De petits corpus pour une grande base de données sur l'anglais oral contemporain : quels enjeux à la lumière du programme PAC ?

Cécile Viollain, Hugo Chatellier

De petits corpus pour une grande base de données sur l'anglais oral contemporain : quels enjeux à la lumière du programme PAC ?

Small corpora for a big database on contemporary oral English: what is at stake in the light of the PAC program?

Cécile Viollain and Hugo Chatellier

Introduction

Au vu du nombre de corpus utilisés aujourd'hui en linguistique en général, et en phonologie en particulier, se demander « pourquoi utiliser un/des corpus ? » pourrait paraître superflu. Il n'en est rien car cette question est le reflet d'un positionnement idéologique sur ce qu'est l'« objet » langue, et d'une réflexion épistémologique et méthodologique sur les outils adaptés à son observation, qui doivent être systématiquement expliqués, défendus, justifiés. Et cette tâche n'est pas des plus aisées car les corpus ont de nombreux détracteurs, qui ne croient ni en leur nécessité, ni en leur pertinence. Nous en voulons pour preuve les propos de Durand (2009 : 26) :

> *That corpora occupy a special place in our field seems undeniable given the number of conferences, articles, monographs, book chapters and journals devoted to this very topic. Indeed, there are even researchers who define their work as framed within 'corpus linguistics'. On the other hand, there are still specialists (mainly but not solely within the Chomskyan tradition) who think that recourse to corpora is not the correct way of addressing the fundamental issues of linguistics.*

Dans ce contexte, défendre l'utilisation et la validité des *petits* corpus spécifiquement pourrait donc ressembler à une mission impossible tant les grands corpus semblent bénéficier, de façon presque automatique, d'une supériorité scientifique, et même d'une certaine forme d'indulgence de la part des sceptiques, car ils sont censés être plus

représentatifs de la langue ou de la variété d'une langue étudiée, et donc fournir des informations plus solides aux chercheurs, ne serait-ce que d'un point de vue statistique. De fait, si les grands corpus, majoritairement écrits, tels que le COCA ou le BNC, se sont fait un nom et ont inspiré de nombreuses recherches récentes, les petits corpus, oraux qui plus est, peinent quant à eux à se faire une place dans la littérature alors même que nous estimons qu'ils constituent une ressource privilégiée pour la communauté scientifique.

Cependant, un tour d'horizon rapide des ressources disponibles, en ligne notamment, dans l'ensemble des disciplines linguistiques, et en particulier en ce qui concerne l'étude de la langue orale, démontre rapidement que les petits corpus sont bel et bien là, et qu'ils font plus que survivre aux côtés des « mégas » ou « gigas » corpus, le plus souvent écrits. La pérennité des petits corpus n'était sans doute pas prévisible avec l'essor de la linguistique de corpus, qui a spectaculairement multiplié les ressources et donc nécessairement engagé une forme de compétition entre corpus, mais surtout avec le développement du *big data* et d'outils, notamment statistiques, toujours plus performants pouvant ingurgiter et traiter des quantités astronomiques de données. McEnery et Wilson (2001 : 191) résument bien cette situation, pour ce qui est des corpus écrits en particulier, dans la deuxième édition de leur introduction à la linguistique de corpus :

> *If there was one development we did not predict it was that, as well as getting larger, corpora would also get smaller! While the individual corpus linguist's handcrafting of their own small corpus in order to address a particular research question was not unusual in the 1980s and early 1990s, one may have assumed that, as corpora have grown, the need for this activity would have faded. This has not proven to be the case. Many linguists are interested in contrasting the language used in large, general-purpose corpora such as the BNC with small corpora representing text types –or the writings of a single author– not available in the BNC.*

Nous pouvons donc nous demander, sur la base de ce constat élémentaire, pourquoi les petits corpus survivraient-ils s'ils n'étaient pas viables des points de vue méthodologique et scientifique. Et, par-là même, quels sont les avantages, quelle est la plus-value des petits corpus, sur les grands corpus notamment, mais également en soi. Il nous semble en effet opportun, puisque l'occasion nous en est offerte, de penser les petits corpus pour ce qu'ils sont, et non systématiquement dans leur rapport aux grands corpus. Aussi, la raison pour laquelle nous avons souhaité contribuer à ce numéro thématique c'est, en creux, de défendre la linguistique de corpus et, en contrepoint de ce qui se fait souvent, défendre les petits corpus et leurs potentialités. Cependant, nous souhaitons le faire sous l'angle spécifique d'un domaine d'étude qui, selon nous, entretient des liens privilégiés avec les corpus, et notamment les petits corpus : la phonologie. Nous devrons par conséquent contextualiser notre réflexion et donc revenir, lorsque nous le jugerons opportun, sur des débats majeurs. Certains sont d'ailleurs toujours d'actualité au sein de la communauté des phonologues en ce qui concerne la définition même de la phonologie et de ce qu'elle est censée observer et penser, de ce qu'est un corpus véritablement phonologique, et de ce que sont un corpus réussi ou optimal et une méthodologie scientifique efficace et adaptée.

Ainsi, dans une première partie (§ 1), nous développerons notre réflexion sur la spécificité de la phonologie par rapport à d'autres disciplines linguistiques dans son rapport aux corpus, et discuterons de la définition de ce qu'est un corpus phonologique réussi ou optimal. Cela nous amènera à interroger et relativiser les notions de petitesse et de représentativité qui pèsent sur la constitution des corpus en général, et des corpus

phonologiques en particulier. Nous détaillerons également le rapport de force qu'entretiennent les corpus dits « généraux » et les corpus dits « spécialisés » et tenterons alors de dépasser l'impératif de représentativité pesant sur les corpus, ainsi que la taille à proprement parler de ces objets, pour nous concentrer sur la question de leur finalité, en tant qu'outils privilégiés pour l'étude d'une multitude de phénomènes phonético-phonologiques. Ceci nous conduira logiquement à réfléchir à la méthodologie adaptée à l'observation et à la compréhension de la structure de la langue orale, et spécifiquement du système de l'anglais oral contemporain.

Par conséquent, dans un deuxième temps (§ 2), nous défendrons les ambitions et la méthodologie propres au programme PAC, dont la base de données sur l'anglais oral contemporain repose sur la constitution et l'exploitation de petits corpus oraux[2] constitués dans les différentes aires géographiques anglophones, selon un protocole commun permettant la comparabilité des données récoltées. La finalité affichée de ce programme est de dresser un portrait précis de l'anglais contemporain tel qu'il est parlé à travers le monde, en intégrant systématiquement la variation à la description phonético-phonologique des systèmes de l'anglais, mais également de mettre à l'épreuve de données authentiques les modèles phonétiques, phonologiques et sociolinguistiques existants. Aussi, les petits corpus dont ce programme dispose sont-ils pertinents à titre individuel pour permettre de passer par exemple d'analyses phonético-acoustiques à des modélisations théoriques de phénomènes phonético-phonologiques. Mais ils le sont également à titre collectif dans la mesure où la jonction des données, des annotations et des analyses effectuées à partir de ces données permettent d'éclairer la tectonique des plaques linguistiques à travers le monde anglophone, l'évolution des systèmes des variétés de l'anglais, mais également les enjeux identitaires et sociolinguistiques liés aux accents régionaux. Nous montrerons que les programmes qui disposent de ce type de ressources tournées vers un objectif similaire ne sont pas si nombreux. Et même si le programme PAC n'est bien évidemment pas le seul, la stratégie qu'il met en place, à savoir la jonction de petits corpus pour l'analyse quantitative et qualitative de phénomènes phonético-phonologiques dans plusieurs variétés d'une même langue, a fait ses preuves pour d'autres programmes, à commencer par le programme PFC (Phonologie du Français Contemporain, Durand & Lyche 2008, Detey *et al.* 2016), dont PAC se veut, à terme, l'équivalent pour l'étude de l'anglais.

Enfin, dans une dernière partie (§ 3), il nous faudra montrer concrètement comment de petits corpus tels que ceux dont dispose le programme PAC peuvent constituer une grande base de données sur l'anglais oral contemporain. Pour ce faire, nous présenterons certains des résultats obtenus grâce à la comparaison des analyses faites à partir de plusieurs corpus sur deux phénomènes phonético-phonologiques majeurs : la rhoticité d'une part (corpus PAC Lancashire, Boston et Nouvelle-Zélande, Navarro 2013, 2016 ; Viollain 2010, 2014), et les changements vocaliques d'autre part (corpus PAC-LVTI Manchester et PAC Nouvelle-Zélande, Chatellier 2016 ; Viollain 2014). Nous montrerons également comment ces résultats alimentent notre réflexion quant à la question des identités nord-sud dans le monde anglophone, d'un point de vue historique et sociolinguistique.

1. Quelle place pour les (petits) corpus en phonologie ?

Si nous suivons la logique que nous avons établie dans notre introduction, avant même de pouvoir défendre les petits corpus et leurs potentialités, il nous faut défendre le recours aux corpus dans le domaine particulier de la phonologie, et par là-même expliquer pourquoi la phonologie entretient un rapport privilégié aux corpus. Pour ce faire, nous souhaitons citer l'introduction magistrale de Scheer (2004 : 30) au numéro thématique dédié aux corpus en phonologie dans cette même revue :

> [...] les phonologues n'ont plus désormais le choix : ils doivent fonder leurs analyses sur des corpus qui donnent un aperçu numérique des forces lexicales en jeu. Cette obligation est la conséquence de la circonstance suivante, néfaste à bien des égards : la phonologie de par sa nature est irrémédiablement et intimement entrelacée avec le lexique, la diachronie et l'analogie ; elle a comme objet d'étude un ensemble fini d'unités. Cette spécificité se révèle donc être un avantage lorsque l'on se penche sur ses conséquences méthodologiques qui concernent, notamment, l'utilisation des corpus.

Ce constat a deux conséquences épistémologiques et méthodologiques majeures, qui seront au cœur de notre propos dans cette première partie. En premier lieu, la phonologie d'une langue, ou d'une variété d'une langue, de par le fait qu'elle repose sur un nombre fini d'oppositions permettant de faire émerger le sens, peut être appréhendée dans son ensemble au sein d'*un* corpus. Autrement dit, il n'y a pas de récursivité (Scheer 2004 : 30) en phonologie : là où il est théoriquement possible de créer des structures syntaxiques à l'infini, en ajoutant par exemple des propositions relatives successives, rien de comparable n'est envisageable en phonologie. Aussi, si l'on souscrit par exemple à la notion de phonème (bien que certains spécialistes ne croient pas en son existence ou lui préfèrent d'autres termes) en tant qu'unité de son au sein des systèmes des langues, on peut établir que l'ensemble des phonèmes d'une langue sera observable au sein d'un seul et même corpus[3]. Il est impossible de postuler la même chose en ce qui concerne la syntaxe ou la morphologie par exemple.

Ceci explique l'existence même d'une approche à part entière dénommée « phonologie de corpus », qui a sa propre entrée dans la *Oxford Research Encyclopedia of Linguistics* (Aronoff 2016). Ainsi, de la même manière que l'on parle de linguistique de corpus pour désigner la démarche scientifique fondée sur l'exploitation de corpus pour l'étude de l'objet langue, on parle de phonologie de corpus pour définir l'étude de la phonologie mettant au centre de sa démarche le recours aux corpus oraux. Sur la base de notre observation précédente, on pourrait même poser que toute phonologie doit être une phonologie de corpus, position justement défendue par Durand (2017) dans cette encyclopédie :

> *As a sub-branch of corpus linguistics it comes in two forms: a strong version that states that the study of spoken corpora should be the aim of phonology; a weaker version which stresses that corpora should occupy pride of place within the set of techniques available (for example, intuitions, psycholinguistic and neurolinguistic experimentation, laboratory phonology, the study of the acquisition of phonology or of language pathology, etc.). Whether one defends a strong or a weak version, corpora are part and parcel of the modern research environment.*

Il explicite d'ailleurs plus loin :

> [...] *are corpora central to phonology or just an element within a large palette of techniques for exploring the phonological structures of given languages? Whether one adopts a strong*

> *or a weak version of corpus phonology, it is argued that off-the-cuff observations and intuitions should play a minimal role in phonology and that systematic data collection and the back and forth cycles between theorization, observations and experimentation necessarily require the use of corpora.*

Nous tenons à souligner qu'en posant deux versions de la phonologie de corpus, l'une plus radicale, et l'autre plus mesurée, Durand ne pèche pas par naïveté épistémologique et ne sous-entend pas que sans corpus, il n'y a point de salut. Au contraire, il suggère que plus l'arsenal du chercheur est fourni en outils complémentaires, plus sa démarche peut se rapprocher de ce que serait un idéal pour l'étude holistique de la langue. Nous souscrivons pleinement à sa définition.

Le choix de la phonologie de corpus peut donc se justifier par la nature même de la discipline et des données qu'elle observe. Mais ce choix épistémologique implique de définir précisément ce qu'est un corpus phonologique, et surtout un corpus phonologique optimal, et donc de dresser un cahier des charges méthodologique permettant de le constituer, de l'atteindre pour ainsi dire. Nous nous référons ici à la définition proposée par Gut et Voorman (2014 : 16) :

> A phonological corpus is thus defined here as a sample of language that contains
> • primary data in the form of audio or video data;
> • phonological annotations that refer to the raw data by time information (time-alignment); and
> • metadata about the recordings, speakers and corpus as a whole.

Nous notons que cette définition reste, à dessein, simple et générale, car Gut et Voorman soulignent eux-mêmes qu'il n'existe, à ce jour, aucune définition d'un corpus phonologique qui ne fasse pas débat au sein de la communauté scientifique.

Aucun mot donc, ici, de la taille à proprement parler du corpus phonologique par excellence, puisque cette définition se concentre plutôt sur la diversité et la complémentarité des données qu'il doit regrouper : données primaires, orales ou audiovisuelles, données secondaires, soit les annotations, potentiellement illimitées, des données primaires, et les métadonnées, c'est-à-dire les informations complémentaires sur les enregistrements en tant que tels et les locuteurs, ou tout autre type d'information pouvant se révéler utile.

Nous en venons ainsi à la deuxième conséquence majeure du rapport privilégié qu'entretient la phonologie avec les corpus. En effet, et là encore contrairement à ce qui se passe en syntaxe ou en morphologie notamment, la phonologie n'exige pas nécessairement de *grands* corpus, justement parce que son domaine d'étude est fini. Il est en effet assez aisé de se figurer combien l'étude de la syntaxe, de la créativité morphologique, ou de la morphologie lexicale et flexionnelle peut bénéficier de ressources écrites quasi-illimitées. Au contraire, des milliers d'heures d'enregistrement de locuteurs anglophones ne changeront rien, pour ainsi dire, à l'inventaire qui pourra être dressé des unités constituant leur système. De nombreux corpus oraux, tel que le *Lancaster / IBM Spoken English Corpus* (53 000 mots), entrent donc dans la catégorie des petits corpus, dont la limite haute est souvent fixée autour de 200 000 mots dans la littérature (Aston 1997 ; Vaughan & Clancy 2013). Toutefois, nous allons vite nous apercevoir que la frontière entre les petits et les grands corpus est extrêmement flexible, et quelque peu arbitraire selon nous, et qu'il existe en réalité autant d'adjectifs de taille que de corpus car la notion de taille, et donc de petitesse, est toute relative.

Allons plus loin, les corpus que la phonologie exploite peuvent se voir systématiquement qualifiés de « petits » étant donné la finitude des unités qu'ils présentent. La phonologie

n'a donc pas peur, pour ainsi dire, des petits corpus, et peut même faire le choix de disposer de petits plutôt que de grands corpus, et transformer la petitesse en potentialité plutôt qu'en contrainte ou en limitation, comme nous allons le voir ultérieurement dans cette même partie. Nous ne souhaitons pas pour autant donner l'impression d'ignorer ou de tourner le dos à tout un pan de la phonologie de corpus qui s'efforce de construire de grands corpus phonologiques compatibles avec les outils modernes de traitement automatique et d'exploitation statistique des données, ou de ne pas voir les bénéfices qu'ils apportent à l'analyse quantitative, notamment, de nombreux phénomènes phonético-phonologiques. Cependant, notre propos consiste ici à nous concentrer sur les bénéfices possibles de l'utilisation de petits corpus, contrairement à ce qui se fait souvent, puisqu'il nous semble plus évident de percevoir les avantages qu'un grand nombre plutôt qu'un petit nombre de données peut représenter.

Toutefois, la proximité qu'entretient la phonologie avec les corpus et l'absence d'une obligation de grandeur pour atteindre une forme de représentativité qui caractérise le domaine n'impliquent pas de tomber nécessairement dans l'extrême inverse : sans requérir de « mégas » corpus, la phonologie, et ses disciplines connexes telles que la phonétique, n'ont vocation ni à n'utiliser que des micros corpus ultraspécialisés ni à s'arrêter à la simple description des données recueillies. Les corpus artificiels de parole nue et les corpus tests permettant de traiter d'une question très précise constituent une ressource précieuse, mais les corpus oraux n'ont bien évidemment pas pour autant d'obligation de « petitesse ».

Par conséquent, nous défendons ici l'idée que la petite taille des corpus oraux, comme ceux exploités au sein du programme PAC, est toute relative étant donné que la représentativité linguistique apparaît dans la littérature comme un objectif inatteignable, comme l'explique Kennedy (1998 : 67) :

> *Sinclair (1991 : 9) was able to suggest that 10-20 million words might constitute 'a useful small general corpus' but 'will not be adequate for a reliable description of the language as a whole'. It was argued that corpora of finite size were inherently deficient because any corpus is such a tiny sample of a language in use that there can be little finality in the statistics. Sinclair (1991 : 9) pointed out that even projected billion-word corpora will show remarkably sparse information about most of a very large word list.*

La citation ci-dessus doit par conséquent permettre, nous semble-t-il, de dédramatiser cette notion de petitesse, souvent perçue et vécue comme une contrainte, un défaut, voire un échec, ou du moins envisagée comme ne pouvant constituer que le stade méthodologique intermédiaire, préliminaire, avant que le petit corpus en question ne soit complété, agrandi, enrichi. En effet, si les corpus en général sont par définition petits au sens où ils ne peuvent saisir la totalité de la langue, la question n'est plus de leur taille en tant que telle mais de la qualité de leur construction, du cahier des charges qu'ils remplissent et des possibilités de traitements multidimensionnels qu'ils peuvent offrir : autrement dit, leur finalité. Nous faisons par conséquent le lien ici entre la réflexion épistémologique que nous avons menée sur le rapport qu'entretient la phonologie avec les corpus, et les conséquences méthodologiques qu'il faut en tirer, qui doivent transcender la simple question de la taille, quant à la démarche à suivre pour constituer un objet pertinent pour analyser en profondeur la langue. Nous en venons donc à la définition de ce que doit être un corpus véritablement phonologique.

Il nous semble alors que la petite taille caractéristique de nombreux corpus oraux devrait être redéfinie selon des critères sans doute plus pragmatiques correspondant à la réalité

des contraintes pesant sur la recherche actuelle. Aussi, nous souscrivons à la conclusion de Gut et Voorman (2014 : 22) :

> *The optimal size of a corpus is therefore one that requires a minimum amount of time, effort, and funding for corpus compilation but that, at the same time, guarantees that the distribution of all linguistic features is faithfully represented.*

Le critère de taille, dans la définition de ce que doit être un corpus réussi ou optimal est ici repensé en termes de coût humain et financier, c'est-à-dire en tenant compte à la fois du temps matériel dont peut disposer le chercheur, et des ressources financières mobilisables pour mener à bien son projet.

Or, ces ressources sont justement loin d'être illimitées. Il suffit de penser aux contraintes qui pèsent sur les chercheurs pour produire des résultats, publier des articles ou des ouvrages, et donc exister au sein de la communauté scientifique au sens large, pour comprendre que le temps est une ressource rare. Dans le cas précis d'une thèse de doctorat par exemple, qui est une mine précieuse pour des publications scientifiques ultérieures dans la carrière d'un chercheur, il faut pouvoir dire quelque chose de pertinent et contribuer au débat, le plus souvent en quatre ans de travail maximum, pour ce qui est des sciences humaines spécifiquement. Il faut donc travailler vite et bien, et dans ces conditions, la constitution du corpus ne peut décemment pas prendre plus d'une année, voire moins.

C'est pourquoi, viser un petit corpus bien fait, nécessitant peu de ressources financières car constitué lors d'un séjour de recherche court, semble la voie la plus sage. Au contraire, s'engager dans la constitution d'un méga corpus peut se révéler très coûteux, à la fois en temps, en énergie, et en ressources financières, et finalement potentiellement coûter sa thèse de doctorat au jeune chercheur qui ne réussirait pas à terminer son projet dans le temps imparti. Hériter d'un grand corpus, ou utiliser un grand corpus que l'on n'a pas constitué, peut également coûter cher si l'on n'a pas accès à toutes les informations nécessaires, notamment aux profils sociolinguistiques détaillés des locuteurs, et si l'ensemble des paramètres tels que l'âge, le sexe, l'ethnicité, l'appartenance à la communauté étudiée, pour n'en citer que quelques-uns, n'a pas été contrôlé.

La citation ci-avant et le propos précédent n'éclipsent ou n'évacuent pas pour autant la question centrale de la représentativité, à laquelle la taille est évidemment intimement liée. De fait, il est aisé de se figurer que plus un corpus sera grand, plus il pourra fournir d'informations pertinentes au chercheur quant à la fréquence d'utilisation des formes, et donc plus il constituera un échantillon viable, plus il sera représentatif de l'objet langue dans son ensemble. Mais, puisque nous parlions de coût précédemment, quel est le prix à payer pour constituer un grand corpus, dont nous avons déjà vu que la représentativité sera nécessairement remise en cause tant il semble vain de vouloir contenir toute la langue en un seul objet fini ?

Il nous semble que le prix à payer par les grands corpus est assez lourd. En premier lieu, là où la constitution d'un petit corpus permet à un individu de superviser personnellement l'ensemble de la démarche de A à Z, et donc de maîtriser un ensemble de paramètres défini, la constitution d'un grand corpus implique généralement de travailler à plusieurs, ce qui a un coût en termes de coordination, et peut donner lieu à des écarts dans l'application du protocole de recherche, voire à des erreurs (Gut & Voorman 2014 : 23-24 ; Delais-Roussarie & Post 2014 : 59). Mais surtout, cela s'avère infiniment chronophage en termes de collecte et de traitement ultérieur des données, ce qui peut finalement conduire à publier des résultats longtemps après la constitution du corpus en

lui-même et faire que ces résultats ne soient plus véritablement jugés pertinents par la communauté scientifique, étant donné la constante mutabilité de l'objet langue et l'éventuelle profusion de publications sur le même sujet.

Enfin, et c'est là toute la relativité de cette notion de taille, un grand corpus, comptant de nombreuses heures d'enregistrement avec de nombreux locuteurs, ne pourra pas forcément donner accès à un grand éventail de contextes d'emploi, ou à un grand nombre de tâches diverses, et donc à une information en profondeur sur le système de ses locuteurs. En effet, pour être appliqué à un grand nombre de locuteurs, et être multipliable à l'envi pour ainsi dire, un protocole plus court, plus simple à mettre en place, se révèle généralement nécessaire pour mener à bien la tâche, contrairement à ce qui se passe pour un projet plus spécialisé et, apparemment, moins ambitieux. De même, les données récoltées pour ce grand corpus ne pourront pas nécessairement être annotées, codées, sur un grand nombre de tires[4], c'est-à-dire pour une variété de phénomènes ou de manière multidimensionnelle, à moins de disposer d'une large équipe capable de traiter de manière coordonnée la masse de données collectée, ou qu'un individu dévoué s'y consacre intégralement pendant plusieurs années.

Au contraire, un petit corpus pourra reposer sur un protocole plus long, composé de tâches plus variées et donnant accès à de plus nombreux contextes d'interaction, et être annoté en profondeur pour répondre à des questions sous de multiples angles ou pour penser des phénomènes divers. Nous y reviendrons dans la deuxième partie du présent article pour le cas précis du protocole commun adopté au sein du programme PAC.

Pour résumer, le dilemme auquel est confronté le linguiste, et par là-même le phonologue, au moment de déterminer la taille de son corpus est donc le suivant : risquer d'un côté la représentativité potentielle de la ressource en choisissant de constituer un petit corpus plus facilement manipulable et exploitable en profondeur, ou risquer d'hériter d'une grande base de données dont on ne maîtrise pas tout et/ou de présenter des résultats plusieurs années après, issus d'une enquête menée à bien au prix de nombreux sacrifices, lorsque ces données et les analyses qui en sont issues ne seront plus nécessairement jugées pertinentes étant donné que la langue est un objet en perpétuel mouvement. Vue sous cet angle, la décision semble moins évidemment pencher en faveur du grand corpus.

On peut également penser que la volonté de créer un corpus représentatif d'une langue ou d'une variété d'une langue données, peu importe l'objectif de taille que l'on se fixe, consiste en fait à projeter des préconceptions et des *a priori* sur ce qu'est la communauté parlant cette langue ou cette variété, à obtenir un corpus qui confirme, et donc représente, ce que l'on sait déjà de son objet d'étude. Parallèlement, ne rien penser en amont de la construction d'un corpus sur le terrain est sans doute la meilleure recette pour que l'objet ne soit pas adapté à la question étudiée. Il semble donc que l'enjeu de la taille doive être dépassé pour penser la question de la finalité de l'objet créé, afin de ne pécher ni par anticipation, par préconception, ni par naïveté ou par impréparation. En d'autres termes, et pour reprendre les propos de Kennedy (1998 : 4), il ne s'agit ni de constituer un corpus tellement petit et spécialisé qu'il n'intéresse que soi et qu'il s'avère être le reflet de la propre vision du chercheur sur son objet d'étude, ni de constituer des archives gigantesques qui n'auraient pas été pensées en amont pour permettre l'étude de phénomènes spécifiques :

> *Whereas a corpus designed for linguistic analysis is normally a systematic, planned and structured compilation of text, an archive is a text repository, often huge and opportunistically collected, and normally not structured.*

Alors, si tout a un coût, les grands comme les petits corpus, si la représentativité semble un idéal impossible à atteindre, et si la démarche scientifique au sens large nécessite de toujours trouver un juste équilibre entre une approche radicalement empiriste et une approche absolument théorique, c'est une position intermédiaire que nous souhaitons défendre ici en ce qui concerne les petits corpus, qui permet de repenser le rapport de force entre corpus généraux et corpus spécialisés, et finalement de montrer l'apport de la complémentarité et de la comparabilité des données.

Nous souhaitons citer à ce sujet les propos de Cheng (2012 : 32) qui nous semblent mettre en perspective de manière pertinente les notions de taille, d'équilibre, de représentativité et de finalité pour ce qui est de la définition de la relation entre corpus généraux et corpus spécialisés :

> *In corpus linguistics, a corpus is often described as being either 'general' or 'specialised'. General corpora are usually much bigger than specialised corpora. For example, the Bank of English is over 600 million words; COCA is more than 400 million words; and the BNC is 100 million words, and all are general corpora. Specialised corpora, on the other hand, can usually be measured in the thousands or low millions of words, although there are some that are very large. However, size is not the main factor distinguishing the two types of corpora. What distinguishes general corpora from specialised corpora is the purpose for which they are compiled. General corpora aim to examine patterns of language use for a language as a whole, and specialised corpora are compiled to describe language use in a specific variety, register or genre. The selection of the contents of a specialised corpus often requires the corpus linguist to seek advice from experts in the field to ensure its representativeness and balance.*

À la lumière de ces propos, il apparaît que la constitution d'un « grand » corpus général, ou d'un « petit » corpus spécialisé est un véritable choix scientifique et méthodologique adapté à ce que l'on souhaite observer. Les deux types d'approche sont complémentaires et peuvent dire des choses pertinentes pour la compréhension de la structure de la langue. Pour ce qui est de la langue orale en particulier, les deux approches ne sont pas nécessairement incompatibles, à savoir que la jonction de petits corpus spécialisés permettant de rendre compte des caractéristiques de l'anglais tel qu'il est parlé dans telle aire géographique, par telle génération, issue de telle catégorie socio-économique, et dans tel contexte d'interaction précis, doit pouvoir constituer finalement un corpus général à même de décrire et modéliser la dynamique des variétés de l'anglais à travers le monde. Cela implique toutefois d'établir un protocole commun solide qui garantisse la comparabilité des données.

En conclusion de cette première partie, nous souhaitons ici retourner le raisonnement classique et poser les questions suivantes : que manque-t-il dans les petits corpus ? Que donneraient à voir de grands corpus que nos petits corpus ne peuvent révéler ? Et inversement, que permettent ces petits corpus spécialisés qui n'aurait pu être entrepris à plus grande échelle ? C'est à ces questions que nous souhaitons fournir des éléments de réponse sur la base des résultats fournis par l'étude des petits corpus dont dispose le programme PAC (voir troisième partie). Pour ce faire, il nous faut maintenant détailler ses ambitions et sa méthodologie propres.

2. Les petits corpus du programme PAC

PAC (Phonologie de l'Anglais Contemporain : usages, variétés et structure) est un programme de recherche lancé en 2000 par Jacques Durand (CLLE-ERSS, Université Toulouse Jean Jaurès) et Philip Carr (EMMA, Université Montpellier 3). Il est aujourd'hui coordonné par Sophie Herment (LPL, Université de Provence Aix-Marseille), Sylvain Navarro (CLILLAC-ARP, Université Paris Diderot), Anne Przewozny (CLLE-ERSS, Université Toulouse Jean Jaurès) et Cécile Viollain (CREA, Université Paris Nanterre). Les objectifs de ce programme sont multiples. Il vise notamment à (Durand & Przewozny, 2015 : 63) :

- décrire l'anglais oral dans son unité et sa diversité géographique, stylistique et sociale ;
- mettre les cadres théoriques existants en phonologie, phonétique et sociolinguistique à l'épreuve de données authentiques ;
- créer une émulation et encourager les collaborations au sein de la communauté des phonologues, phonéticiens et autres chercheurs spécialisés en langue orale ;
- contribuer aux réflexions sur l'enseignement de l'anglais en tant que langue étrangère, à l'aide des données recueillies ainsi que des analyses linguistiques conduites sur ces dernières.

Afin de pouvoir remplir ces objectifs, le programme a adopté une méthodologie commune reposant sur la création et la multiplication de points d'enquête à travers le monde anglophone. Les corpus issus de ce travail de terrain incluent généralement entre 10 et 20 locuteurs et doivent idéalement constituer un échantillon de référence en termes de sexe/genre des locuteurs, d'âge/générations et d'origines/profils socio-économiques. Cette méthodologie, tout à fait semblable à celle adoptée au sein du programme-parent PFC (Phonologie du Français Contemporain, Durand *et al.* 2014), est d'inspiration labovienne, et permet de recueillir des données donnant accès à différents degrés de formalité puisque quatre tâches sont enregistrées avec chaque locuteur : deux listes de mots, un texte à lire à haute voix, un entretien guidé, et une conversation libre.

Les deux listes de mots, respectivement centrées sur les systèmes vocalique et consonantique, incluent un total de 192 items lexicaux et constituent le registre le plus formel du protocole. Elles permettent d'obtenir « simultanément une production de mots isolés et de paires phonologiques minimales, dans un contexte stylistique où l'on peut s'attendre à un style formel ou surveillé produit par un locuteur conscient de sa production linguistique (impliquant des phénomènes stylistiques comme l'hypercorrection) » (Durand & Przewozny 2015 : 70). L'analyse des listes de mots offre donc déjà des informations cruciales qui permettent de décrire la variété d'anglais étudiée selon plusieurs critères. C'est le cas des oppositions vocaliques majeures, comme celles reposant sur la longueur (ou une opposition *tendu/relâché*) en anglais standard (voir troisième partie).

Pour ce qui est de l'étude des consonnes, outre les oppositions de voisement, les listes de mots nous renseignent également sur le caractère rhotique ou non-rhotique de la variété examinée (voir partie 3), mais aussi sur certains phénomènes réalisationnels, tels que la vélarisation de /l/, le battement de /r/ ou la glottalisation. Notons que la présence de nombres devant les mots, qui doivent aussi être lus par les locuteurs, offre un double avantage : elle permet de limiter les erreurs lors de la lecture (ou lors de la phase de transcription des données) d'une part, et d'ajouter des items auxquels les locuteurs prêtent généralement une attention moindre d'autre part. Naturellement, les listes de

mots ne sauraient se substituer à une analyse du discours des locuteurs dans d'autres registres, conversationnels notamment. Néanmoins, elles fournissent déjà des informations précieuses sur la variété étudiée.

En effet, l'expérience montre qu'il existe une asymétrie des oppositions phonologiques dans les listes de mots. En théorie, il pourrait y avoir des cas où des locuteurs feraient une opposition dans les listes de mots qu'ils ne maintiendraient pas en conversation, de la même manière que certaines oppositions, absentes des listes de mots, apparaîtraient en contexte conversationnel. Dans les faits, Durand et Przewozny (2015 : 74) soulignent que si certaines oppositions réalisées dans les listes de mots peuvent disparaître en contexte plus informel, ils n'ont à ce jour jamais observé d'oppositions en conversation qui n'étaient pas déjà présentes dans les listes de mots :

> *il existe une asymétrie dans la fiabilité des lectures à haute voix. La présence d'une opposition dans la lecture de la liste de mots (et en particulier dans les paires minimales) ne prouve pas, convenons-en, que le locuteur fasse cette opposition dans la parole spontanée. En revanche, l'absence d'une opposition dans la lecture à haute voix est un indice très fort que l'enquêté ne pratique pas la distinction en question. Ainsi, la plupart des sujets du Lancashire confrontés à des paires minimales potentielles comme 50. ants et 51. aunts (RP / ænts/ vs /ɑ:nts/) les lisent de manière identique ([ænts]). Or, nous n'avons aucun exemple d'enregistrement où l'un des témoins en question, n'ayant pas fait de distinction dans la lecture à haute voix de tels exemples, en pratiquerait une dans la parole spontanée.*

Le texte, adapté d'un article journalistique, constitue le deuxième niveau le plus formel du protocole. Il permet donc lui aussi de recueillir des données contrôlées, mais dans un exercice en général plus familier des locuteurs que la lecture à haute voix de listes de mots. Il permet également d'approfondir les analyses entreprises grâce à l'étude des listes de mots, et autorise en outre l'observation et le traitement de certains phénomènes de la chaîne parlée, comme les phénomènes de *sandhi* par exemple (voir troisième partie).

Le premier type de données conversationnelles que nous recueillons est souvent désigné sous le nom d'entretien guidé. Il s'agit de données moins contrôlées que celles obtenues lors des tâches de lecture. Cette conversation doit permettre à l'enquêteur de dresser le « profil sociolinguistique » de l'enquêté, qui sera indispensable pour toute analyse intégrant ce type de facteurs et, plus généralement, qui facilitera la réutilisation des données par des chercheurs n'ayant pas participé à la phase de collecte sur le terrain. Il existe à cette fin un formulaire servant de trame à l'entretien guidé qui s'articule autour de questions portant sur l'âge, la profession ou le niveau d'études du locuteur et de ses parents proches. D'autres thèmes abordés lors de l'entretien, tels que les passe-temps ou les voyages, permettent à la fois d'affiner le profil du locuteur, mais aussi d'obtenir des interactions plus naturelles que lors de la récolte d'informations plus « élémentaires ».

Il existe souvent une différence de formalité entre cet entretien guidé et la conversation dite « libre », ne serait-ce que parce qu'enquêté et enquêteur ne se connaissent parfois que peu, voire pas du tout. Bien qu'il puisse s'agir d'un désavantage, notamment si l'on est à la recherche de données purement écologiques, nous pensons au contraire que cette différence est un avantage, puisqu'elle garantit d'accéder à quatre degrés de formalité bien distincts. Quoi qu'il en soit, et malgré l'intérêt des données sociolinguistiques, leur recueil ne doit pas se faire au détriment de l'interaction entre l'enquêteur et le locuteur, qui doit rester la plus naturelle et spontanée possible, et ne pas tourner à l'interrogatoire, surtout dans le cas où l'enquêté ne souhaiterait pas répondre à certaines questions personnelles.

Enfin, le dernier type de données incluses dans les corpus PAC est issu de la conversation « libre », qui se déroule idéalement entre deux locuteurs qui se connaissent bien, et sans la présence de l'enquêteur. Il est également possible de solliciter une connaissance d'un locuteur pour cette tâche, quand bien même cette personne ne remplirait pas nécessairement toutes les conditions pour être considérée comme un véritable locuteur de la variété étudiée. Lorsque les conditions ne peuvent être réunies pour que cette conversation libre ait lieu, et qu'il est impossible de solliciter une tierce personne, c'est l'enquêteur qui se charge d'être l'interlocuteur lors de cette tâche. Contrairement à l'entretien, aucun thème n'est imposé pour la conversation libre, d'où son nom, ce qui doit permettre aux locuteurs de s'exprimer le plus naturellement et spontanément possible.

Notons que les différentes tâches du protocole constituent un socle solide garantissant la comparabilité des données entre les différents corpus constitués, ce qui n'empêche en rien les enquêteurs d'ajouter des tâches supplémentaires si cela leur semble pertinent. Ainsi, des listes de mots complémentaires ou la lecture de courtes phrases (pour l'étude du r de *sandhi* en Nouvelle-Zélande, voir troisième partie) ont pu être ajoutées au protocole commun pour certaines enquêtes (Przewozny 2004 ; Pukli 2006 ; Viollain 2014). Un autre exemple d'étoffement du protocole d'origine est le questionnaire utilisé lors des enquêtes du projet LVTI (Langue, Ville, Travail, Identité) : il s'agit de questions ayant trait à la vie professionnelle des locuteurs, et à leur relation à la langue, à la ville, et à leur identité. Ces questions sont intégrées à l'entretien guidé dans les enquêtes LVTI (Chatellier 2016 : 180-184), et autorisent des investigations sociolinguistiques plus poussées des variétés urbaines que celles traditionnellement permises grâce au protocole PAC classique.

Les données sont ensuite segmentées en fichiers .wav correspondant aux différentes tâches du protocole, puis renommées selon une nomenclature spécifique permettant de conserver certaines informations tout en préservant l'anonymat des locuteurs. La phase d'annotation des données peut alors commencer : celles-ci sont d'abord transcrites en transcription orthographique standard (SOT, *Standard Orthographic Transcription*), sous Praat (Boersma & Weenink 2017) qui permet un alignement des annotations sur le signal sonore. Les annotateurs respectent un ensemble de conventions que nous ne détaillerons pas ici, mais qui sont précisément décrites par Durand et Przewozny (2015 : 79-83). Si la transcription orthographique peut sembler triviale de prime abord, il s'agit en fait d'une véritable analyse du signal sonore, et un soin tout particulier doit lui être apporté puisque de sa qualité et de sa rigueur dépendront la pertinence et la précision d'autres codages ultérieurs. C'est notamment le cas des codages de la rhoticité et du r de *sandhi* qui ont été mis en place au sein du programme PAC (voir Viollain 2014 : 333-335 et troisième partie).

De surcroît, il pourrait sembler étrange que des corpus qui se veulent phonologiques n'incluent pas d'annotations phonémiques par défaut, mais la SOT possède plusieurs avantages par rapport à ce type d'annotations. En effet, l'utilisation d'annotations phonémiques présuppose soit que l'on utilise une variété standard comme référence pour les transcriptions, soit que l'on utilise le système de la variété enregistrée. Si l'on opte pour la première solution, il est probable que certaines caractéristiques de la variété transcrite ne soient pas mises au jour, comme par exemple des différences dans la distribution lexicale des phonèmes. Si l'on choisit la seconde, il devient alors extrêmement difficile de garantir la comparabilité des données. On peut également insister sur le fait que les transcriptions orthographiques, qui demandent déjà beaucoup

de temps, sont relativement peu chronophages par rapport aux transcriptions phonologiques : Gut et Voorman avancent que la transcription phonologique d'une minute de données peut prendre jusqu'à une heure (2014 : 24). La SOT, quant à elle, présente l'avantage de représenter les données sans *a priori* sur le système phonologique des locuteurs, et permet aussi l'utilisation des corpus par tous les chercheurs, quelle que soit leur spécialité ou le type d'analyses qu'ils souhaitent conduire (Durand & Pukli 2004 : 39-40). Qui plus est, elle n'empêche aucunement d'avoir recours à des transcriptions phonologiques ultérieures, manuelles ou automatisées (voir par exemple Chatellier, 2016 : 221).

À la lumière de ces éléments méthodologiques, il apparaît que le programme PAC dispose bien de petits corpus spécialisés, pensés pour l'étude d'une variété régionale spécifique et de phénomènes phonético-phonologiques particuliers, qui constituent bel et bien des corpus phonologiques selon la définition de Gut et Voorman (voir partie 1), au sens où ils incluent des données primaires, des données secondaires variées (annotations prosodiques, segmentales, supra-segmentales, mais également potentiellement syntaxiques, voir Buscail 2013), et des métadonnées en ce qui concerne le corpus en lui-même et les locuteurs qui le composent. Au surplus, les corpus PAC sont accessibles aux chercheurs qui en font la demande, à la différence de nombreuses ressources orales.

De plus, ces petits corpus, dont beaucoup ont été constitués pour des mémoires de Master ou des thèses de doctorat, ont intégré les contraintes de temps et de financement qui s'imposent aux chercheurs. En effet, ces corpus sont le plus souvent le fruit d'une enquête menée individuellement par le chercheur lors d'un séjour de recherche court préparé en amont. Le corpus PAC-LVTI fait figure d'exception au sens où il a été constitué au cours de 4 enquêtes successives entre 2012 et 2014, parfois par un chercheur seul, parfois par une équipe de chercheurs, ce qui en fait d'ailleurs le plus grand corpus en nombre de locuteurs au sein de PAC. Autrement dit, intégrer les contraintes de temps et de financement dans son travail de terrain n'est pas nécessairement une limitation puisque, si les conditions sont réunies, les corpus PAC peuvent bien évidemment inclure plus de locuteurs que la douzaine de locuteurs minimum. Rappelons-le : il n'y a pas d'objectif de petitesse en soi.

Qui plus est, la comparabilité garantie par le protocole commun permet d'envisager les petits corpus spécialisés comme une grande base de données, un corpus général sur l'anglais oral contemporain. Et c'est la mise bout à bout, pour ainsi dire, des analyses issues de ces corpus qui dit quelque chose sur l'évolution contemporaine des variétés de l'anglais et la dynamique des systèmes à travers le monde, comme nous le montrerons justement dans la partie suivante. En ce sens, les ambitions et la méthodologie du programme PAC tentent bien de répondre à la réflexion sur la représentativité et la pertinence des corpus phonologiques que nous avons nous-mêmes développée précédemment. Nous citons Durand (2017) à ce propos :

> *no single corpus (in the sense of collection of recordings with one method) is sufficient to provide data rich enough to allow us to explore the systematic phonological and phonetic features of a given variety. But if we consider that an ideal modern corpus is in fact a large database containing different types of subcorpora we can establish a number of requirements that should ideally be met.*

Nous souhaitons saisir cette occasion de souligner que PAC n'est pas le seul programme à mener un travail de collecte et d'analyse de données authentiques important pour l'étude de la structure de la langue orale au sens large, et du système de l'anglais contemporain en particulier. En revanche, il dispose exclusivement de petits corpus, puisque son plus

grand corpus, le corpus PAC-LVTI Manchester, contient près de 100 000 mots, tandis que le corpus PAC Nouvelle-Zélande atteint pour sa part les 60 000 mots. On est loin du seuil symbolique du million ou du demi-million de mots que Kennedy fixe comme étant le minimum pour des analyses de type syntaxique ou lexicographique (Kennedy, 1998 : 68), ou des cent millions de mots du *British National Corpus*. Même si l'on ne prend en compte que les transcriptions issues de données orales dans le BNC, que l'on estime généralement autour de 10 % du corpus, on atteint les dix millions de mots. De son côté, et même s'il est de taille plus modeste, le *London-Lund Corpus* (500 000 mots) force le respect pour un corpus sur l'anglais oral. Néanmoins, il est encore souvent difficile aujourd'hui d'accéder aux données orales contenues dans ces corpus, ce qui, d'après la définition de Gut et Voorman (2014), et malgré leur intérêt en tant que corpus oraux, n'en fait pas des corpus phonologiques. C'est aussi malheureusement le cas du *Spoken Corpus of British English* ou du *Lancaster/IBM Spoken English Corpus*, même si des projets de numérisation de ses données ont vu le jour, comme le projet Aix-MARSEC (Auran *et al.* 2004). De plus, il s'agit généralement de corpus généraux, bien qu'ils se concentrent sur l'anglais oral, par opposition aux corpus spécialisés, car leurs données incluent des locuteurs de variétés différentes de l'anglais.

Il existe toutefois d'autres corpus phonologiques plus spécialisés, comme le NECTE (*Newcastle Electronic Corpus of Tyneside English*), sur lequel reposent plusieurs études et qui est accessible à la communauté des chercheurs. Durand (2017) le considère d'ailleurs comme un exemple de corpus phonologique réussi :

> *One good example for British English is the Diachronic Electronic Corpus of Tyneside English (DECTE, Beal, Corrigan, Mearns and Moisl 2014), which was built between 2000 and 2005 and is an extension of the Newcastle Electronic Corpus of Tyneside English (NECTE). NECTE is what is called a legacy corpus based on data collected for two sociolinguistic surveys conducted on Tyneside, north-east England, in c.1969-1971 and 1994, respectively. The authors have been pioneers in the construction of a unique electronic corpus of vernacular English which is aligned, tagged for parts of speech and fully compliant with international standards for encoding text, and the continuing work on the subcorpora now included within DECTE is of interest to all projects having to deal with recordings and metadata stretching back in time.*

C'est à cette catégorie qu'appartiennent nos corpus individuellement, puisqu'il s'agit bien de corpus spécialisés, de référence pour une variété de l'anglais spécifique.

Ce type de corpus est précieux, notamment pour retracer l'évolution d'une variété. Nous pensons aussi par exemple aux corpus dont dispose le projet AusTalk[5] sur l'anglais australien, ou le projet ONZE[6] sur l'anglais néo-zélandais. Néanmoins, il nous faut tout de même souligner qu'au-delà de ne traiter qu'une seule variété de l'anglais, les corpus dont disposent ces projets ne sont pas toujours véritablement comparables entre eux, et ces corpus n'ont pas toujours été constitués pour être des corpus linguistiques. C'est le cas justement de ONZE, qui est un projet dialectologique coordonné notamment par Elizabeth Gordon, Jennifer Hay et Margaret Maclagan et soutenu par l'Université de Canterbury à Christchurch en Nouvelle-Zélande. Il dispose de nombreuses archives orales enregistrées avec plusieurs centaines de locuteurs et notamment les premiers locuteurs du *NZE* (*New Zealand English*) nés dans les années 1850. Parmi ces archives, il y a la fameuse *Mobile Unit* qui est en fait un van qui a servi à parcourir la Nouvelle-Zélande afin de recueillir au départ de la musique, puis au fur et à mesure de plus en plus d'enregistrements avec les habitants des régions traversées. C'est en 1989 que l'Université de Canterbury a fait l'acquisition de ces enregistrements et a commencé à les transcrire et à les analyser

(Gordon *et al.*, 2004b) pour retracer l'évolution de l'anglais néo-zélandais depuis les premières vagues de colonisation dans la seconde moitié du XIX[e] siècle, et notamment les changements affectant les voyelles antérieures brèves (Gordon *et al.* 2004a, voir troisième partie). Des travaux majeurs se sont fondés sur ces données, et notamment la thèse de Trudgill (2004) sur l'émergence de nouveaux dialectes en contexte colonial (*new-dialect formation*).

La méthodologie adoptée au sein du programme PAC garantit pour sa part une comparabilité des données issues de variétés différentes de l'anglais. La comparaison et l'analyse de multiples variétés doit permettre à terme de disposer d'une grande base de données générale sur l'anglais oral, dans une démarche qui n'est pas sans rappeler celle des corpus ICE[7] (*International Corpus of English*). Toutefois, là encore, la méthodologie PAC se distingue de celle d'ICE en ce sens qu'elle permet de disposer de petits corpus spécialisés comparables et disponibles pour la communauté des chercheurs autorisant l'analyse en profondeur de nombreux phénomènes linguistiques, et notamment phonético-phonologiques, en lien avec les informations sociolinguistiques récoltées sur les locuteurs et sur la base d'une tire de transcription orthographique standard permettant une recherche automatique de certains phénomènes. Les corpus ICE, bien qu'ils soient constitués selon un protocole commun, ne récoltent que peu de métadonnées sur les corpus en tant que tels, et surtout sur les locuteurs qui les composent. Cela limite grandement l'exploitation sociolinguistique qui peut être entreprise de ces données.

Pour conclure cette partie consacrée aux ambitions et à la méthodologie propres au programme PAC, nous souhaitons dire quelques mots des deux corpus dont certains des résultats seront présentés en détail ci-après. Le premier est le corpus PAC Nouvelle-Zélande, issu d'une enquête menée à Dunedin, la capitale de la région de l'Otago, en décembre 2010. Dans sa version finale, il compte 13 locuteurs : 8 femmes et 5 hommes, ce qui en fait un corpus assez équilibré par rapport à d'autres, quand bien même il n'atteint pas la parité. Sur les 13 locuteurs sélectionnés à Dunedin, on compte 2 femmes et 1 homme âgés de 18 à 20 ans, 3 femmes et 2 hommes âgés de 43 à 51 ans, et 3 femmes et 2 hommes âgés de plus de 60 ans (65 à 76 ans), si bien que le corpus final comprend un échantillonnage générationnel assez équilibré.

Le corpus est en revanche peu représentatif en termes socio-économiques, notamment parce que 10 des 13 locuteurs sont en réalité voisins et ont été enregistrés dans le même quartier résidentiel, à savoir Maori Hill, situé à la frontière nord de la ville de Dunedin. Ce quartier est considéré aujourd'hui comme un quartier exclusif regroupant des gens aisés mais il ne l'a pas toujours été. Nombre de ses résidents, et notamment des locuteurs du corpus PAC Nouvelle-Zélande, s'y sont installés bien avant que Maori Hill ne devienne un quartier huppé, et appartiennent donc plutôt à la *middle class*, voire *lower middle-class*. D'autres appartiennent à l'*upper middle-class*. Quoi qu'il en soit, cela n'offre que très peu de diversité en termes de profils socio-économiques.

Le second est le corpus PAC-LVTI Manchester. Comme son nom l'indique, il s'agit du premier corpus au sein du programme PAC pour lequel le questionnaire LVTI a été intégré à l'entretien guidé des locuteurs. Sa construction a débuté en 2012, et il compte aujourd'hui 67 locuteurs (36 femmes et 31 hommes) du comté métropolitain du Greater Manchester. Il inclut des locuteurs avec des profils variés, que ce soit en termes d'âge ou de profil socio-économique. Les locuteurs retenus pour les analyses, et sur lesquels nous nous concentrerons ci-après, sont au nombre de 31, et appartiennent à trois tranches d'âge distinctes (20-36, 40-50, 55+), deux genres (hommes et femmes) et trois catégories

socio-économiques dénommées G1, G2 et G3 (que l'on peut respectivement associer aux *working-class*, *upper working-class/lower middle-class* et *middle class*, pour reprendre une terminologie souvent employée dans les travaux en sociolinguistique[8]).

3. De petits corpus pour une grande base de données : quels résultats ?

Suite à ces considérations épistémologiques et méthodologiques, il nous faut montrer concrètement comment l'étude phonétique, phonologique et sociolinguistique des corpus PAC, individuellement et collectivement, livre des informations pertinentes quant à la dynamique des systèmes de l'anglais à travers le monde. Pour ce faire, nous avons choisi ici de nous concentrer sur deux phénomènes majeurs pour l'étude du système phonético-phonologique de l'anglais : la rhoticité d'une part, et les changements vocaliques d'autre part.

Nous allons montrer comment la jonction des résultats obtenus à partir des corpus PAC Lancashire, Boston et Nouvelle-Zélande d'une part (Navarro 2013, 2016 ; Viollain 2010, 2014), et des corpus PAC-LVTI Manchester (Chatellier 2016) et PAC Nouvelle-Zélande (Viollain 2014) d'autre part, éclairent ces deux phénomènes respectivement et suggèrent qu'il n'existe pas un système de groupe global pour les locuteurs de ces corpus mais plutôt plusieurs systèmes, et donc des espaces régionaux et nationaux en phonologie, notamment selon des axes nord-sud liés à l'histoire de la diffusion de l'anglais à travers le monde. Nous n'aurons bien évidemment pas le loisir ici de détailler tous nos résultats, et renverrons, le cas échéant, aux travaux concernés. Nous souhaitons surtout donner une idée concrète à notre lecteur des potentialités des petits corpus du programme PAC, et de la variété des angles d'étude qu'ils offrent, en le laissant libre d'explorer plus avant tels ou tels travaux et, peut-être, de demander à avoir accès à tel ou tel corpus pour mener ses propres recherches.

Commençons par le phénomène de rhoticité. Nous citons ici l'introduction de Navarro (2016 : 5) à son ouvrage consacré au /r/ en anglais fondé sur les résultats de l'étude de plusieurs corpus PAC :

> Parmi les caractéristiques phonologiques qui distinguent les différentes variétés d'anglais parlées dans le monde, la distribution du /r/ occupe une place centrale. Il existe une dichotomie majeure entre les variétés dites rhotiques (comme l'accent américain standard, General American), dont les locuteurs prononcent le /r/ partout où il est présent sous forme d'un <r> orthographique, et les variétés non-rhotiques (comme l'accent standard sud-britannique, Received Pronunciation) où le /r/ n'est prononcé que lorsqu'il est immédiatement suivi d'une voyelle. Cette distinction est souvent la première citée par les anglicistes lorsqu'on les interroge sur les différences entre les accents britannique et américain.

Autrement dit, dans les variétés rhotiques, schématiquement, les /r/ se prononcent en position d'attaque et de coda de syllabe (pré-vocalique et non pré-vocalique dans la littérature), c'est-à-dire dans *rap*, *trap*, *strap* mais aussi dans *car*, *cart* et *carving*. Dans les variétés non-rhotiques en revanche, les /r/ en coda de syllabe ne se prononcent pas, c'est-à-dire que les mots *car*, *cart* et *carving* sont réalisés sans [r][9]. Les petits corpus PAC permettent de dresser un portrait actualisé de ces deux grandes familles, ce qui met en perspective l'évolution historique de /r/ dans le monde anglophone et montre la variabilité de ce phénomène et sa sensibilité aux enjeux sociolinguistiques et identitaires.

De fait, historiquement, il n'existait qu'une seule famille de variétés de l'anglais : les variétés rhotiques. Les variétés non-rhotiques que nous connaissons aujourd'hui, comme en Angleterre bien sûr, mais aussi en Nouvelle-Zélande ou en Australie, sont le résultat d'un processus historique de dérhoticisation[10] dont les origines, le scénario et la chronologie exacts font toujours débat au sein de la communauté des phonologues. Il existe toutefois un consensus selon lequel les premiers signes de l'affaiblissement et de la perte du /r/ en coda seraient apparus dès le XV^e^ siècle dans les comtés situés à l'est de Londres. À partir de ce moment-là, l'effacement du /r/ en coda se serait produit par étapes, opérant d'abord en position pré-consonantique (*cart*, *carving*) avant d'affecter la position strictement finale (*car*), et créant par conséquent une forme d'instabilité et d'hybridité au sein du système des locuteurs.

Ce changement se serait ensuite répandu progressivement mais inégalement au cours des siècles suivants, c'est-à-dire sans affecter l'ensemble des territoires de la même manière et au même moment, puisque certaines zones du nord-ouest et du nord-est de l'Angleterre sont restées rhotiques, comme le Lancashire ou le Northumberland, et que l'Écosse et l'Irlande n'ont initialement pas été affectées par ce changement (Wells 1982 : 212-227 ; Navarro 2016 : 49-60). Parallèlement, en Amérique du nord, colonisée depuis le sud-est de l'Angleterre par des locuteurs déjà variablement rhotiques, des normes non-rhotiques se sont imposées à certains endroits, comme à Boston ou à New York notamment, ou dans certains états sudistes comme la Virginie, la Louisiane, le Mississippi ou la Géorgie (Navarro 2016 : 60-92). Au contraire, sur le reste du territoire nord-américain, une norme rhotique s'est imposée, si bien que l'accent de référence américain (*General American*, voir plus haut) est aujourd'hui clairement rhotique.

Ce que les corpus, et notamment les corpus PAC, ont contribué à démontrer est la variabilité et l'instabilité de ce phénomène de rhoticité selon les régions, et donc la pertinence, encore aujourd'hui, d'étudier la dynamique des systèmes de l'anglais en ce qui concerne ce changement. En effet, l'exploitation de corpus à travers le monde anglophone indique qu'il existe une vraie tectonique des plaques rhotique et non-rhotique : des zones historiquement rhotiques semblent céder face à la norme non-rhotique en Angleterre (PAC Lancashire, Navarro 2013) mais également en Écosse (PAC Ayrshire, Pukli 2004, 2015) et en Nouvelle-Zélande (PAC Nouvelle-Zélande, Viollain 2014), tandis que des zones historiquement non-rhotiques semblent, elles, céder face à la norme rhotique aux États-Unis (PAC Boston, Viollain 2010 ; Navarro 2013).

Afin d'étudier la rhoticité dans le système phonologique des locuteurs des corpus PAC, un système de codage de la variable (R) a été élaboré et appliqué de manière systématique aux enregistrements réalisés avec l'ensemble des locuteurs dans l'ensemble des tâches du protocole (Viollain 2014 : 333-334). En effet, toutes ces tâches (listes de mots, texte, phrases courtes, conversations) comprennent des sites potentiels de réalisation de /r/, c'est-à-dire des /r/ en attaque et en coda de syllabe. Rappelons que ce qui fait la rhoticité d'un système phonologique est la réalisation des /r/ en coda.

Deux critères principaux ont présidé à son élaboration : il doit être accessible à la communauté des chercheurs, et notamment aux non-spécialistes du domaine concerné (en l'occurrence la phonologie), et permettre une approche globale des données. C'est pourquoi le codage de la rhoticité est simple et ne note pas de détails trop précis, mais essaie de dégager des contextes ou des cas de figure distincts pouvant se soumettre à des interprétations ultérieures plus poussées. Le codage est alphanumérique et prend en

compte les deux paramètres fondamentaux suivants : (a) présence ou absence d'un [r] et (b) position de ce /r/ en attaque (ex : *raspberry*) ou en coda de syllabe (ex : *more*).

Deux champs supplémentaires (3 et 4) servent à donner plus d'informations sur le contexte droit des /r/ en coda de syllabe et viennent donc compléter les informations élémentaires fournies par les champs 1 et 2. Sont-ils suivis par une frontière de mot (ex : *far*), par une consonne tautosyllabique (ex : *farm*) ou encore par une consonne ou plusieurs consonnes elles-mêmes suivies d'une ou plusieurs voyelles (ex : *forty*) ? C'est précisément ce que code le champ 3. Le champ 4 est quant à lui conçu pour fournir des informations encore plus précises sur les /r/ en coda qui sont suivis par une frontière de mot. En cela, le champ 4 est optionnel car il ne peut être utilisé que si l'indice « 1 » a été attribué à la séquence étudiée dans le champ 3. Il se focalise sur ce qui suit le /r/ analysé : est-il suivi par un mot ayant une attaque vocalique (ex : *far out*), et dans ce cas il s'agit d'un cas de liaison (nous allons y revenir) ? Ou par un mot ayant une attaque consonantique (ex : *for me*) ? Ou encore constitue-t-il une fin de groupe, et dans ce cas il est suivi par une frontière forte (ex : *are you staying here ?*) ? Le champ 4 code ces différents cas de figure. Considérons un exemple concret, soit la séquence *isn't there ?* prononcée ['ɪznt ðeə]. Elle serait codée <isn't there0213 ?> dans laquelle le « 0 » indique que le /r/ n'est pas réalisé, le « 2 » que ce /r/ est en coda, le « 1 » que ce /r/ de coda est suivi d'une frontière de mot (_#), et enfin le « 3 » que cette frontière de mot constitue une frontière forte, en l'occurrence la fin d'une phrase interrogative (_##).

Afin d'illustrer les résultats obtenus par le programme PAC sur cette question, nous allons nous fonder sur le corpus d'enregistrements constitué dans l'isoglosse de la rhoticité en Nouvelle-Zélande, à savoir dans l'extrême sud de l'île du Sud (Dunedin, Otago), historiquement colonisé par une majorité de locuteurs écossais, comme la région du Southland, encore plus au sud (McKinnon 1997, cité dans Gordon *et al.* 2004a : 444-445 ; Viollain 2014 ; Przewozny & Viollain 2015, 2016). Rappelons que l'accent de référence néo-zélandais, le *General NZE*, est lui décrit comme une variété non-rhotique de l'anglais.

Au total, ce sont 13 031 codages de la variable (R) que nous avons extraits du corpus PAC Nouvelle-Zélande, dont 4 370 (soit environ 33,5 %) correspondent aux /r/ en attaque de syllabe. Les 8 661 codages restants correspondent donc aux /r/ en coda dans l'ensemble des tâches du protocole effectuées par nos locuteurs, soit à l'indice « 2 » en deuxième position du codage.

Or, l'indice « 2 » de notre codage spécifique inclut un phénomène qui n'est pas de la rhoticité, à savoir la liaison (*linking r*). C'est là l'un des pièges, pour ainsi dire, des systèmes de codage qui impliquent de traiter les phénomènes de manière binaire : en l'occurrence, le /r/ est-il en position d'attaque de syllabe ou dans une autre position ? Le problème est que dans le cas de la liaison, le mot liaisonnant comporte bien un <r> orthographique en coda (*the ca<r> is*), mais la réalisation d'un [r] dans ce contexte n'est pas une manifestation de la rhoticité du système. En effet, lorsqu'un /r/ en coda est directement suivi par un mot (*car alarm*, Hay & Maclagan 2010 : 41) ou un morphème (*fear ing*, Hay & Maclagan 2010 : 41 ; *hearing*, Hay *et al.* 2008 : 18) ayant une attaque vide et commençant donc par une voyelle, ce /r/ peut être réalisé, en *NZE* comme dans l'ensemble des variétés non-rhotiques. Par conséquent, en extrayant nos codages à partir de l'indice « 2 » en deuxième position (par exemple *in the car0213*), nous extrayons aussi les occurrences de liaison (par exemple *here1211 it is*), alors même que ces deux phénomènes disent des choses différentes sur le système du locuteur étudié[11].

Nous en profitons pour dire quelques mots de ce phénomène de liaison, dont nous n'aurons pas le loisir de discuter plus avant ici, car il est particulièrement intéressant au regard de la notion de petitesse. En effet, on pourrait être tenté de dire qu'il s'agit d'un « petit » phénomène au sens où il est assez rare (environ 2 000 codages du r de *sandhi*[12] au total pour les 3 corpus PAC Lancashire, Boston et Nouvelle-Zélande), contrairement à son « équivalent[13] » en français (plusieurs dizaines de milliers de codages de la liaison dans la base de données PFC, Laks *et al.* 2014), beaucoup plus fréquent. Il est donc difficile de l'étudier systématiquement et à grande échelle, à moins de mettre en place un protocole et un outil de codage adaptés qui permettent de comparer les résultats issus de plusieurs enquêtes.

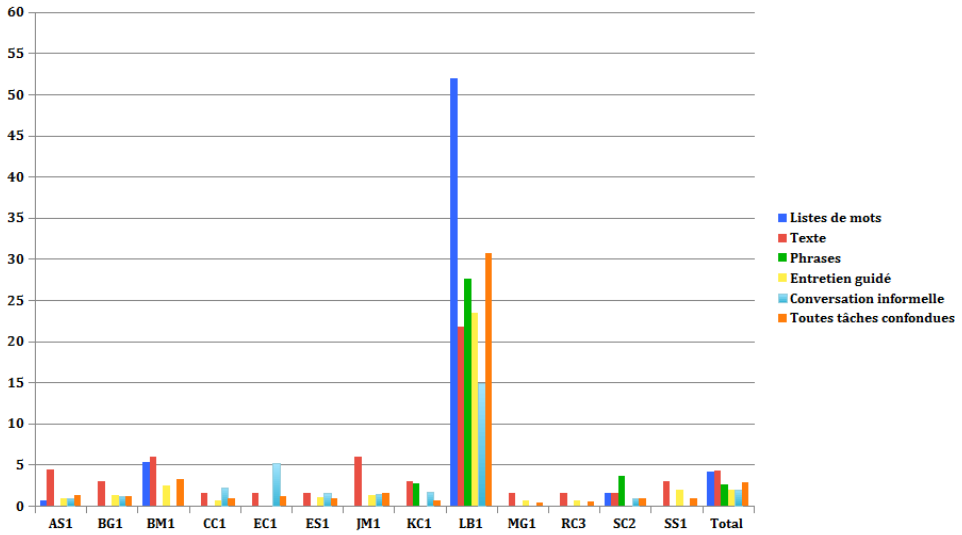
C'est ce qu'a fait le programme PAC, ce qui a permis de mettre à l'épreuve de données authentiques les modélisations théoriques existantes et de montrer non seulement les contraintes morphosyntaxiques et prosodiques particulières qui pèsent sur ce phénomène (voir Durand *et al.* 2014, 2015), mais également l'importance des usages et de la mémorisation de séquences dans sa production, à l'image de ce qu'ont défendu Cox *et al* . (2014) pour le cas particulier de l'anglais australien. D'ailleurs, les potentialités des petits corpus PAC ne s'arrêtent pas aux comparaisons entre variétés de l'anglais puisqu'il est également possible, sur la base d'une comparaison avec les résultats obtenus par le programme PFC sur la liaison en français, de réfléchir à ces deux phénoménologies (voir Navarro & Viollain en préparation), notamment à des fins didactiques pour traiter la production des apprenants francophones de l'anglais (voir Herry-Bénit *et al.* en préparation).

Revenons-en à la rhoticité dans notre corpus néo-zélandais. Il nous faut écarter les codages spécifiques de la liaison (*<r(e)>0211 et <r(e)>1211*) pour ne garder que les contextes pertinents pour l'étude de la rhoticité, à savoir :

- le contexte _C_{1-n}#, soit lorsqu'un /r/ est suivi d'une ou plusieurs consonnes elles-mêmes suivies d'une frontière de mot, comme dans *short*, codé *shor022t* si le /r/ n'est pas réalisé ;
- le contexte _C_{1-n}VX#, soit lorsqu'un /r/ est suivi d'une ou plusieurs consonnes, elles-mêmes suivies d'une ou plusieurs voyelles et d'une frontière de mot, comme dans *forty*, codé *for023ty* si le /r/ n'est pas réalisé ;
- le contexte _#C, soit lorsqu'un /r/ est suivi d'une frontière de mot et que le mot suivant commence par une consonne, comme dans *for me*, codé *for0212 me* si le /r/ n'est pas réalisé ;
- le contexte _##, soit lorsqu'un /r/ est suivi d'une frontière forte (fin de groupe prosodique par exemple) comme dans *is it here?*, codé *is it her0213e?* si le /r/ n'est pas réalisé.

En écartant les codages spécifiques de la liaison, il nous reste 7 055 codages. La figure ci-après (voir figure 1) présente le taux de rhoticité individuel (%) de chacun de nos locuteurs dans chaque tâche du protocole et toutes tâches confondues, ainsi que le taux de rhoticité moyen pour l'ensemble de nos locuteurs pour chaque tâche du protocole et toutes tâches confondues (dernière colonne à droite).

Figure 1.

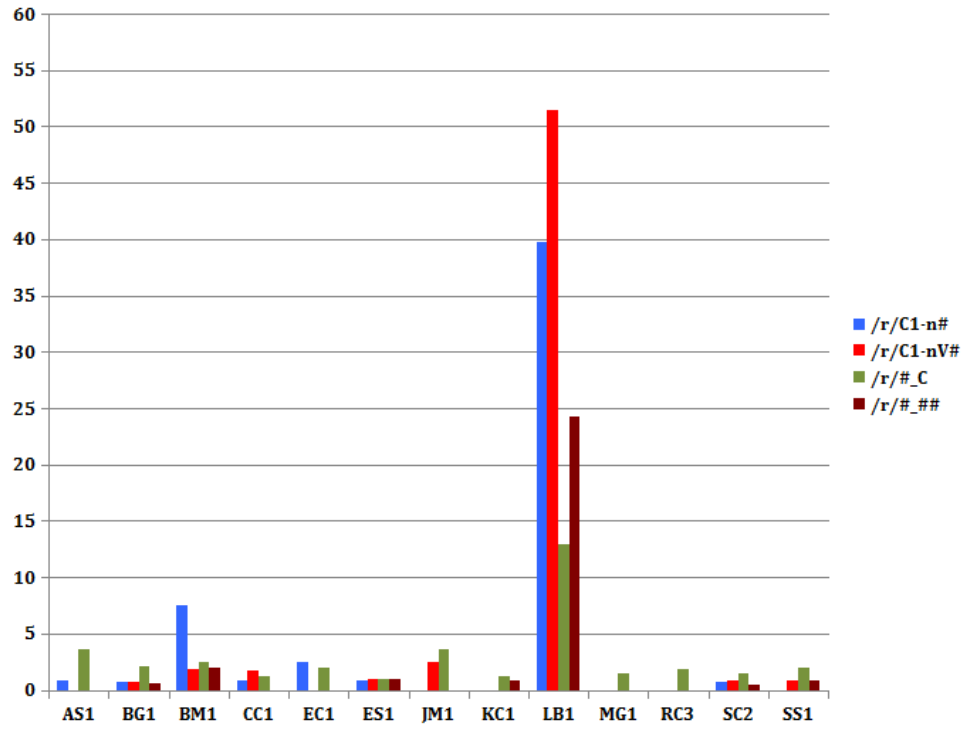


Le premier constat qui s'impose est que la rhoticité est variable selon les locuteurs et qu'aucun locuteur n'a un taux de rhoticité nul. En effet, les taux de rhoticité individuels moyens (en orange dans le diagramme), c'est-à-dire toutes tâches confondues, varient de 0,43 % (MG1) à 30,7 % (LB1) en passant par des valeurs intermédiaires comme 0,95 % (CC1). Le taux de rhoticité moyen toutes tâches confondues pour tous les locuteurs est de 2,9 % (colonne à l'extrême droite du tableau) mais la grande majorité de nos locuteurs, soit 11 locuteurs sur les 13 que compte notre corpus, a un taux de rhoticité largement en dessous de cette moyenne générale (moins de 1,5 %, soit le taux de rhoticité de JM1). Seuls BM1 (3,25 % de taux de rhoticité moyen toutes tâches confondues) et LB1 ont un taux au-dessus de la moyenne générale et il semble d'ailleurs évident que ce sont eux qui la font monter. En comparaison, le taux de rhoticité de BM1 est deux fois supérieur à celui de JM1, tandis que celui de LB1 est vingt fois supérieur à celui de JM1. Ces résultats indiquent une non-rhoticité stable et, par conséquent, une rhoticité résiduelle chez la grande majorité des locuteurs de notre corpus, tandis que nos codages révèlent une rhoticité variable chez BM1 et LB1.

Toutefois, maintenant que nous avons identifié BM1 et LB1 comme des locuteurs variablement rhotiques, ou hybrides, au sein de notre corpus, il nous faut tenter d'expliquer cette variabilité caractéristique de leur système. Un des scénarios les plus plausibles pour en rendre compte est un processus de dérhoticisation en cours qui implique que ces locuteurs, qui étaient à l'origine rhotiques, ou plus rhotiques, soient en train de converger vers la norme non-rhotique néo-zélandaise, incarnée par exemple par les 11 autres locuteurs de notre corpus.

Afin de déterminer si l'environnement morphosyntaxique joue un rôle dans la réalisation ou la non-réalisation des /r/ en coda, et vérifier si, là aussi, nous observons un écart significatif entre BM1 et LB1 d'un côté et les 11 autres locuteurs de notre corpus de l'autre, nous proposons la figure ci-après (voir figure 2) qui indique la proportion de /r/ réalisés et de /r/ non réalisés (%) dans chaque environnement (_$C_{1\text{-}n}$# ; _$C_{1\text{-}n}$VX# ; _#C et _##) pour chaque locuteur individuellement toutes tâches du protocole confondues.

Figure 2.



Il apparaît que BM1 et LB1 ont un taux de réalisation de /r/ dans l'ensemble des contextes supérieur aux autres locuteurs du corpus, mais que la rhoticité de BM1 peut être considérée comme quasi-résiduelle là où celle de LB1 est loin d'être marginale. Nous notons au surplus que, comme LB1, BM1 maintient plus de /r/ en position pré-consonantique (*world*, *party*, en bleu et en rouge) qu'en position strictement finale (*fur*, en marron). Les /r/ en position strictement finale seraient donc, d'après nos données, les premiers à disparaître dans ce processus de dérhoticisation, ce qui fait nécessairement écho à l'enquête fondatrice de Labov sur la réintroduction du /r/ à New York (1966). Dans cette célèbre enquête, Labov étudiait le processus inverse à celui que nous observons, c'est-à-dire le processus de retour à la rhoticité, de réintroduction du /r/ à New York et notait que la réalisation d'un /r/ était plus fréquente en position finale (*floor*, dans l'exemple utilisé par Labov lui-même) qu'en position pré-consonantique (*fourth*). Labov (1972 : 66) postule qu'il existe une contrainte de nature phonologique qui affecte différemment ces deux contextes et qui explique que le /r/ soit plus rapidement réintroduit en position finale qu'en position pré-consonantique, et donc, dans le cas qui nous intéresse ici, qui explique que le /r/ soit plus rapidement effacé en position finale qu'en position pré-consonantique[14].

Il semble en effet que nos données soient l'image miroir des données récoltées et analysées par Labov (1966), ou d'autres chercheurs enquêtant à New York (Mather 2010), à savoir une ville nord-américaine où un retour à la rhoticité est attesté. Ces éléments pointent donc vers des mécanismes parallèles inverses de réintroduction et d'effacement progressifs du /r/ dans les différentes variétés de l'anglais étudiées, que les corpus PAC Boston (Viollain 2010 ; Navarro 2013) et PAC Nouvelle-Zélande (Viollain 2014) nous ont permis de mettre au jour, en s'inscrivant dans la continuité des travaux fournis

notamment par Irwin & Nagy (2007, 2010) et Gordon *et al.* (2004a) et Trudgill (2004) sur ces deux variétés respectivement, qui servent de base diachronique à nos observations.

Qui plus est, en étudiant les profils sociolinguistiques des locuteurs PAC de Boston et de Dunedin, il est apparu que la conscience des enjeux sociolinguistiques liés aux identités régionales et aux accents locaux est un facteur crucial dans ces processus de rhoticisation ou dérhoticisation progressifs. En effet, à Boston, deux locuteurs, DG1 et JT1, ont montré des systèmes assez similaires en ce qui concerne la rhoticité, alors même que leurs profils sociolinguistiques respectifs sont radicalement différents. D'un côté nous avons DG1 (Viollain, 2010 : 161-162), âgée de 54 ans, infirmière dans un *college* de l'agglomération de Boston, fille d'un immigré grec et d'une mère au foyer n'ayant pas fait d'études, qui montre un taux de rhoticité très faible, ce qui fait d'elle une locutrice bostonienne prototypique. De l'autre, nous avons JT1 (Viollain 2010 : 170-171), âgé de 25 ans, qui travaille dans le système judiciaire américain en tant que référent pour les affaires impliquant des mineurs, et qui a donc fait des études supérieures. Il a également cofondé avec deux amis, BH1 et MT1 qui ont tous les deux poursuivi des études supérieures à l'université également, une société dénommée *No-R Lifestyle* qui crée et commercialise des T-shirts revendiquant la spécificité non-rhotique de Boston et reposant sur la fierté d'être bostoniens de ses clients. Ce locuteur présente un taux de rhoticité extrêmement faible également, alors même que ses origines et son parcours auraient pu très certainement le conduire à adopter la norme rhotique américaine. On observe donc un mouvement volontaire de différenciation de la part de JT1 qui n'est pas sans rappeler ce qu'avait observé Labov à Martha's Vineyard (1963).

À Dunedin, des éléments similaires émergent de l'étude approfondie du profil sociolinguistique des locuteurs PAC. En effet, BM1 et LB1 dont nous avons parlé précédemment, comptent parmi les 3 locuteurs du corpus ayant des liens étroits avec le Southland, la région à l'extrême sud de l'île du Sud de la Nouvelle-Zélande, connue nationalement pour sa rhoticité, son *Southland Burr* (Gordon & Maclagan 2008 : 66). Comme nous l'avons vu, BM1 et LB1 sont les deux seuls locuteurs à présenter de la rhoticité variable, et donc les signes d'un ajustement en cours vers la norme non-rhotique locale. De plus, LB1 est le seul locuteur à être né dans le Southland et à y retourner régulièrement pour voir ses parents étant donné qu'au moment de l'enregistrement, il était âgé de seulement 19 ans et encore étudiant. En soi, il n'est donc pas surprenant que BM1 et LB1 soient variablement rhotiques puisque leur profil sociolinguistique indique qu'ils ont très probablement un basilecte rhotique distinct de celui des autres locuteurs du corpus dont le basilecte est plus probablement non-rhotique. Ce qui est intéressant, c'est de constater qu'ils peuvent faire montre de cette caractéristique de manière consciente, comme dans les listes de mots où leur rhoticité s'exprime proportionnellement plus que dans les autres tâches du protocole, pour opérer une distanciation par rapport à la majorité et affirmer par conséquent une identité minoritaire.

Ces éléments suggèrent que le Southland pourrait être la dernière poche de rhoticité encore productive en Nouvelle-Zélande (voir Viollain 2014 : 706-715 pour le détail des calculs statistiques en ce qui concerne l'âge et l'origine géographique des locuteurs). EC1, la troisième locutrice du corpus ayant des liens étroits avec le Southland puisque ses deux parents en étaient originaires n'a, quant à elle, qu'une rhoticité extrêmement résiduelle. Professeure de français à la retraite au moment de l'enquête, et titulaire de nombreux diplômes en langues et en littérature, EC1 a indiqué dans son entretien guidé avec

l'enquêtrice avoir toujours fait attention à sa manière de s'exprimer et avoir le souvenir que ses parents attachaient une grande importance à l'école et aux études afin d'assurer son avenir. Cette conscience du poids de la norme, de la stigmatisation potentielle de certaines caractéristiques régionales, et finalement des enjeux sociolinguistiques liés aux identités locales, sont autant d'éléments qui contribuent à expliquer les trajectoires linguistiques des locuteurs des corpus PAC, surtout quand celles-ci laissent initialement et apparemment perplexe.

Nous souhaitons aborder à présent le second corpus qui nous servira à illustrer notre propos, le corpus PAC-LVTI Manchester, en particulier à travers la question de la dynamique de la variété mancunienne par rapport aux variétés du nord de l'Angleterre, dans un contexte où plusieurs travaux (Beal 2008 ; Watt 1998, 2002) ont avancé que des variantes septentrionales sont en cours de diffusion dans cette région. Il s'agit de variantes monophtonguées de /eɪ/ et /əʊ/ en anglais standard (soit les voyelles des ensembles lexicaux de FACE et GOAT), qui ne sont pas associées avec une aire urbaine en particulier, mais avec le nord de l'Angleterre en général. Cette situation fait écho au phénomène de nivellement dialectal, à savoir l'homogénéisation des variétés parlées dans une zone donnée, ainsi qu'à l'émergence des dialectes modernes de l'anglais, qui sont associés à des zones plus larges (souvent autour d'aires urbaines importantes) que celles des dialectes ruraux traditionnels (Trudgill 2001).

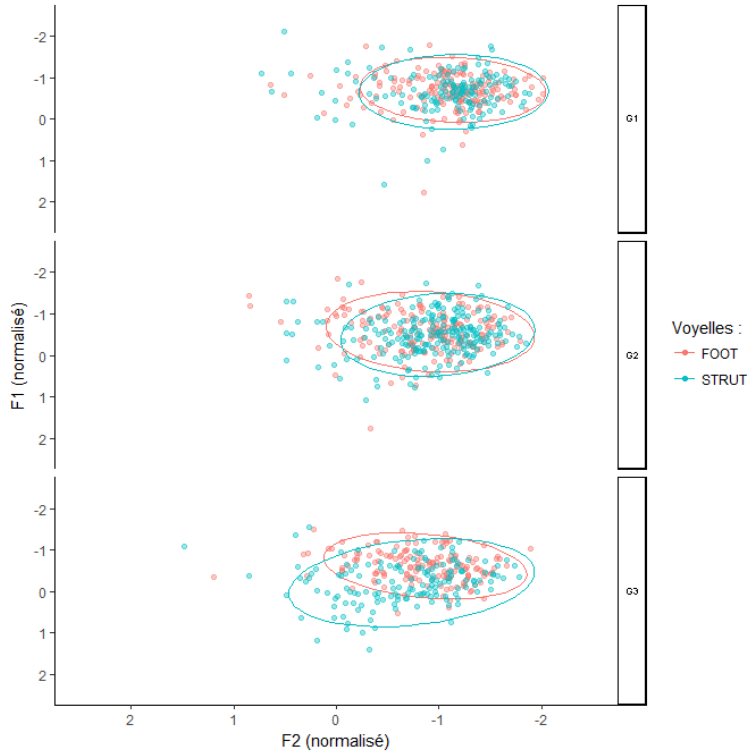
Manchester constitue un terrain de recherche intéressant à plus d'un titre : malgré l'importance de la ville sur les plans culturel, économique, et démographique, peu d'études de grande ampleur y ont été menées jusqu'à récemment. En outre, les travaux récents de Baranowski et Turton (2015) suggèrent que les variantes supralocales septentrionales ne sont pas attestées dans la variété mancunienne, qui ne serait donc pas impliquée dans le même changement vocalique que d'autres variétés du nord de l'Angleterre.

Afin de pouvoir décrire et rendre compte de l'évolution de la variété mancunienne, nous avons entrepris une analyse phonético-acoustique des formants des voyelles de nos locuteurs. Les mesures[15] ont été effectuées automatiquement en deux points de chaque voyelle (1/3 et 2/3 de la durée de la voyelle), à l'aide d'un script sous Praat, puis vérifiées manuellement avant d'être normalisées (procédure Lobanov).

La première caractéristique qu'il nous semble opportun de mentionner ici est la présence ou non d'un phonème /ʌ/, et plus précisément, la présence ou l'absence d'une opposition entre /ʊ/ et /ʌ/, ou FOOT[16] et STRUT. Il s'agit d'un trait que l'on observe dans l'ensemble des variétés du nord de l'Angleterre, puisque le phonème /ʌ/ est une innovation du sud de l'Angleterre, et le résultat de la division phonémique du phonème /u/ ou /ʊ/ du moyen anglais (Wells 1982 : 196-197 ; Beal 2008 : 131). La présence d'une opposition est, dans le nord de l'Angleterre, très fortement corrélée au profil socio-économique du locuteur :

> *broad working-class speakers certainly do not have any control of a FOOT vs. STRUT opposition, which is associated with 'good' speech only* (Wells 1982 : 351-352).

Figure 3.



Nos locuteurs mancuniens suivent cette tendance. La figure 3 représente l'ensemble des réalisations[17] de FOOT et STRUT pour chacune des trois catégories socio-économiques de notre corpus, ainsi que des ellipses de confiance qui contiennent 95 % des occurrences de chaque voyelle. Il apparaît clairement que les réalisations des deux phonèmes sont semblables pour les locuteurs des G1 et G2 alors que ce n'est pas le cas pour les membres du G3 qui correspond, rappelons-le, aux locuteurs les plus aisés de notre stratification socio-économique (voir partie 2). Chez ces derniers, STRUT présente une plus grande variabilité, et on relève un nombre plus important de réalisations ouvertes et centrales, correspondant à des réalisations telles que [ɐ] qu'on observe également en anglais britannique standard (Cruttenden, 2014 : 122). Des calculs statistiques (t-test) confirment ces observations : les différences entre les valeurs des deux premiers formants (F1 et F2) pour FOOT et STRUT ne sont significatives ni pour les locuteurs du G1, ni pour ceux du G2. En revanche, c'est le cas pour les locuteurs du G3, à la fois pour F1 et F2. Ces résultats suggèrent selon nous qu'il n'y a pas d'opposition entre FOOT et STRUT à un niveau basilectal (c'est-à-dire dans la partie « basse » du continuum socio-économique) dans la variété mancunienne[18].

La deuxième grande caractéristique vocalique qu'on relève dans les accents nordiques de l'anglais britannique est la répartition lexicale différente des phonèmes /a/[19] et /ɑ:/ de l'anglais britannique standard. Le second phonème a une répartition plus restreinte dans les variétés du nord, puisque l'allongement de /a/ dans certains contextes (devant une fricative sourde, notamment /f/, /θ/ et /s/) à partir du XVII^{e} siècle, est une fois encore une innovation des variétés du sud de l'Angleterre (*Pre-Fricative Lengthening*, Wells, 1982 : 203-206). De nombreux mots appartenant à l'ensemble lexical de BATH ont donc conservé une voyelle brève dans les accents septentrionaux. En revanche, contrairement à

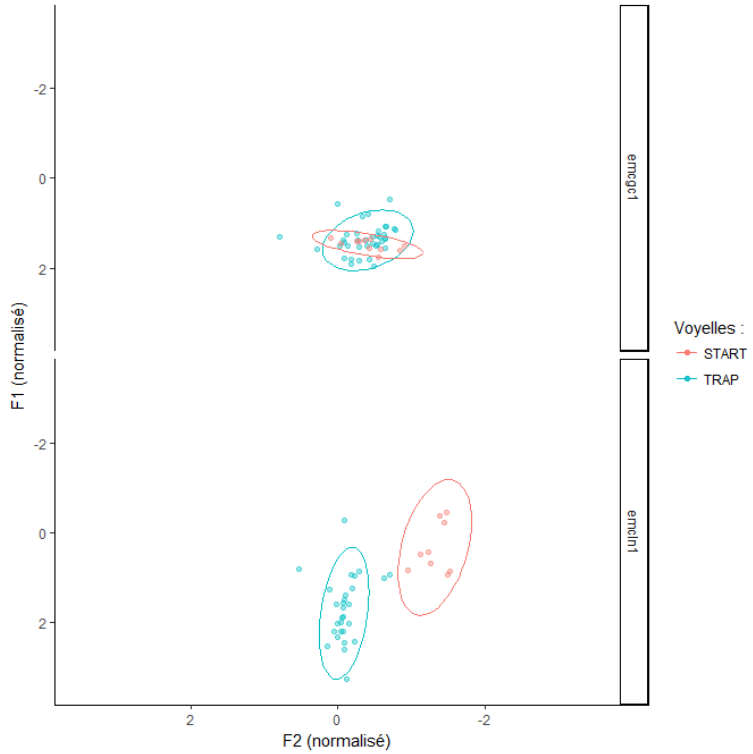
l'absence d'opposition entre FOOT et STRUT, la présence d'une voyelle brève pour les mots de BATH ne semble pas souffrir des mêmes évaluations sociolinguistiques :

> *Retention of a short vowel in BATH words extends much further up the social scale than does the retention of unsplit /ʊ/ ... There are many educated northerners who would not be caught dead doing something so vulgar as to pronounce STRUT words with [ʊ], but who would feel it to be a denial of their identity as northerners to say BATH words with anything other than short [a]* (Wells 1982 : 354).

Il arrive même que des locuteurs originaires du sud de l'Angleterre raccourcissent la voyelle qu'ils utilisent dans les mots appartenant à BATH suite à un séjour de longue durée dans le nord (Beal 2008 : 132).

Naturellement, les mesures formantiques ne pourront pas nous fournir d'informations précises quant à la longueur des voyelles, c'est pourquoi nous nous sommes tournés vers l'écoute des items lexicaux appartenant à TRAP et BATH, en particulier ceux issus des tâches de lecture qui, étant contrôlées, garantissent que chaque locuteur réalise un nombre minimum d'occurrences de chacune de ces voyelles[20]. Parmi nos 31 locuteurs, un seul, qui appartient d'ailleurs à la catégorie socio-économique la plus aisée de notre corpus, présente une distinction claire entre TRAP et BATH, avec une voyelle plus longue et plus postérieure pour ce dernier ensemble lexical. Pour les autres locuteurs, la majorité des items lexicaux de BATH est prononcée avec une voyelle brève, bien qu'une demi-douzaine de locuteurs semblent opérer une distinction entre les items 50 et 51 de la première liste de mots, c'est-à-dire *ants* et *aunts* (respectivement /ˈants/ et /ˈɑːnts/ en *General British*). Néanmoins, une locutrice en particulier semble hésiter et ne pas être certaine de la voyelle à utiliser pour *aunts*, alors qu'une autre prononce cet item avec une voyelle longue, avant de se corriger et de produire une voyelle brève. On remarque que les deux items diffèrent au niveau de la graphie et sont situés directement l'un après l'autre dans la liste, si bien qu'il est probable que ces distinctions soient le fait de la graphie. Malheureusement, ces items n'étant pas présents dans les conversations, il est impossible de vérifier plus avant cette hypothèse.

Figure 4.

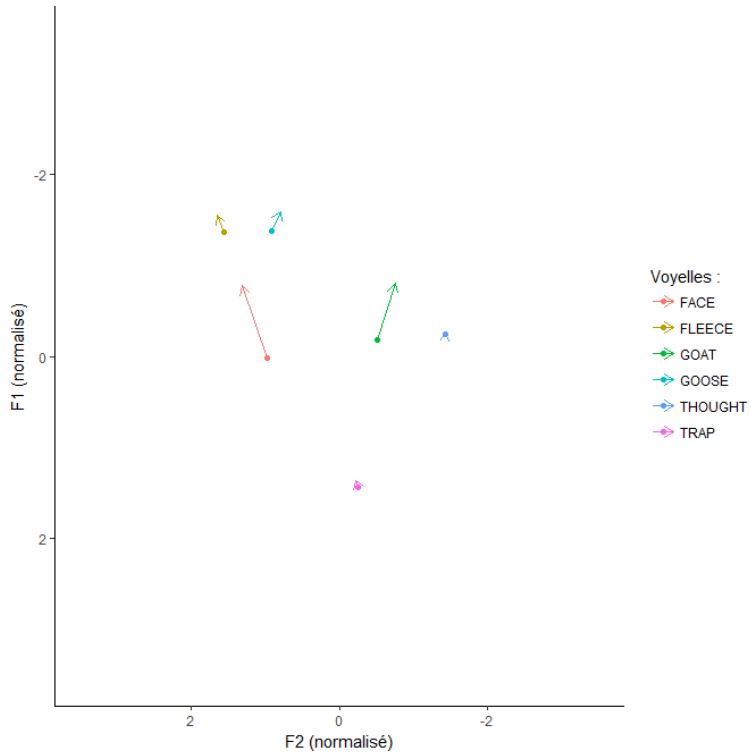


La nature de l'opposition /a/ *vs.* /ɑː/ mérite également notre attention. Même pour les locuteurs qui utilisent une voyelle brève pour les items lexicaux appartenant à BATH, il existe une voyelle ouverte, longue et non-arrondie pour les mots appartenant à PALM et surtout à START. Si cette opposition repose principalement sur la qualité des voyelles plutôt que sur la longueur en anglais standard (Cruttenden 2014 : 120), nos données montrent que les locuteurs de notre corpus se répartissent en deux groupes. D'un côté, parmi les membres du G3, plusieurs locuteurs présentent une distinction claire entre une voyelle brève centrale et une voyelle longue postérieure. C'est par exemple le cas de LN1, dont les réalisations de /a/ et /ɑː/ occupent deux zones bien distinctes dans l'espace vocalique (voir figure 4). De l'autre, la plupart des locuteurs du G1, à un niveau plus basilectal donc, produisent des réalisations des deux voyelles dont la qualité est similaire (voir GC1 dans la figure 4), mais qui diffèrent en termes de longueur. Cela nous encourage à postuler que la longueur peut être envisagée comme pertinente d'un point de vue phonologique en anglais basilectal mancunien.

À la lumière de nos données, la variété mancunienne apparaît donc comme une variété véritablement nordique d'anglais britannique. Les deux caractéristiques majeures communes aux variétés du nord de l'Angleterre y sont présentes et stables. Nous souhaitons donc à présent aborder le cas de la diffusion des variantes supralocales de FACE et GOAT, signe d'un nivellement dialectal à grande échelle dans le nord de l'Angleterre. La figure 5 représente les réalisations moyennes de FACE et GOAT par l'ensemble de nos locuteurs. Les réalisations moyennes de GOOSE (voyelle dont nous reparlerons ci-après), FLEECE, TRAP et THOUGHT sont également incluses à titre de comparaison. S'il apparaît clairement que les trois dernières voyelles sont des monophtongues à l'échelle de notre corpus, il paraît difficile de soutenir la même chose pour FACE et GOAT, qui sont des

diphtongues étant donné les différences de mesure entre les deux points de chaque voyelle (rappelons-le, à 1/3 et 2/3 de la durée de la voyelle).

Figure 5.



Bien sûr, des réalisations moyennes ne sauraient préjuger des comportements individuels des locuteurs, mais l'inspection plus fine de chaque enquêté confirme que les diphtongues constituent l'écrasante majorité des occurrences de FACE et GOAT. Néanmoins, une demi-douzaine de locuteurs, appartenant principalement au G1 et au G2, fait un usage plus fréquent des monophtongues que les autres pour l'ensemble lexical de GOAT, en particulier en contexte conversationnel. Force est de constater toutefois qu'aucun de ces locuteurs n'a moins de 30 ans, ce qui suggère qu'il n'y a pas de nivellement vers des variantes supralocales pour cette voyelle. Les choses sont encore plus claires pour FACE : on ne compte que deux locuteurs (IH1 et VH2) faisant un usage fréquent des monophtongues, les mêmes qui se démarquaient déjà en ce qui concerne GOAT.

Ces résultats peuvent sembler étranges dans un contexte de nivellement dialectal des accents du nord de l'Angleterre vers une variété supralocale, et c'est dans ce genre de situation que les métadonnées sociolinguistiques récoltées au cours du processus d'enregistrement se révèlent une nouvelle fois précieuses. En effet, IH1 et VH2 se démarquent du reste des locuteurs du corpus. Ils habitent actuellement à Horwich, et sont originaires respectivement de Wigan et Westhoughton, villes situées au nord-ouest du Greater Manchester, à l'extérieur de la M60, frontière symbolique utilisée par Baranowski et Turton pour marquer l'étendue de la variété mancunienne. De plus, il s'agit des seuls locuteurs qui se définissent explicitement comme des locuteurs du Lancashire.

La diffusion des variantes supralocales de FACE et GOAT n'est pas présentée dans la littérature comme un changement linguistique dont les locuteurs ne sont pas conscients (voir Watt 2002 ; Haddican *et al.* 2013), et les données de notre enquête ne contredisent

pas ce point : certains locuteurs mentionnent l'existence de ces variantes. En revanche, elles ne sont pas associées à une variété nordique supralocale, mais plutôt à Bolton et au Lancashire. Ces évaluations nous semblent expliquer pourquoi l'adoption de ces variantes ne gagne apparemment pas du terrain à Manchester. Alors que Watt rapporte que les locuteurs du Tyneside choisissent ces variantes afin de s'exprimer comme des locuteurs nordiques « modernes », cette possibilité n'est pas offerte aux locuteurs mancuniens. La variété mancunienne, qui reste donc véritablement une variété nordique d'anglais britannique, ne participe toutefois pas au nivellement dialectal dans le nord de l'Angleterre, qui constitue un mouvement de contre-nivellement (ou d'hétérogénéisation) par rapport au standard britannique.

Cependant, nos données montrent qu'il existe bien des exemples de nivellement dans la variété de Manchester. L'antériorisation de GOOSE, qui est attestée dans de nombreuses variétés de l'anglais britannique (Ferragne & Pellegrino 2010), mais aussi dans des variétés américaines (Fridland 2008) et océaniques (Przewozny 2015 ; Przewozny & Viollain 2015) est également présente à Manchester. La réalisation moyenne est très antérieure, comme indiqué sur la figure 5, et ce quel que soit le contexte phonétique dans lequel se trouve la voyelle (qu'il favorise une valeur élevée de F2 ou non). Ces variantes antérieures s'observent dans tous les groupes socio-économiques de notre corpus. En revanche, des calculs statistiques montrent que l'âge est un facteur extrêmement significatif, les locuteurs les plus jeunes faisant usage de variantes plus antérieures que leurs aînés, ce qui suggère que l'antériorisation de GOOSE est un changement vocalique en cours dans la variété mancunienne. Il ne s'agit donc pas d'une variété isolée des dynamiques des autres variétés de l'anglais, bien que le manque d'évaluations sociolinguistiques associées à ces variantes antérieures doive nous pousser à nous interroger sur le rôle des facteurs internes dans ce changement.

Aussi, en conclusion de cette dernière partie, nous espérons avoir montré que les corpus PAC Nouvelle-Zélande et PAC-LVTI Manchester sont pertinents à titre individuel pour la description phonético-phonologique de ces variétés mais également pour l'étude de phénomènes précis, comme la rhoticité ou les changements vocaliques. Mais nous espérons également avoir établi que la jonction des analyses menées sur plusieurs corpus PAC, grâce à leur comparabilité et aux métadonnées sociolinguistiques récoltées sur les locuteurs qui les composent, contribuent au débat sur des questions théoriques fondamentales comme la dérhoticisation historique progressive de l'anglais, la modélisation du phénomène de r de *sandhi*, ou encore la pertinence de la longueur vocalique d'un point de vue phonologique dans les différentes variétés de l'anglais. Sur ce dernier point, les données issues du corpus mancunien font écho à la réflexion menée sur l'existence de deux sous-catégories de voyelles (simples *vs.* complexes) en anglais néo-zélandais pour rendre compte de ce qui est appelé le *Short Front Vowel Shift* (voir Viollain 2018 ; Durand & Viollain en préparation).

Conclusions

Dans le présent article, nous avons expliqué pourquoi la phonologie a besoin de corpus et pourquoi elle entretient un lien privilégié avec eux, et notamment avec les petits corpus. Nous avons détaillé les avantages et les désavantages respectifs des grands corpus généraux et des petits corpus spécialisés en ce qui concerne notamment la question de la représentativité, pour finalement dépasser la question de la taille et repenser la

définition de ce qu'est un véritable corpus phonologique réussi en termes de finalité et d'adaptation à son objet d'étude. C'est sur la base de ces considérations épistémologiques que nous avons présenté les ambitions et la méthodologie propres au programme PAC pour constituer une grande base de données sur l'anglais oral contemporain grâce à la comparabilité de petits corpus constitués à travers le monde anglophone. Nous avons alors donné un aperçu des résultats obtenus par ce programme sur deux questions phonético-phonologiques majeures, la rhoticité et les changements vocaliques, afin de dévoiler les potentialités de ces petits corpus, des points de vue quantitatif et qualitatif. Nous avons notamment montré comment ils alimentent notre réflexion sur la tectonique des plaques linguistiques nord-sud à travers le monde anglophone, en dévoilant la complexité des mécanismes phonologiques internes, et en soulignant le poids des facteurs externes (sociolinguistiques et identitaires) dans l'évolution respective des variétés étudiées et dans la trajectoire personnelle des locuteurs qui les parlent.

BIBLIOGRAPHY

Aronoff M. (éd.) (2016). *Oxford Research Encyclopedia of Linguistics*. Oxford : Oxford University Press.

Aston G. (1997). « Small and large corpora in language learning », in B. Lewandowska-Tomaszczyk et P. J. Melia (éd.) *PALC '97 Practical Applications in Language Corpora*. Lodz : Lodz University Press, 51-62.

Auran C., Bouzon C. et Hirst D. (2004). « The Aix-MARSEC project : an evolutionary database of spoken British English and automatic tools », in B. Bel et I. Marlien (éd.) *Proceedings of the Second International Conference on Speech Prosody*. Nara : SProSIG, 561-564.

Baranowski M. et Turton D. (2015). « Manchester English », in R. Hickey (éd.) *Researching Northern English*. Amsterdam et Philadelphie : John Benjamins, 293-316.

Beal J. (2008). « English Dialects in the North of England : Phonology », in B. Kortmann et C. Upton (éd.) *Varieties of English. Volume 1 : The British Isles*. Berlin : De Gruyter, 122-144.

Beal J., Corrigan K., Mearns A. et Moisl H. (2014). « The Diachronic Electronic Corpus of Tyneside English : Annotation practices and dissemination strategies », in J. Durand, U. Gut et G. Kristoffersen (éd.) *The Oxford Handbook of Corpus Phonology*. Oxford : Oxford University Press, 517-533.

Boersma P. et Weenink D. (2017). *Praat : doing phonetics by computer*. Logiciel. Version 6.0.29. URL : http://www.praat.org/.

Buscail L. (2013). « Étude comparative des pronoms démonstratifs neutres anglais et français à l'oral : référence indexicale, structure du discours et formalisation en grammaire notionnelle dépendancielle ». Thèse de doctorat. Université Toulouse II.

Chatellier H. (2016). « Nivellement et contre-nivellement phonologique : étude de corpus dans le cadre du projet PAC-LVTI ». Thèse de doctorat. Université Toulouse II.

Cheng W. (2012). *Exploring Corpus Linguistics. Language in Action.* Londres et New York : Routledge.

Cox F., Palethorpe S., Buckley L. et Bentink S. (2014). « Hiatus resolution and linking /ɹ/ in Australian English », *Journal of the International Phonetic Association* 44 : 155-178.

Cruttenden A. (2014). *Gimson's Pronunciation of English*. 8e édition. Londres et New York : Routledge.

Delais-Roussarie É. et Post B. (2014). « Corpus Annotation. Methodology and Transcription Systems », in J. Durand, U. Gut et G. Kristoffersen (éd.) *The Oxford Handbook of Corpus Phonology*. Oxford : Oxford University Press, 46-88.

Detey S., Durand J., Laks B. et Lyche C. (éd.) (2014). *Varieties of Spoken French*. Oxford : Oxford University Press.

Durand J. (2009). « On the scope of linguistics : Data, intuitions, corpora », in Y. Kawaguchi, M. Minegishi et J. Durand (éd.) *Corpus analysis and variation in linguistics*. Amsterdam : John Benjamins, 25-52.

Durand J. (2017). « Corpus Phonology », in M. Aronoff (éd.) *The Oxford Research Encyclopedia of Linguistics*. Oxford : Oxford University Press. En ligne. URL : http://linguistics.oxfordre.com/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-145.

Durand J., Laks B. et Lyche C. (2014). « French Phonology from a Corpus Perspective : the PFC Programme », in J. Durand, U. Gut et G. Kristoffersen (éd.) *The Oxford Handbook of Corpus Phonology*. Oxford : Oxford University Press, 486-497.

Durand J. et Lyche C. (2008). « French liaison in the light of corpus data », *Journal of French Language Studies* 18 : 33-66.

Durand J., Navarro S. et Viollain C. (2015). « R-sandhi in English : how to constrain theoretical approaches », in *Global Communication Studies, vol.2 World Englishes*. Global Communication Institute : Kanda University of International Studies, 103-132.

Durand J. et Przewozny A. (2015). « La variation et le programme PAC : phonologie de l'anglais contemporain », in I. Brulard, P. Carr et J. Durand (éd.) *La prononciation de l'anglais contemporain dans le monde. Variation et structure*. Toulouse : Presses Universitaires du Midi, 55-91.

Durand J. et Pukli M. (2004). « How to construct a phonological corpus : PRAAT and the PAC project », *Tribune Internationale des Langues Vivantes* 36 : 36-46.

Durand J. et Viollain C. (en préparation). « On the New Zealand Short Front Vowel Shift », in A. Przewozny, S. Navarro et C. Viollain (éd.) *Advances in the phonology of contemporary English : Variation, change and spoken corpora*.

Ferragne E. et Pellegrino F. (2010). « Formant frequencies of vowels in 13 accents of the British Isles », *Journal of the International Phonetic Association* 40.01 : 1-34.

Fridland V. (2008). « Patters of /uw/, /ʊ/ and /ow/ fronting in Reno, Nevada », *American Speech* 83.4 : 432-454.

Gordon E., Campbell L., Lewis G., Maclagan M., Sudbury A. et Trudgill P. (2004a). *New Zealand English : Its origins and evolution*. Cambridge et New York : Cambridge University Press.

Gordon E. et Maclagan M. (2008). « Regional and social differences in New Zealand : phonology », in K. Burridge et B. Kortmann (éd.) *Varieties of English. Volume 3 : The Pacific and Australasia*. Berlin : De Gruyter, 64-76.

Gordon E., Maclagan M. et Hay J. (2004b). « The ONZE corpus », in J. Beal, K. Corrigan et H. Moisl (éd.) *Models and methods in the handling of unconventional digital corpora : vol. 2 Diachronic corpora*. Houndmills : Palgrave Macmillan, 82-104.

Gut U. et Voormann H. (2014). « Corpus Design », in J. Durand, U. Gut et G. Kristoffersen (éd.) *The Oxford Handbook of Corpus Phonology.* Oxford : Oxford University Press, 13-26.

Haddican B., Foulkes P., Hughes V. et Richards H. (2013). « Interaction of social and linguistic constraints on two vowel changes in northern England », *Language Variation and Change* 25 : 371-403.

Hay J. et Maclagan M. (2010). « Social and phonetic conditioners on the frequency and degree of 'intrusive /r/' in New Zealand English », in D. Preston et N. Niedzielski (éd.) *A reader in sociophonetics*. New York : De Gruyter Mouton, 41-69.

Hay J., Maclagan M. et Gordon E. (2008). *New Zealand English*. Édimbourg : Edinburgh University Press.

Herry-Bénit N., Kamiyama T. et Tennant J. (en préparation). « ICE-IPAC (Interphonology of Contemporary English) project : methodological issues », in A. Przewozny, S. Navarro et C. Viollain (éd.) *Advances in the phonology of contemporary English : Variation, change and spoken corpora*.

Irwin P. et Nagy N. (2007). « Bostonians /r/ speaking : A quantitative look at (R) in Boston », *University of Pennsylvania Working Papers in Linguistics* 13 (2) : 135-147.

Irwin P. et N. Nagy (2010). « Boston (r) : Neighbo(r)s nea(r) and fa(r) », *Language Variation and Change* 22 : 241-278.

Kennedy G. (1998). *An Introduction to Corpus Linguistics*. Londres et New York : Longman.

Labov W. (1963). « The social stratification of a sound change », *Word* 19 : 273-309.

Labov W. (1966). *The Social Stratification of English in New York City.* Washington D.C. : Center for Applied Linguistics.

Labov W. (1972). *Sociolinguistic Patterns*. Philadelphie : University of Pennsylvania Press.

Lobanov B. M. (1971). « Classification of Russian vowels spoken by different speakers », *Journal of the Acoustical Society of America* 49.2B : 606-608.

Mather P.-A. (2010). « Phonetic changes in New York City : 1962-2009 » Communication présentée à la *Annual Conference of the International Linguistic Association*, avril 2010, SUNY-New Paltz, New York.

McEnery T. et A. Wilson (2001). *Corpus Linguistics. An Introduction.* 2e édition. Édimbourg : Edinburgh University Press.

McKinnon M. (éd.) (1997). *New Zealand historical atlas.* Auckland : Bateman.

Navarro S. (2013). « Rhoticité et 'r' de sandhi en anglais : du Lancashire à Boston ». Thèse de doctorat. Université Toulouse II.

Navarro S. (2016). *Le /r/ en anglais. Histoire, phonologie et variation.* Dijon : Éditions Universitaires de Dijon.

Navarro S. et Viollain C. (en préparation). « R-sandhi in English and liaison in French : two phenomenologies in the light of the PAC and PFC data », in A. Przewozny, S. Navarro et C. Viollain (éd.) *Advances in the phonology of contemporary English : Variation, change and spoken corpora*.

Przewozny A. (2004). « Variation in Australian English », *Tribune Internationale des Langues Vivantes* 36 : 74-86.

Przewozny A. (2015). « L'Australie », in I. Brulard, P. Carr et J. Durand (éd.) *La prononciation de l'anglais contemporain dans le monde. Variation et structure*. Toulouse : Presses universitaires du Midi, 331-348.

Przewozny A. et Viollain C. (2015). « La Nouvelle-Zélande », in I. Brulard, P. Carr et J. Durand (éd.) *La prononciation de l'anglais contemporain dans le monde. Variation et structure*. Toulouse : Presses universitaires du Midi, 349-369.

Przewozny A. et Viollain C. (2016). « On the representation and evolution of Australian English and New Zealand English », *Anglophonia* 21. En ligne. URL : http://anglophonia.revues.org/727.

Pukli M. (2006). « Investigation sociophonétique de l'anglais en Écosse : le cas de Ayr ». Thèse de doctorat. Université Toulouse II.

Pukli M. (2015). « Scots : la variation des formes basilectales en Écosse – Annbank (Ayrshire) », in I. Brulard, P. Carr et J. Durand (éd.) *La prononciation de l'anglais contemporain dans le monde. Variation et structure*. Toulouse : Presses universitaires du Midi, 167-182.

Scheer T. (2004). « Présentation du volume. En quoi la phonologie est vraiment différente », *Corpus* 3. En ligne. URL : http://corpus.revues.org/193.

Scheer T. (2015). *Précis de structure syllabique, accompagné d'un apparat critique*. Lyon : ENS Éditions.

Sinclair J. (1991). *Corpus, Concordance, Collocation*. Oxford : Oxford University Press.

Sloetjes H. et Wittenburg P. (2008). « Annotation by category - ELAN and ISO DCR », *Proceedings of the 6th International Conference on Language Resources and Evaluation* : 816-820.

Trudgill P. (2001). « Received pronunciation : sociolinguistic aspects », *Studia Anglica Posnaniensia* 36 : 3-13.

Trudgill P. (2004). *New-dialect formation : The inevitability of colonial Englishes*. Édimbourg : Edinburgh University Press.

Vaughan E. et Clancy B. (2013). « Small corpora and pragmatics », *The Yearbook of Corpus Linguistics and Pragmatics*, 53-73.

Viollain C. (2010). « Sociophonologie de l'anglais à Boston : Une étude de la rhoticité et de la liaison ». Mémoire de master. Université de Toulouse II.

Viollain C. (2014). « Sociophonologie de l'anglais contemporain en Nouvelle-Zélande : corpus et dynamique des systèmes ». Thèse de doctorat. Université Toulouse II.

Viollain C. (2018). « Variation et changement : le *Short Front Vowel Shift* et le NEAR/SQUARE *merger* à la lumière du corpus PAC Nouvelle-Zélande », in O. Glain et M. Jobert (éd.) *Phonologies de l'anglais : théories et applications*. Limoges : Lambert Lucas.

Watt D. (1998). « Variation and Change in the Vowel System of Tyneside English ». Thèse de doctorat. University of Newcastle.

Watt D. (2002). « 'I Don't Speak with a Geordie Accent, I Speak, like, the Northern Accent': Contact-Induced Levelling in the Tyneside Vowel System », *Journal of Sociolinguistics* 6.1 : 44-64.

Wells J. C. (1982). *Accents of English*. 3 volumes. Cambridge : Cambridge University Press.

NOTES

1. Nous employons ici le terme de « finalité » dans son acception bel et bien téléologique, et non théologique, à savoir le fait pour un corpus de se voir assigner un but précis par celui qui le construit, d'être pensé comme moyen pour atteindre une fin spécifique et donc de constituer un outil adapté à ce qu'il est censé chercher, et trouver.

2. Nous utiliserons dans cet article le terme générique de « corpus oral/oraux » pour désigner les corpus dont les données primaires sont orales et non écrites. Nous opposerons dans la suite de cet article ce terme à celui de « corpus phonologique(s) » que nous définirons précisément. Aussi, nous voulons poser dès à présent que tous les corpus oraux ne sont pas nécessairement phonologiques, alors qu'à l'inverse, les corpus phonologiques sont nécessairement oraux.

3. Bien évidemment, nous ne sous-entendons pas que l'ensemble des phénomènes phonético-phonologiques caractéristiques d'une langue ou d'une variété sont nécessairement observables sur la base de n'importe quel corpus, quelle que soit sa taille ou sa raison d'être. De la même façon, nous ne sous-entendons pas que les intuitions des locuteurs et des chercheurs doivent être nécessairement disqualifiées, mais considérons au contraire qu'elles peuvent constituer, tout comme les corpus, un outil précieux d'analyse.

4. Nous faisons référence ici aux différents niveaux d'annotations qu'il est possible de créer sous Praat (Boersma & Weenink 2017) ou sous d'autres logiciels tels qu'ELAN par exemple (Sloetjes & Wittenburg 2008).

5. https://austalk.edu.au.

6. http://www.nzilbb.canterbury.ac.nz/onze.shtml.

7. http://www.ucl.ac.uk/english-usage/projects/ice.htm.

8. Voir Chatellier (2016 : 103-200) pour une description plus exhaustive du processus de sélection et une réflexion sur la définition du profil socio-économique des locuteurs du corpus PAC-LVTI Manchester.

9. Nous utilisons ici le symbole [r] pour signifier qu'un /r/ est réalisé phonétiquement, en surface, et non pour lui assigner une qualité phonétique spécifique.

10. Nous employons ce terme ici pour référer aux trois étapes décrites par Wells (1982 : 213-222) qui mènent à l'effacement du /r/ en coda, à savoir le *Pre-R Breaking*, le *Pre-Schwa Laxing* et le *R Dropping*.

11. De nombreuses analyses rendent compte du phénomène de liaison dans *here it is* en termes de re-syllabification d'un /r/ en position d'attaque de la syllabe suivante.

12. Nous employons le terme de r de *sandhi* au sein de PAC pour renvoyer non seulement à la liaison que nous avons définie précédemment, mais également à l'intrusion, à savoir l'émergence d'un [r] entre un morphème ou un mot à finale vocalique (et non pas avec un <r> orthographique) et un morphème ou un mot à initiale vocalique, comme dans *vanilla*[r] *icecream* (Navarro 2016 : 93-97).

13. Nous utilisons le terme d'« équivalent » pour souligner le fait que deux phénomènes de *sandhi* , en anglais et en français, portent le même nom. Nous renvoyons à Navarro & Viollain (en préparation) pour ce qui est d'une comparaison approfondie de ces deux phénomènes qui remet en cause justement cette apparente équivalence.

14. Cette situation n'est pas surprenante dans la mesure où la littérature postule l'existence de deux types de codas distincts : les codas dites « internes », c'est-à-dire pré-consonantiques, et les codas dites « finales », c'est-à-dire en position de coda stricte. Plusieurs travaux soulignent que ces deux types de codas ne se comportent pas toujours de la même façon dans les langues du monde. Nous renvoyons ici à Scheer (2015).

15. Pour le détail des contextes phonétiques exclus des mesures, voir Chatellier (2016 : 220).

16. Nous utilisons ici les noms des ensembles lexicaux (voir Wells 1982) comme raccourcis pour faire référence au phonème qu'ils comportent, lorsque cela est possible.
17. FOOT et STRUT étant clairement des monophtongues dans la variété mancunienne, nous n'avons fait figurer que la première mesure pour chaque voyelle.
18. Notons que ces conclusions pourraient être appuyées par une étude de la perception de ces voyelles par ces mêmes locuteurs, ce qu'en l'état notre protocole ne permet malheureusement pas. Cela pourrait faire l'objet d'une tâche supplémentaire, comme cela est envisagé au sein de la méthodologie PAC.
19. Nous adoptons ici les symboles utilisés par Cruttenden (2014) pour le *General British*.
20. Bien entendu, il serait plus précis d'analyser la durée des voyelles de TRAP et BATH dans les mêmes environnements phonétiques. Il s'agit justement d'un travail en cours dont les résultats préliminaires vont dans le sens de l'analyse acoustique présentée ici.

ABSTRACTS

The present article takes a simple observation as a starting point for its subsequent reflection on the epistemological and methodological issues surrounding the resort to small corpora: the latter have survived when the rise of corpus linguistics and the development of big data have multiplied resources. To us, this survival is proof of the material and logistical necessity as well as the scientific relevance of small corpora, which we explore from the particular angle of spoken language and within the specific domain of phonology. We explain why this field has a unique connection to its linguistic object and consequently a privileged relation to corpora and rethink the viability of corpora not in terms of size but in terms of purpose. We then defend the methodology and ambitions put forward by the PAC program whose database on contemporary oral English exclusively relies on small specialized corpora built in the various English-speaking areas around the world. Not only do we demonstrate how relevant these small corpora are individually for the quantitative and qualitative study of specific phenomena, such as levelling in the Mancunian variety of northern English for instance, but also show how they collectively allow for the comparative and cumulative analysis as well as the theoretical modelling of such phenomena as rhoticity and r-sandhi in the different varieties of English spoken worldwide. Therefore, we illustrate how the results yielded by the in-depth study of the small PAC corpora help us take into account the system dynamics of the varieties of English and the sociolinguistic weight of local identities, especially along the north-south axis, in the description and understanding of their phonetic and phonological characteristics. In addition, by supporting small corpora, we implicitly defend corpus phonology and question the automatic superiority that large corpora enjoy as far as the in-depth analysis of spoken language and of the evolution of the varieties of English spoken worldwide is concerned.

Le présent article prend pour point de départ de sa réflexion sur les enjeux épistémologiques et méthodologiques des petits corpus un constat simple, à savoir que ceux-ci ont survécu à l'heure où l'essor de la linguistique de corpus et le développement du *big data* ont multiplié les ressources. Cette survie est, selon nous, la preuve d'une nécessité à la fois matérielle et logistique et d'une pertinence scientifique des petits corpus que nous explorons sous l'angle particulier de la langue orale et du domaine spécifique de la phonologie. Nous montrons en quoi cette discipline entretient un rapport unique à l'objet langue et donc privilégié aux corpus et

repensons la validité des corpus non pas en termes de taille, mais en termes de finalité[1]. Nous défendons alors la méthodologie et les ambitions du programme PAC dont la base de données sur l'anglais oral contemporain repose exclusivement sur de petits corpus spécialisés constitués dans les différentes aires géographiques anglophones. Nous démontrons non seulement la pertinence individuelle de ces petits corpus pour l'étude quantitative et qualitative de phénomènes précis, comme le nivellement de la variété mancunienne dans le nord de l'Angleterre ; mais également l'apport concret de la comparabilité et de la jonction possible des analyses faites à partir de ces petits corpus pour la modélisation théorique des phénomènes de rhoticité et de r de *sandhi* dans les variétés d'anglais parlées à travers le monde. Ainsi, nous illustrons la manière dont les résultats obtenus grâce à l'étude en profondeur des petits corpus PAC permettent de penser la dynamique des systèmes de l'anglais ainsi que le poids sociolinguistique des identités, notamment nord-sud, dans la description et la compréhension de leurs caractéristiques phonético-phonologiques. Aussi, en défendant les petits corpus, nous défendons en creux la phonologie de corpus, et nous remettons en cause la supériorité automatique des grands corpus pour ce qui est de l'analyse en profondeur de la langue orale et de l'évolution des variétés d'anglais parlées à travers le monde.

INDEX

Mots-clés: Phonologie, corpus phonologiques, variétés de l'anglais, programme PAC, changement vocalique, rhoticité
Keywords: Phonology, phonological corpora, varieties of English, PAC program, vowel shift, rhoticity

AUTHORS

CÉCILE VIOLLAIN

Université Paris Nanterre, CREA EA370

HUGO CHATELLIER

Université Paris Nanterre, CREA EA370