

198

avril-juin 2020

Les corpus oraux et leur didactisation

Coordonné par Véronique Bourhis et Roxane Gagnon

études de
linguistique appliquée
revue de didactologie
des langues-cultures
et de lexiculturologie

éla

Didier Érudition
Klincksieck

198

avril-juin 2020

Les corpus oraux et leur didactisation

Coordonné par Véronique Bourhis
et Roxane Gagnon

études de
linguistique appliquée
revue de didactologie
des langues-cultures
et de lexiculturologie

éla

Didier Érudition
Klincksieck

Les corpus oraux et leur didactisation

Les corpus parlés et leur didactisation : quelle parole (ap)prise dans l'espace de la classe ? Introduction.....	133
par Roxane GAGNON et Véronique BOURHIS	
Pour un français parlé de référence à partir de corpus	137
par Christophe BENZITOUN	
D'une parole individuelle à une dynamique argumentative collective	149
par Sylvie PLANE	
S'approprier les constructions des verbes <i>se souvenir</i> et <i>oublier</i> : prendre en compte la parole des élèves allophones	163
par Marie-Noëlle ROUBAUD et Marie-Laure ELALOUF	
Le synopsis comme annotation d'un corpus filmique. De l'objet enseigné à l'objet de recherche.....	181
par Véronique BOURHIS	
Des corpus vidéo pour (re)construire l'intention de la parole multimodale de l'enseignant.....	195
par Brahim AZAOUI	
Des corpus pour travailler la compréhension de l'oral.....	207
par Jean-François DE PIETRO, Roxane GAGNON et Christian REHM	
Didactiser des corpus de documents authentiques pour enseigner à interagir en langue étrangère	225
par Élodie OURSEL	

Les enjeux de la compréhension du français oral quotidien en FLE : création d'une base de données de français parlé annoté par Christian SURCOUF	241
BIOBIBLIOGRAPHIE DES CONTRIBUTEURS.....	257

LES ENJEUX DE LA COMPRÉHENSION DU FRANÇAIS ORAL QUOTIDIEN EN FLE : CRÉATION D'UNE BASE DE DONNÉES DE FRANÇAIS PARLÉ ANNOTÉ

Résumé : Depuis environ deux décennies, le développement des corpus oraux et des outils informatiques en linguistique et didactique des langues a ouvert de nouvelles perspectives en français langue étrangère (FLE). Les corpus oraux restent toutefois sous-représentés, à fortiori lorsqu'il s'agit d'entreprendre de familiariser les apprenants aux caractéristiques du français oral quotidien. Dans cet article, après avoir présenté les enjeux liés à la compréhension orale en FLE, nous offrons un aperçu des défis que représente l'élaboration d'une interface informatique permettant l'exploration d'un corpus de français parlé annoté à des fins pédagogiques.

Apprend-on vraiment à faire du vélo en roulant sur un tricycle ou à nager en pratiquant les mouvements allongé sur la pelouse à côté de la piscine ? (Holec 1990 : 72)

1. DE LA NÉCESSITÉ DE TRAVAILLER LA COMPRÉHENSION ORALE EN FLE

Comme Holec (1990) le suggère malicieusement, il paraît difficile de prétendre développer une compétence sans la travailler concrètement. En ce qui concerne la compréhension orale en langue étrangère, le bon sens dicterait en effet de recourir à « l'utilisation de documents authentiques¹ plutôt que construits : leur authenticité accroît 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 : 68). Pourtant, bien qu'emblématiques d'une certaine conception contemporaine

1. « tout document non-construit à des fins d'enseignement/apprentissage de langue » (Holec 1990 : 67).

de l'enseignement-apprentissage des langues, les manuels de FLE n'offrent guère d'occasions de travailler la compréhension orale du français tel qu'il est parlé au quotidien par les francophones :

Les discours authentiques oraux sont [...] très rares dans les méthodes de FLE avant le niveau B2. Ce phénomène peut paraître d'autant plus paradoxal que l'approche communicative accorde, comme les méthodologies précédentes, une place importante à l'oral en début d'apprentissage. (Parpette 2018 : 19)²

Quelles pourraient être les répercussions d'une telle conception de la compréhension orale ? Prenons un exemple concret³, observé dans le cadre d'un cours de FLE de niveau B2 en milieu universitaire homoglotte en Suisse romande. Vingt étudiants⁴ étaient invités à transcrire en orthographe des énoncés authentiques, extraits de reportages radiophoniques. Il s'agissait en l'occurrence de vérifier leur degré de compréhension fine du français parlé, et de détecter les éventuelles difficultés. À cet effet, leur était présenté, parmi d'autres, l'énoncé authentique [1], extrait d'une émission des *Pieds sur terre* (France Culture, 14/02/2017) :

[1] [eœtāfeœtāpøzystəplykwa] (*et euh et tu en fais un et tu en peux juste plus quoi*)

L'énoncé était diffusé à quatre reprises, avec une pause conséquente entre chaque écoute. Il était par ailleurs précisé par écrit que « le locuteur parle de ses enfants », information complétée par la transcription exacte des propos précédant le segment à transcrire :

ils ont cinq ans dans deux semaines quatre ans et au mois d'avril trois ans dix-huit mois d'écart entre les deux le temps que la maman récupère tu vois et puis quatorze mois entre le deuxième et le troisième ben après je me suis dit au moins j'ai fini tu vois j'aurais pas pu attendre trente-six trente-sept ans pour en avoir

Sur les vingt étudiants, un seul est parvenu à comprendre et transcrire convenablement l'énoncé, et seulement quatre étudiants ont réussi à identifier les éléments clés que constituent les deux *tu* génériques de l'énoncé. On pourrait bien sûr objecter qu'en l'occurrence la difficulté de compréhension était double dans la mesure où il s'agissait de *tu* génériques, apparaissant de surcroît sous leur forme réduite [t] (« et euh et *tu en* [=tā] fais un et *tu en* [=tā] peux juste plus quoi »). Toutefois, l'usage générique de *tu* est désormais courant en français conversationnel (voir Barbéris 2010), et la réduction de [ty] à [t]

2. L'origine des difficultés de compréhension du français oral quotidien, constatées même chez des apprenants de niveau relativement avancé (B2) est multifactorielle (voir Surcouf & Ausoni 2018 : 76-78), mais relève en partie d'une représentation de la langue cible puisant dans les normes de l'écrit : « L'apprenant écrit plus qu'il ne parle [...]. La puissance du modèle de l'écrit fait donc largement partie de l'imaginaire des apprenants et des enseignants » (Weber 2006 : 31).

3. Il ne s'agit là que d'une illustration sans prétention scientifique. Une étude approfondie sur la question requerrait l'intégration de données relatives à la fréquence d'utilisation du français, le réseau de relations francophones, le degré de motivation, les raisons du séjour en Suisse romande, etc.

4. Le groupe était composé de locuteurs de dix langues différentes (allemand (1), anglais (3), arabe (1), chinois (1), espagnol (2), italien (4), khmer (2), marathi (1), portugais (2), russe (3)). La durée moyenne de séjour en milieu francophone était de 3 ans (minimum : 3 mois, maximum : 8 ans) avec une durée moyenne d'apprentissage du français de 7 ans (minimum : 1 an 8 mois, maximum : 15 ans).

devant voyelle très fréquente, s'élevant à 87,1 % (196/225) au sein de notre corpus de français parlé de quatre heures (voir 2.4). Par ailleurs, avec un débit de 5,8 syllabe/s (13 syllabes articulées en 2,23s), cet énoncé est conforme aux résultats de Léon (1996 : 104) sur le français spontané. Dès lors, pourquoi un tel énoncé authentique à priori pragmatiquement, phonétiquement, lexicalemment et morphosyntaxiquement accessible à des apprenants de ce niveau suscite-t-il autant de problèmes de compréhension, alors qu'une fois proposé sous sa forme écrite, il leur était immédiatement compréhensible ? Pourquoi des apprenants de FLE, même de niveau avancé (ici B2), sont-ils si souvent déstabilisés par certaines caractéristiques fréquentes du français parlé ? Un tel constat ne serait-il pas imputable à un enseignement-apprentissage en partie inadéquat à cet égard, comme l'évoque clairement l'un des étudiants :

c'est aussi un choque pour nous des fois d'apprendre que les francophones natifs ne parlent pas comme on a été enseigné dans les manuels et cours de langues⁵. (voir également Durán & McCool 2003 ; Weber 2006)

C'est pourquoi nous avons entrepris de constituer un corpus informatisé de français parlé (voir Surcouf & Ausoni 2018), annoté à des fins pédagogiques pour le FLE, permettant à l'apprenant d'accéder à une multitude d'exemples sonores pour se familiariser avec le « français ordinaire » (c'est-à-dire « la langue de tous les jours » selon Gadet 1996 : V), tel qu'il est parlé par les francophones dans leur quotidien. Si en soi, l'idée est simple, sa mise en œuvre soulève de nombreux défis à la fois didactiques, linguistiques et informatiques, dimensions qui, s'enchevêtrant, doivent être pensées simultanément lors de la conception de la base de données et de l'interface. Aussi présenterons-nous les dimensions linguistiques et informatiques en relation avec la dimension didactique⁶, première dans notre réflexion.

2. LES ENJEUX DERRIÈRE LA CRÉATION D'UNE BASE DE DONNÉES INFORMATIQUE POUR L'APPRENTISSAGE SUR CORPUS EN FLE

2. 1. L'importance de la compréhension de l'oral

Dans la continuité de l'approche communicative, où « l'oral occupe une place de choix » (Cuq & Gruca 2002 : 247), de nos jours, pratiquer la compréhension orale est d'autant plus aisé qu'internet offre une grande diversité de documents sonores authentiques. Si une telle pratique peut se révéler souhaitable en milieu hétéroglotte, elle s'avère indispensable en milieu homoglotte dans la mesure où elle contribue à une meilleure intégration de l'apprenant

5. La graphie et la syntaxe sont celles de l'étudiant (anglophone ayant suivi un programme d'immersion de 14 ans au Canada et résidant en Suisse romande depuis 4 mois). Ce commentaire provient d'un travail d'analyse d'une transcription de deux minutes de français conversationnel dans le cadre d'une évaluation en fin de semestre.

6. Dans cet article, « didactique » renvoie à la *réflexion* en amont sur les stratégies et les moyens d'enseignement, « pédagogique » se rapporte à ce qui est/sera concrètement mis en œuvre durant le face-à-face entre l'enseignant et l'apprenant.

dans la société cible. Signalons que notre attention ne se portera pas ici sur la production, mais sur l'entraînement à la compréhension du français oral quotidien à l'aide d'une base de données permettant la recherche informatisée de traits caractéristiques du français parlé (voir 2.5). En effet, en accord avec Vandergrift, il semble indispensable de sensibiliser le plus tôt possible à l'écoute du français parlé ordinaire :

The ultimate goal of listening instruction is to help L2 listeners understand the target language in everyday situations. Authentic listening materials are best suited to achieve this goal because they reflect real-life listening, they are relevant to the learners' lives, and they allow for exposure to different varieties of language. (Vandergrift 2007 : 199)

2. 2. Compréhension et production orales

Aussi incontournable soit-elle, la compréhension orale constitue un redoutable défi pour l'apprenant dans la mesure où, schématiquement (voir Cutler 2012 : 304s ; Vandergrift & Goh 2012 : 38s), en dehors de connaissances sociopragmatiques, elle requiert la mise en œuvre de processus cognitifs complexes, qui devront lui permettre *en temps réel* de :

- (1) discriminer et identifier tous les phonèmes de la langue cible
- (2) segmenter le flux sonore en unités significantes
- (3) repérer l'agencement morphosyntaxique et textuel de ces unités significantes

La difficulté est d'autant plus grande qu'une telle activité impose l'analyse d'environ dix à quinze sons par seconde (en 1), soit la reconnaissance de deux à trois mots durant ce même intervalle (en 2) (Levelt 1994 : 90), le tout devant s'organiser en (3) pour constituer des agencements syntaxiques et sémantiques cohérents de quelques secondes seulement, sachant que ces trois processus ne se déroulent pas de façon linéaire, mais s'enchevêtrent (Vandergrift & Goh 2012 : 43).

Par ailleurs, « la capacité de diversification stylistique » étant « une propriété des langues en usage » (Gadet 2004 : 1), la prise en compte de la variation s'avère fondamentale, et à fortiori dans les approches pédagogiques actuelles, où la communication occupe une place centrale (Tyne 2012 : 107). Qu'en production, la maîtrise de la « compétence sociolinguistique⁷ » représente l'un des derniers obstacles de l'apprentissage (« the final hurdle », Tyne 2009 : 243), qu'en est-il en compréhension ? L'apprenant est-il en mesure de comprendre ce qu'il ne produit pas lui-même ? En effet, si la production fréquente d'un trait langagier semble impliquer sa compréhension⁸, sa non-production n'est pas pour autant synonyme de non-compréhension. Voyons brièvement quels sont les enjeux.

7. « C'est-à-dire le fait de pouvoir produire et reconnaître un style de langage relatif à une situation de communication donnée [...] impliqu[ant] de fait la considération de traits renvoyant aux dimensions stylistique (diaphasique) et sociale (diastatique) » (Tyne 2012 : 104).

8. On peut ainsi supposer que l'étudiant utilisant fréquemment *tu* sous sa forme réduite aurait correctement segmenté et compris [etâfeẽ] dans l'énoncé [1].

On sait que, dans la continuité des pratiques d'enseignement en français langue maternelle (à cet égard, voir Gagnon & Benzitoun, 2020), la norme écrite occupe une place importante en FLE⁹, tendant à masquer l'existence des variantes du français parlé. Or, du point de vue de la compréhension, fondamentale s'avère l'exposition de l'apprenant aux variantes courantes du français oral quotidien. Comment l'apprenant pourrait-il effectivement inférer par lui-même que les énoncés transcrits en phonétique ci-dessous sont sémantiquement équivalents¹⁰ ? En effet, en raison du statut privilégié de la variante écrite normée dans l'apprentissage guidé caractéristique des systèmes scolaires et universitaires, il semble naturel qu'en l'absence de sensibilisation aux variantes parlées ordinaires, la lecture s'impose à l'apprenant comme étalon de l'oral.

VARIANTE ÉCRITE	VARIANTE LUE	VARIANTE ORDINAIRE	DIVERGENCE ¹¹
<i>quand j'étais petit</i>	[kãzetepəti]	[kãʃtepti]	phonétique
<i>tu habites où ?</i>	[tyabitu]	[tabitu]	phonétique
<i>je le lui ai dit</i>	[zələlɥiɛdi]	[zɥiɛdi]	phonétique/syntaxique
<i>je ne sais pas</i>	[zənəsɛpa]	[ʃɛpa] ¹²	phonétique/syntaxique
<i>son père ne la voit plus</i>	[sɔpɛʁnələvɔiply]	[sɔpɛʁilavɔiply]	phonétique/syntaxique

Tableau 1 – Exemples de variantes courantes et des dimensions qu'elles affectent

Rappelons tout d'abord une évidence probablement constitutive du développement de la compréhension orale durant l'apprentissage. Intuitivement, l'apprenant sait que, quelle que soit la langue, il a tout intérêt à postuler que deux énoncés dissemblables acoustiquement seront sémantiquement différents. De manière flagrante, il en serait par exemple ainsi de [ləbebedɔk] comparé à [ilplø] (*le bébé dort ≠ il pleut*). Plus subtilement, conformément au système phonologique du français, il devra discriminer [ilzɔʃo] de [ilsɔʃo] (*ils ont chaud ≠ ils sont chauds*). En d'autres termes, il semble logique et productif d'adopter le principe selon lequel une différence acoustique entraîne une différence sémantique. Cependant, aussi rentable soit-il, ce principe se voit contredit par l'existence de variantes stylistiques, inhérentes au fonctionnement même de la langue. L'apprenant devra par exemple apprendre que [zəpɑʁl] et [ʃpɑʁl] sont deux énoncés sémantiquement identiques (*je parle*) en dépit des différences acoustiques. Se pose alors la question de l'appropriation de telles variantes stylistiques dont la connaissance est fondamentale en compréhension du fait de leur fréquence dans le français oral quotidien. Le tableau ci-dessous récapitule les quatre configurations, dont la (iv) se trouve être la plus complexe

9. « L'apprenant est en effet habitué à situer la norme par rapport à une seule référence (l'écrit) » (Weber 2010 : 178).

10. Nous n'aborderons pas ici la question délicate posée par Gadet (2004 : 4) : « est-ce qu'on peut dire la même chose en disant différemment, selon différents styles ? » (pour une brève réflexion à ce propos, voir Gadet 2007 : 153-154).

11. Les variantes lexicales (non-présentées ici) restent les plus faciles à identifier notamment parce qu'on les trouve dans les dictionnaires. Les phénomènes phonétiques, syntaxiques ou discursifs font quant à eux rarement l'objet d'une présentation dans les ouvrages de référence destinés aux apprenants.

12. Voir les déboires relatés par Durán & McCool (2003).

en ce qu'elle contrevient au principe de base de la configuration (i) – et de son corolaire (ii) –, qui invite à envisager comme sémantiquement différents des énoncés acoustiquement différents.

	ACOUSTIQUEMENT		SÉMANTIQUEMENT	REMARQUE
(i)	[зэравлЕ] ≠ [зЕравлЕ]	⇒	<i>je parlais ≠ j'ai parlé</i>	principe de base
(ii)	[зэравлЕ] = [зэравлЕ]	⇒	<i>je parlais = je parlais</i>	
(iii)	[зларовт] = [зларовт]	⇒	<i>je l'apporte ≠ je la porte</i>	homophonie (occasionnel)
(iv)	[зэравлЕ] ≠ [ʃравлЕ]	⇒	<i>je parlais = je parlais</i>	cas difficile (courant)

Tableau 2 – Exemplification des quatre configurations en compréhension orale

En raison de l'incomplétude de ses connaissances sur la langue cible – par la nature même du processus d'apprentissage dans lequel il s'inscrit –, l'apprenant éprouvera nécessairement des difficultés à déterminer s'il se trouve dans la configuration (i) ou (iv). Par ailleurs, même pour des variantes phonotactiquement motivées comme [зэкув]/[ʃкув] (*je cours*)¹³, il est peu probable qu'en l'absence de connaissances linguistiques, l'apprenant puisse inférer leur équivalence sémantique. En définitive, afin de garantir une meilleure compréhension orale, il semble souhaitable de sensibiliser les apprenants à l'existence de telles variantes, et ainsi répondre à l'une des exigences d'une étudiante du groupe¹⁴ :

j'ai commencé à parler français selon les règles de l'écriture [...] J'estime que les enseignants doivent souligner le fait que la structure des phrases diffère à l'écrit et à l'oral.

Comme l'illustre le tableau 1, toutes les dimensions (phonétiques, morphologiques, syntaxiques, lexicales, discursives) peuvent donner lieu à des variantes stylistiques et générer des problèmes de compréhension orale chez l'apprenant. Or, on ne peut que constater « l'absence de réflexion sur le rôle du style dans l'enseignement du français » (Paternostro 2017 : 285) et le « décalage important entre la langue présentée dans les manuels de cours et la langue « authentique » [...] [d]es corpus, surtout pour la langue parlée » (Debaisieux & Boulton 2007 : 31)¹⁵. Comment dès lors permettre à l'apprenant de pratiquer cette compétence ? En quoi l'apprentissage informatisé sur corpus pourrait-il contribuer à l'améliorer ?

2. 3. Les corpus oraux informatisés : focaliser sur l'oral, certes, mais lequel ?

Force est de constater qu'en didactique des langues, l'apprentissage informatisé sur corpus demeure marginal (Boulton 2017 : 483), et l'enseignant de FLE désireux de faire travailler ses apprenants sur le français parlé se heurte à une première difficulté :

Les corpus disponibles sont aujourd'hui majoritairement des corpus d'écrits (ou mixtes dans certains cas). L'oral est considérablement à la traîne. Or, en ce qui concerne la

13. Ici, l'assimilation de [з] à [ʃ] devant [k] après la disparition de [ə] dans *je*.

14. Pour la provenance de ce commentaire, voir note 4.

15. Voir également Étienne & Sax (2009) ; Giroud & Surcouf (2016) ; Surcouf & Giroud (2016) ; Vialleton & Lewis (2014).

didactique des langues, force est de constater que l'oral occupe une place majeure. (Boulton & Tyne 2014 : 46)

Debaixieux (2009 : 43) déplorait pour sa part qu'« il n'existe pratiquement aucun outil de consultation convivial, adapté à un usage d'apprentissage et surtout libre d'utilisation ». Aussi avons-nous entrepris de constituer FLORALE¹⁶, une base de données à visée pédagogique dans le but de sensibiliser les apprenants aux caractéristiques du français oral quotidien tout en tirant profit de l'apprentissage informatisé sur corpus et des avantages qu'il pourrait offrir. Boulton & Cobb (2017 : 349-350) évoquent ainsi : l'authenticité des données, la saillance des formes les plus fréquentes, le développement de la démarche inductive, l'apprentissage basé sur l'exemple et la possibilité de travailler en autonomie.

S'il existait déjà des corpus oraux conçus pour la recherche en linguistique et détournés par la suite pour servir des fins pédagogiques (pour PFC-EF, voir Detey *et al.* 2009 ; pour Clapi-FLE, voir Ravazzolo *et al.* 2015), la création d'un corpus à visée expressément pédagogique soulève des questions différentes, dont, en premier lieu, celle du choix des documents sonores. En effet, si l'on reconnaît l'importance que revêt l'authenticité dans la poursuite de notre objectif pédagogique, ce terme n'en reste pas moins ambigu et continue de faire débat en didactique des langues (voir par exemple Chambers 2009 ; Dubois *et al.* 2010). Ainsi, dans sa réflexion sur l'importance d'intégrer la dimension sociolinguistique en FLE, Paternostro constate que, ces dernières années :

Le recours aux documents authentiques [...] ne semble guère avoir changé la donne, du fait que ces documents – le plus souvent issu des médias – relèvent d'un genre écrit oralisé et manifestent des traits parfois éloignés de la langue quotidienne. Le recours aux corpus oraux et aux enregistrements de situations réelles demeure encore marginal. (Paternostro 2017 : 281)

Que la source soit orale n'est effectivement pas un critère suffisant pour garantir la possibilité de travailler la compréhension des variantes stylistiques du français oral quotidien. En effet, si, par définition, en accord avec le *Petit Robert*, l'oral concerne tout ce qui « se transmet par la parole », il peut néanmoins revêtir plusieurs formes, comme l'illustre l'aperçu ci-dessous :

	Auteur	Finalité prévue par l'auteur	Encodeur :	sa tâche	Décodeur	Exemple
①	Écrit	sera lu en silence	différent	lit à voix haute	écoute	roman lu
②						
③		sera lu et interprété		interprète		doublage de film
④		sera mémorisé et interprété				
⑤	Parle	est écouté en direct	identique	parle		conversation quotidienne

Tableau 3 – Aperçu des configurations prototypiques de quelques types de productions orales

16. Français langue orale pour le FLE : <https://florale.unil.ch/> (voir Surcouf & Ausoni 2018).

Seule la configuration ⑤ représente l'oral non-préparé et correspond dans sa version la plus fréquente au français oral quotidien. C'est aussi celle qui suscite le plus de difficultés de compréhension chez les apprenants, paradoxalement, davantage familiarisés au cours de leur apprentissage à des écrits oralisés¹⁷ de type ② ou ④, notamment au travers des dialogues (plus ou moins réalistes) des manuels de FLE (voir note 15). Aussi notre attention s'est-elle essentiellement portée sur des documents sonores de type ⑤, permettant de travailler la compréhension des traits caractéristiques du français oral quotidien. Quels documents avons-nous choisis ?

2. 4. Le choix des documents sonores

Pour l'heure, l'essentiel de nos données provient de l'émission radiophonique *Les Pieds sur Terre* (France Culture), qui « s'organis[e] [...] autour de récits, d'histoires vraies [...] racontées à la première personne et nourries d'éléments de reportage »¹⁸. Ces enregistrements présentent l'avantage d'être de qualité professionnelle, assurant donc une excellente qualité acoustique malgré la présence occasionnelle de bruits de fond ou de musique, ajoutée au montage. En effet, à l'instar de beaucoup d'œuvres radiophoniques proposées en différé, ces émissions sont mixées et peuvent donner lieu à l'insertion de musique ou à la suppression – non systématique – de la parole du journaliste durant l'interaction. En ce sens, même si elles répondent à notre objectif pédagogique premier consistant à présenter dans leurs multiples dimensions les variantes stylistiques du français parlé, ces émissions ne se prêteraient guère à l'observation des interactions en français (à cet égard, voir Clapi-Fle). À ce stade de développement, notre base de données comporte quatre heures d'émissions sous format .wav, importées dans le logiciel Elan¹⁹, où nous les avons segmentées manuellement en unités morphosyntaxiquement cohérentes d'environ trois secondes. Elan permet non seulement d'aligner la transcription sur le signal sonore, mais offre de surcroît la possibilité d'inclure de nombreuses strates d'annotations (voir plus loin l'illustration 2).

Une fois le choix des sources sonores effectué, s'est posée la question beaucoup plus délicate de la sélection des phénomènes du français parlé à annoter dans la transcription, pour leur présentation ultérieure dans l'interface, où ils deviennent recherchables et écoutables (voir Illustration 6).

2. 5. Le choix des phénomènes

S'il nous a paru fondamental de mettre en évidence les traits langagiers du français oral quotidien se démarquant notamment des usages de l'écrit normé (ou de l'écrit oralisé qui en émane), la sélection des phénomènes s'est effectuée conformément aux critères suivants (susceptibles de se recouper) :

17. Envisagé comme « type de réalisation dans lequel une phase d'émission orale succède à une phase de production écrite non spontanée » (Boullanger *et al.* 1972 : 6).

18. <https://www.franceculture.fr/emissions/les-pieds-sur-terre> (consulté le 25/02/2019).

19. Max Planck Institute for Psycholinguistics (Nijmegen) disponible ici : <https://tla.mpi.nl/tools/tla-tools/elan/>.

- ① Le phénomène diverge de la norme écrite (souvent prise comme référence)
- ② L'explication du phénomène apparaissant dans les manuels ou les ouvrages de référence de FLE ou de FLM n'est accompagnée d'aucun exemple sonore et demanderait à l'être pour devenir compréhensible
- ③ La description du phénomène est inexistante ou rare dans lesdits ouvrages
- ④ La description du phénomène est incomplète et/ou à connotation prescriptive :
 - de manière explicite : « Ne dites pas *t'aimes*, dites *tu aimes* » (Delatour *et al.* 1991 : 163)
 - de manière implicite : « *Ils, elles* se prononcent [il] [ɛl] devant une consonne » (Poisson-Quinton *et al.* 2003 : 34)
- ⑤ Le phénomène est fréquent en français oral quotidien, mais en partie ignoré en raison de ③ ou ④, et requerrait par conséquent une sensibilisation spécifique pour être mieux perçu et compris
- ⑥ Le phénomène fait typiquement l'objet de problèmes de compréhension et/ou de production
- ⑦ Le phénomène est difficile à expliquer pédagogiquement (ou linguistiquement) et l'observation de nombreux exemples pourrait alors aider l'apprenant à s'en développer une compréhension intuitive

À titre d'illustration, nous fournissons quelques exemples concrets extraits de notre corpus, et recherchables via l'interface :

PHÉNOMÈNE	EXEMPLE	①	②	③	④	⑤	⑥	⑦
Assimilation	ben du coup <i>je suis</i> [fui] partie		x	x		x		
[ty] prononcé [t]	<i>tu as</i> [ta] quel âge ?			x	x	x	x	
disparition d'un son	<i>c'était</i> [ste] pas douloureux			x				x
intonation de liste	un type un peu <i>terne</i> et un peu euh <i>vouté</i> euh <i>maigre</i>		x	x				
négation sans <i>ne</i>	il pouvait <i>pas</i> retourner chez lui	x		x	x	x		
<i>tu</i> générique	<i>tu</i> peux pas être célibataire en Iran <i>tu</i> dois <i>te</i> marier	x		x		x	x	
<i>quoi</i>	du coup voilà c'était une préparation à l'exil <i>quoi</i>	x		x		x	x	x
<i>quand même</i> ²⁰	<i>c'est quand même</i> bien agréable <i>quoi</i>							x

Tableau 4 – Quelques exemples issus du corpus et leur correspondance aux critères de sélection

Ces critères nous ont permis de retenir à ce jour environ 280 phénomènes répartis dans les quatre grandes catégories suivantes :

LIBELLÉ DANS L'INTERFACE	DIMENSION	NOMBRE
Prononciation	phonétique	30
Structures de phrase	morphosyntaxique	130
Outils de la conversation	discursive/conversationnelle	70
Mots et expressions	lexicale	50

Tableau 5 – Les quatre catégories présentées dans l'interface

20. Souvent prononcé [kœmmœm] et erronément écrit « comme même ».

L'étiquetage de ces 280 phénomènes a été effectué manuellement dans Elan sur les items « tokenisés²¹ » de la transcription orthographique à l'aide des vocabulaires contrôlés (associés à trente strates d'annotation – « tiers »²²), comme illustré ci-dessous :

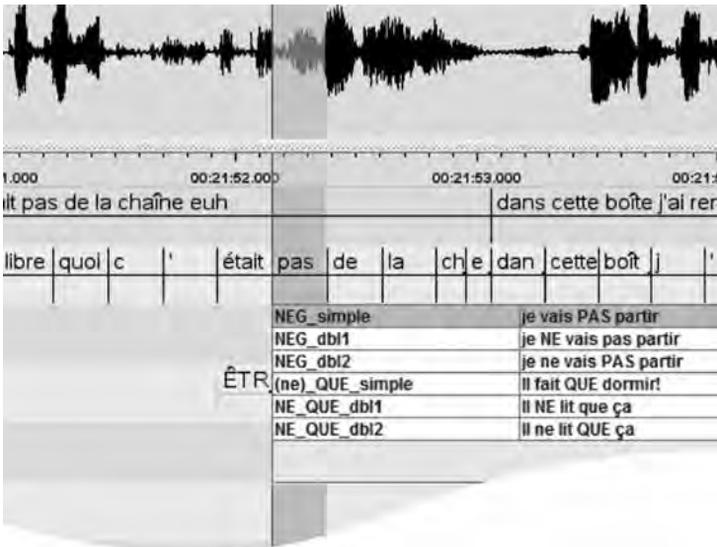


Illustration 1 – Aperçu d'une liste déroulante d'un vocabulaire contrôlé dans Elan durant l'annotation

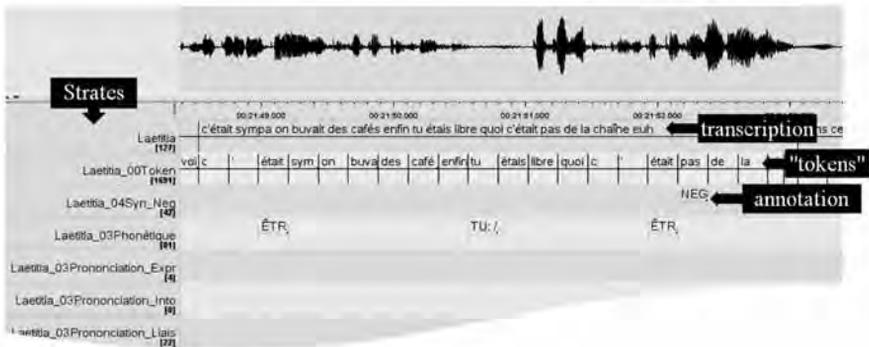


Illustration 2 – Aperçu de quelques annotations de segments audio dans Elan

21. Il s'agit d'une sous-segmentation automatique isolant tout item compris entre deux espaces.
 22. Comme le montre l'illustration 2, le logiciel Elan permet la segmentation et l'alignement sur le signal sonore de la transcription, laquelle peut alors faire l'objet d'annotations à l'aide de strates, au sein desquelles un item peut être étiqueté manuellement en recourant à un vocabulaire contrôlé (liste déroulante d'options prédéfinies par les annotateurs, voir l'illustration 1). Par exemple, l'item *pas* de la strate « Token » permet l'étiquetage de la négation simple du segment « c'était *pas* de la chaîne » en sélectionnant « NEG_simple » du vocabulaire contrôlé de la strate « Syn_Neg ».

Pour l'heure, notre corpus comprend environ 12 000 annotations, soit une cinquantaine par minute. L'étiquetage manuel permet de garantir la qualité des résultats obtenus dans l'interface, qu'il faut désormais évoquer pour les redoutables défis d'organisation qu'elle a posés.

2. 6. La présentation des contenus du corpus : l'organisation de l'interface

Si un linguiste est disposé à consacrer plusieurs heures pour se familiariser avec l'interface d'une base de données qu'il considère fondamentale pour sa recherche, un apprenant doit en revanche pouvoir trouver rapidement ce qu'il cherche. Qu'entraîne alors la poursuite d'un tel objectif pédagogique (ambitieux) dans la conception de l'interface ?

Rappelons notre finalité première. L'apprenant doit pouvoir prendre conscience des caractéristiques du français oral quotidien en accédant à l'écoute d'une multitude de courts extraits sonores (d'environ 3s) et de leur transcription (tout en conservant la possibilité d'écouter un contexte plus étendu pour chacun d'entre eux). L'utilisateur ordinaire n'étant ni linguiste ni informaticien, les exigences les plus élevées apparaîtront dans le cas d'un apprenant de niveau A2²³, qui devra pouvoir :

- trouver le trait langagier recherché sans passer par l'appropriation d'outils informatiques spécifiques d'exploration de corpus
- se repérer facilement à tous les niveaux de l'interface malgré la multitude des traits langagiers répertoriés
- effectuer n'importe quelle recherche au sein du corpus sans connaissances préalables en linguistique
- comprendre toutes les instructions verbales en français apparaissant dans l'interface

La conjonction de ces divers critères soulevant de redoutables défis, s'est vite imposée comme fondamentale la création d'une interface intuitive, facile à utiliser, en dépit de la complexité résultant de l'organisation en amont des 280 phénomènes, regroupés en l'occurrence en quatre grandes catégories sur la page d'accueil de l'interface :



Illustration 3 – La page d'accueil de l'interface

23. Cette exigence est en soi trop sévère dans la mesure où dans le parcours d'apprentissage, certains traits langagiers sont nécessairement prioritaires, et pour cette raison devront être plus aisément accessibles.

Chacune de ces quatre dimensions est caractérisée par une couleur spécifique se répercutant dans chaque sous-catégorie. Par exemple, un premier clic sur le pavé « Outils de la conversation » offre quatre choix reprenant la même couleur que la catégorie mère, et proposés sous forme de questions dans une langue se voulant le plus accessible possible à l'apprenant :

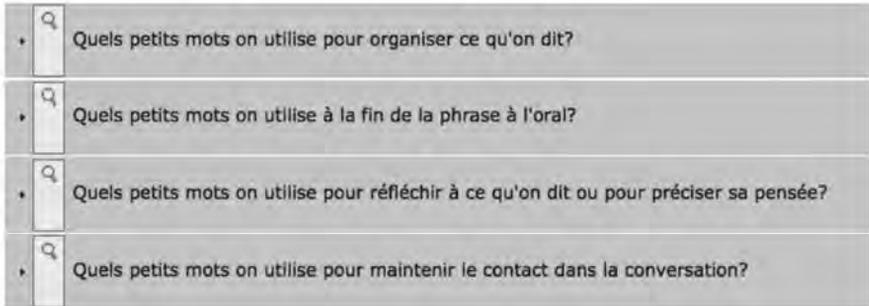


Illustration 4 – Le premier niveau d'accès après un clic sur le pavé « Outils de la conversation » de la page d'accueil

Les questions sont censées remplir les trois fonctions essentielles suivantes :

- leur formulation sert à contourner l'usage du métalangage technique
- elles permettent le regroupement des phénomènes relevant d'une même sous-catégorie
- elles dynamisent l'interface en interpellant l'utilisateur

Si l'utilisateur entend savoir « quels petits mots on utilise pour organiser ce qu'on dit »²⁴, alors un clic sur le premier pavé lui offre une liste de marqueurs :

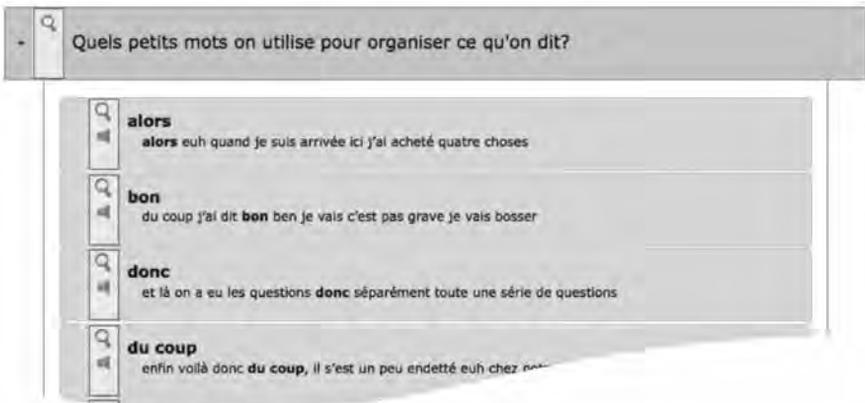


Illustration 5 – Un aperçu du début de la liste des « petits mots » correspondant à la question

24. Les formulations présentes sur le site sont toutes susceptibles d'être modifiées pour mieux répondre aux besoins des usagers.

Pour chaque pavé de couleur, l'utilisateur peut écouter un exemple sonore en cliquant sur le hautparleur, tout en lisant la transcription où ce trait langagier est mis en évidence en gras. Si l'apprenant désire accéder à l'intégralité des occurrences annotées, il lui suffit de cliquer sur la loupe, qui, dans le cas de *du coup*, lui présentera 85 exemples, surlignés en couleur :



Illustration 6 – Aperçu de l'affichage des premiers résultats du marqueur « du coup »

Précisons que les résultats des requêtes ne s'affichent pas sous la forme conventionnelle d'un concordancier et sa forme pivot au centre (à ce propos, voir Tribble 2013). Chaque ligne présente en effet le segment sonore tel qu'il a été découpé, transcrit et annoté dans Elan, comme l'illustre la capture d'écran du premier exemple tel que traité dans Elan :

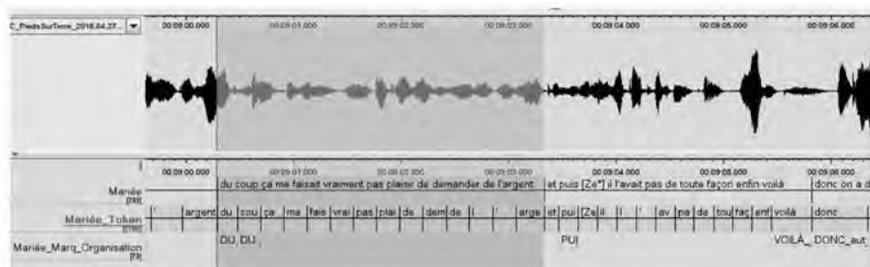


Illustration 7 – La segmentation dans Elan constitue l'unité minimale d'écoute dans l'interface

Chaque exemple peut être écouté isolément à l'aide du bouton ▶ (voir Illustration 6), avec la possibilité d'accéder à un contexte plus étendu dans « Extrait » (« Tout » donnant accès à l'intégralité de l'émission). Tous les traits langagiers recherchables via l'interface donnent accès à des exemples sonores écoutables et transcrits, leur nombre pouvant varier selon le phénomène et sa fréquence d'apparition au sein de notre corpus. Ainsi, grâce à cette base de données, l'apprenant dispose-t-il de la possibilité unique qu'offre l'informatique d'écouter une multitude d'exemples sonores similaires afin de nourrir son intuition de la langue cible et mieux se préparer à la compréhension du français oral quotidien. Ardents défenseurs de l'apprentissage sur corpus, Boulton & Cobb relèvent en effet :

DDL [data-driven learning] further provides access to the massive amounts of authentic language needed [...] but, crucially, it organizes it to make patterns salient, as is necessary for noticing. There is also increasing psycholinguistic evidence for the

importance of chunking, which DDL helps to highlight. (Boulton & Cobb 2017 : 350-351)

Si notre objectif majeur a consisté, tout au long de la conception, à construire une interface conviviale permettant à l'utilisateur d'accéder facilement à des exemples sonores, un tel accomplissement n'est devenu possible qu'à l'issue d'un travail considérable en amont dans l'élaboration de l'architecture informatique, complexe, de l'interface administrative, que nous n'aborderons pas ici. La question sera désormais de savoir comment les apprenants s'approprient ce nouvel outil, et les bénéfices qu'ils en tireront.

3. CONCLUSION

On peut par exemple envisager qu'en dehors des avantages que procurerait la possibilité d'accéder facilement à de nombreux exemples sonores authentiques, l'apprenant puisse également prendre conscience de la « fragilité » du locuteur natif (qui hésite et se trompe lui aussi) » (Paternostro 2017 : 284). L'authenticité des extraits disponibles via l'interface contraste en effet avec la « perfection » des dialogues fabriqués des manuels de FLE (voir note 15), qui non seulement ne préparent guère l'apprenant à la compréhension des usages réels de la langue, mais pourraient en définitive s'avérer contreproductifs :

Learners in the classroom are given the impression that spoken discourse is neat and tidy, with interlocutors who say exactly what they intended to say, and nothing more. It gives a model of language which is both unrealistic and unattainable, and might serve to demoralise students who feel they will never reach the lofty heights of perfect speech. And of course, they would be right, since no-one ever does. (Gilmore 2004 : 368)

Notre base de données incitera-t-elle les apprenants à davantage prêter attention au français oral quotidien ? Leur permettra-t-elle d'améliorer leur compréhension orale et éviter les écueils relevés dans la transcription de l'énoncé 1 ? De nombreux défis restent encore à relever. Le projet étant encore à ses débuts, il faudra observer comment les apprenants s'approprient l'interface, qui sera progressivement enrichie de fiches explicatives sur les traits langagiers, de parcours pédagogiques, et d'exercices clés en main. À défaut, il est probable que ce projet ne rencontre guère son public. Souvenons-nous de l'enthousiasme de Sinclair (1997 : 30) qui, mu par les possibilités qu'offraient les nouvelles technologies, écrivait il y a plus de vingt ans : « Quite young learners [...] will become self-taught data-driven learning students. [...], corpus data will become central to the daily concerns of language teachers »...

Christian SURCOUF

*Maitre d'enseignement et de recherche
Université de Lausanne, École de français langue étrangère*

RÉFÉRENCES

- BARBÉRIS, J.-M. 2010. « “Quand t’es super bobo”... La deuxième personne générique dans le français parisien des jeunes », *Congrès Mondial de Linguistique Française - CMLF 2010*, Neveu, F. *et al.* (Eds.), 1839-1857.
- BOULANGER, C., ETIENNE, R., HEDDESHEIMER, C., HOLEC, HENRI, MULLER, M. S., ROUSSEL, F., TOWEL, R. & ZOPPIS, C. 1972. « Le discours oral », *Mélanges pédagogiques*, p. 1-14.
- BOULTON, A. & TYNE, H. 2014. *Des documents authentiques aux corpus. Démarches pour l'apprentissage des langues*, Paris : Didier.
- BOULTON, A. 2017. « Corpora in language teaching and learning », *Language Teaching*, 50-4, p. 483-506.
- BOULTON, A. & COBB T. 2017. « Corpus Use in Language Learning: A Meta-Analysis », *Language Learning*, 67-2, p. 348-393.
- CHAMBERS, A. 2009. « Les corpus oraux en français langue étrangère: authenticité et pédagogie », *Mélanges Crapel*, 31, p. 15-33.
- CUQ, J.-P. & GRUCA, I. 2002. *Cours de didactique du français langue étrangère et seconde*, Grenoble : Presses Universitaires de Grenoble.
- CUTLER, A. 2012. *Native Listening: Language Experience and the Recognition of Spoken Words*, Cambridge : MIT Press.
- DEBAISIEUX, J.-M. & BOULTON, A. 2007. « Alors la question c’est... ? Questions pragmatiques et annotation pédagogique des corpus », *Cahiers de l’AFLS*, 13-2, p. 31-59.
- DEBAISIEUX, J.-M. 2009. « Des documents authentiques oraux aux corpus: un défi pour la didactique du FLE », *Mélanges Crapel*, 31, p. 35-56.
- DELATOUR, Y., JENNEPIN, D., LÉON-DUFOUR, M., MATTÉLÉ-YEGANEH, A. & TEYSSIER, B. 1991. *Grammaire du français*, Paris : Hachette FLE.
- DETEY, S., LYCHE, C., TCHOBANOV, A., DURAND, J. & LAKS, B. 2009. « Ressources phonologiques au service de la didactique de l’oral : le projet PFC-EF », *Mélanges Crapel*, 31, p. 223-236.
- DUBOIS, M., KAMBER, A. & DEKENS SKUPIEN, C. 2010. « L’exploitation didactique des documents authentiques audio et vidéo dans l’enseignement des langues étrangères », *Bulletin Vals-Asla*, 92, p. 1-7.
- DURÁN, R. & McCOOL, G. 2003. « If This Is French, Then What Did I Learn in School? », *The French Review*, 77-2, p. 288-299.
- ÉTIENNE, C. & SAX, K. 2009. « Stylistic Variation in French: Bridging the Gap between Research and Textbooks », *The Modern Language Journal*, 93-4, p. 584-606.
- GADET, F. 1996. *Le français ordinaire*, Paris : Armand Colin.
- GADET, F. 2004. « Le style comme perspective sur la dynamique des langues. Introduction », *Langage et société*, 109, p. 1-8.
- GADET, F., 2007. *La variation sociale en français*, Paris : Ophrys.
- GAGNON, R., BENZITOUN, C. 2020. « Le français parlé comme objet d’enseignement ? Regards croisés d’une didacticienne et d’un linguiste », *Revue des HEP*, 25, p. 37-51.
- GILMORE, A. 2004. « A comparison of textbook and authentic interactions », *ELT Journal*, 58-4, p. 363-374.
- GIROUD, A. & SURCOUF, C. 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 », *Congrès Mondial de Linguistique Française*, 2016, p. 1-18.

- HOLEC, H. 1990. « Des documents authentiques, pour quoi faire ? », *Mélanges Crapel*, 20, p. 65-74.
- LÉON, P. 1996. *Phonétisme et prononciations du français*, Paris : Armand Colin.
- LEVELT, W. J. M. 1994. *The skill of speaking*, dans *International Perspectives on Psychological Science*. Volume I: *Leading themes*, Bertelson, P. et al. (Eds.), Hillsdale, Lawrence Erlbaum Associates, p. 89-104.
- PARPETTE, C. 2018. « Quelle relation entre discours oral naturel et document oral authentique en FLE ? », *Action Didactique*, 1, p. 18-30.
- PATERNOSTRO, R. 2017. « Peut-on enseigner la variation ? », dans Tyne, H. et al. (Eds.), *La variation en question(s). Hommages à Françoise Gadet*, Bruxelles : Peter Lang, p. 279-290.
- POISSON-QUINTON, S., HUET-OGLE, C., BOULET, R. & VERGNE-SIRIEYS, A. 2003. *Grammaire expliquée du français. Niveau débutant*, Paris : CLE International.
- RAVAZZOLO, E., TRAVERSO, V., JOUIN, É. & VIGNER, G. 2015. *Interactions, dialogues, conversations : l'oral en français langue étrangère*, Paris : Hachette.
- SINCLAIR, J. 1997. « Corpus evidence in language description », dans Wichmann, A. et al. (Eds.), *Teaching and language corpora*, London : Routledge, p. 27-39.
- SURCOUF, C. & GIROUD, A. 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*, 78-4, p. 11-27.
- SURCOUF, C. & AUSONI, A. 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)*, 31, p. 71-91.
- TRIBBLE, C. 2013. « Concordancing », dans Chapelle, C. A. (Ed.), *The Encyclopedia of Applied Linguistics*, Chichester : Wiley-Blackwell, p. 2-10.
- TYNE, H. 2009. « Style in L2 : The Icing on the Cake? », dans Labeau, E. & Myles, F. (Eds.), *The Advanced Learner Variety: The Case of French*, Bern : Peter Lang, p. 243-268.
- TYNE, H. 2012. « La variation dans l'enseignement-apprentissage d'une langue 2 », *Le français aujourd'hui*, 176, p. 103-112.
- VANDERGRIFT, L. 2007. « Recent developments in second and foreign language listening comprehension research », *Language Teaching*, 40-03, p. 191-210.
- VANDERGRIFT, L. & GOH C. C. M. 2012. *Teaching and learning second language listening: Metacognition in action*, New York : Routledge.
- VIALLETON, É. & LEWIS, T. 2014. « Reconsidering the authenticity of speech in French language teaching: theory, data, methodology, and practice », dans TYNE, H. et al. (Eds.), *French through Corpora: Ecological and Data-Driven Perspectives in French Language Studies*, Newcastle upon Tyne, Cambridge : Scholars Publishing, p. 293-316.
- WEBER, C. 2006. « Pourquoi les Français ne parlent-ils pas comme je l'ai appris ? », *Le français dans le monde*, 345, p. 31-33.
- WEBER, C. 2010. « Quelle Place pour la variation et l'oralité dans l'enseignement du français ? », dans Bertrand, O. & Schaffner I. (Eds.), *Quel français enseigner ? La question de la norme dans l'enseignement/l'apprentissage*, Paris : Éditions de l'École Polytechnique, p. 169-184.

Les corpus parlés et leur didactisation :
quelle parole (ap)prise dans l'espace de la classe?

Roxane GAGNON et Véronique BOURHIS

Pour un français parlé de référence à partir de corpus
Christophe BENZITOUN

D'une parole individuelle
à une dynamique argumentative collective

Sylvie PLANE

S'approprier les constructions des verbes *se souvenir* et *oublier* :
prendre en compte la parole des élèves allophones

Marie-Noëlle ROUBAUD et Marie-Laure ELALOUF

Le synopsis comme annotation d'un corpus filmique.

De l'objet enseigné à l'objet de recherche

Véronique BOURHIS

Des corpus vidéo pour (re)construire l'intention
de la parole multimodale de l'enseignant

Brahim AZAOUI

Des corpus pour travailler la compréhension de l'oral

Jean-François DE PIETRO,

Roxane GAGNON et Christian REHM

Didactiser des corpus de documents authentiques
pour enseigner à interagir en langue étrangère

Élodie OURSEL

Les enjeux de la compréhension du français oral quotidien
en FLE : création d'une base de données de français parlé annoté

Christian SURCOUF



ISBN 978-2-252-04508-4
ISSN 0071-190X

éla

59^e année