Arbach, N. (2015) « Historique des corpus oraux », In *Constitution d'un corpus oral de FLE : enjeux théoriques et méthodologiques,* (p. 33-38)

1.3 Corpus et acquisition du langage

Nous allons discuter dans cette section d'un autre champ d'application et d'exploitation des corpus qui émergea bien avant toute technologie, celui de **l'acquisition du langage chez l'enfant** ou **ontogénèse linguistique**. L'étude de l'ontogénèse est un des domaines précurseurs où les chercheurs s'intéressèrent à la langue orale pour des raisons évidentes : il va de soi que durant les premières années de l'enfant, la langue écrite n'est pas maîtrisée et que le seul moyen d'analyser la langue en cours d'acquisition et ses mécanismes ne peut se faire que par le biais de données orales de productions d'enfants. Le plus ancien document pouvant être apparenté à un corpus est le journal que tint Jean Héroard sur la vie quotidienne de Louis XIII, *Le journal d'un roi*, dès la naissance de ce dernier en 1601 et jusqu'à la mort de l'auteur en 1628. L'ouvrage était essentiellement tenu pour des raisons médicales, cependant Jean Héroard y consigna tous les événements de la vie du dauphin : ses horaires et habitudes détaillés, ainsi que tout épisode public ou privé. Ce « procès-verbal d'expérience » rapporte également les productions langagières de l'enfant, transcrites (pour certains cas une transcription semi-phonétique) et parfois commentées. Ce journal, unique en son genre pour son époque, constitue un trésor pour les historiens, les psychologues et les linguistes, et l'intérêt d'un tel corpus est tel que, quatre siècles après sa création, il est encore étudié et exploité. Mais ce sont deux siècles plus tard que des savants s'intéressèrent de manière plus précise aux productions orales de leurs enfants, phénomène qui a été suivi par la constitution de larges corpus **transversaux** et **longitudinaux** tout au long du XXème siècle. Nous présenterons ces travaux et terminerons par la présentation de la situation actuelle en France.

1.3.1 Les premiers *baby books* ou *diary's note*

Au XIXème siècle, les expériences du type *baby books* ou *diary's note*, dans lesquels le chercheur tenait un journal sur le développement de son ou de ses enfants, s'inscrivent dans le courant des sciences naturelles et évolutionnistes. Charles Darwin en est le représentant le plus célèbre, et tint lui-même un journal sur son fils aîné et qui fera l'objet d'un article, « A biographical sketch of an infant » (Darwin, 1877), publié dans *Mind*. Il avait en cela été inspiré par Hyppolite Taine qui, un an plus tôt, avait publié « Note sur l'acquisition du langage chez les enfants et dans l'espèce humaine dans Revue Philosophique de la France et de l'étranger » (Taine, 1876). Après ces pionniers, nombreux prirent le relais, et nous citerons notamment les allemands Wilhelm Preyer et William Stern, pour les détails riches qu'ils nous ont laissés sur

la constitution de leurs corpus Preyer observe son fils quotidiennement et note avec précision toutes ses remarques. Il est le premier à transcrire phoniquement et de manière très détaillée les productions langagières de son fils, dont il fera un compte rendu dans « Die Seele des Kindes » (Preyer, 1884). L'étude couvre les productions langagières de l'enfant dès ses premières semaines et jusqu'à la fin de sa troisième année. Les travaux de Preyer restent néanmoins ceux d'un psychologue, qui s'est principalement intéressé aux processus cognitifs de l'acquisition du langage. Ce n'est que quelques années plus tard qu'une étude focalisée principalement sur l'acquisition du langage a eu lieu : William Stern et son épouse Clara ont tenu des journaux sur leurs trois enfants durant dix-huit ans, et ont publié « Die Kindersprache » (Stern & Stern, 1907), le premier journal entièrement consacré au langage de l'enfant. Nous rapportons ici quelques-unes des remarques faites par Morgenstern (2009) qui montrent l'importance et la qualité des travaux des Stern. Tous les deux faisaient la distinction entre observation et interprétation et autant que faire se peut, le couple tentait de tenir le journal sans que les enfants ne s'en rendissent compte en faisant déjà la distinction entre rapport d'événement et commentaire. En cela, leur travail met en relief le rapport, encore flou de nous jours, entre transcription et annotation. William Stern envisagea également très tôt que le « phonographe » et la « photographie » changeront sans doute la recherche sur le développement de l'enfant en permettant aux chercheurs de faire des enregistrements sans que l'enfant ne s'en rende compte.

1.3.2 Corpus transversaux et corpus longitudinaux

Nous avons constaté dans le paragraphe qui précède qu'il a existé **deux types d'approches** dans les corpus d'enfants. La première approche consiste à rechercher des universaux en collectant les données d'un grand nombre d'enfants, puis à les comparer entre elles ; le corpus constitué est alors de type transversal. Un corpus transversal est donc un corpus regroupant les productions de plusieurs enfants à un seul moment de leur développement, sans les suivre individuellement. La seconde approche consiste à étudier l'ontogénèse d'un enfant en particulier, afin de pouvoir suivre les mécanismes d'acquisition et leur évolution dans le temps ; le corpus constitué est alors un corpus de type longitudinal. Les corpus constitués furent alors auprès d'un nombre restreint d'enfants mais s'inscrivant dans la durée en enregistrant régulièrement les enfants participant à l'expérience. Si les premières études de Taine et des Stern que nous avons présentées étaient européennes, les études transversales en acquisition du langage qui ont suivi ont été principalement américaines, entre 1926-1927 et jusqu'en 1957 (McEnery & Wilson, 2001 : 3). Les Américains (voir tableau ci-dessous) prirent le relais car ils considérèrent ces études européennes comme étant aléatoires, peu scientifiques, peu fiables et

comme décrivant des enfants qui ne reflétaient pas un standard. Les travaux américains de la première moitié du XXème siècle ont donc voulu remédier à ces lacunes en adaptant de nouvelles méthodologies cherchant à octroyer une scientificité à ces corpus en instaurant des critères tels l'échantillonnage, la prise en compte de critères métalinguistiques (situation d'énonciation, sexe, milieu socio-économique, âge et enfants spécifiques comme les jumeaux ou les enfants doués) et l'homogénéité des données. Ces critères, comme nous le verrons tout au long du 2ème chapitre, feront partie des critères de constitution des corpus scientifiques modernes. L'objectif formel des corpus transversaux est l'établissement de normes dans l'acquisition du langage grâce à de larges études quantitatives et comparatives. Voici un tableau regroupant les principales études transversales menées entre 1926 et 1957 :

Tableau 1 : Études transversales sur l'acquisition du langage entre 1926 et 1957

Auteur	Date	Nombre d'enfants	Âge	Échantillons	Thème de recherche
Smith	1926	124	2 à 5	1 heure de conversation	Longueur d'énoncé/ développement
McCarthy	1930	140	1;5 à 4;6	50 énoncés	Longueur d'énoncé/ développement
Day	1932	160	2 à 5	50 énoncés	Langage des jumeaux
Fisher	1934	72	1;6 à 4;6	3 heures par échantillon	Enfants doués
Davis	1937	173/166	5;6 à 6;6	50 énoncés	Jumeaux/ enfants uniques
Young	1941	74	2;6 à 5;5	6 heures de conversations	Classe sociale
Templin	1957	430	3 à 8	50 énoncés	Longueur d'énoncé/ développement

Comme nous pouvons le voir, la principale caractéristique des corpus transversaux est la collecte de données auprès d'un grand nombre d'enfants, selon un échantillonnage qui dépend d'un protocole scientifique (échantillons égaux en taille, critères de sélection des enfants à enregistrer). La période des corpus transversaux dura jusqu'en 1957. Les chercheurs ont alors délaissé le souci quantitatif et comparatif pour s'intéresser à l'évolution dans le temps des compétences linguistiques d'un ou de plusieurs sujets. Les raisons principales du délaissement des études transversales au profit d'études longitudinales furent de deux ordres, l'un technique, l'autre théorique. Le premier fut la démocratisation du magnétophone qui permit le suivi d'un enfant plus facilement qu'auparavant. Le second fut la parution de *Syntactic structures* (Chomsky, 1957) qui amena les chercheurs à orienter leurs travaux sur la naissance de la syntaxe ; pour cela, ils eurent besoin de corpus longitudinaux leur permettent de suivre

l'évolution de la syntaxe d'un seul et même enfant et donc, théoriquement, de comprendre le processus universel d'acquisition du langage. Les principaux projets entre 1963 et 1970 sont les suivants :

Tableau 2 : Principales études longitudinales sur l'acquisition du langage entre 1957 et 1973

Auteur	Année	Nombre d'enfants	Âge en début de projet	Durée de l'étude	Intermittence de recueil des données
Braine	1963	3	Entre 19 et 23 mois		Continue + 12 sessions de 4 heures
Miller & Ervin	1964	5	Entre 21 et 24 mois	2 ans	Deux à trois séances de quatre à cinq heures tous les deux mois.
Bloom	1970	3	Entre 21 et 27 mois		Deux heures toutes les deux semaines + une demi-heure toutes les semaines
Brown	1973	3	Premiers mots de l'enfant	Fin 3ème année	Hebdomadaire ou bimensuel

En comparaison avec le tableau précédant, nous constatons le nombre beaucoup plus restreint d'enfants enregistrés, mais le suivi des enfants sur des périodes relativement plus étendues par rapport aux données transversales. Ce type de collecte s'étala jusqu'à la moitié des années 1970 environ ; depuis, la situation est une synthèse des deux approches comme nous allons le voir dans le paragraphe qui suit.

1.3.3 Situation actuelle

Nous avons vu que les études transversales américaines avaient pour principales motivations le souci de rigueur ainsi que la recherche de scientificité des analyses et des résultats, éléments qui faisaient défaut dans les premières approches européennes. Les données représentaient les productions d'un grand nombre d'enfants sans avoir pu opérer de suivis longitudinaux en raison de l'absence des techniques appropriées entre 1927 et 1957. Les études longitudinales qui ont suivi à partir de 1957 ont délaissé la représentativité et l'échantillonnage en faveur du suivi du développement de l'enfant ; là encore, en raison de l'absence de moyens techniques et humains à mener des études longitudinales d'envergure. Les corpus uniquement transversaux ou uniquement longitudinaux ont montré leurs limites. Or les valeurs transversales ou longitudinales d'un corpus ne sont pas exclusives. Depuis les années 1970, les chercheurs désirent avoir à disposition des corpus à la fois transversaux et longitudinaux afin de pouvoir à la fois comprendre les processus d'acquisition en étudiant les corpus longitudinaux, mais aussi de vérifier leurs résultats, de les comparer et de les compléter en ayant sous la main des données transversales, soit des études longitudinales effectuées sur un grand nombre d'enfants. C'est dans ce contexte que, les moyens techniques aidant, le projet « Child Language Data Exchange

System » dit CHILDES (Brian MacWhinney & Snow, 1985) vit le jour en 1984. Il s'agit de la première base de données orale numérique participative internationale de langue orale et les premières données à y être intégrées sont celles du projet Brown. Elle contient aujourd'hui un grand nombre de productions d'enfants collectées de par le monde, et représente un corpus longitudinal, transversal et multilingue. Notre corpus a été transcrit au moyen de l'un des outils de CHILDES, le logiciel CLAN, que nous présenterons en détail au 3ème chapitre. En outre, CLAN est le logiciel de transcription de l'un des corpus français les plus importants en ce qui concerne le développement du langage entre 1 et 3 ans, à savoir le corpus du projet ANR Colaje, qui regroupe des corpus audio-visuels, leurs transcriptions et des tests de langage. Les enfants sont filmés dans leurs familles et l'ensemble des documents est mis à disposition. L'apport de l'étude de l'ontogénèse à la linguistique de corpus, en prenant en compte le potentiel des données à représenter des universaux, et donc en instaurant une base scientifique aux bases de données, est ainsi considérable au niveau méthodologique. Il faut également retenir que l'idée de suivi de cohorte est née dans ce domaine, et que les corpus longitudinaux sont aujourd'hui fréquents dans l'étude de l'interlangue.