

La collecte de corpus d'interprétation : un défi permanent

Caterina Falbo

Volume 63, numéro 3, décembre 2018

Traductologie de corpus : 20 ans après

URI : <https://id.erudit.org/iderudit/1060167ar>

DOI : <https://doi.org/10.7202/1060167ar>

[Aller au sommaire du numéro](#)

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)

1492-1421 (numérique)

[Découvrir la revue](#)

Citer cet article

Falbo, C. (2018). La collecte de corpus d'interprétation : un défi permanent. *Meta*, 63(3), 649–664. <https://doi.org/10.7202/1060167ar>

Résumé de l'article

Depuis l'article programmatique de Shlesinger (1998), beaucoup de corpus d'interprétation ont vu le jour. Il s'agit de corpus constitués d'interprétations effectuées dans des contextes de conférences, dans le domaine médical, ou encore dans des situations institutionnelles comme la télévision. Malgré les différences portant sur le contexte situationnel, ils présentent des traits communs sur le plan de leur dimension, qui est en général limitée, et de leur exploitation, qui est, quant à elle, conditionnée par la présence ou l'absence de l'alignement des transcriptions avec l'audio ou l'image, de même que par le niveau d'étiquetage choisi. Nous nous proposons de réfléchir à ces aspects, en illustrant brièvement les caractéristiques de deux corpus collectés à l'Université de Trieste tels que CorIT (corpus d'interprétation à la télévision italienne) et CorILex (corpus d'interprétation judiciaire) et en mettant en évidence les obstacles qui ont entravé ou qui conditionnent leur collecte. Nos remarques conclusives porteront enfin sur les traits qui différencient ces deux corpus dans le but de les articuler avec les réflexions qui se sont développées au sein de la linguistique de corpus afin de déterminer la place des corpus d'interprétation à l'intérieur d'un domaine qui se montre de plus en plus diversifié, hétérogène et interdisciplinaire.

La collecte de corpus d'interprétation : un défi permanent

CATERINA FALBO

Università di Trieste, Trieste, Italie
cfalbo@units.it

RÉSUMÉ

Depuis l'article programmatique de Shlesinger (1998), beaucoup de corpus d'interprétation ont vu le jour. Il s'agit de corpus constitués d'interprétations effectuées dans des contextes de conférences, dans le domaine médical, ou encore dans des situations institutionnelles comme la télévision. Malgré les différences portant sur le contexte situationnel, ils présentent des traits communs sur le plan de leur dimension, qui est en général limitée, et de leur exploitation, qui est, quant à elle, conditionnée par la présence ou l'absence de l'alignement des transcriptions avec l'audio ou l'image, de même que par le niveau d'étiquetage choisi. Nous nous proposons de réfléchir à ces aspects, en illustrant brièvement les caractéristiques de deux corpus collectés à l'Université de Trieste tels que CorIT (corpus d'interprétation à la télévision italienne) et CorLex (corpus d'interprétation judiciaire) et en mettant en évidence les obstacles qui ont entravé ou qui conditionnent leur collecte. Nos remarques conclusives porteront enfin sur les traits qui différencient ces deux corpus dans le but de les articuler avec les réflexions qui se sont développées au sein de la linguistique de corpus afin de déterminer la place des corpus d'interprétation à l'intérieur d'un domaine qui se montre de plus en plus diversifié, hétérogène et interdisciplinaire.

ABSTRACT

Since Shlesinger's (1998) programmatic article, many interpreting corpora have been developed. Generally, they include interpreted communication instances taking place in various settings, such as conferences, healthcare, and other institutional contexts, such as television. Notwithstanding differences concerning setting and communicative situation, all these corpora show common traits as far as size and interrogation aspects are concerned: more specifically, they have limited size, and their interrogation possibilities are influenced by the presence – or absence – of transcript-audio alignment and tagging level. This paper focuses on these aspects by presenting two corpora collected at the University of Trieste, namely CorIT (Italian Television Interpreting Corpus) and CorLex (Legal Interpreting Corpus). The main differences distinguishing these two corpora will be discussed in relation to the latest developments of research in corpus linguistics. Our goal is to identify the place interpreting corpora currently occupy within corpus linguistics, a domain that is becoming increasingly diverse, multifaceted, and multidisciplinary.

RESUMEN

Desde la publicación del artículo programático de Shlesinger (1998), muchos corpus de interpretación se han desarrollado. Generalmente, estos corpus incluyen interpretaciones llevadas a cabo en distintos entornos, como las conferencias, las interacciones en el ámbito socio-sanitario y otros contextos institucionales, como la televisión. A pesar de las diferencias relativas al contexto y a la situación comunicativa, todos estos corpus presentan características comunes en términos de tamaño y aspectos de búsqueda. En particular, se trata de corpus de dimensiones generalmente reducidas y cuyas posibilidades de búsqueda dependen de la presencia o ausencia de alineaciones de textos transcritos a partir de grabaciones y del etiquetado elegido. La presente comunicación

se centra en estos aspectos presentando dos corpus compilados en la Universidad de Trieste: CorIT (corpus de interpretación televisiva) y CorILex (corpus de interpretación jurídica). Se presentarán las diferencias más relevantes entre estos dos corpus en relación a los últimos desarrollos en la investigación sobre lingüística de corpus. El objetivo es delimitar el espacio ocupado por los corpus de interpretación en la lingüística de corpus, una disciplina cada vez más variada, polifacética y multidisciplinar.

MOTS-CLÉS/KEYWORDS/PALABRAS CLAVE

corpus d'interprétation, transcription, alignement, interprétation à la télévision, interprétation judiciaire

interpreting corpora, transcription, alignment, television interpreting, legal interpreting corpus de interpretación, transcripciones, alineación, interpretación televisiva, interpretación jurídica

Tout le monde reconnaît que la constitution de corpus consomme un temps considérable, et tout le monde n'a pas le goût d'y consacrer son temps et ses forces. De fait, nombre de linguistes se demandent où «trouver» des corpus, comme s'il s'agissait d'une opération purement utilitaire. (Cappeau et Gadet 2007 : 108)

1. Propos préliminaires

La construction de corpus d'interprétation¹ est un projet ambitieux mais qui gagne en popularité depuis environ deux décennies, car ces corpus permettent aux chercheurs d'observer et de mettre au jour, à une échelle écologique, les caractéristiques propres à l'interprétation. Il existe déjà un certain nombre de corpus, tirés de divers contextes d'interprétation :

1. Interprétation de conférence
 - a. European Parliament Interpreting Corpus (EPIC) (Russo, Bendazzoli, *et al.* 2012)
 - b. European Parliament Interpreting Corpus-Ghent (EPICG) (Bernardini, Ferraresi, *et al.* 2018)
 - c. Directionality in Simultaneous Interpreting Corpus (DIRSI-C) (Bendazzoli 2012)
 - d. Corpus Football in Europe (FOOTIE) (Sandrelli 2012)
 - e. CoSI-corpus (House, Meyer, *et al.* 2012)
2. Interprétation de dialogue en contexte médical
 - a. DiK-corpus (Schmidt et Meyer 2009)
 - b. AIM corpus (Gavioli 2015; Niemants, dans le présent volume)
 - c. HCIQ.1415 (Dal Fovo 2017, 2018)
3. Interprétation de dialogue en contexte judiciaire
 - a. Corpus d'audiences judiciaires italiennes (Biagini 2012)
4. Interprétation télévisuelle
 - a. Corpus of Television Interpreting (CorIT) (Falbo 2012a), qui comprend, à la fois, de l'interprétation de conférence et de dialogue.

Si on observe les caractéristiques de ces corpus et le chemin théorique et pratique parcouru par leurs auteurs, nous pouvons identifier au moins deux aspects qui, d'une part, ont déterminé – sous différentes formes – la nature de ces corpus et qui, d'autre part, constituent un défi permanent. Il s'agit de la taille, qui renvoie, d'une certaine

manière, à la question de la représentativité, et de la transcription, qui exige que l'on prenne en compte les enjeux d'interrogation des corpus.

Si les grands corpus oraux ont du mal à atteindre la taille des grands corpus de textes écrits pour des raisons bien connues, la dimension des corpus d'interprétation est, quant à elle, bien inférieure à celle des grands corpus oraux. Une explication possible réside, sans aucun doute, dans le décalage temporel qui sépare la collecte de corpus de langue orale de celle de corpus d'interprétation. En effet, ce n'est qu'à partir de la première décennie des années 2000 qu'on a commencé à réfléchir et à travailler concrètement à la collecte de corpus au sein des études sur l'interprétation (*interpreting studies*, selon la dénomination courante). Toutefois, d'autres facteurs conditionnent la formation de corpus d'interprétation, par exemple la difficulté d'accéder aux lieux de l'interprétation et, à quelques exceptions près, l'impossibilité d'enregistrer les performances des interprètes dans des contextes sensibles, comme le domaine médical ou judiciaire ou encore le refus des interprètes d'être enregistrés². Très souvent c'est grâce à une relation de confiance tissée au fil des mois, voire des années, et ancrée aussi bien dans une série d'autorisations de la part des institutions impliquées que dans des garanties fournies par le chercheur, que ce dernier est autorisé à enregistrer des interactions avec la présence d'un interprète. La taille réduite de ces corpus met en cause leur caractère représentatif en ce qui a trait à l'interprétation dans sa globalité mais aussi au domaine spécifique qu'ils essaient de saisir. Sans aucun doute le débat sur la représentativité des corpus d'interprétation peut tirer profit des réflexions qui ont animé le domaine de la linguistique de corpus et qui ont mis en cause la capacité des grands corpus de représenter la langue dans sa globalité (Cappeau et Gadet 2007; Arbach et Ali 2013). Nous y reviendrons.

La transcription demeure une phase délicate et fondamentale. Le choix du logiciel à utiliser et la mise au point des conventions de transcription sont désormais soumis à l'exigence de faciliter le repérage automatique de certains phénomènes linguistiques (voire interactionnels); encore faut-il que les choix opérés n'altèrent pas la nature des données à cause d'une standardisation plus ou moins poussée ou d'une représentation caricaturale (Ochs 1979/1999; Jefferson 1983; Blanche-Benveniste et Jeanjean 1987; Blanche-Benveniste 1997; Gadet 2008), qui pourrait avoir des retombées considérables et un effet de distorsion sur l'analyse elle-même. Tout cela exige par conséquent une démarche réfléchie, basée sur plusieurs essais et des projets pilotes³. La phase de transcription comporte également le choix entre la mise au point de *spoken corpora* ou de *speech corpora* (Zanettin 2009) (ce qui pourrait donner en français *corpus parlé* vs *corpus parlant*). Dans le premier, la transcription des données n'est pas alignée avec leur enregistrement audio ou vidéo⁴. Cela comporte l'écrasement, voire l'effacement, d'un aspect fondamental et constitutif de l'interprétation, à savoir sa dimension spécifique orale. C'est donc vers la création de *speech corpora*, c'est-à-dire de corpus oraux où la transcription est alignée avec la source audio ou vidéo correspondante, que la recherche s'oriente désormais, avec tout ce que cela comporte en termes d'efforts, d'énergie et de ressources.

À tout cela il faut ajouter la présence ou l'absence d'étiquetage, soit-il grammatical et/ou syntaxique. Très souvent, cela est le résultat d'un choix réfléchi dû aux intérêts spécifiques des chercheurs. Ceux qui travaillent sur l'interaction, par exemple, sont moins attirés par les caractéristiques morphosyntaxiques ou lexicales que par « les procédures (qui ne sont pas exclusivement verbales) à travers lesquelles

les interlocuteurs produisent et comprennent leurs propres comportements sociaux ainsi que ceux d'autrui» (Zanettin 2009: 330, notre traduction⁵). Ce sont, en effet, ces aspects-là qui constituent l'objet de recherche des analystes de la conversation (Sacks, Schegloff, *et al.* 1974) et qui posent problème quant à leur repérage automatique à l'intérieur d'un corpus (Niemants, dans le présent volume).

Tous ces aspects, qui, d'une manière ou d'une autre, nous ramènent aux caractéristiques des corpus mentionnés plus haut, sont à la base d'une question portant sur le statut de ces *recueils de discours interprétés*: s'agit-il vraiment de corpus au sens propre du mot? La réponse dépend de la définition choisie pour le mot *corpus* et, bien entendu, de l'approche adoptée en rapport avec des *faits de langue*. D'après une définition stricte du mot – selon laquelle un corpus est un recueil de textes écrits, oraux ou multimodaux informatisé et étiqueté de façon adéquate – certains de ces *corpus* ne seraient que des « pré-corpus » ou des « corpus préhistoriques » (Barbera, Corino, *et al.* 2007: 70), c'est-à-dire des archives multimodales qui pourraient, à la suite d'interventions ciblées, se transformer en de véritables corpus. C'est grâce à des définitions plus inclusives et, entre autres, axées sur la dimension orale, telles que celles données par Blanche-Benveniste (2000a: 12) – « choix organisé » de données orales – ou Baude (2006) – « collections ordonnées d'enregistrements de productions linguistiques orales et multimodales » – que nous nous approprions le mot et la notion de *corpus*.

Au vu de ces préalables, nous essayerons ici de mettre en évidence ce qui dans la compilation de corpus d'interprétation constitue un véritable défi. Pour ce faire, nous nous pencherons sur les spécificités de deux corpus que nous connaissons de près et qui présentent des caractéristiques très différentes, CorIT (section 2.1) et CorILex (section 2.2). Parmi les corpus d'interprétation collectés jusqu'à présent – et que nous avons énumérés plus haut –, ils représentent, à notre avis, deux typologies assez (proto)typiques se situant, si l'on veut, aux deux pôles d'un continuum sur lequel il serait possible de mesurer le degré de facilité/difficulté résultant de l'enchevêtrement de différentes caractéristiques qui concernent la collecte de corpus d'interprétation. C'est à travers la description de ces deux corpus que nous nous focaliserons sur les éléments qui entrent en jeu lors de la collecte d'un corpus d'interprétation et qui très souvent apparaissent comme des entraves insurmontables. Autrement dit, nous illustrerons et essayerons d'expliquer les choix opérés, ou subis, quant à la sélection des objets qui composent ces deux corpus, à leur taille et représentativité, au système de transcription. Nous nous arrêterons en particulier sur le deuxième, CorIT ayant déjà fait l'objet de nombreuses publications (Straniero Sergio et Falbo 2012; Dal Fovo 2011). Nous terminerons par des remarques portant sur les caractéristiques qui différencient CorIT et CorILex en les articulant avec les réflexions qui se sont développées à l'intérieur du domaine de la linguistique de corpus (section 3). En d'autres mots, ce que nous nous fixons comme but ici, c'est de montrer que, au-delà des analogies ou des différences qui caractérisent les corpus d'interprétation, il est possible de faire émerger un faisceau de contraintes qui, *mutatis mutandis*, sont propres à la collecte de corpus d'interprétation.

2. CorIT et CorILEX : un objectif commun, deux parcours spécifiques

2.1. CorIT : caractéristiques et conditions de collecte

CorIT a été collecté au cours des années 2000 et compte maintenant à peu près 2700 enregistrements⁶ d'interprétations effectuées à la télévision italienne depuis les années 50 jusqu'à nos jours. CorIT a été collecté grâce au dépouillement des archives de la RAI (Radio-Télévision italienne) et à l'enregistrement en direct d'émissions passées aussi bien sur les chaînes nationales que sur des chaînes commerciales italiennes. CorIT est donc composé d'enregistrements d'interprètes traduisant d'une langue étrangère⁷ vers l'italien et, en partie, par les productions correspondantes en langue originale. La présence réduite des productions en langue originale est due à deux ordres de facteurs, l'un externe et l'autre interne. D'une part, il faut considérer qu'il n'a pas toujours été possible d'obtenir la version originale d'un débat ou d'un discours prononcé par une personnalité politique étrangère et passée sur des chaînes étrangères; d'autre part, ce sont les caractéristiques mêmes des interactions avec l'interprète à la télévision qui englobent et *cachent* ce qu'un locuteur dit en langue étrangère. C'est le cas par exemple de l'interview d'un invité étranger lors d'un talk-show : chaque fois que l'interprète traduit en langue étrangère les propos de l'intervieweur au bénéfice de l'invité, la voix de l'interprète est cachée par celle de l'intervieweur qui s'adresse, bien entendu, à l'invité mais surtout au public italo-phoné; lors du tour de l'invité qui s'exprime en langue étrangère, au contraire, c'est la voix de l'interprète traduisant en italien qui l'emporte sur celle de l'interviewé afin que le public puisse suivre l'interview⁸. Cela explique pourquoi dans ces cas il est impossible de récupérer les tours de parole en langue étrangère aussi bien de l'interprète que de l'invité. C'est la raison pour laquelle nous avons défini CorIT comme un corpus *partiellement parallèle* (Falbo 2012a). Enfin, CorIT se présente sous forme de corpus *ouvert* puisque de nouveaux enregistrements y sont régulièrement ajoutés. Cela arrive surtout pour ce qui est des débats présidentiels américains ou français passés à la télévision italienne avec interprétation simultanée. CorIT se compose des sous-corpus figurant dans le tableau suivant, dans lequel nous indiquons la durée approximative totale des enregistrements transcrits (et alignés avec le fichier son correspondant⁹) jusqu'à présent :

TABLEAU 1
Composition de CorIT

Sous-corpus	Durée (heure) des transcriptions effectuées avec alignement texte/son
«talk shows»	≈ 40
événements médiatiques	≈ 90
débats présidentiels	≈ 40
«breaking news»	-
discours de guerre	-
discours (victoire, état de l'Union, etc.)	≈ 20
festivals (Cannes, Oscars, etc.)	≈ 30

Les données en termes d'heures transcrites ne représentent qu'une petite partie de la totalité des objets qui constituent ce corpus.

2.1.1. *Le repérage des objets composant le corpus*

Au début de la collecte de CorIT, aucune distinction n'a été faite concernant la typologie d'émission ou la modalité d'interprétation, l'objectif étant le repérage de la totalité des performances des interprètes à la télé. Si on peut supposer avoir atteint cet objectif lors du dépouillement des archives RAI, il est pratiquement impossible d'affirmer qu'aucune performance n'a échappé aux procédures d'enregistrement en direct, à cause notamment de la multiplication des chaînes, surtout commerciales, et des émissions à partir des années 1980 à peu près. Ce sont plutôt des critères arbitraires et contingents qui ont réglé la sélection des performances à enregistrer et précisément la disponibilité du chercheur et des moyens d'enregistrement ainsi que la valeur médiatique de certains événements, par exemple les débats présidentiels américains. Cette démarche ne répond certainement pas à la démarche qu'il faudrait suivre lors de la conception d'un corpus (*corpus design*), autrement dit, l'objectif de repérage de performances d'interprètes à la télévision n'avait pas été fixé sur la base d'une connaissance fine des différents contextes télévisuels, ce qui revient à une sélection basée sur des critères externes (Clear 1992). En effet, ce n'est qu'à posteriori que nous avons mis de l'ordre dans le cumul de données collectées. Le besoin de catégorisation nous a amenée à examiner soigneusement les caractéristiques *extérieures* de chaque objet, à savoir de chaque enregistrement, ce qui nous a permis d'identifier les traits en mesure de définir une seule performance¹⁰, c'est-à-dire un seul item. Cette opération a acquis, au fur et à mesure, une valeur heuristique insoupçonnée, puisqu'elle a permis de mettre en évidence des points fondamentaux concernant les modalités d'interprétation, les types d'interaction et les genres télévisuels qui caractérisent l'interprétation à la télévision et, par là, CorIT¹¹. C'est grâce à ce travail à la fois de distinction et de regroupement des différents objets sur la base de leurs traits extérieurs que des sous-corpus se sont dessinés, selon la caractéristique choisie (modalité d'interprétation, type d'interaction, genre télévisuel, caractéristiques de l'interprète, etc.).

2.1.2. *La transcription d'un corpus multimodal et son exploitation*

Rendre un corpus tel que CorIT exploitable est apparu, dès le début, comme un défi énorme. Le mode de transcription – étape incontournable – s'est tout de suite imposé comme un choix capable d'orienter, de façon plus ou moins définitive, la qualité et l'exploitabilité des données à étudier. À la fin de la première décennie des années 2000, peu de logiciels offraient la possibilité d'aligner la transcription avec le son ; un seul, à notre connaissance, permettait de gérer, en même temps, la vidéo. Notre choix est donc tombé sur *WinPitch* (Martin 2009) qui permet, entre autres, d'obtenir une représentation fine des éléments prosodiques, ce qui s'avère essentiel si on souhaite analyser la prosodie des interprètes ou le décalage du discours produit par l'interprète par rapport au discours original (Ahrens 2004 ; Falbo 2017, 2018), dans le but, entre autres, d'en mesurer le caractère télégénique. S'agissant de la transcription et des enjeux d'interrogation qui en découlent, nous nous bornerons à remarquer que, tout en incluant un système de recherche au niveau du mot, *WinPitch* n'offre pas la possibilité de mener une interrogation à la fois sur deux ou plusieurs transcriptions

alignées avec les traces audio/vidéo correspondantes. Les transcriptions – dans leur forme (trans-)écrite, c'est-à-dire dépourvues du support audio/vidéo correspondant – peuvent, bien entendu, être passées sur n'importe quel concordancier, ce qui équivaut malheureusement à renoncer à l'objectif d'une analyse fondée sur le rôle essentiel du support audio/vidéo en mesure de rétablir, à tout moment, tous les aspects propres à l'oral (et à la mimique, à la gestuelle) et qui échappent à une représentation écrite, c'est-à-dire à une représentation qui n'offre qu'une image estompée de la parole.

Il faut toutefois signaler que la précision et les possibilités d'analyse et de traitement des données orales offertes par *WinPitch* ont permis jusqu'à présent de mener des études qualitatives fines de certaines caractéristiques de l'oral des interprètes à la télévision (Papa 2010).

2.2.3 *CorILEX: caractéristiques et conditions de collecte*

CorILEX est un corpus en cours de composition et compte à présent six audiences audio-enregistrées auprès de la Section GIP/GUP (en Italie, juge pour l'enquête préliminaire / juge pour l'audience préliminaire, procédure pénale) d'une ville du nord de l'Italie¹² et dont nous fournissons des indications sur les langues impliquées, outre l'italien et la durée dans le tableau suivant :

TABLEAU 2

Composition de CorILEX

Langue	Durée
Lituanien	15'40''
Moldave	30'10''
Punjabi	20'15''
Roumain	27'40''
Serbe	33'26''
Plusieurs langues	1h12'59''

L'accès à ce contexte spécifique est le fruit de la disponibilité et de l'esprit de collaboration du personnel de justice travaillant dans ce service. Notre projet initial, qui consistait à collecter un corpus capable de répondre au besoin d'étudier l'interaction avec interprète en milieu juridique¹³, s'est réduit au fur et à mesure qu'augmentait le silence, voire le refus, des institutions que nous avons invitées à participer à notre projet.

La connaissance préalable des caractéristiques propres à la situation de communication qui se déroule au tribunal devant un juge (audience) et des participants potentiels ainsi que leur positionnement dans la salle nous a permis de bien organiser la collecte des données sur le terrain. À ce travail de collecte des items suit la phase de transcription qui, elle, ne cesse de poser problème.

2.2.1. *Le travail de terrain: enregistrement et notation des éléments extralinguistiques*

La saisie vidéo étant interdite en salle d'audience, notre choix est tombé forcément sur l'enregistrement audio. Les audiences ont été enregistrées grâce à deux magnétophones : l'un positionné près de l'interprète et de la personne (témoin, accusé) ne

parlant pas la langue de la procédure (c'est-à-dire l'italien), l'autre près du chercheur-observateur; le recours à un deuxième magnétophone a le but de fournir, le cas échéant, l'enregistrement des voix des autres participants à l'audience au cas où elles ne seraient pas captées par le premier magnétophone. Ce double enregistrement constitue en outre une sorte de bouée de sauvetage en cas de problèmes techniques qui pourraient compromettre le bon fonctionnement de l'un des deux magnétophones.

Ces modalités d'enregistrement ont été mises au point après qu'on a remarqué qu'il était impossible d'utiliser les enregistrements effectués par des techniciens du tribunal conformément aux dispositions de la loi, et dont la transcription est considérée comme partie intégrante du procès-verbal de l'audience. Comme la langue de la procédure est l'italien, seuls les propos prononcés en italien seront transcrits et insérés dans le procès-verbal et, par conséquent, seuls les participants s'exprimant en italien sont censés parler dans un micro qui permettra de les enregistrer. Il en va de même pour l'interprète. C'est la raison pour laquelle il est pratiquement impossible de repérer, dans les fichiers audios du tribunal, les propos prononcés par le témoin/accusé ou par l'interprète en langue étrangère. Quant aux transcriptions effectuées par le tribunal, elles sont inutilisables, puisqu'elles ne sont pas alignées avec le son et répondent à d'autres critères que ceux qui règlent le processus de transcription visant la préparation de données destinées à l'analyse linguistique et/ou interactionnelle. Dans les transcriptions insérées dans les procès-verbaux d'audience, les conventions de transcription ne sont pas explicitées, mais en les lisant on s'aperçoit que ces textes ont été soumis à un processus de normalisation selon les règles de l'écrit. En effet, il n'y a aucune trace de mots tronqués, de pauses, d'hésitations et d'autres phénomènes qui caractérisent la langue parlée; il y a au contraire la ponctuation qui, elle, établit des frontières et oriente l'interprétation (au sens herméneutique du mot) des énoncés. Le magnétophone placé tout près du témoin/accusé et de l'interprète permet en outre de capter assez bien ce que l'interprète traduit en chuchotage à son client lors d'un échange verbal qui occupe les autres participants à l'interaction, ainsi que d'éventuels échanges dyadiques entre l'interprète et son client.

La participation du chercheur à l'audience est essentielle (Capeau et Gadet 2007: 102) non seulement pour gérer au mieux les opérations d'enregistrement, mais aussi pour noter dans les détails tout ce qui se passe dans la salle d'audience et qui échappe à l'enregistrement audio: le nombre de personnes présentes, leur rôle, leur positionnement, leurs déplacements éventuels, leurs gestes, par exemple la remise de papiers de la part d'un interlocuteur à un autre pour que ce dernier les examine ou les lise à haute voix. La notation de ces éléments permettra ensuite une interprétation (au sens herméneutique du mot) adéquate de certains tours de parole produits par les différents interlocuteurs. Par exemple, c'est grâce à cette démarche que nous avons été à même de comprendre que des bruits de feuilles et de pas étaient à attribuer à la greffière qui s'était levée, s'était approchée de l'interprète et lui avait remis une feuille sur laquelle campait la formule que cette dernière était censée lire à voix haute et par laquelle elle s'engageait à traduire correctement et fidèlement. De plus, le chercheur qui assiste à l'audience aura moins de mal à mettre en relation une voix avec la personne à laquelle cette voix appartient. Rien n'est plus difficile en effet que de distinguer les différentes voix qui se relaient ou se chevauchent dans une interaction et de les attribuer à leurs propriétaires légitimes quand on ne peut compter que sur un enregistrement audio et que l'on ne connaît pas ou du moins pas suffisamment les

données extralinguistiques qui enveloppent et forment l'interaction elle-même. Nous nous approprions là un principe fondateur de la sociolinguistique – ou disons, d'une branche de la sociolinguistique – qui exige une approche « caractérisé[e] davantage par le processus de regroupement de données contextualisées et rapportées à des locuteurs bien identifiés que par la référence à un contexte spécifiquement social ou écologique » (Cappeau et Gadet 2007 : 100; Wörner 2012, sur le rôle essentiel des métadonnées).

2.2.2. *La variété linguistique et les défis de la transcription*

Les flux migratoires d'un côté et la mobilité des personnes au sein d'une Union européenne élargie de l'autre ont transformé profondément le panorama linguistique également à l'intérieur des palais de justice. Très souvent l'anglais, le français et l'espagnol ne sont utilisés qu'avec leur statut de grandes langues véhiculaires, alors que d'autres langues qualifiées de rares ou exotiques par le personnel de justice occupent le devant de la scène. Bien qu'étant à un état embryonnaire, CorILex est d'ores et déjà en mesure de fournir une image de la variété linguistique qui peuple les salles d'audience : roumain/moldave, lituanien, serbe, punjabi, espagnol. Dans la plupart des cas, cette variété ne correspond pas aux connaissances et compétences des personnes impliquées dans la collecte d'un corpus, ce qui demande l'intervention d'experts linguistiques capables de comprendre et écrire les langues en question. Un besoin apparemment simple à satisfaire. Notre expérience nous a montré que parfois les choix théoriques et pratiques qui sous-tendent la collecte d'un corpus multilingue comme CorILex, à savoir prévoir la transcription en langue originale et une traduction interlinéaire en mesure de donner à l'analyste – qui ne peut pas accéder directement à cette langue – une image le plus possible près de la réalité, peuvent être remis en question ou tout simplement modifiés de façon ponctuelle à cause de situations inattendues. C'est ce qui a caractérisé la transcription d'une audience avec la présence d'un témoin de langue punjabi. La personne de langue punjabi qui nous a aidée dans la transcription, nous a expliqué, à notre grand étonnement, qu'aucun des ressortissants de la région du Punjab (Pakistan) présents sur le territoire où a été menée la recherche n'était à même d'écrire le punjabi, puisque l'écriture de cette langue n'est apprise que par les étudiants qui choisissent de suivre des études universitaires en langue et littérature punjabi. Pour le reste de la population du Punjab, le punjabi est la langue du quotidien mais pas une langue officielle enseignée à l'école dans sa version orale et écrite. Il va de soi que prétendre avoir une personne capable d'écrire le punjabi est un objectif démesuré par rapport aux ressources que nous avons à notre disposition pour ce projet¹⁴.

À ces aspects imprévus s'ajoutent souvent des difficultés liées au fait que ces soi-disant experts linguistiques n'ont bien entendu aucune formation en linguistique et sont dépourvus de toute capacité de réflexion métalinguistique. Il n'est pas toujours évident de leur faire comprendre que l'objectif de leur apport est d'aboutir à une transposition qui nous révèle la composition des énoncés au-delà du bien-fondé des affirmations qui y sont contenues ; il est même très difficile de leur apprendre à s'abstenir de toute (sur-)interprétation et à signaler la possibilité d'une double interprétation – toujours au sens herméneutique du mot. Cette tentative de démêler ce qui, dans le processus de compréhension et transposition – dans ce cas, du punjabi à l'italien – relève du sens de l'énoncé de ce qui est ajouté, voire imaginé, par l'expert

linguistique, nous fait percevoir toute la fragilité de cette opération et nous renvoie aux mécanismes qui règlent l'écoute, la compréhension et qui obéissent avant tout à la recherche du sens selon les attentes de chacun (Blanche-Benveniste 1997; Bilger, Blasco, *et al.* 1997).

L'absence de la transcription en punjabi se traduit par un vide sur le plan de la transcription. Sur la base de l'expérience faite sur CorIT, nous avons décidé de transcrire sous ELAN, un logiciel qui, à côté de l'alignement entre le son (vidéo) et le texte transcrit, offre des instruments d'interrogation fondamentaux pour l'exploitation du corpus (Niemants, dans le présent volume). Dans une optique d'homogénéité dans la notation des rôles assumés par les différents interlocuteurs, nous avons mis au point un modèle unique à appliquer lors de la transcription de chaque audience, dans lequel figurent tous les interlocuteurs qui pourraient intervenir dans une audience de la section considérée¹⁵: du ministère public, au juge, à l'interprète, en passant par la personne qui s'exprime en langue étrangère. Pour les langues auxquelles l'analyste ne peut pas accéder directement, il a été prévu de transcrire sur la ligne assignée à un interlocuteur (parent tiers) les mots prononcés en langue originale; sur une deuxième ligne, différente et dépendante de la première (child tier), figurera leur traduction, dans notre cas, en italien¹⁶. Si on revient à l'exemple du punjabi, la ligne consacrée à la langue originale reste vide mais la présence simultanée du support audio (ainsi que la présence de la traduction en italien) nous signale l'existence de tours de parole qui n'ont pas pu être transcrits et qui nous rappellent, encore une fois, les limites qui définissent la collecte de ce type de corpus.

3. Deux parcours différents, des défis partagés

Dans les paragraphes qui précèdent, nous avons décrit le travail de collecte de deux corpus d'interprétation très différents en nous arrêtant en particulier sur les modes de repérage et d'acquisition de leurs composants et sur le processus de transcription. Dans la partie qui suit, nous nous focaliserons sur les différences qui les caractérisent en essayant de les articuler avec les réflexions qui se sont développées au sein de la linguistique (ou, si l'on veut, des linguistiques) de corpus afin d'identifier la place des corpus d'interprétation à l'intérieur d'un domaine qui se montre de plus en plus diversifié, hétérogène et interdisciplinaire (Eshkol-Taravella et Lefeuve-Halftermeyer 2017).

Le repérage des items qui composent CorIT se rapproche beaucoup de la collecte de corpus de textes (qu'ils soient écrits ou oraux) tirés de l'internet¹⁷ ou encore en libre accès comme ceux qui forment EPIC (Russo, Bendazzoli, *et al.* 2012). En effet, aucune autorisation n'est nécessaire du fait du caractère public des émissions dans lesquelles ces textes s'inscrivent¹⁸. Par contre, CorILex est le fruit d'un travail de terrain qui recoupe, dans ses traits généraux, la démarche propre à une certaine branche de la sociolinguistique. La présence du chercheur dans la salle d'audience n'est néanmoins pas assimilable à l'observation participante qui caractérise certaines études sociolinguistiques. En effet, le chercheur assiste à l'audience comme s'il assistait à une sorte de rituel public. Si sa présence est bien entendu relevée et acceptée par les différents acteurs de l'interaction, elle n'en est pas moins oubliée au fur et à mesure que se développe l'interaction hautement ritualisée qu'est l'audience. Chacun joue son rôle selon le scénario prévu sans prêter aucune attention à l'enregistrement

en cours. Les moyens (personnels) d'enregistrement se confondent avec l'appareillage technique installé dans la salle et donné pour acquis. La présence de l'observateur-chercheur est donc comparable et identifiable à celle du technicien du son, qui surveille le processus d'enregistrement en vue de la rédaction du procès-verbal ou à celle d'un public possible. Ces *overhearers*, selon le schéma participatif mis au point par Erving Goffman (1976), n'ont, par conséquent, aucun pouvoir sur l'interaction en cours. Leur présence est perçue en tant que partie du décor¹⁹.

CorIT et CorILex présentent deux formes très différentes. Le premier, avec sa quantité énorme d'items, ressemble plutôt à une « base textuelle » ou un « réservoir à corpus » (Habert 2000 : 12-13) à partir duquel il est possible de sélectionner le ou les sous-corpus souhaités ; par exemple, les débats présidentiels américains ou français, les talk-shows avec invité anglophone, etc. Cette « opération de choix raisonné parmi les composants disponibles » (Habert 2000 : 13) peut porter sur les traits extérieurs des composants eux-mêmes (par exemple genre télévisuel ou interaction monologale vs interaction dialogale), sur les traits internes (la langue parlée par l'invité étranger tous types d'interaction confondus) ou sur une combinaison des deux (interaction dialogale – interview – avec invité lusophone). Il en va autrement pour CorILex qui présente une homogénéité en ce qui a trait aux composants, puisqu'il a été collecté à partir d'une sélection faite à priori, basée sur des facteurs extérieurs (audiences auprès de la Section GIP/GUP). L'identification de sous-corpus ne peut se faire qu'à partir des différences inhérentes à chaque audience, en sélectionnant par exemple les interprètes pour une langue étrangère donnée dans le but d'étudier les caractéristiques de l'un et de l'autre sur le plan des pratiques discursives et traductives.

D'après ces caractéristiques, CorIT semble rentrer dans la typologie des monitor corpus (corpus de suivi, Habert 2000 : 13). Étant donné ses dimensions actuelles, la question se pose de savoir jusqu'à quand il faudra continuer à y ajouter de nouveaux composants. En fait, cette grande archive multimodale ou, si l'on veut, ce réservoir à corpus, s'est avérée ingérable à plusieurs points de vue. Le manque de financements et de ressources humaines nous a contrainte à privilégier certains sous-corpus – qui pourraient être traités de corpus spécialisés, si l'on adoptait un regard différent – comme celui des débats présidentiels ou d'une typologie de talk-show. Autrement dit, le corpus design manqué, qui avait profondément marqué le début de l'aventure CorIT, fait désormais ressentir tout son poids en termes de gestion et de préparation des données en vue de leur exploitation à l'aide d'outils informatiques. Le mode de transcription choisi montre aujourd'hui ses limites quant aux procédures d'interrogation, qui pourraient néanmoins être surmontées grâce à des interventions technologiques et informatiques appropriées soutenues par des moyens financiers adéquats. Mais au-delà des entraves que nous avons évoquées, nous nous interrogeons sur la nécessité de faire de CorIT un corpus au vrai sens du mot. Si on passe en revue les études qualitatives (Straniero Sergio 2007, 2011, 2012 ; Falbo 2012a ; Dal Fovo 2011) qui ont été menées sur des sous-corpus réunis et transcrits manuellement, dirait-on, on a l'impression que tout a été dit sur l'interprétation à la télévision italienne. Mais en revenant au questionnement qui a inspiré la collecte de CorIT, on s'aperçoit que ses données ont encore beaucoup à révéler et qu'une étude systématique portant sur une quantité plus élevée de données et se focalisant sur des aspects communs ou distinctifs des différents sous-corpus formant CorIT permettrait d'élargir nos

connaissances du domaine, ne serait-ce que d'un point de vue statistique. De plus, un traitement des données qui tienne compte de leur dimension spécifiquement orale, par exemple un alignement texte transcrit-son, pourrait montrer ce qu'une présentation écrite cache très souvent aux yeux et aux oreilles. Par exemple, les phénomènes prosodiques, tels que le rythme d'élocution, les chevauchements ou encore l'émotivité qui colore la parole (Fonagy 1983), sont assez difficiles à représenter, mais parfois indispensables pour saisir correctement le sens d'un énoncé ou tout simplement pour le rendre intelligible (Blanche-Benveniste 2008, 2010). La possibilité d'accéder au son et à la transcription en même temps permettrait en outre de vérifier, de corriger ou tout simplement de peaufiner le texte transcrit pendant l'analyse, avec des conséquences remarquables sur les résultats de l'analyse elle-même.

CorILex, quant à lui, apparaît comme un corpus spécialisé capable de représenter la réalité d'une section particulière du domaine juridique. Encore faut-il qu'il puisse dépasser la quantité d'audiences qui le composent en ce moment. Ce petit corpus que nous voudrions développer représente à nos yeux une précieuse rareté. Dans le domaine juridique, il est très difficile de sortir de l'étude de cas. Seuls les ouvrages de Berk-Seligson (1990/2002) et Hale (2004/2010) se fondent sur un nombre significatif d'audiences qui ont permis d'étudier de façon plus approfondie l'interprétation dans les tribunaux des États-Unis et de l'Australie. Nul n'ignore la finalité ou, du moins, les retombées en termes d'applications possibles de ces études²⁰. La connaissance des spécificités de l'interaction dans la salle d'audience, en particulier, et dans le domaine juridique, en général, obtenue grâce à une étude scientifique du matériel collecté sur le terrain, ne peut qu'aider les chercheurs à dépasser le seuil de l'impression personnelle et le caractère subjectif, parfois prescriptif, de certains apports. C'est donc à partir d'un questionnement fort que nous nous sommes attelée à la collecte de CorILex, en ayant pleine conscience du caractère instable des liens tissés avec les autorités judiciaires. Il suffit qu'une seule personne travaillant dans le service qui collabore au projet soit mutée ou parte à la retraite pour empêcher CorILex de grandir. La taille de ces deux corpus, on l'aura compris, pose, *mutatis mutandis*, toujours problème. Un speech corpus tel CorIT est un défi en termes de gestion, de transcription et, par là, d'interrogation; face à un speech corpus comme CorILex, on s'interroge sur son caractère représentatif. C'est là aussi que réside la fragilité de la collecte de corpus d'interprétation dans certains domaines.

Au fil de ces quelques lignes, nous avons essayé de mettre au jour les aspects qui, à nos yeux, font émerger toute la particularité des corpus d'interprétation. Sur le plan pratique aussi bien que théorique, la collecte de corpus d'interprétation passe par une prise en compte des caractéristiques propres à l'interprétation inscrite dans ses différents contextes. Cela exige une approche différenciée et interdisciplinaire aussi bien en ce qui concerne le corpus design que l'interrogation du corpus et la préparation des données. Et s'il reste encore beaucoup à faire pour obtenir des résultats écologiquement viables, ne serait-ce qu'au sein d'un domaine et d'une situation de communication donnée, il est possible d'affirmer que les corpus d'interprétation réunis jusqu'à présent ont le mérite d'avoir lancé le débat, envisagé des buts à atteindre et identifié les obstacles et les limites à surmonter ou tout simplement à accepter. C'est là, peut-être, le défi le plus difficile à relever.

NOTES

1. L'interprétation, c'est-à-dire la traduction orale, répond aux besoins de communication d'interlocuteurs qui ne partagent pas la même langue. Elle s'effectue dans des situations de communication variées et à travers des modalités différentes telles que l'interprétation consécutive, l'interprétation simultanée et le chuchotage. Au vu du contexte, de la modalité et du type d'interaction (monologal/dialogal, selon Kerbrat-Orecchioni 2005: 16) qui s'instaure entre les interlocuteurs, il est possible de distinguer, par exemple, l'interprétation de conférences de l'interprétation de service public (en milieu scolaire, médical, pédagogique ou en milieu juridico-judiciaire). Nous signalons, par souci de clarté, que la dénomination « interprétation de dialogue », quant à elle, met en valeur la dimension dialogale de l'interaction dans laquelle intervient l'interprète et regroupe, en gros, toutes les situations caractérisées par le fait que les interlocuteurs, interprète y compris, se parlent en face-à-face, par exemple lors d'une consultation médicale ou au tribunal. Pour un approfondissement, nous renvoyons aux différentes entrées de la *Routledge Encyclopedia of Interpreting Studies* (Pöchhaker 2015).
2. Ces contraintes caractérisent la collecte de corpus de langue parlée en général, mais elles ne pèsent sûrement pas aussi lourd que dans le domaine de l'interprétation.
3. GAVIOLI, Laura, BARALDI, Claudio et NIEMANTS, Natacha (2016): From archives to corpora: extracting data for analysis, from a collection of interpreter-mediated interactions. CL8: 8th International Critical Link Conference. Édimbourg, 29 juin-1 juillet 2016.
4. L'absence physique du support audio dans les corpus oraux risque de biaiser le regard du chercheur qui pourrait tomber dans le piège de considérer le texte transcrit comme un texte écrit, une tentation à laquelle ont cédé pas mal d'études sur l'interprétation (Falbo 1999).
5. Texte original: «le procedure (non esclusivamente verbali) attraverso cui i parlanti producono e comprendono i propri e gli altrui comportamenti sociali».
6. Le mot *enregistrement* est utilisé ici comme synonyme de *performance* ou d'*interprétation*, en tant que produit du travail de l'interprète. Les interprétations qui composent le corpus ont des durées très variées: de quelques minutes (typique des « breaking news ») à plus de deux heures (débats présidentiels).
7. Les langues étrangères les plus représentées sont l'anglais et le français, suivies du russe, de l'espagnol, de l'allemand, du portugais, du serbe, du bosniaque, de l'albanais et du japonais.
8. Pour une description plus détaillée, nous renvoyons à Falbo (2012a, 2012b).
9. Francesco Straniero Sergio, qui a eu l'initiative d'assembler CorIT, dans son ouvrage de 2007 consacré au « talk show interpreting », a transcrit et analysé les interprétations de 107 interprètes intervenant en consécutive ou en simultanée dans des émissions de type « talk-show » depuis 1990 jusqu'à 2005/2006 environ. Malheureusement, les transcriptions n'ont pas été alignées avec la vidéo correspondante et, de plus, les fichiers contenant les transcriptions en Word ont été perdus. Cela donne une idée assez claire de la difficulté de reprendre le travail sur ce corpus en essayant de le rendre utilisable, ne serait-ce que pour compléter les transcriptions.
10. Les éléments capables d'identifier un seul item sont les suivants: nom de l'interprète, modalité d'interprétation, type d'interaction, noms des participants à la communication médiatique, date, émission italienne/chaîne de télévision, chaîne étrangère, macro-genre (télévisuel), genre télévisuel, type de texte (interaction en langue originale ou en italien) (Falbo 2012a).
11. Pour une description détaillée de ces aspects, nous renvoyons à Falbo (2012a).
12. L'autorisation à enregistrer les audiences et à analyser des données récoltées est assujettie à l'obligation de ne pas révéler des éléments qui pourraient contribuer à l'identification des personnes impliquées dans la procédure.
13. En adoptant le mot *juridique*, nous accueillons la proposition du Forum de réflexion sur le multilinguisme et la formation des interprètes (2009: 8): « Le forum de réflexion a opté pour le terme "interprète/interprétation juridique", plus inclusif que celui "d'interprète judiciaire" [...]. L'interprétation juridique couvre les prestations effectuées dans tous les services présentant des composantes juridiques, depuis les enquêtes policières et douanières jusqu'aux commissions rogatoires, en passant par la phase pré-procédurale, les entretiens entre avocat et client, le procès, la phase post-procédurale, l'immigration, les procédures relevant du mandat d'arrêt européen, etc. » Voir FORUM DE RÉFLEXION SUR LE MULTILINGUISME ET LA FORMATION DES INTERPRÈTES (2009): *L'interprétation juridique au sein de l'Union européenne. Recommandations en vue de meilleures pratiques*. Bruxelles: Commission des Communautés européennes. Consulté le 8 décembre 2018, <<https://eulita.eu/wp/wp-content/uploads/files/Reflection%20Forum%20Final%20Report-f.pdf>>.
14. Depuis toujours, les corpus coûtent cher (voir par exemple Zanettin 2012: 40).

15. Cela permettra ensuite un repérage automatique sur une ou plusieurs transcriptions (c'est-à-dire audiences) d'une même fonction (par exemple ministère public) et, éventuellement – grâce aux données extralinguistiques ou métadonnées –, la comparaison entre personnes physiques différentes qui, dans les différentes audiences, ont eu cette même fonction. Par exemple, on pourrait étudier le comportement verbal/interactionnel des différents ministères publics étant intervenus dans les différentes audiences.
16. Des traductions en d'autres langues pourront être ajoutées avec l'insertion de child tiers supplémentaires, selon le public auquel l'analyste s'adresse.
17. Il faut préciser que l'accès aux archives RAI et l'acquisition des enregistrements a coûté très cher, ce qui n'est pas le cas de la plupart des enregistrements qui circulent sur l'internet et que l'on peut télécharger gratuitement.
18. L'exploitation de ces textes pourrait toutefois être soumise, selon les cas, à des règles strictes (Baude 2006; Allora et Barbera 2007; Zanni 2007).
19. Il en irait autrement dans le contexte médical où la présence du chercheur pourrait inhiber l'interaction entre personnel médical et patient (Niemants, dans le présent volume).
20. PERRY, Melissa, BLOKLAND, Jenny, BURDON-SMITH, Susan, *et al.* (2017) : *Recommended National Standards for Working with Interpreters in Courts and Tribunals*. Canberra: Judicial Council on Cultural Diversity. Consulté le 10 décembre 2018, <<https://jccd.org.au/wp-content/uploads/2018/02/JCCD-Interpreter-Standards.pdf>>. Ces recommandations constituent l'exemple le plus récent des aboutissements d'une recherche et d'un engagement qui durent depuis très longtemps.

RÉFÉRENCES

- AHRENS, Barbara (2004) : *Prosodie beim Simultandolmetschen* [La prosodie en interprétation simultanée]. Frankfurt am Main: Peter Lang.
- ALLORA, Adriano et BARBERA, Manuel (2007) : Il problema legale dei corpora. Prime approssimazioni [Le problème légal des corpus. Premières réflexions]. In: Manuel BARBERA, Elisa CORINO et Cristina ONESTI, dir. *Corpora e linguistica in rete* [Corpus et linguistique en ligne]. Torino: Guerra Edizioni, 109-118.
- ARBACH, Najib et ALI, Saandia (2013) : Aspects théoriques et méthodologiques de la représentativité des corpus. In: Laurence VINCENT-DURROUX et Philip CARR, dir. *Statut et utilisation des corpus en linguistique. Corela*. HS-13. Consulté le 10 septembre 2018, <<http://corela.revues.org/3029>>.
- BARBERA, Manuel, CORINO, Elisa et ONESTI, Cristina (2007) : Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup [Qu'est-ce qu'un corpus? Pour une définition plus rigoureuse de corpus, token, markup]. In: Manuel BARBERA, Elisa CORINO et Cristina ONESTI, dir. *Corpora e linguistica in rete* [Corpus et linguistique en ligne]. Torino: Guerra Edizioni, 25-88.
- BAUDE, Olivier, dir. (2006) : *Corpus oraux. Guide des bonnes pratiques*. Paris: CNRS Éditions.
- BENDAZZOLI, Claudio (2012) : From international conferences to machine-readable corpora and back: an ethnographic approach to simultaneous interpreter-mediated communicative events. In: Francesco STRANIERO SERGIO et Caterina FALBO, dir. *Breaking Ground in Corpus-based Interpreting Studies*. Berne: Peter Lang, 91-117.
- BERK-SELIGSON, Susan (1990/2002) : *The Bilingual Courtroom: Court interpreters in the judicial process*. Chicago: University of Chicago Press.
- BERNARDINI, Silvia, FERRARESI, Adriano, RUSSO, Mariachiara, *et al.* (2018) : Building interpreting and intermodal corpora: a *how-to* for a formidable task. In: Mariachiara Russo, Claudio Bendazzoli et Bart Defrancq, dir. *Making Way in Corpus-based Interpreting Studies*. Singapore: Springer, 21-42.
- BIAGINI, Marta (2012) : Data collection in the courtroom: challenges and perspectives for the researcher. In: Francesco STRANIERO SERGIO et Caterina FALBO, dir. *Breaking Ground in Corpus-based Interpreting Studies*. Berne: Peter Lang, 231-251.
- BILGER, Mireille, BLASCO, Mylène, CAPPEAU, Paul, *et al.* (1997) : Transcription de l'oral et interprétation. Illustration de quelques difficultés. *Recherches sur le français parlé*. 14:57-86.

- BLANCHE-BENVENISTE, Claire (1997): *Approches de la langue parlée en français*. Paris: Ophrys.
- BLANCHE-BENVENISTE, Claire (2000a): Introduction. In: Mireille BILGER, dir. *Corpus. Méthodologie et applications linguistiques*. Perpignan: Champion/Presses universitaires de Perpignan, 11-15.
- BLANCHE-BENVENISTE, Claire (2000b): Corpus de français parlé. In: Mireille BILGER, dir. *Corpus. Méthodologie et applications linguistiques*. Perpignan: Champion/Presses universitaires de Perpignan, 15-25.
- BLANCHE-BENVENISTE, Claire (2008): Les unités de langue écrite et de langue parlée. In: Mireille BILGER, dir. *Données orales. Les enjeux de la transcription*. Perpignan: Presses universitaires de Perpignan, 192-217.
- BLANCHE-BENVENISTE, Claire (2010): *Le français. Usages de la langue parlée*. Paris/Louvain: Peeters.
- BLANCHE-BENVENISTE, Claire et JEANJEAN, Colette (1987): *Le français parlé. Transcription et édition*. Paris: INALF/Didier Érudition.
- CAPPEAU, Paul et GADET, Françoise (2007): L'exploitation sociolinguistique des grands corpus. *Revue française de linguistique appliquée*. 12(1):99-110.
- CLEAR, Jeremy (1992): Corpus sampling. In: Gerhard LEITNER, dir. *New Directions in English Language Corpora: Methodology, results, software, developments*. Berlin/New York: Mouton de Gruyter, 21-31.
- DAL FOVO, Eugenia (2011): Through the CorIT looking-glass and what MA students found there. *The Interpreters' Newsletter*. 16:1-20.
- DAL FOVO, Eugenia (2017): Good health across languages: how access to healthcare by non-Italian speaking patients is ensured in Italy. A case study. *Lingue Culture Mediazioni*. 4(1):33-55.
- DAL FOVO, Eugenia (2018): The use of dialogue interpreting corpora in healthcare interpreter training: taking stock. *The Interpreters' Newsletter*. (23):83-113.
- ESHKOL-TARAVELLA, Iris et LEFEUVRE-HALFTERMEYER, Anaïs (2017): Linguistique de corpus: vues sur la constitution, l'analyse et l'outillage. In: Iris ESHKOL-TARAVELLA and Anaïs LEFEUVRE-HALFTERMEYER, dir. *Linguistique de corpus: vues sur la constitution, l'analyse et l'outillage*. Corela. HS-21. Consulté le 10 septembre 2018, <<https://corela.revues.org/4800>>.
- FALBO, Caterina (1999): L'interprétation: une forme particulière d'oralité. *Revue française de linguistique appliquée*. 4(2):99-112.
- FALBO, Caterina (2012a): CorIT (Italian Television Interpreting Corpus): classification criteria. In: Francesco STRANIERO SERGIO et Caterina FALBO, dir. *Breaking Ground in Corpus-based Interpreting Studies*. Berne: Peter Lang, 155-185.
- FALBO, Caterina (2012b): L'interprète dans la communication interculturelle à la télévision. In: Nathalie AUGER, Christine BÉAL et Françoise DEMOUGIN, dir. *Interactions et interculturelité: Variétés des corpus et des approches*. Berne: Peter Lang, 347-364.
- FALBO, Caterina (2017): Les oraux des interprètes: un terrain particulier d'observation. In: Enrica GALAZZI et Marie-Christine JAMET. *Les z'oraux – Les français parlés entre sons et discours*. Repères DoRiF. 12. Consulté le 6 septembre 2018, <http://dorif.it/ezine/ezine_articles.php?dorif_ezine=20d58da01f93a17d17f6c2961c3d7c61&art_id=338>.
- FALBO, Caterina (2018): Insegnare la simultanea dal francese all'italiano: la prosodia come rivelatore dell'apprendimento. In: Laurie ANDERSON, Laura GAVIOLI and Federico ZANETTIN, dir. *Translation And Interpreting for Language Learners (TAIL)*. inTRALinea. Consulté le 4 novembre 2018, <<http://www.intralinea.org/specials/article/2313>>.
- FÖNAGY, Ivan (1983): *La vive voix: essais de psycho-phonétique*. Paris: Payot.
- GADET, Françoise (2008): L'oreille et l'œil à l'écoute du social. In: Mireille BILGER, dir. *Données orales. Les enjeux de la transcription*. Perpignan: Presses universitaires de Perpignan, 35-48.
- GAVIOLI, Laura (2015): On the distribution of responsibilities in treating critical issues in interpreter-mediated medical consultations: the case of "le spieghi(amo)." *Journal of Pragmatics*. 76:169-180.
- GOFFMAN, Erving (1976): Replies and responses. *Language in Society*. 5:257-313.

- HABERT, Benoît (2000): Des corpus représentatifs: de quoi, pour quoi, comment? *In*: Mireille BILGER, dir. *Linguistique sur corpus. Études et réflexions*. Perpignan: Presses universitaires de Perpignan, 11-58.
- HALE, Sandra B. (2004/2010): The discourse of court interpreting: Discourse practices of the law, the witness and the interpreter. Amsterdam/Philadelphia: John Benjamins.
- HOUSE, Juliane, MEYER, Bernd et SCHMIDT, Thomas (2012): CoSi – A Corpus of Consecutive and Simultaneous Interpreting. *In*: Thomas SCHMIDT et Kai WÖRNER, dir. *Multilingual Corpora and Multilingual Corpus Analysis*, Amsterdam/Philadelphia: John Benjamins, 295-304.
- JEFFERSON, Gail (1983): Issues in the transcription of naturally-occurring talk: caricature versus capturing pronunciation particulars. *Tilburg Papers in Language and Literature*. 34:1-12.
- KERBRAT-ORECCHIONI, Catherine (2005): *Le discours en interaction*. Paris: Armand Colin.
- MARTIN, Philippe (2009): *Intonation du français*. Paris: Armand Colin.
- OCHS, Elinor (1979/1999): Transcription as theory. *In*: Adam JAWORSKI et Nikolas COUPLAND, dir. *The Discourse Reader*. London/New York: Routledge, 167-182.
- PAPA, Carlo (2010): *Il fenomeno del self-repair in interpretazione simultanea* [Le phénomène du self-repair en interprétation simultanée]. Mémoire de maîtrise, non publié. Trieste: Università di Trieste.
- PÖCHHACKER, Franz, dir. (2015): *Routledge Encyclopedia of Interpreting Studies*. London/New York: Routledge.
- RUSSO, Mariachiara, BENDAZZOLI, Claudio, SANDRELLI, Annalisa, et al. (2012): The European Parliament interpreting corpus (EPIC): Implementation and developments. *In*: Francesco STRANIERO SERGIO et Caterina FALBO, dir. *Breaking Ground in Corpus-based Interpreting Studies*. Berne: Peter Lang, 53-90.
- SACKS, Harvey, SCHEGLOFF, Emanuel A. et JEFFERSON, Gail (1974): A simplest systematics for the organization of turn-taking for conversation. *Language*. 50:696-735.
- SANDRELLI, Annalisa (2012): Introducing FOOTIE (Football in Europe): Simultaneous interpreting in football press conferences. *In*: Francesco STRANIERO SERGIO et Caterina FALBO, dir. *Breaking Ground in Corpus-based Interpreting Studies*. Berne: Peter Lang, 119-153.
- SHLESINGER, Miriam (1998): Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta*. 43(4):486-493.
- STRANIERO SERGIO, Francesco (2007): *Talkshow Interpreting. La mediazione linguistica nella conversazione spettacolo* [Talkshow interpreting. La médiation linguistique au sein de la conversation-spectacle]. Trieste: Edizioni Università di Trieste.
- STRANIERO SERGIO, Francesco (2011): Language mediation in news making: From simultaneous interpreting to other (hybrid) transfer modes. *The Interpreters' Newsletter*. 16:175-196.
- STRANIERO SERGIO, Francesco (2012): Using corpus evidence to discover style in interpreters' performances. *In*: Francesco STRANIERO SERGIO et Caterina FALBO, dir. *Breaking Ground in Corpus-based Interpreting Studies*. Berne: Peter Lang, 211-230.
- STRANIERO SERGIO, Francesco et FALBO, Caterina, dir. (2012): *Breaking Ground in Corpus-based Interpreting Studies*. Berne: Peter Lang.
- WÖRNER, Kai (2012): Finding the balance between strict defaults and total openness: Collecting and managing metadata for spoken language corpora with the EXMARaLDA Corpus Manager. *In*: Thomas SCHMIDT et Kai WÖRNER, dir. *Multilingual Corpora and Multilingual Corpus Analysis*, Amsterdam/Philadelphia: John Benjamins, 383-400.
- ZANETTIN, Federico (2009): Corpora multimediali e analisi dell'interazione. Osservazioni su strumenti e metodologie [Corpus multimédia et analyse de l'interaction. Remarques sur les instruments et les méthodologies]. *In*: Laura GAVIOLI, dir. *La mediazione linguistico-culturale: una prospettiva interazionista* [La médiation linguistico-culturelle: une perspective interactionniste]. Perugia: Guerra Edizioni, 325-350.
- ZANETTIN, Federico (2012): *Translation-driven Corpora*. Manchester: St. Jerome.
- ZANNI, Samantha (2007): Corpora elettronici e copyright. Lo status legale della questione [Corpus électroniques et droits d'auteur. Le statut légal de la question]. *In*: Manuel BARBERA, Elisa CORINO et Cristina ONESTI, dir. *Corpora e linguistica in rete* [Corpus et linguistique en ligne]. Torino: Guerra Edizioni, 119-126.