Studii de lingvistică 7, 2017, 89 - 112

L'apport de la linguistique de corpus à l'étude des situations cliniques : l'utilisation de ressources écologiques

The contribution of corpus linguistics to the study of clinical situations: using ecological resources

Christine da Silva Genest[1]
Caroline Masson[2]

Abstract: Studies in the field of language acquisition and language disorders are part of different and complementary theoretical frameworks. These approaches vary from experimental to natural resource methods. The tools developed in the branch of corpus linguistics can be used to describe and analyse clinical situations. Furthermore, they can be used by speech and language pathologists in the context of their professional practices. After explaining methods for the transcription of natural data, we show tools enabling the quantitative and qualitative analysis of language practices. To this end, the focus will be placed on CLAN software and on the relevance of linguistic expertise as a means of evaluating professional practices.

Key words: corpus linguistics, language acquisition, language disorders, interaction, analysis tools.

1. Introduction

La linguistique de corpus, par opposition à d'autres approches de la linguistique telles que le générativisme, s'est développée en se centrant sur la description de faits linguistiques attestés. Pour ce faire, elle a dû se doter d'outils de traitement statistique (Teubert 2009) permettant le repérage, automatique ou semi-automatique, de phénomènes langagiers à partir de données authentiques. Le recours à des données tant observables que vérifiables et la recherche d'outils lui assurant une rigueur scientifique (Lejeune 2010, Leech

[1] EA 3450 DevAH, Université de Lorraine, et membre associée à l'EA 7345 CLESTHIA, Université Sorbonne Nouvelle ; christine.da-silva-genest@univ-lorraine.fr.
[2] EA 7345 CLESTHIA, Université Sorbonne Nouvelle ; caroline.masson@sorbonne-nouvelle.fr.

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

1992) sont l'essence même du travail en linguistique de corpus. La valorisation et la diffusion des *corpora* oraux ou écrits sont des problématiques tout aussi centrales pour les chercheurs que celles de leur constitution et diffusion. En effet, depuis longtemps, la question de la mise à disposition de données constituées par eux-mêmes est soulevée, que ce soit les recherches s'intéressant au français parlé (Blanche-Benveniste 1997, Debaisieux *et al.* 2016), à la didactique des langues (Boulton 2009), à l'acquisition (Canut & Vertalier 2008, Morgenstern & Parisse 2007) ou aux pathologies du langage (Goodwin 2000 ; MacWhinney *et al.* 2010, 2011).

Les résultats obtenus à partir d'analyses de *corpora* permettent de contribuer au développement des connaissances sur des phénomènes langagiers. Dans le champ des études sur l'acquisition et les pathologies du langage (dans lequel nous nous inscrivons), il existe différentes méthodes pour analyser les phénomènes langagiers. La manière d'appréhender les processus d'acquisition peut être placée sur un continuum allant d'épreuves contrôlées dans des situations induisant des formes et des structures verbales ciblées (Karmiloff & Karmiloff-Smith 2001) au recueil de données en situation la plus écologique possible (Morgenstern & Parisse 2007). De par notre positionnement socio-interactionniste (Bruner 1983, Vygotski 1997, François 1993, de Weck & Salazar-Orvig 2010, Veneziano 2000), nous considérons que la meilleure façon de décrire le fonctionnement langagier est d'avoir accès à des données en situation naturelle. Sans le recours aux productions réelles des enfants, il est impossible de rendre compte des processus acquisitionnels, qu'ils soient typiques ou atypiques. Or, le développement des connaissances sur l'acquisition et les pathologies du langage est un enjeu de santé publique (Masson 2014) qui dépasse le cadre de la recherche. Elles contribuent à l'amélioration des prises en charge, notamment orthophoniques, de ces pathologies. Le travail sur corpus présente un intérêt principal, celui de passer d'une approche déductive à une approche empirique, fondée sur des données attestées. La seule intuition du chercheur, comme celui du praticien, ne peut rendre compte à elle seule de la réalité des formes langagières. Il est nécessaire d'avoir recours à une méthodologie fiable et reconnue (Boulton *et al.* 2013, Parisse & Le Normand 1998), telle qu'elle est proposée dans la linguistique de corpus. Ainsi, la constitution de données authentiques, l'élaboration d'une méthodologie et la création d'outils d'analyses spécifiques sont tout aussi primordiales pour la recherche fondamentale que pour les pratiques professionnelles.

Nous nous proposons d'étudier les productions orales attestées des professionnels en situation d'interactions naturelles et de présenter le logiciel CLAN, fréquemment utilisé par la communauté des chercheurs en acquisition et pathologie du langage. Notre objectif

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

est de déterminer de quelle façon la linguistique peut aider les professionnels à décrire leurs gestes et leur degré d'efficience ainsi qu'à apprécier l'effet sur leur interlocuteur. Avant de présenter les outils linguistiques de traitement de données orales (cf. sections 3 et 4), nous exposerons les principes et les méthodes de la constitution d'un corpus (cf. section 2).

2. Constituer un corpus : principes et méthodologie

La constitution d'un corpus doit nécessairement être en accord avec les principes méthodologiques inhérents au cadre théorique dans lequel s'inscrit le chercheur (cf. 2.1.) ainsi qu'avec des aspects légaux et juridiques (cf. 2.2.). Il convient donc de définir ces points avec précision.

2.1. Présentation des méthodes de recueil de données

La démarche de recueil et d'analyse de données orales dans laquelle nous nous inscrivons implique d'adopter plusieurs principes. Il est nécessaire :

- de déterminer et définir avec précision son cadre d'étude, les conditions de recueil de ses données, afin d'anticiper d'éventuels problèmes. Le choix de l'activité observée (récit d'expériences personnelles, conversation ordinaire ou thématisée, description d'images, jeux libres, etc.), la disposition du/des locuteur(s) ou encore la maîtrise technique du matériel sont des points que l'informateur devra examiner avant de commencer son travail de collecte ;

- d'avoir recours à l'enregistrement audio et/ou vidéo ; ce choix se fait en fonction de ses objectifs, des conditions de recueil et de la situation. Ainsi, des données vidéo seront indispensables pour une étude portant sur la multimodalité du langage. La vidéo offre, en outre, de meilleures conditions d'interprétation des productions des participants, car elle permet de lever certaines ambiguïtés. Quelle que soit la méthode de recueil, la nature des données aura un effet sur la forme et la complexité de la transcription ;

- de s'interroger sur ce qu'on entend par « données réelles ». On retrouve plusieurs termes utilisés de façon synonymique dans la littérature, comme « données écologiques », « naturelles » ou « spontanées ». Pourtant, l'expression « données écologiques » ne doit pas être réduite aux situations informelles et aux productions spontanées des locuteurs. Les situations cliniques, où les formes langagières sont semi-dirigées par le praticien, peuvent aussi être qualifiées de contexte naturel. Le terme « écologique » renvoie à la situation et au contexte de recueil, tandis que celui de « spontané » représente des situations où les productions verbales des locuteurs

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

ne sont pas totalement induites et provoquées et permettent donc aux participants d'avoir une certaine liberté créative (de Weck 2003).

L'intérêt des chercheurs en acquisition pour les situations naturelles date des premières descriptions du langage enfantin, situées entre la fin du 19e siècle et le début du 20e (pour une revue de la question, voir Morgenstern 2009). Les outils techniques ne permettant pas à l'époque de collecter un grand nombre de données, des échantillons de formes orales produites par les enfants sont constitués par le biais de prises de notes du chercheur[3]. A partir des années 60, grâce à l'apparition des magnétophones, on commence à enregistrer le langage des enfants. Les travaux de Brown (1973) sur les données longitudinales de trois enfants (rebaptisés dans l'étude Adam, Eve et Sarah) montrent le changement de méthodologie qui s'opère à cette époque. En effet, l'enregistrement sur bandes permet, non seulement, de conserver les données mais, en plus, d'y revenir afin d'analyser ce qui a été produit et non ce que l'on pense avoir perçu. Les possibilités d'étude s'en trouvent également élargies, puisque les enregistrements permettent de travailler sur plusieurs aspects du développement langagier. Cette période correspond également à l'émergence de travaux portant sur le rôle de l'*input* (Veneziano 2014, Onnis 2014) : on s'intéresse alors à la corrélation entre le langage reçu par l'enfant et le langage produit par celui-ci, ce qui suppose de recueillir des données en situation réelle. Par la suite, et parallèlement au développement de nouvelles technologies, les chercheurs passeront d'enregistrements audio à des enregistrements vidéo, qui offrent des possibilités de traitement plus riches, en considérant la dimension multimodale du langage (gestes, regards, contextes, etc.).

Les évolutions techniques, telles que la miniaturisation des appareils d'enregistrement et la place prise par ces outils dans les foyers à partir de la fin des années 2000, impliquent également de repenser la place de l'observateur et de son matériel. La question le plus souvent posée porte sur la façon de recueillir des données écologiques en introduisant un objet étranger dans la situation d'interaction verbale. Labov (1972) avait ainsi formulé que l'observateur se trouvait pris dans un « paradoxe » : en cherchant à rendre compte de formes réelles et/ou spontanées, son statut d'examinateur a pour effet de modifier le cadre initial. La présence d'un dispositif d'enregistrement n'est pas tout à fait transparente et il est indéniable que le locuteur observé ne l'oublie jamais totalement (Lallier 2011). Il appartient donc au chercheur de déterminer le degré de modification induite et d'en tenir compte (Morgenstern 2016). Bien qu'elle soit étrangère dans un premier temps, la caméra est rapidement acceptée par les locuteurs en présence au fur et à mesure des séances et au cours même de

[3] Nous pouvons citer ici comme exemples de recherches s'appuyant sur des journaux parentaux Darwin (1877), Cohen (1993) ou François (1977).

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

l'enregistrement. Dans le cadre de recueils de *corpora* longitudinaux auprès de jeunes enfants, la présence de l'informateur avec sa caméra peut même être totalement intégrée à la vie quotidienne des enfants, habitués à la visite régulière de celui-ci (Morgenstern & Parisse 2012). Dans le cabinet du praticien, la caméra peut être considérée de prime abord comme intrusive mais, à l'instar des enregistrements réalisés dans le cadre familial, si elle est présente systématiquement, sans qu'elle ne soit forcément en marche, alors elle suscitera sans doute moins d'intérêt ou d'inquiétude chez le patient, qui agira ainsi de façon plus naturelle. Le praticien qui désire appréhender dans sa globalité sa pratique peut concevoir la caméra comme un outil de remédiation semblable aux autres outils utilisés lors de la prise en charge, au sens où elle va l'aider à évaluer les effets de son intervention et à créer les ajustements nécessaires.

Il est évident que l'enregistrement de données vidéo n'est qu'un point de vue et n'est pas « un reflet fidèle d'un événement réel » (Mondada 1998 : 61). Il est donc important de connaître les limites et les biais de chaque mode de recueil de données et d'en tenir compte au moment de leur traitement.

2.2. Aspects réglementaires et juridiques de la recherche impliquant la personne en France

La constitution de corpus et le recours à des enregistrements audio et/ou vidéo sont donc devenus des pratiques de plus en plus courantes. Dans le cadre de la recherche, celles-ci doivent avoir un objectif et être justifiées. C'est pourquoi les questions d'éthique liées à la création de corpus sont devenues centrales. En 2006, un guide des « bonnes pratiques » pour la constitution, l'étude et la conservation de données orales (Baude 2006) a été élaboré, pour que les *corpora* oraux soient considérés comme un objet construit au même titre qu'une œuvre littéraire.

Récemment, le cadre juridique français a évolué et son application ne se restreint plus aux recherches dites biomédicales. En effet, depuis le 16 novembre 2016, la loi Jardé régit le cadre des recherches impliquant la personne humaine. Trois types de recherches sont alors distingués : celles de catégorie 1, qui sont interventionnelles et non justifiées par une prise en charge habituelle ; celles de catégorie 2, « qui ne portent pas sur des médicaments et ne comportent que des risques et des contraintes minimes » et s'inscrivent dans le cadre d'une pratique habituelle ; et, enfin, celles de catégorie 3 dite non interventionnelles ou observationnelles. Quel que soit le type, toutes les recherches doivent être présentées à un Comité de Protection de la Personne, qui devient ainsi la seule instance éthique pour toute recherche portant sur l'être humain.

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

Les formes de recherche présentées dans le cadre de cet article sont pour la plupart observationnelles et s'inscrivent dans le champ linguistique. Plus récemment, le décret n°2017-884 du 9 mai 2017 simplifie certaines dispositions réglementaires de la loi Jardé, en précisant le champ des recherches impliquant la personne humaine soumises à l'avis des comités de protection des personnes. Ainsi, conformément à l'article 2, ne sont pas considérées comme des recherches impliquant la personne humaine celles qui, entre autres, visent à réaliser des expérimentations en sciences humaines et sociales dans le domaine de la santé et qui n'ont pas pour objet le développement des connaissances biologiques ou médicales.

Toute collecte de données doit être réalisée dans le respect de la loi *Informatique et libertés*. Le consentement des participants doit être éclairé. Selon le Centre National de l'Informatique et des Libertés (CNIL), tous les participants doivent être informés de l'identité du responsable de la recherche, du cadre, des finalités, des modalités pratiques, de l'existence d'un droit d'accès, de rectification voire d'opposition à la collecte (cf. article 6 de la loi Informatique et Libertés)[4]. Bien évidemment, les connaissances apportées sur les finalités de la recherche peuvent être un frein au recueil de données écologiques ou naturelles (cf. 2.1.) et aux pratiques habituelles. En outre, d'autres questions doivent être abordées et exposées clairement aux participants, notamment celles concernant l'anonymat, le cryptage, le « floutage », le traitement des données, leur éventuelle diffusion ou non diffusion, etc.

Les questions d'éthique et de déontologie de la recherche sont de plus en plus soulevées par les chercheurs en linguistique fondamentale ou appliquée et en sciences humaines et sociales de manière plus générale. L'intégration des règles de « bonnes conduites scientifiques » est centrale, quel que soit le champ d'investigation.

3. Transcrire des données : méthodes et outils

3.1. Transcrire : principes de base

Tout comme le mode de recueil des données, le travail de transcription relève de choix théoriques et interprétatifs du chercheur (Mondada 2000, Mondada 2008). Il est ainsi indispensable de réfléchir en amont à ce que l'on veut voir apparaître dans sa transcription et aux objectifs que l'on souhaite atteindre (Ochs 1979, Sinclair 1996), au risque de passer par une étape supplémentaire de vérifications ou d'ajouts.

[4] Le consortium CORLI (Corpus, Langues, Interactions), porté par l'Institut de Linguistique Française, travaille notamment sur ces questions pour assurer le respect des droits dans le cadre des recherches en linguistique (https://corli.huma-num.fr.).

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

La transcription des données représente la première étape de l'analyse, au sens où plus elle sera complète et proche des éléments produits, plus les résultats et les conclusions seront riches et précis. C'est pourquoi le travail de transcription ne doit pas être influencé par les hypothèses formulées ou les représentations du transcripteur. Quels que soient la pratique, l'orientation théorique et les objectifs de recherche, il convient de respecter la règle suivante : transcrire seulement ce qui a été réellement produit par les locuteurs. Il n'en reste pas moins que le fait de transcrire par écrit la langue parlée est paradoxal (Blanche-Benveniste & Jeanjean 1987, Blanche-Benveniste 1997, Gadet 2003) et entraîne de nombreuses difficultés dont certaines sont propres au transcripteur. Pour reprendre les termes de Blanche-Benveniste & Jeanjean (1987 :102), « l'oreille est un traître ; on écoute ce qu'on s'attend à écouter » ou ceux de Bilger *et al.* (1997 : 58) « [l'oreille est] surtout asservie à la recherche de signification ». Une transcription présente donc toujours une part de subjectivité (Ochs 1979) qui doit influencer le moins possible le transcripteur. Ainsi, celui-ci pourra être tenté de « corriger » une production en fonction de la norme de l'écrit (Falbo 2005). C'est pourquoi une connaissance approfondie des caractéristiques de l'oral (pour une revue, voir Blanche-Benveniste & Bilger 1999) est indispensable pour appréhender la diversité des pratiques du locuteur et tendre à une certaine objectivité. Ces recommandations valent autant pour le chercheur qui a pour objectif de mener un travail exhaustif sur une ou plusieurs problématiques que pour le professionnel qui collecte et travaille des données pour sa pratique, qu'il soit orthophoniste ou enseignant (Canut *et al.* 2017).

Un autre aspect important à considérer est celui du temps consacré à la transcription des données. La réalisation d'une transcription complète est une activité assez chronophage : on considère généralement que la transcription d'un enregistrement vidéo d'une heure peut aller de dix heures (Tomasello & Stahl 2004) à cinquante heures de travail (Morgenstern & Parisse 2007), en fonction de la finesse des descriptions. Or, le temps de la recherche et celui de la pratique sont différents. Il est difficilement envisageable de demander à un professionnel de santé d'accorder autant de temps à l'étape de la transcription. De ce fait, on peut s'interroger sur la façon dont un professionnel de santé va procéder pour réaliser cette étape inévitable et se demander si des adaptations sont envisageables sans nuire à la qualité ultérieure des analyses.

Tout d'abord, comme pour le chercheur, des choix doivent être effectués, car on ne peut pas rendre compte de tous les aspects. Les adaptations envisagées dépendront fortement des objectifs préalables et des analyses ultérieures. S'il s'agit d'évaluer la participation du patient aux divers échanges, alors il sera nécessaire de transcrire le discours de tous les participants. S'il s'agit d'apprécier l'évolution des

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

productions du patient sur un niveau linguistique (morphosyntaxique, par exemple) à un temps T+1 par rapport à T lors d'une même activité, alors seuls les énoncés de ce dernier seront transcrits, par exemple.

Pour savoir quelles adaptations effectuer, il est nécessaire, d'une part, d'avoir une connaissance approfondie des logiciels et des conventions de transcription et, d'autre part, de se poser trois questions principales :

- quelles seront les analyses effectuées et quelles informations seront nécessaires pour les réaliser ?

- est-ce que le fait de ne transcrire que certains comportements aura des conséquences pour les analyses ultérieures ?

Face aux outils dont il dispose (cf. 3.2.), le praticien s'interrogera, par exemple, sur la pertinence d'une transcription intégrale de l'enregistrement, de la notation des pauses, des chevauchements de parole, des reprises à l'intérieur d'un même énoncé, de la description fine des comportements non verbaux ou encore de la réduction du nombre de conventions. Un même extrait d'interaction peut être transcrit différemment selon les objectifs et l'objet d'analyse (Traverso 2016). La transcription *via* un simple logiciel de traitement de texte, en écoutant et en visionnant en parallèle le support audio ou vidéo pour y revenir si besoin, est envisageable. Toutefois, le logiciel n'offrira pas ensuite à l'utilisateur la possibilité de générer des analyses automatiques qui faciliteront la description ultérieure des données recueillies. Or, certains logiciels de transcription sont conçus justement pour réaliser, à partir de conventions de transcription établies par leurs créateurs et utilisateurs, des analyses automatiques.

3.2. Utiliser des logiciels de transcription et des outils de traitement de données

Développé par Brian MacWhinney & Catherine Snow en 1984 aux USA dans le cadre du projet international CHILDES[5], CLAN est un logiciel de transcription et d'analyses automatiques en accès libre. Il offre la possibilité de transcrire des données de manière très fine et d'aligner l'audio ou la vidéo à la transcription.

L'utilisation de CLAN est proche du traitement de texte : parmi ses fonctionnalités, la plupart de celles proposées par les logiciels de traitements de texte sont disponibles (copier/coller, choix de police, etc.). Plus précisément, un fichier transcrit sous CLAN est constitué de la façon suivante (cf. image 1) :

- des lignes d'en-têtes, commençant par le symbole @ et contenant les métadonnées du corpus telles que la langue, les participants, les informations sur la date, le lieu, le nom de la vidéo associée, etc. ;

[5] http://childes.talkbank.org.

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

- la transcription : chaque production d'un locuteur est associée à une balise qui isole le passage audio ou vidéo concerné en vue d'une lecture seule ou totale de la transcription. Un ensemble de conventions de transcription doit être appliqué pour rendre compte des différents phénomènes linguistiques des locuteurs (ex. élision, vocalisations…). Des lignes dites « dépendantes » peuvent également être ajoutées aux lignes dites « principales » afin d'apporter des informations supplémentaires (gestes, regards, interprétations, etc.)[6]. Leur choix et leur nombre sont à déterminer par le chercheur et/ou le praticien en fonction des objectifs de l'analyse.

```
@Begin
@Languages:    fra
@Participants:  CHI Rayan Target_Child, THE Therapist, AMA Amandine, OBS
      Observatrice
@ID:   fra|MAKATON|CHI|5;00.09|male|||Target_Child|||
@ID:   fra|MAKATON|THE|||female||Therapist|||
@ID:   fra|MAKATON|AMA|||female||Student|||
@ID:   fra|MAKATON|OBS|||female||Camera|||
@Birth of CHI:  08-DEC-2009
@Date:  23-NOV-2012
@Number:      1
@Location:     Therapist's room
@Comment:      Perrin Camille (août 2016) - Caroline Masson (oct 2016)
@Media: MAKATON-05-15_11_2013suite, video
@G:    séance autour d'une phrase MAKATON
%com:   la séance reprend après une coupure liée à l'intervention d'une
      personne extérieure
*THE:   on recommence allez on est (re)partis ! •
%sit:   AMA tend la boîte à picto à CHI
*AMA:   oh ! •
*THE:   oh non non mais oh c'est quoi ? •
%sit:   CHI pioche deux pictos
*CHI:   e~@fs (p)ain. •
%pho:   e~ e~
*THE:   ouais c'est du pain t(u) as bien reconnu super ! •
%gpx:   [PAIN]
*CHI:   en+haut. •
%pho:   a~o
%sit:   se lève et place le picto sur le mur
*THE:   ah ouais comment on fait pain ? •
*THE:   pain. •
%gpx:   [PAIN]
*CHI:   0. •
%gpx:   [PAIN]
```

[6] Le détail des conventions de transcription et des lignes dépendantes sont disponibles sur le site de CHILDES, en anglais. Des groupes de recherche français ont également élaboré des guides, comme celui, très complet, du projet ANR CoLaJE (http://colaje.scicog.fr/images/stories/PDF/Guide-CLAN_colaje-juin-11.pdf). Le lecteur intéressé pourra s'y reporter en gardant à l'esprit que le logiciel est constamment mis à jour et amélioré.

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)



Image1 : Illustration d'une transcription réalisée avec le logiciel CLAN

Le logiciel CLAN offre également la possibilité d'exporter des données vers d'autres logiciels, en vue de générer des analyses complémentaires. Les fichiers peuvent ainsi être convertis pour être utilisés avec des logiciels tels que PRAAT[7] (Boersma & Weenink 2009), conçu pour analyser le traitement du signal acoustique, PHON[8] (Rose & MacWhinney 2014), élaboré pour des analyses phonologiques fines, et ELAN[9] (Brugman *et al.* 2004), pour l'analyse des gestes et du non verbal. En ce qui concerne la transcription de données audio, elles peuvent être réalisées avec l'outil Transcriber[10] (Barras *et al.* 2000), qui associe texte et son et crée des balises d'annotations.

4. Types d'analyses appliqués aux situations cliniques

Le développement des savoirs théoriques portant sur les processus développementaux et les caractéristiques langagières des troubles peut avoir une application dans d'autres domaines, comme celui de l'éducation ou de la santé. Par exemple, dans le cadre de l'enseignement, le travail sur des données orales attestées constitue le point de départ de la prise de conscience des enseignants sur le lien entre les interactions langagières et l'apprentissage du langage (Canut *et al.* 2013). Il nous semble que ces outils développés par la linguistique de corpus peuvent être adaptées également aux besoins des professionnels de santé. Au sein de nos recherches actuelles (da

[7] https://www.phon.ca/phontrac.
[8] http://www.fon.hum.uva.nl/praat/.
[9] https://tla.mpi.nl/tools/tla-tools/elan/download/.
[10] http://trans.sourceforge.net/en/presentation.php.

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

Silva Genest 2014, Canut *et al.* 2017)[11], nous tentons précisément de montrer les ponts envisageables entre la recherche et la pratique orthophonique.

Pour le professionnel orthophoniste, la constitution de données et leur traitement peuvent être motivés par des objectifs thérapeutiques et par l'adéquation avec les recommandations de la profession. Selon le référentiel des activités des orthophonistes (cf. Annexe 1, BO n°32 du 5 septembre 2013), ces professionnels doivent être en mesure d'utiliser des outils d'observation clinique permettant d'évaluer les fonctions du langage et de la communication en utilisant du matériel spécifiquement adapté (grilles, questionnaires, tests standardisés), des enregistrements audio et/ou vidéo. Ils doivent également pouvoir transcrire les *corpora* recueillis. Le plus souvent, ces professionnels ont recours à des tests standardisés mais, dès lors qu'il s'agit d'évaluer les compétences pragmatiques et communicationnelles, leur utilisation s'avère plus difficile. Par conséquent, l'appropriation des outils spécifiques au recueil de données dites spontanées ou semi-spontanées est essentielle. Bien qu'indispensables pour réaliser ces objectifs, les analyses linguistiques et interactionnelles sont relativement coûteuses en temps et demandent un haut niveau d'expertise. Elles peuvent, cependant, être appréhendées plus facilement en utilisant le logiciel CLAN, pour lequel un guide d'utilisation destiné aux orthophonistes a été élaboré (Ratner & Brundage 2016). Ce guide présente toutes les étapes, du téléchargement du logiciel aux commandes d'analyses automatiques. Ce sont ces commandes que nous présentons par la suite. Nous nous proposons de montrer la pertinence de l'utilisation par les orthophonistes des outils développés par la linguistique à deux niveaux au moins : d'une part, pour analyser finement les caractéristiques langagières des productions de leur patient en vue, par exemple, d'un bilan ou d'un suivi thérapeutique de manière complémentaire à l'utilisation d'autres outils (cf. 4.1) et, d'autre part, pour décrire leurs conduites professionnelles et appréhender leur efficience (cf. 4.2.).

4.1. Décrire les productions langagières des locuteurs

CLAN offre la possibilité de traiter les données transcrites en utilisant diverses commandes. Parmi celles-ci, on peut distinguer quatre types de commande :

- celles qui permettent de réaliser des analyses automatiques sur le fichier .cha telles que MOR ou RETRACE ;
- celles qui calculent automatiquement une mesure comme

[11] Le numéro 18 de la revue *CORPUS* portera spécifiquement sur cette thématique (da Silva Genest, C. et Masson, C. (2018), « Corpus et pathologies du langage : du recueil à l'analyse de données pour une linguistique clinique et appliquée »).

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

la commande MLU, MLT, FREQ correspondant respectivement à l'analyse de la longueur moyenne des énoncés, des tours de parole ou de la fréquence lexicale (Type Token Ratio ou VOCD) ;

- commandes qui regroupent un ensemble d'analyses comme EVAL, KIDEVAL ou FLUCALC en générant un fichier *output* qui se présentent sous la forme d'un tableur ;

- enfin, les commandes qui permettent de convertir le fichier .cha dans un nouveau format, utilisé par un autre logiciel de traitement de données (ex. ELAN, PHON, PRAAT ; cf. 2.2.1.).

De nouvelles commandes ont été développées récemment par les chercheurs du projet CHILDES. Leur apparition est fortement liée à une réflexion plus générale sur les outils d'évaluation du langage oral. Si les professionnels de santé privilégient les batteries de tests de langage pour situer le niveau d'un patient par rapport à une norme établie (notamment pour justifier une prise en charge d'un sujet auprès des institutions), les compétences d'un sujet ne peuvent toutefois pas se réduire aux résultats obtenus par ce biais. En effet, ces derniers peuvent être critiqués sur de nombreux aspects tels que : la non prise en compte des différences socioculturelles et des variantes langagières des sujets (Rondal 1997, Grégoire 2006) ; l'évaluation totalement décontextualisée des productions linguistiques, impliquant une évaluation des connaissances et non des usages de la langue ; ou la mise en évidence des déficits et non des compétences des sujets (de Weck & Marro 2010). Les professionnels doivent donc avoir recours à des méthodes complémentaires, telles que celle de l'évaluation du langage en situations dites « spontanées », pour considérer l'ensemble des compétences d'un locuteur. Même si des recommandations existent concernant le recueil et le traitement de données spontanées[12], ces pratiques sont encore rares et difficiles à réaliser par les professionnels. Les principaux freins concernent le temps de traitement des données, la formation à l'analyse des interactions verbales et aux outils disponibles pour les traiter.

Face à ces contraintes, nous pouvons mentionner l'utilisation de deux commandes, EVAL et KIDEVAL, qui ont été élaborées et développées pour évaluer les productions verbales de locuteurs et faciliter la réalisation d'analyses linguistiques. Ces commandes s'exécutent à partir d'un ou de plusieurs fichiers de transcription CLAN (.cha). La présentation des résultats est réalisée sous forme d'un tableur au sein duquel chaque colonne expose les résultats d'une analyse automatique particulière et chaque ligne correspond à un fichier de transcription. Ainsi, une évaluation des compétences d'un patient à un moment T ou à différents moments de la prise en charge (T+1, T+2, etc.) peut être

[12] American Speech-Language-Hearing Association (2004), *Preferred practice patterns for the profession of speech-language pathology [Preferred Practice Patterns]*, www.asha.org/policy.

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

envisagée. Celle-ci permettra également d'offrir, de manière indirecte, des indices de l'efficience de l'intervention orthophonique.

La commande EVAL a été développée pour les cliniciens et les chercheurs cliniciens dans le but de pouvoir avoir recours à l'analyse simple et efficace de productions langagières de patients aphasiques (MacWhinney *et al.* 2010, 2011 ; MacWhinney & Fromm 2016). Elle permet de calculer automatiquement (à partir de la ligne %mor, le plus souvent) 30 mesures dont la durée, le nombre total d'énoncés, la mesure de la MLU (Mean Length of Utterances), les types, tokens ainsi que le type-token ratio (TTR), le nombre total de mots, le nombre total et le pourcentage de chaque partie du discours, etc. Lors de l'application de la commande EVAL, il est envisageable de comparer les données de son corpus avec celles d'une base de données (AphasiaBank[13]), en sélectionnant certains critères : le type d'aphasie (fluente ou non fluente, Broca, Wernicke, mixte…), l'âge du locuteur, le sexe, le type d'activité (langage spontané, discours dirigé…), etc. AphasiaBank est constituée de données de 290 patients aphasiques et 190 locuteurs contrôles en langue anglaise accomplissant diverses tâches de langage (discours spontané, narration d'histoires à partir d'un support imagé, narration d'un conte et de discours procédural). Certains travaux portant sur le français sont actuellement en cours dans le but notamment d'une éventuelle adaptation à la langue française (Sahraoui & Nespoulos 2012 ; Colin & Le Meur 2016). La commande KIDEVAL[14], créée spécifiquement pour l'évaluation des productions enfantines (Ratner & Brundage 2016), permet, quant à elle, d'extraire un tableur constitué de 82 colonnes de résultats dont le nombre total d'énoncés produits, le MLU, les types, tokens et leur ratio (TTR), le nombre moyen de propositions par énoncés, le VocD (indice de diversité lexicale basé sur une estimation probabiliste), la valeur absolue ou le pourcentage de chaque partie du discours (nom, verbe, adjectif, *etc.*).

Ces commandes génèrent les résultats sur la base de la ligne dépendante %mor. Ainsi, il est indispensable de réaliser deux étapes préliminaires, l'une d'analyse automatique en utilisant la commande MOR, l'autre de vérification par le chercheur ou le praticien de l'analyse automatique obtenue. En effet, même si les analyses sont automatiques, il est important de procéder à une relecture voire une révision des analyses proposées. Par exemple, la commande MOR peut ne pas reconnaître un mot et, dans ce cas, celui-ci sera noté « *?|* » associé au mot non reconnu sur la ligne %mor générée automatiquement. Le chercheur ou le praticien devra alors remplacer le point d'interrogation en procédant lui-même à l'analyse morphosyntaxique.

---

[13] Le protocole AphasiaBank ainsi que les informations complémentaires sont disponibles en ligne à l'adresse suivante : http://aphasia.talkbank.org/.

[14] La commande est la suivante : *KIDEVAL +t*CHI +lfra @* où « +l » est utilisé pour indiquer la langue cible (le français ici), « @ » correspond au(x) fichier(s) sur le(s)quel(s) la commande sera utilisée.

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

Par ailleurs, selon les besoins et les objectifs scientifiques et/ou thérapeutiques, la transcription doit être adaptée. En effet, les conventions de transcription et l'utilisation de CLAN incluent le marquage de certaines conduites verbales telles que la répétition de mots, les reformulations sémantiques ou morphosyntaxiques, l'emploi de fillers, l'inintelligibilité, etc. Par exemple, les chercheurs et orthophonistes en aphasiologie coderont au niveau du mot diverses modifications (*e.g.* phonologiques, morphologiques, néologismes), appelées le plus souvent « erreurs », en six catégories, en utilisant le codage suivant [*] (cf. MacWhinney *et al.* 2011). Au niveau de l'énoncé, ces erreurs sont regroupées en cinq catégories (énoncés agrammaticaux, discours syntaxiquement grammatical mais dénué de sens, circonlocutions, jargon et persévérations). Ainsi, si le patient présente des troubles de la fluence, les analyses portant spécifiquement sur ces aspects (*e.g.* répétitions) devront apparaître dans le fichier de transcription. Si, lors de la transcription manuelle, les marques de répétition représentée par '[/]' n'ont pas été annotées, la commande RETRACE réalise l'annotation automatiquement. D'autres codes peuvent être annotés, que ce soit pour la commande KIDEVAL ou EVAL (cf. tableau 1) :

Niveaux	Conventions	Significations	Exemples[15]
Mot	[*] ou [* p]	Modifications phonologiques	[a]*CHI: comment [*] on peut appeler nos personnages [*] ? %pho:tomɑ̃ ɔ̃ pø apəle no pɛsonaʒ
	[* n]	Néologismes	[b]*EXA: racontez-nous ce qui vous arrive. *PAT : mais quoi (..) je suis complètement noge [n*]. *PAT : je n'ai pas nogé [n*].
	0	Omissions d'un mot/argument + fonctions ou natures de l'élément omis (0obj)	[a]*CHI: je nettoie 0obj . [+ gram]
Enoncé	[+ gram]	Enoncés agrammaticaux	[a]*CHI: elle adore [*] 0obj ! [+ gram] %mor: pro:subj|elle v|adorer-PRES&SUB&13s ! %pho: ɛl adɔ
	[+ es]	Enoncés syntaxiquement grammaticaux mais dénués de sens	[b]*PAT : l'argent ça sonne. [+ es]
	[+ jar]	Jargon	[b]*PAT : et il y a une locomotive c'est de l'orthographe. [+ jar]

Tableau 1 : Quelques exemples de conventions utilisées dans le système CHAT en vue d'une évaluation du langage

[15] [a]Exemples extraits d'interactions mère-enfant dysphasique en situation de jeu symbolique (da Silva 2014) ; [b]Exemples extraits de Chomel-Guillaume *et al.* (2010) et adaptés aux conventions de CHILDES. Dans les exemples : *EXA= examinateur, *PAT = patient, *CHI = enfant.

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

Il est important de se conformer aux conventions données (cf. 3.2.) pour que la communauté dispose de connaissances partagées et que les utilisateurs soient sûrs de leur analyse. Toutefois, selon les besoins, certains codes seront préférés à d'autres. D'éventuelles adaptations, qui impliquent la maîtrise du logiciel et une prise de conscience des limites que cela peut engendrer, sont envisageables.

4.2. Analyser les pratiques professionnelles

Parallèlement aux mesures permettant d'évaluer le langage oral des patients, les praticiens peuvent également recueillir des données dans le but de rendre compte de la dynamique des interactions et des effets de leurs pratiques sur le patient. Quel que soit le cadre professionnel et le domaine, l'analyse des postures professionnelles ne peut être menée sans une attitude réflexive. Celle-ci permet de mettre en évidence des actes professionnels, des compétences, une identité construite ou encore le sens des actions menées en contexte. Cela renvoie au concept de « conscientisation des pratiques » (Schön 1993, Canut *et al.* 2013, pour le cadre éducatif) dans lequel il s'agit d'amener les professionnels à atteindre une plus grande expertise sur les modalités du langage adressé (aux élèves, aux patients, etc.). L'une des voies d'accès à cette conscientisation pour les professionnels est de réaliser des enregistrements audio et/ou vidéo dans le but de mesurer l'adéquation de leurs pratiques avec les objectifs langagiers visés. Le professionnel prend ainsi de la distance par rapport à ce qu'il pense avoir produit pour tenter de comprendre ce qui fonctionne ou ne fonctionne pas dans ses échanges avec le patient et pour y apporter des modifications. L'utilisation de la vidéo, en tant qu'outil, devient alors une pratique indispensable pour décrire les gestes professionnels et les évaluer.

Pour illustrer ce point, nous avons sélectionné trois extraits de *corpora* recueillis dans le cadre d'interactions orthophoniste/patient (enfant). Ces trois exemples visent à présenter au lecteur, d'une part, un modèle de transcription adapté pour les professionnels orthophonistes, et, d'autre part, une réflexion générale sur l'évaluation des pratiques professionnelles à partir de données attestées. Dans les extraits suivants, un choix a été opéré pour restreindre la transcription aux informations pertinentes pour l'objectif visé (l'évaluation des pratiques professionnelles) :

- les productions verbales de chacun (lignes principales) ;
- la correspondance phonétique (ligne %pho), pour l'enfant uniquement, en SAMPA[16] ;

[16] Il s'agit d'un jeu de caractères phonétiques qui utilise les symboles d'un clavier classique. Ce système, très simple à mémoriser, évite d'avoir à installer une police phonétique spécifique.

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

- les éléments contextuels, comme les actions (%act) et les gestes (%gpx et %xpnt) ;
- les interprétations possibles en cas d'ambiguïté (%int) ;
- les précisions utiles sur les formes produites (%com).

En outre, conformément aux conventions de transcription de CHILDES, 'yy' (segment sans correspondant orthographique ou soumis à interprétation) est noté sur la ligne principale.

Les deux premiers exemples sont extraits d'une même interaction et d'une même activité en situation d'intervention orthophonique. L'orthophoniste (SLT) propose à l'enfant (CHI) une activité narrative au sein de laquelle il doit trouver des solutions pour résoudre les problèmes rencontrés par les protagonistes. Le support est composé de cartes que l'enfant doit piocher et associer à d'autres.

Ces exemples ont été choisis car ils illustrent une technique interactionnelle courante, à savoir l'usage de reformulations sur un ou plusieurs niveaux linguistiques. Les nombreuses études sur le sujet, en situation clinique comme familiale (voir Clark & Chouinard 2000 et da Silva Genest 2014), décrivent les reformulations comme un moyen efficace pour établir une intercompréhension entre les participants et une forme d'étayage permettant à l'enfant de se saisir des formes linguistiques en vue d'une appropriation dans son propre système. Pour soutenir les processus développementaux, offrir un modèle phonologique et morphosyntaxique, y compris de manière implicite, fait partie des stratégies auxquelles le professionnel doit avoir recours. C'est pourquoi l'orthophoniste doit être en mesure d'identifier le ou les moment(s) opportun(s), d'apprécier l'effet produit sur le discours du patient et de déterminer la façon dont il est attentif à ces formes d'intervention.

Exemple 1 : Extrait d'une interaction entre un orthophoniste (SLT) et un enfant (CHI), âgé de 7 ans 7 mois : cas de reformulation phonologique

*CHI:	0.
%act:	pioche une carte représentant un cadeau
*CHI:	<un yy mais non> [= ! rit] +/!
%pho:	e~ gato mE no~
%act:	montre la carte à SLT
*SLT:	comment [/] comment tu dis qu(e) ça s'appelle ?
*CHI:	un yy ?
%pho:	e~ gato
*SLT:	cadeau.
%com:	insiste sur la première syllabe
*CHI:	cadeau.
%pho:	kado

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

Exemple 2 : Extrait d'une interaction entre un orthophoniste (SLT) et un enfant (CHI), âgé de 7 ans 7 mois : cas d'absence de reformulation

*CHI:	y+a pas un (pe)tit tr(uc) le soleil.
%pho :	ja pa e~ ti tR l@ solEj
%act :	fouille dans les cartes
*SLT:	j(e) sais pas on va voir +…
*CHI :	je sais i(l) nous faut quoi.
%pho :	Z@ sE i nu fo kwa
*SLT :	qu'est+ce qu'il nous faudrait ?
*CHI :	une casquette ou un chapeau.
%pho :	yn@ kasKEt u e~ Sapo
*SLT :	mmh

Ces deux exemples montrent la façon dont cet orthophoniste construit son geste professionnel. Dans l'exemple 1, la production non conventionnelle de *cadeau* (prononcé [gato]) pouvait créer une confusion lexicale entre deux lexèmes existants qui ont des signifiés différents (*cadeau* vs *gâteau*). L'orthophoniste reformule alors les propos de l'enfant en apportant la forme conventionnelle du terme cible *cadeau*, qui est reprise à l'identique par l'enfant. A l'inverse, dans l'exemple 2, l'orthophoniste ne reformule pas les propos de l'enfant alors même que l'énoncé est non conventionnel, comme dans le premier extrait. Cela s'explique par le fait que la production de l'enfant dans le premier énoncé (*y+a pas un (pe)tit tr(uc) le soleil*) ne génère pas une difficulté de compréhension et n'entrave donc pas la communication entre les deux participants, contrairement à l'énoncé produit par l'enfant dans l'exemple 1.

Ces deux exemples isolés illustrent donc deux conduites récurrentes du professionnel au cours de cette activité, à savoir que la reformulation des propos de l'enfant intervient lorsqu'il y a un obstacle à une bonne compréhension (exemple 1) mais pas lorsqu'il n'y a pas de risque d'entrave (exemple 2). Les conduites de l'orthophoniste visent davantage l'intercompréhension. On peut donc affirmer que la façon dont l'orthophoniste déploie ou non une stratégie d'étayage de type « reformulation » a un sens dans ce contexte. Cette affirmation ne peut se faire qu'à partir d'une analyse interactionnelle *a posteriori* et en dégageant la régularité de ces gestes. Ces observations sont à la base d'une conscientisation des conduites du professionnel et pourra faire l'objet d'un questionnement sur celles-ci, comme par exemple sur les occasions les plus opportunes pour avoir recours à des stratégies de reformulations.

Dans le troisième exemple, issu d'un corpus différent par rapport aux exemples précédents, le praticien présente à l'enfant des images décrivant des objets et des scènes de la vie quotidienne qu'il associe à des gestes issus du répertoire de la Langue des Signes Française (entre crochets).

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

Exemple 3 : Extrait d'une interaction entre un orthophoniste (SLT) et un enfant (CHI) âgé de 3 ans 10 mois : cas de reformulations syntaxiques

*SLT:	maman des fois elle prend peut+être un parapluie .
%gpx:	[MAMAN]
%com:	insiste sur maman et parapluie
*CHI:	<oui [=! fait oui de la tête] et papa> [>] non[=! fait non de la tête] .
%pho:	i e papa na~
*SLT:	<hein quand i(l)> [<] +/.
*SLT:	<et papa non> [=! fait non de la tête] bah ouais c'est les mamans des fois qui ont des parapluies .
%gpx:	[MAMAN]
*SLT:	et les bottes [//] qui met les bottes c'est toi hein ?
%xpnt:	show une image
%gpx:	[TOI]
*CHI:	oui [=! en souriant] .
%pho:	wi
*SLT:	<oui je sais qu(e) t(u) as des bottes> [=! en riant] .
*CHI:	+< xxx bottes .
%pho:	X pOt
%xpnt:	show ses pieds
*SLT:	ah <ouais> [/] ouais des bottes !
*SLT:	tiens .
%act:	donne une image à CHI
*CHI:	maman xxx à nous .
%pho:	mama~ X a nu
%int:	nous/3/
%act:	pose l'image sur l'image correspondante
*SLT:	il pleut regarde il pleut.
%gpx:	[PLEUVOIR]
*CHI:	0 .
%gpx:	[PLEUVOIR]
*CHI:	et yyy +/.
%pho:	e ma
%xpnt:	show ses pieds
*SLT:	et maman elle met les bottes voilà .
*SLT:	quand il pleut hein quand il fait mauvais il pleut hof@i !
%gpx:	[PLEUVOIR]
*CHI:	et maman e@fs met bottes .
%pho:	e mama~ E me bOke

L'exemple montre un échange dans lequel le praticien cherche à déclencher, par un rapprochement de la situation actuelle à des connaissances réelles ou supposées de l'enfant, une verbalisation de la part de celui-ci. Il reprend, en le validant, l'essai de l'enfant et en y ajoutant des éléments lexicaux et grammaticaux (par exemple, il insère l'élément lexical *bottes* dans un énoncé qui établit un lien avec le vécu

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

de l'enfant). Il associe également des gestes (MAMAN, PLEUVOIR) et des pointages qui vont renforcer les éléments saillants des énoncés. Du côté de l'enfant, on constate une reprise d'éléments présents dans les énoncés de l'adulte pour parvenir à former l'énoncé *et maman e@ fs met bottes*, ainsi que la reprise d'éléments non verbaux. Dans cet échange, la mise en contexte effectuée par l'adulte entraîne une prise de parole plus longue de l'enfant et une mobilisation des ressources non verbales dont il dispose.

Amener les professionnels à une conscientisation de leurs pratiques professionnelles nécessite des outils d'analyse et le développement de « savoir-analyser ». Si l'une des questions principales du professionnel est de déterminer comment évaluer l'efficacité de son intervention, alors il pourra trouver dans la linguistique outillée des « instruments » pour répondre à cette question. Si on pense au troisième extrait présenté ci-dessus, il peut être pertinent, par exemple, d'apprécier l'occupation de l'espace discursif en mesurant le nombre de mots produits par les locuteurs (calcul automatique *via* le TTR) et la qualité des échanges (*e.g.* les chevauchements de parole), d'analyser finement les actes de langage réalisés par les locuteurs en utilisant la ligne dépendante %spa, etc. Quelle que soit l'analyse quantitative effectuée, un traitement qualitatif des échanges langagiers complémentaires est nécessaire pour rendre compte de leur richesse et de leurs effets sur le discours du patient et ses acquisitions.

5. Discussion et conclusion

Notre propos était de montrer en quoi l'étude sur la base de corpus constitue un atout à la fois pour la description de faits linguistiques et pour la réflexion des professionnels sur leurs pratiques langagières. Si nous avons centré notre réflexion sur l'utilisation du logiciel CLAN, qui est notre principal outil de travail pour nos recherches, c'est en raison des potentialités qu'il offre pour l'étude des productions en situation spontanée. Toutefois, d'autres outils – évoqués en 3.2. – contribuent aussi à l'étude des données orales spontanées.

Dans le cadre des ressources disponibles pour l'acquisition du langage, la construction collaborative de CHILDES, cité précédemment, a permis de mettre en relation des chercheurs de divers horizons sur des *corpora* présentés avec des conventions unifiées. En France, le développement d'ORTOLANG (Pierrel & Parisse 2016) permet à présent aux chercheurs qui le souhaitent de déposer sur la plateforme des ressources et des outils issus de leurs travaux, en vue de les diffuser et de les partager avec d'autres chercheurs. Ainsi, le chercheur travaillant dans le champ des pathologies du langage peut avoir accès à quelques *corpora* constitués dans le cadre de projets de recherche,

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

parmi lesquels nous pouvons citer (sans être exhaustives) PLPNat (Bassano 2007), Paroles Disfluentes (Praxiling) ou encore DySpoLec (Lalain *et al.* 2012).

Malgré ce développement, la diffusion de données de productions linguistiques de locuteurs (enfants ou adultes) présentant des troubles du langage reste encore parcellaire : les chercheurs s'accordent sur le fait que les données recueillies sont peu nombreuses et très rarement partagées au sein de la communauté scientifique, notamment pour les données francophones. Or, rendre accessible ces données en interaction contribuerait à la mise en place de recherches portant sur le développement langagier pathologique, en proposant aux chercheurs et aux cliniciens de travailler sur des données écologiques, recueillies dans un cadre scientifique et totalement annotées.

La recherche sur des données orales montre que les *corpora* de langue parlée permettent de compléter et d'objectiver les intuitions du chercheur sur la langue (Blanche-Benveniste 1996). Cette démarche n'est pas spécifique au cadre de la recherche, elle est tout aussi indispensable pour le professionnel. En effet, la création de grands *corpora* permettrait aux professionnels de comparer leurs données à d'autres, en sélectionnant des critères pertinents (activités, âges, pathologies ou non, etc.) et de déterminer si l'écart relevé entre deux temps d'évaluation/observation est significatif par rapport au patient lui-même ou à ce qui est attendu compte-tenu de son âge, de son niveau linguistique et/ou de sa pathologie. En outre, l'accès à des *corpora* attestés est tout aussi important en amont de la pratique, dans le cadre de la formation initiale, car ils permettent aux futurs professionnels de s'approprier et de maîtriser les outils linguistiques mis à disposition, ainsi que de s'exercer à l'analyse de situations cliniques diverses.

Le recours à l'enregistrement audio et à des formes de transcriptions est courant en orthophonie. L'étude de Kemp & Klee (1997), réalisée auprès de 250 orthophonistes américains, montre ainsi que 85% d'entre eux s'appuient sur le langage spontané dans leurs pratiques et qu'ils transcrivent presque tous les données recueillies (au cours des séances le plus souvent), en se basant sur 50 énoncés ou 15 minutes. Les professionnels de santé opposent souvent à cette pratique de recueil, de transcription et d'analyses une limite de temps. Or, transcrire ses données sous CLAN plutôt que sur un traitement de texte ne demande pas plus de temps, même si des adaptations sont nécessaires, et assure l'obtention de résultats fiables et fidèles. Le praticien pourra ainsi adopter quelques principes généraux, adaptés à ses objectifs et ses besoins, comme, par exemple, ne pas réaliser l'alignement texte/vidéo que permet CLAN ou limiter le nombre de lignes dépendantes.

Les analyses automatiques proposées par CLAN offrent en outre de nombreuses possibilités. Toutefois, les résultats obtenus,

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

notamment dans le cas du français, ne peuvent pas être comparés à une population de sujets contrôles. Par conséquent, d'autres questions doivent être soulevées : est-ce un frein à une démarche professionnelle ? Comment un praticien peut-il interpréter les résultats obtenus ?

Cet article s'inscrit dans deux perspectives : une première qui est tournée vers la pratique clinique et une seconde vers la recherche clinique. En effet, il s'agit à la fois : d'une première piste de réflexion pour les praticiens orthophonistes qui cherchent de nouveaux outils pour évaluer les compétences langagières de patients et/ou leurs pratiques professionnelles et d'un objet d'étude pertinent pour le chercheur en linguistique en vue d'approfondir les connaissances actuelles sur les interactions verbales, les troubles du langage et les processus de remédiation.

Références bibliographiques

Barras, C., Geoffrois, E., Wu, Z., Liberman, M. (2000), "Transcriber: development and use of a tool for assisting speech corpora production", *Speech Communication*, 33/1-2), p. 1-28.

Bassano, D. (2007), « Emergence et développement du langage : enjeux et apports des nouvelles approches fonctionnalistes », in Dumont, E., Metz-Lutz, M. N. (éds), *L'acquisition du langage et ses troubles*, Solal Editeurs, Marseille, p. 13-46.

Baude, O. (2006), *Corpus oraux: guide des bonnes pratiques*, Presses Universitaires d'Orléans.

Bilger, M., Blasco, M., Cappeau, P., Pallaud, B., Sabio, F., Savelli, M. J. (1997), « Transcription de l'oral et interprétation. Illustrations de quelques difficultés », *Recherches sur le français parlé*, 14, p. 57-86.

Blanche-Benveniste, C. (1996), « De l'utilité du corpus linguistique », *Revue française de linguistique appliquée*, I/2, p. 25-42.

Blanche-Benveniste, C. (1997), *Approches de la langue parlée en français*, Ophrys, Paris.

Blanche-Benveniste, C. & Bilger, M. (1999), « Français parlé – oral spontané. Quelques réflexions », *Revue française de linguistique appliquée*, 4/2, p. 1-30.

Blanche-Benveniste, C. & Jeanjean, C. (1987), *Le français parlé. Transcription et édition,* INALF/Didier Érudition, Paris.

Boersma, P., Weenink, D. (2009), *Praat: doing phonetics by computer* (Version 5.1. 05) [Computer program], Retrieved May 1, 2009.

Boulton, A. (2009), « Documents authentiques, oral, corpus », *Mélanges CRAPEL*, 31, p. 5-13.

Boulton, A., Canut, E., Guerin, E., Parisse, C., Tyne, H. (2013), « Corpus et appropriation de L1 et L2 », *Linx*, 68-69, p. 9-32.

Brown, R. W. (1973), *A first language: the early stages*, Harvard University, Cambridge.

Brugman, H., Russel, A., Nijmegen, X. (2004), "Annotating multi-media/ multimodal resources with ELAN", *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, p. 2065-2068.

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

Bruner, J. S. (1983), *Le développement de l'enfant : savoir faire, savoir dire*, Presses Universitaires de France, Paris.
Canut, E., Espinosa, N. Vertalier, M. (2013), « Corpus et prise de conscience des processus interactionnels d'apprentissage du langage pour repenser les pratiques enseignantes en maternelle », *Linx*, 68-69, p. 69-93.
Canut, E., Masson, C., Leroy-Collombel, M. (2017), « Efficience de l'étayage dans les pratiques professionnelles : effets sur le développement du langage d'enfants atteints de déficience intellectuelle », *Langage & Pratiques*, 59, p. 69-86.
Canut, E., Vertalier, M. (2008), « Des données *représentatives*... De quoi en acquisition du langage ? Constitution de données à observer et objectifs d'analyse », *Verbum*, 30/4, p. 299-312.
Chomel-Guillaume, S., Leloup, G., Bernard, I. (2010), *Les aphasies. Evaluation et rééducation*, Masson, Issy-les-Moulineaux.
Clark, E. V., Chouinard, M. (2000), « Enoncés enfantins et reformulations adultes dans l'acquisition du langage », *Langages*, 140, p. 9-23.
Cohen, M. (1925/1993), « Sur les langages successifs de l'enfant », *L'Acquisition du Langage Oral et Ecrit*, 31, p. 33-43.
Colin, C., Le Meur, C. (2016), *Adaptation du projet APHASIABANK à la langue française – Contribution pour une évaluation informatisée du discours oral de patients aphasiques*, Mémoire d'orthophonie, Université Paul Sabatier – Toulouse III.
Darwin, C. (1877), « A biographical sketch of an infant », *Mind*, 7, p. 285-294.
da Silva Genest, C. (2014), « Les reformulations en situation de rééducation orthophonique », *Travaux Neuchâtelois de Linguistique*, 60, p. 137-148.
Debaisieux, J.-M., Benzitoun, C., Delofeu, H. J. (2016), « Le projet ORFEO : Un corpus d'études pour le français contemporain », *Corpus*, 15, p. 91-114.
de Weck, G. (2003), « Pratiques langagières, contextes d'interaction et genres de discours en logopédie / orthophonie », *TRANEL*, 38-39, p. 25-48.
de Weck, G., Marro, P. (2010), *Les troubles du langage chez l'enfant. Description et évaluation,* Elsevier Masson, Issy-les-Moulineaux.
de Weck, G., Salazar-Orvig, A. (2010), « Les interactions mère-enfant dysphasique : qu'y a-t-il encore à comprendre ? », *Langage & Pratiques*, 46, p. 7-16.
Falbo, C. (2005), « La transcription: une tâche paradoxale », *The interpreters' Newsletter*, 13, p. 25-38.
François, D. (1977), « Du pré-signe au signe », in François, F., François, D., Sabeau-Jouannet, E., Sourdot, M. (éds), *La syntaxe de l'enfant avant 5 ans,* Larousse, Paris, p. 53-89.
François, F. (1993), *Pratiques de l'oral*, Nathan, Paris.
Gadet, F. (2003), *La variation sociale en français,* Ophrys, Paris.
Goodwin, C. (2000), "Gesture, aphasia, and interaction", in McNeill, D. (éd.), *Language and Gesture*, Cambridge University Press, Cambridge, p. 84-98.
Grégoire, J. (2006), « Propriétés métriques des tests de langage et leurs implications pratiques », in Estienne, F., Piérart, B. (éds), *Les bilans de langage et de voix,* Masson, Paris, 14-26.
Kemp, K. & Klee, T. (1997), "Clinical language sampling practices: results of a survey of speech-language pathologists in the United States », *Child Language Teaching and Therapy*, 13, p. 161-176.

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

Karmiloff, A. & Karmiloff-Smith, K. (2001), *Comment les enfants entrent dans le langage*, Retz, Paris.
Labov, W. (1972), *Sociolinguistic patterns*, University of Pennsylvania Press, Philadelphia.
Lalain, M., Mendonça-Alves, L. Espesser, R., Ghio, A., De Looze, C., Reis, C. (2012), « Lecture et prosodie chez l'enfant dyslexique, le cas des pauses », Besacier, L., Lecouteux, B., Sérasset, G. (éds), *Actes de la conférence conjointe JEP-TALN-RECITAL*, vol. 1, p. 41-48.
Lallier, C. (2011), « L'observation filmante. Une catégorie de l'enquête ethnographique », *L'Homme*, 198-199, p. 105-130.
Leech, G. (1992), "Corpora and theories of linguistic performance", in Svartvik, J. (éd.), *Directions in corpus linguistics: proceedings of Nobel symposium 82*, Mouton de Gruyter, Berlin / New York, p. 125-148.
Lejeune, C. (2010), « Montrer, calculer, explorer, analyser. Ce que l'informatique fait (faire) à l'analyse qualitative », *Recherches Qualitatives*, 9, p. 15-32.
MacWhinney, B., Fromm, D. (2016), "AphasiaBank as BigData", *Seminars in Speech and Language*, 37, p. 10-22.
MacWhinney, B., Fromm, D., Forbes, M., Holland, A. (2011), "AphasiaBank: Methods for studying discourse", *Aphasiology*, 25/11, p. 1286-1307.
MacWhinney, B., Fromm, D., Holland, A., Forbes, M., Wright, H. (2010), "Automated analysis of the Cinderella story", *Aphasiology*, 24/ 6-8, p. 856-868.
Masson, C. (2014), « Repérage précoce des dysfonctionnements langagiers: enjeux et élaboration d'une action de prévention des troubles du langage au sein d'un Centre d'Action Médico-Sociale Précoce (CAMSP) », *Enfance*, 2, p. 171-187.
Mondada, L. (1998), «Technologies et interactions dans la fabrication du terrain du linguiste », *Cahiers de l'ILSL*, 10, p. 39-68.
Mondada, L. (2000), « Les effets théoriques des pratiques de transcription », *Linx*, 42, p. 131-146.
Mondada, L. (2008), « La transcription dans la perspective de la linguistique interactionnnelle », in Bilger, M. (éd.), *Données orales, les enjeux de la transcription,* Presses Universitaires de Perpignan, Perpignan, p. 78-109.
Morgenstern, A. (2009), *L'enfant dans la langue*, Presses de la Sorbonne Nouvelle, Paris.
Morgenstern, A. (2016). « Pratiques langagières et comportements du patient en milieu familial : apport des méthodes ethnographiques multimodales pour la recherche en médecine », *Ethics, Medicine and Public Health*, 2, p. 641-649.
Morgenstern, A., Parisse, C. (2007), « Codage et interprétation du langage spontané d'enfants de 1 à 3 ans », *Corpus*, 6, p. 55-78.
Morgenstern, A., Parisse, C. (2012), "The Paris Corpus", *French language studies*, 22, p. 7-12.
Ochs, E. (1979), « Transcription as theory », *Developmental pragmatics*, 10/1, p. 43-72.
Onnis, L. (2014), "Corpus-based method", in Brooks, P. J., Kempe, V. (éds), *Encyclopedia of Language development*, SAGE Publications, p. 111-114.
Parisse, C., Le Normand, M.T. (1998), « Traitement automatique de la morphosyntaxe chez le petit enfant », *Glossa*, 61, p. 9-22.
Pierrel, J. M., Parisse, C. (2016), « ORTOLANG: a French infrastructure for Open Resources and Tools for LANGuage", *5th CLARIN Annual Conference*, Aix-en-Provence.

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)

Ratner, N. B., Brundage, S. B. (2016), *A Clinician's Complete Guide to CLAN and PRAAT* (on line: talkbank.org/manuals/Clin-CLAN.pdf).
Rondal, J. A. (1997), *L'évaluation du langage,* Mardaga, Sprimont.
Rose, Y., MacWhinney, B. (2014), « The PhonBank initiative », in Durand, J., Gut, U., Kristoffersen, G. (éds), *The Oxford Handbook of Corpus Phonology*, Oxford University Press, Oxford, p. 380-401.
Sahraoui, H., Nespoulous, J. (2012), "Across-task variability in agrammatic performance", *Aphasiology*, 26/6, p. 785-810.
Schön, D. (1993), *Le praticien réflexif. À la recherche du savoir caché dans l'agir professionnel*, Éditions Logiques, Montréal.
Sinclair, J. (1996), *Preliminary recommendations on corpus typology*, Technical report EAGLES (Expert Advisory Group on Language Engineering Standards).
Teubert, W. (2009), « La linguistique de corpus : une alternative », *Semen*, 27 (en ligne : http://semen.revues.org/8923).
Tomasello, M., Stahl, D. (2004), "Sampling children's spontaneous speech: how much is enough?", *Journal of child language*, 31/1, p. 101-121.
Traverso, V. (2016), *Décrire le français parlé en interaction*, Editions Ophrys, Paris.
Veneziano, E. (2000), « Interaction, conversation et acquisition du langage dans les trois premières années », in Kail, M., Fayol, M. (éds), *L'acquisition du langage. Le langage en émergence de la naissance à trois ans*, Presses Universitaires de France, Paris, p. 231-265.
Veneziano, E. (2014), « Interactions langagières, échanges conversationnels et acquisition du langage », *Contraste*, 39, p. 31-49.
Vygotski, L. S. (1997 [1934]), *Pensée et langage*, La Dispute, Paris.

BDD-A27748 © 2017 Facultatea de Litere din Oradea
Provided by Diacronia.ro for IP 216.73.216.239 (2026-11-05 11:25:18 UTC)