Meta
Journal des traducteurs
Translators' Journal

META

Du texte aux ressources multimodales : faire avancer la recherche en interprétation à partir d'un corpus déjà existant

Claudio Bendazzoli, Michela Bertozzi and Mariachiara Russo

New Contexts in Discourse Analysis for Translation and Interpretation
Volume 65, Number 1, April 2020

URI: https://id.erudit.org/iderudit/1073643ar
DOI: https://doi.org/10.7202/1073643ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Bendazzoli, C., Bertozzi, M. & Russo, M. (2020). Du texte aux ressources multimodales : faire avancer la recherche en interprétation à partir d'un corpus déjà existant. *Meta, 65* (1), 211–236. https://doi.org/10.7202/1073643ar

Article abstract

Corpus linguistics generally relies on textual sources to analyse representative instances of language use, be it written or spoken. However, the fundamental role of nonverbal and multimodal resources has come to the fore as they contribute tremendously to meaning-making processes, in both direct and mediated communication. This is all the more relevant in Interpreting Studies, where transcripts of mediated interactions can only reveal a partial picture of the communicative exchange. That is why attempts have been made to include multimodal resources (for instance, extralinguistic information, video and audio recordings) in corpus projects such as the European Parliament Interpreting Corpus (EPIC), which set the basis for further interpreting corpora. This paper focuses on such developments and illustrates three more interpreting corpora, namely the Directionality in Simultaneous Interpreting (DIRSI) Corpus, the European Parliament Translation and Interpreting Corpus (EPTIC), and the Anglintrad Corpus and platform. These linguistic resources take advantage of multimodality to a different extent, offering textual and multimedia materials either separately or aligned to each other. All these examples are evidence of how important it is to maintain flexible formats and structures when developing an interpreting corpus, so that existing resources can be the springboard for further progress in the study of interpreting beyond the textual level.

Du texte aux ressources multimodales : faire avancer la recherche en interprétation à partir d'un corpus déjà existant¨

CLAUDIO BENDAZZOLI
*Università degli Studi di Torino, Turin, Italie**
claudio.bendazzoli@unito.it

MICHELA BERTOZZI
*Università di Bologna, Forlì, Italie***
michela.bertozzi6@unibo.it

MARIACHIARA RUSSO
*Università di Bologna, Forlì, Italie***
mariachiara.russo@unibo.it

RÉSUMÉ

En règle générale, la linguistique de corpus repose sur des sources textuelles et vise l'analyse d'instances représentatives de la langue orale ou écrite. Toutefois, le rôle des ressources non verbales et multimodales y est crucial, car ces dernières contribuent considérablement aux processus de création de sens, tant dans le cas de la communication directe que dans celui de la communication médiée. Ce rôle est d'autant plus essentiel pour les études en interprétation, où les transcriptions des interactions médiées ne peuvent refléter qu'une partie des échanges communicationnels. Des projets comme EPIC (European Parliament Interpreting Corpus) tentent donc d'inclure des ressources multimodales dans leur corpus (p. ex. les informations extralinguistiques, la vidéo et les enregistrements audios) et définissent de nouvelles bases pour les corpus d'interprétation conçus ultérieurement. Nous aborderons ici les développements de cet ordre et donnerons trois autres exemples de tels corpus : DIRSI (Directionality in Simultaneous Interpreting), EPTIC (European Parliament Translation and Interpreting Corpus), le corpus et la plateforme Anglintrad. Ces ressources linguistiques tirent de nouveaux avantages de la multimodalité et offrent des ressources textuelles et multimédias indépendantes ou alignées. Ces exemples montrent l'importance de conserver des formats et des structures souples lorsque l'on crée un corpus d'interprétations, de façon que les ressources ainsi constituées puissent faire progresser les études en interprétation au-delà d'un niveau strictement textuel.

ABSTRACT

Corpus linguistics generally relies on textual sources to analyse representative instances of language use, be it written or spoken. However, the fundamental role of nonverbal and multimodal resources has come to the fore as they contribute tremendously to meaning-making processes, in both direct and mediated communication. This is all the more relevant in Interpreting Studies, where transcripts of mediated interactions can only reveal a partial picture of the communicative exchange. That is why attempts have been made to include multimodal resources (for instance, extralinguistic information, video and audio recordings) in corpus projects such as the European Parliament Interpreting Corpus (EPIC), which set the basis for further interpreting corpora. This paper focuses on such developments and illustrates three more interpreting corpora, namely the Directionality in Simultaneous Interpreting (DIRSI) Corpus, the European Parliament

Translation and Interpreting Corpus (EPTIC), and the Anglintrad Corpus and platform. These linguistic resources take advantage of multimodality to a different extent, offering textual and multimedia materials either separately or aligned to each other. All these examples are evidence of how important it is to maintain flexible formats and structures when developing an interpreting corpus, so that existing resources can be the springboard for further progress in the study of interpreting beyond the textual level.

RESUMEN

La lingüística del corpus se basa generalmente en fuentes textuales y está dirigida al análisis de instancias representativas del lenguaje oral o escrito. Sin embargo, el papel de los recursos no verbales y multimodales es crucial, ya que estos últimos contribuyen considerablemente al proceso de creación de significado, tanto en el caso de la comunicación directa como en el de la comunicación mediada. Este papel es aún más esencial para los estudios de interpretación, donde las transcripciones de interacciones mediadas solo pueden reflejar parte de los intercambios comunicativos. Por lo tanto, proyectos como EPIC (European Parliament Interpreting Corpus) intentan incluir recursos multimodales en un corpus (por ejemplo, información extralingüística, grabaciones video y audio) y definen nuevas bases para los corpus de interpretación diseñados posteriormente. Analizaremos aquí su evolución en ese sentido y presentaremos tres ejemplos de tales corpus: el corpus DIRSI (Directionality in Simultaneous Interpreting), el corpus EPTIC (European Parliament Translation and Interpreting Corpus), y el corpus y la plataforma Anglintrad. Estos recursos lingüísticos obtienen nuevas ventajas de la multimodalidad y ofrecen recursos de texto y multimedia independientes o alineados. Estos ejemplos muestran la importancia de preservar formatos y estructuras flexibles al construir un corpus de interpretaciones, de modo que los recursos así creados permitan avanzar en los estudios de interpretación más allá de un nivel estrictamente textual.

MOTS CLÉS/KEYWORDS/PALABRAS CLAVE

corpus d'interprétation, EPIC, EPTIC, DIRSI-C, Anglintrad
interpreting corpora, EPIC, EPTIC, DIRSI-C, Anglintrad
corpus de interpretación, EPIC, EPTIC, DIRSI-C, Anglintrad

1. Introduction

En linguistique, on peut généralement définir un corpus comme : « a collection of (1) *machine-readable* (2) *authentic* texts (including transcripts of spoken data) which is (3) *sampled* to be (4) *representative* of a particular language or language variety » (McEnery, Xiao *et al*. 2006 : 5). Dans les cas de la traductologie et des études d'interprétation, on peut également définir un corpus comme un ensemble structuré d'instances de communication dans un cadre spécifique, qui peut comporter des instances écrites, parlées ou signées (ou une combinaison de ces dernières). Ces instances coexistent avec des services de médiation langagière (comme la traduction et l'interprétation). Il existe donc de nombreuses configurations de corpus, comme les corpus comparables (constitués de textes sources et de textes cibles dans la même langue) et les corpus parallèles (par exemple, constitués de textes sources en anglais et de textes cibles en italien). Les corpus prenant en compte des traductions dans les deux directions sont appelés corpus réciproques (Zanettin 2012). Les corpus intermodaux incluent quant à eux des textes cibles obtenus à partir du même texte source par le biais de modes de traduction différents, comme la traduction écrite et l'interprétation simultanée (Shlesinger 2008 ; Shlesinger et Ordan 2012). De plus, du fait

du lien entre traduction et situation, les corpus d'interprétation peuvent également être multimodaux, car la création du sens des messages échangés y contient plusieurs couches sémiotiques (Baldry et Thibault 2001 ; Gao et Wang 2017). Cependant, l'annotation et la représentation de couches sémiotiques (comme les informations extralinguistiques sur les participants, les actes de discours et le cadre, ou les instances non verbales comme l'intonation, le langage corporel, le contact visuel, etc.) sont obtenues de différentes manières dans les corpus d'interprétation. Dans les faits, les données vidéo ou audio peuvent se présenter sous forme de fichiers distincts ou alignés avec les transcriptions, et il se peut que le comité de recherche n'ait pas accès à la totalité des données, pour des raisons de confidentialité ou pour des raisons techniques.

Les linguistes de corpus visent à inclure autant d'instances que possible, afin que leur corpus soit suffisamment fourni pour être représentatif d'une instance de communication particulière (Biber 1993 ; Halverson 1998). On obtiendra différents degrés de représentativité selon divers facteurs, dont la disponibilité de données écrites. En effet, dans le cas des corpus oraux, tout particulièrement les corpus d'interprétation, la transcription est une étape essentielle dans la constitution d'un corpus qui pourra être analysé subséquemment[1]. La transcription étant une activité chronophage, elle a un impact significatif sur la taille des corpus et sur l'approche analytique adoptée (Bendazzoli 2018 ; Bernardini, Ferraresi *et al.* 2018). Dans les faits, l'annotation des éléments paralinguistiques et cinétiques et des informations extralinguistiques implique déjà une forme d'analyse de la part de la personne chargée de la transcription (O'Connel et Kowal 1994, 1999).

Étant donné l'évolution de la conception des corpus d'interprétation au fil des ans, de nombreux projets se sont limités à des corpus de petite taille, analysés manuellement. Bendazzoli et Sandrelli (2009) distinguent trois étapes dans l'évolution des corpus : les corpus « manuels » et les premiers corpus lisibles par des machines ; les corpus entièrement lisibles par des machines, uniquement accessibles aux chercheurs qui les développent ; les corpus accessibles en ligne, qui offrent la possibilité de requêtes par le biais de divers logiciels permettant l'extraction et l'analyse d'occurrences.

Dans le présent article, nous présentons quatre projets dont les données sont déjà accessibles à l'ensemble de la communauté scientifique. Nous commencerons par le projet European Parliament Interpreting Corpus (EPIC, section 2), qui a été décrit pour la première fois dans *Meta* il y a plus de quinze ans (Monti, Bendazzoli *et al.* 2005). EPIC a inspiré les corpus d'interprétation ultérieurs que nous présenterons par la suite, en guise d'exemples de développement de ressources plus sophistiquées : le Directionality in Simultaneous Interpreting Corpus (DIRSI, section 3) ; le European Parliament Translation and Interpreting Corpus (EPTIC, section 4) ; le corpus et la plateforme Anglingtrad (section 5)[2].

Bien que les transcriptions soient les pierres angulaires de tout corpus, ces ressources linguistiques comportent bien plus que des mots. Les informations que l'on peut y enregistrer dépassent en effet les stricts éléments linguistiques, on y trouve par exemple les attributs et les tags qui catégorisent les éléments non verbaux et extralinguistiques (comme les informations sur la situation de communication, sur les participants et leurs *actes de discours*). Bien que l'on ne puisse extraire d'un corpus que les informations qu'on y a rentrées, certaines méthodes permettent l'observation

systématique et la combinaison de paramètres descriptifs qui peuvent enrichir le point de vue des analystes, en leur offrant non plus de simples observations linguistiques, mais une approche discursive plus complète (Bowker et Pearson 2002 ; Partington, Morley *et al.* 2002).

2. Projet European Parliament Interpreting Corpus (EPIC)

Le projet EPIC (2004-2006) a été inspiré par l'article de référence de Shlesinger (1998) et mené à bien au Département d'interprétation et de traduction de l'Université de Bologne (campus de Forlì), par un groupe de recherche interdisciplinaire composé d'interprètes, de linguistes de corpus et d'experts en nouvelles technologies, qui ont conçu et développé une archive multimédia et un corpus de transcriptions en ligne, lisibles par des machines (respectivement *EPIC Multimedia Archive* et *EPIC corpus*). La principale visée du projet était la collecte d'une grande quantité de données authentiques d'interprétation simultanée, afin de produire une recherche empirique bien nécessaire sur les caractéristiques des discours interprétés, d'informer et d'améliorer les pratiques de formation (Monti, Bendazzoli *et al.* 2005 ; Sandrelli, Bendazzoli *et al.* 2010 ; Russo, Bendazzoli *et al.* 2012).

EPIC inclut des enregistrements de la chaîne d'information *EbS* (*Europe by Satellite*) et des discours originaux, interprétés et enregistrés lors de séances plénières au Parlement européen. Les données enregistrées comprennent des extraits des sessions de février, de mars, d'avril et de juillet 2004. Les enregistrements ont été numérisés et découpés en fichiers audio individuels. Ces derniers ont été transcrits en incluant des éléments linguistiques (transcription orthographique suivant les directives du *Code de rédaction interinstitutionnel* de l'U.E.) ; paralinguistiques (pauses remplies ou vides, mots tronqués ou mal prononcés) ; et extralinguistiques (métadonnées). La conception des éléments extralinguistiques était fondée sur les informations disponibles sur le site du Parlement européen et sur les caractéristiques spécifiques aux débats qui s'y tiennent. Le principe de transparence des institutions européennes est avantageux pour les chercheurs en quête de données sur les sujets participant aux débats. En revanche, les règlements stricts sur les procédures d'allocation du temps de parole influencent considérablement les types d'actes de discours pendant les débats, en termes de durée, de mode de délivrance et de vitesse. Le tableau 1 recense tous les attributs appliqués aux transcriptions d'EPIC, sous forme de titres.

Tableau 1

EPIC : titres et descripteurs de métadonnées

Date	jj-mm-aa-matin/après-midi
Ordre de discours	000 (comme indiqué dans le rapport textuel *in extenso*)
Langue	it/en/es
Type	org-it/en/es ou int-it/en/es-it/en/es
Longueur du discours	court (< 120 s) moyen (de 121 à 360 s) long (> 360 s)
Durée	nombre total de secondes

Volume de mots	petit (< 300 mots) moyen (de 301 à 1000 mots) grand (> 1000 mots)
Nombre de mots	nombre total de mots
Vitesse	lente (<130 mots/min) moyenne (de 131 à 160 mots/min) rapide (> 160 mots/min)
Nombre de mots par minute	nombre de mots par minute
Délivrance du texte source	impromptu/lu/mixte
Intervenant	nom de famille, prénom
Genre	F/M
Pays	
Première langue	oui/non
Fonction politique	Eurodéputé Eurodéputé président de la session Président du Parlement européen Vice-président du Parlement européen Commission européenne Conseil de l'Europe Invité
Groupe politique	(nom du groupe politique en question)
Sujet	Agriculture et pêche Économie et finance Emploi Environnement Santé Justice Politique Procédures et formalités Société et cultures Sciences et technologies Transports
Sujet spécifique	(tel que mentionné dans le rapport textuel *in extenso*)
Commentaires	Réf. à Division du Conseil Réf. à Direction générale Réf. à rôle de l'invité Accent (p. ex. écossais, andalous, etc.) Problèmes techniques Autre

Le corpus est lemmatisé, indexé, étiqueté et se divise en 9 sous-corpus : 3 sous-corpus en langues sources (italien, anglais, espagnol) et 6 sous-corpus d'interprétation simultanée, dans toutes les directions et dans toutes les combinaisons possibles des trois langues du corpus, ce qui fait d'EPIC un corpus à la fois parallèle, réciproque et comparable d'environ 180 000 mots (tableau 2).

Tableau 2
Taille et composition d'EPIC

Sous-corpus	Nombre de discours	Nombre total de mots	% d'EPIC
ORG-EN	81	42 705	25
INT-EN-IT	81	35 765	20
INT-EN-ES	81	38 066	21
ORG-IT	17	6 765	4
INT-IT-EN	17	6 708	4
INT-IT-ES	17	7 052	4
ORG-ES	21	14 406	8
INT-ES-IT	21	12 833	7
INT-ES-EN	21	12 995	7
TOTAL	357	177 295	100

Légende : ORG = texte source ; INT = texte cible ; IT = italien ; EN = anglais ; ES = espagnol.

Depuis 2018, il est possible d'accéder à EPIC via une nouvelle interface[3] et d'y effectuer des requêtes libres grâce au moteur *NoSketch Engine* (Rychlý 2007). La totalité des transcriptions, des fichiers audio et des fichiers vidéo est ainsi librement accessible depuis le catalogue de l'ELRA (European Language Resources Association[4]). Les chercheurs peuvent ainsi exploiter le plein potentiel d'EPIC.

Au fil des ans, les contenus d'EPIC ont fait l'objet de nombreuses études qualitatives et quantitatives, menées par des étudiants à la maîtrise pour leurs mémoires (Ghiselli 2015, 2018 ; Lobascio 2015 ; Russo 2010) et par des universitaires (Sandrelli et Bendazzoli 2005 ; Russo, Bendazzoli *et al.* 2006 ; Spinolo et Garwood 2010 ; Bendazzoli, Sandrelli *et al.* 2011 ; Russo 2011, 2018). Ces derniers ont pu utiliser les contenus transcrits sous une forme lisible par des machines, avec des annotations pertinentes.

L'avantage d'un grand corpus d'interprétation lisible par des machines, par rapport aux études de cas, réside dans la possibilité d'extraire automatiquement de grandes quantités des phénomènes que l'on souhaite étudier, s'ils y sont correctement annotés. Comme prévu, EPIC est étiqueté et des conventions de transcription spécifiques y ont été appliquées pour marquer les mots tronqués ou mal prononcés : la fin des mots tronqués est signalée par un tiret (-) (p. ex. : Pre- President it is a pleasure to be here…) et les mots mal prononcés ou tronqués au milieu ont d'abord été « normalisés » pour permettre aux étiqueteurs de les reconnaître, puis transcrits tels qu'ils ont réellement été prononcés, entre des chevrons (p. ex : **il Parlamento </parlomento/> ha deciso che**…). Ces caractéristiques ont permis aux chercheurs d'étudier la densité lexicale, la variété lexicale et toutes les disfluences précédemment mentionnées dans les 357 discours d'EPIC. Les chercheurs ont ainsi pu avoir accès à un aperçu inédit des productions linguistiques des interprètes de conférence, d'un point de vue multilingue (p. ex. les interprètes anglais et espagnols) et multidirectionnel (p. ex. de l'italien vers l'anglais et l'espagnol et vice-versa).

Les études sur la densité et la variété lexicales (Sandrelli et Bendazzoli 2005 ; Russo, Bendazzoli *et al.* 2006) fournissent de nombreuses preuves des capacités

expressives et de la richesse linguistique des interprètes. Ces études ont été inspirées par les travaux de Laviosa (1998), qui a comparé la densité et la variété lexicales de la prose narrative anglaise et des textes traduits depuis diverses langues européennes vers l'anglais (Translational English Corpus, TEC), et qui a trouvé que la densité et la variété lexicales des textes traduits étaient inférieures à la prose originale en anglais. Les études mentionnées ci-dessus visaient à vérifier si l'on pouvait remarquer des tendances similaires dans EPIC, un corpus de discours (parlés) interprétés, ou si ces conclusions s'appliquaient uniquement à la traduction écrite. De plus, comme prévu, la recherche dans EPIC concernait trois langues : l'italien, l'espagnol et l'anglais. En conséquence, l'étude visait également à établir s'il existait des schémas lexicaux dans les textes interprétés et si ces schémas variaient selon la paire de langues et les langues sources et cibles. Les recherches sur EPIC et celles de Laviosa diffèrent aussi, car TEC est un corpus comparable, alors qu'EPIC est à la fois comparable et parallèle. Cette caractéristique a permis aux chercheurs d'observer des tendances dans les schémas lexicaux, pas uniquement par la comparaison de discours originaux délivrés en anglais, en italien et en espagnol avec des textes interprétés dans ces langues, mais aussi par la comparaison des textes originaux dans les trois langues avec des textes cibles correspondants dans les trois langues. Effectuer de telles recherches manuellement aurait été très complexe (voire impossible), mais leurs résultats prouvent que la densité lexicale des discours interprétés a tendance à être supérieure à celle des discours originaux (à seulement deux exceptions près). Cet effet est contraire aux observations de Laviosa sur les textes traduits. Quant à la variété lexicale, elle s'est avérée généralement inférieure dans les discours interprétés que dans les discours originaux dans la même langue, tout comme Laviosa l'avait remarqué pour l'anglais traduit. Cependant, les interprètes italiens échappent à cette tendance, car le degré de variété lexicale de l'italien interprété s'est avéré supérieur à celui des discours originaux en italien, et ce, quelle que soit la langue source (anglais ou espagnol). Les résultats ainsi obtenus indiquent comment le mode de traduction (traduction écrite ou interprétation simultanée), la combinaison linguistique, la langue source et la langue cible peuvent influencer la densité et la variété lexicales des textes.

Une approche centrée sur le corpus permettrait l'étude systématique de deux autres types de disfluences dans la langue parlée : les mots mal prononcés et les mots tronqués (non terminés) (Bendazzoli, Sandrelli *et al.* 2011). Les conventions de transcription d'EPIC (voir ci-dessus) ont offert aux chercheurs la possibilité d'extraire automatiquement ces deux disfluences. Plus spécifiquement, les chercheurs ont pu extraire et compter ces disfluences pour déterminer si elles étaient plus fréquentes dans les textes sources ou dans les textes cibles. De plus, grâce à l'analyse des données, il a été possible de vérifier si les locuteurs et les interprètes réussissaient à rectifier leur production, c'est-à-dire s'ils parvenaient à bien prononcer les mots mal prononcés et à terminer les mots tronqués.

Notre hypothèse de départ était que les discours interprétés auraient tendance à contenir un plus grand nombre des deux disfluences et un nombre moindre de rectifications, du fait des contraintes propres à l'interprétation simultanée, comme le temps. Notre étude a démontré que la fréquence des mots mal prononcés et tronqués était supérieure dans les textes cibles (TC) que dans les textes sources (TS), à deux exceptions : les TS anglais comportent plus de mots tronqués que leurs TC en italien et en espagnol ; et les interprètes anglais semblent rencontrer moins de problèmes de

prononciation que leurs homologues italiens ou espagnols. De plus, ni les locuteurs originaux ni les interprètes ne corrigent généralement les mots mal prononcés, une tendance particulièrement marquée dans les discours interprétés, quelles que soient la combinaison linguistique, la langue source et la langue cible. Quant aux mots tronqués, les locuteurs originaux les terminent plus souvent que les interprètes.

Le troisième objet de la recherche basée sur les corpus est la récente comparaison des tendances et des schémas linguistiques relatifs au genre (Russo 2018) entre des interprètes, femmes et hommes, espagnols, anglais et italiens. Une approche quantitative a été suivie, comme amorce de recherches qualitatives ultérieures basées sur les métadonnées d'EPIC. Le mode de délivrance du locuteur, la rapidité d'élocution des locuteurs originaux, les combinaisons linguistiques ont été étudiés, ainsi que leur rapport avec la longueur du discours cible (DC) dans 200 discours. Les performances des interprètes, femmes et hommes, de l'anglais vers l'italien et entre l'italien et l'espagnol ont fait l'objet d'analyses. En sont ressorties les différences statistiques suivantes entre les femmes et les hommes ($p < 0,05$) : pour des discours lus de l'anglais vers l'espagnol, les femmes délivrent en moyenne les discours plus rapidement que les hommes (143 mots/min pour les femmes contre 124 mots/min pour les hommes) ; les hommes ont tendance à produire des DC plus courts que les femmes (respectivement, 16 % en moyenne par rapport aux discours sources, contre 8 %) ; les discours cibles sur les sujets « Politique » et « Procédures et formalités » sont plus courts quand ils sont délivrés par des hommes que par des femmes (respectivement, 18 % contre 4 %, et 21 % contre 0,3 %). Enfin, l'étude des contenus d'EPIC a permis de repérer une tendance inverse de fond entre la vitesse de délivrance des discours sources et cibles, principalement du fait des femmes qui assurent les interprétations de l'anglais vers l'espagnol et l'italien. Cette étude a révélé que d'importantes tendances liées au genre des interprètes semblent émerger, et que l'approfondissement des recherches sur ce sujet est prometteur pour les études en interprétation. De plus, il serait nécessaire d'étudier la réduction significative de la taille du DC quand la vitesse d'élocution du locuteur augmente, pour déterminer son influence sur de potentielles pertes sémantiques, sur les éventuelles stratégies de compensation réussies par les interprètes et sur les conditions idéales de communication.

Aujourd'hui, EPIC fait l'objet d'ajouts de contenus et d'alignements entre des transcriptions et des vidéos et des fichiers audio : ces dernières années, 278 376 mots de plus ont été transcrits, et 462 discours (un total de 1 269 minutes) sont en attente de transcription. La version 2.0 d'EPIC, dans sa taille finale, offrira certainement de nombreuses nouvelles opportunités de recherche.

3. Directionality in Simultaneous Interpreting Corpus (DIRSI)

Le corpus DIRSI a été créé juste après EPIC, dans le cadre d'un projet doctoral (Bendazzoli 2010). Les interprètes d'EPIC travaillant uniquement d'autres langues vers leur langue A, le principal objectif de DIRSI était de collecter un nouveau corpus d'interprétations simultanées, dans lequel les interprètes professionnels travailleraient dans les deux sens, soit d'une langue B vers leur langue A et vice-versa[5]. DIRSI contient les discours d'ouverture, les communications et les discours de clôture de trois conférences médicales (les questions-réponses n'y figurent pas), en anglais et en italien. DIRSI se compose de quatre sous-corpus : deux sous-corpus de discours

sources (en anglais et en italien) et deux sous-corpus des discours cibles correspondants. Comme EPIC, DIRSI est un corpus à la fois parallèle, réciproque et comparable. DIRSI contient un total de 136 000 mots, répartis dans quatre sous-corpus, comme détaillé dans le tableau 3.

Tableau 3
Composition et taille de DIRSI

Sous-corpus	Nombre d'actes de discours	Nombre de mots	% de DIRSI-C
ORG-IT	63	33 412	24,6
INT-IT-EN	63	31 510	23,2
ORG-EN	16	37 249	27,4
INT-EN-IT	16	33 664	24,8
TOTAL	158	135 835	100

Légende : ORG = discours source ; INT = discours cible ; IT = italien ; EN = anglais

La mise en œuvre du projet EPIC a fourni des outils méthodologiques pour créer des corpus d'interprétations simultanées et a servi de base à des projets ultérieurs similaires, comme le corpus DIRSI. Cependant, il n'a pas été possible d'appliquer toutes les caractéristiques propres au Parlement européen ni les choix imposés par son contexte à de situations différentes, comme les conférences internationales privées (qui sont la cible de DIRSI). Comme mentionné dans les sections précédentes, la disponibilité des métadonnées et les paramètres de classification utilisés comme attributs pour refléter les caractéristiques des discours sources et cibles du Parlement européen (comme la durée, la longueur, la vitesse de délivrance) sont uniquement pertinents dans le cadre de la communication médiée par l'interprétation au Parlement européen. Dans un contexte différent, comme celui des conférences médicales internationales, les informations équivalentes ne sont pas forcément disponibles et on doit modifier celles qui le sont pour refléter le contexte communicationnel desdites conférences.

Pendant la première phase du projet DIRSI, des enregistrements audio de certaines conférences ont été effectués et ont constitué une archive multimédia. De plus, les observations sur le terrain, pendant les enregistrements, ont servi à concevoir ou à modifier les attributs à appliquer aux données du corpus et à classifier tous les contenus (Bendazzoli 2012). En plus de modifier les attributs inspirés des titres d'EPIC, de nouvelles caractéristiques ont été prises en compte afin de pouvoir aligner les contenus textuels et audio, et d'effectuer des requêtes spécifiques sur la directionnalité de la traduction (p. ex. vers quelle langue, A ou B, l'interprète a traduit le texte source). Les titres ainsi conçus et tous les attributs figurent dans le tableau 4.

Tableau 4
Titres et descripteurs des métadonnées de DIRSI

Titre de la conférence	(titre complet)
Référence de la conférence	CFF4 CFF5 ELSA
Sujet de la conférence	Santé
Date de la conférence	année-mois-jour
Lieu de la conférence	Vérone Cesena
Séance de la conférence	ouverture communication discussion clôture
Titre de la séance	(selon le programme officiel)
Acte de discours	remarques sur l'ouverture ou la clôture communication ou cours temps de parole annonces concernant la procédure ou l'administratif question réponse commentaire
Ordre du discours	000
Type de discours	org-it org-en int-it-en int-en-it
Titre du discours	(selon le programme officiel et les présentations PowerPoint)
Longueur du discours	court (< 900 s) moyen (de 900 à 1 800 s) long (> 1 800 s)
Durée	nombre total de secondes
Volume de mots	petit (< 1 650 mots) moyen (de 1 650 à 3 300 mots) grand (> 3 300 mots)
Nombre de mots	nombre total de mots
Vitesse	faible (< 100 mots/min) moyenne (de 100 à 120 mots/min) élevée (> 120 mots/min)
Mots par minute	nombre total de mots par minute
Type de discours	impromptu lu mixte
Support audiovisuel	oui non
Participation à la conférence	organisateur commanditaire président répondant conférencier ou professeur membre du public interprète

ID du participant à la conférence	Prénom, Nom IT-01 IT-02 IT-03 IT-04 UK-01
Genre	m f
Pays	(spécifier le pays d'origine)
Langue	it en
Locuteur natif	oui non
Directionnalité	A B
Ressources fournies aux interprètes	à l'avance sur place aucune
Lien audio	(titre complet du fichier audio associé)
Commentaires	

En plus d'offrir un grand nombre d'attributs, la conception de DIRSI permet d'aligner les fichiers sources et cibles du corpus ainsi que ses enregistrements audio et ses transcriptions. L'alignement des textes a été manuel, il ne reflète donc pas le décalage temporel entre le discours source des locuteurs et les discours cibles des interprètes. Les interprètes professionnels commencent en effet généralement à traduire après avoir entendu quelques unités informationnelles (Schweda-Nicholson 1987). L'alignement des contenus de DIRSI a simplement été conçu pour faciliter la gestion et les analyses qualitatives des données. L'alignement des textes et des sons a été obtenu en ajoutant manuellement des codes temporels, ou étiquettes temporelles, dans les transcriptions grâce à *Transana*[6], un logiciel de transcription conçu pour les recherches qualitatives, et en ajoutant le titre complet du fichier audio associé dans les attributs disponibles dans les titres des transcriptions (voir plus haut, l'avant-dernier attribut du tableau 4). Les transcriptions ont alors été converties en fichiers XML, dont chaque attribut est devenu un tag XML[7].

On accède au corpus grâce à une interface en ligne[8] dans laquelle les transcriptions s'affichent en mode parallèle, accompagnées de lecteurs média qui permettent de lancer les fichiers audio associés. Les tags temporels sont intégrés dans les transcriptions, et les utilisateurs peuvent s'en servir pour écouter l'enregistrement au bon moment. Cette combinaison pratique de données écrites et audio s'est avérée utile pour prendre en compte les éléments métalinguistiques, car on peut en effet en prendre connaissance en écoutant directement les données audio et elle a permis de désambiguïser des occurrences qui auraient pu n'être analysées que sur la base des représentations verbales des communications (comme dans les transcriptions écrites).

Parmi les exemples de prise en compte de ces caractéristiques non verbales et extralinguistiques, on compte l'utilisation de l'anglais comme langue de travail dans les conférences internationales (Bendazzoli 2017) et l'usage du marqueur so par les interprètes traduisant de l'italien vers l'anglais comme langue active (Bendazzoli 2019). Ces études montrent le potentiel des ressources linguistiques telles que les

corpus, qui permettent d'analyser les communications médiées par des interprètes, bien au-delà de l'expression verbale, en tirant profit des métadonnées et de la multimodalité. La première étude (Bendazzoli 2017) sur le temps de parole et la longueur des textes cibles produits en anglais par des locuteurs natifs ou non natifs a mis en évidence des différences de pouvoir expressif entre les participants (Albl-Mikasa 2013). La seconde étude (Bendazzoli 2019) était basée sur le recueil automatique de toutes les occurrences du marqueur discursif so, dont chacune était vérifiée et désambiguïsée grâce à un alignement texte-son intégré pour traiter les données du corpus. Cette fonctionnalité a non seulement rendu possible le calcul du taux de génération du marqueur discursif des interprètes (+30 %), mais aussi l'étude des différentes fonctions de l'utilisation du marqueur discursif so dans les discours cibles, par exemple simplifier une syntaxe complexe dans le discours source, réitérer des informations précédemment fournies, en introduire de nouvelles, etc.

Le corpus DIRSI se base sur une sélection de seulement trois conférences. Cependant, comme mentionné plus haut, au stade de la collecte de données, de nombreuses autres conférences ont été enregistrées et des données ont été recueillies sur le terrain. On peut donc envisager d'étendre le corpus ou d'en créer un nouveau, comparable, à l'avenir.

4. European Parliament Translation and Interpreting Corpus (EPTIC)

Le projet EPTIC a été initié en 2009, avec comme ambition la collecte et la mise à disposition des comptes rendus anglais-italien dans leurs langues source et cible, ainsi que les transcriptions des discours et leurs traductions, le tout dans EPIC. En 2009, lors de discussions informelles à l'Aston Corpus Symposium, Miriam Shlesinger a fait remarquer qu'un corpus comme EPTIC (qui ne portait pas encore de nom à l'époque) fournirait de précieuses données sur l'intermodalité, ou encore sur les différences et les similitudes entre les différents modes de médiation linguistique (Shlesinger 2008). Le soutien de cette universitaire visionnaire a stimulé la confiance dans le projet et a transformé un projet secondaire en un effort collectif bien plus ambitieux, aujourd'hui nourri par la collaboration de plusieurs équipes internationales[9].

EPTIC est un corpus multilingue qui contient des échantillons en anglais, en français, en italien, en slovène et en polonais. L'ampleur de cette collaboration s'accroît suivant les priorités des diverses équipes de recherches impliquées. Ceci signifie que toutes les paires linguistiques ne sont pas représentées équitablement dans toutes ses versions : la première version du corpus comprenait uniquement des interprétations italien<>anglais ; les combinaisons français<>anglais ont suivi, quant aux plus récentes interprétations slovène<anglais et polonais>anglais, elles sont pour l'instant unidirectionnelles.

Ce corpus multilingue est parallèle, réciproque, intermodal, et multimodal. Il est *parallèle,* car il comprend des échantillons dans une langue et leurs cibles alignées de manière interlinguistique dans au moins une autre langue. Du fait qu'il contient plus d'une paire de langues source/cible et que la même langue peut s'y trouver à la fois comme langue source ou cible, le corpus, grâce à ses éléments en anglais, en français et en italien, est également *multiréciproque*. En tant que tel, il permet des comparaisons parallèles dans les deux directions de chaque paire linguistique, ainsi que des comparaisons comparables entre divers textes monolingues ou bilingues sur

le même sujet ou du même genre dans les trois langues. EPTIC est également *intermodal*, car il présente côte à côte les produits de deux modes de médiation interlinguistique, soit des échantillons d'interprétations simultanées et des traductions correspondantes. Ses composants *multimodaux* sont ses métadonnées et ses vidéos alignées temporellement avec les discours et leurs interprétations, et accessibles depuis les lignes de concordance. Chaque ligne de concordance s'affiche avec un hyperlien vers le fichier multimédia associé, et donne accès à l'enregistrement vidéo du discours source ou cible en question.

Chaque événement d'EPTIC est donc disponible dans au moins six « versions » différentes : la transcription du discours original tel que délivré, la transcription de son interprétation, la version écrite officielle, sa traduction officielle, la vidéo du discours original et celle de son interprétation. Chacun des 16 sous-corpus d'EPTIC comprend entre 15 000 et 20 000 mots, composant un total de plus de 400 000 mots (voir tableau 5 pour les détails). Bien que la taille intégrale du corpus soit substantielle (tout du moins pour un corpus d'interprétation), EPTIC contient relativement peu de références représentatives, car tous ses sous-corpus ne pourront certainement pas être utilisés ensemble.

Tableau 5
Taille et composition d'EPTIC

	Sources		Cibles	
	Orales	**Écrites**	**Interprétations**	**Traductions**
Anglais	24 136	22 782	53 615	58 561
Français	27 713	26 674	23 185	25 855
Italien	20 016	19 591	20 352	23 234
Polonais	11 011	10 616	–	–
Slovène	–	–	18 082	20 762

En matière de procédure de constitution de corpus (Ferraresi et Bernardini 2019), les textes bruts (les comptes rendus et leurs traductions) et les vidéos comportant plusieurs pistes audio ont d'abord été téléchargés sur le site Web du Parlement européen, et les informations contextuelles sur les discours et les locuteurs ont été enregistrées. L'étape suivante a été la transcription orthographique des discours et de leurs interprétations, suivant les conventions du *Code de rédaction interinstitutionnel* de l'U.E. relatives à l'orthographe, à l'usage des majuscules, aux acronymes et aux titres. Les transcriptions ont alors été segmentées en unités proches de phrases et l'on a ajouté la ponctuation en tenant compte des indices prosodiques et syntaxiques. Une telle segmentation ne rend pas parfaitement le caractère oral de ces événements, elle était cependant essentielle pour aligner des textes écrits entre eux ou avec des vidéos, et pour l'étiquetage morphosyntaxique. L'étape finale de cette préparation des textes a été l'insertion des applaudissements, des rires et d'autres bruits de fond notoires, ainsi que des mauvaises prononciations, des mots tronqués, des faux départs, des silences et des pauses remplies. Des métadonnées contenant des informations sur les textes, leurs contextes de production et sur les locuteurs du Parlement européen qui les avaient délivrés ont alors été ajoutées, ce qui a permis d'effectuer des requêtes basées sur les attributs associés aux locuteurs et aux actes de discours. On a ensuite

procédé à l'alignement automatique des textes entre eux, et à sa correction manuelle, avant d'aligner les textes et les vidéos grâce à l'utilisation d'un logiciel de sous-titrage[10], puis on a converti les débuts et les fins de segments en valeurs d'attributs XML. Enfin, on a procédé à l'étiquetage morphosyntaxique, à la lemmatisation et à l'indexation du corpus pour sa consultation via le moteur *NoSketch Engine* (Rychlý 2007), via des utilitaires de lignes de commandes et des scripts Perl spécifiques.

Comme nous décrivons les applications des corpus d'interprétation en recherche et en enseignement dans d'autres sections du présent article, nous fournirons ici deux exemples d'applications basées sur les composants multimodaux. On a très récemment utilisé EPTIC pour étudier la simplification lexicale dans différents modes de médiation, en comparant des interprétations et des traductions du français et de l'italien vers l'anglais à des discours originaux comparables en anglais et à leurs versions écrites modifiées (Ferraresi, Bernardini *et al*. 2018). EPTIC offrant la possibilité de combiner les perspectives monolingue et multimodale et différentes langues sources, les auteurs ont pu conclure que l'hypothèse de la simplification comme tendance universelle de la médiation interlinguistique n'est pas réalisée inconditionnellement. On observe effectivement une simplification lexicale dans l'anglais médié, mais elle est plus importante en français et en italien, dans les interprétations que dans les traductions. D'un point de vue appliqué, Bernardini (2016) suggère que l'on pourrait utiliser EPTIC dans des programmes de traduction et d'interprétation, comme source de preuve de l'existence de variantes en traduction, au même titre que des ressources plus connues comme des corpus de traductions multiples (Malmkjaer 2003) et des corpus d'apprentissage de la traduction (Castagnoli 2016). En se concentrant sur le rendu des collocations dans les traductions et les interprétations italien>anglais et anglais>italien, Bernardini a pu observer une expansion dans les deux modes de médiation et dans les deux directions, alors que d'autres phénomènes, comme la contraction, dépendent de la direction (italien>anglais). On pourrait utiliser ce type de preuves en classe, avec une approche socioconstructiviste de l'enseignement de la traduction et de l'interprétation, pour encourager une réflexion sur les processus de prise de décisions professionnelles et pour que les futurs traducteurs et interprètes aient plus conscience des effets de la multimodalité sur la médiation interlinguistique.

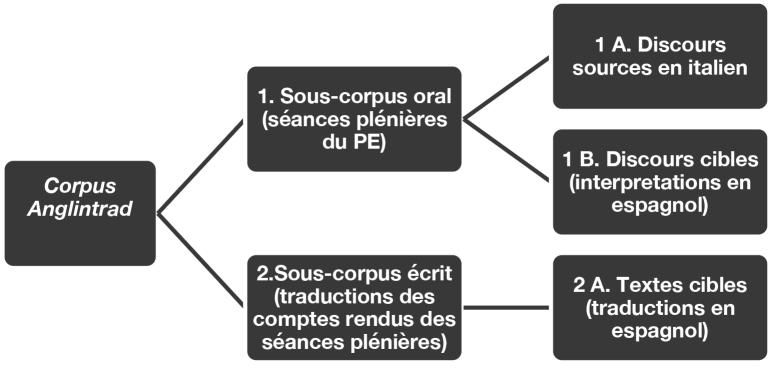
En bref, EPTIC, malgré sa petite taille, s'avère être un corpus multilingue, parallèle, intermodal et multimodal extrêmement complexe, qui se prête à divers types de recherche et diverses applications pédagogiques. Son principal inconvénient est la petite taille de ses sous-corpus, qui limite considérablement le type d'études que l'on peut y mener. Un effort communautaire serait nécessaire pour étoffer les contenus d'EPTIC, car l'étendue de l'expertise et des ressources requises semble plus relever d'une collaboration que du travail d'une équipe unique. À cette fin, une plateforme est en cours de développement, avec la double visée de favoriser des collaborations de longue distance et de simplifier les processus de préparation des textes, d'alignement des textes et des vidéos et d'indexations, qui sont fastidieux et prêtent aux erreurs. À court terme, on envisage d'autres améliorations, dont l'ajout de composants anglais>finlandais et la désambiguïsation automatique de la voix des interprètes grâce à une technologie de diarisation des locuteurs.

5. Le corpus et la plateforme Anglintrad

Le quatrième corpus d'interprétation inspiré par le projet EPIC est une ressource organisée dans une plateforme en ligne en libre accès[11]. Sa création a été motivée par : le besoin pratique de révéler un phénomène particulièrement problématique en interprétation simultanée de l'italien vers l'espagnol, la présence significative d'emprunts linguistiques intégraux à l'anglais[12] dans les discours politiques et institutionnels ; la nécessité d'un nouvel outil facilement accessible à des fins de recherche ou d'enseignement. Le corpus et la plateforme Anglintrad sont les principaux résultats d'un projet doctoral (Bertozzi 2018a, 2018b). Ils ont été développés pour étudier les stratégies des interprètes de conférence face à une difficulté linguistique potentielle quand ils traduisent de l'italien vers l'espagnol, en comparant leurs discours cibles avec les traductions officielles des comptes rendus in extenso. Il s'agissait donc d'observer le même phénomène (les emprunts intégraux à l'anglais) de deux perspectives différentes (interprétation et traduction), en tirant parti de l'intermodalité et de la multimodalité du corpus.

Les données sources compilées dans le corpus Anglintrad étaient encore les séances plénières du Parlement européen (PE), plus spécifiquement, une sélection de 26 séances, tenues en 2011. Après lecture des comptes rendus disponibles sur le site Internet du registre des documents du PE, on a dénombré 143 discours sources en italien qui contenaient des emprunts intégraux à l'anglais. Les discours cibles en espagnol ont alors été analysés, et tous les discours (originaux et interprétés) ont été transcrits pour permettre une comparaison immédiate entre les deux composantes du sous-corpus oral (figure 1) ; les traductions officielles associées ont été alignées manuellement, phrase par phrase, aux discours originaux et traduits. Les traductions vers l'espagnol constituent le sous-corpus écrit, composé des comptes rendus officiels publiés sur le site Internet du registre des documents du PE. Anglintrad a été conçu comme illustré dans la figure 1 et compte 249 emprunts intégraux à l'anglais, identifiés dans 143 textes sources en italien.

FIGURE 1
Structure du corpus Anglintrad



Le corpus a été compilé dans des feuilles de calcul et enrichi d'un ensemble de métadonnées collectées sous un titre spécifique basé sur le corpus EPIC. Le titre contient des informations sur chaque discours et l'agenda associé, le locuteur (coordonnées, genre et affiliation politique), le sujet abordé, la vitesse (nombre de mots par minute), la durée (en minutes), la longueur (nombre total de mots), le mode de délivrance (discours lu, impromptu, mixte) et quelques caractéristiques pertinentes du mot emprunté à l'anglais (nom commun, nom propre, mot unique ou chaîne de mots, acronyme et éventuelles difficultés liées à la prononciation). Outre les titres, le corpus contient un tableau de bord qui offre des informations sur l'usage de chaque emprunt intégral à l'anglais en espagnol (définitions du dictionnaire, présence de l'entrée dans les bases de données de l'U.E. si applicable) et sur les stratégies spécifiques adoptées par les interprètes et les traducteurs face au même phénomène linguistique (voir tableau 6).

Tableau 6

Taxonomie des stratégies d'Anglintrad

Stratégie	Définition et exemples (italien/espagnol)	
1) SUPPRESSION	Le phénomène disparait dans le texte cible (Pym 2008; Korpal 2012)	
	IT	/il principio del paese d'origine criticato come precursore del **dumping** sociale è stato soppresso e sostituito dal principio del paese di destinazione/
	ES	/el principio del país de origen criticado **…** ha sido suprimido y reemplazado por el principio de país de destino/
2) TRANSPOSITION EXACTE	Le phénomène est traduit par le même mot, sans modification (a) ou avec une simple adaptation morphologique (b) (Giménez-Folqués 2012)	
	IT	a) /ehm sento dire che mandano-hanno mandato due esperti…/ beh ehm a Lampedusa c'è un dramma…anche umanitario di enorme rilievo/ qui bisogna affrontarlo con una **task force** </taske fors/> adeguata…ehm ehm…rimediando agli errori e alle inadempienze del recente passato/ b) il governo italiano da tempo… con lo **slogan** Immigrazione Zero ha infatti smantellato il centro di accoglienza esistente e ha ridotto la-le strutture e ha tolto all'Italia la possibilità di fronteggiare l'immigrazione clandestina/
	ES	a) /han enviado dos expertos/ pero la situación en Lampedusa es dramática también desde un punto de vista humanitario/ ehm… hay que dotarse de una **task force** con… una composición adecuada y no podemos utilizar recetas del pasado para nada/ b) /el **eslogan** en Italia es Inmigración Cero…y por eso se ha desmantelado el centro de acogida…por lo tanto Italia ya no dispone de ninguna estructura que permita…luchar contra la…inmigración.
3) GÉNÉRALISATION	L'intention communicative ou le concept de base est rendu de manière générique (Al-Khanji, El-Shiyab *et al.* 2000; Bartlomiejczyk 2006).	
	IT	/questo approccio ovviamente favorirà chi si presenta con un **business plan** ragionevole/
	ES	/ y así ehm claro las empresas tienen que… presentar un **proyecto** que sea… razonable/
4) SUBSTITUTION	Reformulation lexicale (utilisation d'un synonyme) ou syntaxique du phénomène (Riccardi 1999; Li 2013).	
	IT	/i finanziamenti degli Stati europei non sono arrivati/ quelli della Commissione restano in **standby**/
	ES	/tenían que llegar los fondos/ por lo tanto… **todo ha quedado detenido**/

5) TRADUCTION	L'entrée est soit adaptée aux normes morphologiques et lexicales de la langue cible, soit traduite par un équivalent lexicalisé dans la langue cible (Hurtado Albir 2001).	
	IT	/non abbiamo votato contro perché ci sono dei diritti positivi che vengono tutelati e a questo proposito vanno segnalati quelli dei portatori di **handicap** e delle persone a ridotta m-mobilità/
	ES	/no votamos en contra porque hay derechos positivos bien plasmados y tutelados y amparados… sobre todo entonces los **discapacitados** o personas con movilidad reducida/
6) EXPANSION	L'interprète ou traducteur procède à des ajouts au texte source (Kalina 1998; Bartlomiejczyk 2006)	
	IT	/attraverso quali meccanismi/ attraverso nuovi meccanismi finanziari anzi noi ehm qua diciamo e abbiamo ripetuto più volte i **project bond**/
	ES	/pues bien con los nuevos ehm instrumentos y mecanismos financieros el BEI tiene que intervenir/se ha hablado aquí de esas **obligaciones de proyectos o bonos de proyectos**/

Enfin, pour une analyse en profondeur des emprunts repérés dans le corpus, une base de données spécifique a été développée et inclut une feuille d'analyse pour chaque phénomène. Cette feuille contient des informations détaillées sur l'usage de chaque emprunt à l'anglais en italien, selon les paramètres suivants : caractéristique grammaticale ; genre ; nombre ; référence lexicographique anglaise (issue de l'*Oxford English Dictionary*) ; sources lexicographiques et terminologiques italiennes ; contexte, années ; productivité du lexème ; informations phonétiques ; références ; et notes. On peut consulter des exemples de feuilles d'analyse conçues pour chaque anglicisme dans la section « Indice schede analitiche » de la plateforme Anglintrad. La quantité très importante de données collectées dans une feuille de calcul par emprunt à l'anglais, ainsi que le discours source, le discours interprété et la traduction associée ont été organisés dans une structure cohérente et facile d'accès, d'où le besoin d'inclure toutes ces données dans une plateforme multimédia en ligne, accessible à tout visiteur enregistré. L'adaptation du corpus multimodal en plateforme d'apprentissage a présenté des défis techniques liés au besoin de réunir divers contenus multimédias dans une plateforme en ligne tout-en-un, de constituer une base de données lexicales interrogeable et de supporter différents types de requêtes (recherche par mot ou par item dans le corpus et la base de données). Après un examen des options disponibles, dont plusieurs exigeaient une connaissance poussée des langages de programmation, c'est la solution open source WordPress 4.9.4 qui a été retenue. Cet outil de gestion de contenu rend possible la création d'un site Internet qui rassemble des contenus textuels et multimédias, qui peuvent être mis à jour de manière dynamique sans connaissance particulière d'aucun langage de programmation.

Le statut courant du projet de plateforme en matière de taille et d'accessibilité est le suivant : elle contient toutes les instances de données inclues dans le corpus Anglintrad, ainsi que les vidéos originales des locuteurs italiens, les versions audio des interprétations, le titre décrit ci-dessus, la transcription des textes sources et cibles, la traduction des rapports officiels, les stratégies adoptées pour chaque emprunt à l'anglais et un lien vers la fiche analytique qui renferme les informations sur l'usage de chacun des mots empruntés à l'anglais en italien. La plateforme en

ligne contient donc 249 emprunts à l'anglais organisés en 233 fiches (dont 16 contiennent plus d'une occurrence), toutes connectées à la base de données lexicales en italien par des hyperliens (figure 2).

Figure 2
Capture d'écran de la plateforme Anglintrad

1 – Standby

Tema specifico dell'intervento	Dichiarazioni del Presidente del Parlamento Europeo sulla situazione in Tunisia
Oratore	Pier Antonio Panzeri
Gruppo	S&D
Sesso	Uomo
Argomento	Politica
Velocità d'eloquio	media – 155 parole/min (4 minuti, 620 parole)
Tipo di delivery	letto
Tipo di lessema	Comune (C) Singolo (U)
Problemi di pronuncia nel testo originale	NO
Acronimo	NO

Comme on peut le voir d'après la description du corpus et de la plateforme Anglintrad, le projet repose sur une double perspective : il ne se fonde pas uniquement sur un corpus intermodal qui combine deux modes de traduction (l'interprétation simultanée et la traduction écrite), mais il est surtout un projet plus large dont la multimodalité est un élément crucial, intégré pour soutenir les apprentissages de l'interprétation et de la traduction.

Les applications possibles d'Anglintrad sont multiples, tant en matière de recherche en interprétation que de pédagogie (Bertozzi 2018b : 500). Par exemple, il est possible d'utiliser les discours sources du corpus dans le cadre de la pratique de l'interprétation, à des fins d'(auto) évaluation. De plus. La plateforme Anglintrad est utilisable a posteriori pour étudier les stratégies adoptées par les interprètes en formation et les comparer à celles des interprètes du PE en contexte professionnel. La plateforme peut aussi servir à se préparer avant une mission d'interprétation, car elle met à disposition des interprètes en formation des informations contextuelles. Enfin, la base de données lexicales peut être interrogée pour étudier et analyser les emprunts intégraux à l'anglais et leur usage en italien.

Les projets de développement d'Anglintrad incluent la collecte de nouveaux échantillons de discours, l'ajout de nouvelles variables ainsi que la création de requêtes avancées pour l'interface. De la même manière, l'analyse de stratégies d'interprétation pourrait s'étendre aux disfluences et à l'impact des emprunts intégraux à l'anglais dans les discours politiques institutionnels dans d'autres langues.

6. Discussion

Le présent aperçu des corpus inspirés d'EPIC (sections 3 à 5) montre le grand potentiel des corpus d'interprétation comme ressources linguistiques pour la communauté des interprètes et des traducteurs. Les nombreuses études effectuées sur ces corpus mettent en lumière des caractéristiques spécifiques des discours et des textes sources et cibles dans les séances plénières du PE et dans les conférences internationales. De plus, des applications pédagogiques sont en cours de développement, car l'accès aux fichiers multimédias (comme les vidéos et les enregistrements audios), aux transcriptions et aux métadonnées est pratique.

Ce sont l'annotation des caractéristiques verbales et non verbales et des informations extralinguistiques qui enrichissent les corpus et permettent aux analystes de dépasser le stade de l'étude des simples occurrences textuelles. Cet enrichissement a été rendu possible, dans des mesures diverses, en travaillant sur la conception des corpus, en ajustant les titres des transcriptions et en alignant les transcriptions aux fichiers multimédias correspondants. EPIC, le premier, a montré que la structure complexe d'un corpus d'interprétation trilingue réciproque exigeait d'organiser chaque fichier (comme les vidéos + les transcriptions de discours sources et les fichiers audio + les transcriptions des discours cibles) suivant des conventions de nommages fonctionnelles, qui offrent des informations sur la date de la session parlementaire, le type de discours (original ou interprétation) et la combinaison linguistique. Surtout, les attributs extralinguistiques annotés dans les titres de chaque transcription (tableau 1) rendent possible de rechercher et de gérer tous les fichiers efficacement et de dépasser le niveau verbal. Ces attributs ont été conçus pour prendre en compte les caractéristiques les plus saillantes des séances plénières du PE, et se fondent sur une documentation pertinente et l'observation des données collectées dans le corpus. Ces mêmes attributs ont fait l'objet d'ajustements dans les projets de corpus subséquents, comme dans DIRSI, où l'on trouve des informations sur le type de séance des conférences, sur les types spécifiques d'actes de discours (présentations d'articles, remarques d'introduction ou de conclusion), et le rôle communicatif des participants (qui ne sont plus uniquement désignés comme des locuteurs). Il a été possible d'utiliser certains attributs dans les différents corpus, comme « longueur du texte », « durée » et « vitesse de délivrance », mais des ajustements ont été nécessaires pour refléter avec exactitude les situations de communications en question (la vitesse moyenne de délivrance des discours aux séances plénières du PE est généralement supérieure à celle des conférences internationales). Des attributs supplémentaires ont été conçus, suivant la visée de chaque projet, comme dans le cas d'Anglintrad et de ses informations lexicales sur les emprunts intégraux trouvés dans les textes sources et les stratégies de traduction des textes cibles.

Quant aux annotations des occurrences verbales et non verbales contenues dans les transcriptions, elles étaient généralement plutôt limitées dans les projets de corpus étudiés. EPIC comprend uniquement des annotations sur les pauses vides ou remplies, les mots mal prononcés, tronqués, et les unités de sens. On trouve encore moins de telles annotations dans DIRSI, où seuls les mots mal prononcés, tronqués et les unités de sens sont annotés. En revanche, EPTIC a réintroduit les marques de ponctuation pour refléter les schémas d'intonation des locuteurs et pour aligner les textes sources et les textes ou discours cibles avec plus de précision. Comme aux tout débuts

du projet EPIC, il était clair que pour faciliter les processus de transcription et d'annotation, le nombre d'occurrences verbales et non verbales annotées serait limité, pour un usage intuitif. Dans les faits, les annotations dépendent strictement des objectifs de recherche, car on peut considérer la transcription comme une étape d'analyse en soi. Produire une transcription non enrichie peut néanmoins fournir une base facile à utiliser pour y ajouter de futures annotations, à des stades subséquents du projet, ou pour ajuster les tags existants pour de nouveaux usages des mêmes données.

Enfin, l'alignement des textes sources et des textes cibles, des transcriptions et des fichiers audio ou vidéo s'est désormais amélioré. Du fait que dans EPIC, toutes les données de corpus (comme les transcriptions et les enregistrements audio ou vidéo) se présentent encore sous la forme de fichiers séparés, les autres corpus offrent un accès pratique aux informations multimédias. L'alignement des textes a été soit manuel (DIRSI) soit automatique (EPTIC). En revanche, l'alignement texte-vidéo/son a été obtenu par l'annotation de marqueurs temporels dans les transcriptions, ou par l'inclusion d'un lien dans le titre de la transcription. Bien que ces processus puissent être chronophages et exiger des compétences technologiques, les récents progrès en matière de reconnaissance vocale et de diarisation des locuteurs en amélioreront certainement l'accessibilité.

Le tableau 7 montre un sommaire des caractéristiques principales des quatre projets de corpus décrits dans le présent article. Il est encourageant de remarquer qu'une même séance plénière du Parlement européen peut faire l'objet d'approches diverses avec des méthodes différentes de plus en plus sophistiquées. De la même manière, des événements différents, comme les conférences internationales représentées dans DIRSI, peuvent aussi être étudiés en tirant parti des méthodes initialement développées pour le projet EPIC.

Tableau 7

Sommaire des principales caractéristiques des quatre corpus étudiés

	EPIC	DIRSI	EPTIC	ANGLINTRAD
Données sources	Séances plénières du PE	Conférences médicales	Séances plénières du PE	Séances plénières du PE
Mode de traduction	Interprétations simultanées	Interprétations simultanées	Interprétations simultanées + traductions écrites	Interprétations simultanées + traductions écrites
Langues	EN, IT, ES	EN, IT	EN, IT, FR (PL, SL)*	IT, ES
Direction de la traduction	EN>IT/ES IT>EN/ES ES>IT/EN	EN<>IT	EN<>IT EN<>FR EN>SL* PL>EN*	IT>ES
Structure du corpus	Parallèle + Comparable (réciproque)	Parallèle + Comparable (réciproque)	Parallèle + Comparable (réciproque) *monodirectionnel	Parallèle, monodirectionnel
Alignement texte source-texte cible	non	oui	oui	oui

Alignement texte-vidéo/audio	non	oui	oui	oui
Metadonnées	titre (locuteur et acte de discours)	titre (locuteur et acte de discours)	titre (locuteur et acte de discours)	titre (locuteur et acte de discours + emprunt et stratégie d'interprétation)
Accès	Portail CoLiTec (*NoSketch Engine*), uniquement des transcriptions; catalogue de l'ELRA (transcriptions + fichiers multimédias)	Interface Web: transcriptions + fichiers multimédias	Portail CoLiTec (*NoSketch Engine*), transcriptions + fichiers multimédias	Plateforme Anglintrad: transcriptions + fichiers multimédias + détails lexicaux

7. Remarques en conclusion

Plus de quinze ans se sont écoulés depuis la première publication sur le corpus EPIC dans *Meta* (Monti, Bendazzoli *et al*. 2005). Depuis, la même méthodologie a été utilisée, avec des ajustements nécessaires, pour concevoir de nouvelles ressources linguistiques, comme des corpus d'interprétation, qui deviennent de plus en plus multimodaux. Les transcriptions de discours sont accompagnées de données multimédias associées, au format audio ou vidéo. La multimodalité et les attributs des métadonnées conçus pour les types d'événements étudiés dans chaque corpus fournissent désormais aux chercheurs des outils analytiques qui leur permettent de dépasser le cadre verbal et d'obtenir une vision plus large du discours interprété situé.

Le présent article fournit un aperçu de projets de corpus d'interprétation qui ont certainement utilisé les connaissances acquises pour le développement d'EPIC (section 2). L'ajout de textes sources et cibles écrits a mené à la création d'EPTIC (section 4), un corpus intermodal qui fournit des liens vers les vidéos ou les fichiers audio associés aux discours oraux cibles et sources. Toujours en relation avec les débats parlementaires, Anglintrad (section 5) a été conçu avec la visée spécifique d'étudier la fréquence des anglicismes dans les discours sources en italien et les stratégies adoptées par les interprètes de conférence pour les traduire en espagnol. Le corpus dispose d'une plateforme en ligne qui offre de très nombreuses ressources terminologiques et qui peut être facilement exploitée à des fins de formation. Des conférences médicales internationales (en anglais et en italien), tenues pour le marché italien, fournissent les données sources du corpus DIRSI (section 3), qui a été développé juste après EPIC et qui peut être enrichi d'alignements de textes sources et cibles et d'alignement texte-audio. Les caractéristiques particulières des discours médiés par des interprètes ont requis quelques ajustements des attributs des métadonnées conçues pour capturer et enregistrer des informations sur la situation de communication, les participants et leurs actes de discours.

Malgré les progrès considérables effectués dans le domaine des études d'interprétation basées sur des corpus depuis quinze ans (Bendazzoli, Russo *et al*. 2018; Russo, Bendazzoli *et al*. 2018), les projets de corpus présentés dans cet article

demeurent de taille limitée, leur potentiel pour la recherche reste à explorer et leur potentiel pour l'enseignement et la formation professionnelle est encore moins exploité. Les sources de données n'ayant jamais été aussi accessibles que de nos jours (y compris celles des Nations Unies, voir Dayter 2018), et les outils informatiques pour la transcription et l'encodage étant de plus en plus intuitifs, on peut espérer que cette profusion de ressources continue d'inspirer de nouvelles recherches en interprétation, avec une approche plus large du discours, en renforçant la multimodalité et en informant mieux les membres intéressés des communautés scientifiques et professionnelles concernées.

REMERCIEMENTS

Les auteurs remercient Silvia Bernardini et Adriano Ferraresi pour leur conseil et leur soutien inestimables.

NOTES

† Le présent article a été traduit de l'anglais par Audrey Canalès. Bien qu'il soit le résultat d'une collaboration, Claudio Bendazzoli peut être identifié comme l'auteur des sections 1, 3, 4, 6 ; Mariachiara Russo comme l'auteure de la section 2 ; Michela Bertozzi comme l'auteure de la section 5 ; la section 7 a été coécrite.

* Dipartimento di Scienze economico-sociali e matematico-statistiche.

** Dipartimento di Interpretazione e Traduzione.

1. Pour un aperçu spécifique des transcriptions en recherche en interprétation, voir Leech (1997) et Niemants (2012).
2. De nombreux autres corpus d'interprétations ont été collectés depuis. Pour en avoir un aperçu, Setton (2011) et Bendazzoli (2018).
3. Corpora, Linguistics, Technology Research Centre (CoLiTec) (Dernière mise à jour : 4 avril 2018) : *Corpora and tools*. Forlì : Dipartimento di Interpretazione e Traduzione, Università di Bologna. Consulté le 30 novembre 2019, <https://corpora.dipintra.it>.
4. European Language Resources Association (Dernière mise à jour : 18 novembre 2016) : European Parlament Interpretation Corpus (EPIC). *ELRA*. Consulté le 30 novembre 2019, <http://catalog.elra.info/product_info.php?products_id=1145>.
5. L'Association internationale des interprètes de conférence (AIIC) classifie les langues de travail des interprètes comme suit : langue A (première langue), langue B (langue active, les interprètes peuvent traduire de et vers leur langue B) et langue C (langue passive, les interprètes peuvent traduire de leur langue C mais pas vers elle).
6. Woods, David K. (7 novembre 2017) : *Transana*. Version 3.21. Consulté le 20 mai 2019, <https://www.transana.com>.
7. Les mêmes transcriptions ont été traitées pour être utilisées avec la suite d'outils *Corpus Work Bench* (Christ 1994), ce qui prouve la compatibilité de leur format.
8. Laboratorio de Lingüistíca Informática (Dernière mise à jour : 20 septembre 2018) : *Directionality in Simultaneous Interpreting*. Madrid : Universidad Autónoma de Madrid. Consulté le 30 mai 2019, <http://cartago.lllf.uam.es/static/dir-si/dir-si.html>.
9. Miličević Petrović, Maja, Bernardini, Silvia, Ferraresi, Adriano *et al.* (Dernière mise à jour : 22 novembre 2018) : *European Parliament Translation and Interpreting Corpus (EPTIC)*. Forlì : Dipartimento di Interpretazione e Traduzione, Università di Bologna. Consulté le 30 mai 2019, <https://corpora.dipintra.it/eptic/?section=about>.
10. Dans ce cas, *AegiSub* a été utilisé, mais tout autre logiciel de sous-titrage aurait convenu. Martin Hansen, Niels et Braz Monteiro, Rodrigo (7 décembre 2014) : *Aegisub*. Version 3.2.2. Consulté le 30 mai 2019, <http://www.aegisub.org/>.
11. Bertozzi, Michela (Dernière mise à jour : 23 mai 2018) : *Anglintrad Corpus*. Consulté le 30 mai 2019, <http://anglintradcorpus.altervista.org/>.
12. L'expression *emprunt intégral* est une référence aux catégorisations de Bombi (2005) et Furiassi (2010) des anglicismes en italien, et désigne un emprunt qui passe dans la langue cible sans subir aucune modification morphologique ou phonétique.

RÉFÉRENCES

ALBL-MIKASA, Michaela (2013) : ELF speakers' restricted power of expression. Implications for interpreters' processing. *In* : Maureen EHRENSBERGER-DOW, Birgitta ENGLUND DIMITROVA, Séverine HUBSCHER-DAVIDSON *et al*, dir. *Describing cognitive processes in translation : Acts and events. Translation and Interpreting Studies*. 8(2):191-210.

AL-KHANJI, Rajai, EL-SHIYAB, Said et HUSSEIN, Riyadh (2000) : On the use of compensatory strategies in simultaneous interpretation. *Meta*. 45(3):548-557.

BALDRY, Anthony et THIBAULT, J. Paul (2001) : Towards multimodal corpora. *In* : Guy ASTON et Lou BURNARD, dir. *Corpora in the Description and Teaching of English. Papers from the 5th ESSE Conference*. Bologne : Clueb, 277-305.

BARTLOMIEJCZYK, Magdalena (2006) : Strategies of simultaneous interpreting and directionality. *Interpreting*. 8(2):149-174.

BENDAZZOLI, Claudio (2010) : *Corpora e interpretazione simultanea* [Les corpus et l'interprétation simultanée]. Bologne : Asterisco.

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*. Francfort-sur-le-Main : Peter Lang, 91-117.

BENDAZZOLI, Claudio (2017) : Benefits and drawbacks of English as a Lingua Franca and as a working language : The case of conferences mediated by simultaneous interpreters. *In* : Cecilia BOGGIO et Alessandra MOLINO, dir. *English in Italy : Linguistic, Educational and Professional Challenges*. Milan : FrancoAngeli, 119-141.

BENDAZZOLI, Claudio (2018) : Corpus-based interpreting studies : Past, present and future developments of a (wired) cottage industry. *In* : Mariachiara RUSSO, Claudio BENDAZZOLI et Bart DEFRANCQ, dir. *Making Way in Corpus-based Interpreting Studies*. Singapour : Springer, 1-19.

BENDAZZOLI, Claudio (2019) : Discourse markers in English as a target language : The use of *so* by simultaneous interpreters. *Textus*. 32(1):183-201.

BENDAZZOLI, Claudio et SANDRELLI Annalisa (2009) : Corpus-based interpreting studies : Early work and future prospects. *Tradumatica*. 7:9p. Consulté le 30 mai 2019, <http://webs2002.uab.es/tradumatica/revista/num7/articles/08/08art.htm>.

BENDAZZOLI, Claudio, RUSSO, Mariachiara et DEFRANCQ, Bart, dir. (2018) : *New Findings in Corpus-based Interpreting Studies. inTRAlinea*. Numéro spécial. Consulté le 30 mai 2019, <http://www.intralinea.org/specials/cbis>.

BENDAZZOLI, Claudio, SANDRELLI, Annalisa et RUSSO, Mariachiara (2011) : Disfluencies in simultaneous interpreting : A corpus-based analysis. *In* : Alet KRUGER, Kim WALLMACH et Jeremy MUNDAY, dir. *Corpus-based Translation Studies : Research and Applications*. Londres/New York : Continuum, 282-306.

BERNARDINI, Silvia (2016) : Intermodal corpora : A novel resource for descriptive and applied translation studies. *In* : Gloria CORPAS PASTOR et Miriam SEGHIRI, dir. *Corpus-based Approaches to Translation and Interpreting : From Theory to Applications*. Francfort-sur-le-Main : Peter Lang, 129-148.

BERNARDINI, Silvia, FERRARESI, Adriano et MILICEVIC, Maja (2016) : From EPIC to EPTIC – Exploring simplification in interpreting and translation from an intermodal perspective. *Target*. 28(1):61-86.

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*. Singapour : Springer, 21-42.

BERTOZZI, Michela (2018a) : ANGLINTRAD : Towards a purpose specific interpreting corpus. *In* : Claudio BENDAZZOLI, Mariachiara RUSSO et Bart DEFRANCQ, dir. *New Findings in Corpus-based Interpreting Studies. inTRAlinea*. Numéro special. Consulté le 20 mai 2019, <http://www.intralinea.org/specials/article/2317>.

Bertozzi, Michela (2018b) : L'anglicismo in interpretazione e in traduzione dall'italiano allo spagnolo : uno studio sperimentale attraverso il corpus Anglintrad [L'anglicisme en interprétation et en traduction de l'italien vers l'espagnol : une étude expérimentale basée sur le corpus Anglintrad]. Thèse de doctorat non publiée. Bologne : Université de Bologne.

Biber, Douglas (1993) : Representativeness in corpus design. *Literary and Linguistic Computing.* 8(4):243-257.

Bombi, Raffaella (2005) : *La linguistica del contatto. Tipologie di anglicismi nell'italiano contemporaneo e riflessi metalinguistici* [La linguistique de contact. Typologie des anglicismes en italien contemporain et réflexions métalinguistiques]. Rome : Il Calamo.

Bowker, Lynne et Pearson, Jennifer (2002) : *Working with Specialized Language. A Practical Guide to Using Corpora*. Londres/New York : Routledge.

Castagnoli, Sara (2016) : Investigating trainee translators' contrastive pragmalinguistic competence : A corpus-based analysis of interclausal linkage in learner translations. *The Interpreter and Translator Trainer.* 10(3):343-363.

Christ, Oliver (1994) : A modular and flexible architecture for an integrated Corpus Query System. *In* : Ferenc Kiefer, Gábor Kiss et Júlia Pajzs, dir. *Papers in computational lexicography, COMPLEX '94*. (COMPLEX '94 : 3rd Conference on Computational Lexicography and Text Research, Budapest, 7-10 juillet 1994). Budapest : Budapest : Research Institute for Linguistics, Hungarian Academy of Sciences, 23-32.

Dayter, Daria (2018) : Describing lexical patterns in simultaneously interpreted discourse in a parallel aligned corpus of Russian-English interpreting (SIREN). *Forum.* 16(2):241-264.

Ferraresi, Adriano et Bernardini, Silvia (2019) : Building EPTIC : A many-sided, multipurpose corpus of EU Parliament proceedings. *In* : Irene Doval et María Teresa Sánchez Nieto, dir. *Parallel Corpora for Contrastive and Translation Studies. New Resources and Applications*. Amsterdam/Philadelphie : John Benjamins, 123-139.

Ferraresi, Adriano, Bernardini, Silvia, Miličević Petrović, Maja *et al.* (2018) : Simplified or not simplified ? The different guises of mediated English at the European Parliament. *Meta*. 63(3):717-738.

Furiassi, Cristiano (2010) : *False Anglicisms in Italian*. Monza : Polimetrica International Scientific Publisher.

Gao, Fei et Wang, Binhua (2017) : A multimodal corpus approach to dialogue interpreting studies in the Chinese context : Towards a multi-layer analytic framework. *The Interpreter's Newsletter.* 22:17-38.

Ghiselli, Serena (2015) : *Le sfide traduttive dei sintagmi nominali con modificatori in posizione prenominale nell'interpretazione simultanea dall'inglese in italiano : uno studio sul corpus EPIC* [Les défis de la traduction des syntagmes nominaux avec modificateurs en position prénominale en interprétation simultanée de l'anglais vers l'italien : une étude sur le corpus EPIC]. Mémoire de maîtrise non publié. Forlì : Université de Bologne.

Ghiselli, Serena (2018) : The translation challenges of premodified noun phrases in simultaneous interpreting from English into Italian - A corpus-based study on EPIC. *In* : Claudio Bendazzoli, Mariachiara Russo et Bart Defrancq, dir. *New findings in corpus-based interpreting studies. inTRAlinea*. Numéro special. Consulté le 20 mai 2019, <http://www.intralinea.org/specials/article/2322>.

Giménez-Folqués, David (2012) : Los extranjerismos en el español académico del siglo XXI. *Normas Revista de Estudios Lingüísticos Hispanicos.* Annexe 3:9-79.

Halverson, Sandra (1998) : Translation studies and representative corpora : Establishing links between translation corpora, theoretical/descriptive categories and a conception of the object of study. *Meta*. 43(4):494-514.

Hurtado Albir, Amparo (2001) : *Traducción y traductología*. Madrid : Cátedra.

Kalina, Sylvia (1998) : *Strategische Prozesse beim Dolmetschen : Theoretische Grundlagen, empirische Fallstudien, didaktische Konsequenzen* [Processus stratégiques en interprétation : principes théoriques, études de cas empiriques et leurs conséquences pour l'enseignement]. Tubingue : Gunter Narr.

Korpal, Pawel (2012) : Omission in simultaneous interpreting as a deliberate act. *In* : Anthony Pym et David Orrego Carmona, dir. *Translation Research Projects*. Vol. 4. Tarragone : Intercultural Studies Group, 103-111.

Laviosa, Sara (1998) : Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta* 43(4):557-570.

Leech, Geoffrey (1997) : Introducing corpus annotation. *In* : Roger Garside, Geoffrey Leech et Anthony McEnery, dir. *Corpus Annotation. Linguistic Information from Computer Text Corpora*. Londres/New York : Longman, 1-18.

Li, Xiangdong (2013) : Are interpreting strategies teachable ? Correlating trainees' strategy use with trainers' training in the consecutive interpreting classroom. *The Interpreter's Newsletter.* 18:105-128.

Lobascio, Marco (2015) : *Genitive variation and unique items hypothesis in simultaneous interpreting from Italian into English. An intermodal study based on EPIC*. Mémoire de maîtrise non publié. Forlì : Université de Bologne.

Malmkjaer, Kirsten (2003) : On a pseudosubversive use of corpora in translator training. *In* : Federico Zanettin, Silvia Bernardini et Dominic Stewart, dir. *Corpora in Translator Education*. Manchester : St. Jerome, 119-134.

McEnery, Tony, Xiao, Richard et Yukio, Tono (2006) : *Corpus-based Language Studies. An Advanced Resource Book*. Londres/New York : Routledge.

Monti, Cristina, Bendazzoli, Claudio, Sandrelli, Annalisa *et al.* (2005) : Studying directionality in simultaneous interpreting through an electronic corpus : EPIC (European Parliament Interpreting Corpus). *Meta*. 50(4):16p. Consulté le 17 mai 2019, <https://www.erudit.org/fr/revues/meta/2005-v50-n4-meta1024/019850ar/>.

Niemants, Natacha (2012) : The transcription of interpreting data. *Interpreting.* 14(2):165-191.

O'Connell, C. Daniel et Kowal, Sabine (1994) : Some current transcription systems for spoken discourse : A critical analysis. *Pragmatics.* 4(1):81-107.

O'Connell, C. Daniel et Kowal, Sabine (1999) : Transcription and the issue of standardization. *Journal of Psycholinguistic Research.* 28(2):103-120.

Partington, Alan, Morley, John et Haarman, Louann (2004) : *Corpora and Discourse*. Berne : Peter Lang.

Pym, Anthony (2008) : On omission in simultaneous interpreting : Risk analysis of a hidden effort. *In* : Gyde Hansen, Andrew Chesterman et Heidrun Gerzymisch-Arbogast, dir. *Efforts and Models in Interpreting and Translation Research. A Tribute to Daniel Gile.* Amsterdam/Philadelphie : John Benjamins, 83-105.

Riccardi, Alessandra (1999) : Interpretazione simultanea : strategie generali e specifiche [Interprétation simultanée : strategies générales et spécifiques]. *In* : Caterina Falbo, Mariachiara Russo et Francesco Straniero Sergio, dir. *Interpretazione simultanea e consecutiva : Problemi teorici e metodologie didattiche* [Interprétation simultanée et consécutive : problèmes théoriques et méthodologies didactiques]. Milan : Hoepli, 161-174.

Russo, Mariachiara (2010) : Reflecting on interpreting practice : Graduation theses based on the European Parliament Interpreting Corpus (EPIC). *In* : Lew Zybatow, dir. *Translationswissenschaft – Stand und Perspektiven* [Traductologie – *État de l'art et perspectives]*. Francfort-sur-le-Main : Peter Lang, 35-50.

Russo, Mariachiara (2011) : Text processing patterns in simultaneous interpreting (Spanish-Italian) : A corpus-based study. *In* : Pöck Wolfgang, Onhneiser Ingeborg et Peter Sandrini, dir. *Translation – Sprachvariation – Mehrsprachigkeit. Festschrift für Lew Zybatow zum 60. Geburtstag.* Francfort-sur-le-Main : Peter Lang, 83-103.

Russo, Mariachiara (2018) : Speaking patterns and gender in the European Parliament Interpreting Corpus. A quantitative study as a premise for qualitative investigations. *In* : Mariachiara Russo, Claudio Bendazzoli et Bart Defrancq, dir. *Making Way in Corpus-based Interpreting Studies*. Singapour : Springer, 115-131.

Russo, Mariachiara, Bendazzoli, Claudio et Defrancq, Bart (2018) : *Making Way in Corpus-based Interpreting Studies*. Singapour : Springer.

Russo, Mariachiara, Bendazzoli, Claudio et Sandrelli Annalisa (2006) : Looking for lexical patterns in a trilingual corpus of source and interpreted speeches : Extended analysis of EPIC (European Parliament Interpreting Corpus). *Forum*. 4(1):221-254.

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*. Francfort-sur-le-Main : Peter Lang, 53-90.

Rychlý, Pavel (2007) : Manatee/Bonito - A modular corpus manager. *In* : Petr Sojka et Aleš Horák, dir. *RASLAN 2007 - Recent advances in Slavonic natural language processing*. (RASLAN2007 : First workshop on recent advances in Slavonic natural language processing. Karlova Studánka, 14-16 décembre 2007). Brno : Université Masaryk, 65-70.

Sandrelli, Annalisa et Bendazzoli, Claudio (2005) : Lexical patterns in simultaneous interpreting : A preliminary investigation of EPIC (European Parliament Interpreting Corpus). *Proceedings from the Corpus Linguistics Conference Series*. 1(1):18p. Consulté le 20 Mai 2019, <https://www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2005-conf-e-journal.aspx>.

Sandrelli, Annalisa, Bendazzoli, Claudio et Russo, Mariachiara (2010) : European Parliament Interpreting Corpus (EPIC) : Methodological issues and preliminary results on lexical patterns in simultaneous interpreting. *International Journal of Translation*. 22(1-2):165-203.

Schweda-Nicholson, Nancy (1987) : Linguistic and extralinguistic aspects of simultaneous interpretation. *Applied Linguistics*. 8(2):194-205.

Setton, Robin (2011) : Corpus-based interpreting studies (CIS) : Overview and prospects. *In* : Alet Kruger, Kim Wallmach et Jeremy Munday, dir. *Corpus-based Translation Studies : Research and Applications*. Londres/New York : Continuum, 31-75.

Shlesinger, Miriam (1998) : Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta*. 43(4):486-493.

Shlesinger, Miriam (2008) : Towards a definition of Interpretese. An intermodal, corpus-based study. *In* : Gyde Hansen, Andrew Chesterman et Heidrun Gerzymisch-Arbogast, dir. *Efforts and Models in Interpreting and Translation Research. A Tribute to Daniel Gile*. Amsterdam/Philadelphie : John Benjamins, 237-253.

Shlesinger, Miriam et Ordan, Noam (2012) : More spoken or more translated ? Exploring a known unknown of simultaneous interpreting. *Target*. 24(1):4360.

Spinolo, Nicoletta et Garwood, Christopher (2010) : To kill or not to kill : Metaphors in simultaneous interpreting. *Forum*. 8(1):181-211.

Zanettin, Federico (2012) : *Translation-driven Corpora : Corpus Resources for Descriptive and Applied Translation Studies*. Manchester : St. Jerome.