CORPUS

Corpus

22 | 2021
Du recueil à l'outillage des corpus oraux : comment accéder à la variation ?

Vers un outillage informatique optimisé pour corpus langagiers oraux en vue d'une exploitation textométrique : le cas des interrogatives partielles dans ESLO

Flora Badin, Loïc Liégeois, Gabriel Thiberge et Christophe Parisse

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/5752
DOI : 10.4000/corpus.5752
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Référence électronique
Flora Badin, Loïc Liégeois, Gabriel Thiberge et Christophe Parisse, « Vers un outillage informatique optimisé pour corpus langagiers oraux en vue d'une exploitation textométrique : le cas des interrogatives partielles dans ESLO », *Corpus* [En ligne], 22 | 2021, mis en ligne le 28 janvier 2021, consulté le 16 février 2021. URL : http://journals.openedition.org/corpus/5752 ; DOI : https://doi.org/10.4000/corpus.5752

Ce document a été généré automatiquement le 16 février 2021.

Vers un outillage informatique optimisé pour corpus langagiers oraux en vue d'une exploitation textométrique : le cas des interrogatives partielles dans ESLO

Flora Badin, Loïc Liégeois, Gabriel Thiberge et Christophe Parisse

1. Introduction

1 Si la disponibilité croissante des corpus oraux participe à l'essor que vit actuellement la linguistique de corpus, elle oblige les chercheurs à prendre en main les problématiques liées à la structuration, la diffusion et le partage des données. En effet, alors que l'hétérogénéité des méthodologies de structuration des corpus oraux apparaît comme une marque de la vitalité du domaine, celle-ci peut également représenter un frein certain dans la réutilisation des données par la communauté. Il n'est pas toujours facile de jongler entre les formats de structuration des données, fortement dépendants des logiciels utilisés au départ pour transcrire et/ou annoter et/ou structurer le corpus. Ainsi, nous sommes partis d'une problématique simple : comment réutiliser le corpus ESLO (Eshkol-Taravella *et al.*, 2011), qui apparaît actuellement comme l'un des corpus oraux les plus volumineux librement disponibles, pour mener une étude sociolinguistique à la croisée des axes syntaxiques et pragmatiques ?

2 Alors que le corpus est à ce jour majoritairement distribué dans son format d'origine, celui de l'outil de transcription et d'annotation Transcriber (Barras, Geo & Wu, 2001), l'étude que nous souhaitions mener nécessitait l'utilisation d'un outil de textométrie puissant, capable de gérer une annotation morphosyntaxique ainsi que la richesse des métadonnées du corpus. Nous nous sommes dirigés vers l'outil TXM (Heiden, Magué, & Pincemin, 2010), le plus à même selon nous de répondre à ce cahier des charges. Si la

possibilité d'import de données structurées à l'aide de l'outil Transcriber est déjà prévue par TXM, cet import ne permet pas l'utilisation de certaines fonctionnalités de TXM comme la création de partitions de corpus ou de sous-corpus fondés sur les informations disponibles dans les métadonnées locuteurs, particulièrement riches pour le corpus ESLO (âge, profession, niveau d'étude etc.).

3 L'objectif de notre article est de présenter la méthodologie mise en œuvre pour passer d'une série de fichiers de transcription au format Transcriber accompagnés de leurs métadonnées correspondantes (au niveau locuteur et enregistrement) à une base exploitable à l'aide de l'outil TXM. Dans une première partie, nous reviendrons sur les problématiques liées à la mise à disposition des corpus oraux. Nous exposerons les solutions actuellement disponibles pour la conversion des données et leur réutilisation dans différents logiciels du domaine de la linguistique de corpus. Nous présenterons également différentes facettes du projet ESLO : le corpus et ses spécificités ouvrant la voie à l'analyse de phénomènes linguistiques dans une perspective micro-diachronique, ainsi que la plateforme d'interrogation associée. Dans une deuxième section, nous détaillerons le cahier des charges qui a guidé notre travail de restructuration du corpus ESLO. Nous verrons comment celui-ci a été conditionné par des besoins généraux, liés à l'analyse de la langue orale, et d'autres plus spécifiquement liés à la thématique de l'analyse linguistique que nous souhaitions effectuer sur le corpus. La partie suivante se focalisera quant à elle sur les traitements réalisés sur le corpus de départ afin d'obtenir des données finement analysables au moyen de TXM. Enfin, nous verrons comment nous avons réussi à exploiter l'outil afin d'analyser l'usage des interrogatives partielles dans une perspective micro-diachronique en français parlé de la région Orléanaise par les différents locuteurs de la section entretien du corpus ESLO. Nous présenterons une série d'analyses qui s'appuie sur la richesse des métadonnées disponibles ainsi que sur une lemmatisation et une annotation en parties du discours des données transcrites.

2. Des corpus oraux diffusés pour être réutilisés. Oui, mais comment ?

2.1. Gérer l'hétérogénéité des formats

4 Depuis quelques années, le développement d'infrastructures nationales de stockage et de partage des corpus (oraux, écrits ou multimodaux) a grandement facilité les échanges de ressources entre chercheurs – voir par exemple ORTOLANG (https://www.ortolang.fr) ou Cocoon (https://cocoon.huma-num.fr/). La circulation des données entre chercheurs s'avère essentielle, particulièrement pour les données orales et multimodales. Au vu du temps consacré à la constitution d'un corpus lors des tâches de recueil, de transcription et d'annotation des interactions, il apparaît précieux de pouvoir réutiliser des données existantes. Cette option permet ainsi de mener des analyses linguistiques sans consacrer un temps important à toutes ces tâches. Toutefois, ce travail ne peut pas être réalisé sans prendre en considération la méthodologie de recueil et de structuration du corpus que l'on réutilise, ce qui oblige le chercheur à se confronter aux problématiques liées aux formats des données. Nous distinguons ici deux types de format : le format de transcription et d'annotation, lié aux conventions définies en amont du projet, et le format de structuration, généralement dépendant du logiciel choisi pour transcrire et annoter les interactions. Dans les

domaines de la linguistique orale et multimodale, plusieurs logiciels standards ont été adoptés par la communauté, chacun d'eux étant (au moins à l'origine) lié à un axe d'analyse linguistique comme Praat (Boersma & Weenink, 2019) pour la phonétique et la prosodie ou ELAN (Wittenburg *et al.*, 2006), pour la gestualité et l'étude des langues peu décrites. Bien que ciblé la plupart du temps sur un type d'usage, il semble normal qu'un corpus puisse être réutilisé à plusieurs fins scientifiques et par des communautés scientifiques diverses, et donc avec différents logiciels.

5 C'est dans cet objectif que Parisse a développé, à partir de 2015, un logiciel de conversion de données dans le cadre de ses travaux effectués au sein de l'équipement ORTOLANG (Outils et Ressources pour un Traitement Optimisé de la LANGue, Équipements d'Excellence ANR–11–EQPX–0032) et du consortium CORLI (CORpus, Langues et Interactions) de la Très Grande Infrastructure de Recherche Huma-Num. La réalisation technique des logiciels de conversion se présente de deux manières. La première est celle d'une bibliothèque programmée en langage Java par Myriam Majdoub et Christophe Parisse. Ce logiciel est disponible en ligne (http://ct3.ortolang.fr/teicorpo/) et ses sources sont sous licence BSD 2-Clause et sont librement disponibles sur GitHub (https://github.com/christopheparisse/teicorpo). Une interface sous forme de service web est également disponible (http://ct3.ortolang.fr/teiconvert/). Les sources de cette version sont distribuées sous la même licence (https://github.com/christopheparisse/teiconvert).

6 Ces outils font partie des maillons d'une chaîne de traitement nommée TEI-CORPO et utilisant le format TEI (Text Encoding Initiative) comme format pivot (Parisse *et al.*, sous presse). Ce format pivot partage les informations issues des différentes sources ou logiciels, sans perte d'informations, et permet en retour d'utiliser d'autres outils. Si ce format respecte complètement les consignes de la norme TEI/ISO pour ISO 24624:2016, celui-ci va plus loin puisqu'il permet le codage de fichiers de transcription issus des logiciels les plus utilisés dans la communauté de l'oral. Les logiciels ont été choisis en fonction des usages constatés lors de réunions de travail du consortium IRCOM de la TGIR Huma-Num (2012-2015) : CLAN (MacWhinney, 2000), ELAN, Praat et Transcriber. De plus, il faut être capable de récupérer, pour un corpus, toutes les métadonnées fournies par les outils d'origine pour les intégrer dans le format TEI et permettre leur conservation dans le format de sortie. TEICORPO a utilisé la TEI et la norme TEI/ISO comme un format souple s'adaptant aux caractéristiques des formats d'origine des données. Ceci a amené à élargir les usages de la TEI présentés dans la norme TEI/ISO en ajoutant des informations supplémentaires, sans modifier toutefois le schéma de la TEI (utilisation de champs commentaires et notes). Sans l'insertion de ces informations, une conversion des logiciels source vers la TEI serait possible, mais une conversion inverse de la TEI vers ces logiciels ne le serait plus. TEI-CORPO garantit ainsi l'existence d'un aller-retour entre un logiciel source et la TEI. En revanche, les conversions croisées (d'un format logiciel à un autre) sont limitées par les caractéristiques des logiciels visés.

7 Dès l'origine, TEI-CORPO a aussi été développé pour permettre une conversion des données TEI depuis ou vers des formats autres que ceux des logiciels d'annotation de corpus oraux. Il est ainsi possible d'importer et d'exporter des données au format texte UNICODE et aux formats Microsoft Word ou Microsoft Excel (uniquement les variantes du format Office Open XML, norme ISO/CEI 29500, extensions usuellement appelées .docx ou .xlsx) et d'exporter vers les formats de données utilisés par des logiciels de

textométrie comme TXM, Iramuteq (Ratinaud, 2009), Lexico (Lamalle *et al.*, 2003) ou Le Trameur (Zimina & Fleury, 2015).

2.2. Cas pratique : réutiliser le corpus ESLO

8 Le corpus ESLO apparaît comme le corpus oral de langue française le plus volumineux actuellement librement accessible. Au total, il se compose d'environ 4,9 millions de mots transcrits à partir de près de 422 heures de données audio recueillies entre 1960 et aujourd'hui. Le corpus peut être divisé en deux parties, généralement nommées ESLO1 (1968-1974) et ESLO2 (2008 à nos jours). Si plusieurs décennies séparent les deux périodes de collecte de données, les objectifs scientifiques sont identiques pour les deux parties du corpus. Il s'agit de recueillir des paroles de la vie quotidienne pour obtenir des données témoignant de la variété et la diversité des langues parlées quotidiennement dans la ville d'Orléans, tout en apportant un témoignage précieux sur la ville en elle-même. Dans cet objectif, le protocole mis en place pour le recueil du corpus ESLO varie les situations d'interaction. Si la majeure partie de ces dernières correspondent à des entretiens entre un enquêteur et une personne enquêtée, le corpus se compose également de paroles recueillies dans les commerces ou dans la rue, à la sortie d'un cinéma par exemple, d'interviews de personnalités orléanaises, des conférences universitaires et des communications téléphoniques. Pour chaque enregistrement, une série de métadonnées permet d'avoir accès à des informations importantes comme la durée et la date de l'enregistrement ou la qualité de celui-ci.

9 Certains contextes d'interaction se retrouvent dans ESLO1 et ESLO2 et fournissent des données recueillies suivant le même protocole. Cette spécificité du corpus ouvre la voie à des analyses micro-diachroniques de phénomènes langagiers et permet l'observation de l'évolution du français oral. C'est par exemple dans cette perspective qu'Abouda et Skrovec (2017) ont observé l'évolution des usages des différentes formes d'expression du futur (futur simple et futur périphrastique) entre ESLO1 et ESLO2. Outre la diversité des situations d'interaction, celle des profils des locuteurs enregistrés favorise les études sociolinguistiques. Si les membres du projet ESLO ont veillé à faire varier les locuteurs au niveau du genre, de l'âge, de la zone d'habitation ou encore du niveau d'étude et de la catégorie socioprofessionnelle, une des grandes forces du projet réside dans le fait que l'ensemble de ces métadonnées ont été renseignées de façon exhaustive et quasi-systématique. Ainsi, il est par exemple non seulement possible d'observer l'évolution globale d'un phénomène oral comme la (non) production du schwa sur un plan micro-diachronique en prenant en compte l'ensemble des locuteurs du corpus mais également de constituer des sous-corpus en fonction des profils de locuteurs afin d'analyser l'évolution de l'usage du schwa au sein d'une classe d'âge ou d'une catégorie socioprofessionnelle particulière (Liégeois *et al.*, 2018).

10 La richesse et le volume des (méta)données du projet ESLO nous ont amenés à le sélectionner afin d'étudier la production des interrogatives partielles en français. Plusieurs structures phrastiques sont disponibles pour les francophones (Coveney, 2011 ; Delaveau, 2021, pour un panorama) pour construire ces interrogatives, en particulier celles illustrées en (1a-d) :

(1) a. Tu pars quand ? → in situ (IS)
b. Quand tu pars ? → antéposition (« fronting », F)
c. Quand pars-tu ? → antéposition + inversion (FINV)
d. Quand est-ce que tu pars ? → antéposition + est-ce que (FESK)

11 Ces variantes ont déjà été comparées, par exemple sous l'angle d'une influence de la situation informationnelle (Boeckx, 1999 ; Beyssade, 2006 ; Déprez *et al.,* 2013), de la phonotaxe (Hamlaoui, 2009) ou encore, notamment dans les études sur l'acquisition de ces structures, en termes de complexité structurelle (Jakubowicz, 2011). Ces travaux n'évitent cependant pas le prisme de l'équivalence sémantique entre les différentes variantes, vues comme un même contenu propositionnel, partiellement spécifié mais dont la variable à spécifier est unique et identique à travers toutes les formes. Dans les exemples (1a-d), le but est ainsi toujours de recueillir l'information temporelle spécifiant le moment de l'arrivée du ou de la destinataire.

12 En parallèle, d'autres études expérimentales (Thiberge, 2018) ont montré un meilleur jugement porté par les francophones sur les interrogatives FINV et la projection d'indices sociaux sur les utilisateurs de ce type de phrase (richesse, éducation, lecture fréquente...) par contraste avec les types IS et F. L'objectif de notre analyse du corpus ESLO, décrite en détail dans Thiberge, Badin et Liégeois (soumis), était d'enrichir la compréhension des facteurs sociolinguistiques pouvant influer sur ces préférences dégagées expérimentalement, par une analyse de production spontanée. Nous avons analysé cette variation sous un angle triple :

• micro-diachronique : y a-t-il un eu un changement dans les utilisations des différentes variantes disponibles entre la première période de recueil (ESLO1, 1968-1974) et la seconde (ESLO2, 2008-) ?

• diastratique : y a-t-il des différences observables entre les francophones d'une même époque selon leur tranche d'âge, comme déjà observé dans Thiberge (2018) ? L'âge étant une donnée démographique qu'on peut en ce sens rapprocher d'autres données sociales (Gadet, 1996), nous avons ici comparé les productions des tranches d'âge 15-25 ans et 35-55 ans.

• diaphasique : y a-t-il des différences observables selon les contextes interactionnels ? Nous avons ici comparé les productions en interviews, à l'école et lors de repas.

13 Si le corpus ESLO, de par sa méthodologie de constitution et la richesse de ses métadonnées associées, apparaît comme la source de données la plus appropriée pour mener à bien cette analyse de la variation dans des perspectives micro-diachronique, diastratique et diaphasique, les possibilités actuellement disponibles pour interroger les données semblent très limitées. Avant ce travail il n'existait en effet, à notre connaissance, qu'une plateforme d'interrogation en ligne du corpus élaborée par les membres du projet ESLO (http://eslo.huma-num.fr/index.php/pagecorpus/pageaccescorpus). Cette plateforme permet une interrogation des données à 4 niveaux :

• du fichier son : un formulaire de requête permet de sélectionner des enregistrements précis en fonction du titre du fichier, de la date de recueil ou de la qualité sonore.

• le locuteur : un formulaire permet d'obtenir toutes les données répondant à un locuteur en particulier (par saisie de sa « référence ») ou à un ensemble de locuteurs en fonction de leurs sexe, tranche d'âge, catégorie professionnelle ou niveau d'étude.

• du fichier de transcription : un formulaire de requête permet de sélectionner des transcriptions en fonction de leur titre ou encore de la personne qui s'est chargée de la transcription.

• du texte transcrit : le formulaire disponible pour cette section permet d'effectuer des requêtes au niveau d'une forme (une forme précise ou en fonction du début ou de la fin d'une forme) ou d'une série de formes, appelée « motif ». Il est ainsi possible, par

exemple, de rechercher toutes les constructions débutant par « est-ce que » ou se terminant par « quoi ».

14 Cette plateforme affiche certaines qualités comme la possibilité de pouvoir croiser ces quatre niveaux d'interrogation. Cependant, elle s'avère limitée sur plusieurs points. Nous les avons relevés dans le cahier des charges qui nous a amené à construire une version TXM du corpus ESLO décrite dans la section suivante.

3. Cahier des charges pour une version TXM d'ESLO

15 L'étude que nous souhaitions effectuer à partir des données du corpus ESLO nous a amené à mettre en place une chaîne de traitement permettant d'obtenir une version du corpus utilisable avec TXM et qui réponde à plusieurs critères : visualisation, partitionnement et interrogation du corpus.

3.1. Visualisation

16 L'objectif est de fournir à l'utilisateur une visualisation du texte facilitant la lecture continue de la transcription. TXM étant originellement conçu pour traiter des corpus écrits, le principal souci de la visualisation de données issues de l'oral concerne la définition de l'unité de segmentation à utiliser pour le découpage de la transcription (tour de parole ou énoncé) et l'insertion d'un indice visuel permettant de voir quel locuteur produit un énoncé. Il conviendra alors de trouver une solution permettant que chaque mention du code locuteur ne soit pas comptabilisée par l'outil comme un token appartenant à la transcription.

3.2. Partitionnement et sous-corpus

17 L'un des principaux atouts de l'outil TXM est la possibilité de partitionner un corpus ou de le diviser en sous-corpus en s'appuyant sur les métadonnées. Pour pouvoir profiter au maximum cette fonctionnalité, la version TXM du corpus ESLO devra donc contenir un maximum de métadonnées. Cette ambition pose un double défi. Tout d'abord, il conviendra de définir une unité de segmentation adéquate des transcriptions du corpus ESLO. En effet, un tour peut englober les productions de deux locuteurs différents si celles-ci se chevauchent temporellement. Les productions simultanées de deux locuteurs se trouvent ainsi intégrées à la même unité de segmentation. Cette méthode ne peut pas être conservée car elle empêcherait un partitionnement en fonction des métadonnées locuteurs (identifiant, âge, catégorie socioprofessionnelle etc.). Il conviendra de regrouper ensemble les productions d'un seul et même locuteur. Le deuxième défi technique consistera à intégrer au niveau de cette unité de segmentation des métadonnées stockées indépendamment des transcriptions.

3.3. Interrogation

18 Dans sa forme de départ, le corpus ESLO ne permet qu'une interrogation des données au niveau de la forme orthographique transcrite, c'est-à-dire du token. Pour mener à bien nos analyses, une segmentation en tokens et un étiquetage minimal au niveau du lemme et de la catégorie morphosyntaxique est indispensable. L'outil TXM proposant

une tokenisation et un étiquetage de ce type via TreeTagger (Schmid, 1995) au moment de l'import des données, deux solutions s'offrent à nous : nous appuyer sur cette fonctionnalité de TXM ou effectuer ces opérations en amont avec TreeTagger ou un autre outil.

4. Chaine de traitement de Transcriber à TXM en passant par TEICORPO

19 TEICORPO désigne un ensemble d'outils utilisant le format TEI comme pivot pour :

• la conversion d'un format vers le format pivot TEI ;
• l'ajout de métadonnées (enregistrements et locuteurs), réalisé à l'intérieur du format TEI ;
• la conversion du format pivot TEI vers des formats logiciels (XML-TXM par exemple).

20 Les sections suivantes présentent la façon dont nous avons appliqué cette chaîne de traitement au corpus ESLO pour son utilisation avec l'outil de textométrie TXM.

4.1. Première étape : de Transcriber à la TEI

21 Le format Transcriber (.trs) est un format XML. Il comporte peu d'information de métadonnées générales. Dans le cas du corpus ESLO, quasiment aucune métadonnée autre que le nom des locuteurs n'est renseignée dans les fichiers. Il faut donc se fier au nom du fichier pour savoir de quel enregistrement il s'agit.

Exemple 1. Format Transcriber original

```
<Section type="report" topic="to6" startTime="85.552" endTime="171.323">
<Turn speaker="spk1" startTime="85.552" endTime="89.302">
<Sync time="85.552"/>
est-ce que vous pouvez me décrire une journée de travail euh
<Sync time="88.392"/>
</Turn>
<Turn speaker="spk1 spk2" startTime="89.302" endTime="92.601">
<Sync time="89.302"/>
<Who nb="1"/>
euh normale enfin typique si vous voulez euh oui
<Who nb="2"/>
euh
<Event desc="rire" type="noise" extent="instantaneous"/>
euh oui c'est ça eh bien voilà
</Turn>
<Turn speaker="spk2" startTime="92.601" endTime="95.077">
<Sync time="92.601"/>
euh j'ai un aut- euh j'ai un autre fils qui va au lycée
</Turn>
```

22 Les transcriptions Transcriber sont divisées en sections et en tours de parole. Les sections pouvant comporter le thème de plusieurs tours de parole sont représentées dans le format TEI par des éléments que nous n'exploitons pas dans ce travail. Le format de Transcriber est basé sur des tours de parole (élément <Turn>) découpés en un ou plusieurs segments ou énoncés (<Sync>) et modélise les chevauchements de parole

en regroupant les productions des locuteurs (attribut *speaker*) dans un même tour (cf. exemple 1).

23 Les limites imposées par le système de chevauchement ne permettent pas de maintenir une limite de tour correcte. Ainsi dans exemple 1, le segment « euh normale enfin typique si vous voulez euh oui » du locuteur « spk1 » devrait appartenir au tour de parole précédent.

24 Pour ces raisons, la division en segments du corpus ESLO n'est pas parfaite. Nous ne chercherons pas à modifier ici les caractéristiques des alignements temporels. Cette opération pourrait en revanche être envisagée à une autre étape en effectuant des traitements spécifiques à l'intérieur de la TEI.

25 La conversion vers la TEI respecte les indications de la norme TEI/ISO (ISO 24624:2016). Cette conversion se réalise à l'aide du programme java disponible sur la page de TEICORPO (http://ct3.ortolang.fr/teicorpo/teicorpo.jar).

26 Un fichier TEI peut être créé à partir d'un fichier Transcriber en effectuant la commande suivante (le résultat obtenu est visible en Exemple 2) :

java -cp teicorpo.jar fr.ortolang.teicorpo.TeiCorpo nom_du_fichier_transcriber

Exemple 2. Format TEI de transcriptions des données

```
<annotationBlock end="#T64" start="#T61" who="spk1" xml:id="au47">
<u>
<seg>est-ce que vous pouvez me décrire une journée de travail euh</seg>
<anchor synch="#T65"/>
<seg/>
</u>
</annotationBlock>
<annotationBlock end="#T66" start="#T64" who="spk1" xml:id="au48">
<u>
<seg>euh normale enfin typique si vous voulez euh oui </seg>
</u>
</annotationBlock>
<annotationBlock end="#T66" start="#T64" who="spk2" xml:id="au49">
<u>
<seg>euh<incident subtype="instantaneous" type="noise"><desc>rire</desc>
</incident>euh oui c'est ça eh bien voilà</seg>
</u>
</annotationBlock>
<annotationBlock end="#T67" start="#T66" who="spk2" xml:id="au50">
<u>
<seg>euh j'ai un aut- euh j'ai un autre fils qui va au lycée</seg>
</u>
```

27 Nous pouvons remarquer que, dans le fichier de sortie, la notion de tour de parole n'est pas conservée et que le découpage se fait sur la base des segments pour respecter les indications du format TEI. Les événements du format Transcriber sont conservés. Aucun nettoyage de données n'est réalisé lors de cette conversion, dans le but de garantir une conversion inverse.

4.2. Deuxième étape : ajout des métadonnées

28 Dans le cadre d'ESLO, les métadonnées sont éditées et sauvegardées au format tableur. Les métadonnées sont de deux types : celles qui concernent tout un enregistrement et celles qui concernent des locuteurs. Le procédé d'importation automatique est légèrement différent dans les deux cas.

4.2.1. Ajout de métadonnées « Enregistrement »

29 Les métadonnées « enregistrement » contiennent des informations de date, de lieu, de conditions d'enregistrement, de droits d'utilisation, etc. On dispose d'informations qui sont associées à un nom de fichier TEI et à une étiquette qui les caractérise. Toutes les informations de métadonnées sont entrées et éditées à l'aide d'un tableur et sauvegardées dans un format CSV (virgule comme séparateur). L'exemple 3 illustre les premières lignes d'un tel fichier. On note par exemple que pour le fichier ESLO1_ENT_001, la valeur « excellente » est associée à la métadonnée « acoustique » (cf. Exemple 3).

Exemple 3. Fichier CSV des métadonnées Enregistrement (seules les 4 premières colonnes sont présentées)

id	acoustique	type	categorie	date
id	recording/media/desc[@type='quality']	recordingStmt/recording/media/@type	profileDesc/textDesc/domain/@nature	profileDesc/settingDesc/setting/date/@when-iso
ESLO1_ENT_001_C.tei_corpo.xml	Excellente	audio	Entretien	1969-04-01
ESLO1_ENT_002_C.tei_corpo.xml	Bonne	audio	Entretien	1969-04-12

30 La première ligne indique le nom des métadonnées décrites dans la colonne correspondante. Ce nom est un aide-mémoire qui n'est pas utilisé dans l'import de métadonnées. La deuxième ligne du fichier CSV contient un chemin au format XPATH (https://www.w3.org/TR/1999/REC-xpath-19991116/) qui est utilisé par TEICORPO pour insérer automatiquement des informations dans les fichiers TEI. Si les nœuds de destination n'existent pas, ils sont alors créés par le programme.

Exemple 4. Ajout d'information dans la partie « recording »

```
<recording>
<media mimeType="audio/x-wav" type="audio" url="ESLO1_ENT_012_C.wav">
<desc type="quality">Excellente</desc>
</media>
<date dur="4718.501">101129</date>
</recording>
```

31 Pour produire un tel fichier, la commande est la suivante :

java -cp "teicorpo.jar:lib/*" fr.ortolang.teicorpo.TeiInsertCsv *metadonnees_enregistrement.csv* -o repertoire_optionel_des_resultats

32 Si l'option « -o » n'est pas utilisée, les fichiers sont renommés avec une extension supplémentaire. Ce programme utilise des librairies java supplémentaires qui doivent être téléchargées depuis la page http://ct3.ortolang.fr/teicorpo et déposées dans un dossier « lib ».

4.2.2. Ajout des métadonnées « Locuteurs »

33 Les métadonnées « locuteurs » fonctionnent sur un principe différent. Au lieu d'être renseignées pour chaque enregistrement, elles le sont pour chaque locuteur, certains locuteurs figurant dans plusieurs enregistrements. Le format de structuration reste identique : il s'agit d'un fichier tableur (cf. exemple 5).

Exemple 5. Fichier CSV des métadonnées Locuteurs (seules les premières colonnes sont présentées)

idlocuteur	reference	niveau-etude	situation	profession-insee	annee-naiss	tranche-age
none	none	education	state	occupation	birth/date	age
	1 BA725	CEP	Marié	Artisans, commerçants et chefs d'entreprise	1912	55/65
	2 UG393	Secondaire Incomplet	Célibataire	Professions intermédiaires	1897	+ de 65

34 Dans ce fichier la première colonne correspond à l'identifiant numérique du locuteur dans la base de données, et la deuxième à l'identifiant utilisé dans les transcriptions. Le reste des informations suit le même modèle que précédemment. Ainsi, on retrouve une information XPATH qui indique où insérer les valeurs des métadonnées qui, au lieu d'être calculées par rapport à la racine du fichier, le sont par rapport au nœud « *person* » du schéma TEI.

35 La commande à lancer est la suivante :

```
java -cp "teicorpo.jar:lib/*" fr.ortolang.teicorpo.TeiInsertCsv metadonnees_locuteur.csv -o repertoire_optionel_des_resultats -userinfo nom_du_fichier_TEI
```

Exemple 6. Information insérée dans les métadonnées TEI pour un locuteur

```
<person>
<persName>LD386</persName>
<education>Bac + 5 et plus</education>
<state>Marié</state>
<birth>
<date>1915</date>
<placeName>Corbie (Somme)</placeName>
</birth>
</person>
```

4.3. Troisième étape : export vers TXM

36 La conversion vers TXM transforme les informations de la TEI pour permettre l'exploitation du moteur d'interrogation CQP (Corpus Query Processor). En effet, bien que TXM soit capable de manière native d'indexer des fichiers TEI, seuls les éléments de l'arborescence XML des champs indexés peuvent être interrogés. Il convient donc de préparer les données en intégrant les métadonnées dans les éléments , <u> et/ou <w> de la TEI. Cette pratique est contraire aux recommandations de la TEI puisque les métadonnées sont normalement regroupées en en-tête du document XML (élément

<teiHeader>) et donc séparées des données situées dans un élément <text> (qui regroupent les éléments , <u>, <w>).

37 La conversion réalisée par TEICORPO consiste donc à créer une version « simplifiée » du fichier TEI contenant des éléments <u> (les segments) et éventuellement des éléments . À l'intérieur de ces éléments sont ensuite projetées, sous forme d'attributs XML, toutes les informations métadonnées.

38 Pour cela, on utilise un paramètre de la commande de conversion : « -mv » (pour métadonnée et valeur) :

• pour indiquer des métadonnées indépendantes du locuteur : -mv champ_txm:chemin_xpath
• pour indiquer des métadonnées dépendantes du locuteur : -mv champ_txm:chemin_xpath:nom_du_locuteur

39 Pour un traitement multiple, le nom du locuteur peut être remplacé par « * ». La commande est alors de ce type :

```
java -cp teicorpo.jar fr.ortolang.teicorpo.TeiToTxm nom_de_fichier_TEI -mv domain:domain/@nature ... -mv educ:education:* -utt -spk pers -rawline
```

40 Avec :

• -utt : pour obtenir une sortie segmentée en énoncés (par défaut en mots).
• -spk pers : indique qu'on veut utiliser les champs « persName » (nom du locuteur) et non « alt » (code du locuteur).
• -rawline : pour ne convertir que le texte sans les codes spéciaux de l'oral (silences, pauses, etc.). Un exemple de résultat obtenu est présenté dans l'exemple 7 ci-dessous.

Exemple 7. Format XML importable par TXM

```
<p>
  <meta>
    <br/>
    [BA725]
  </meta>
  <u AM="D" age="55/65" birth="1912" birthloc="loiret" date="1969-04-01" educ="CEP" end="1.317" lang
  ="Français" loc="loiret" nature="Entretien" occ="Artisans, commerçants et chefs d'entreprise" prof=
  "boucher, gérant boucherie supermarché" qualite="Excellente" sexe="homme" start="0.449" statut=
  "Marié" type="audio" who="BA725">vous savez euh</u>
</p>
```

4.4. Quatrième étape : intégration dans TXM

41 L'intégration dans TXM se fait via l'import XML-TEI Zero + CSV. Il n'y a plus besoin d'indiquer de CSV comme fichier de métadonnées puisque les informations ont déjà été intégrées aux données. Nous sélectionnons en revanche l'option « Annoter les données » puisque l'étiquetage morphosyntaxique n'a pas été réalisé via TEICORPO au moment de la conversion.

42 Nous proposons une version du corpus en projetant notamment les indications concernant le nom du locuteur, les éléments paraverbaux (comme les rires par exemple) et l'empan temporel (optionnel) dans un élément XML spécifique de type <meta>. En procédant ainsi, ces informations particulièrement pertinentes pour l'interprétation des données et la lecture de la transcription ne sont pas indexées par l'outil (paramètre de l'import : plans textuels > hors texte à éditer = meta) mais visibles dans la partie d'édition du logiciel. Les textes sont ainsi visualisés par TXM sous la forme présentée dans l'exemple 8. Grâce à l'utilisation des balises <p>, chaque segment

est affiché dans une ligne distincte. L'absence de cette balise entraînerait un affichage au kilomètre, plus compact mais beaucoup moins lisible.

Exemple 8. Présentation de la visualisation des textes sous TXM

[OU] est-ce que vous pouvez me dire euh en quoi ça consistait le tour de France ?
[BA725]
[BA725] eh bien
[BA725] euh
[OU] *[=!pi /instantaneous/PHO]*
[BA725] à l'époque
[BA725]
[BA725] vous comprenez on cherchait à s'améliorer

43 L'avantage de ce format est qu'il permet l'écoute des fichiers média (voir extension MediaPlayer dans https://groupes.renater.fr/wiki/txm-users/public/extensions_beta) grâce aux attributs « start » et « end » de la balise <u>. Cette écoute peut se faire dans le concordancier (voir Exemple 9) au moyen d'un clic droit qui proposera l'option « jouer le média ».

Exemple 9. Visualisation de texte oral sous TXM

44 L'autre avantage du choix de ce format est qu'il permet d'utiliser la fonctionnalité d'annotation, non disponible au moment de la réalisation de notre étude via le module d'import de fichiers Transcriber proposé par défaut.

5. Exploitation pour une analyse fine des interrogatives partielles en français

45 Grâce au travail effectué en amont, nous avons extrait du double corpus ESLO (1&2) les interrogatives partielles produites par des francophones de deux tranches d'âge : 15-25 ans et 35-55 ans, afin de comparer les emplois de phrases comme (1a-1d). La liste des mots interrogatifs francophones a été intégrée à une expression régulière dans le

moteur de recherche TXM, puis les extractions obtenues ont été filtrées manuellement pour ne conserver que les interrogatives non-sujet, finies, et racines (N=617). Le choix des variantes observées, le détail des critères d'extraction, des annotations syntaxiques effectuées puis de l'analyse statistique menée sont décrits dans Thiberge, Badin & Liégeois (soumis) ; nous en résumons les principaux résultats pour mieux illustrer l'intérêt de la méthodologie présentée ici.

5.1. Une évolution diachronique

46 En croisant les données des corpus ESLO1 et ESLO2, une évolution importante des pratiques linguistiques peut s'observer. Alors que dans les années 1960 (ESLO1), moins d'un quart des interrogatives partielles sont *in situ* (IS, *Tu pars quand ?*), leur proportion d'emploi atteint près de 60% dans les années 2000 (ESLO2), toutes personnes et contextes confondus. En parallèle, la part d'interrogatives antéposées de manière générale, mais en particulier d'antéposées avec inversion verbe-sujet (FINV, *Quand pars-tu ?*) baisse drastiquement, passant de 22% d'emploi à 8% environ.

5.2. Variation diastratique

47 Cette observation est nuancée par un facteur d'âge, déjà observé lors d'expériences (Thiberge, 2018). Et c'est ici que tout le travail effectué pour projeter les métadonnées locuteurs sur les extractions du corpus ESLO prend son véritable sens.

48 La différence d'emploi entre IS et antépositions n'est en surface pas très importante entre groupes générationnels : 42% d'*in situ* et 15% de FINV chez les 35-55 ans contre 51% d'*in situ* et 9.5% de FINV chez les 15-25 ans, tous corpus et contextes confondus. Si l'on regarde en détail en contrastant à la fois les corpus et les groupes d'âge, cependant, les différences générationnelles se font plus fortes, différemment pour chaque époque.

49 Ainsi, dans les années 1960 (ESLO1), alors que les deux groupes d'âge emploient la même proportion d'*in situ* (25% environ), les 35-55 ans emploient presque deux fois plus de FINV que les 15-25 ans (25% contre 13%). Dans les années 2000 (ESLO2), alors que les 15-25 ans emploient près de 70% d'*in situ* et 18% d'antéposition simple (F, *Quand tu pars ?*), les 35-55 ans emploient « seulement » 55% d'*in situ* et près de 27% d'antéposition simple.

50 Les phénomènes de variation sociolinguistique sont souvent observés sous un tel prisme sociolectal : à tel groupe de personnes (défini par son âge, sa catégorie socioprofessionnelle, son éducation, etc.) correspondrait une variante principale. C'est la perspective adoptée dans la plupart des travaux menés en sociolinguistique sur l'interrogation partielle en français jusqu'ici (Pohl, 1965 ; Terry, 1970 ; Behnstedt, 1973 ; Ashby, 1977 ; Söll, 1982 ; Coveney, 1996). Cette observation doit cependant elle aussi être nuancée.

5.3. Variation diaphasique

51 La variation peut en effet aussi être vue comme un outil dont les locuteurs et locutrices se servent, plus ou moins consciemment, pour influer sur le masque social (*persona*) que les personnes à qui ils ou elles s'adressent construisent à leur égard. Cette conception est à l'origine d'un renouvellement de la littérature sociolinguistique depuis la fin du

XX^e siècle (Eckert, 2012, pour une description de ces changements). Dans la lignée de ce cadre théorique, les stratégies interactionnelles étant nécessairement différentes selon le contexte d'interaction (influence du public, du niveau de formalité jugé approprié/nécessaire), nous avons contrasté les trois contextes principaux où se trouvaient les interrogatives partielles extraites (interviews, N=336 ; école, N=130 ; repas, N=94).

52 Un net contraste apparaît alors, tous groupes d'âge confondus, avec par exemple 72% d'*in situ* lors de repas, contre 50% à l'école et à peine 31% en interviews. En contraste, seulement 3.2% d'antépositions avec inversion sont observées lors des repas, contre près de 15% à l'école, et 18% en interviews. Ceci s'explique sans doute par les contraintes sociales pesant sur ces différents contextes : les repas sont des situations informelles entre proches alors que les interviews, même si elles sont menées à la maison, sont un type d'interaction très particulier où un ou une scientifique extérieur.e au cercle social familier pose des questions à l'interrogé.e sur son histoire et ses habitudes de vie. Les contextes d'école quant à eux semblent un milieu presque intermédiaire, où un adulte s'adresse à des enfants dans un cadre semi-formel (classes d'enfants, mais milieu institutionnel). Il est à noter aussi une forte proportion d'antéposition en « est-ce que » (FESK, *Quand est-ce que tu pars ?*) en interviews (18%) par rapport aux deux autres contextes (5% à l'école, 6% en repas).

5.4. À la croisée des phénomènes diastratiques et diaphasiques

53 Au vu de la répartition des extractions d'interrogatives partielles effectuées, il ne faisait sens, statistiquement, que de croiser les interviews, les tranches d'âge, et les contextes interactionnels au sein d'ESLO2 même. Ce faisant, il apparaît que les stratégies d'interrogation partielle sont bien la conséquence de plusieurs facteurs interagissant entre eux, plutôt que d'un seul dominant. Ainsi, en repas, les 15-25 ans et les 35-55 ans des années 2000 ont un comportement linguistique similaire (environ 70% d'*in situ* contre 30% d'antépositions diverses, avec quelques nuances dans les types d'antépositions favorisés). En contexte d'interviews cependant, le tableau est différent : les 35-55 ans utilisent 46% d'*in situ*, alors que les 15-25 ans en utilisent toujours plus de 58%. Une vision qui associerait les interrogatives partielles *in situ* à un contexte informel (repas), par exemple par contraste avec les interrogatives FINV qui seraient plutôt associées à un contexte formel ou semi-formel, doit donc être nuancée, sinon écartée. Ou alors, il faut lui associer l'idée que la perception de ce qu'est un milieu formel et un milieu informel varie selon le groupe d'âge (ou groupe défini par un autre critère social), ou encore que la variante jugée adaptée à un milieu formel/informel ne sera pas la même selon ce groupe.

54 En parallèle des considérations plus directement syntaxiques ou sémantiques, les données sociolinguistiques explorées grâce au corpus ESLO permettent ainsi d'approfondir les travaux expérimentaux qui dégagent un « sens social » différent aux différentes variantes disponibles (Thiberge, 2018 ; voir aussi Thiberge & Hemforth, 2019, pour une prolongation). Une personne employant une variante plutôt qu'une autre, dans un contexte plutôt que dans un autre, véhiculera des informations sociales différentes vis-à-vis d'elle-même.

6. Conclusion et perspectives

55 L'objectif principal de notre travail était de proposer une méthode de compilation du corpus ESLO dans le but de le rendre exploitable avec l'outil de textométrie TXM et, à partir des données obtenues, de présenter une brève analyse de l'usage des énoncés interrogatifs par les locuteurs du corpus. Notre étude s'est donc focalisée sur la problématique de l'interopérabilité des données et rend compte des possibilités offertes en la matière par l'outil TEICORPO. Si, avant cette étude, l'outil proposait déjà une série d'opérations de conversion permettant d'exploiter et d'explorer un même corpus au moyen d'une suite d'outils variés, les développements effectués dans le cadre de ce travail permettent aujourd'hui de fournir à la communauté une version du corpus ESLO totalement inédite et particulièrement efficace pour mener à bien des travaux dans des perspectives sociolinguistiques et micro-diachroniques.

56 Exploitant toutes les potentialités de l'outil TXM, la version du corpus ESLO que nous proposons aujourd'hui pourrait favoriser la réutilisation de données riches qui, selon nous, n'ont que trop peu été explorées par les chercheurs du domaine. Ainsi, nous fournissons une version « clé en main » du corpus ESLO permettant d'effectuer des requêtes multi-niveaux (forme, lemme, catégorie morphosyntaxique) paramétrables en fonction de la richesse des métadonnées préalablement disponibles au niveau des enregistrements (époque de recueil et situation d'interaction par exemple) et des locuteurs (âge, catégorie socioprofessionnelle et niveau d'étude par exemple). L'analyse des énoncés interrogatifs que nous avons présentée révèle non seulement la pertinence de l'approche méthodologique mise en place mais illustre également que le corpus ESLO peut fournir des données particulièrement pertinentes pour l'étude du langage oral, et ce quel que soit l'axe d'analyse privilégié (sémantique, syntaxique, sociolinguistique...).

57 Grâce aux développements continuels de l'équipe de TXM et aux retours utilisateurs, des améliorations internes à TXM sont prévues pour faciliter l'interrogation des corpus oraux. En effet, l'écoute pourra se faire également au niveau de l'édition et non pas uniquement en passant par les résultats d'une recherche de concordances, ce qui permettra par exemple une écoute de l'ensemble du contexte entourant l'élément recherché. Il ne sera alors plus nécessaire d'utiliser le logiciel VLC, la lecture du signal sonore étant intégrée à l'outil. Enfin, les fichiers sonores pourront être lus depuis un serveur et n'auront plus nécessairement à être stockés sur un poste de travail ce qui, pour un corpus particulièrement volumineux comme ESLO, est un avantage certain.

BIBLIOGRAPHIE

Abouda L. & Skrovec M. (2017). « Du rapport micro-diachronique futur simple / futur périphrastique en français moderne. Étude des variables temporelles et aspectuelles ». *Corela*, *HS-21*, 1-25.

Ashby W. (1977). « Interrogative forms in Parisian French ». *Semania* 4 : 35-52.

Barras C., Geo E. & Wu Z. (2001). « Transcriber : Development and Use of a Tool for Assisting Speech Corpora Production ». *Speech Communication*, 33 : 5-22.

Behnstedt P. (1973). *Viens-tu ? Est-ce que tu viens ? Tu viens ? Formen und Strukturen des direkten Fragesatzes im Französischen.* Tübingen, Narr.

Beyssade C. (2006). « La structure de l'information dans les questions : quelques remarques sur la diversité des formes interrogatives en français ». *Linx, Revue des linguistes de l'université Paris X Nanterre,* 55 : 173-193.

Boeckx C. (1999). « Decomposing french questions ». *University of Pennsylvania Working Papers in Linguistics*, 6(1) : 6.

Boersma P. & Weenink D. (2019). « Praat : doing phonetics by computer ». **[Computer program].** Version 6.0.37, retrieved 14 october 2019 from http://www.praat.org/.

Coveney A. (1996). *Variability in spoken french : a sociolinguistic study of interrogation and negation.* Exeter, Elm Bank Publication.

Coveney A. (2011). « L'interrogation directe ». *Travaux de linguistique 2011/2*, 63 : 112-145.

Déprez V., Syrette K. & Kawahara S. (2013). « The interaction of syntax, prosody, and discourse in licensing french wh-in-situ questions », *Lingua*, 124 : 4-19.

Delaveau A. (2021). « Les phrases interrogatives ». In Abeillé A. et Godard D. (éd.), *La grande grammaire du français*, Actes sud.

Eckert P. (2012). « Three waves of variation study : The emergence of meaning in the study of sociolinguistic variation ». *Annual review of Anthropology* 41 : 87-100.

Eshkol-Taravella I., Baude O., Maurel D., Hriba L., Dugua C. & Tellier I. (2012). « Un grand corpus oral "disponible" : le corpus d'Orléans 1968-2012 ». *TAL*, 52(3) : 17-46.

Gadet F. (1996). « Niveaux de langue et variation intrinsèque ». *Palimpsestes* 10 : 17-40.

Hamlaoui F. (2009). « A prosodic study of wh-questions in french natural discourse », *Proceedings of the LangUE*, 27-38.

Heiden S., Magué J.-P. & Pincemin B. (2010). « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement ». In S. Bolasco, I. Chiari, & L. Giuliano (éd.), *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010,* 2 : 1021-1032. Rome, Italy : Edizioni Universitarie di Lettere Economia Diritto.

Jakubowicz C. (2011). « Measuring derivational complexity : New evidence from typically developing and SLI learners of L1 French ». *Lingua*, 121(3) : 339-351.

Lamalle C., Martinez W., Fleury S., Salem A., Fracchiolla B., Kuncova A. & Maisondieu A. (2003). Lexico 3 version 3.41. Outils de statistique textuelle. Manuel d'utilisation. Laboratoire SYLED-CLA2T, Université de la Sorbonne nouvelle - Paris 3.

Liégeois L., Skrovec M., Abouda L. & Belhoum S. (2018). « Usage du schwa au sein des constructions de type *je vais* : une marque d'un processus de grammaticalisation du futur périphrastique ? », In Colloque de la Société Internationale de Diachronie du Français, Neuchâtel.

MacWhinney B. (2000). The CHILDES Project : Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ : Lawrence Erlbaum Associates.

Parisse C. & Le Normand M.-T. (2006). « Une méthode pour évaluer la production du langage spontané chez l'enfant de 2 à 4 ans ». *Glossa*, 97 : 20-41.

Parisse C., Etienne C. & Liégeois L. (sous presse). « TEICORPO : A Conversion Tool for Spoken Language Transcription with a Pivot File in TEI ». *Journal of the Text Encoding Initiative*.

Pohl J. (1965). « Observations sur les formes d'interrogation dans la langue parlée et dans la langue écrite non littéraire ». *Actes du Xe Congrès International de Linguistique et de Philologie Romanes*, Tome 2, Paris, Klincksieck, 501-513.

Ratinaud P. (2009). « Iramuteq : Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires ». Téléchargeable à l'adresse : http://www. iramuteq. org.

Schmid H. (1995). *Improvements in Part-of-Speech Tagging with an Application to German. In Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

Söll L. (1982). « L'interrogation directe dans un corpus en langage enfantin ». In F.-J. Haussman (éd.), *Études de grammaire française descriptive.* Heideberg, Groos.

TEI/ISO (2016). ISO 24624:2016 - Language resource management - Transcription of spoken language. Retrieved from https://www.iso.org/obp/ui/#iso:std:iso:24624:ed-1:v1:en.

Terry R. M. (1970). *Contemporary French interrogative structures.* Montréal et Sherbrooke, Éd. Cosmos.

Thiberge G. (2018). « Position du syntagme Wh- en français : réelle optionnalité ou biais sociolinguistique ? », *ELIS, Échanges de Linguistique en Sorbonne* 5 : 64-91.

Thiberge G. & Hemforth. B. (2019). « Variation in French Partial Interrogatives : Social meaning as a key factor ». Poster présenté à la *8th Experimental Pragmatics conference (XPrag 2019)*. Edinburgh, Scotland.

Thiberge G., Badin F. & Liégeois L. (soumis). « French partial interrogatives : a microdiachronic study of variation and new perspectives in a refined pragmatics framework ». *Faits de Langue.*

Wittenburg P., Brugman H., Russel A., Klassmann A. & Sloetjes H. (2006). « ELAN : a Professional Framework for Multimodality Research ». In *Proceedings of the Fifth International conference on Language Resources and Evaluation*, 1556-1559.

Zimina M. & Fleury S. (2015). « Perspectives de l'architecture Trame/Cadre pour les alignements multilingues ». *Nouvelles Perspectives En Sciences Sociales*, 11(1) : 325-353. https://doi.org/https://doi.org/10.7202/1035940ar.

RÉSUMÉS

Pour répondre aux problématiques engendrées par la diffusion de plus en plus massive des corpus linguistiques et à l'hétérogénéité de leurs formats, nous proposons une méthode permettant de prendre en main des corpus langagiers oraux et de les convertir dans un format permettant leur exploitation outillée. Pour cette recherche, le corpus ESLO nous sert d'exemple par sa licence de diffusion, son format, son volume et ses atouts sociolinguistiques et diachroniques. Notre travail se fonde sur la compilation de ce corpus pour le rendre compatible avec l'outil de textométrie TXM. Nous opérons un ensemble de transformations des données pour l'utiliser au mieux. Enfin, pour illustrer les apports de ces avancées méthodologiques, nous proposons une analyse fine et multidimensionnelle de l'usage des interrogatives dans le corpus ESLO.

To answer the increasing trend of corpora sharing and data format heterogeneity, we present a method for converting spoken language corpora to several tool formats in order to facilitate linguistic analysis. For this research, we take as an example the ESLO corpus for several reasons: its open-source licence, its standard format used for its construction, its size, and its sociolinguistic and micro-diacronic characteristics. Our study is based on a compilation of the ESLO corpus in order to make it compatible with the textometric tool TXM. We operate a set of operations to use all the possibilities the tool offers. Finally, we present a fine-grained and multidimensional analysis of the interrogatives utterances used in the ESLO corpus.

INDEX

Mots-clés : corpus oraux, interopérabilité, textométrie, XML, interrogatives
Keywords : spoken language corpora, interoperability, text analysis, XML, interrogatives

AUTEURS

FLORA BADIN

LLL, UMR 7270, Université d'Orléans, COMUE Centre-Val de Loire,
CNRS, F- 45065 Orléans, France

LOÏC LIÉGEOIS

CLILLAC-ARP, EA 3967, Université de Paris, Sorbonne Paris
Cité, F-75013 Paris, France
LLF, UMR 7110, Université de Paris, Sorbonne Paris Cité,
CNRS, F-75013 Paris, France

GABRIEL THIBERGE

LLF, UMR 7110, Université de Paris, Sorbonne Paris Cité,
CNRS, F-75013 Paris, France

CHRISTOPHE PARISSE

MODYCO, INSERM, CNRS/Université Paris Nanterre, F 92000,
Nanterre, France