Histoire Épistémologie Langage
38/2 (2016), p. 9-21

DOI : 10.1051/hel/2016380202

Disponible en ligne sur :
www.hel-journal.org

La constitution d'un corpus de geordie parlé : choix épistémologiques et réalisations empiriques. Retour sur un demi-siècle de sociophonétique anglaise

Maëlle Amand

Université Paris Diderot, Sorbonne Paris Cité / Université de Newcastle

Résumé

Cet article traite de la constitution d'un corpus parlé de l'anglais du Tyneside (plus connu sous le nom de Geordie) ou *Diachronic Electronic Corpus of Tyneside English* (DECTE, Corrigan *et al.* 2010-2012). Ce dernier consigne près de cinquante ans de recherche en linguistique de corpus. À la fin des années soixante, au moment où débute l'enquête linguistique du Tyneside (ou TLS), deux enquêtes linguistiques s'achèvent. L'une en Grande-Bretagne, l'autre aux États-Unis. Or, si Labov recommandait le recours à un nombre limité de variables lors d'études de données linguistiques, la TLS avait pour but original d'inclure le maximum de variables afin de permettre des analyses plus précises sur des données plus conséquentes grâce aux nouveaux outils informatiques à la disposition des chercheurs (Pellowe *et al*. 1972, Jones-Sargent 1983). Lors de la conception de l'enquête TLS, la possibilité de traitement par la machine avait toute son importance. Après une présentation de l'école anglaise en dialectologie, souvent promue par les universités du nord de l'Angleterre (malgré un grand nombre d'études sur les dialectes dans les années 1970), nous traitons de la particularité des approches méthodologiques du DECTE, des projets à la fois de conservation des données afin de numériser des enregistrements jusqu'alors sur bande magnétique, et de diffusion auprès du grand public.

Mots-clés

dialectologie anglaise, corpus oraux, anglais de Tyneside, Geordie, sociophonétique, dialectologie urbaine

Abstract

This paper investigates the creation of the *Diachronic Electronic Corpus of Tyneside English* (henceforth, DECTE, Corrigan *et al*. 2010-2012) a spoken corpus of Tyneside English, more commonly known as Geordie English. It comprises more than fifty years of research in corpus linguistics. At the end of the 1960s, as the Tyneside Linguistic Survey (henceforth TLS) was about to start, two linguistic surveys had just been completed, the former in Great-Britain and the other one, in the USA. But while Labov recommended the use of a smaller number of variables in the study of linguistic data, the TLS's original aim was to use as many variables as possible so as to enable more detailed analyses on bigger data thanks to the new computing tools that were increasingly available for research (Pellowe *et al*. 1972, Jones-Sargent 1983) and to make it machine-readable. After an overview of the English school of dialectology, often propelled by northern universities (despite a great number of studies on southern dialects in the 1970s), we highlight the specificities of DECTE regarding its methodological approaches, the various preservation projects to transfer the sounds from magnetic tapes to digital files along with their accessibility to the community of linguistics as well as to the public.

Keywords

English dialectology, spoken corpora, Tyneside English, Geordie English, sociophonetics, urban dialectology

INTRODUCTION

Vers la seconde moitié du XIX^e^ s., lors de la création de la Société des dialectes de l'anglais, l'étude des dialectes est encore fortement associée à celle du folklore et non à une culture d'élite. À l'instar du *connoisseur* méprisé au XVIII^e^ s. par les érudits en art ou en science (Smentek 2014), les dialectologues sont perçus par les théoriciens de la linguistique comme « de vulgaires collectionneurs de papillons [qui] s'attardent sur des futilités et s'attachent à des vétilles » (Chevillet 1991, p. 23). Il faudra près d'un siècle pour qu'il en soit autrement, et ce, grâce à deux grandes enquêtes dialectales nationales menées entre 1873 et 1896 (Ellis 1890) et la *Survey of English Dialects* ou SED, entre 1948 et 1971 (Orton et Dieth 1952). C'est bien dans la lignée de cette tradition dialectale que Firth appelle « l'école anglaise de phonétique » (1946), ajoutée à la croissance quasi exponentielle d'une linguistique de corpus à partir des années 1960 (Johansson 1991), qu'est né le corpus de Geordie parlé DECTE ou *Electronic Corpus of Tyneside English* (Corrigan *et al*. 2001) à l'aube du XXI^e^ siècle. Les deux premières sections sont un sous-corpus clos et stabilisé. Ce dernier comprend deux enquêtes mises en œuvre à une décennie d'intervalle : la *Tyneside Linguistic Survey* ou TLS en 1967-72, et la *Phonological Variation and Change in Contemporary Spoken English* ou PVC en 1991-1994. Le troisième et dernier sous-corpus, dit NECTE (*Newcastle Electronic Corpus of Tyneside English* 2007), est un corpus ouvert et fréquemment enrichi de nouveaux enregistrements. La TLS, tout comme le NECTE, rassemble une série d'entretiens de trente minutes menée par un enquêteur né à Newcastle, comprenant la lecture d'une liste de mots. Les locuteurs devaient également indiquer si oui ou non il connaissaient et/ou utilisaient une série d'expressions typiques du dialecte Geordie (TLS uniquement). La PVC comprend des entretiens entre deux personnes se connaissant ainsi que la lecture d'une liste de mots.

Quelle est donc la spécificité de ce premier corpus sur l'anglais du nord et dans quelles traditions linguistiques et méthodologiques s'inscrit-il ? Dans un premier temps, nous traitons d'une tradition dialectale anglaise promue par les universités du nord de l'Angleterre. Nous soulevons ensuite des questions d'ordre méthodologique ayant trait au corpus tel que le passage de l'hégémonie géolinguistique en dialectologie à un intérêt pour la sociolinguistique urbaine. Nous abordons enfin la pérennisation de ces archives sonores par leur transfert sur bandes magnétiques vers des supports numériques en ligne, accessibles aussi bien à la communauté de chercheurs en linguistique qu'aux locuteurs de cette variété d'anglais du nord.

1. LA TRADITION DU NORD EN DIALECTOLOGIE :

La tradition dialectale en Angleterre se développe principalement dans les centres intellectuels universitaires du nord de l'Angleterre. Bien que la Société de

Dialectologie Anglaise soit fondée à Cambridge par le célèbre philologue et étymologiste W. W. Skeat, en 1876, son siège est relocalisé à Manchester, dans le nord de l'Angleterre, jusqu'à sa dissolution en 1893. Cela représente une grande partie de son existence (17 sur 23 ans) et rend vraisemblablement compte du fait qu'un grand nombre de dialectologues britanniques firent leurs débuts en travaillant sur une variété d'anglais du nord entre 1870 et 1960 : Joseph Wright, J. D. O'Connor, Harold Orton ou encore le fameux phonéticien Daniel Jones. La grande enquête nationale sur les dialectes de l'anglais est également dirigée depuis l'université de Leeds, suite à la mort du chercheur suisse Eugen Dieth. Enfin, le projet titanesque de l'enquête du Tyneside sur le Geordie est lui aussi piloté à partir d'une autre université du nord par Barbara Strang (Newcastle). Durant la première moitié du XX^e^ siècle, les universités du nord se positionnent en promoteurs de la recherche en dialectologie, le sud étant davantage tourné vers la linguistique générale ou les langues orientales (Léon 2011).

Aujourd'hui encore, c'est bien le nord qui fonctionne comme promoteur de recherche en dialectologie anglaise puisque le succès de l'atelier de recherches sur les variétés de l'anglais du nord (*Workshop on Northern Englishes*), qui en est à sa septième édition en 2016, a fait naître un atelier équivalent sur l'anglais du sud. La présentation de ce dernier, organisé à Brighton, énonce clairement l'hégémonie du nord en dialectologie : "following the success of the Northern Englishes series, this conference lays the focus on any aspect of English in the linguistic South of the United Kingdom"[1].

2. L'enquête du Tyneside : les motivations du projet et le passage de la dialectologie rurale à la sociolinguistique urbaine

Lors d'une communication au Congrès international de dialectologie à Marburg, Strang (1968) explique la nécessité imminente de mener à bien une enquête urbaine : la stratification sociale par quartier, encore bien délimitée à Newcastle, va être bouleversée par une restructuration de l'aménagement urbain de Gateshead, au sud de la Tyne. C'est là une manifestation de la contre-urbanisation qui débute à la fin des années 1960 (Chaline 1982) : d'anciens quartiers de mineurs vont laisser place à des tours d'immeuble modernes. Il devient alors beaucoup trop complexe de retrouver les locuteurs de ce quartier, dont le parler était très représentatif du Geordie, une fois relogés aux quatre coins de la ville.

Outre les problèmes soulevés par la future déstratification sociolinguistique d'un quartier de Newcastle, les motivations à l'origine de l'enquête sont principalement d'ordre méthodologique. Dirigée par Orton et Dieth, l'enquête nationale sur les dialectes de l'anglais (SED) suscite un certain nombre de déceptions chez

1 https://sites.google.com/site/sewbrighton/ [consultation le 30 juin 2015].

les dialectologues : seuls les locuteurs masculins sont retenus car ils refléteraient un parler plus pur que celui des femmes : “[I]n this country, men speak vernacular more frequently, more consistently, and more genuinely than women” (Orton 1962, p. 15). Chaque zone urbaine est soigneusement contournée par les enquêteurs, alors que de l’autre côté de l’Atlantique, la dialectologie urbaine, inspirée de la sociologie, gagne en importance (Kretzschmar et Schneider 1996).

En Grande-Bretagne, le « virage massif » vers cette nouvelle sous-discipline est impulsé par Siverstein, Gregg et Viereck – sur le cockney (1960), le gaélique écossais d’Ulster en zone urbaine (1958) et l’anglais du Tyneside (1968) respectivement. Cependant, la taille de l’échantillon est souvent bien trop petite avec seulement quatre informatrices pour le cockney et douze hommes âgés pour l’anglais de la Tyne[2]. On constate le même problème de « non représentativité des dialectes ruraux » (Chevillet 1991, p. 49) : seul un genre y est représenté, la variation est souvent passée sous silence et la quête de « pureté dialectale » en milieu urbain conditionne très largement le mode d’enquête et les résultats. Barbara Strang, l’instigatrice du projet TLS, saura tirer parti de ces imperfections et s’inspirera de ce qu’il y a de mieux dans la SED : la rigueur de sa méthodologie et son recours systématique à l’enregistrement des locuteurs. Elle étudie avec soin l’enquête sur l’anglais écossais dirigée par McIntosh (1948-1961) car cette dernière prend en compte aussi bien les zones rurales que les zones urbaines. Il est décidé que la TLS sélectionnera des locuteurs féminins et masculins issus de la municipalité de Newcastle et dont l’âge variera entre 17 et 80 ans. Le but de la TLS est bien de combler le manque d’études en zone urbaine en Grande-Bretagne, de proposer une représentation globale d’un parler local en prenant en compte une grande diversité de locuteurs, d’accroître le nombre d’échantillonnages, à l’instar de l’enquête linguistique de Détroit (Shuy *et al*. 1968). Il n’est plus question de cartographier les variétés rurales mais bien d’étudier les stratifications sociales d’un territoire géographique très limité.

3. La mise en œuvre de la TLS : la démarche méthodologique et l’encodage des données

Malgré l’absence d’ordinateur au sein du laboratoire de linguistique en 1972[3], Pellowe, Nixon et McNeany surent anticiper la prédominance future de

2 Dans un compte rendu publié dans le *Journal of Linguistics* Strang déplore de la part de Viereck, le choix d’un titre aussi général que “A Diachronic-Structural Analysis of a Northern English Urban Dialect” (1968) car il ne concerne que le quartier de Gateshead et : “Viereck’s results are incorrectly taken as representing Tyneside generally” (Strang 1975, p. 140).

3 Cela dit, un système d’exploitation en temps partagé connu sous le nom de *Michigan Terminal System* est disponible pour la communauté des chercheurs de Newcastle et de Durham dès 1969. L’équipe envisage d’y traiter ses données futures. C’est d’ailleurs là que les données de la TLS sont traitées pour la première fois (Jones-Sargent 1983, p. 63).

l'informatique pour le traitement des données linguistiques. Ils n'étaient pas les seuls : un an plus tôt, le *Maître Phonétique*, jadis entièrement rédigé en API, devenait le *Journal of the International Phonetic Association* et dans son éditorial « Future Phoneticians », D. B. Fry prédit la future hégémonie du traitement informatique chez les phonéticiens :

> It would be very surprising if the future does not see […] the widespread use of digital computers […] For the phonetician it has the immense advantage that it can place the control of experimentation in his hands in return for a very small amount of learning. (Fry 1971, p. 9)[4]

Il donne ainsi une orientation nouvelle à ce périodique qui, jusqu'alors, reflétait une approche phonétique d'avant-guerre, davantage qualitative que quantitative.

Dès la conception du projet, Pellowe et Strang tentent de se démarquer de la démarche labovienne qui gagne alors en popularité auprès des chercheurs britanniques, même si chacune des démarches revendique une approche quantitative : « these aims [latitudinal and longitudinal studies] distinguish our investigation rather sharply from those of other Workers on urban speech variation […] (Houck 1969, Labov 1966) » (Pellowe *et al*. 1972, p. 1). Si, lors de ses premières publications, Labov s'oriente vers la sociologie quantitative et s'inspire du sociologue Barber (*Social Stratification* 1957), les dialectologues de Newcastle sont plutôt orientés vers la biologie et les mathématiques. L'ouvrage de Sokal intitulé *Numerical Taxonomy* (1966), restera pendant longtemps l'œuvre de référence pour les spécialistes de ce corpus oral. Il énumère un ensemble de techniques de classification numérique des espèces grâce à une base de données exhaustive et informatisée sur les caractéristiques de chaque individu.

Outre deux maîtres à penser distincts, la méthodologie diffère de l'approche labovienne puisqu'un très grand nombre de variables linguistiques furent dès le début prises en compte dans la TLS. Ainsi, Jones-Sargent, dont les travaux sont dans la continuité de ceux de Pellowe *et al.*, critique sévèrement certaines études de sociolinguistique urbaine, jugées trop schématiques, et propose un modèle plus complexe, fondé sur des analyses multivariées et de la taxonomie numérique :

> The initial dramatic findings and claims of sociolinguistic surveys of urban speech now need to undergo critical methodological assessment if the theoretical contribution of the subject is not to be seriously vitiated. … Multivariate techniques [however], are appropriate to apply to sociolinguistics data. (1983, p. 21-22)

4 Des problèmes similaires se posent chez les historiens français au moment où l'informatique gagne du terrain dans les différents domaines de recherche : « l'historien de demain sera programmeur ou il ne sera plus » déclara en 1967 l'historien Emmanuel Le Roy Ladurie (publié ensuite dans son ouvrage *Le territoire de l'historien* 1973, p. 14). L'informatique et les méthodes quantitatives semblent un futur outil incontournable, non seulement en phonétique mais dans l'ensemble des sciences humaines et sociales.

L'auteur rapporte que, dans le modèle statistique d'analyse des données de la TLS, aucune variable explicative n'est privilégiée par rapport à une autre afin de rendre compte d'une stratification sociolinguistique globale entre les locuteurs, les données étant censées parler d'elles-mêmes : "no variables are predicted by the model as being key, or defining, characteristics or groups. Rather, the natural groups emerge from the classification process" (1983, p. 24). Elle ajoute : " [h]ence the need for a model which exhaustively characterises social and linguistic differenciation" (p. 29). Afin de prendre un compte en nombre exhaustif de variables, l'étiquetage linguistique se veut le plus diversifié possible et comprend initialement neuf catégories, soit près de 300 variables listées ci-dessous (tableau 1) :

TABLEAU 1
Liste des variables linguistiques pour l'encodage de la TLS (Pellowe *et al.* 1972, p. 17-19)

Catégorie linguistique	**Nbr. variables**	**Catégorie linguistique**	**Nbr. variables**
Suprasegmental	58	Voyelles (accentuées, réduites, suivies d'un <r>)	113
Consonnes	45	Structure des mots et des syllabes en parole continue	33
Complexité grammaticale	36	Phénomènes de pause et d'hésitation	9
Lexique (connaissance passive et active)	2	Syntaxe (acceptabilité et usage)	14
Ressource lexicale	1	**Nombre total de variables linguistiques prévues**	**311**

L'étiquetage phonologique est un véritable défi pour l'équipe de chercheurs : il s'agissait non seulement de rendre compte de toutes les subtilités de prononciation d'un phonème en parole spontanée (Pellowe *et al*. 1972, p. 2) mais aussi de donner du sens à cette variation phonétique en proposant une norme à partir de laquelle une prononciation dialectale varie. Pour traiter des voyelles de l'anglais du Tyneside, Pellowe et McNeany proposent un codage de chaque variété de phonème selon une structure hiérarchique précise, comprenant trois strates allant d'une prononciation standard à celle d'un individu en parole spontanée, en passant par des variantes régionales – les mesures acoustiques ne sont cependant pas prises en compte dans le modèle.

La première strate ou *overall unit* est définie par Pellowe *et al.* (1972a) comme un « symbole phonologique abstrait, choisi de façon arbitraire, qui englobe la totalité des ensembles lexicaux dans lesquels il est présent » ([traduction MA],

cité dans Jones-Sargent, p. 39A). Cela dit, cette unité n'est qu'une étiquette catégorielle dont la fonction est de faciliter la comparabilité. Elle ne servira cependant pas à la classification des données phonétiques » (Pellowe *et al*. 1972, p. 21). La seconde strate correspond à la variation régionale, *putative diasystemic variation* ou PDV, et comprend un sous-ensemble, la variation individuelle ou *state* : « a PDV … is a class of phonetic states which is sociolinguistically discriminable as a class from all other such classes, if any, within a particular overall unit … [while a state is] a symbol representing a phonetic realization which is auditorily discriminable from all other states » (Pellowe *et al*. 1972, cité dans Jones-Sargent p. 39B).

Afin de faciliter le traitement de ces données par ordinateur, chaque unité est associée à une codification numérique (Pellowe *et al*. 1972). Par exemple, les cinq PDV de l'unité d'ensemble /i:/ sont respectivement codés par des nombres pairs (fi.1) : 0002 /i:/, 0004 /ɪ/, 0006 /ɛ/, 0008 /eɪ/, 0010 /ɪə/ et 0012 /ɪi/. Les *states*, relatifs à une transcription fine complétée par un grand nombre de diacritiques, sont différenciés par un chiffre supplémentaire à droite du codage PDV : la variation /i̽/ est ainsi codée : 00025. Il est à noter que la numérotation des *states* et des PDV n'est nullement une classification ordinale mais énumère simplement les possibilités de prononciation d'un individu en parole spontanée.

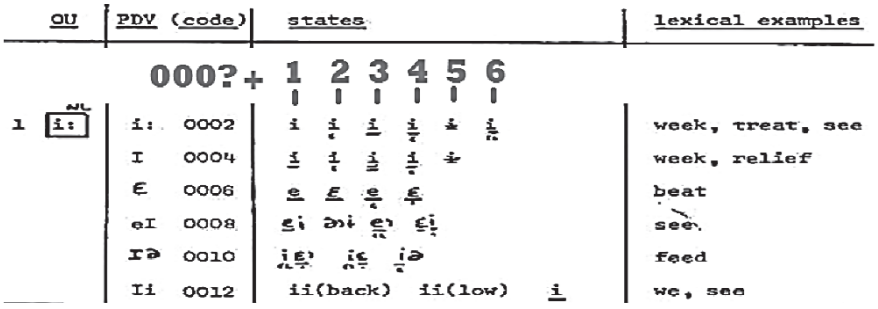

Figure 1 : Codification de Pellowe *et al* (1972) : possibilités de réalisations de l'ensemble lexical FLEECE (Wells 1982). Adapté de Jones-Sargent (1983, p. 295).

C'est à l'enquêteur lui-même, Vince McNeany, que revient la lourde tâche d'encoder chaque énoncé. Cela permet d'éviter les écueils de la *SED*. Orton collaborait avec un certain nombre de co-éditeurs et le travail d'encodage était souvent confié à des étudiants, ce qui multipliait les incohérences de transcription (Viereck 1997, p. 80). La TLS n'est pas dénuée de toute difficulté dans ce domaine et l'encodage aurait dû comprendre une meilleure structuration hiérarchique car cela aurait facilité le traitement informatique. Ainsi que le fait remarquer Jones-Sargent :

> The organisation of codes does not completely reflect the hierachical nature of the segmental variables. If it did, the accumulation of state score could be efficiently managed … [and] the use of the PL/1 [programming language] structure could have been an effective programming tool. (Jones-Sargent 1983, p. 69)

Jones-Sargent propose une amélioration de l'encodage dont les 2 premiers chiffres représenteraient l'unité d'ensemble ou OU, le suivant, le PDV et encore les deux derniers, les *states* (fig. 2). Cette hiérarchisation en structure arborescente aurait pu faciliter la recherche et la classification par le programme informatique PL/1.

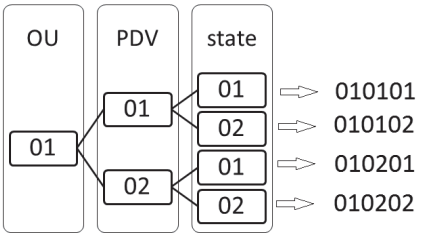

Figure 2 Proposition d'optimisation du codage TLS avec une structure arborescente.

Une telle approche aurait ainsi facilité le traitement des données phonétiques et aurait permis la mise en relation de chacune des strates, mais elle n'est proposée que bien après la conception de l'encodage par Pellowe *et al*. Aujourd'hui, un simple ordinateur personnel suffit pour traiter ce genre de données, malgré le manque de structure hiérarchique entre les PDV et les OU. Moisl *et al.* ont ainsi pu utiliser les codes tels quels, jusqu'au niveau PDV pour illustrer que la stratification linguistique est étroitement liée au statut social des locuteurs. Ils utilisèrent la classification hiérarchique en ayant recours à la méthode de Ward, une méthode de classification minimisant la variation intra-classe (Murtagh et Legendre 2014). Les auteurs concluent cependant qu'utiliser des données numériques à l'aide d'extractions acoustiques permettrait de donner davantage de poids à ces résultats préliminaires (Moisl et Maguire 2008, p. 68). En effet, l'annotation phonétique de ces anciennes bandes sonores, qui permettrait l'extraction de données numériques, est une tâche difficile au vu de la qualité sonore de certains enregistrements – les particularités de l'accent du Tyneside sont toutefois bien audibles et il serait dommage de ne pas exploiter un tel corpus.

4. Le traitement et la pérennité des données : un défi technologique ?

Un projet de grande ampleur comme la TLS n'a pu se construire en un jour, le plus gros frein étant la lenteur des traitements informatiques. De plus, la transcription et l'encodage des enregistrements représentent un coût humain considérable. Malgré

son intérêt pour la TLS, Milroy souligne les faiblesses du projet : « … the very punctiliousness of the Tyneside Linguistic Survey researchers has led to an imbalance in favour of methodology and theory and a relative weakness on results » (Milroy 1984, p. 207). Les conclusions de Milroy sont quelque peu hâtives puisque la publication des résultats du premier traitement informatique des données sociophonétiques et syntaxiques coïncide presque avec celle du linguiste (Jones-Sargent 1983 et 1984 respectivement). Jones-Sargent souligne toutefois la difficulté de faire traiter simultanément l'intégralité des données par des programmes informatiques (PL/1 et CLUSTAN) et de traiter l'encodage phonétique et l'étiquetage phonétique, en raison du manque de puissance des ordinateurs et de l'excès de variables sociolinguistiques présentes dans le modèle.

Suite aux critiques de la chercheuse, il faut attendre une décennie avant que le projet ne refasse surface : l'archivage des données sonores devenait problématique car celles-ci commençaient à se détériorer. Entre 1994 et 1995, Joan Beal reçoit un financement de la part de la fondation Catherine Cookson afin de sauvegarder les bandes magnétiques et de les transférer vers un support plus durable (aujourd'hui au format WAV), d'y associer les métadonnées, de numériser l'étiquetage linguistique ainsi que toute documentation concernant le projet (Allen *et al.* 2007). Tout doit être conforme aux recommandations de la *Text Encoding Initiative*, mise en place à peine un an avant le début de la restauration des données de la TLS. Sans ce financement, cette collecte d'anglais parlé des années 1970 se serait volatilisée en « oxyde de fer » (Widdowson 2003, p. 84). Si le nombre réel d'enregistrements effectués par l'équipe de la TLS reste encore difficile à déterminer, aujourd'hui, 37 enregistrements d'une trentaine de minutes chacun, comprenant l'étiquetage linguistique et les métadonnées, ont pu être restaurés[5]. Une seconde série d'entretiens fut traitée de la même manière (*Phonological Variation and Change in Contemporary Spoken English*) et comprend 18 enregistrements de parole spontanée d'environ une heure. On constate que l'équipe de chercheurs a su s'adapter à chaque avancée majeure de la recherche, tant au niveau de la pérennisation des archives de la parole qu'au passage à la linguistique de corpus oraux.

5. D'une enquête linguistique à un corpus diachronique (DECTE) : de la conception à la mise à disposition

Dans les années 1990, la linguistique de corpus gagne du terrain avec une augmentation de 20 publications au début de la TLS à 320 en 1991 (Johansson 1991). Il semble logique de passer de « collectes échantillonnées » à « un corpus de référence » (Cori *et al*., p.5) sur l'anglais du Tyneside. Pour transformer ces enquêtes

5 En 2014, un certain nombres de bandes magnétiques ont été retrouvées à l'université de York et numérisées en vue d'un traitement

en un véritable corpus, l'équipe a recours à l'étiqueteur grammatical CLAWS 4[6], qui permit également l'étiquetage du BNC (*British National Corpus*). Ce type d'étiquetage n'est cependant pas très adapté aux corpus oraux puisqu'il fut initialement conçu pour traiter des données écrites (Valli et Véronis 1999, p. 3), bien que cette version améliorée de l'étiqueteur CLAWS soit mieux adaptée aux données orales (Leech *et al*. 1994, p. 624). Il a fallu trouver un autre moyen d'intégrer l'encodage phonétique d'origine et c'est le format XML ou *extensible markup language* qui semblait le plus porteur et le plus durable aux yeux de l'équipe (Allen *et al*. 2007).

L'intuition des chercheurs de la TLS fut la bonne puisque ce format, lisible aisément tant par une machine que par un humain, est encore très largement utilisé aujourd'hui et sa transférabilité vers d'autres formats ou programmes phares en linguistique tel qu'ELAN ou CHAT le rend d'autant plus usité (Gries et Berez à paraître, p. 12). Christophe Parisse a récemment proposé un transfert de cet étiquetage en XML vers Praat en passant par CHAT grâce aux outils d'ORTOLANG (http://ct3.ortolang.fr/tei-corpo/). Outre l'alignement de l'étiquetage du son par intervalle de 20 secondes, le recours à Praat ouvre un large champ de possibilités d'extraction de mesures acoustiques.

Afin de mettre en ligne le corpus, l'université de Newcastle reçoit un financement du *Arts and Humanities Research Council* et met le corpus à la disposition de la communauté de chercheurs au début des années 2000 (Corrigan *et al*. 2001)[7]. Il englobe près d'un demi-siècle de données sur l'anglais du Tyneside, incluant les fichiers son, la transcription au format texte ainsi que l'étiquetage grammatical et phonétique. Le corpus est régulièrement mis à jour et de nouveaux enregistrements y sont ajoutés. La mise en ligne du DECTE a fonctionné comme un véritable émulateur de recherche puisque le nombre de publications sur l'anglais du Tyneside s'est considérablement accru : de quelques publications entre 1960 et 1990 le nombre de chercheurs et doctorants publiant sur le Geordie, s'accroît considérablement avec pour point d'orgue les années 2005 à 2007 pour lesquelles plus d'une trentaine d'articles et communications sont répertoriées. La mise en corpus de ces archives de la parole est bien l'exemple d'une réussite majeure en dialectologie.

CONCLUSION

Si Trudgill est connu pour être le pionnier britannique de la sociolinguistique grâce à son enquête sur l'anglais de Norwich en 1971 (Brown et Law 2002), l'enquête de Strang sur le Tyneside fut l'une des premières en dialectologie urbaine en Angleterre. Son cheminement jusqu'à sa mise à disposition dans le corpus DECTE

6 *Constituent Likelihood Automatic Word-tagging System*.
7 http://research.ncl.ac.uk/decte/index.htm

reflète celui de la recherche en dialectologie urbaine qui sait progressivement trouver sa place au sein des sciences de l'homme (Léonard 2012), aussi bien grâce à la popularité des ouvrage de Labov qu'aux outils informatiques toujours plus performants et à une collaboration interdisciplinaire croissante. Cette enquête porte également sur le parcours de la linguistique des corpus oraux et sur la pérennité des archives de la parole. Les données du DECTE surent également être transposées vers un support accessible à la communauté linguistique de Newcastle par le site *Talk of the Toon*[8] (Corrigan 2012) où chaque témoignage enregistré devient un ethnotexte (Joutard 1980), texte oral énoncé par et pour une communauté qui devient actrice de sa propre histoire linguistique ou dialectale. Le DECTE est aussi bien une ressource utile au chercheur qu'une aide à la construction de la conscience linguistique (Agha 2003) des locuteurs de l'anglais du Tyneside.

Bibliographie

Agha, Asif, 2003. "The Social Life of Cultural Value", *Language and Communication* 23, 231-73.

Allen, Will, Beal, Joan, Corrigan, Karen, Maguire, Warren, & Moisl, Hermann, 2007. "A Linguistic 'Time Capsule': The Newcastle Electronic Corpus of Tyneside English", Corrigan, Karen, Beal, Joan et Hermann, Moisl (éd.), *Creating and Digitizing Language Corpora*, vol. 2, Basingstoke, Palgrave Macmillan, 16-48.

Barber, Bernard. 1957. *Social Stratification: A comparative analysis of structure and process*, New York, Brace and Company.

Beal, Joan, Corrigan, Karen, & Moisl, Hermann, 2013. "The Newcastle Electronic Corpus of Tyneside English: Annotation Practices and Dissemination Strategies", Jacques Durand, Ulrike, Gut et Gjert, Kristoffernson (éd.), *Handbook of Corpus Phonology*, Oxford, Oxford University Press.

Brown, Keith, & Law, Vivien. (eds), 2002. *Linguistics in Britain: Personal Histories*, Oxford, Wiley-Blackwell.

Chaline, Claude, 1982. « Ville et urbanisme en Grande-Bretagne », *Annales de Géographie* 91, 145-153.

Chevillet, François, 1991. *Les variétés de l'anglais*, Paris, Nathan Université.

Cori, Marcel, David, Sophie, & Léon, Jacqueline, 2008. « Présentation : éléments de réflexion sur la place des corpus en linguistique », *Langage* 17, 5-11.

Corrigan, Karen, Moisl, Hermann, Beal, Joan, Buchstaller, Isabelle, & Mearns, Adam, 2001. *Diachronic Electronic corpus of Tyneside English.*

Corrigan, Karen, Buchstaller, Isabelle, Mearns, Adam, & Moisl, Hermann, 2012. *The Talk of the Toon*.

Ellis, Alexander, 1890. *English dialects, their sounds and homes*, London, Published for the English Dialect Society by K. Paul, Trench, Trübner & co.

Firth, John, 1946. "The English School of Phonetics", *Transactions of the Philological Society,* 45, 92-132.

Fry, Dennis, 1971. "Future Phoneticians", *Journal of the International Phonetic Association* 1, 2-10.

Gregg, Robert, 1958. "Notes on the Phonology of a Co. Antrim Scotch-Irish Dialect", *Orbis*, 7, 392-406.

8 *Toon* rend compte orthographiquement de la prononciation du mot *Town* en anglais de Tyneside (/u:/).

Gries, Stephan, & Berez, Andrea, (à paraître). “Linguistic annotation in/for corpus linguistics”, N. Ide et J. Pustejovsky (éd.), *Handbook of Linguistic Annotation*, Berlin, Springer.
Houck, Charles, 1969. *A Statistical and Computerized Methodology for Analyzing Dialect Materials*, University of Iowa, United States.
Johansson, Stig, 1991. *Times change, and so do corpora*. Karin, Aijmer et Bengt, Altenberg (éd.), *English Corpus Linguistics*, London, Longman, 305-314.
Jones, Daniel, 1911. “English: Tyneside dialect (Northumberland)”, *Le Maître Phonétique* 26, 184.
Jones, Val, 1984. “A presentation of some data from the Tyneside Linguistic Survey”, W. Viereck (éd.), *Focus on: England and Wales*, Philadelphia, John Benjamins, 163–177.
Jones-Sargent, Val, 1983. *Tyne bytes: a Computerised Sociolinguistic Study of Tyneside*, Frankfurt am Main, Lang.
Joutard, Philippe, 1980. « Un projet régional de recherche sur les ethnotextes », *Annales. Économies, Sociétés, Civilisations*, 35, 176-182.
Kretzschmar, William, & Schneider, Edgar, 1996. *Introduction to Quantitative Analysis of Linguistic Survey Data: An Atlas by the Numbers*, Thousand Oaks, SAGE.
Labov, William, 1966. *The Social Stratification of English in New York City*, Washington D. C., Center for Applied Linguistics.
— 1972. “Some principles of linguistic methodology”, *Language in Society* 1, 97-120.
Leech, Geoffrey, Garside, Roger, & Bryant, Michael, 1994. “CLAWS4: The tagging of the British National Corpus”, *the 15th International Conference on Computational Linguistics* 94, Kyoto.
Léon, Jacqueline, 2011. « De la linguistique descriptive à la linguistique appliquée dans la tradition Britannique : Sweet, Firth et Halliday », *Histoire Épistémologie Langage* 33-1, 69-81.
Léonard, Jean-Léo, 2012. *Éléments de dialectologie générale*, Paris, Michel Houdiard.
Le Roy Ladurie, Emmanuel, 1973. *Le territoire de l’historien*, Paris, Gallimard.
McIntosh, Angus, 1952. *An Introduction to a Survey of Scottish Dialects*, Edinburgh, Thomas Nelson.
Milroy, Lesley, 1984. “Urban dialects in the British Isles”, Peter Trudgill (eds.), *Language in the British Isle*s, Cambridge, Cambridge University Press, 199–218.
Moisl, Hermann & Maguire, Warren, 2008. “Identifying the Main Determinants of Phonetic Variation in the Newcastle Electronic Corpus of Tyneside English”, *Journal of Quantitative Linguistics*, 15, 49-69.
Murtagh, Fionn & Legendre, Pierre, 2014. “Ward’s hierarchical agglomerative clustering method: which algorithms implement Ward’s criterion?”, *Journal of Classification*, 31, 274-295.
O’Connor, Joseph, 1947. “The phonetic system of a dialect of Newcastle-upon-Tyne”, *Le Maître Phonétique*, 87, 6-8.
Orton, Harold, 1933. *The Phonology of a South Durham Dialect: Descriptive, Historical, and Comparative*, London, Trubner & Company.
— 1962. *Survey of English Dialects A: Introduction*, Leeds, E. J. Arnold & Son.
Orton, Harold & Dieth, Eugen, 1952. *A Questionnaire for a Linguistic Atlas of England*, Leeds, Leeds Philosophical and Literary Society.
Orton, Harold & Halliday, Wilfrid (eds), 1963. *Survey of English Dialects B: The Basic Material. Volume I: The Six Northern Counties and the Isle of Man*, Leeds, E.J. Arnold & Son.
Pellowe, John, Nixon, Graham, Strang, Barbara & McNeany, Vince, 1972. “A Dynamic Modelling of Linguistic Variation: the Urban (Tyneside) Linguistic Survey”, *Lingua* 30, 1-30.
Shuy, Roger, Wolfram, Walt & Riley, William, 1968. *Field Techniques in an Urban Language Study*, Michigan State University, Center for Applied Linguistics.
Siverstein, Eva, 1960. *Cockney Phonology*, Bergen, Oslo University Press.

Smentek, Kristel, 2014. *Mariette and the Science of the Connoisseur in Eighteenth-Century Europe*, Farnham, Ashgate.
Sokal, Robert, 1966. "Numerical Taxonomy", *Scientific American* 215, 106-116.
Strang, Barbara, 1967. "The Tyneside Linguistic Survey", *Verhandlungen des Zweiten Internationalen Dialektologenkongress*, Marburg/Lahn, 5-10 septembre 1965.
Trudgill, Peter, 1971. *Social differentiation of English in Norwich*, Thèse de doctorat, Université d'Edinbourg.
Valli, André & Véronis, Jean, 1999. « Étiquetage grammatical des corpus de parole : problèmes et perspectives », *Revue française de linguistique appliquée* 122, 113-133.
Viereck, Wolfgang, 1997. "The Computer Developed Linguistic Atlas of England, Volumes 1 1991 and 2 1997", *Dialectological, Computational and Interpretative Aspects. ICAME Journal* 21, 79-90.
— 1968. "A Diachronic-Structural Analysis of a Northern English Urban Dialect", *Leeds Studies in English* 2, 65-79.
Widdowson, John, 2003. "Hidden depths: exploiting archival resources of spoken English", *Lore and Language* 171, 81-92.
Wright, Joseph, 1892. *A Grammar of the Dialect of Windhill in the West Riding of Yorkshire*, London, Kegan Paul.