. L'information y sera distribuée du fait de la décomposition des informations, et redondante – une même information pouvant apparaitre dans plusieurs éléments. La redondance viendra également des différences de granularité dans la

ressource d'éléments d'information qu'elle réunit. Cette ressource devra permettre que certaines informations soient manquantes ou incomplètes. Enfin, elle indiquera explicitement l'origine de chacune des informations qu'elle contient pour reconnaitre le crédit de leurs auteurs, permettre la citation de leur travaux et éventuellement la sélection ou le masquage de certaines des descriptions. L'objectif premier est donc de créer une ressource où les morphologues puissent intégrer leurs collections d'exemples, obtenir des données pour de nouvelles études et à terme l'utiliser comme un outil intégré de constitution et de stockage des jeux de données morphologiques.

5.2 Le réseau morphologique *Démonette*

Une ressource qui satisfait en partie aux spécifications présentées ci-dessus est en cours de développement dans le cadre d'une collaboration avec NAMER (HATHOUT & NAMER 2014a, 2014b). Ce nouveau réseau lexical du français, baptisé *Démonette*, se caractérise par la variété des relations morphologiques qu'il décrit, à la fois directes (entre ascendants et descendants) et indirectes (au sein de la même famille dérivationnelle), simples et complexes, et par le nombre des traits morphologiques, phonologiques, catégoriels et sémantiques dont sont munis les sommets, qui représentent les lexèmes, et les arcs, qui représentent les relations dérivationnelles. Démonette est conçu pour articuler des informations provenant de deux systèmes fondés sur des principes totalement opposés. Le premier est DériF (NAMER 2009, 2013), un analyseur morphologique dérivationnel qui implémente une vingtaine de règles de construction de lexèmes définies et mises au point par des linguistes comme la suffixation en *-age*, la préfixation en *dé-* ou la composition savante. Ce système prend en entrée des formes de citation de lexèmes (construits) munies de catégories grammaticales. Pour chaque lexème construit,

DériF calcule un lexème de base, le procédé dérivationnel utilisé pour le construire, la liste de ses antécédents dérivationnels et une glose de son sens construit (35). Les analyses sont réalisées en appliquant récursivement les règles implémentées. DériF dispose en outre de listes d'exceptions qui permettent de prendre en compte les irrégularités lexicales.

(35) enneigement/NOM ==> [[en [neige NOM] VERBE] ment NOM] (enneigement/NOM, enneiger/VERBE, neige/NOM) "(Action - résultat de l'action) de enneiger"

Démonette contient également des analyses provenant de Morphonette (HATHOUT 2011) un réseau lexical du français basé sur une conception relationnelle et paradigmatique de la morphologie. Dans ce lexique, les propriétés morphologiques sont décrites par la position des lexèmes dans le réseau, position identifiée par les paradigmes qui les contiennent. Par exemple, la position d'un dérivé comme *modifiable* est décrite par sa famille dérivationnelle qui rassemble les lexèmes *modifier*, *modification*, *modificateur*, *modificatif*, *modifiant*, *modifieur, immodifiable*, etc. et par sa série qui contient l'ensemble des dérivés en *-able* : *agaçable*, *agitable*, *chevauchable*, *définissable*, *différenciable*, *rechargeable*, *réconciliable*, *soutenable*, etc. Morphonette est composé de filaments, c'est-à-dire de triplets (m, p, $s_p(m)$) où m est une entrée, p est un membre de la famille dérivationnelle de m et $s_p(m)$ est la sous-série dérivationnelle de m relativement à p. $s_p(m)$ est l'ensemble des mots du lexique qui se trouvent dans une relation similaire à celle que m entretient avec p. En d'autres termes, un mot u appartient à $s_p(m)$ s'il existe un mot v tel que $m{:}p{=}u{:}v$ (i.e. tel que m, p, u, v forment une analogie). L'exemple (36) présente le filament de l'adjectif m = *modifiable* pour p = *modificateur*. Ce filament illustre l'une des caractéristiques originales de Morphonette,

à savoir qu'il décrit à la fois des relations directes et des relations indirectes, comme ici entre deux dérivés du verbe *modifier*.

(36) (modifiable, modificateur, {amplifiable, glorifiable, identifiable, justifiable, clarifiable, mystifiable, rectifiable, sanctifiable, simplifiable, spécifiable, unifiable, vérifiable})

Les informations issues de DériF et de Morphonette ont été décomposées et intégrées dans le réseau Démonette, qui décrit les relations dérivationnelles entre des couples de mots. Ces relations sont caractérisées par cinq types de propriétés : caractéristiques des entrées (graphies, catégories et types sémantiques) ; description « topologique » de la relation (orientation et complexité) ; description constructionnelle de la relation (types et exposants des procédés dérivationnels et thèmes dérivationnels) ; gloses concrètes et abstraites ; descriptions phonologiques (transcriptions API). Ces informations sont illustrées dans le tableau 1 (page suivante) pour la relation entre AMORTIR et le nom de résultat AMORTISSEMENT (2e colonne ; cette relation est issue de DériF) et entre le nom d'agent MODIFICATEUR et le nom d'action MODIFICATION (3e colonne ; cette relation est issue de Morphonette). Dans les entrées, la 1re information est la graphie, la 2e, l'étiquette morphosyntaxique au format Grace et la 3e, le type morpho-sémantique (@ = prédicat, @AGM = agent masculin, @RES = résultat, @ACT = nom d'action).

	amortir ← amortissement	modificateur ← modification
Entrée 1	amortir/Vmn----/@	modificateur/Ncms/@AGM
Entrée 2	amortissement/Ncms/@RES	modification/Ncfs/@ACT
Relation 1 ← 2	descendant/simple	transversale/simple
Construction 1		suf/eur/modificat
Construction 2	suf/ment/amortiss	suf/ion/modificat
Glose concrète	réaliser l'action dont le résultat est un amortissement	(agent masculin OR instrument) de la modification
Glose abstraite	réaliser l'action dont le résultat est @RES	(agent masculin OR instrument) de @ACT
Phono 1	amɔʁtiʁ	mɔdifikatœʁ
Phono 2	amɔʁtisəmɑ̃	mɔdifikasjɔ̃

Tableau 1 – Descriptions des relations morphologiques dans Démonette

Dans le tableau 1 (page précédente), les relations décrivent l'entrée 1 relativement à l'entrée 2. On observe que Démonette contient des relations descendantes qui, comme dans la 2e colonne connectent une base à l'un de ses dérivés, mais aussi des relations transversales entre des mots qui appartiennent à la même famille dérivationnelle. Les autres particularités sont la présence de descriptions morphosémantiques (types), la redondance des informations (chaque dérivation donne lieu à deux relations ; un lexème a autant de descriptions qu'il y a de relations dans lesquelles il apparait), la conception cumulative du sens (chaque mot a autant de gloses sémantiques concrètes et abstraites qu'il a de relations dérivationnelles) et l'indication de l'origine de chacune des informations (omises dans le tableau 1)[6]. Par ailleurs, certaines informations peuvent ne pas être renseignées si elles ne sont pas fournies par la ressource originale. Grâce à son architecture « ouverte » Démonette peut être complétée par une variété de ressources morphologiques dérivationnelles. La liste des traits morphologiques, phonologiques, catégoriels et sémantiques peut être étendue pour inclure celles d'une nouvelle ressource que l'on souhaiterait y ajouter. L'enrichissement de Démonette se fait au moyen des programmes de transfert spécifiques à chaque ressource. Démonette est distribuée sous une licence du domaine publique classique (Creative Commons).

Nous envisageons dans un proche avenir de construire une interface d'interrogation et d'extraction de collections

[6] Les graphies, les étiquettes morphosyntaxiques et les transcriptions phonologiques sont celles de la forme de citation du lexème car c'est son identifiant standard. Dans une version ultérieure, Démonette sera couplée à un lexique flexionnel et phonologique similaire à GLÀFF qui liste l'ensemble des graphies/formes phonologiques / étiquettes morpho-syntaxiques de chaque lexème.

d'exemples adaptée aux pratiques des morphologues. À terme, Démonette sera intégrée dans une application Web plus ambitieuse permettant de réaliser une grande partie des descriptions morphologiques : collecte des exemples, stockage, nettoyage, annotation et analyse.

6. Conclusion

Cette courte histoire de la morphologie extensive présente les évolutions de cette nouvelle approche, des questions qu'elle pose et des perspectives qu'elle ouvre. La question principale est assurément celle de la nature des données que la morphologie doit décrire. Quelle est la place des constructions spontanées en français non-normé, qui aujourd'hui constituent la plus grande partie des productions écrites : la multitude des communications personnelles comme les SMS, tweets, messages instantanés sur Facebook, posts sur des blogs, etc. dépassent de très loin la quantité des textes professionnels, commerciaux, journalistiques, etc. rédigés dans un français plus conforme aux normes institutionnelles. L'irruption de ces nouvelles formes de communication et de la langue qui y est utilisée est généralement totalement ignorée par les linguistes et notamment les morphologues, du fait de leur formation académique mais aussi parce qu'ils sont souvent totalement démunis face à la créativité de ces nouveaux locuteurs-rédacteurs. La seconde évolution qu'apporte la morphologie extensive concerne la place des données et l'importance de la constitution de collections étendues dans le travail des morphologues. Ces derniers doivent apprendre à manipuler et exploiter des données en grand nombre et acquérir de nouvelles méthodes de travail, plus expérimentales et probablement, à terme, plus quantitatives. Ces évolutions auront également une conséquence sur la nature des études qui deviendront plus techniques, plus longues, plus couteuses et qui imposeront de travailler en équipe.

La morphologie extensive n'est naturellement pas une évolution isolée en linguistique. La place des données change aussi dans d'autres sous-disciplines avec un recours plus important aux corpus annotés au niveau syntaxique comme le Penn TreeBank (MARCUS *et al.* 1993) ou le French TreeBank (ABEILLÉ *et al.* 2003) ou discursif comme le Penn Discourse TreeBank (PRASAD *et al.* 2008) ou le corpus Annodis (PÉRY-WOODLEY *et al.* 2009). Ces sources de données riches ont fait l'objet de nombreuses publications et peuvent sur certains aspects nous servir d'exemples pour une meilleure rentabilité de la création de collections d'exemples en morphologie. Une autre direction dans laquelle une solution serait la bienvenue est la création d'outils de moissonnage lexical du Web capables de créer des collections d'attestations génériques. La mise en place de dispositifs de ce type ne peut cependant être réalisée qu'avec un soutien important des agences nationales et européennes de financement de la recherche.

Références

ABEILLÉ Anne, CLÉMENT Lionel & TOUSSENEL François. (2003). Building a Treebank for French. In *Treebanks*, 165-187. Dordrecht: Kluwer.

ANSCOMBRE Jean-Claude & LEEMAN Danielle. (1994). La Dérivation des Adjectifs en *-ble* : Morphologie ou Sémantique ? *Langue Française 103*, 32-44.

BAAYEN R. Harald, PIEPENBROCK Richard & GULIKERS Leon. (1995). *The CELEX lexical database (release 2). CD-ROM*. Philadelphia: Linguistic Data Consortium.

BARONI Marco, BERNARDINI Silvia, FERRARESI Adriano & ZANCHETTA Eros. (2009). The WaCky Wide Web: a Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation 43-3*, 209-226.

BAUER Laurie. (2014). Grammaticality, acceptability, possible words and large corpora. *Morphology 24-2*, 83-103.

CHURCH Kenneth W. & MERCER Robert L. (1993). Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational linguistics 19-1*, 1-24.

DAL Georgette, LIGNON Stéphanie, NAMER Fiammetta & TANGUY Ludovic. (2004). Toile contre Dictionnaires : Analyse Morphologique en Corpus de Noms Déverbaux Concurrents. In *Colloque international sur les noms déverbaux*, Villeneuve-d'Ascq.

DUBOIS Jean. (1969). *Grammaire Structurale du Français : La Phrase et les Transformations*. Paris : Larousse.

FRADIN Bernard. (2003). *Nouvelles Approches en Morphologie*. Paris : Presses Universitaires de France.

FRANCOPOULO Gil, MONTE George, CALZOLARI Nicoletta, MONACHINI Monica, BEL Nuria, PET Mandy & SORIA Claudia. (2006). Lexical Markup Framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, LREC (2006), Genova, Italy.

GAWELKO Marek. (1977). *Évolution des Suffixes Adjectivaux en Français*. Wrocław : Zakład Narodowy im. Ossolińskich.

GREFENSTETTE Gregory. (1999). The World Wide Web as a Resource for Example-Based Machine Translation Tasks. In *Proceedings of the 21st ASLIB Conference on Translating and the Computer*.

HATHOUT Nabil. (2009). *Contributions à la Description de la Structure Morphologique du Lexique et à l'Approche Extensive en Morphologie*. Toulouse : Université Toulouse le Mirail - Toulouse II. Habilitation à diriger des recherches.

HATHOUT Nabil. (2011). Morphonette: a Paradigm-Based Morphological Network. *Lingue e Linguaggio 10-2*, 245-264.

HATHOUT Nabil, PLÉNAT Marc & TANGUY Ludovic. (2003). Enquête sur les Dérivés en *-able*. *Cahiers de Grammaire 28*, 49-90.

HATHOUT Nabil & TANGUY Ludovic. (2003). Webaffix : Une Boite à Outils d'Acquisition Lexicale à Partir du Web. *Revue québécoise de linguistique 32-1*, 61-84.

HATHOUT Nabil, MONTERMINI Fabio & TANGUY Ludovic. (2008). Extensive Data for Morphology: Using the World Wide Web. *Journal of French Language Studies 18-1*, 67-85.

HATHOUT Nabil, NAMER Fiammetta, PLÉNAT Marc & TANGUY Ludovic. (2009a). La Collecte et l'Utilisation des Données en Morphologie. In FRADIN Bernard, KERLEROUX Françoise & PLÉNAT Marc (Eds), *Aperçus de Morphologie du Français*. Saint-Denis : Presses Universitaires de Vincennes, 267-287.

HATHOUT Nabil, SAJOUS Franck & TANGUY Ludovic. (2009b). Looking for French Deverbal Nouns in an Evolving Web (a Short History of WAC). In *Proceedings of Web as Corpus (2009) (WAC5)*, San Sebastián.

HATHOUT Nabil & NAMER Fiammetta. (2014a). Démonette, a French Derivational Morpho-Semantic Network. *Linguistic Issues in Language Technology 11-5*, 125-168.

HATHOUT Nabil & NAMER Fiammetta. (2014b). La Base Lexicale Démonette : Entre Sémantique Constructionnelle et Morphologie Dérivationnelle. In *Actes de la 21^{e} Conférence sur le Traitement Automatique des Langues Naturelles TALN (2014)*, Marseille, 208-219.

HATHOUT Nabil, SAJOUS Franck & CALDERONE Basilio. (2014). GLÀFF, a large versatile French lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland: European Language Resources Association (ELRA).

KILGARRIFF Adam & GREFENSTETTE Gregory. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational linguistics 29-3*, 333-347.

KOEHL Aurore. (2012a). *Altitude*, *négritude*, *bravitude* ou la Résurgence d'une Suffixation. In *Actes du 3e Congrès Mondial de Linguistique Française (CMLF 2012), vol. 1*, 1307-1323.

KOEHL Aurore. (2012b). *La Construction Morphologique des Noms Désadjectivaux Suffixés en Français*. Nancy : Université de Lorraine Thèse de doctorat.

KOEHL Aurore & LIGNON Stéphanie. (2014). Property Nouns with *-ité* and *-itude* : Formal Alternation and Morphopragmatics or the sad-itude of the Aité$_{N}$. *Morphology 24-4*. 351-376.

LEEMAN Danielle. (1992). Deux Classes d'Adjectifs en *-ble*. *Langue Française 96*, 44-64.

LEEMAN Danielle & MELEUC Serge. (1990). Verbes en tables et Adjectifs en *-able*. *Langue Française 87*, 30-51.

MACWHINNEY Brian. (2000). *The CHILDES project: Tools for Analyzing Talk*. Mahwah: Lawrence Erlbaum.

MARCUS Mitchell P., MARCINKIEWICZ Mary Ann & SANTORINI Beatrice. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational linguistics 19-2*, 313-330.

NAMER Fiammetta. (2003). WaliM : Valider les Unités Morphologiques Complexes par le Web. In *Les unités morphologiques. Actes du 3e Forum de morphologie (Silexicales 3)*. Villeneuve d'Ascq : Presses Universitaires de Lille, 142-150.

NAMER Fiammetta. (2009). *Morphologie, Lexique et Traitement Automatique des Langues : L'Analyseur DériF*. Paris : Hermès Science-Lavoisier.

NAMER Fiammetta. (2013). A Rule-Based Morphosemantic Analyzer for French for a Fine-Grained Semantic Annotation of Texts. In MAHLOW Cerstin & PIOTROWSKI Michael (Eds), *SFCM 2013 CCIS 380*. Heidelberg : Springer, 93-115

NAMER Fiammetta & DAL Georgette. (2000). GÉDÉRIF: Automatic Generation and Analysis of Morphologically Constructed Lexical Resources. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens: ELRA.

NEW Boris. (2006). Lexique 3 : Une Nouvelle Base de Données Lexicales. In *Verbum ex machina. Actes de la 13e conférence sur le Traitement automatique des langues naturelles*, Louvain-la-Neuve : Presses Universitaires de Louvain.

PÉRY-WOODLEY Marie-Paule. (1995). Quels Corpus pour quels Traitements Automatiques ? *Traitement Automatique Des Langues 36-1/2*, 213-232.

PÉRY-WOODLEY Marie-Paule, ASHER Nicholas, ENJALBERT Patrice, BENAMARA Farah, BRAS Myriam, FABRE Cécile, FERRARI Stéphane, HO-DAC Lydia-Mai, LE DRAOULEC Anne, MATHET Yann, MULLER Philippe, PRÉVOT Laurent, REBEYROLLE Josette, TANGUY Ludovic, VERGEZ-COURET Marianne, VIEU Laure & WIDLÖCHER ANTOINE. (2009). ANNODIS : une Approche Outillée de l'Annotation de Structures Discursives. In *Actes de la 16e conférence sur le*

traitement automatique des langues naturelles (TALN-2009), Senlis, France.

PLÉNAT Marc. (1988). Morphologie des Adjectifs en *-able*. *Cahiers de Grammaire 13*, 101-132.

PLÉNAT Marc. (2000). Quelques Thèmes de Recherche Actuels en Morphophonologie Française. *Cahiers de Lexicologie 77*, 27-62.

PLÉNAT Marc. (2011). Enquête sur Divers Effets des Contraintes Dissimilatives en Français. In ROCHÉ Michel, BOYÉ Gilles, HATHOUT Nabil, LIGNON Stéphanie & PLÉNAT Marc (Eds), *Des Unités Morphologiques au Lexique*, Paris : Hermès, 145-190.

PLÉNAT Marc, TANGUY Ludovic, LIGNON Stéphanie & SERNA Nicole. (2002). La Conjecture de Pichon. *Corpus 1*, 105-150.

PLÉNAT Marc & ROCHÉ Michel. (2003). Prosodic Constraints on Suffixation in French. In BOOIJ Geert E., DECESARIS Janet, RALLI Angela & SCALISE Sergio (Eds), *Topics in Morphology. Selected Papers from the third Mediterranean Morphology Meeting*. Barcelone : Universitat Pompeu Fabra, 285-299.

PRASAD Rashmi, DINESH Nikhil, LEE Alan, MILTSAKAKI Eleni, ROBALDO Livio, JOSHI Aravind K. & WEBBER Bonnie L. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the sixth international conference on Language Resources and Evaluation* (LREC 2008), Marrakech.

SAJOUS Franck, HATHOUT Nabil & CALDERONE Basilio. (2013). GLÀFF, un Gros Lexique À tout Faire du Français. In *Actes de la 20e Conférence sur le Traitement Automatique des Langues Naturelles (TALN-2013)*. Les Sables d'Olonne, France, 285-298.

TALMY Leonard. (2000). *Toward a Cognitive Semantics*. Cambridge: MIT Press.

TANGUY Ludovic. (2012). *Complexification des Données et des Techniques en Linguistique : Contributions du TAL aux Solutions et aux Problèmes*. Toulouse : Université de Toulouse 2 - Le Mirail, Habilitation à diriger des recherches.

TANGUY Ludovic. (2013). La Ruée Linguistique vers le Web. Texto! *Textes et Cultures 18-4*.

TANGUY Ludovic & HATHOUT Nabil. (2002). Webaffix : Un Outil d'Acquisition Morphologique Dérivationnelle à partir du Web. In

PIERREL Jean-Marie (Ed.), *Actes de la 9e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2002)*, Nancy : ATALA, 245-254.

TANGUY Ludovic & HATHOUT Nabil. (2007). *Perl pour les Linguistes. Programmes en Perl pour Exploiter les Données Langagières*. Paris : Hermès Science-Lavoisier.

TRIBOUT Delphine. (2010). *Les Conversions de Nom à Verbe et de Verbe à Nom en Français*. Université Paris 7. Thèse de doctorat.

Cahiers de l'ILSL-45 (2016 : 161-196)

Ce que les corpus de production orale ne peuvent montrer : apports de l'oculométrie (*eye-tracking*) dans la recherche sur le bilinguisme et sur la dysphasie[1]

Elena TRIBUSHININA
Willem M. MAK
Universiteit Utrecht (NL)
Correspondance : e.tribushinina@uu.nl

1. Introduction

Nous commencerons par revenir sur le point soulevé par Steven GILLIS dans sa conférence, à savoir que dans la recherche sur l'acquisition du langage par l'enfant, on examine fréquemment le langage spontané, c'est-à-dire des corpus.

Mirjam ERNESTUS (voir page 65) a montré la nécessité d'indices convergents. Si l'étude de corpus nous apprend quelque chose, elle doit nécessairement être complétée par des recherches sur la compréhension, par exemple, et par une approche expérimentale.

Dans cette contribution nous présenterons deux études de cas portant sur un matériel linguistique précis, russe en l'occurrence, qui montrent ce qui peut se passer quand on utilise uniquement des données de corpus – c'est-à-dire à l'instar de 99% des études qui s'intéressent aux différences entre populations bilingues et populations atteintes de dysphasie. Nous montrerons ce qui arrive lorsqu'on n'utilise que ce type de données et à quelles conclusions erronées on risque de parvenir.

---

[1] *What speech production corpora cannot tell us: Insights from eye-tracking in research on bilingualism and SLI (specific language impairment).* La conférence a été présentée par Elena TRIBUSHININA. Transcription, traduction et adaptation par Guillaume Feigenwinter et Marianne Kilani-Schoch.

Pendant longtemps, on a cru à tort que le bilinguisme ou le multilinguisme étaient préjudiciables ; heureusement, on ne pense plus de cette façon aujourd'hui. Le nombre de bilingues est en augmentation, alors que dans les années 80, donner une éducation bilingue à ses enfants était considéré par certains comme nuisible. Aujourd'hui, les choses ont changé, mais pas complètement. Les parents d'enfants bilingues remarquent souvent que les enseignants, ou bien des membres du personnel médical, leur conseillent d'élever leurs enfants dans une seule langue. Ce type de conseil est fréquemment prodigué à des familles dont le niveau de formation est peu élevé. Il s'avère très dommageable, parce qu'ensuite les parents se mettent à parler à leurs enfants une langue qui n'est pas leur langue maternelle, faisant ainsi beaucoup d'erreurs.

2. Bilinguisme et dysphasie

Le bilinguisme n'est évidemment pas négatif en soi, mais présente un certain nombre de problèmes spécifiques. Des études ont montré que plus on entend une langue, plus on l'apprend vite[2]. Si on est bilingue, on n'entend donc qu'une « moitié » de chaque langue, soit deux fois moins qu'un enfant monolingue, et l'apprentissage risque de prendre un peu plus de temps. En outre, très souvent, dans les familles mixtes, les parents décident de ne parler qu'une seule langue à leur enfant, le plus souvent la langue du pays où ils habitent. Ceci veut dire qu'un des parents en tout cas ne parle pas sa langue maternelle.

Un cerveau bilingue connait évidemment des influences entre les deux langues, influences dont on ignore encore l'extension. Certains linguistes considèrent que deux systèmes linguistiques autonomes coexistent dans le cerveau bilingue, c'est-à-dire qu'on aurait affaire à deux

[2] NdE : HART & RISLEY (1995, 2003).

monolinguismes. Aujourd'hui, on sait qu'il y a en fait des influences interlinguistiques, mais leur statut n'est pas défini : s'agit-il plus ou moins d'un autre système linguistique ou s'agit-il plutôt d'un problème d'inhibition ? C'est un point qui n'est pas encore résolu.

Comme les bilingues sont différents, prennent parfois plus de temps pour acquérir un phénomène, et que leurs performances dans certains tests sont différentes, ils sont souvent diagnostiqués à tort pour un trouble du langage ou dysphasie. Le terme employé dans notre domaine est « trouble spécifique du langage » bien qu'aujourd'hui plus personne ne croie qu'il soit véritablement spécifique, me semble-t-il. Il y a quelques années, on trouvait très commode – particulièrement dans le paradigme générativiste – de montrer qu'il existe une catégorie de troubles spécifiques du langage. Il allait de soi qu'il s'agissait d'enfants « normaux » dans tous les domaines, avec un « QI normal », un « développement social normal » et un « développement émotif normal » : tout était normal, si ce n'est leur langage. Maintenant on sait que la réalité est différente : ces enfants ont en fait des problèmes au niveau des fonctions exécutives, de l'attention et des capacités motrices ; l'enfant qui souffre d'un trouble du langage ne sera jamais comme les autres enfants du point de vue du développement. Ce trouble n'est donc pas si spécifique. Néanmoins nous continuerons à utiliser le terme pour respecter la tradition. De nos jours, certains spécialistes parlent seulement d'un « trouble du langage » pour qualifier ce phénomène. Peut-être qu'un changement interviendra dans ce sens.

Ces enfants étant souvent très semblables les uns aux autres en ce qui concerne la production du langage, les logopédistes ont du mal à les différencier et à établir s'il s'agit d'un bilingue « typique » (juste bilingue, donc différent des monolingues) ou si l'enfant a un réel problème, une dysphasie. C'est un sujet très discuté dans la recherche

aujourd'hui, motivé par un vrai besoin – dans la vraie vie ! – de disposer d'outils permettant de différencier ces populations, parce que la plupart des tests de dépistage ont été développés pour des enfants monolingues. Or on ne peut pas tester les bilingues avec les mêmes instruments. Les recherches en cours sont nombreuses, notamment de grands projets COST (*European Cooperation in Science and Technology*). En général, elles s'intéressent à l'acquisition de la langue seconde, par exemple un enfant venant d'une famille turque, avec le turc pour langue maternelle, qui apprend le néerlandais ou l'allemand comme deuxième langue en entrant à l'école. Ces enfants ont parfois des problèmes linguistiques qui s'apparentent à la dysphasie, et par conséquent la morphosyntaxe de cette deuxième langue est souvent au centre de l'attention. Car les enfants atteints de dysphasie ont beaucoup de problèmes avec la morphosyntaxe, par exemple en anglais avec le *-s* final de la troisième personne du singulier, qu'ils omettent souvent. Cette omission est un symptôme de dysphasie. Depuis le début des recherches sur le sujet, l'intérêt a porté principalement sur des problèmes de ce type.

3. Méthodologie

Ici intervient le point critique du point de vue méthodologique : la production du langage. Presque toutes les conclusions concernant la dysphasie et le bilinguisme sont basées exclusivement sur la production du langage. Or il s'agit beaucoup moins de données spontanées que dans l'étude du développement « typique » du langage de l'enfant. Ce sont souvent des tâches narratives qui sont examinées. Steven GILLIS (voir page 95) a mentionné le fait que CHILDES contient une base de données entièrement constituée de récits. Pourquoi ? Ce genre de production permet d'avoir un certain contrôle sur ce que les enfants disent, et facilite la comparaison. Il y a évidemment la célèbre histoire de la

grenouille, mais on utilise aussi d'autres récits. Les enfants regardent les images, puis racontent l'histoire. C'est une méthode que nous suivons aussi, et comme les autres chercheurs, nous regardons quels types d'erreurs sont produits par les enfants, et combien. Tous les résultats qui aboutissent à la conclusion qu'il y a une similarité entre le multilinguisme et la dysphasie se basent sur ce type de données, c'est-à-dire sur des récits. Regardons ce qui se passe lorsqu'on adopte une telle approche. Un exemple, très célèbre, est *l'histoire du chat* de HICKMAN (2003).

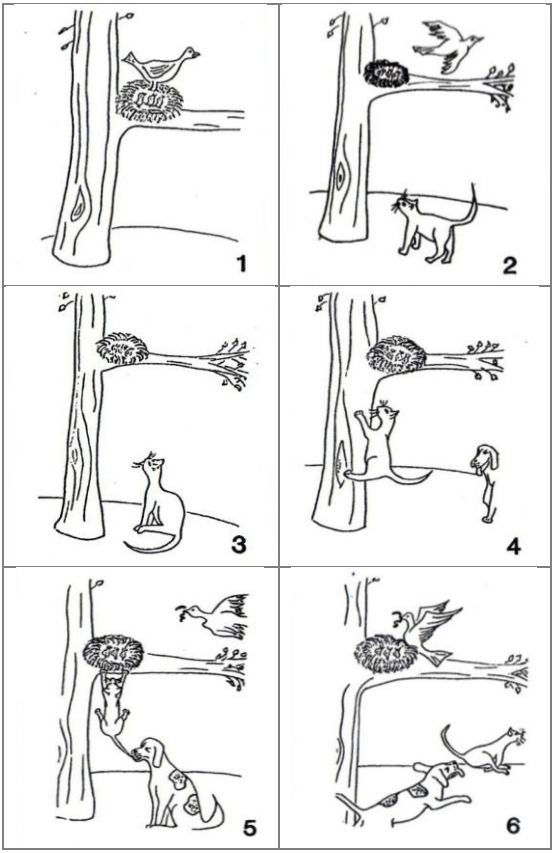

Figure 1 – L'histoire du chat (HICKMAN 2003)

Les enfants adorent cette histoire. Elle a été utilisée dans plusieurs langues, et c'est un très bon outil pour étudier la référence. En général, on commence par montrer aux enfants les six images, pour qu'ils puissent comprendre l'intégralité de l'histoire ; ensuite on montre les images une à une, et les enfants racontent ce qui se passe dans chacune des images, en établissant éventuellement des liens entre les différentes séquences. La transcription et l'analyse des données de cette histoire sont standardisées.

La comparaison de récits en russe par des bilingues 2L1 ou simultanés[3], et des monolingues dysphasiques donne une bonne idée des similitudes entre ces deux populations. Les deux groupes font des erreurs dans la conjugaison des verbes, ce qui est tout à fait étrange, en russe, pour des enfants de sept ans. Les cas et l'accord en genre constituent également des difficultés. Enfin, un problème que nous développerons ensuite est celui des connecteurs discursifs. Le russe a deux sortes de 'et' (pas seulement le russe, d'ailleurs, mais beaucoup de langues slaves), qui donnent lieu à un résultat intéressant : on les trouve très fréquemment dans la production des enfants, mais la distribution dans les deux populations est en quelque sorte inversée. Nous avons affaire ici à une différence et pas à une similitude. Or nous ne nous attendions pas à trouver ce résultat quand nous avons commencé notre recherche.

4. Production des connecteurs additifs en néerlandais et en russe par des enfants monolingues, monolingues dysphasiques et bilingues

Nous avons collecté une quantité de données, très importante pour une étude sur l'acquisition du langage

[3] Les bilingues 2L1 ou simultanés sont des enfants qui acquièrent simultanément deux langues premières, c'est-à-dire à qui on a parlé en deux langues dès leur naissance et dont on ne peut pas dire qu'une des deux langues est première et l'autre seconde.

puisque nous avons testé au total environ 1500 enfants sur deux récits, en Allemagne, en Russie et aux Pays-Bas (TRIBUSHININA, MAK, ANDREIUSHINA, DUBINKINA & SANDERS 2015). L'âge de ces enfants, bilingues ou monolingues, atteints ou non de dysphasie, s'étendait entre trois et neuf ans.

Nous allons maintenant présenter quelques-uns des résultats de cette étude sur les connecteurs discursifs, qui montrent quelles conclusions on peut espérer obtenir lorsqu'on n'utilise que des récits comme matériel de recherche.

Voyons d'abord comment fonctionne le système des connecteurs en néerlandais et en russe :

		RU		
		а	и [i]	но [no]
NL	en (*et*)	+	+	–
	maar (*mais*)	+	–	+

Tableau 1 – Connecteurs additifs en néerlandais et en russe

En néerlandais, la situation est simple : l'opposition est équivalente à l'opposition anglaise entre *and* et *but*. En russe, par contre, il y a non pas deux mais trois connecteurs qui se divisent le même espace sémantique. 'но' [no] est une sorte de *mais* négatif, utilisé uniquement dans un contexte argumentatif comme par exemple un déni d'attente : « cette bague est belle, mais chère ». 'и' [i] est positif. À première vue il parait semblable à *et* mais s'en distingue, comme on va le voir. Et finalement, 'а' est un genre de combinaison entre *et* et *mais*. Les deux connecteurs sont très courants. Mais 'а' apparait un peu plus tôt dans le langage de l'enfant, pour des raisons de saillance phonologique. Ce que nous avons découvert, c'est que 'а' et 'и' [i] sont relativement énigmatiques pour beaucoup d'enfants ; on ne s'y attendait pas.

Quelle différence y a-t-il entre ces deux connecteurs ? 'и' [i] est un marqueur de maintien du thème, réalisé dans la plupart des cas par le maintien de la référence : pour relier deux parties d'un discours en maintenant la référence, on utilise donc 'и' [i].

Voici une production d'un enfant dysphasique (1) :

(1) **И** : maintien de la référence :
Жили-были мама-птичка, **и** она родила цыплята. (SLI-075)[4]
'Il y avait une maman-oiseau *et* elle donna naissance à des oisillons'

Comme le second segment concerne le même référent, c'est-à-dire la « maman-oiseau », il faut effectivement utiliser 'и' [i] ici, l'emploi est correct. Cependant la relation n'est pas biunivoque : parfois, on peut employer 'и' [i] alors qu'il y a un changement de référence, à une condition très stricte : les deux parties du discours doivent entretenir un lien de causalité. 'и' [i] est un marqueur du maintien du thème, et dans la plupart des cas, on a aussi le maintien de la référence ; si elle change, alors il faut quelque chose d'autre pour maintenir le thème et c'est la causalité. Ainsi dans la phrase (2) le référent change, mais ce changement est possible puisque 'и' [i] signifie « *et donc* il[5] eut peur » :

(2) **И** : changement de référence (lecture causale obligatoire) :
А орёл её прямо за спину вот своим клювом схватил, **и** она испугалась. (L1-062)
'Et l'aigle piqua le renard sur le dos avec son bec *et* il eut peur.'

En anglais, en néerlandais et dans bien d'autres langues, on trouve la même présupposition pour *et,* mais celle-ci ne

[4] Comme il s'agit d'un enfant dysphasique, sa production comporte une erreur dans la flexion verbale, qui est au pluriel au lieu du singulier.
[5] NdE : traduction littérale : *elle*, *renard* étant féminin en russe.

représente pas une contrainte obligatoire, c'est-à-dire que dans ces langues, on peut changer de référence et utiliser *et* sans qu'il y ait de relation de causalité entre les propositions ; en russe, ce n'est pas possible. Un changement de référence dans ce cas est considéré comme une erreur.

Pour 'а', c'est l'inverse : en général, on trouve ce connecteur dans des contextes de changement de référence, dans des contrastes du type « A fait ceci, B fait cela » :

(3) **а** : changement de référence :
Лиса гоняется, **а** птица улетает. (L1-063)
'Le renard lui court après *et/mais* l'oiseau s'envole'

Si on veut utiliser 'а' pour maintenir la référence, c'est plus compliqué, il faut un contraste – en fait, nous n'avons aucun exemple correct dans notre corpus ; les énoncés suivants (4a et b) sont des énoncés d'adulte :

(4) **а** : maintien de la référence (contraste obligatoire) :
(4a) Утром папа читает газеты, **а** вечером он смотрит телевизор.
'Le matin papa lit les journaux, *et/mais* le soir il regarde la télévision.'
(4b) *Были птичка, ?**а** улетела птичка. (SLI-052)
'*Il y avait un oiseau, *et/mais* l'oiseau s'envola.

Souvent, le contraste est d'ordre temporel. Dans nos données, nous avons des exemples de maintien de la référence, mais sans contraste temporel, et ces énoncés sont incorrects. Si l'enfant dit « Il y avait un oiseau et l'oiseau s'est envolé », il s'agit de la même référence, il faut donc employer 'и' [i] et pas 'а'.

En résumé et en simplifiant quelque peu, 'и' [i] vaut donc pour le maintien de la référence et 'а' pour le changement, à moins d'une relation de cause à effet ou d'un contraste (voir tableau 2 page suivante).

И	А
maintien	changement
à moins d'un lien causal obligatoire	à moins d'un contraste

Tableau 2 – Connecteurs russes И et А

Comme un tel contraste ne se rencontre pas fréquemment dans le discours adressé à l'enfant et en tout cas moins fréquemment que l'usage causal, 'а' est fortement associé au changement et 'и' [i], qu'on utilise à la fois pour le maintien et le changement, est plus complexe.

En somme, ces connecteurs constituent un phénomène difficile : on a affaire à des mots très courts, très fréquents, mais qui sont quasi-synonymes. Comprendre la distinction nécessite un peu de temps. Nous avons remarqué que des populations différentes rencontrent des problèmes dans l'emploi de ces connecteurs. Des enfants de sept ou huit ans ont de la difficulté à les utiliser régulièrement sans faire d'erreur, ce qui est tard pour l'acquisition du langage puisque la morphosyntaxe est en place à trois ou quatre ans. Nous avons observé les mêmes problèmes chez des enfants dont le bilinguisme russe-allemand est successif, à savoir des enfants dont la langue maternelle est le russe et qui ont commencé l'allemand à trois ans (TRIBUSHININA, VALCHEVA & GAGARINA, à paraitre). Nous les avons relevés également chez des bilingues russe-néerlandais, éduqués simultanément dans les deux langues ainsi que chez des monolingues dysphasiques (TRIBUSHININA, MAK, ANDREIUSHINA, DUBINKINA & SANDERS, 2015). D'ailleurs, les erreurs que les enfants commettent sont très semblables. Par exemple, l'utilisation de 'и' [i] dans l'exemple (5) :

(5) Собака погнала за кошкой, ?**и** птица червячков в гнездо к птичкам. (2L1-012)
'le chien a poursuivi le chat *et* l'oiseau a ramené des vers dans son nid'

Il s'agit d'un cas de changement de référence sans relation de causalité entre les propositions : l'oiseau n'a pas ramené de ver parce que le chien poursuivait le chat. L'usage de 'и' [i] ici est une erreur. En fait, les bilingues semblent employer le connecteur russe comme s'il s'agissait du connecteur néerlandais.

Avec les enfants dysphasiques, on rencontre exactement le même type d'erreurs (voir exemple 6) :

(6) Потом пришла кошка и хотела достать, ?**и** птичка улетела.
'le chat voulait les attraper *et* (*donc*) l'oiseau s'est envolé'

Ce n'est pas ce qui se passe dans l'histoire : l'oiseau ne s'est pas envolé parce que le chat est arrivé mais parce qu'il allait chercher de la nourriture (voir figure 1-2 et 6 page 165).

On pourrait supposer que cet usage correspond à ce que l'enfant pense de l'histoire, mais nous avons vérifié ce point au préalable. Nous avons contrôlé la compréhension des liens de causalité dans l'histoire, et les enfants dysphasiques ne sont pas différents des enfants sans trouble du langage : ils comprennent parfaitement ce qui se passe. Chaque fois qu'un enfant commettait une erreur dans un récit, nous l'interrogions sur la raison de l'action qu'il venait d'évoquer. Dans 94% des cas les enfants avaient compris l'histoire. Leurs réponses aux questions étaient correctes, ce qui veut dire qu'ils avaient compris les liens conceptuels de l'histoire, mais qu'ils ne les avaient pas exprimés correctement dans le récit. Le problème est un problème d'appariement avec la forme linguistique.

Considérons maintenant plus en détail cette étude portant sur la production des connecteurs (TRIBUSHININA, MAK, ANDREIUSHINA, DUBINKINA & SANDERS 2015). Comme toute étude typique d'acquisition du langage chez les bilingues et les enfants dysphasiques, celle-ci inclut, en plus

des deux catégories d'enfants du même âge, un groupe d'enfants monolingues au développement typique (voir tableau 3 ci-dessous).

GROUPE	N	ÂGE MOYEN	TRANCHE D'ÂGE
2L1	20	8;5	8;0-8;10
L1-DT*	20	8;5	8;0-8;11
L1-D**	20	8;5	8;0-8;11

Tableau 3 – Production des connecteurs : participants

* Développement Typique ; ** Dysphasie

La méthode utilisée consiste donc à susciter des récits (deux par enfant) à partir de l'histoire du chat (HICKMANN 2003) et de l'histoire du renard (GÜLZOW & GAGARINA 2007).

L'histoire du renard, développée par nos collègues de Berlin, est une sorte de contrepartie à *l'histoire du chat*. Elle lui ressemble sur plusieurs points, comme le nombre de personnages, le genre grammatical des personnages – en allemand et en russe – ou la structure de l'information (voir les illustrations de la figure 2 de la page suivante). Nos collègues ont fait de leur mieux pour que les deux récits concordent.

Les données ont été transcrites selon CHILDES (MACWHINNEY 2000) et annotées morphologiquement au moyen de MORCOMM (GAGARINA, VOEIKOVA, & GRUZINCEV 2003).

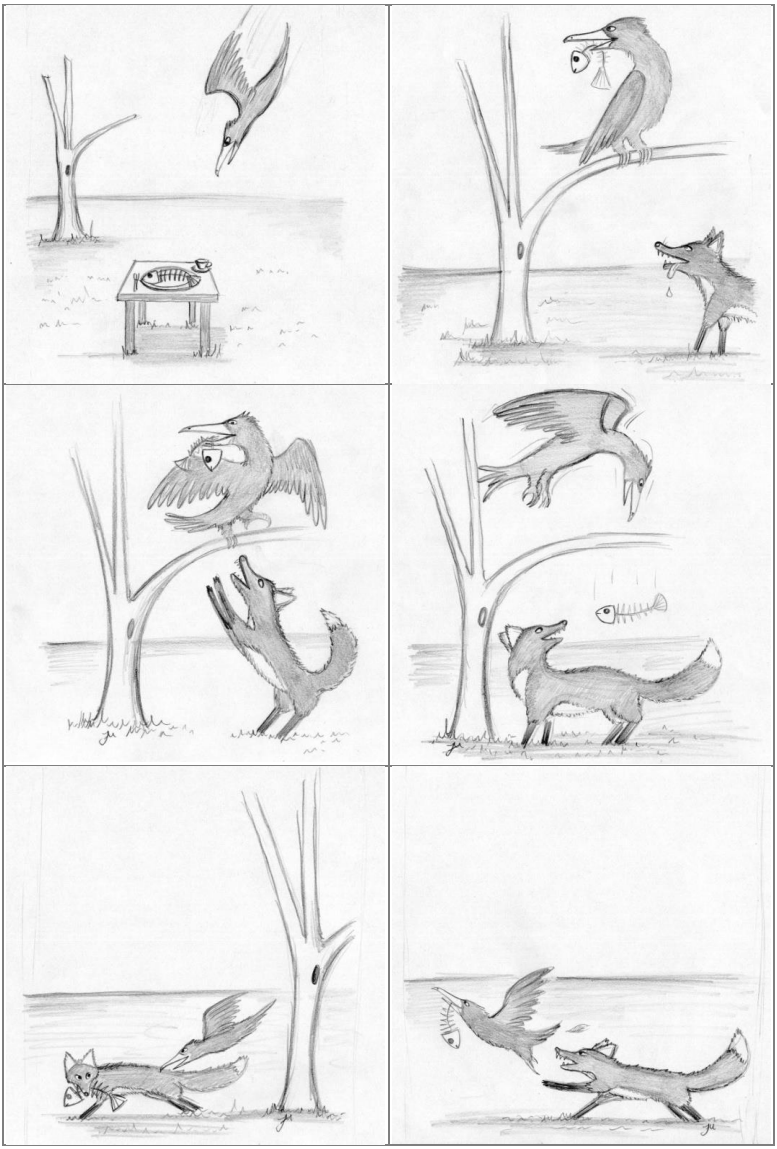

Figure 2 – L'histoire du renard (GÜLZOW & GAGARINA 2007)

Les résultats sur l'usage des connecteurs discursifs 'и' [i] et 'а' montrent une proportion semblable d'erreurs chez les enfants bilingues et chez les enfants dysphasiques. Quant aux réponses des monolingues au développement typique, elles ne sont pas non plus parfaites. Il s'agit donc bien d'un phénomène exigeant pour les enfants (voir figure 3 ci-dessous, et tableau 4, page suivante).

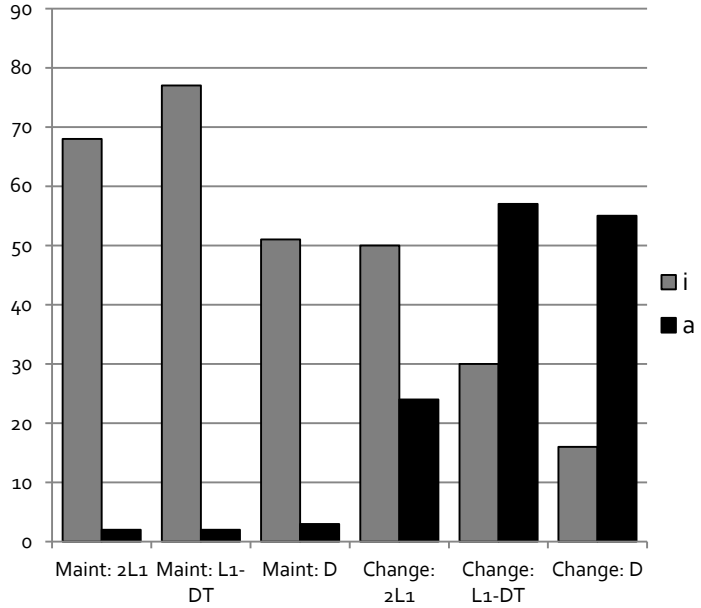

Figure 3 – Production des connecteurs : Résultats[6]

DT : Développement Typique ; D : Dysphasie ; Maint : maintien ; Change : changement

Comme nous l'avons vu plus haut, l'usage de 'и' [i] représente une difficulté pour les bilingues parce qu'ils l'emploient de la même manière qu'en néerlandais et l'étendent aux contextes de changement sans lien de causalité, ce qui n'est pas acceptable en russe.

[6] NdE : voir aussi TRIBUSHININA, MAK, ANDREIUSHINA, DUBINKINA & SANDERS (2015).

De même, les enfants dysphasiques (D) utilisent plus souvent 'и' [i] dans le cas d'un changement sans lien causal que les enfants monolingues au développement typique (DT). Signalons par ailleurs que ces enfants dysphasiques montrent la même fréquence de distribution des deux connecteurs que les enfants monolingues au développement typique.

Ainsi, lorsque l'on considère le nombre et le type des erreurs, les deux groupes d'enfants bilingues et d'enfants dysphasiques ne peuvent être distingués :

GROUPE	% ERREURS
2L1	21
L1-D	26
L1-DT	12

Tableau 4 – Production des connecteurs : taux d'erreurs (TRIBUSHININA, MAK, ANDREIUSHINA, DUBINKINA & SANDERS 2015)

Dans la mesure où les enfants bilingues font le même genre d'erreurs que les enfants monolingues dysphasiques (voir exemples 7 et 8), on comprend que les logopédistes qui utilisent souvent des récits pour leurs dépistages soient amenés à suspecter qu'un enfant bilingue est atteint d'un trouble du langage.

(7) Там птица приземлилась, ?**и** собака за кошкой побежала. (BR-006)
'L'oiseau se posa *et* le chien se mit à poursuivre le chat.'

(8) И собака ее прогнала. ?**И** птичка … она накормила птенцов. (SLI-082)
'Et le chien le poursuivit. *Et* l'oiseau nourrit les oisillons.'

Sur la base de telles données, les deux groupes ne peuvent être différenciés ni quantitativement, ni qualitativement. Or, répétons-le, c'est sur ce type de données que se fondent la plupart des recherches dans le domaine de la dysphasie. Ajoutons que nous n'avons choisi ici qu'un exemple, mais que l'examen des récits dans leur intégralité conduit au même constat.

On pourrait se demander ce qui rend une telle situation possible. Celle-ci ne révèle-t-elle pas quelque chose du système linguistique des bilingues, à savoir qu'ils auraient un système différent dans leur cerveau, à cause de l'interaction entre les langues ? Ou s'agit-il plutôt d'un problème très commun, consistant dans la difficulté à désactiver la ou les autres langues connues. Très souvent, dans l'usage d'une langue étrangère ou d'une troisième langue, le locuteur est tout à fait conscient d'une erreur commise, et, en tant que linguiste, est même surement capable d'attribuer cette erreur à une propriété particulière de la langue dominante, en général la langue maternelle ; néanmoins il est incapable de l'empêcher. Cela vient du fait que les langues sont en compétition dans le système linguistique et qu'aucune n'est jamais tout à fait inactive. Ainsi l'autre langue du répertoire (ou les autres) peut-elle causer des erreurs. Il s'agit alors d'un problème d'inhibition : cette autre langue n'a pas été désactivée (ARGYRI & SORACE 2007, MONTGOMERY & LEONARD 1998, UNSWORTH & HULK 2008).

Si on regarde la façon dont les enfants bilingues traitent le langage, on peut faire l'hypothèse qu'ils connaissent les différences entre 'и' [i] et 'а' mais que parfois ils recourent au néerlandais en même temps qu'ils parlent russe. Car comme on l'a vu, en néerlandais, on a deux options pour *et*, tandis qu'en russe on en a qu'une seule et elle est très strictement contrainte. Donc ce que font les bilingues dans leur production russe, c'est en quelque sorte de se calquer sur le connecteur néerlandais qui est polyvalent. Le russe devient alors souvent agrammatical. Ce problème n'apparait probablement que dans la production du langage et représente une sorte de cout de traitement ou de compétition, autrement dit un problème d'inhibition.

La littérature fournit seulement des indices très indirects des différences entre les deux groupes d'enfants. Il faudrait tester une tâche de traitement du langage qui permette de

les faire apparaitre. Une récente étude sur la flexion verbale en anglais L2 portant sur des bilingues (turc L1, anglais L2) a comparé les résultats obtenus avec ceux d'une étude plus ancienne sur les enfants dysphasiques, effectuée à partir des mêmes tests (MARINIS & CHONDROGIANNI 2011).

L'étude de MONTGOMERY & LEONARD (1998) montrait que les enfants dysphasiques font des fautes dans l'usage des morphèmes de temps, mais qu'en plus ils sont incapables en réception d'identifier des éléments grammaticalement incorrects. Dans un exercice où on leur demandait de dire quelle phrase était correcte et quelle phrase ne l'était pas, ils n'étaient pas sensibles à la différence. Il n'est donc pas surprenant qu'ils fassent des erreurs en production. Dans l'étude de MARINIS & CHONDROGIANNI (2011) sur des enfants bilingues, on trouve aussi des erreurs de production, mais ces enfants sont capables de reconnaitre les phrases grammaticalement incorrectes. Il y a donc une différence : les bilingues se rendent compte qu'ils font des erreurs, alors que les enfants dysphasiques n'en ont probablement pas conscience. Juger de la grammaticalité d'un énoncé est une tâche métalinguistique complexe pour les enfants, et encore plus pour des enfants dysphasiques. Très souvent, ce type d'expérimentation exige une réponse motrice, comme appuyer sur tel ou tel bouton selon que l'item est correct ou incorrect. Or comme nous l'avons vu plus haut, les enfants dysphasiques ont également des déficits moteurs et sont plus lents dans leurs réactions. Ce n'est donc pas une manière adéquate et subtile de procéder.

5. Traitement des connecteurs et oculométrie

Voilà pourquoi nous avons décidé de faire une expérience dans le paradigme du monde visuel en utilisant l'oculométrie, le but étant de voir si on peut distinguer des profils différents entre les groupes. Expliquons très brièvement ce qu'est le paradigme du monde visuel et quelles sont ses hypothèses de

base : dans ce paradigme, on suppose qu'à l'écoute d'un discours, une sorte de monde se dessine dans l'esprit, et que si l'on transpose visuellement certains aspects de ce monde mental sur un écran, on peut mesurer l'attention et comment elle se porte sur certains points de ce monde. Un autre principe de base est que lorsqu'on entend quelque chose, l'attention est captée et l'on est mentalement focalisé sur cet élément, de telle sorte que si celui-ci est présent dans le monde visuel, alors il est très difficile de ne pas le regarder. Comme il nous arrive dans nos expériences de prendre des étudiants qui servent de contrôles, l'un d'entre eux nous a dit une fois après l'expérience qu'il n'avait pas regardé là où il devait, qu'il avait pensé qu'un enfant de trois ans ne connaissait pas ces mots et donc porté son attention sur d'autres. Néanmoins, quand nous avons examiné son test, nous avons constaté que, alors même qu'il faisait de son mieux pour ne pas regarder la cible, il commençait d'abord par elle, et ensuite seulement en regardait une autre. Donc si on pense à un élément, et que celui-ci est présent dans le monde visuel, la tendance est de le regarder. C'est la raison pour laquelle on peut aussi employer ce procédé avec les bébés.

Nous n'entrerons pas dans le détail des différents types d'oculomètres comme ceux que l'on monte sur la tête et qui sont utilisés pour des recherches précises sur la lecture. Pour les expériences sur le monde visuel avec des enfants – et des adultes – on utilise d'ordinaire un oculomètre à distance qui capte la lumière réfléchie par les yeux, et qui enregistre les fixations et les focales.

L'oculométrie repose sur la présomption que l'on regarde un endroit spécifique ou que l'on tend à le regarder, car ce n'est pas toujours le cas. Des scores de 100% ne sont jamais atteints dans les données d'oculométrie. Dans les expériences avec les enfants, ceux-ci ne sont pas astreints à des tâches particulières, ils ne font qu'écouter et regarder. Ils

peuvent regarder ce qu'ils veulent, mais s'ils sont réactifs, alors apparaissent des différences entre groupes en fonction des conditions de l'expérience. Cela ne veut pas dire non plus que s'ils ne regardent pas tel élément dans une proportion de 100%, il y ait un problème de compréhension. Ce n'est pas une tâche de compréhension, mais une expérience de traitement portant sur la relation entre l'attention, le regard et les stimulus linguistiques qui guident cette attention. Un tel paradigme présente l'avantage de pouvoir être utilisé avec des enfants très jeunes ; ils n'ont rien à faire, ils restent assis sur les genoux de leurs parents et regardent des images. Par ailleurs, c'est une sorte d'étude longitudinale, puisque nous pouvons voir comment leur regard se développe au fil du temps.

Nous avons commencé par étudier des adultes (MAK, TRIBUSHININA & ANDREIUSHINA 2013) parce que la question du 'и' [i] et du 'a' est bien connue en sémantique russe, mais en sémantique des années 70, ce qui veut dire sans le support d'études de corpus. Il nous a donc fallu d'abord vérifier si ces distinctions sémantiques existent bel et bien et c'est ainsi que nous avons considéré ce que font les adultes. Nous avons tout d'abord examiné les mots 'en' et 'maar' dans le Corpus Oral du Néerlandais, puis les mots 'и' [i] et 'a' dans le Corpus National Russe. Il est intéressant de noter que nous avons trouvé les mêmes fréquences. La proportion de maintiens après 'и' [i]/'en' (*et*) ainsi que la proportion de changements de référence après 'a'/'maar' (*mais*) est la même dans les deux langues. Si nous étions de purs adeptes de la théorie basée sur l'usage, ce que nous sommes presque, nous dirions, en nous basant sur la fréquence, que dans les deux langues, on peut prédire un changement de référence après 'a'/'maar' (*mais*) et un maintien de référence après 'и' [i]/ 'en' (*et*). Mais ce n'est pas ce que l'expérience a mis en évidence.

Considérons les conditions du test. D'abord les sujets voient deux images d'animaux. En raison de l'absence

d'articles en russe, ces animaux sont désignés par des noms propres pour permettre la comparaison entre les langues. Ensuite, on leur fournit soit un contexte de maintien, soit un contexte de changement de référence, avec 'и' [i]/'en' (*et*) ou avec 'a'/'maar' (*mais*), par exemple : « Girafe lit un livre et...elle écoute de la musique ». Une petite pause semblable à celles que l'on trouve dans la production naturelle est introduite après la conjonction pour donner au sujet le temps de regarder. L'exemple précédent est un cas de maintien de la référence. Il en est de même de la phrase « Le matin, Chien boit du café, mais le soir il boit du thé ». Dans cette phrase, après le mot *Chien*, les sujets regardent le chien, évidemment ; mais que font-ils lorsqu'ils entendent le connecteur ? Dirigent-ils leur regard vers Girafe avant que la seconde partie de la phrase ne commence ? Cette probabilité est-elle plus élevée quand le connecteur est 'a'/'maar' (*mais*) que lorsqu'il s'agit de 'и' [i]/'en' (*et*) ?

Dans les phrases caractérisées par un contraste mais avec une continuité au niveau de la référence, les sujets adultes dirigent leur regard vers l'autre référent. Qu'il n'y ait qu'un contraste dans la continuité ne semble pas important ; ce qui compte le plus, c'est la fenêtre temporelle entre la quatrième seconde correspondant à la prononciation du connecteur et la cinquième seconde, début de la seconde partie de la phrase fournissant un stimulus explicite du référent. La question que nous avons soulevée est la suivante : que se passe-t-il durant cette fenêtre temporelle cruciale entre le début du connecteur et le début de la seconde partie de la phrase ?

Les résultats du néerlandais montrent qu'il n'y a pas de différence entre les deux connecteurs 'en' (*et*) et 'maar' (*mais*) et qu'ils sont traités de façon tout à fait identique. On ne s'attend donc pas à ce que les locuteurs du néerlandais marquent une différence au niveau de la fixation du regard.

Précisons que la tendance dans les deux groupes d'adultes était de détourner le regard aussitôt que le connecteur était prononcé. Mais en néerlandais, ce mouvement était exactement le même avec les deux connecteurs, alors qu'en russe, nous avons observé une nette différence. Les sujets tendaient effectivement à détourner très vite leur regard vers l'autre image, et cette tendance était significative, alors que pour 'и' [i], les sujets ne détournaient leur regard qu'au début de la seconde partie de la phrase. Il y a donc une différence dans le traitement de ces deux connecteurs.

6. Traitement des connecteurs par des enfants monolingues dysphasiques et bilingues

Ces résultats nous ont amenés à tenter la même expérience avec les enfants pour voir ce qui se passe au niveau du traitement dans les deux populations d'enfants bilingues et dysphasiques (voir tableau 5 ci-dessous), qui, comme nous l'avons dit, ont beaucoup de problèmes avec ces connecteurs (MAK, TRIBUSHININA, LOMAKO, GAGARINA, ABROSOVA & SANDERS soumis) (voir également TRIBUSHININA, DUBINKINA & SANDERS 2015).

GROUPE	N	ÂGE MOYEN	TRANCHE D'ÂGE
2L1	23	5;9	5;0-6;11
L1-DT	29	5;10	5;2-6;7
L1-D	20	6;3	5;3-7;0

Tableau 5 – Traitement des connecteurs : participants

Considérons d'abord les résultats des monolingues russes au développement linguistique typique (voir figure 4 page suivante)[7].

[7] NdE : voir également TRIBUSHININA, MAK, ANDREIUSHINA, DUBINKINA & SANDERS (2015).

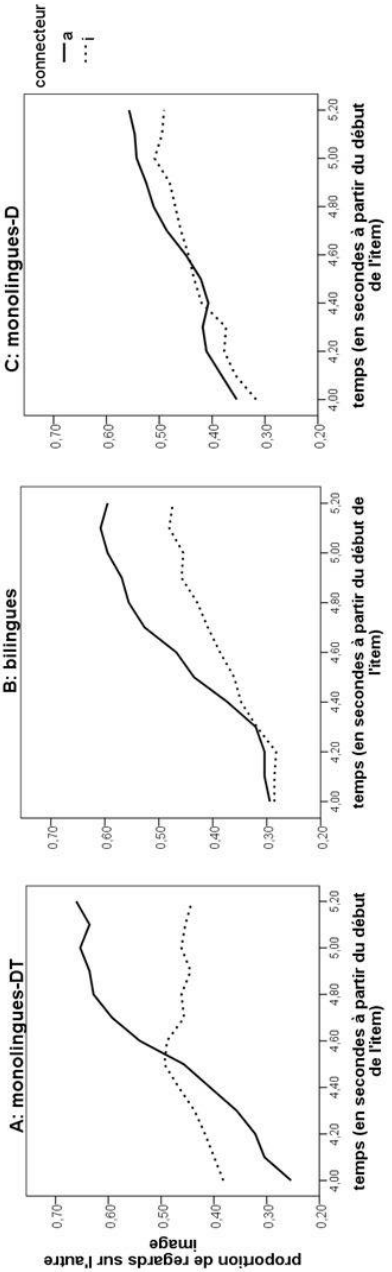

Figure 4 – Traitement des connecteurs : Résultats

La ligne continue représente 'a' (*mais*), c'est-à-dire le changement de référence. On constate que l'attention visuelle des enfants s'est portée sur l'autre image. Ils ont passé beaucoup plus rapidement à l'autre image avec 'a' (*mais*) qu'avec 'и' [i] (*et*) ; l'accélération est la plus forte dans la zone critique comprise entre quatre et cinq secondes, entre le début du connecteur et le début de la seconde proposition. Le même patron apparait chez les bilingues. Avec 'и' [i] (*et*), le mouvement en direction de l'autre référent n'est pas du tout aussi rapide et l'augmentation de la courbe sur le graphique est plus faible. Ces deux groupes se révèlent donc sensibles aux différences sémantiques entre les connecteurs. Il y a une interaction entre le type de connecteur et la fixation du regard.

Ce n'est pas le cas du groupe d'enfants russes dysphasiques. Ces enfants ne sont pas sensibles aux différences sémantiques. Ils se comportent comme les sujets néerlandais monolingues de l'expérience et ne perçoivent pas de différence sémantique entre les deux connecteurs.

On peut déduire de ces résultats la conclusion suivante : lorsqu'on ne considère que la production du langage, c'est-à-dire des récits, comme cela se pratique généralement dans le domaine aujourd'hui, on ne peut pas différencier les bilingues des enfants dysphasiques. En revanche, en recourant à des mesures plus fines du traitement réceptif, une différence apparait clairement : en termes de production, les bilingues sont très similaires aux enfants atteints de dysphasie, mais en termes de traitement, ils fonctionnent comme des monolingues (sans trouble du langage).

7. Genre pronominal

Maintenant, abordons brièvement un autre exemple qui révèle exactement les mêmes résultats chez les bilingues. Il s'agit du genre pronominal. Tout comme l'allemand, le russe comporte trois genres pronominaux, grammaticalement

déterminés, par exemple 'рыба' (*poisson*) requiert un pronom féminin (она), стол (*table*) un pronom masculin (он), et солнце (*soleil*) un pronom neutre (оно). En néerlandais des Pays-Bas[8], le genre – qui a connu des changements – tend aujourd'hui à être plus ou moins sémantiquement déterminé. En général, on utilise le pronom masculin 'hij' pour référer à des objets, quel que soit le genre grammatical. Si le référent est un être animé dont on sait qu'il est de sexe féminin – le poisson rouge d'un dessin animé, par exemple – alors on utilise le pronom féminin 'zij'. Et s'il s'agit du poisson que l'on a dans l'assiette, alors c'est en général le neutre 'het'. Le système est donc différent.

Nous avons regardé ce que font les bilingues en néerlandais et en russe, en nous basant sur les histoires présentées plus haut (*l'histoire du renard* : renard, oiseau, poisson ; *l'histoire du chat* : maman-oiseau, chat, chien). Septante-sept enfants monolingues néerlandais et septante-quatre bilingues néerlandais-russe ont participé à cette étude. Les résultats en néerlandais sont très intéressants. Comme nous l'avons vu, il y a une maman-oiseau dans cette histoire. Mais les enfants monolingues néerlandais de notre recherche, entre quatre et six ans, utilisent partout 'hij' (*il*), même pour la maman-oiseau. À six ans, l'usage du pronom masculin est généralisé, bien qu'ils sachent qu'avec une fille, on utilise 'zij' (*elle*) : dans ces histoires, la maman-oiseau tend à être traitée avec un pronom masculin. D'ailleurs pour les adultes, qui ne sont pas présentés ici, il y aurait environ 90% de masculin 'hij' (*il*), soit presque pour tous les référents, sauf la maman-oiseau.

[8] Le néerlandais de Belgique est différent.

GROUPE	PRONOMS MASCULINS EN %
Monolingues	99
Bilingues	90

Tableau 6 – Genre pronominal en néerlandais : production

La production des bilingues est à cet égard meilleure que celle des monolingues, probablement parce que le russe leur procure l'avantage de savoir ce qu'est le genre : les bilingues font en néerlandais ce que font les adultes monolingues : ils utilisent le féminin 'zij' (*elle*) pour parler de la maman-oiseau (voir tableau 6 ci-dessus). En fait, ils utilisent aussi quelques féminins de manière étrange, pour des personnages qui devraient être masculins en néerlandais. Ces erreurs peuvent venir du russe ou du néerlandais parlé par leurs parents.

Voyons ce qui se passe pour le russe en production. Les résultats des enfants monolingues russes concernant le genre pronominal montrent 100% de réponses correctes, dès l'âge de quatre ans (voir la ligne noire continue sur la figure 5 ci-dessous). La littérature dirait que c'est parce que le genre pronominal est aussi saillant qu'il est acquis à partir de trois ou quatre ans.

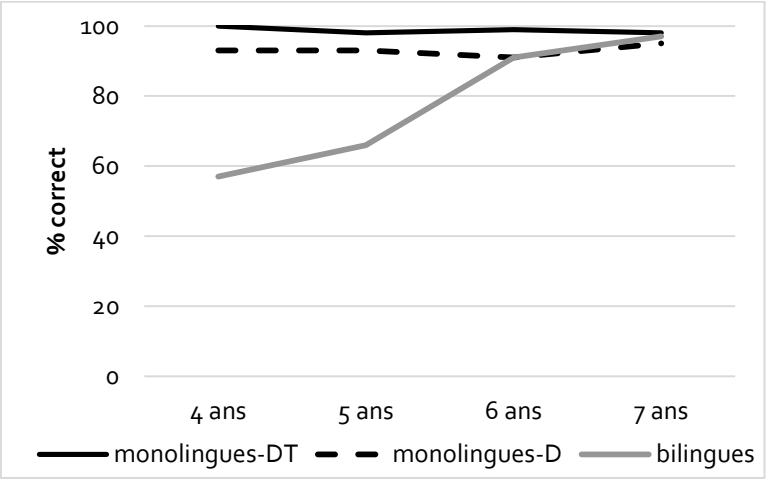

Figure 5 – Genre grammatical : production

Les enfants dysphasiques (D) font également de bons scores (ligne pointillée de la figure 5 page précédente) ; dans 90% des cas le genre est correct. Mais par la suite ils restent à ce niveau et on ne voit aucun développement.

La partie la plus intéressante de ces résultats vient des bilingues (ligne grise de la figure 5 page précédente). Quand les bilingues ont quatre ou cinq ans, ils semblent plutôt deviner les réponses et n'obtiennent que 50% d'items corrects. Par la suite, toutefois, ils atteignent la cible. Le message aux parents d'enfants bilingues peut être optimiste : si cela leur prend un peu de temps, les enfants bilingues parviennent néanmoins à la maitrise du genre. Pour la recherche, la période la plus intéressante à étudier est celle où ils tâtonnent dans leur production. La question est de savoir si à ce moment-là ils sont déjà sensibles à la nature grammaticale de la distinction de genre.

Nous avons vérifié ce point en utilisant le même paradigme du monde visuel dans un autre type d'expérience : deux personnages sont présentés aux enfants dans la première partie de la phrase – par exemple, un singe et un serpent ; parfois ces personnages sont du même genre, ce qui rend le pronom ambigu, par exemple s'ils sont tous deux féminins, le genre est ambigu et l'enfant ne sait pas où regarder – en fait, il devrait regarder les deux personnages :

(9) Змея навещает обезьяну. **Она** играет с мячом.
'Serpent rend visite à Singe. Il[9] joue à la balle.'

Dans d'autres cas, un personnage est féminin, l'autre masculin ; la distinction de genre est donc informative. On s'attend alors à ce qu'en entendant le pronom, les enfants regardent le référent du nom féminin.

[9] NdE : féminin en russe.

Dans l'exemple suivant figurent un poisson, nom féminin, et une girafe, nom masculin :

(10) Рыба приветствует жирафа. **Он** надувает красный шарик.
'Poisson salue Girafe. *Il* souffle dans un ballon rouge.'

Les enfants entendent le pronom masculin et s'ils sont sensibles au genre, ils fixeront leur regard sur la girafe ; la proportion de regards portés à la girafe doit être supérieure à celle des regards portés au poisson.

Les résultats préliminaires que nous avons obtenus pour ces enfants montrent que dans la fenêtre temporelle qui nous intéresse, c'est-à-dire, comme nous l'avons dit, celle qui intervient avec le début de la réalisation du pronom, dans la condition informative, les monolingues commencent par regarder le référent correct. Et c'est exactement ce que font aussi les bilingues. Nous avons vu avant qu'à cet âge (quatre ou cinq ans), 50% de la production de ces enfants seulement est grammaticalement correcte en russe et qu'elle semble assez aléatoire. Les résultats de cette expérience d'oculométrie montrent qu'en réalité la connaissance de ces enfants ne l'est pas. Parfois, ces enfants procèdent à la manière néerlandaise parce que le néerlandais est leur langue dominante, la langue qu'ils parlent à l'école avec leurs camarades, notamment. Mais cela n'empêche pas qu'ils sachent très bien ce qu'est le genre grammatical et soient sensibles aux indices de genre, comme on le voit très clairement dans ces tâches qui mesurent le traitement de manière plus subtile.

On peut se demander ce qui se passe quand deux noms féminins suivis d'un pronom féminin sont présentés. C'est un fait connu dans la littérature, et aussi dans la littérature qui concerne les adultes, que la tendance est de regarder le référent du sujet ; le sujet est plus activé que l'objet, si bien qu'on le regarde davantage. Dans la condition informative,

les deux groupes d'enfants ont regardé la cible. Ils utilisent donc bien des indices de genre dans le traitement réceptif, alors même que leur performance en production apparait comme plus aléatoire.

8. Conclusion

Ces deux études montrent que s'il est très important de savoir comment les enfants s'expriment dans des contextes plus ou moins spontanés, dans ce cas particulier, limiter la recherche à la production de récits conduit à des conclusions erronées sur le bilinguisme. La combinaison des deux méthodes est une nécessité pour bien comprendre ce qui se passe avec ces enfants. De toute évidence, l'oculométrie est particulièrement désignée pour cette tâche puisque, comme nous l'avons vu, c'est une méthode qui ne demande aucune action de la part des enfants. Les sujets de notre recherche avaient entre quatre et sept ans, on peut donc recourir à une telle méthode avec des enfants vraiment très jeunes.

Un autre point, très important, concerne le problème de la comparaison d'individus dysphasiques avec des monolingues ou des bilingues qui ne sont pas atteints d'un trouble du langage. Comme on l'a vu, lorsqu'ils doivent appuyer sur un bouton, les enfants dysphasiques réagissent lentement. Il existe même une hypothèse de « ralentissement général ». On pourrait donc dire qu'il est évident que ces enfants ne peuvent pas réagir à ces connecteurs aussi vite que des enfants au développement typique parce qu'ils seraient lents. Le recours à l'oculométrie permet d'établir que cette hypothèse n'est pas correcte. Nous avons regardé à quelle vitesse les enfants dysphasiques regardent le sujet, par exemple lorsqu'ils entendent « Girafe ». Il n'y a aucune différence avec les enfants qui n'ont pas de trouble du langage. Si l'expérience avait demandé une réponse motrice, ils auraient effectivement été plus lents, mais l'oculométrie montre qu'ils dirigent leur regard vers la bonne image aussi

rapidement que les enfants qui ne sont pas atteints d'un trouble du langage. Donc la vitesse, dans ce cas, est véritablement un problème lié aux connecteurs. D'autres études comparables ont montré que les enfants dysphasiques étaient capables de prédire un nom d'après un verbe. Par exemple si on leur présente la séquence « traire leur… » et ensuite une vache et une chaussure, ces enfants regardent la vache aussi rapidement que les autres enfants (ANDREU, SANZ-TORRENT & TRUESWELL 2013).

Le paradigme du monde visuel, l'oculométrie en l'occurrence, est donc une très bonne mesure pour ces enfants dysphasiques dont la vitesse de regard est normale. On ne peut avoir la même comparabilité entre les groupes en se basant sur des réponses motrices. Or, étonnamment, l'hypothèse du ralentissement se base exclusivement sur des mesures motrices.

Remerciements

Cette recherche a été subventionnée par un fonds *Marie Curie International Research Staff Exchange Scheme* dans le cadre du 7e programme cadre de la Communauté européenne (subvention numéro 269173).

Questions

Légende : « Q » pour « Question », « ET » pour « Elena TRIBUSHININA »

Q : Vous avez parlé des connecteurs, du genre, mais est-ce que vous avez des données sur des phénomènes linguistiques qui reflètent des différences typologiques majeures entre le russe et le néerlandais, comme par exemple les cas, les aspects, les articles, etc. Ou bien est-ce que ça ne faisait pas partie de vos objectifs ?

ET : Non, ce n'était effectivement pas notre but, parce que nous nous sommes principalement intéressés à la cohérence du discours.

Comparer les cas n'aurait pas eu beaucoup de sens, puisqu'il n'y a pas de cas en néerlandais, du moins en synchronie. Historiquement c'est autre chose. Pour les articles, on ne peut pas non plus comparer puisqu'il n'y a pas d'articles en russe.

Q : Oui, mais c'est ma question : est-ce qu'il y a une influence dans ces cas ? Est-ce qu'un élément spécifique à une langue est remplacé dans l'autre, par exemple ?

ET : Oui, nous nous sommes intéressés à cette question. Ce que l'on remarque, c'est qu'il y a des influences. Je dois d'ailleurs vous dire que le néerlandais des bilingues est vraiment très bon, à tel point qu'il est difficile de percevoir des différences avec les enfants monolingues. Mais parfois, on en voit d'intéressantes, en effet. Par exemple, en russe, les bilingues tentent parfois de trouver des moyens d'exprimer le fait qu'un nom est défini. Mais à vrai dire, cela n'arrive pas souvent. D'autres fois, ils tentent d'utiliser *ces* ou *un* plus souvent, parce qu'ils n'ont pas accès aux articles, mais ce n'est pas non plus constant ; il existe une différence statistiquement significative, mais il faut bien la chercher. Et dans l'autre sens, il arrive aussi, chez des enfants plus jeunes, qu'ils omettent l'article en néerlandais ; nous avons remarqué cette tendance et je crois que c'était aussi statistiquement significatif. Pour l'acquisition des cas en russe, on sait que cela leur prend beaucoup de temps, et tout le monde n'y arrive pas, pour autant que l'on sache.

Q : Et l'aspect ? D'ordinaire, c'est ce qu'il y a de plus difficile à apprendre.

ET : À vrai dire, je n'ai pas du tout considéré ce point. Nous n'avons pas regardé les cas non plus ; nous avons des collègues à Amsterdam qui s'y intéressent, mais nous nous sommes occupés des articles, et là il y a une différence. Nous avons aussi vu des différences dans l'ordre des mots. En russe, on peut se servir de l'ordre des mots pour définir un nom, et c'est une chose à laquelle les enfants bilingues sont moins sensibles, apparemment. Dans la production de ces récits, ils étaient moins enclins à utiliser l'ordre des mots à cet effet. Donc il y a une différence.

Q : Y a-t-il d'autres techniques que vous avez considérées ou employées pour étudier le traitement et la réception, et qui présentent peut-être des défauts par rapport à ce que vous utilisez.

ET : Nous sommes très satisfaits de ce paradigme du monde visuel. Nous l'avons employé avec des objectifs variés, et nous avons aussi la chance d'avoir un laboratoire d'oculométrie performant[10] et des oculomètres mobiles. Pour ce que nous en faisons, nous en sommes très satisfaits. J'ai essayé une fois une expérience de traitement chez l'adulte avec un électroencéphalogramme, mais sans beaucoup de succès. Je pense que c'était une question sémantique trop spécifique. Mais vous parlez d'expériences en temps réel ? Parce que nous avons fait une série d'expériences sur la compréhension, dont Steven GILLIS a parlé, mais il s'agissait d'expériences en temps différé. Dans notre laboratoire certains collègues pratiquent la méthode de la 'lecture à son propre rythme' (*self-paced reading*) et observent aussi comment les enfants apprennent à lire, comment les marqueurs de cohésion affectent leur lecture, s'ils reviennent en arrière quand ils rencontrent un marqueur de cohésion, ou bien la vitesse à laquelle ils lisent. On peut également combiner la lecture à son propre rythme avec l'oculométrie et plus spécifiquement suivre le trajet du regard sur certains mots. Ce type de méthode donne aussi des résultats intéressants, mais pas pour la tranche d'âge qui m'intéresse, elle s'applique surtout à des enfants plus âgés. Enfin, je connais des chercheurs qui travaillent avec des bébés en utilisant l'écoute préférentielle et le paradigme de la tomographie, c'est-à-dire l'imagerie spectroscopique proche infrarouge (*near-infrared spectroscopy* ou NIR). Je crois que celle-ci gagne en popularité, parce qu'elle représente une alternative moins couteuse à l'imagerie par résonance magnétique (IRM). La lumière rouge devient plus transparente, et on mesure l'activité du cerveau. Mais on utilise cette méthode surtout dans la recherche impliquant des bébés.

Q : Au début, il me semblait que vous suggériez que les enfants dysphasiques ne sont pas seulement atteints au niveau linguistique mais aussi au niveau moteur et cognitif. Pourtant, vous nous avez dit ensuite que leurs yeux bougent comment ceux des autres enfants. Est-ce que le fait qu'ils sont plus lents à saisir certains aspects difficiles d'une langue veut dire qu'ils ont juste besoin de

[10] NdE : "Eye-tracking lab of the Utrecht Institute of Linguistics".

plus d'exposition à la langue et de plus de temps, mais qu'ils finissent par arriver au même niveau que les gens normaux, moyens, ou bien est-ce que c'est une différence cognitive permanente ?

ET : Personne ne sait. Impossible d'en dire davantage aujourd'hui parce que les chercheurs d'habitude s'intéressent aux enfants et seules une ou deux études ont été réalisées sur des adultes diagnostiqués comme présentant un trouble du langage. J'aurais tendance à dire que ces différences subsistent. Mais ils apprennent à compenser. J'ai entendu un jour l'exemple d'un cas typique de dysphasie en la personne d'un célèbre joueur de football néerlandais. D'après ce qu'on dit, il est atteint de dysphasie autant en néerlandais qu'en espagnol. Avec une de mes collaboratrices logopédistes, nous tentons actuellement de retrouver de jeunes adultes, d'une vingtaine d'années, qui étaient ses patients dans leur enfance. Je crois que nous en avons trouvé dix-huit, ce qui a parfois été difficile, parce qu'ils se sont éloignés géographiquement. Je n'ai pas encore regardé les données ; nous sommes en train de les récolter, mais je suis vraiment intéressée de voir ce qui leur est arrivé. À mon avis, la dysphasie ne disparait pas.

Q : J'ai travaillé en tant que logopédiste, une fois par semaine. Les enseignants nous envoient des enfants et ils veulent que nous leur disions si les problèmes de ces enfants sont spécifiques ou bien dus au bilinguisme. C'est probablement parce que beaucoup de ces enfants viennent de catégories sociales défavorisées, ils ont des difficultés à acquérir le français et l'autre langue. Alors, il est vrai que l'oculométrie pourrait être une manière de faire la différence entre les deux groupes. Vous avez mentionné le jugement de grammaticalité, et je pense que c'est quelque chose qu'on pourrait pratiquer dans notre cas.

ET : Oui, c'est faisable, c'est juste que l'oculométrie est chère.

Q : Est-ce que vous auriez une idée d'autres configurations ou d'autres tâches qui pourraient aider à mettre en évidence si le trouble est spécifique ou pas ?

ET : C'est difficile. Ce que vous voyez ici dans le cas du bilinguisme, c'est que la compréhension est probablement meilleure que celle des enfants dysphasiques. Le jugement de

grammaticalité ne fonctionne que jusqu'à un certain point. Comme vous le savez, les bilingues, en général, ont une plus grande conscience métalinguistique, parce qu'ils connaissent déjà plus d'une langue ; ils savent déjà qu'il peut y avoir plus d'un nom pour une chose. Ils sont plus conscients des aspects métalinguistiques simplement parce qu'ils connaissent mieux le système linguistique. Pour les enfants monolingues, ce serait plus compliqué, en particulier s'ils ont un trouble du langage. Dans notre cas, l'équipement d'oculométrie a couté quarante mille euros ; on en trouve de plus en plus de meilleur marché, et de toute évidence notre équipement est trop cher pour une utilisation quotidienne en cabinet. Il vous faut donc une solution plus abordable. Mais je pense que le jugement de grammaticalité est une solution ; la compréhension, d'une manière générale, aussi. Mais de nouveau, très souvent les bilingues ont une compréhension différente : je ne sais pas pour quelle raison mais quand on leur fait passer les tests de dépistage normaux, standardisés, ils se comportent parfois différemment dans les tâches actionnelles, mais souvent ce n'est pas parce qu'ils ne savent pas. En réalité, je ne sais vraiment pas ce qui se passe.

Q : Est-ce qu'ils réfléchissent trop ? Ils ont trop de possibilités, donc il leur faut plus de temps pour le traitement ?

ET : En fait, les tests ne mesurent même pas le temps de réaction. Très souvent, les logopédistes demandent aux enfants de « mettre le cheval derrière la barrière », et parfois les bilingues le placent ailleurs, et je suis presque sure que ce n'est pas parce qu'ils ne savent pas, mais il se passe quelque chose. En tout cas, ce que je peux dire pour le moment, c'est que tous les types de mesures réceptives sont plus appropriés pour distinguer les enfants atteints ou non d'un trouble du langage. À vrai dire, les enfants bilingues sont souvent envoyés en logopédie parce que leurs enseignants sont inquiets, mais souvent les monolingues provenant de milieux socioéconomiques défavorisés ne montrent pas de performances supérieures. Pendant la conférence de Steven GILLIS, j'ai déjà mentionné le fait qu'à sept mois seulement, on voit déjà des différences d'activité cérébrale entre les catégories socioéconomiques.

Q : Vous avez dit que vous attribuiez la différence que vous avez observée au bilinguisme, mais est-ce un problème de bilinguisme ou est-ce le problème d'avoir une langue dominante et une langue non-dominante ? Parce que vous avez dit que vous observiez des choses différentes en néerlandais et en russe.

ET : Je pense que dans notre cas, il s'agit véritablement d'un problème de dominance, parce que nous ne voyons pas tellement de bizarreries en néerlandais qui est clairement leur langue dominante. C'est la situation que l'on rencontre en général : ces enfants vivent dans un pays particulier, ils ont donc une langue dominante. Le Canada est probablement le meilleur endroit pour faire des recherche sur le bilinguisme, parce qu'autant l'anglais que le français – bien sûr à nouveau, vous vous trouverez dans une certaine région et il y aura une préférence – mais en tout cas au niveau du prestige, les deux langues sont plus ou moins équivalentes.

Q : Les logopédistes suisses ne peuvent-ils pas nous dire quelque chose sur ce sujet ?

Une autre personne dans le public : En fait, si on compare la situation avec le Canada, où on laisse les enfants parler leur langue maternelle, en Suisse, c'est une toute autre façon de penser. Et c'est un réel problème, parce que la plupart du temps, comme vous l'avez dit, on envoie aux logopédistes des enfants bilingues, probablement parce qu'ils ne parlent pas bien le français. Alors qu'au Canada, d'abord on apprend sa propre langue, puis l'anglais ou le français, donc les situations ne sont pas les mêmes.

ET : Oui, mais le bilinguisme sans dominance n'existe pas. Il serait empiriquement intéressant de tester des locuteurs néerlandophones en Russie, ou en Pologne mais il n'y en a pratiquement pas !

Références

ANDREU Llorenç, SANZ-TORRENT Mònica & TRUESWELL John C. (2013). Anticipatory Sentence Processing in Children with Specific Language Impairment: Evidence from Eye Movements during Listening. *Applied Psycholinguistics 34*, 5-44.

ARGYRI Efrosyni & SORACE Antonella. (2007). Crosslinguistic Influence and Language Dominance in Older Bilingual Children. *Bilingualism*: *Language and Cognition 10-01*, 77-99.

GAGARINA Natalia, VOEIKOVA Maria & GRUZINCEV Sergej. (2003). New Version of Morphological Coding for the Speech Production of Russian Children. In KOSTA Peter, BLASZCZAK Joanna, FRASEK Jens, GEIST Ljudmila & ZYGIS Marzena (Eds), *Investigations into Formal Slavic Linguistics*. Berne: Peter Lang, 243-258.

GÜLZOW Insa & GAGARINA Natalia. (2007). Noun Phrases, Pronouns and Anaphoric Reference in Young Children Narratives. *ZAS Papers in Linguistics 48*, 203-223.

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

HART Betty & RISLEY Todd R. (2003).The Early Catastrophe: The 30 Million Word Gap. *American Educator*, 4-9.

HICKMANN Maya. (2003). Children's Discourse. Person, Space, and Time across Languages. Cambridge: Cambridge University Press.

MACWHINNEY Brian. (2000). *The CHILDES Project: Tools for Analyzing Talk.* Mahwah: Lawrence Erlbaum.

MAK Willem M., TRIBUSHININA Elena & ANDREIUSHINA Elizaveta (2013). Semantics of Connectives Guides Referential Expectations in Discourse: An Eye-Tracking Study of Dutch and Russian. *Discourse Processes 50-8*, 557-576.

MAK Willem M., TRIBUSHININA Elena, LOMAKO Julia, GAGARINA Natalia, ABROSOVA Ekaterina & SANDERS Ted (soumis) Connective Processing by Simultaneous Bilingual Children and Monolinguals with SLI: Similar Profiles? Submitted to *Journal of Child Language*.

MARINIS Theodoros & CHONDROGIANNI Vasiliki. (2011). Comprehension of Reflexives and Pronouns in Sequential Bilingual Children: do they Pattern Similarly to L1 Children, L2 Adults, or Children with Specific Language Impairment? *Journal of Neurolinguistics 24*, 202-212.

MONTGOMERY James W. & LEONARD Laurence B. (1998). Real-Time Inflectional Processing by Children with Specific Language Impairment: Effects of Phonetic Substance. *Journal of Speech, Language, and Hearing Research 41*, 1432-1443.

TRIBUSHININA Elena, DUBINKINA Elena & SANDERS Ted. (2015). Can Connective Use Differentiate between Children with and without Specific Language Impairment? *First Language 35-1*, 3-26.

TRIBUSHININA Elena, MAK Willem M., ANDREIUSHINA Elizaveta, DUBINKINA Elena & SANDERS Ted. (2015). Connective Use by Bilinguals and Monolinguals with SLI. *Bilingualism: Language and Cognition.* Online first, doi:10.1017/S1366728915000577.

TRIBUSHININA Elena, VALCHEVA Eva & GAGARINA Natalia. (à paraitre). Acquisition of Additive Connectives by Russian-German Bilinguals: A Usage-Based Approach. In EVERS-VERMEUL Jacqueline & TRIBUSHININA Elena (Eds), *Usage-Based Approaches to Language Acquisition and Language Teaching*.

UNSWORTH Sharon & HULK Aafke. (2008). Early Successive Bilingualism: Disentangling the Relevant Factors. *Zeitschrift für Sprachwissenschaft 28-1*, 69-77.