Éléments de catalogage
Des corpus numériques à l'analyse linguistique en langues de spécialité. Sous la direction de Cécile Frérot et Mojca Pecman. 374 p. : couv. ill. en coul. ; 21,5 cm. Collection « Langues, Gestes, Paroles », ISSN 2105-9497 — ISBN 978-2-37747-261-1

Ce travail a bénéficié du soutien du Centre de Linguistique Inter-langues, de Lexicologie, de Linguistique Anglaise et de Corpus-Atelier de Recherche sur la Parole (CLILLAC-ARP) de l'Université de Paris, de l'ILCEA4 et du programme IDEX Université Grenoble Alpes

Université Grenoble Alpes
CS 40700
38058 GRENOBLE CEDEX 9

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

Chapitre 4

Le français oral quotidien, un objectif spécifique en FLE ? Retour sur les défis de la création d'un corpus de français parlé annoté à visée pédagogique

Christian Surcouf
École de Français Langue Étrangère, Faculté des Lettres, Université de Lausanne

Résumé : Apprendre une langue étrangère est un processus long et complexe, au cours duquel la compréhension orale joue un rôle fondamental, *a fortiori* depuis la place qu'a pris l'oral dans les années soixante-dix avec l'approche communicative. Pourtant, bien qu'elle constitue la plus exigeante des compétences langagières, la compréhension orale ne fait pas l'objet de l'attention à la hauteur des défis qu'elle soulève. Même à un niveau avancé, les apprenants de français langue étrangère affichent encore de nombreuses difficultés à comprendre le français tel qu'il est parlé dans les situations quotidiennes (i.e. le plus souvent) qui plus est dans les contextes socioprofessionnels spécifiques où ils sont amenés à utiliser le Français sur Objectifs Spécifiques (FOS). Une grande partie de ces difficultés semble imputable au manque de sensibilisation aux caractéristiques du français parlé. Aussi avons-nous créé FLORALE, une base de données de français parlé à visée pédagogique utilisant des documentaires radiophoniques (par ex. *Les Pieds sur Terre*), que nous avons transcrits, segmentés et annotés pour permettre, dès le niveau A2, d'écouter nombre d'exemples authentiques révélant les fonctionnements effectifs du français oral quotidien. Dans ce chapitre, après avoir rappelé l'importance de la compréhension orale du français parlé en FLE, nous évoquerons les défis informatiques, linguistiques et pédagogiques qu'il nous a fallu relever pour constituer notre base de données et son interface-usager, et la manière dont notre expérience pourrait être étendue à l'enseignement du français parlé dans le contexte du FOS[1].

Mots-clés : annotation, compréhension orale, corpus de français parlé, FLE, interface

1. Je tiens à remercier chaleureusement Mojca Pecman et Cécile Frérot pour leurs suggestions déterminantes concernant la problématique du FOS.

1 Introduction : les enjeux de la compréhension orale

Ce chapitre clôt la première partie de cet ouvrage, consacrée à la constitution, l'annotation et l'exploitation de corpus en vue de la création de ressources pour les études en langues de spécialité. Il s'agira ici d'illustrer les problématiques relatives à l'annotation d'un corpus de français parlé destiné à une exploitation dans le cadre de l'enseignement-apprentissage du Français Langue Étrangère (FLE). À bien des égards, la compréhension du français oral quotidien par les étudiants allophones s'apparente à un « objectif spécifique » de FLE dans la mesure où les problématiques soulevées sont proches des objectifs de compréhension du français de spécialité, également important en classe de FLE et plus particulièrement de Français sur Objectifs Universitaires (FOU). En quoi le français parlé demanderait-il qu'on lui accorde une attention spécifique en FLE ?

Déplorant que « lorsqu'il est question de la langue française, de sa grammaire et de son lexique, c'est en général de la langue écrite qu'il s'agit », Blanche-Benveniste (Blanche-Benveniste, 2003, p. 317) rappelle à juste titre que « c'est pourtant sous sa forme parlée que la langue est le plus largement partagée. Tous les gens bien portants parlent ; mais combien écrivent ? » Les recherches présentées par Worthington et Fitch-Hauser (Worthington & Fitch-Hauser 2018, p. 6, tableau 1-1) portant sur le temps consacré quotidiennement aux quatre activités langagières – parler, écouter, lire, écrire – permettent d'établir qu'écouter occupe en moyenne 50 % du temps éveillé, parler 24 %, lire 14 %, et écrire 12 %[2]. Quelles qu'en soient les proportions exactes, en situation naturelle, écouter constitue le premier accès à la langue et son fonctionnement. En ce sens, la compréhension orale s'avère incontournable dans l'activité langagière, et l'apprentissage du FLE n'échappe pas à cet impératif, *a fortiori* depuis les années 1970 et l'avènement de l'approche communicationnelle dans laquelle « l'oral occupe une place de choix » (Cuq & Gruca, 2002, p. 247). Or, comme le rappelle Porcher (Porcher, 1995, p. 45) « la compétence de réception orale est de loin la plus difficile à acquérir et c'est pourtant la plus indispensable ». Par ailleurs, si en accord avec Vandergrift (Vandergrift, 2007, p. 199), « *the ultimate goal of listening instruction is to help L2 listeners understand the target language in everyday situations* », force est de constater que les

2. Ces calculs ont été effectués sur la base de huit des dix études mentionnées par les auteurs.

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

documents sonores utilisés en FLE ne sont pas toujours en adéquation avec un tel objectif. Parpette (Parpette, 2018, p. 19) relève ainsi que « les discours authentiques oraux sont […] très rares dans les méthodes de FLE avant le niveau B2 […] phénomène […] d'autant plus paradoxal que l'approche communicative accorde […] une place importante à l'oral en début d'apprentissage. » Il semblerait effectivement logique d'entrainer à l'écoute de tels documents pour la simple raison que « leur authenticité accroit la probabilité qu'ils offriront bien à l'apprenant les moyens d'acquérir les savoirs dont il aura besoin pour fonctionner langagièrement en situation "réelle" » (Holec, 1990, p. 68). Or, qu'il s'agisse des enregistrements sonores ou des descriptions apparaissant dans les manuels ou les grammaires de FLE, l'apprenant est la plupart du temps insuffisamment confronté[3] au « français ordinaire », « celui dont chacun est porteur dans son fonctionnement quotidien, dans le minimum de surveillance sociale : la langue de tous les jours » (Gadet, 1996, p. v).

L'enseignement de la compréhension du français oral quotidien pourrait-il constituer un « objectif spécifique » de la didactique du FLE ? Probablement pas dans le sens courant d'« objectif spécifique ». Cependant, si l'on considère avec Cuq et Gruca (Cuq & Gruca, 2002, p. 327) que « le français sur objectifs spécifiques dépend […] de l'analyse des objectifs et des besoins », deux points de vue sont envisageables selon la perspective adoptée sur la notion de *besoins*, appréhendés d'une part comme « les attentes des apprenants (ou "besoin ressenti") » et de l'autre comme « les "besoins objectifs" (mesurés par quelqu'un d'autre que l'apprenant) » (Cuq, 2003, p. 35). Dans le premier sens, le « besoin ressenti » par les apprenants suffirait à lui seul à justifier l'élaboration, par exemple, d'un programme d'enseignement du français sur objectif universitaire dont l'objectif consisterait précisément à répondre à ces « besoins ressentis ». En revanche, du point de vue des « besoins objectifs », il se peut que l'offre didactique, bien qu'elle poursuive un « objectif spécifique », ne résulte plus d'une demande explicite de la part des apprenants mais d'un constat établi par les enseignants.

Or, en ce qui concerne la compréhension orale du français ordinaire, force est de constater que les apprenants affichent d'importantes lacunes, même s'ils ne formulent pas nécessairement d'attentes à cet égard. Les « besoins objectifs » sont réels comme le souligne Wagner avec emphase :

3. Pour une présentation de ces problèmes, voir notamment Giroud et Surcouf (2016), Surcouf et Giroud (2016), Surcouf et Ausoni (2018).

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

> *It is an only too well-known phenomenon – a learner has studied a foreign language for years, diligently attending class; spending hundreds of classroom hours reading, writing, speaking, and listening to the target language; interacting with other learners; studying word lists; and completing workbook after workbook of grammar activities. Yet when that learner finds herself in a "real-world" language context trying to converse with a speaker of the target language, she is unable to comprehend almost anything of what that speaker says.* (Wagner, 2014, p. 288)

Dès lors, sans pour autant ériger la compréhension du français oral quotidien en « objectif spécifique », il semble néanmoins légitime de lui accorder une attention particulière en proposant des enseignements ou des outils spécifiques permettant de sensibiliser les apprenants à ses caractéristiques. Il parait en effet paradoxal que des étudiants même de niveau avancé et séjournant de surcroit en milieu homoglotte éprouvent de telles difficultés alors que dans la vie courante, les échanges se déroulent majoritairement à l'oral, et que cet oral tend précisément à relever du français ordinaire. Quelles sont donc les difficultés que rencontrent les apprenants de FLE confrontés au français oral quotidien ?

2 Une illustration des difficultés de compréhension de l'oral spontané en FLE

Prenons un exemple concret, issu, parmi d'autres, d'un exercice de compréhension orale fine en FLE mené en milieu universitaire homoglotte, à l'université de Lausanne. Dix-neuf étudiants de niveau B2[4] étaient invités à transcrire[5] l'énoncé 1, extrait d'un reportage radiophonique de l'émission *Les Pieds sur Terre* diffusée le 14 février 2017 sur France Culture.

> Énoncé 1) [dõkɛvzesɛ̃kœʁdymatɛ̃øʃepamwakatɔʁzœʁ] (donc elle faisait cinq heures du matin euh je sais pas moi quatorze heures)

Les étudiants disposaient de plusieurs minutes avant l'exercice d'écoute pour prendre connaissance des consignes et des informations suivantes, présentées sous leur forme écrite :

4. Durée moyenne du séjour en milieu francophone : 3 ans (minimum : 3 mois, maximum : 8 ans) ; durée moyenne d'apprentissage du français : 7 ans (minimum : 1 an 8 mois, maximum : 15 ans).
5. Sur l'intérêt de la transcription en FLE, voir les réflexions de Paternostro (2016, p. 133-144).

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

Thème de la conversation : la locutrice parle de sa mère.

Transcription du contexte précédant le segment à transcrire :

moi ma mère elle travaillait beaucoup elle était seule avec quatre enfants elle faisait des petits trucs euh au black euh des ménages euh et puis en même temps elle travaillait à Carrefour

Quatre écoutes successives étaient ensuite proposées, espacées d'une pause conséquente pour permettre aux apprenants de réfléchir et d'écrire leur transcription. Remarquons en premier lieu que, à l'exception de *euh* – absent –, et de *quatorze* (en 578e position), tous les lemmes de cet extrait figurent parmi les 200 premiers mots du *français fondamental* (Gougenheim, Michéa, Rivenc & Sauvageot, 1964, p. 69-89) (ci-dessous « Rang FF »). Par ailleurs, avec ses 15 syllabes articulées en 2,6 secondes, le débit de 5,8 est conforme aux résultats relevés « dans un corpus de français parlé spontané » par Léon (Léon, 1996, p. 104). En somme, par sa banalité, un tel énoncé reflète l'usage ordinaire du français parlé et ne présente *a priori* aucune difficulté majeure. Pourtant, après quatre écoutes, seuls deux des dix-neuf étudiants sont parvenus à le transcrire convenablement[6]. Le Tableau 1 recense le nombre d'étudiants (en gras) qui, parmi les 19, ont transcrit l'item répertorié dans la première ligne (voir page suivante).

Bien qu'il n'ait aucune prétention scientifique, un tel décompte donne des indices sur les zones de vulnérabilité, et par conséquent les points à travailler avec les apprenants. Si, d'une manière générale, la compréhension de chaque mot n'est pas en soi indispensable pour accéder au sens global des énoncés, il est cependant fort probable que l'identification quasi intégrale des unités composant l'énoncé garantit une compréhension globale de meilleure qualité, alors que des difficultés locales risquent au contraire de gêner le bon déroulement de la suite du processus de compréhension (Goh, 2000, p. 63-64).

6. La transcription ne nous intéressait qu'en ce qu'elle témoigne du niveau de compréhension de l'étudiant, quels que soient les écarts orthographiques par rapport à la norme.

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

	donc	elle	faisait	cinq	heures	du	matin	euh	je	sais	pas	moi	quatorze	heures
Rang FF	162	26	19	137	82	36	173	–	4	45	8	51	578	82
Lecture	[dõk]	[ɛl]	[fəzɛ]	[sɛ̃k]	[œʁ]	[dy]	[matɛ̃]	[ø]	[ʒə]	[sɛ]	[pa]	[mwa]	[katɔʁz]	[œʁ]
Articulé ici	=	[ɛ]	[vze]	=	=	=	=	=	[ʃe]		=	=	=	=
Transcrit ?	**5**	**4**	**2**	**19**	**19**	**18**	**18**	**4**	**16**	**14**	**13**	**12**	**15**	**18**

Tableau 1 – Un aperçu des points forts et faibles dans la compréhension de l'énoncé

Comme le montre le Tableau 1, la séquence *cinq heures du matin* n'a guère suscité de problème, ni, dans une moindre mesure, *quatorze heures*. Outre le caractère quasi figé en *n heures du matin*, cette première séquence présente l'avantage d'avoir été articulée conformément à la norme de lecture. Tel n'est pas le cas en revanche pour les séquences *je sais pas* et *elle faisait*. Si la première se voit prononcée sous sa forme courante [ʃepa], la seconde – identifiée par deux apprenants sur 19 – soulève davantage de difficultés en raison de la réduction de [ɛl] à [ɛ] – signalée par Salins (Salins, 1996, p. 21) dans sa *Grammaire pour l'enseignement-apprentissage du FLE* – et l'assimilation [vze] engendrée par la disparition du [ə], transformant donc l'initiale du lemme, dès lors plus difficile à identifier. Courantes, de telles difficultés de compréhension pourraient en partie s'expliquer par le décalage de ces formes avec celles résultant de la lecture ou de l'écoute d'écrits oralisés (journaux télévisés ou radiophoniques) ou interprétés (dialogues de manuels, films, etc.) (voir l'analyse de manuels par Vialleton & Lewis, 2014 ; Giroud & Surcouf, 2016 ; Surcouf & Giroud, 2016). Rappelons que d'une manière générale, au niveau purement phonétique :

> *the acoustic realization of continuous speech is severely degraded when compared to the maximally distinct isolated utterances often used in laboratory situations. The acoustic cues that accompany words spoken in isolation are often simply not present in fluent speech; segments and syllables are omitted, vowel color is significantly changed by consonantal environment* (Bond & Garnes, 1980, p. 116)

Si les changements phonétiques semblent pouvoir expliquer les incompréhensions face à *elle faisait* et *je sais pas*, plus étonnante s'avère en revanche la difficulté d'identification du marqueur *donc* pourtant d'un usage courant en français parlé. En somme, même après autant d'années d'apprentissage (voir note 445), nombre d'apprenants éprouvent encore des difficultés à l'écoute de l'oral spontané, et semblent davantage inscrire leurs attentes dans le prolongement des structures de l'écrit et de la lecture qui s'ensuit, comme en témoignent les commentaires[7] suivants de deux étudiantes de ce groupe :

> Personnellement […], je n'ai jamais appris qu'on pouvait éviter de dire « ne » dans un discours oral. C'est pourquoi, j'ai commencé à parler

7. Ces commentaires sont reproduits tels qu'ils ont été écrits par les étudiants dans leur compte rendu.

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

> français selon les règles de l'écriture. J'utilisais ainsi toujours la double négation lorsque je parlais. […] J'estime que les enseignants doivent souligner le fait que la structure des phrases diffère à l'écrit et à l'oral.

Les remarques ci-dessous soulignent quant à elles les répercussions possibles de telles difficultés en milieu homoglotte :

> je sentais que j'avais parfois de la difficulté à comprendre le discours des francophones natifs. […] il est notable que les réductions phonétiques, l'omission de *ne* dans la négation et les variantes formes de réductions phonétiques font partie du répertoire langagier des francophones natifs. Donc, il est primordial que les apprenants de FLE apprennent ces phénomènes de français qui sont assez courants à l'oral.

En dépit de la qualité de rédaction de ces deux témoignages et leur excellent niveau de français écrit, aucune de ces deux étudiantes n'est parvenue à transcrire les séquences *elle faisait* et *je sais pas* de l'Énoncé 1. Comment en définitive expliquer de telles difficultés, alors que « depuis une quarantaine d'années, à travers l'approche communicative, la notion de document authentique s'est imposée comme un des éléments structurants de l'enseignement du FLE » (Parpette, 2018, p. 19), et que « les programmes de FLE sont quasi systématiquement organisés selon une progression qui traite d'abord, aux niveaux élémentaires, les situations quotidiennes » ? En effet « les discours oraux sont indiscutablement l'outil de communication du quotidien en face-à-face, dans les situations privées au sens strict (la famille, les amis) ou au sens large (les relations sociales dans les commerces, l'école, les administrations, etc.) » (Parpette, 2018, p. 24)[8]. Dès lors, comme le préconisent Vialleton et Lewis :

> *when it comes to getting students to cope with the complexity of naturally-occurring speech they need to be confronted with that complexity to start being able to make sense of it.* […] *such an ecological approach is currently far from the norm in published course materials. This is problematic because it precludes empowering learners: they cannot become independent users of a language if they are not presented with the right representation of that language and if they are not given the right tools to understand it.* (Vialleton & Lewis, 2014, p. 312)

8. Voir également Biber (1988, p. 161) : « *Face-to-face conversation is a stereotypically oral genre, having the characteristic situational features that are most typical of speech* ».

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

Partageant les mêmes constats à la fois sur les causes et leurs effets, nous avons donc entrepris de créer une base de données permettant aux apprenants d'accéder aisément à de nombreux exemples sonores authentiques illustrant des traits langagiers caractéristiques du français parlé[9]. Avant d'en venir aux nombreux défis soulevés par la constitution d'un tel corpus, présentons un aperçu de l'affichage des résultats d'une recherche portant sur des altérations phonétiques, parmi lesquelles apparait notamment *faisait* (articulé [vze]), l'un des écueils à la compréhension de l'Énoncé 1 :

Figure 1 – Aperçu des résultats d'occurrences de mots phonétiquement « altérés »

Comme le révèle la Figure 1, l'affichage des résultats n'apparait pas sous la forme d'un concordancier, et s'explique par les choix effectués en amont lors de la conception de la base de données, sur laquelle nous allons maintenant revenir.

3 Les défis de la construction d'un corpus de français parlé à visée pédagogique pour le FLE

Bien que, selon Boulton et Tyne (Boulton & Tyne, 2014, p. 6 & 46), au cours des deux dernières décennies, l'apprentissage sur corpus en « est venu à s'imposer en tant que véritable nouvelle approche en didactique des langues », les auteurs relèvent néanmoins qu'« il arrive à peine en France » et que « les corpus disponibles sont […] majoritairement des corpus d'écrits (ou mixtes dans certains cas). L'oral

9. Les documents audio utilisés proviennent d'entretiens radiophoniques ou de reportages sur le vif comme *Les Pieds sur Terre* (France Culture) (voir Surcouf, 2020).

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

[étant] considérablement à la traine » en dépit du rôle central qu'il joue dans l'enseignement/apprentissage du FLE depuis les années soixante. Pourtant, dans la problématique de sensibilisation au français parlé qui est la nôtre, l'apprentissage sur corpus présente des atouts manifestes, notamment :

> *Data-Driven Learning helps learners to recognize the fuzzy nature of authentic language use in context, essential if they are to deal with it.* […] *DDL further provides access to the massive amounts of authentic language needed (input flood) but, crucially, it organizes it to make patterns salient, as is necessary for noticing.* (Boulton & Cobb, 2017, p. 350-351)

Toutefois, Vyatkina et Boulton évoquent certains écueils[10] :

> *One obvious obstacle is the non-transparent user interface of many available corpora which were designed by corpus linguists for specialists like themselves, and not with teachers and students in mind, requiring considerable levels of linguistic and technological sophistication.* (Vyatkina & Boulton, 2017, p. 2)

Effectivement, pour le français, bien qu'il existe déjà certains corpus oraux à visée pédagogique (par ex. PFC-EF, voir Detey, Lyche, Tchobanov, Durand & Laks 2009 ; Clapi-FLE, voir Ravazzolo, Traverso, Jouin & Vigner, 2015), ils constituent néanmoins un développement ultérieur de corpus originellement compilés pour la recherche, dont les préoccupations premières s'avèrent différentes[11]. Si le scientifique est disposé à consacrer de nombreuses heures à s'approprier la manipulation des outils d'exploration de la base de données dont dépend son travail de recherche, en tant que non-spécialiste, l'apprenant doit quant à lui pouvoir immédiatement se repérer dans l'interface. Wang (Wang, 2017, p. 91) rapporte les efforts récents de simplification à cet égard, consistant à rendre l'interface conviviale, sobre, tout en réduisant le nombre d'étapes sollicitant l'usager pour atteindre son objectif. Pour notre part, dès l'origine (récente) du projet[12], en nous concentrant sur

10. Rappelons qu'une telle approche reste marginale dans l'enseignement-apprentissage des langues (Boulton, 2017, p. 483), *a fortiori* dès qu'il s'agit d'oral (Paternostro, 2017, p. 281).
11. Ce qui n'est pas le cas du projet Fleuron, qui, dès l'origine avait des objectifs pédagogiques (voir André, 2018).
12. Si l'idée originelle remonte à 2012, sa mise en œuvre n'a véritablement commencé qu'en 2016.

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

l'usager potentiel de notre future interface (et malgré les limites budgétaires), nous avons essayé de garantir :

a) un usage simple et si possible intuitif, compatible avec des compétences numériques élémentaires ;
b) des instructions écrites compréhensibles dès le niveau A2 ;
c) un métalangage linguistique minimal et si possible transparent (par rapport à l'anglais) ;
d) un accès facile à tous les exemples sonores de la base de données, avec leur transcription en orthographe conventionnelle ;
e) la possibilité, pour chaque trait langagier répertorié, d'écouter, à titre d'illustration, un exemple sonore représentatif des autres exemples sonores de sa catégorie ;
f) l'écoute de segments très courts (2-4 secondes), rendant théoriquement la compréhension plus aisée ;
g) la possibilité, pour chaque exemple sonore, d'élargir le contexte d'écoute ;
h) la possibilité de ralentir le débit ;
i) idéalement, la possibilité d'un travail en autonomie.

En raison du nombre élevé de traits langagiers retenus pour l'annotation, la navigation n'est cependant pas aussi aisée qu'on l'aurait souhaité. Venons-en précisément aux défis qu'ont soulevé l'annotation et l'organisation de ces traits langagiers au sein de la base de données.

4 Les traits langagiers, leur étiquetage et leur organisation dans l'interface-administrateur

Entreprendre, comme « objectif spécifique », de familiariser les apprenants de FLE aux caractéristiques du français parlé au travers d'une base de données informatisée accessible via une interface en ligne nécessite des compétences multiples. Aussi en tant que concepteurs (2) avons-nous dû collaborer étroitement avec un informaticien (1), en ciblant l'apprenant en (3). Signalons que, bien que l'interface soit opérationnelle, le projet est toujours en cours de construction, et l'ouverture récente du site FLORALE ne nous a pas encore permis de collecter les commentaires des usagers.

Notre présentation ne concernera que les niveaux (2) et (3), et nous n'aborderons pas ici les critères linguistiques et pédagogiques de sélection des traits langagiers (voir Surcouf & Ausoni, 2018 ; Surcouf, 2020), mais les problématiques relatives à la transcription,

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

la segmentation, et l'annotation des émissions radiophoniques dans le logiciel Elan (Brugman & Russel, 2004)[13], avant le dépôt sur le site de l'interface-administrateur de FLORALE, où l'ensemble des traits langagiers est organisé pour en permettre la recherche et l'écoute via l'interface-usager en (3) (voir Figure 2).

(1)	informaticien	architecture informatique	connaissances informatiques
	↑↓		
(2)	concepteurs	**interface-administrateur**	connaissances linguistiques et didactiques
	↑↓		
(3)	utilisateurs (apprenants de FLE)	**interface-usager**	connaissances minimales en français (A2) et bases en littératie numérique

Figure 2 – Schéma des interactions entre les usages et les concepteurs de la base de données

4.1 La mise en place de la structure d'accueil des données dans les interfaces administrateur et usager

Pour assurer la cohérence informatique, linguistique et pédagogique des annotations sur le long terme, un modèle (format .etf) a été conçu dans Elan, et sert pour toutes les transcriptions. Ce modèle comprend, pour chaque locuteur, 30 strates[14] (« tiers »), dont 28 servent aux annotations des dimensions phonétiques, syntaxiques, discursives et lexicales, réparties comme suit :

13. Elan : Max Planck Institute for psycholinguistics, Nijmegen, disponible en ligne sur http://tla.mpi.nl/tools/tla-tools/elan [consulté en décembre 2020].
14. La première strate est dédiée à la transcription, la deuxième aux tokens, lesquels servent de base pour l'annotation des traits langagiers repérés à l'aide des vocabulaires contrôlés associés aux 28 strates restantes.

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

Dimension	Strates	Traits langagiers
phonétique	6	32
syntaxique	12	130
discursive	5	75
lexicale	5	54

Tableau 2 – Les quatre catégories présentées dans l'interface-usager

Si le nombre de dimensions linguistiques, quatre, répertoriées dans la colonne de gauche, a été arrêté dès l'élaboration de l'architecture informatique sous-jacente à l'interface-administrateur, le nombre de strates et de traits langagiers relève de notre responsabilité et peut si nécessaire être modifié dans le modèle d'Elan. Comme il apparait dans le Tableau 2, c'est la dimension syntaxique qui comporte le plus de traits langagiers, en raison d'une volonté délibérée de discriminer des phénomènes parfois regroupés dans une seule catégorie par les linguistes, ce qui, pédagogiquement, peut poser problème. Ainsi en est-il par exemple de toutes les dislocations, différenciées ici selon qu'elles sont par exemple à gauche, à droite, en fonction sujet. L'ensemble des vocabulaires contrôlés de ces 28 strates comporte à ce jour 291 étiquettes. À chaque strate est associé un vocabulaire contrôlé allant de 2 à 64 items[15], simples ou doubles, les deux types apparaissant dans l'exemple ci-dessous, où l'énoncé « la semaine il y avait rien » fait l'objet de deux annotations. La première sur « *il* y *avait* » sert à indiquer la prononciation [javɛ] à l'aide de l'étiquette double |IL Y A (tous temps)1| (balise ouvrante) et |IL Y A (tous temps)2| (balise fermante), la deuxième sur *rien* est signalée à l'aide de l'étiquette simple |NEG_simple| :

15. L'annotation d'autant de traits langagiers dans Elan est loin d'être évidente en raison de la longueur des listes déroulantes de chaque vocabulaire contrôlé. Le problème, délicat, consiste à trouver un compromis entre le nombre de strates et le nombre d'items par vocabulaire contrôlé pour chaque strate.

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

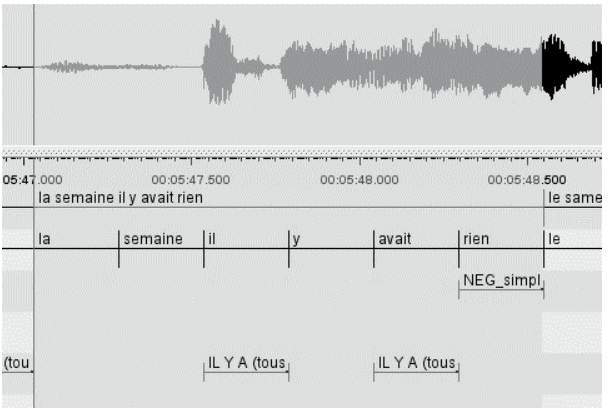
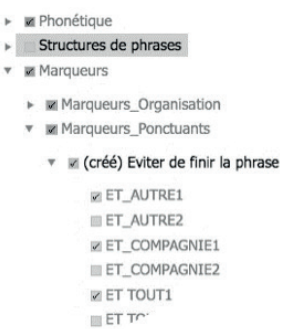


Figure 3 – Exemples d'annotations *double* sur « il y avait » et *simple* sur « rien »

L'ensemble des vocabulaires contrôlés d'Elan peut être exporté sous format .ecv vers l'interface-administrateur, où ils peuvent être répartis dans les quatre grandes catégories (Phonétique, Structures de phrase, Marqueurs, et Lexique), se développant au maximum en quatre niveaux de profondeur hiérarchiques, comme l'illustre l'aperçu ci-dessous :



Les quatre niveaux de profondeur de la grande catégorie des marqueurs discursifs, et leurs étiquettes doubles :

1. Marqueurs
2. Marqueurs_Ponctuants
3. Éviter de finir la phrase
4. *et autre* (1 & 2)
4. *et compagnie* (1 & 2)
4. *et tout* (1 & 2)

Figure 4 – Aperçu des quatre niveaux de profondeur de la catégorie *Marqueurs* dans l'interface-administrateur

Les informations inhérentes à chaque étiquette du vocabulaire contrôlé ou à chaque nœud de l'organisation hiérarchique peuvent être renseignées dans une fenêtre dédiée, permettant de :

a) faire apparaitre ou non le trait langagier dans l'interface-usager ;

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

b) changer le libellé technique en libellé pédagogique pour l'interface usager (voir Figure 5, ci-dessous) ;
c) intégrer un exemple sonore et sa transcription en guise d'illustration du trait langagier dans l'interface-usager ;
d) insérer un document d'explications sous format .pdf, qui sera téléchargeable à partir de l'interface usager (inexploité à ce jour) ;

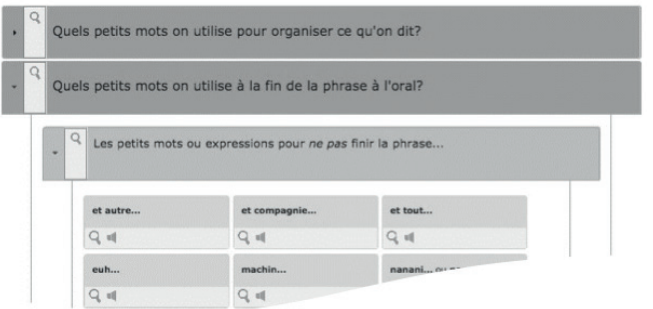


Figure 5 – Aperçu des « traductions pédagogiques » dans l'interface-usager

La structure d'accueil des données étant en place, il faut alors l'alimenter à l'aide des émissions transcrites et annotées.

4.2 Le traitement des données audio dans Elan

L'émission radiophonique est importée au format .wav dans Elan, où elle est découpée manuellement en segments sémantiquement et morpho-syntaxiquement cohérents d'environ 2-4 secondes. Une fois tous les segments transcrits en orthographe conventionnelle, ils sont subdivisés automatiquement par mot à l'aide du « tokenizer » d'Elan (voir par ex. les 6 tokens de « la semaine il y avait rien » dans la Figure 3). Les 28 strates du modèle sont alors importées et attribuées à chaque locuteur « tokenizé », autorisant désormais l'étiquetage manuel[16] des traits langagiers à l'aide des vocabulaires contrôlés associés à chacune de ces 28 strates. L'annotation effectuée[17], les fichiers .eaf et .wav sont conjointement exportés vers l'interface-administrateur et les métadonnées rentrées dans une page spécialement prévue à cet effet, où sont précisés le nom et la date de l'émission, le thème, les mots-clés, l'âge des locuteurs, etc.

16. Environ une cinquantaine d'annotations par minute de transcription.
17. Chaque transcription est supervisée deux fois, l'annotation une.

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

L'apprenant pourra alors, via l'interface-usager, accéder à tous les traits langagiers annotés dans Elan. À titre d'illustration, nous prendrons un exemple du type de recherche le plus complexe dans ce mode sachant qu'elle requiert quatre clics de la part de l'apprenant, soit le maximum.

En mode de recherche simple, l'apprenant arrive sur la page d'accueil suivante, où chaque dimension langagière est figurée par une couleur (se répercutant à tous les niveaux hiérarchiques de la catégorie explorée) :

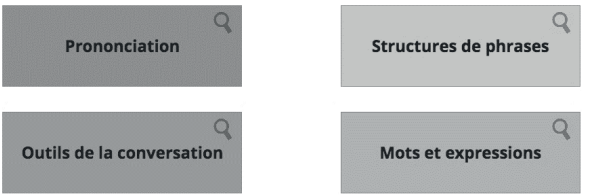


Figure 6 – La page d'accueil de l'interface-usager

En cliquant sur la loupe, l'apprenant affichera tous les segments sonores annotés relevant de la catégorie choisie. S'il désire en revanche affiner sa recherche par exemple sur les « Outils de la conversation », il devra cliquer sur le pavé en question, et accèdera aux questions suivantes :

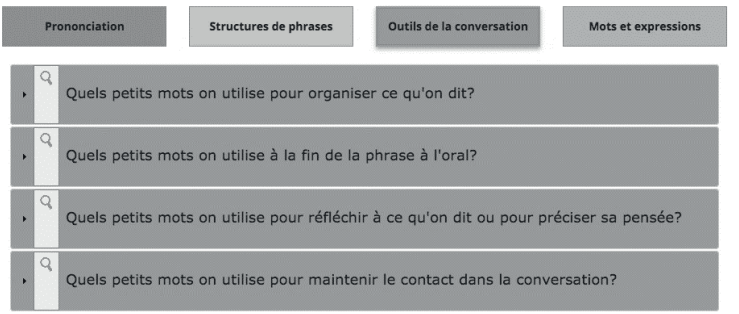


Figure 7 – Les quatre questions concernant les « Outils de la conversation »

Si l'apprenant entend savoir « Quels petits mots on utilise à la fin de la phrase à l'oral », un clic sur le pavé concerné lui offrira ici une liste de deux choix :

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

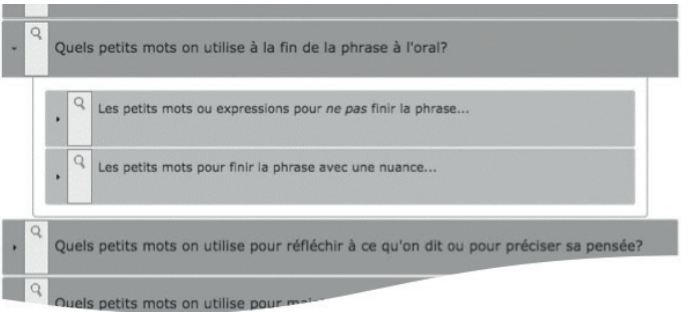


Figure 8 – Le dernier stade avant l'affichage des résultats

Un clic sur le pavé de la première option lui donnera accès aux traits langagiers correspondant à la question :

Figure 9 – L'affichage des traits langagiers annotés dans la base de données

À ce stade, l'apprenant est donc face à tous les traits langagiers annotés de cette sous-catégorie de « petits mots ou expressions pour *ne pas* finir la phrase… ». Il peut soit a) cliquer sur le hautparleur pour écouter l'exemple sonore (une infobulle apparait lui donnant la transcription de l'exemple sonore), soit b) cliquer sur la loupe et obtenir tous les exemples sonores illustrant ce trait langagier au sein de la base de données, soit, pour « et tout », à ce jour, 26 segments :

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

Figure 10 – L'affichage terminal des résultats d'une requête

Tous les exemples peuvent être écoutés isolément à l'aide du bouton ▶, avec la possibilité d'accéder à un contexte plus étendu dans « Extrait », et à l'intégralité de l'émission via « Tout ».

5 Les défis de la construction d'un corpus de français parlé à visée pédagogique pour le Français sur Objectifs Spécifiques (FOS)

Si la manipulation de cette « langue de tous les jours » (Gadet, 1996, p. v) – visée par FLORALE – s'avère fondamentale pour s'intégrer dans tout milieu socioprofessionnel ou académique, elle reste néanmoins insuffisante pour appréhender les discours de spécialité propres à ces milieux. Comme le montrent les recherches sur le FOS ou le FOU (Français sur Objectifs Spécifiques/Universitaires), l'apprenant a en effet tout intérêt à être expressément initié à la compréhension et la production des discours relevant de son champ d'activité. C'est d'ailleurs ce que fait ressortir l'interprétation comme « lignes directrices aux enseignants et aux concepteurs de programmes » des descripteurs des « niveaux communs de référence » du niveau C1 du *Cadre européen commun de référence pour les langues* :

> Peut comprendre une grande gamme de textes longs et exigeants, ainsi que saisir des significations implicites. Peut s'exprimer spontanément et couramment sans trop apparemment devoir chercher ses mots. *Peut utiliser la langue de façon efficace et souple dans sa vie sociale,* professionnelle *ou* académique. Peut s'exprimer sur des sujets complexes de façon claire et bien structurée et manifester son contrôle des outils d'organisation, d'articulation et de cohésion du discours. (2001, p. 25) *(c'est nous qui soulignons en italique et en gras)*

Limitons ici notre réflexion au monde académique. Bien que l'écrit

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

y joue incontestablement un rôle déterminant, les contraintes temporelles de sa production et de sa compréhension diffèrent largement de celles de l'oral. Rappelons que, pour comprendre l'oral (voir section 1), l'auditeur doit toujours s'adapter *en temps réel* au locuteur, à son débit, son accent, ses habitudes lexicales, discursives, etc. Fugace par nature, l'énoncé, unique et irréversible, doit être compris dans l'instant même de sa production. En ce sens, l'oral présente des difficultés particulières sur lesquelles l'apprenant a tout intérêt à travailler en FLE comme en FOS ou en FOU.

Alors que pour l'écrit, il est possible d'utiliser des ressources initialement produites sous format numérique, en dehors des difficultés logistiques et juridiques soulevées par l'enregistrement audio ou vidéo des données sur le terrain, la constitution d'un corpus oral ne peut se concevoir sans l'aide de transcriptions « très couteuses » en temps (Mangiante & Meneses-Lerín, 2016, p. 34). À cet obstacle se greffe éventuellement, selon l'exploitation informatique envisagée, le problème de la qualité de l'étiquetage grammatical automatique, peu adapté au français parlé (Benzitoun, Fort & Sagot, 2012 ; Habert, 2005, p. 64-65). Toutefois, quelles que soient les difficultés, d'un point de vue didactique, « le traitement du langage parlé à travers des corpus oraux s'avère nécessaire pour l'enseignement du français sur objectif spécifique » (Mangiante & Meneses-Lerín, 2016, p. 25). À ce jour, les corpus écrits restent très largement majoritaires, et leur grande taille rend propice le développement d'outils d'exploration plus sophistiqués que le simple concordancier. À titre d'illustration, le LEXICOSCOPE, développé par Kraif et Diwersy, présente un intérêt particulier pour le FOU en ce qu'il permet « d'accéder à différents types de corpus (littéraire, journalistique, académique, institutionnel, Web) » (Kraif, 2019, p. 73), déjà étiquetés[18], et au sein desquels il est possible d'effectuer non seulement des requêtes de « cooccurrences » mais de manière plus intéressante de « cooccurrence[s] étendue[s] », c'est-à-dire de « cooccurrents de l'expression formée du mot pôle associé à son cooccurrent » (Kraif, 2019, p. 74). Ainsi peuvent être mises en évidence des routines discursives propres à certaines disciplines. Si l'interface de LEXICOSCOPE donne accès à une variété de corpus écrits relevant de discours spécialisés des sciences humaines et sociales (géographie, sociologie, anthropologie,

18. « Le LEXICOSCOPE peut fonctionner avec n'importe quel type de corpus annoté en dépendances » (Kraif, 2019, p. 73).

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

psychologie, linguistique, sciences politiques, etc.), l'oral n'y apparait *a priori* que sous la forme d'écrit oralisé dans le corpus Europarl (Koehn, 2005), constitué de comptes rendus d'interventions au Parlement européen, dont la plupart résultent en fait de la lecture de textes rédigés au préalable par les orateurs. D'où, en définitive, pour le FOS et le FOU, « l'importance de constituer des corpus oraux afin d'analyser, étudier et enseigner les discours professionnels et leurs caractéristiques linguistiques » (Mangiante & Meneses-Lerín, 2016, p. 25). En ce qui concerne le FOU, Mangiante (Mangiante, 2017, § 38) rappelle effectivement que « l'oral et la prise de notes des étudiants constituent à l'université française le mode privilégié de transmission des connaissances », relevant par ailleurs que « parmi les corpus les plus représentatifs en FOU pour la préparation aux études supérieures en français figurent les enregistrements de cours magistraux, discours peu utilisés en cours de langue ». En somme, il s'avère d'autant plus fondamental de sensibiliser les apprenants à ces discours oraux que « les corpus oraux professionnels relèvent […] de règles spécifiques de genre et s'avèrent inhabituels pour les enseignants de langue et déroutants pour certains apprenants plus sensibles à "un enseignement conventionnel" » (Mangiante, 2017, § 19). Et c'est en ce sens qu'il pourrait être intéressant de s'inspirer de l'architecture informatique et pédagogique d'une plateforme comme FLORALE pour construire de nouveaux corpus oraux exploitables dès les niveaux B2 et C1 pour la pratique du FOS ou du FOU.

En nous inspirant des défis auxquels nous avons été confrontés lors de la construction de FLORALE, en guise de réflexion exploratoire, nous proposons de redéfinir les divers critères retenus plus haut (voir section 3), en imaginant cette fois-ci la constitution d'une base de données orales pour le FOS, dont l'interface serait utilisable en autonomie.

	FLE	**FOS**
a)	un usage simple et si possible intuitif, compatible avec des compétences numériques élémentaires	un usage compatible avec des compétences numériques avancées
b)	des instructions écrites compréhensibles dès le niveau A2	des instructions écrites compréhensibles au niveau C1
c)	un métalangage linguistique minimal et si possible transparent (par rapport à l'anglais)	un métalangage linguistique clair mais élaboré

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

d)	un accès facile à tous les exemples sonores de la base de données, avec leur transcription en orthographe conventionnelle	un accès facile à tous les exemples audio et vidéo de la base de données, leur transcription en orthographe conventionnelle, et toutes les métadonnées se rapportant à la situation et aux acteurs de l'interaction
e)	la possibilité, pour chaque trait langagier répertorié, d'écouter, à titre d'illustration, un exemple sonore représentatif des autres exemples sonores de sa catégorie	
f)	l'écoute de segments très courts (2-4 secondes), rendant théoriquement la compréhension plus aisée	l'écoute et le visionnage d'extraits très courts (2-4 secondes), rendant théoriquement la compréhension plus aisée
g)	la possibilité, pour chaque exemple sonore, d'élargir le contexte d'écoute	
h)	la possibilité de ralentir le débit	
i)	idéalement, la possibilité d'un travail en autonomie	la possibilité d'un travail en autonomie

Tableau 3 – Mise en regard des défis dans la construction de corpus avec interface-usager à destination des apprenants du FLE et du FOS

Si l'on considère avec Mangiante (Mangiante, 2017, § 22) que « le travail de conception didactique de l'enseignant de FOS consist[e] […] en partie à dégager la "part langagière du travail" […] ou plus exactement la part langagière de "l'action au travail" », alors, des recoupements existent nécessairement entre les problématiques développées par FLORALE et celles du FOS. Toutefois, aussi intéressante soit-elle comme base de réflexion, cette mise en parallèle de certains des défis occulte la disparité des phénomènes à étiqueter dans les deux types de ressource. En premier lieu, pour le FOS, l'intégration de la vidéo semble fondamentale en raison du caractère déterminant pour la compréhension du message du cadre de l'interaction et des actions qui s'y déroulent[19]. Surgiront dès lors de nouveaux défis informatiques et pédagogiques entrainés par la dimension visuelle, inexistante dans FLORALE. Ensuite, s'il était

19. Dans le cas particulier du FOU, que deviendrait par exemple un cours magistral si l'apprenant n'accède qu'à l'audio sans accès à l'image, le privant ainsi de ce qu'écrit l'enseignant au tableau, ou de ce qu'il commente de son diaporama ?

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

« relativement » aisé de puiser dans les recherches en linguistique pour définir la plupart des étiquettes des phénomènes à annoter dans FLORALE, il est probablement plus difficile de cerner avec certitude ce qui *doit* être étiqueté pour le FOS. En effet, pour prétendre sensibiliser aux caractéristiques des discours professionnels par le biais d'une interface à l'image de celle de FLORALE, il est nécessaire d'identifier un ensemble bien déterminé de catégories, dont chacune devra être étiquetée pour permettre l'annotation informatique, et, en aval, le requêtage de la part de l'usager. À titre d'illustration, prenons le cas du traitement des cours magistraux en FOU. Dans la lignée de Mangiante (Mangiante, 2017), on pourrait par exemple imaginer la création de deux grandes catégories correspondant aux deux types de dialogisme définis ci-dessous :

> Le cours magistral constitue un discours complexe de transmission des connaissances comportant deux dimensions énonciatives […] : un dialogisme interlocutif avec la prise en compte dans le discours des étudiants co-actants, réagissant, et un dialogisme interdiscursif dans la mesure où le discours de l'enseignant convoque d'autres discours de référence. (Mangiante, 2017, § 39)

Supposons admises ces deux grandes catégories, et nommons-les A et B. Dans son corpus, le concepteur doit alors :

1) repérer, selon les critères pédagogiques retenus, tous les phénomènes langagiers relevant de A et de B, établissant ainsi un nombre x et y de sous-catégories de A et de B ;
2) étiqueter de manière précise et unique chacune de ces sous-catégories au sein de la base de données ;
3) s'assurer – pour des raisons informatiques notamment – du caractère discret de toutes ces sous-catégories (il ne doit y avoir aucun chevauchement) ;
4) pour chaque sous-catégorie, trouver une étiquette pédagogique pour l'interface-usager ;
5) organiser l'ensemble pour que l'usager puisse aisément accéder via l'interface à l'information qu'il recherche.

Quelle que soit leur ampleur, entreprendre de répondre conjointement à ces défis informatiques et pédagogiques impose au concepteur une réflexion approfondie sur la nature même de ses données et la manière de les rendre aisément disponibles aux apprenants. On peut alors supposer

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

que, aussi difficile soit-elle, cette réflexion ne pourra que bénéficier à l'ensemble de la discipline.

6 Conclusion

Si déjà en soi la création d'une base de données de français parlé à des fins de recherches en linguistique soulève de redoutables défis, toute visée pédagogique ajoute de nombreuses contraintes en raison du public ciblé. En effet, dans l'absolu, aussi pertinente puisse-t-elle paraitre pour remplir l'« objectif spécifique » consistant à sensibiliser les apprenants de FLE au français oral quotidien, une telle base de données ne se révèlera intéressante qu'à partir du moment où elle rencontrera le public qu'elle cible, c'est-à-dire dans la configuration la plus exigeante pour le concepteur, l'apprenant de niveau A2 dépourvu de connaissances linguistiques et sans aucune expérience des outils d'exploration de corpus. En somme, bien que la brièveté de ce chapitre et la complexité des enchevêtrements entre contraintes informatiques, linguistiques et pédagogiques ne nous aient pas permis de l'aborder véritablement, une première question, fondamentale, s'est posée dès le début de la conception de l'interface : *comment faire en sorte que l'apprenant dont le niveau en français est encore relativement faible (A2) puisse comprendre les informations présentées sur l'interface ?* À ce premier défi, s'en est greffé un autre, sous-jacent à tout notre exposé : *comment parvenir à présenter près de 300 traits langagiers en les répartissant en quatre grandes catégories ?* En ce qui concerne la première interrogation, comme l'ont laissé entrevoir les captures d'écran précédentes, nous avons contourné l'usage du métalangage en recourant essentiellement à des questions formulées dans une langue simple, non seulement pour orienter la recherche, mais également interpeler la curiosité de l'apprenant tout en dynamisant l'interface. Quant au deuxième défi, il ne touche pas uniquement à la problématique informatique de la relation entre les interfaces administrateur et usager, mais prend ses racines dans la manière dont les 291 traits langagiers peuvent être manuellement annotés dans Elan, avec toutes les contraintes résultant parfois de la présence de plusieurs locuteurs dans le même enregistrement, sachant que pour chacun 30 strates sont nécessaires pour l'annotation.

À ce stade de développement, en dépit du caractère intuitif de l'interface, il est clair que l'usager à la recherche d'un trait langagier particulier

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

devra se montrer patient et curieux pour parvenir à ses fins[20]. Or, s'il travaille en autonomie, il est plus probable que sa frustration l'emporte sur sa motivation. En d'autres termes, en dehors de quelques améliorations ergonomiques et esthétiques, des parcours pédagogiques et des fiches explicatives en libre accès devront être proposés en ligne aux apprenants et enseignants afin de leur permettre de mieux tirer profit de cette base de données. Comme nous l'avons vu, la compréhension de l'oral spontané s'avère exigeante et continue de poser problème même en milieu homoglotte. En ce sens, un corpus de français oral quotidien comme le nôtre offrirait aux apprenants l'occasion d'améliorer cette compétence :

> *Most of the findings in the literature suggest that unscripted spoken language is generally more difficult for L2 learners, especially for those who have had little exposure to this type of language, but the literature also suggests that exposure to and drawing learners attention to these characteristics of unscripted spoken language leads to statistically significant learner gains.* (Wagner & Ockey, 2018, p. 26)

Mais en dehors de l'intérêt théorique que pourrait avoir notre base de données à cet égard – *a priori* novatrice dans sa finalité –, si l'on veut s'assurer qu'elle ait un quelconque avenir, il faudra la pourvoir d'activités pédagogiques en libre accès, à défaut, le risque est grand de la voir devenir victime d'une obsolescence pourtant non programmée.

Références bibliographiques

André Virginie, 2018, « Nouvelles actions didactiques : faire de la sociolinguistique de corpus pour enseigner et apprendre à interagir en français langue étrangère », *Action Didactique*, n° 1, p. 71-88.

Biber Douglas, 1988, *Variation across speech and writing*, Cambridge, Cambridge University Press.

Blanche-Benveniste Claire, 2003, « La langue parlée », dans M. Yaguello (dir.), *Le Grand Livre de la Langue française*, Paris, Seuil, p. 317-344.

Bond Z. S. & Garnes Sara, 1980, « Misperceptions of fluent speech », dans R. A. Cole (éd.), *Perception and production of fluent speech*, Hillsdale, Lawrence Erlbaum, p. 115-132.

20. Pour l'heure, bien que ce développement soit prévu dès que des ressources budgétaires seront disponibles, il n'est pas possible de savoir si un trait langagier comporte des illustrations sonores dans notre base de données (c'est-à-dire a été 1/ inclus dans le vocabulaire contrôlé, et 2/ annoté dans Elan).

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

Boulton Alex & Tyne Henry, 2014, *Des documents authentiques aux corpus. Démarches pour l'apprentissage des langues*, Paris, Didier.

Boulton Alex, 2017, « Corpora in language teaching and learning », *Language Teaching*, vol. 50, nº 4, p. 483-506.

Boulton Alex & Cobb Tom, 2017, « Corpus Use in Language Learning: A Meta-Analysis », *Language Learning*, vol. 67, nº 2, p. 348-393.

Brugman Hennie & Russel Albert, 2004, « Annotating Multi-media/Multi-modal Resources with ELAN », (proceedings of the *Fourth International Conference on Language Resources and Evaluation (LREC 2004),* 26-28 May), Lisbon, Portugal, European Language Resources Association, p. 2065-2068, disponible en ligne sur http://www.lrec-conf.org/proceedings/lrec2004 [consulté en decembre 2020].

Cadre européen commun de référence (CECR) pour les langues : apprendre, enseigner, évaluer. 2001, Unité des Politiques linguistiques, Strasbourg, disponible sur www.coe.int/lang-CECR [consulté 16 octobre 2020].

Cuq Jean-Pierre & Gruca Isabelle, 2002, *Cours de didactique du français langue étrangère et seconde*, Grenoble, Presses universitaires de Grenoble.

Cuq Jean-Pierre (éd.) 2003, *Dictionnaire de didactique du français langue étrangère et seconde*, Paris, CLE international.

Detey Sylvain, Lyche Chantal, Tchobanov Atanas, Durand Jacques & Laks Bernard, 2009, « Ressources phonologiques au service de la didactique de l'oral : le projet PFC–EF », *Mélanges Crapel*, nº 31, p. 223-236.

Gadet Françoise, 1996, *Le français ordinaire*, Paris, Armand Colin.

Giroud Anick & Surcouf Christian, 2016, « De "Pierre, combien de membres avez-vous ?" à "Nous nous appelons Marc et Christian" : réflexions autour de l'authenticité dans les documents oraux des manuels de FLE pour débutants », (actes du 5e Congrès Mondial de Linguistique Française (CMLF 2016), 4-8 juillet), Tours, *SHS Web of Conferences,* vol. 27, EDP Sciences, p. 1-18.

Goh Christine C. M., 2000, « A cognitive perspective on language learners' listening comprehension problems », *System*, nº 28-1, p. 55-75.

Gougenheim Georges, Michéa René, Rivenc Paul & Sauvageot Aurélien, 1964, *L'élaboration du français fondamental : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris, Didier.

Habert Benoît, 2005, *Instruments et ressources électroniques pour le français*, Gap, Ophrys.

Holec Henri, 1990, « Des documents authentiques, pour quoi faire ? », *Mélanges Crapel*, nº 20, p. 65-74.

Koehn Philipp, 2005, « Europarl : A parallel corpus for statistical machine translation » (proceedings of 10th Machine Translation Summit, AAMT, 12-16 December), Phuket, Thailand, p. 79-86.

Kraif Olivier, 2019, « Explorer la combinatoire lexico-syntaxique des mots et expressions avec le Lexicoscope », *Langue française*, nº 203, p. 67-82.

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

Léon Pierre, 1996, *Phonétisme et prononciations du français*, Paris, Armand Colin.

Mangiante Jean-Marc, 2017, « Discours et action(s) en milieux professionnel et universitaire : d'une norme d'usage à une contextualisation didactique en FOS et FOU », dans H. Tyne (éd.), *Le français en contextes*, Perpignan, Presses universitaires de Perpignan, disponible en ligne sur https://books.openedition.org/pupvd/2816 [consulté en décembre 2020].

Mangiante Jean-Marc & Parpette Chantal, 2011, « Le Français sur Objectif Universitaire : de la maitrise linguistique aux compétences universitaires », *Synergies monde*, n° 8, p. 115-134.

Mangiante Jean-Marc & Meneses-Lerín Luis, 2016, « Analyse des données et élaboration des contenus de formation en FOS : des corpus aux ressources », *Points Communs – Recherche en didactique des langues sur objectif(s) spécifique(s)*, n° 3, p. 25-43.

Parpette Chantal, 2018, « Quelle relation entre discours oral naturel et document oral authentique en FLE ? », *Action Didactique*, n° 1, p. 18-30.

Paternostro Roberto, 2016, *Diversité des accents et enseignement du français. Les parlers jeunes en région parisienne*, Paris, L'Harmattan.

Paternostro Roberto, 2017, « Peut-on enseigner la variation ? », dans H. Tyne, M. Bilger, P. Cappeau and E. Guerin (dir.), *La variation en question(s). Hommages à Françoise Gadet*, Bruxelles, Peter Lang, p. 279-290.

Porcher Louis, 1995, *Le français langue étrangère : émergence et enseignement d'une discipline*, Paris, Hachette.

Ravazzolo Elisa, Traverso Véronique, Jouin Émilie & Vigner Gérard, 2015, *Interactions, dialogues, conversations : l'oral en français langue étrangère*, Paris, Hachette.

Salins (de) Geneviève-Dominique, 1996, *Grammaire pour l'enseignement/apprentissage du FLE*, Paris, Didier-Hatier.

Surcouf Christian & Giroud Anick, 2016, « À quelle langue accède l'apprenant ? Examen critique du traitement de l'oral dans les premières leçons de manuels de français langue étrangère pour débutants », *Linguistik Online*, n° 78-4, p. 11-27.

Surcouf Christian & Ausoni Alain, 2018, « Création d'un corpus de français parlé à des fins pédagogiques en FLE : la genèse du projet FLORALE », *EDL (Études en didactique des langues)*, n° 31, p. 71-91.

Surcouf Christian, 2020, « Les enjeux de la compréhension du français oral quotidien en FLE : atouts possibles d'un corpus de français parlé annoté à des fins pédagogiques », *Études de linguistique appliquée*, n° 198, p. 241-256.

Vandergrift Larry, 2007, « Recent developments in second and foreign language listening comprehension research », *Language Teaching*, n° 40-03, p. 191-210.

Vialleton Élodie & Lewis Tim, 2014, « Reconsidering the authenticity of speech in French language teaching: theory, data, methodology, and practice », dans H. Tyne, V. André, Ch. Benzitoun, A. Boulton, Y. Greub (dir.), *French*

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.

through Corpora: Ecological and Data-Driven Perspectives in French Language Studies, Newcastle upon Tyne, Cambridge Scholars Publishing, p. 293-316.

Vyatkina Nina & Boulton Alex, 2017, « Corpora in language learning and teaching », *Language Learning & Technology*, n° 21-3, p. 1-8.

Wagner Elvis, 2014, « Using unscripted spoken texts in the teaching of second language listening », *TESOL Journal*, n° 5-2, p. 288-311.

Wagner Elvis & Ockey Gary J., 2018, « An overview of the use of authentic, real-world spoken texts on L2 listening tests », dans G. J. Ockey et E. Wagner (dir.), *Assessing L2 listening: moving towards authenticity*, Amsterdam/philadelphia, John Benjamins Publishing Company, p. 13-28.

Wang Ilaine (2017), *Syntactic Similarity Measures in Annotated Corpora for Language Learning: Application to Korean Grammar*, thèse de doctorat, université Paris Nanterre, Nanterre.

Worthington Debra L. & Fitch-Hauser Margaret E., 2018, *Listening: processes, functions, and competency*, Oxon, Routledge.

Cécile Frérot et Mojca Pecman (dir.), Des corpus numériques à l'analyse linguistique en langues de spécialité, UGA Éditions, coll. « Langues, Gestes, Paroles », 2021, ISBN 978-2-37747-261-1.