Linx

Revue des linguistes de l'université Paris X Nanterre

13 | 2020

Collocations et traditions discursives

Mettre en place une méthodologie pour l'étude de phénomènes morphophonologiques rares : l'exemple du pataquès

Gasparde Coutanson

Édition électronique

URL : <http://journals.openedition.org/linx/4066>

DOI : 10.4000/linx.4066

ISSN : 2118-9692

Éditeur

Presses universitaires de Paris Nanterre

Référence électronique

Gasparde Coutanson, « Mettre en place une méthodologie pour l'étude de phénomènes morphophonologiques rares : l'exemple du pataquès », *Linx* [En ligne], 13 | 2020, mis en ligne le 30 mai 2020, consulté le 01 juin 2020. URL : <http://journals.openedition.org/linx/4066> ; DOI : <https://doi.org/10.4000/linx.4066>

Ce document a été généré automatiquement le 1 juin 2020.

Département de Sciences du langage, Université Paris Ouest

Mettre en place une méthodologie pour l'étude de phénomènes morphophonologiques rares : l'exemple du pataquès

Gasparde Coutanson

- ¹ Nous nous interrogeons ici sur la méthodologie à adopter pour travailler sur des phénomènes linguistiques rares en corpus, plus précisément morphophonologiques. En effet, nous nous intéressons dans le cadre de nos recherches au pataquès, un phénomène morphophonologique produit de manière sporadique notamment en français et qui est apparenté au phénomène de liaison, un sandhi externe consonantique également propre entre autres au français. Crystal définit le terme sandhi de la sorte : “phonological modification of grammatical forms which have been juxtaposed” (Crystal 2008 : 422). On distingue souvent le sandhi interne, un sandhi qui se produit entre des morphèmes à l'intérieur de l'unité mot, du sandhi externe, un sandhi qui se produit à la frontière entre des mots. On peut trouver des sandhis tonaux, vocaliques ou encore consonantiques, et ce sont ces derniers qui constituent ici notre objet d'étude. Le terme de liaison correspond pour Bürki et Laganaro à un « phénomène selon lequel certains mots du français ont deux réalisations (ou variantes), la seconde différant de la première par la présence d'une consonne additionnelle dans certains contextes » (Bürki & Laganaro 2014 : 169). Ainsi, on trouvera en français *des gens* mais *des [z] enfants* devant initiale vocalique. Il existe des cas, par exemple dans *huit [z] enfants*, où la consonne graphique du mot précédant la liaison (ici, <t>) ne correspond pas au segment consonantique de liaison réellement produit (ici, [z]). On peut nommer ces cas *pataquès*. Comment étudier de tels phénomènes sur corpus en linguistique? Nous avons adopté une méthodologie en quatre temps (voir Figure 1), qui sera développée point par point par la suite. Pour commencer, il semble nécessaire de réaliser un va-et-vient entre état de l'art et collection d'exemples attestés. On peut ensuite sélectionner un ou des corpus disponibles et variés qui nous semblent pertinents pour l'étude de notre phénomène, qu'on carottera. Si un ou des corpus nous semblent intéressants

mais font défaut, on peut passer par une étape facultative de constitution de ces nouveaux corpus, qui seront de même carottés. Enfin, il convient d'analyser plus finement les corpus qui se seront révélés les plus pertinents.

Figure 1 : Méthodologie adoptée pour étudier des phénomènes morphophonologiques rares en corpus

1. Première étape : va-et-vient entre état de l'art et collection d'exemples attestés

- 2 Dans le cadre d'une thèse, cette première étape de va-et-vient entre état de l'art et collection d'exemples attestés peut être appréhendée dès un mémoire de master. Pour le chercheur confirmé, cette étape essentielle ne doit pas être négligée. En effet, ce va-et-vient paraît indispensable car de la réflexion théorique dépendra le rattachement des exemples au phénomène étudié tout comme les exemples rencontrés viendront nourrir l'acceptation que l'on peut se faire du phénomène en question.
- 3 Au cours de cette étape, une des premières difficultés à laquelle est confronté le linguiste dans sa revue de littérature d'un phénomène rare est d'accéder aux diverses appellations qui ont pu lui être données. En ce qui nous concerne (voir Figure 2), on peut voir que ces termes sont pléthore¹. Ici, la taille relative des termes dans le nuage de mots est liée au nombre d'écrits scientifiques utilisant le terme en question, leur couleur dépend quant à elle de leur catégorisation comme hyperonymes seulement (noir), hyperonymes ou hyponymes (gris médian) ou hyponymes seulement (gris clair). Les items les plus communément employés dans la communauté linguistique pour dénommer le phénomène en question pourraient, notamment pour des phénomènes peu étudiés, nous renseigner sur la transmission des termes linguistiques entre les différents auteurs s'y étant intéressés, ce qui, couplé aux références bibliographiques citées par les auteurs, permettrait de tisser un véritable réseau d'influence. Il est important de distinguer hyperonyme d'hyponyme car cela nous informe sur l'objet de travail des chercheurs précédents, qui peuvent avoir étudié le phénomène dans son ensemble ou en partie seulement.

Figure 2 : Termes utilisés pour dénommer le phénomène en question selon le nombre d'écrits scientifiques utilisant le terme

- 4 Peut-on aller plus loin qu'une collection d'exemples pour étudier un phénomène rare ? Jusqu'à l'avènement d'une linguistique sur corpus informatisée, les linguistes utilisant des données écologiques travaillaient à partir de collections d'exemples attestés (entendus ou lus). Gilquin et Gries rangent notamment dans cette catégorie de linguistes les comparatistes européens et les structuralistes américains (Gilquin & Gries 2009 : 2). Selon de Fornel et Verdier, une collection d'exemples constitue un « agrégat de cas singuliers » (de Fornel et Verdier 2018 : 5). On réunit donc des occurrences d'un même phénomène disjointes temporellement ou spatialement qui peuvent être produites par divers locuteurs dans des contextes communicatifs variés. Ces collections d'exemples ont pu être dépassées grâce aux avancées technologiques aussi bien pour les corpus oraux que les corpus écrits, que ce soit au niveau de leur collecte (enregistrement des données orales, ocrisation ou glanage des données écrites) comme de leur gestion (transcription, annotation, fouille outillée des données orales, annotation et fouille outillée des données écrites). Pour l'oral, O'Keeffe & McCarthy soulignent en particulier l'importance de la miniaturisation de la technologie utilisée pour la collecte de données (O'Keeffe & McCarthy 2010 : 5). Il est vrai que l'introduction de la cassette audio compacte dans les années 1960 a révolutionné le travail du collecteur en offrant à la fois un gain de place et en favorisant la mobilité du chercheur par rapport aux bandes magnétiques utilisées au préalable tout comme la miniaturisation progressive des enregistreurs digitaux et caméscopes ont à nouveau grandement facilité le travail de collecte de données du chercheur. Pour ce qui est de la gestion des données, les auteurs notent que "it was the revolution in hardware and software in the 1980s and 1990s which really allowed corpus linguistics as we know it to emerge" et soulignent notamment les améliorations de la puissance de calcul des ordinateurs ainsi que de la vitesse de téléchargement des données comme décisives pour la constitution et la diffusion des corpus. L'utilisation de corpus pour étudier des phénomènes rares permet, outre l'analyse du phénomène qui pourrait être réalisée également à partir d'une collection d'exemples, à la fois de vérifier la présence de ces

occurrences rares, de balayer l'ensemble du corpus à leur recherche ou encore de se faire une idée statistique de la fréquence du phénomène afin de confirmer leur rareté. Pourtant, certains linguistes hésitent à travailler sur des phénomènes rares en corpus, notamment pour les phénomènes oraux, et ne souhaitent pas dépasser la collection d'exemples, comme on peut le voir dans la citation suivante de Cappeau et Gadet en lien avec notre objet d'étude :

Faudra-t-il dès lors récuser toute autre méthode d'approche de données, comme la prise de notes à la volée ? Pourtant, y a-t-il d'autres moyens d'accéder à des phénomènes dont on ne peut pas savoir à l'avance si et quand ils vont émerger ? Ainsi [...] d'erreurs de liaison, etc. que l'on rencontre en trop faible quantité dans les corpus oraux, qui deviennent finalement trop petits ! (Cappeau & Gadet 2007 : 102)

- 5 En effet, le problème majeur est celui de la taille des corpus. Travailler sur un phénomène rare implique pour des questions de significativité statistique d'utiliser des corpus massifs, ce qui peut demander un travail colossal selon l'état d'avancement du corpus (état de transcription pour les corpus oraux, de repérage et d'annotation du phénomène recherché...), ce sur quoi nous reviendrons par la suite, et qui semble plutôt démesuré par rapport à la rareté du phénomène. Néanmoins, si travailler seulement à partir d'une collection d'exemples présente l'avantage de permettre de constituer une première typologie du phénomène en question, et s'il est clair qu'un travail sur corpus permettra potentiellement de mettre à jour des occurrences prototypiques comme moins prototypiques du phénomène recherché, une collection d'exemples a le désavantage majeur de ne pas mettre en lumière tous les contextes dans lesquels le phénomène étudié n'apparaît pas, pour cela le passage par une analyse sur corpus semble indispensable pour comprendre le fonctionnement du phénomène, d'où les trois étapes suivantes, même si les données ne sont pas assez massives pour permettre une analyse quantitative.

2. Deuxième étape : choix et carottage de corpus disponibles variés

- 6 Pour le linguiste de corpus se pose toujours la question du choix des corpus de travail et dans un premier temps du type de corpus nécessaire. Pour commencer, la langue dans laquelle on étudiera le phénomène orientera notre recherche de corpus vers des corpus de cette langue (il peut parfois y avoir plusieurs langues d'étude), mais on restera ouvert aux corpus de langues génétiquement ou typologiquement proches qui pourraient apporter un éclairage nouveau au phénomène analysé. Pour notre étude du pataquès nous sélectionnerons ainsi des corpus en français, langue phare du phénomène associé au pataquès, la liaison, mais nous resterons ouverts à des corpus notamment en francoprovençal, une langue génétiquement proche où on trouve aussi de la liaison. Suite à notre état de l'art, nous avons aussi eu connaissance de langues telles que l'anglais ou le russe dans lesquelles s'observent à la fois des sandhis externes consonantiques historiques et leurs pendants non historiques. Dans l'état actuel de notre recherche, nous nous limiterons à la première étape de notre méthodologie pour ces cas, c'est-à-dire une combinaison entre état de l'art et collection d'exemples attestés. Ainsi, nous ne sélectionnerons pas de corpus pour étudier ces phénomènes, excepté pour le recours à des exemples attestés. On peut opter ensuite au besoin pour un corpus synchronique ou un corpus diachronique, cependant, pour notre recherche,

même si nous préférons un corpus synchronique, un corpus diachronique pourrait également se révéler intéressant. On peut encore choisir un corpus parlé, chanté ou écrit selon qu'on travaille en linguistique de l'oral ou en linguistique de l'écrit. Les premiers peuvent être seulement audio ou seulement vidéo (notamment pour les langues des signes) ou multimodaux (audio et vidéo). Enfin, on peut souhaiter travailler à partir de données primaires ou à partir de données secondaires. Pour l'oral par exemple, on parle de données primaires pour les enregistrements en opposition aux données secondaires qui sont des données de seconde main telles que des transcriptions ou annotations. Les premières sont vues comme plus proches de la production réelle des locuteurs alors que les secondes comprennent déjà une part d'interprétation. Les données primaires sont toujours de meilleure qualité pour le chercheur, mais face à des contraintes temporelles ou si les données secondaires découlent d'une méthodologie partagée ou faisant référence dans un milieu donné, et si l'annotation existante est réutilisable pour travailler sur le phénomène qu'il analyse, le chercheur peut se contenter de données secondaires. L'étude du pataquès nous amènera à favoriser des données primaires, car l'étape de carottage des corpus existants annotés pour la liaison (abordée par la suite) révèle que ceux-ci ne sont pas sans sous-catégorisation (ou silence : oublis) et sur-catégorisation (ou bruit : l'occurrence n'a pas été prononcée, elle correspond à une liaison typique et non à un pataquès ou n'est ni de la liaison, ni du pataquès) d'occurrences comme pataquès. Ces sur- et sous-catégorisations des transcrip-teurs ou annotateurs constituent un risque majeur pour l'étude de phénomènes rares, en effet la rareté des phénomènes ne permet pas de maintenir une attention constante de la part des annotateurs, au risque de sous-catégorisations, et inversement les annotateurs se focalisant trop sur le phénomène rare recherché peuvent vite sur-catégoriser. Ce n'est pas pour autant que nous excluons les données secondaires car si celles-ci existent nous les retravaillerons en repartant des données primaires pour un gain de temps.

- 7 Une fois le type de corpus recherché en tête, il est temps de se demander comment y accéder. La question du signalement des corpus est essentielle, car la disponibilité de corpus (constitués, transcrits voire annotés) et leur réutilisation permet un gain de temps considérable pour le chercheur. Sur un continuum, on pourrait distinguer quatre cas. Les corpus qui semblent les plus facilement utilisables sont les corpus scientifiques référencés et librement accessibles en ligne (notons que certains corpus sont partiellement libres d'accès, mais l'accès à l'ensemble des données peut parfois être autorisé au chercheur qui en fait la demande). Pour le français², on peut en trouver à trois niveaux : au niveau international, européen et français. Les différents corpus peuvent être répertoriés à plusieurs niveaux mais ne le sont pas systématiquement. On trouvera notamment au niveau international de nombreux corpus répertoriés au sein de l'Open Language Archives³, au niveau européen dans le Virtual Language Observatory de CLARIN (Common Language Resources and Technology Infrastructure)⁴ et au niveau français sur Ortolang (Outils et Ressources pour un Traitement Optimisé de la LANGue)⁵ ou encore COCOON (Collections de Corpus Oraux Numériques)⁶ pour des corpus seulement audio. Ces corpus qui semblent les plus facilement accessibles et réutilisables comprennent notamment les corpus les plus répandus dans le champ académique, ce qui implique de vérifier que l'analyse que l'on veut mener n'a pas déjà été auparavant réalisée, mais permet une réutilisation de corpus répandus parmi les chercheurs, corpus de référence pour lesquels il ne sera pas nécessaire de trop détailler la méthodologie de constitution, de transcription et d'annotation. Des corpus similaires

mais moins accessibles en raison d'une méconnaissance des corpus disponibles dans d'autres domaines de recherche liée au manque d'interdisciplinarité qui a pu et peut encore exister au sein du champ académique, sont les corpus issus d'autres disciplines ou encore d'institutions patrimoniales. Par exemple, la Base inter-régionale du Patrimoine oral (BIPO)⁷ est une banque de données contenant des archives notamment ethnomusicologiques qui pourraient intéresser les linguistes s'intéressant à la langue chantée. La BNF (Bibliothèque Nationale de France) dispose elle aussi d'archives orales qui pourraient être exploitées par les linguistes. Une part importante de ces ressources est accessible via le portail Gallica⁸. Encore un peu moins faciles d'accès, des corpus encore non compilés voire même non recueillis feront l'objet de notre point suivant. En 2000, Habert répertorie les variétés de français sous-représentées : le français non hexagonal, le français scientifique ou technique et le français non standard contrairement à celles sur-représentées : le français d'hier (valable seulement pour l'écrit en raison des difficultés juridiques à compiler des corpus à partir de textes récents), le français écrit et le français suivant les règles du bon usage (Habert 2000 : 4). Même si la situation a certainement évolué, certaines tendances se confirment, comme la masse de données écrites par rapport aux données orales (qui sont plus longues à traiter car nécessitant une transcription préalable qui ne peut être automatisée), et d'autres semblent déjà s'être inversées, comme la prévalence de textes anciens sur les textes récents en raison des changements juridiques liés aux données de la recherche. C'est bien souvent pour ces variétés de français sous-représentées qu'il faudra compiler voire recueillir de nouveaux corpus. Pour finir, les corpus qui paraissent les plus compliqués à atteindre sont des corpus non référencés, notamment les corpus produits par des chercheurs mais non partagés à la communauté scientifique ou restreints à un petit groupe de recherche, voire privés (par exemple des enregistrements familiaux, des comptes-rendus associatifs...). L'absence de répertoriage de ces corpus en fait des corpus fantômes dont les chercheurs ignorent l'existence, même si celle-ci peut être supposée et mener à tenter d'accéder à ces données, demande qui n'est pas sûre d'aboutir. Les corpus non issus de la discipline linguistique peuvent en outre présenter des divergences méthodologiques (l'absence de métadonnées des documents sources peut en être une) qui pourront les rendre inutilisables selon les besoins du chercheur. Enfin, ces données peuvent nécessiter un temps de travail considérable avant de pouvoir être fouillées (par exemple, numérisation puis transcription des supports audio et vidéo, océrisation des textes, annotation des données transcrites ou océrisées).

- 8 Nous avons sélectionné différents corpus pour l'étude du pataquès. Le premier est le fonds Jean Dumas, un corpus ethnomusicologique provenant de la BIPO⁹ et constitué de chansons traditionnelles francophones enregistrées dans les années 1960 dans l'actuelle région Auvergne-Rhône-Alpes (départements de la Haute-Loire, de la Loire et du Puy-de-Dôme) par un collecteur nommé Jean Dumas. Le second est le corpus Acquisition de la Liaison et Interaction Parents Enfants (ALIPE)¹⁰, un corpus acquisitionnel interactionnel de français recueilli dans les années 2010 (Laboratoire de Recherche sur le Langage, Université Clermont Auvergne : voir Chabanal, Liégeois & Chanier 2015 ainsi que LRL, 2017 pour la ressource). Le troisième est l'Atlas Linguistique Audiovisuel du francoprovençal VALaisan (ALAVAL)¹¹, composé de données semi-spontanées produites par des locuteurs de francoprovençal du canton du Valais en Suisse romande et développé entre les années 1990 et 2000 (Centre de dialectologie et d'étude du français régional de l'Université de Neuchâtel : voir Diémoz & Kristol, 2018). Le dernier corpus est le corpus PFC, un corpus variationniste issu du projet Phonologie

du Français Contemporain¹² initié par Jacques Durand (Université de Toulouse Jean Jaurès), Bernard Laks (Université Paris Nanterre) & Chantal Lyche (Université d'Oslo) et enregistré depuis les années 2000 dans toute la Francophonie : France métropolitaine, Europe francophone, France d'outre-mer, Afrique du Nord, Amérique du Nord, Afrique subsaharienne (voir Durand, Laks & Lyche, 2009 et MoDyCo & RUG, 2017 pour la ressource). Un autre corpus, le fonds Charles Joisten¹³, issu de la BIPO a aussi été carotté pour ce qui est des chansons traditionnelles, néanmoins, étant redondant avec le fonds Jean Dumas il n'a pas été utilisé par la suite et nous n'y ferons donc plus allusion. Les quatre corpus sélectionnés diffèrent surtout par les diverses populations auprès desquelles ils ont été enregistrés.

- 9 Nous avons déjà entraperçu une certaine variation diatopique dans le lieu de vie des locuteurs enregistrés : si PFC veut être représentatif de toute la Francophonie, les trois autres corpus se limitent à des zones géographiques plus restreintes. Si on prend à présent l'exemple de l'âge ou du sexe des locuteurs, on peut noter de grandes divergences (voir Figure 3 et Figure 4). Pour le fonds Jean Dumas & le corpus de l'ALAVAl, les locuteurs sont assez âgés. Il est à noter pour l'ALAVAl que le francoprovençal est une langue en danger (LED), cas dans lequel les linguistes se tournent souvent vers les locuteurs les plus âgés. Bert & Grinevald expliquent ce comportement de la sorte : « [...] dans les situations de LED, [...] l'âge et le niveau de compétences sont presque toujours corrélés, même si l'on peut observer des exceptions individuelles » (Bert & Grinevald 2010 : 126). C'est la même logique qui a poussé Jean Dumas à collecter des chansons auprès de locuteurs-chanteurs âgés. Le corpus acquisitionnel, par nature, est quant à lui composé de très jeunes locuteurs (sans oublier les parents de ce corpus interactionnel, mais la priorité sera mise sur les productions enfantines du corpus). Le corpus PFC est plutôt équilibré même s'il comprend beaucoup de jeunes dans la vingtaine et presque aucun enfant. Il faut noter que les corpus n'ayant pas été enregistrés à la même époque (voir Figure 3), l'âge relatif des sujets varie, un sujet de quarante ans en 1960 du fonds Jean Dumas par exemple serait comparable à un sujet de quatre-vingt-dix ans en 2010 du corpus PFC (par rapport à leurs dates de naissance). Pour ce qui est du sexe des locuteurs, l'ALAVAl est quasiment équilibré hommes-femmes dans une optique d'échantillonnage sociolinguistique représentatif de la population étudiée (un homme et une femme ont été enregistrés pour chaque point d'enquête, pour un seul village il n'y avait plus de locutrice, deux hommes ont alors été enregistrés). Les trois autres corpus comprennent plus de femmes que d'hommes (pour l'ALIPE, si on s'intéresse principalement aux enfants).

Figure 3: Âge et sexe des locuteurs dans les différents corpus sélectionnés pour l'étude du pataqués

- 10 La Figure 3 met à jour également le déséquilibre numérique de l'échantillon de locuteurs retenus pour représenter une population donnée : si le corpus ALIBE n'admet que trois locuteurs enfants (ce qui est habituel pour un corpus acquisitionnel longitudinal dense) et six locuteurs adultes, le fonds Jean Dumas présente quarante-huit locuteurs et l'ALAVALL cinquante, alors que PFC répertoriait quatre cent trente-quatre locuteurs lors de la rédaction de ces lignes. Rappelons que plus l'échantillon de locuteurs retenus pour représenter une population donnée est large, plus l'échantillon est censé être représentatif de cette population.

Figure 4 : Âge des enfants de l'ALIBE à chaque session d'enregistrement

- 11 Après une réflexion sur le type de corpus nécessaire à l'étude du phénomène étudié, sur la façon d'accéder à des corpus disponibles, après une sélection de corpus variés notamment en raison de leurs diverses populations d'étude associées, il est désormais temps de passer à une phase de carottage de ces corpus. Cette notion de carottage peut

être appréhendée comme un échantillonnage non représentatif (à l'opposé de l'échantillonnage qui se doit d'être représentatif). Le procédé est simple : on opte pour une durée de carottage qui sera la même pour tous les corpus. Nous avons choisi pour notre étude du pataquès d'adopter une durée de carottage de vingt minutes, ce qui nous semblait assez conséquent pour se faire une première idée quantitative de la présence du phénomène dans les corpus sélectionnés.

- 12 Pour l'ALAVAl, seules dix minutes ont été carottées, mais une comparaison reste possible avec les résultats des autres carottages. Le Tableau 1 présente par ordre croissant les résultats obtenus dans les différentes carottes des corpus, soit le nombre de pataquès obtenus par carotte selon la durée de carottage du corpus. On part systématiquement des données primaires, et, en s'aidant des données secondaires si elles existent, on liste et comptabilise les occurrences rencontrées du phénomène rare recherché. La phase de carottage prend du temps mais ne doit pas être chronophage et plutôt orienter le chercheur plus rapidement vers des corpus pertinents. C'est pour cela que la durée de carottage doit rester limitée, même si les carottages s'orienteront ainsi plutôt vers une analyse quantitative des données (voir si les corpus présentent le phénomène en question, et si oui dans quelle mesure, notons au passage que l'absence de données en elle-même est significative) que vers une analyse qualitative des données, analyse qui aura partiellement pu être faite lors de la première étape de va-et-vient entre état de l'art et collection d'exemples attestés et qui le sera à nouveau dans la dernière étape au cours de l'analyse plus fine des corpus les plus pertinents. C'est le corpus de chansons traditionnelles, le fonds Jean Dumas qui présente le plus de pataquès lors du carottage (dix occurrences en vingt minutes), suivi de près par l'atlas linguistique de francoprovençal (quatre occurrences en dix minutes) et dépassant de loin le corpus variationniste PFC (deux occurrences en vingt minutes) ainsi que le corpus acquisitionnel ALIPE (une occurrence en vingt minutes). Il est bon de garder à l'esprit que la carotte n'est pas forcément représentative du corpus dont elle est issue et que ce n'est qu'une analyse plus poussée qui pourra confirmer ou infirmer ces résultats. Néanmoins, tous les corpus présentent dès l'état de carottage des pataquès, ce qui constitue déjà une indication très intéressante.

	ALIPE	PFC	ALAVAl	Jean Dumas
Durée de carottage	20 min	20 min	10 min	20 min
Nombre d'occurrences	1	2	4	10

Tableau 1 : Nombre de pataquès par carotte (ordre croissant)

- 13 La dernière phase de cette deuxième étape est une potentielle première fouille des données qui viendra compléter les résultats obtenus pour le carottage. Celle-ci ne peut s'effectuer qu'à condition que des critères spécifiques soient remplis : pour des données orales, le corpus doit être transcrit, pour des données orales ou écrites, le corpus doit en outre être annoté pour le phénomène recherché à l'aide d'un codage pouvant être fouillé par un outil d'exploration existant. Pour notre cas, les critères à valider sont listés dans le Tableau 2.

Corpus	Jean Dumas	ALAVAL	PFC	ALIPE
Transcrit	Non	Oui	Oui	Oui
Codé pour la liaison	Non	Non	Oui	Oui
Outil d'exploration ¹⁴	Non	Non	Oui	Oui

Tableau 2: Critères de sélection des corpus pouvant bénéficier d'une première fouille de leurs données

- 14 Les corpus répondant à tous les critères sont PFC et ALIPE. Deux analyses quantitatives vont pouvoir être effectuées lors de cette première fouille des corpus pour ce qui est du pataquès : tout d'abord une analyse du pourcentage de locuteurs produisant des pataquès dans les corpus (Tableau 3), ensuite du pourcentage de pataquès réalisés pour chaque corpus (Tableau 4). Précisons que les résultats peuvent être vérifiés pour ce qui est de la sur-catégorisation d'occurrences par les annotateurs précédents comme appartenant au phénomène recherché, ce qui n'est pas le cas pour les sous-catégorisations qui demanderaient un passage en revue de l'ensemble du corpus.

PFC						ALIPE					
Lecture		Conversation		Total		Parents		Enfants		Total	
%	n	%	n	%	n	%	n	%	n	%	n
6,86	28/408	4,57	19/416	10,24	43/420	50	3/6	100	3/3	66,66	6/9

Tableau 3: Première fouille des corpus PFC et ALIPE : taux de locuteurs produisant des pataquès

- 15 Comme le mentionne le Tableau 3, une des spécificités du corpus PFC est de disposer à la fois de données de lecture et de données conversationnelles de par son protocole réparti entre ces diverses tâches (seront regroupées ici les conversations libres et les conversations guidées). On peut constater en observant les résultats présentés dans le Tableau 3 qu'un corpus comme PFC où chaque locuteur est enregistré une cinquantaine de minutes en conversation (libre et guidée) plus tâches de lecture se distingue d'un corpus dense comme celui de l'ALIPE où les locuteurs sont enregistrés plusieurs heures chacun (notons au passage que seules dix minutes de conversation par locuteur sont annotées pour la liaison dans le corpus PFC). En effet, cette première fouille des données fait ressortir que deux tiers des locuteurs de l'ALIPE, parents compris, produisent au moins un pataquès, chiffre largement supérieur aux quelques 4,57% de locuteurs de PFC produisant des pataquès en conversation, modalité d'enregistrement comparable à celle de l'ALIPE. On pourrait émettre l'hypothèse que tous les locuteurs francophones produisent des pataquès, mais que ceux-ci ne sont visibles en corpus que dans des corpus denses.

PFC						ALIPE					
Lecture		Conversation		Total		Enfants		Parents		Total	
%	n	%	n	%	n	%	n	%	n	%	n
0,24	29/12296	0,06	25/41946	0,10	54/54242	2,61	60/2296	0,19 ¹⁵	10/5302	0,92	70/7598

Tableau 4 : Première fouille du corpus : part des pataquès par corpus

- 16 En comparant à présent le taux de pataquès dans les deux corpus Tableau 4, il semblerait que ce soit le corpus acquisitionnel ALIPE qui en concentre le plus (0,92% contre 0,06% en conversation, modalité comparable dans PFC). Ces taux très faibles nous confortent dans l'idée que le phénomène étudié est rare, en tout cas dans ces corpus selon les annotations existantes (vérifiées simplement pour la sur-annotation). Cette différence est due au fait que dans le corpus ALIPE, les enfants, catégorie de locuteurs absente de PFC, produisent plus de pataquès que leurs parents (2,61% contre 0,19%). On note également que dans PFC, c'est la tâche de lecture qui présente le plus de réalisations de pataquès (0,24% contre 0,06%).

3. Troisième étape : constitution de nouveaux corpus

- 17 La constitution de nouveaux corpus est une solution possible suite à la deuxième étape de choix et carottage de corpus disponibles variés si les corpus existants ou accessibles ne sont pas adaptés pour l'étude que l'on veut mener et peut s'effectuer de deux façons. La première est la moins impliquante pour le chercheur et consiste à compiler des données déjà existantes. Si notre objet d'étude est l'écrit, on pourra par exemple se tourner vers des copies d'étudiants, des articles de presse ou scientifiques, des livres, des comptes-rendus de réunion, des lettres, SMS, mails, posts sur les réseaux sociaux, commentaires suite à des publications sur internet, messages instantanés sur des chats, messages laissés sur des forums, pages web, textes de lois, transcriptions de discours... Au contraire, si notre objet d'étude concerne l'oral, on pourra utiliser des livres audio, des chansons, des données télévisuelles, radiophoniques, des discours d'hommes politiques, des cours enregistrés (que ce soit dans le cadre d'un MOOC (Massive Open Online Course) ou pour un podcast), des vidéos de Youtubeurs... Pour des données francophones il est possible d'avoir accès à des données orales sur des sites aussi divers que Youtube¹⁴, l'INA (Institut National de l'Audiovisuel)¹⁵ ou encore le site de l'Élysée¹⁶. Notons que l'INA possède certaines données qui ne sont accessibles que depuis ses locaux.
- 18 La seconde est d'obtenir soi-même de nouvelles données pour constituer un corpus. Les données écrites sont bien souvent préexistantes à la volonté de constituer un corpus, car intrinsèquement l'écrit, contrairement à l'oral et jusqu'à l'avènement de l'enregistrement sonore, était le médium par excellence de conservation dans le temps des productions langagières. Ainsi ce ne sont principalement que des données élicitées, c'est-à-dire provoquées par le chercheur, qui peuvent être demandées à être produites : copie d'un texte, rédaction d'un texte selon une consigne spécifique... Pour les données orales, on peut enregistrer son propre corpus. Celui-ci peut se composer de données spontanées soit monologiques, avec par exemple des récits de vie, soit dialogales, avec des interactions (médiées ou non par téléphone, ordinateur). Ce corpus peut aussi être composé de données ayant un degré de spontanéité moindre, qu'elles soient à tendance monologique avec des récitation (poésies, comptines, karaoké), des conférences ou encore des cours magistraux ou à tendance dialogale comme avec des récitation (théâtre, film). Enfin, ce corpus peut être constitué à partir de données élicitées, c'est-à-dire provoquées par le chercheur. Ces données peuvent résulter d'une tâche de lecture, de description suite à des stimuli visuels, sonores ou encore olfactifs, ou d'une tâche de traduction.

- 19 Ces différents types de données : spontanées, moins spontanées, élicitées pourraient toutes être considérées comme des corpus. C'est ce que nous proposent Gilquin et Gries dans la Figure 5 (Gilquin & Gries 2009 : 5). Dépassant une dichotomie stricte entre données écologiques et données élicitées, allant même jusqu'à remettre en cause une dichotomie entre données de corpus et données issues de méthodes expérimentales, les auteurs nous proposent un continuum novateur entre divers types de données linguistiques selon le degré du contexte écologique de leur production ou collection. L'ordonnement du continuum pourrait néanmoins être remis en cause. Par exemple, les auteurs n'explicitent pas ce que constitue pour eux une collection d'exemples (notamment si ces collections se limitent aux données écrites ou peuvent aussi prendre en compte des données orales), mais il semble surprenant selon notre conception (en particulier de la collection d'exemples oraux entendus mais non enregistrés) de placer ceci en seconde place sur le continuum. Pour l'oral, même si cette méthodologie est moins intrusive que l'enregistrement des locuteurs, la fiabilité et la représentativité des données obtenues reste discutable et de cette façon leur caractère écologique de même. Pour ce qui est du troisième point, une distinction aurait encore pu être faite pour les productions orales réalisées en contexte professionnel qui sur un tel continuum semblent plus écologiques que des productions orales enregistrées dans des sociétés/communautés ou la prise de note/l'enregistrement n'est pas particulièrement spectaculaire/invasif mais dans des situations de communication qui ne sont habituellement pas enregistrées. Ces données prises dans un cadre professionnel semblent aussi écologiques que le seraient des productions écrites réalisées dans un contexte professionnel qui elles seraient classées sous le premier point : corpus de textes écrits. Gilquin et Gries modifient néanmoins notre façon d'appréhender les données de la recherche. En effet, leur proposition de continuum entre données écologiques, prototypiquement utilisées pour constituer un corpus, et données élicitées, résultant prototypiquement de méthodes expérimentales ouvre le champ des possibles sur les données potentiellement utilisables par les

chercheurs lors de la constitution de corpus, tout en permettant une réflexion quant au caractère écologique ou non des données mobilisées.

Kinds of linguistic data (sorted according to naturalness of production/collection)	
Data source	
1. corpora with written texts (e.g., newspapers, webblogs)	
2. example collections	
3. corpora of recorded spoken language in societies/communities where note-taking/recording etc is not particularly spectacular/invasive	
4. corpora with recorded spoken language from fieldwork in societies/communities where note-taking/recording etc is spectacular/invasive	
5. data from interviews (e.g., sociolinguistic interviews)	
6. experimentation requiring subjects to do something with language they usually do anyway, e.g.,	
– sentence production as in answering questions in studies on priming	
– picture description in studies on information structure	
7. elicited data from fieldwork (e.g., responses to “how do you say X in your language?”)	
8. experimentation requiring subjects to do something with language they usually do not do, on units they usually interact with, e.g.,	
– sentence sorting	
– measurements of reaction times in lexical decision tasks	
– word associations	
9. experimentation requiring subjects to do something with language they usually do not do on units they usually interact with, involving typical linguistic output, e.g.,	
– measurements of event-related potentials evoked by viewing pictures	
– eye-movements during reading idioms	
– acceptability/grammaticality judgments on units they usually do not interact with, involving the production of linguistic output, e.g.,	
– phoneme monitoring	
– gating	
– ultrasound tongue-position videos	

Figure 5 : Proposition de classification des données linguistiques en fonction de l'aspect écologique de leurs conditions de production ou collection selon Gilquin et Gries (2009 : 5)

- 20 Il va sans dire qu'une fois les corpus manquants constitués, par compilation ou création et cette troisième étape terminée, il est nécessaire comme présenté sur le schéma initial de la méthodologie adoptée de repasser par la deuxième étape afin de carotter les corpus nouvellement retenus et d'effectuer une première fouille des nouvelles données. Dans notre cas, nous n'avons pas besoin de constituer de nouveau corpus, nous passerons donc cette étape. D'autres corpus auraient toutefois pu se révéler intéressants pour l'étude du pataquès, tels des corpus de créoles à base française, des corpus dans d'autres langues d'oïl, des corpus de locuteurs bègues, des corpus de célébrations religieuses ou de prières, des corpus de journalistes ou encore des corpus de contes traditionnels. Certains de ces corpus n'existent pas à l'heure actuelle, ou la méthodologie de constitution des corpus existants ne nous convient pas. Mais étant donné que ces corpus n'ont ni été sélectionnés lors de la seconde étape de la méthodologie que nous avons mise en place, ni été carottés, ils ne seront pas plus constitués lors de cette troisième étape. Il a en effet fallu faire des choix pour éviter un surplus de travail et nous avons préféré choisir de premiers corpus les plus pertinents possibles pour une première analyse sur corpus de notre phénomène.

4. Quatrième étape : analyse fine des corpus les plus pertinents

- 21 Une fois les corpus les plus pertinents sélectionnés, c'est-à-dire après avoir eu accès à des corpus disponibles variés et/ou une fois un ou de nouveaux corpus compilés ou créés, on peut passer à une étape d'analyse plus fine. Un nombre limité de corpus

pourra être analysé dans cette dernière étape, plus spécifiquement il faudra souvent analyser moins de corpus que ceux carottés aux étapes précédentes. On agira par ordre de priorité selon un critère de pertinence résultant à la fois des résultats du carottage et, si réalisée, de la première fouille des données effectuée, c'est-à-dire de la pertinence des corpus pour ce qui est du phénomène rare étudié, combiné à l'état d'avancement des corpus présélectionnés. Cet état d'avancement du corpus comprendra divers critères, parmi lesquels le fait de disposer d'un accès non limité aux données primaires comme secondaires, d'une transcription (pour les corpus oraux seulement), d'une annotation du phénomène étudié, voire mieux d'un codage du phénomène, qui, combiné à un outil d'exploration permettra une fouille rapide des données. Nous reprenons ici nombre de critères utilisés précédemment lors de la deuxième étape et plus précisément la première fouille des données. Notons cette fois-ci pour une analyse plus fine des corpus que la qualité de la transcription (si corpus oral) et de l'annotation existantes peut impliquer un surplus de travail considérable pour le chercheur. En outre, tous les formats des données de seconde main (transcriptions, annotations) ne sont pas accessibles de la même façon pour le chercheur en fonction de sa maîtrise des logiciels de transcription associés et peuvent nécessiter un temps de formation. Cela nous questionne quant à l'interopérabilité entre les corpus existants ou constitués car si une conversion vers un format mieux maîtrisé peut aussi être envisagée, cela ne sera pas sans conséquence sur les données obtenues (ceci sera évoqué par la suite).

- 22 Le Tableau 5 détaille les caractéristiques principales propres à chaque corpus présélectionné lors des étapes précédentes pour étudier le pataquès. Les corpus sont classés de gauche à droite du moins accessible au plus accessible. Il est évident que ces critères sont valables pour un instant donné et qu'ils peuvent être amenés à évoluer. Pour cela il faudra se tenir informé des mises à jour liées aux corpus, la situation pouvant évoluer très vite. Pour nous par exemple l'ALAVAL s'est doté en 2018 d'une interface de recherche développée pour le site de présentation mais ne permettant pas encore l'extraction des pataquès lors de la fouille des données. Étant donné que le pataquès constitue un cas de sandhi externe, une transcription orthographique peut être suffisante pour son étude, si celui-ci (ou la liaison) a été annoté. Une transcription phonétique nous conviendrait également, cependant certaines analyses notamment lexicales ou encore syntaxiques seront difficilement applicables, et on sait que ces niveaux d'analyse influencent le comportement de notre phénomène. Nous avons fait le choix de travailler sur des corpus au format CHAT, un format lié au logiciel de transcription CLAN (Computerized Language ANalysis, voir MacWhinney, 2000) utilisé tout d'abord par les chercheurs en acquisition. Des données non transcrites dans ce format seront converties (voir ci-dessous) et celles n'étant pas encore transcrites le seront dans ce format. Le codage se fait quant à lui directement dans la tire principale avec le segment consonantique de liaison entre crochets (par exemple *des [z] étoiles*) et l'annotation [o] peut être utilisée pour la non-liaison (par exemple *il est [o] arrivé*). Ce codage résulte d'une proposition d'un groupe de chercheurs s'étant réunis en 2017 (voir Dugua et al. 2017) afin de faire converger les pratiques en matière de codage de la liaison. Cette proposition n'étant destinée qu'à coder des cas typiques, nous avons en outre dû proposer une façon d'annoter les pataquès, avec un astérisque avant le segment consonantique de pataquès entre crochets (du type *huit [*z] enfants*). De la même façon, des annotations n'utilisant pas ce format seront converties et celles n'étant pas encore codées le seront selon ce codage.

- 23 Le Tableau 5 met en avant trois corpus qui seront analysés plus finement en priorité dans la suite de nos recherches : le fonds Jean Dumas et les corpus PFC et ALIPE. Le premier nécessitera le plus de temps de travail car il faudra transcrire et annoter ce corpus pour la liaison et le pataquès (pour localiser à la fois les cas typiques : la liaison, les cas atypiques : le pataquès, et les contextes sans réalisations) avant de pouvoir l'analyser. Le deuxième corpus a l'avantage d'être transcrit et annoté pour la liaison et le pataquès, transcription et annotation qui devront tout de même être vérifiées et surtout converties du format TEXTGRID (.TextGrid), format de transcription du logiciel Praat, vers le format CHAT (.cha), format de transcription du logiciel CLAN, en gardant à l'esprit que toute conversion comporte forcément une perte d'informations.

Corpus	ALAVAL	Jean Dumas	PFC	ALIPE
Accessibilité des données primaires	Oui	Oui	Oui	Oui
Transcrit	Oui	Non	Oui	Oui
Mode de transcription	Phonétique, traduction	Orthographique adapté	Orthographique	Orthographique
Format	.pdf	.pdf (manuscrit numérisé)	.TextGrid (Praat)	.cha (CLAN)
Transcrit au format CHAT	Non	Non	Non	Oui
Codé pour la liaison	Non	Non	Oui	Oui
Codé comme adopté	Non	Non	Non	Non
Accessibilité des données secondaires	Non	-	Oui	Oui
Outil d'exploration pour le pataquès ¹⁷	Non	Non	Oui	Oui

Tableau 5 : État d'avancement des divers corpus sélectionnés pour l'étude du pataquès

- 24 Une conversion peut s'effectuer de deux façons : les logiciels de transcription eux-mêmes peuvent proposer des commandes spécifiques (comme la commande PRAAT2CHAT du logiciel CLAN pour notre cas) ; il existe de plus un outil de conversion, le TEICONVERT proposé par la plateforme Ortolang¹⁷ (dans ce contexte nous choisirons comme fichier source un .TextGrid et comme format de destination CHA (chat – childes)). Comme le notaient Daoust et Marcoux en 2006, « il est possible de concevoir des programmes externes de conversion agissant comme passerelles entre chacun des formats propriétaires et le format pivot. On a donc besoin de deux types de passerelles : du format propriétaire au format pivot, et du format pivot vers le format propriétaire » (Daoust & Marcoux 2006 : 5). C'est ce que l'outil TEICONVERT (voir Étienne, Parisse & Liégeois 2018) a réussi à concrétiser avec une conversion entre les divers formats utilisés par les logiciels de transcription majoritairement répandus dans la communauté linguistique en passant par le format pivot TEI. Le dernier corpus, ALIPE, nécessitera le moins de travail : il est transcrit au format CHAT, annoté pour la liaison

et le pataquès, seul le codage proposé devra être transformé en direction du codage adopté.

- 25 Cette proposition de méthodologie pour l'étude de phénomènes rares en corpus, qui résulte du constat que le chercheur s'attaquant à ce type d'étude ne peut pas forcément répliquer les méthodologies traditionnellement utilisées en linguistique de corpus, est ici présentée pour fournir matière à réflexion au chercheur et ne veut pas être rigide. De cette manière, les étapes proposées pourront être supprimées, modifiées tandis que d'autres pourront être ajoutées. Même si la méthodologie que nous avons développée est principalement inspirée de notre propre expérience, elle ne semble pas se limiter aux phénomènes morphophonologiques rares ou phénomènes rares dans les corpus oraux qui nous intéressent tout particulièrement.
- 26 Lors de la première étape, soit le va-et-vient entre état de l'art et collection d'exemples attestés, nous nous sommes concentrés sur trois aspects, soit les diverses appellations données aux phénomènes rares pouvant ralentir l'étape de revue de la littérature, la question de savoir si une collection d'exemple peut être dépassée pour mener une analyse sur corpus et enfin nous avons abordé le problème de la nécessité de bénéficier de données massives pour une analyse quantitative statistiquement pertinente des phénomènes rares. Durant la seconde étape de choix et carottage de corpus disponibles variés, il nous a semblé essentiel d'aborder les questions du type de corpus recherché et des endroits où trouver des corpus. Une fois les corpus sélectionnés, une attention particulière a été accordée aux populations ayant produit ces données (et plus largement aux métadonnées associées aux corpus). Toujours au sein de la même étape, les corpus sélectionnés ont été carottés et une première fouille des données a pu être réalisée. Lors de la troisième étape de constitution de nouveaux corpus qui n'étaient pas disponibles à l'étape précédente, deux solutions ont été présentées : la compilation de données existantes et la création de nouveaux corpus avec recueil de données primaires. Au cours de cette étape l'accent a été mis sur le type de données qui peuvent être utilisées en corpus en posant la question du caractère écologique des données auxquelles le chercheur a recours. L'émergence de nouveaux corpus entraîne ensuite selon notre méthodologie un retour à la deuxième étape pour un carottage et une possible première fouille des données. Finalement, au cours de la dernière étape, les corpus retenus comme les plus pertinents ont été analysés plus finement par ordre de priorité. Ceci dépend des résultats du carottage et de la première fouille éventuelle des données, tout comme de l'état d'avancement du corpus permettant sa réutilisation. Nous avons en outre présenté le format de transcription retenu pour notre étude (lié à un logiciel de transcription), le codage adopté, ainsi que des outils de conversion à destination de la communauté scientifique.
- 27 Après une étude poussée des données lors de la dernière étape, il sera par la suite intéressant de voir si un corpus analysé finement reflétait déjà des tendances lors de la collection d'exemples, du carottage et de la première fouille des données. De plus, utiliser cette méthodologie avec différents phénomènes linguistiques rares nous permettra de démontrer dans quelle mesure les phases de recueil d'une collection d'exemples et de carottage des données sont nécessaires voire pertinentes. Pour l'heure, nous ne saurions nous prononcer étant donné que nous n'avons pas achevé la dernière étape, soit celle d'une analyse plus fine de nos données.

BIBLIOGRAPHIE

- ARMSTRONG, Nigel & Timothy POOLEY (2010) : *Social and linguistic change in European French*. New York : Palgrave Macmillan.
- BERT, Michel et Colette GRINEVALD (2010) : « Proposition de typologie des locuteurs de LED ». *Faits de langues* 35 (1) : 117-32.
- BLANCHE-BENVENISTE, Claire (2000) : *Approches de la langue parlée en français*. Paris : Ophrys.
- BLANCHE-BENVENISTE, Claire & Philippe MARTIN (2010) : *Le français : usages de la langue parlée*. Les langues du monde. Leuven : Peeters.
- BÜRKI, Audrey et Marina LAGANARO (2014) : « Représentation et encodage de la liaison : approches psycholinguistiques et neuropsycholinguistiques ». *La liaison, approches contemporaines* éd. par Christiane Soum-Favaro, Annelise Coquillon & Jean-Pierre Chevrot. Sciences pour la communication. Bern : Peter Lang, 169-85.
- BYBEE, Joan (2005) : « La liaison : effets de fréquence et constructions ». *Langages* 2 : 24-37.
- CAPPEAU, Paul et Françoise GADET (2007) : « L'exploitation sociolinguistique des grands corpus. » *Revue française de linguistique appliquée*, vol 12, n°1 : 99-110.
- CHABANAL, Damien, Loïc LIÉGEOIS et Thierry CHANIER (2015) : « Acquisition de la variation phonologique et recueil de corpus d'interactions naturelles parents-enfants : nouvelle méthode, nouveaux enjeux ». *LIDIL - Revue de linguistique et de didactique des langues* 51 (Acquisition du langage et phonologie : méthodes, objets, enjeux théoriques) : 65-88.
- CÔTÉ, Marie-Hélène (2017) : « La liaison en diatopie : esquisse d'une typologie ». *Journal of French Language Studies* 27 (1) : 13-25.
- CRYSTAL, David (2008) : *A dictionary of linguistics and phonetics*. 6e éd. The language library. Malden, Massachusetts, Oxford : Blackwell Pub.
- DAOUST, François et Yves MARCOUX (2006) : « Logiciels d'analyse textuelle : vers un format XML-TEI pour l'échange de corpus annotés ». *Les Cahiers de la MSH Ledoux* 3, Actes des JADT 1 : 327-40.
- DESROCHERS, Richard (1994) : « Les liaisons dangereuses : le statut équivoque des erreurs de liaison », *Lingvisticae Investigationes* 18 (2) : 243-84.
- DIÉMOZ, Federica et Andres KRISTOL (2018) : « Atlas linguistique audiovisuel du francoprovençal valaisan. Questions de morphologie et de syntaxe francoprovençales (version provisoire) ». Centre de dialectologie et d'étude du français régional de l'Université de Neuchâtel. Consultable à l'adresse suivante : http://alaval.unine.ch/uploads/alaval_docs/Pre%CC%81face-Intro%20V2%202018.pdf (20.03.2019).
- DUGUA, Céline, Olivier BAUDE, Flora BADIN, Marie-Hélène CÔTÉ, Jennifer GANAYE, Loïc LIÉGEOIS, Christophe PARISSÉ et Anne SICCARDI (2017) : « La liaison dans l'environnement langagier des enfants : Vers une annotation commune ? » présenté à FLORAL 2017 : Accessibilité, représentations et analyses des données, Orléans.
- DURAND, Jacques, Bernard LAKS, Basilio CALDERONE & Atanas TCHOBANOV (2011) : « Que savons-nous de la liaison aujourd'hui ? » *Langue française* 169 (1) : 103-135.

- DURAND, Jacques, Bernard LAKS et Chantal LYCHE (2009) : « Le projet PFC (phonologie du français contemporain) : une source de données primaires structurées ». *Phonologie, variation et accents du français* éd. par Jacques Durand, Bernard Laks & Chantal Lyche. Paris : Hermès, 19-61.
- ETIENNE, Carole, Christophe PARISSE et Loïc LIÉGEOIS (2018) : « Spoken Language : Tools and Workflow for Creating and Editing Data and Metadata ». Workshop à TEI 2018, Tokyo.
- FORNEL, Michel de & Maud VERDIER (2018) : « Corpus, classes d'exemples et collections en analyse de conversation ». *Corpus* 18.
- GAATONE, David (1978) : « Forme sous-jacente unique ou liste d'allomorphes ? (à propos des consonnes de liaison en français) ». *Linguistics* 16 (214) : 33-54.
- GADET, Françoise (1992) : *Le français populaire*. 2. éd. corr. Que sais-je? n°1172. Paris : Presses Université de France.
- GADET, Françoise (1997) : *Le français ordinaire*. Vol. 2. Paris : Armand Colin.
- GILQUIN, Gaëtan et Stefan Th. GRIES (2009) : « Corpora and experimental methods : A state-of-the-art review ». *Corpus Linguistics and Linguistic Theory* 5, n°1 : 1-26.
- HABERT, Benoît (2000) : « Des corpus représentatifs : de quoi, pour quoi, comment ». *Cahiers de l'Université de Perpignan* 31 : 11-58.
- HORNSBY, David (2012) : « Getting It Wrong : Liaison, Pataquès, and Repair in Contemporary French ». *On Linguistic Change in French : Socio-Historical Approaches. Studies in Honour of R. Anthony Lodge. Le Changement Linguistique En Français : Aspects Socio-Historiques. Études En Hommage Au Professeur R. Anthony Lodge* éd. par Timothy Pooley et Dominique Lagorgette, éd. par T. Pooley et D. Lagorgette. Chambéry : Éditions de l'Université de Savoie, 69-84.
- JONES, Mari C. (2012) : « Liaison patterns and usage in Jersey Norman French ». *Probus* 24 (2) : 197-232.
- LEE, David Y. W. (2010) : « What corpora are available ». *The Routledge handbook of corpus linguistics* éd. par Anne O'Keefe et Michael McCarthy. Abingdon : Routledge, 107-21.
- LÉON, Monique (1984) : « Erreurs et normalisation : les liaisons fautives en français contemporain », *Revue de Phonétique Appliquée* 69 : 1-9.
- LÉON, Pierre (2011) : *Phonétisme et prononciations du français : avec travaux pratiques d'application et corrigés*. 6e éd. Cours Lettres. Paris : Colin.
- LÉON, Pierre R. (1993) : *Précis de phonostylistique : parole et expressivité*. Paris : Nathan.
- LRL (2017) : *ALIPE (Acquisition de la Liaison et Interactions Parents Enfants)*. <https://hdl.handle.net/11403/alipe-000853/v1.1> (20.03.2019).
- MACWHINNEY, Brian (2000) : « The CHILDES Project : Tools for Analyzing Talk ». Manuels consultables aux adresses suivantes : <https://talkbank.org/manuals/CHAT.pdf>, <https://talkbank.org/manuals/CLAN.pdf> (20.03.2019).
- MoDyCo & RUG (2017). *PFC - Phonologie du Français Contemporain*. <https://hdl.handle.net/11403/pfc/v1> (20.03.2019).
- MORIN, Yves-Charles & Jonathan D. KAYE (1982) : « The Syntactic Bases for French Liaison ». *Journal of Linguistics* 18 (2) : 291-330.
- MÜLLER, Bodo (1985) : *Le français d'aujourd'hui*. Éd. rév. et augm. Paris : Klincksieck.

O'KEEFE, Anne et Michael MCCARTHY : « Historical perspective : what are corpora and how have they evolved? ». In *The Routledge handbook of corpus linguistics*, édité par Anne O'Keeffe, et Michael McCarthy. Abingdon : Routledge, 31-41.

SPITZER, Leo (1941). « French Pataquès ; Cuir », *Linguistic Society of America*, 17 (3) : 253-255.

NOTES

1. Cette visualisation se fonde sur le relevé manuel des termes usités dans un petit corpus de textes scientifiques abordant notre phénomène d'étude (Spitzer, 1941 ; Gaatone, 1978 ; Morin & Kaye, 1982 ; Léon, M., 1984 ; Müller, 1985 ; Gadet, 1992 ; 1997 ; Léon, P., 1993 ; 2011 ; Desrochers, 1994 ; Blanche-Benveniste, 2000 ; Blanche-Benveniste & Martin, 2010 ; Bybee, 2005 ; Armstrong & Pooley, 2010 ; Durand et al., 2011 ; Hornsby, 2012 ; Jones, 2012 ; Côté, 2017).
2. Pour un bon aperçu des corpus existants pour l'anglais, même si déjà daté, voir Lee, 2010.
3. Accessible à l'adresse suivante : www.language-archives.org
4. Accessible à l'adresse suivante : <https://vlo.clarin.eu/>
5. Accessible à l'adresse suivante : <https://www.ortolang.fr/>
6. Accessible à l'adresse suivante : <https://cocoon.huma-num.fr/>
7. Accessible à l'adresse suivante : <http://patrimoine-oral.org/>
8. Accessible à l'adresse suivante : <https://gallica.bnf.fr/>
9. Accessible à l'adresse suivante : <http://patrimoine-oral.org/dyn/portal/index.seam?aloId=11898&page=alo&fonds=&cid=448>
10. Accessible à l'adresse suivante : <https://www.ortolang.fr/market/corpora/alipe-000853>. Dans la Figure 4, l'âge des enfants est donné en années;mois.jours.
11. Accessible à l'adresse suivante : <http://alaval.unine.ch/>
12. Accessible à l'adresse suivante : <https://www.ortolang.fr/market/corpora/pfc>
13. Accessible à l'adresse suivante : <http://patrimoine-oral.org/dyn/portal/index.seam?page=alo&aloid=31770&fonds=8&nat=3&cid=423>
14. Accessible à l'adresse suivante : <https://www.youtube.com/>
15. Accessible à l'adresse suivante : <https://www.ina.fr/>
16. Accessible à l'adresse suivante : <https://www.elysee.fr/>
17. Disponible à l'adresse suivante : <http://ct3.ortolang.fr/teiconvert/#>

RÉSUMÉS

Cette contribution constitue une proposition de méthodologie pour travailler sur des phénomènes linguistiques rares et est à destination de chercheurs qui voudraient se lancer à leur tour dans ce type de recherches. Issue de notre expérience d'étude du pataquès, un phénomène propre notamment au français, cette méthodologie est orientée vers l'étude de données orales et plus particulièrement de phénomènes morphophonologiques rares. Pour commencer, nous proposons d'approcher le phénomène par un va-et-vient entre état de l'art et collection d'exemples attestés, débouchant sur le choix de corpus disponibles variés qui sont par la suite carottés (échantillonnage non représentatif). Nous abordons en outre la question de la constitution de nouveaux corpus pour les corpus qui semblent pertinents mais qui n'existent pas

encore. Après avoir été à leur tour carottés, il est alors possible de sélectionner les corpus à analyser plus finement. Plus généralement, les chercheurs intéressés par l'utilisation de corpus en linguistique retrouveront ici abordées des questions essentielles (délimitation du phénomène étudié, recherche de corpus, utilisation de bases de données, d'outils de conversion, importance des métadonnées, évaluation du temps de travail sur un corpus, du caractère écologique des données en présence, masse de données nécessaires...).

This paper proposes a methodology to work on rare linguistic events. Based on our research experience with *pataquès*, a phenomenon occurring among others in French, we mostly focus here on oral data and rare morphophonological events. This paper aims to provide other researchers an example of how to proceed with such research interests. In order to first address this issue, we recommend navigating between a state of the art and collecting attested examples. Once done, it is thus possible to choose various corpora between available ones. We then propose to “core drill” selected corpora (contrary to sampling, core drilling does not implicate representativeness). When interesting corpora are lacking, new corpora can also be constituted before getting core drilled. Finally, we select the corpora that seem the most appropriate and proceed to a finer-grained data analysis. Many issues raised here would interest linguists working with corpora (delimitating the phenomenon studied, searching corpora, using databases and conversion tools, understanding how important are the metadata, evaluating how much work is needed on a corpus, how much ecological are the data we use, the amount of data needed...).

INDEX

Mots-clés : phénomènes linguistiques rares, méthodologie, corpus oraux, morphophonologie, pataquès

Keywords : rare linguistic events, methodology, oral corpora, morphophonology, pataquès

AUTEUR

GASPARDE COUTANSON

Laboratoire MoDyCo (UMR 7114, Université Paris Nanterre)