Histoire Épistémologie Langage
38/2 (2016), p. 23-40
© SHESL/EDP Sciences
DOI : 10.1051/hel/2016380203

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

Du dictionnaire lexico-phonétisé aux corpus oraux, quelques problèmes épistémologiques pour l'école de Guierre

Nicolas Ballier
Université Paris Diderot Sorbonne Paris Cité, CLILLAC-ARP (EA 3967)

Résumé

Cette contribution examine le changement à l'œuvre dans une partie de la recherche des phonologues formés dans le cadre de l'école de Guierre et se propose de revenir sur cinquante ans de cette tradition d'analyse du placement accentuel de l'anglais, en exposant le déplacement de certaines problématiques, d'un questionnement de l'institutionnalisation de la variation à partir de sa consignation dans les dictionnaires de prononciation à son exploration dans les corpus oraux.

Mots-clés

phonologie anglaise, troisième révolution de la grammatisation, modèles LNRE, tokenisation des corpus oraux

Abstract

This paper discusses some of the theoretical issues attending the on-going changes in English phonology as analysed in France over the last fifty years, from a corpus-driven investigation of stress patterns and vowel realisations in pronouncing dictionaries to a corpus-based investigation of variants in phonetic corpora. Section 1 presents the research agenda set out by Lionel Guierre and pursued by other French phonologists that he inspired. Section 2 discusses the kind of data analysed, from dictionaries to spoken corpora. Section 3 exemplifies three issues of this 'paradigm shift': tokenization of phonetic forms, LNRE distributions and conditional probabilities of occurrences of spoken phenomena.

Keywords

English phonology, *SPE*, Stress rules, LNRE models, phonetic corpora, tokenisation

Article published by EDP Sciences and available at http://www.hel-journal.org or http://dx.doi.org/10.1051/hel/2016380203

INTRODUCTION

Lionel Guierre (1921-2001) a été, dans les années 1970, l'un des pionniers d'une certaine forme de phonologie de corpus, en constituant une version électronique du dictionnaire de prononciation de l'anglais de Jones (dans sa douzième édition, celle de 1968). Le dictionnaire de Jones n'est pas le premier dictionnaire de prononciation de l'anglais, que l'on doit à James Buchanan et à son *Linguae Britannicae Vera Pronunciatio* (1757). On assiste au XVIII^e^ siècle à un véritable « moment phonologique », qui culmine sans doute en 1775 avec le dictionnaire de Spence, *The Grand Repository of the English Language*, premier à proposer un système de transcription véritablement cohérent (Beal 1999), et qui voit en l'espace de cinquante ans s'imposer toute une série de révolutions techniques, tels la notation de l'accent, la coupe syllabique et un système de représentation de la prononciation. Pour simplifier, le XIX^e^ siècle a surtout vu de très nombreuses rééditions des dictionnaires du XVIII^e^, dont celui de Walker, le *Critical Pronouncing Dictionary* (1791), qui s'impose comme la bible de la bourgeoisie en quête de modèle dans son ascension sociale (Trapateau 2015). À l'inverse, le dictionnaire de Jones est le premier dictionnaire de l'anglais à appliquer les symboles phonétiques de l'association de phonétique internationale, que Jones a contribué à fonder avec Paul Passy.

Lionel Guierre a systématisé et théorisé (Guierre 1979) l'interrogation des régularités accentuelles à partir des séquences graphématiques. Cette contribution se propose de revenir sur cinquante ans (Guierre 1966 pose les préalables de codage de ses données) de cette tradition d'analyse du placement accentuel de l'anglais, en exposant le déplacement de certaines problématiques, d'un questionnement de l'institutionnalisation de la variation à partir de sa consignation dans les dictionnaires de prononciation à son exploration dans les corpus oraux.

Dans un premier temps, j'exposerai quelques-uns des résultats de cette école théorique, en montrant l'importance quantitative des données lexicales étudiées. J'examinerai ensuite les problèmes théoriques rencontrés par la deuxième génération de disciples de Guierre (Martin 2011, Videau 2013), qui s'efforcent de confronter les variantes institutionnalisées répertoriées dans les dictionnaires de prononciation à des analyses sous Praat[1] de réalisations phonétiques collectées. Deux ordres de problématiques seront envisagés. J'examinerai la question de l'échantillonnage du « corpus oral » phonétique qu'il convient de constituer pour observer la variabilité et surtout des phénomènes discursifs à analyser. La prise en compte du contexte et de la sémantique met en tension certains traits retenus par

1 Praat (Boersma 2001) est un logiciel gratuit de visualisation de la parole qui permet des analyses acoustiques.

la lexicographie (par exemple, la possibilité d'un *stress shift*[2], notée pour certaines unités lexicales). Au-delà des problèmes posés par les modélisations statistiques de la fréquence (en particulier dans les modèles LNRE, cf. Baayen 2001), la question de la représentativité des *tokens* dans les corpus écrits est peut-être réglée par les giga-corpus, mais elle est balbutiante pour la représentation des variantes phonétiques (au-delà de nos rassurantes formes lemmatisées des corpus écrits) dans les corpus oraux. Je déploierai plus spécifiquement deux ordres de problématiques à l'œuvre dans cette approche des occurrences phonétiques du lexique à partir de corpus : la minoration de l'allophonie, ce que j'appellerai les problématiques de tokenisation phonétique du lexical et du sublexical, et l'échantillonnage des données lexicales dans des corpus non-contraints (comment s'assurer de la présence des formes que l'on souhaite analyser dans le corpus ?).

1. Une linguistique de corpus fondée sur des données écrites et transcrites

Cette première section va caractériser l'école de Guierre, tant dans la contextualisation de son émergence, dans la synthèse de ses principes que dans le déploiement des problématiques de ses 'disciples'. Le programme de recherche peut se simplifier en le présentant comme la recherche de règles de placement accentuel en anglais à partir des transcriptions données dans le dictionnaire de prononciation de Jones. Son entreprise de numérisation du dictionnaire de Jones lui permet de disposer d'un dictionnaire lexico-phonétisé, où chaque entrée lexicale est associée à sa prononciation, à son schéma accentuel et à la graphie inverse, innovation de Lionel Guierre qui lui permet de recenser les régularités liées à la terminaison des mots. Il peut ainsi systématiser les inventaires lexicaux qui vérifient tel ou tel placement accentuel pour une séquence finale donnée.

1.1. L'école de Guierre

Lionel Guierre commence sa thèse en 1961 et, à partir de 1964, entreprend la numérisation de la huitième édition du dictionnaire de Daniel Jones, soit 34 056 mots, qui tiendront sur 60 000 cartes perforées. Cette entreprise pionnière est largement méconnue, car elle est pour l'essentiel conduite en français. Elle est surtout contemporaine de la révolution considérable qu'a constitué *Sound Patterns of English (SPE)* en 1968, qui lui est strictement contemporaine. Dans cet ouvrage dédié à Jakobson, Chomsky & Halle prolongent les travaux de la phonologie pragoise, de Troubetskoï et de Jakobson, en reprenant la notion de la marque.

2 Échange de proéminences ou remontée de l'accent en cas de conflit accentuel : l'accent remonte en cas de succession d'accents : *thir'teen* (accent sur la seconde syllabe), mais *'thirteen men*.

Il s'agit de donner une représentation structuraliste de la langue, articulant des formes profondes et des formes de surface. Des règles de réécriture de type A → B /X__ Y, où *X* et *Y* symbolisent les contextes avant et après *A* font intervenir des matrices de traits (nasalité, sonorité, aperture de la voyelle, etc.) pour rendre compte tant des systèmes phonologiques des langues que des contextes d'application (Y) ou des résultats (B) de telle ou telle règle. Halle & Keyser publient leur *English Stress* en 1971. Goyvaets et Pullum font paraître en 1978 les *Essays on the Sound Patterns of English*. Lionel Guierre soutient sa thèse d'État en 1979, *Essai sur l'accentuation en anglais contemporain.* Il fait paraître en 1984 un premier manuel, en anglais, alors que la tradition de *SPE* est solidement établie dans le monde anglophone. Il en donne une version française en 1987. Dans les années 1990, Guierre travaille avec une version d'imprimeur du fichier ayant servi à produire le dictionnaire de prononciation coordonné par John Wells pour les éditions Longman, le *Longman Pronouncing Dictionary (LPD)*[3]. On trouve dans la thèse de Franck Zumstein une étude de la variation du placement accentuel dans les dissyllabes fondée sur ce fichier historique, enrichi par Guierre d'une annotation systématique du schéma accentuel, de la catégorie grammaticale et de la graphie inverse (Zumstein 2007, p. 110 détaille les propriétés du codage du fichier).

S'il convient d'interroger la postérité intellectuelle de la méthode d'analyse du placement accentuel et des réalisations des voyelles accentuées, elle a, institutionnellement, considérablement influencé l'épreuve de phonologie à l'agrégation d'anglais, qui a lieu depuis 2000 à l'écrit du concours et qui fait la part belle aux problématiques guierriennes (Ballier & Fournier 2007). La postérité intellectuelle tient surtout en une filiation importante chez les anglicistes français ; on peut tracer une forme de généalogie de l'école de Guierre.

Lionel Guierre a soutenu sa thèse avec Antoine Culioli, tout comme Claude Boisson, qui devait donner à sa curiosité intellectuelle d'autres objets que l'accentuation des noms composés dans la suite de sa carrière. Je rattache à la première génération de guierriens les travaux d'Alain Deschamps et de Jean-Louis Duchet en accentuation et en phonographématique, ainsi que les recherches en accentologie de Jean-Michel Fournier et en productivité des règles et des procédés morphologiques constructionnels d'Ives Trevian. Je compte parmi cette génération Michel Ginésy, formé à l'école d'Aix de Georges Faure, mais qui a consacré en 2000 un manuel empruntant aux méthodes et au formalisme de Lionel Guierre.

La question déterminante des travaux de la deuxième génération des guierriens est qu'elle oblige à poser l'analyse de la variation : le schéma accentuel

3 Dans cette édition concurrente du dictionnaire de prononciation de Jones, Wells propose une coupe syllabique et des transcriptions des variantes britanniques et américaines, s'efforçant même de rendre compte de la variation au sein de chaque variété de références par un sondage à grande échelle.

des composés (Moore 2002), des dissyllabes en britannique (Zumstein 2007) ou en australien (Martin 2011). Susan Moore (2002), Marjolaine Martin (2011) et Nicolas Videau (2013) ont ainsi cherché des corrélats du signal pour les placements accentuels. Appartenant à la même génération, Anne Talbot, Isabelle Girard, Véronique Abasq et Pierre Fournier ont plutôt cherché à explorer les propriétés posées par Guierre pour expliquer le système phonologique de l'anglais, en particulier l'isomorphisme[4], en étudiant des préfixés ou des terminaisons. L'analyse prend enfin un tour diachronique, les méthodes étant étendues aux dictionnaires de prononciation plus anciens, qu'il s'agisse de la thèse en cours de Jérémy Castanier sur les éditions du dictionnaire de Daniel Jones ou des travaux de Nicolas Trapateau (2015) sur les voyelles réduites, qui appliquent ce type d'analyse aux travaux des orthoépistes du XVIII^{e} siècle. On voit ainsi la portée des travaux de Lionel Guierre à l'aune de la recherche qu'elle a pu inspirer.

1. 2. Les principes théoriques de Guierre

Je renvoie à la contribution de Deschamps & O'Neil (2007) pour un hommage et une caractérisation plus précise du parcours intellectuel de Lionel Guierre. Je retiens les points saillants de son cadre théorique : il est pionnier d'une certaine forme de phonologie de corpus par l'importance qu'il confère aux données, sur la base de transcriptions dictionnairiques.

1.2.1. Importance de la graphie.

Elle est au cœur des critiques qu'il adresse à la phonologie générative. Guierre ne se priva pas de dénoncer dans sa thèse le caractère *ad hoc* du statut conféré à la graphie dans l'analyse phonologique de Chomsky & Halle :

> Si, selon nous, les auteurs de *SPE* échouent dans leur tentative pour construire une grammaire de la langue parlée, ce n'est pas tant que leurs règles comportent une proportion très élevée d'exceptions, ce à quoi il faut peut-être se résoudre, que parce qu'ils les justifient en se référant à l'écrit, c'est-à-dire en appelant "phonologiques" les transcriptions graphiques. Si le recours à la graphie était systématique, il s'agirait non d'une phonologie, mais d'une **phonographématique**. Puisqu'il s'agit d'une "phonologie", le recours, même non-systématique, donc arbitraire, à la graphie, est injustifié. La graphie, dans *SPE*, vole au secours de la phonologie quand et seulement quand celle-ci est défaillante. (Guierre 1979, p. 34).

De manière générale, Lionel Guierre n'est pas tendre avec *SPE* :

> La lecture de *SPE*, une fois franchi ou déplacé l'obstacle d'une « formalisation » excessivement pesante, révèle dans la théorie mise à nu, des failles, des fragilités, des habiletés qui ne concernent pas que des points mineurs ou inintéressants. (Guierre 1979, iv).

4 Tendance à une réalisation identique à celle du dérivant. Voir le développement *infra* en 1.2.4.

Au contraire, Lionel Guierre mettait en avant un programme de recherche empirique, que l'on retrouve dans cette remarque liminaire :

> Nous avions alors la conviction (nous l'avons encore) que rien n'est acquis, qu'il faut tout reprendre du début, vérifier les règles d'où qu'elles viennent, nous limiter dans un premier temps, à une description raisonnée, puisqu'elle fait défaut, présenter l'intégralité des données dites « superficielles », et n'avancer d'interprétations que fondées sur l'analyse exhaustive de chaque classe et le relevé des exceptions. (Guierre 1979, p. iv)

1.2.2. La prise en compte de la fréquence des phénomènes dans les analyses

Le contraste est effectivement saisissant entre les formulations de règles de type guierrien, qui fonctionnent par classes lexicales, et qui sont susceptibles de préciser l'effectif de la classe concernée par une règle, à partir d'un système de tri des données et d'inventaires des classes. Certaines des règles données dans *SPE*, et plus encore, des exemples pris comme paradigmes de la règle, sont d'une fréquence très faible dans les corpus, qu'il s'agisse des exemples choisis dans *SPE*, tels *Apalachicola* (« le pays au-delà » en langue creek) ou des formes illustrant des règles dans le chapitre III. La règle (106) de *Sound Patterns of English*, qui illustre l'absence de réduction vocalique, porte davantage sur les xénismes (les emprunts) que sur le fonds germanique ou latin du lexique (distinction opératoire qu'opèrent régulièrement Guierre et ses disciples).

(106) Kalamazoo, Tatamagouchi, Winnipesaukee, mulligatawny

De même, les mots servant d'illustration emblématique des règles (159) *molluscoid*, ou (162) *recondite*, et, pour le placement accentuel dépendant des terminaisons (60) *indemnification* ou (59) *theatricality* ne brillent pas par leur fréquence (ce qui, incidemment, problématise également la forme citationnelle de la règle que l'on associe à un élément du lexique). À l'inverse, parmi les résultats empiriques fondés sur des données, parfois singulièrement plus étayés que dans la phonologie générative, Guierre peut quantifier la productivité de la règle (son efficacité) : ainsi, par exemple, le placement accentuel des dissyllabiques non préfixés vaut-il pour 93% de la classe de l'inventaire estimé à 6 500 lexèmes (Guierre 1984, p. 26).

1.2.3. Une analyse potentiellement hybride entre graphie et phonie

Les règles proposées font intervenir des propriétés graphiques et phonologiques, je vois un avantage à ce régime hybride qui manipule des entités d'ordres différents : graphique (notion de terminaison), morphologique (préfixation), phonologique (nombre de syllabes), il fait apparaître des régularités inédites, telle celle associée

aux mots « italiens ». Les mots « italiens », qui se terminent par une alvéolaire et une voyelle graphique non muette autre que <y>, ce que Guierre note ainsi :

Rule STR Dent Pen : -V[+dent]V# (V ≠ y) (Guierre 1984, p. 75)

Un mot comme *sonatina*, qui est bien terminé par une séquence formée d'une alvéolaire et d'une voyelle <a> non-muette, est accentué sur l'avant-dernière syllabe. On dirait aujourd'hui qu'il confère un statut quasi-cognitif à ces xénismes dont Guierre montre qu'ils s'accentuent sur la pénultième (comme en italien). Cette notion me paraît éminemment féconde dans le traitement de la phonologie des emprunts (*loanword phonology*). Des tests psycholinguistiques de familiarité de ces unités lexicales donneraient vraisemblablement des aperçus 'cognitifs' intéressants sur cet inventaire lexical qui est allogène sans être strictement italien[5]. Point non-négligeable, la prise en compte de la graphie dans l'analyse participe aussi de tout un courant actuel de la phonologie de la langue seconde.

1.2.4. Deux grands principes structurants

Sans se livrer à une compétition stérile avec la phonologie *SPE*, notons que les données quantifiées font notamment apparaître la pertinence des distinctions catégorielles N *vs* V pour les dissyllabes en *–ate* (*ˈsenate* vs. *creˈate*), la pertinence de la terminaison plutôt que le noyau lourd pour le placement accentuel (*-ic*). Lionel Guierre voit dans le système phonologique de l'anglais deux forces qui le structurent :

- Les deux grands principes qui gouvernent l'accentuation en anglais nous semblent être :
    - la rétraction germanique,
    - l'isomorphisme. Tantôt ces deux principes concordent, tantôt ils s'opposent. (Guierre 1979, p. 769)

Il applique ce principe en 1982 dans un article aux réalisations des voyelles. Guierre explique en substance que *creative* maintient une voyelle tendue /eɪ/ pour le <a> par isomorphisme avec le dérivant *create*, mais que la réalisation du <a> en schwa dans *relative* s'explique par la rétraction germanique de l'accent, qui portait dans le dérivant sur la dernière syllabe. La dérivation en *–ive* ne se traduit pas en ce cas par un isomorphisme mais par une remontée de l'accent. Je n'ai pas abordé ici la conception exclusivement pré-tonique de l'accent secondaire (et les quatre règles pour rendre compte de son placement) ou les questions internes à la théorisation du placement accentuel, tel que le statut du C2, agrégat consonantique, qui, chez Guierre reçoit une définition particulière (notamment telle que Cr n'est pas un C2,

5 Pour un traitement particulièrement détaillé de cette notion des « mots de type italien », voir (Castanier 2016, 326-338) , qui aborde, selon les ensembles lexicaux considérés, la variation du placement accentuel dans des dictionnaires de prononciation, de 1727 à 2008.

mais rC en est un), ou sa conception de la syllabe. Il ne s'agit pas ici de proposer une critique interne mais bien plutôt une problématisation du rapport à son objet (les données). Cette école subordonne la formulation des règles aux observables, ce qui conduit à examiner plus avant le statut de ces observables.

2. LE STATUT DES DONNÉES ANALYSÉES

On peut chercher à caractériser les mutations de l'*observatoire* (Milner 1989). Après avoir dû batailler plusieurs siècles contre un certain arbitraire notationnel (et une quasi-absence de correspondance univoque graphie/phonie), les linguis-tes ont dû développer des préventions contre des décisions arbitraires (conserva-tisme, prescriptivisme) des lexicographes dans leurs transcriptions phonétiques. Cela pose plus généralement la question de la représentativité du dictionnaire pour analyser la langue.

2.1. Conséquences

Dans le domaine traditionnellement dévolu à l'anomie[6], le lexique, l'école de Guierre tranche vigoureusement avec la tradition anomaliste (de Jones précisant dans la première préface de son dictionnaire l'impossibilité des règles à Bolinger qui explique que l'accent n'est prévisible que si l'on est devin). Disposer d'un dictionnaire numérisé permet de contrôler l'application de la règle (et de contester, par exemple, la thèse du noyau lourd, d'un placement accentuel fondé exclusivement sur les agrégats consonantiques) mais également de tester l'efficacité des règles proposées. C'est la notion de productivité de la règle, inégalement envisagée au sein de cette école. L'importance de la règle est évaluée à l'aune du nombre d'items lexicaux pour laquelle elle vaut. On peut estimer qu'une règle est productive si elle vaut pour un maximum des mots de la classe qui en relèvent (par exemple, les mots qui se terminent en *–ic* sont accentués sur l'avant-dernière, à moins de dix exceptions près). En termes stricts, elle désigne un quotient de l'ensemble des éléments lexicaux qui relèvent de la règle sur l'ensemble des éléments lexicaux dont les propriétés graphématiques, morphologiques, syllabiques et catégorielles en relèvent. Pour bien faire, il conviendrait de factoriser la fréquence des différents éléments lexicaux, pour rendre compte de l'efficience de la règle dans l'emploi de la langue, et pas seulement au sein d'un inventaire lexical considéré (voir, pour une analyse allant dans ce sens, les illustrations tirées du N-Gram Viewer de Google dans Castanier 2016). Disposer d'un inventaire des fréquences (fût-il réduit à un dictionnaire) permet de penser la règle phonologique en terme de productivité, en quelque sorte d'affecter la règle d'un coefficient implicite de

6 Croyance en une absence complète de règles.

vraisemblance. Elle permet ainsi de se doter d'une procédure de validation, une sorte de pourcentage de fiabilité de la règle.

Conséquence de ce type d'observation, la règle n'est pas nécessairement conçue comme absolue et imperméable à la moindre exception, mais comme des règles ayant une plus ou moins grande productivité. Je ne détaille pas ici les conséquences de cette grammaire tendancielle. Je fais valoir les points suivants : elle se prête bien à l'analyse en synchronie du changement diachronique et permet de justifier des locus de variabilité observée pour le placement accentuel ou la valeur vocalique. Il me semble même que l'on peut trouver dans cette analyse en termes de tendances, la possibilité de débusquer des tropismes accentuels, des attracteurs de placement accentuel, mais ceci est un autre débat.

2.2. Hybridité entre phonétique et phonologique

Un dictionnaire dit de « prononciation » est au cœur de cette problématique. Instrument de référence conçu également pour les non-natifs, il consigne un certain nombre de variantes qui relèvent clairement du phonétique, mais qui sont indispensables à l'intelligibilité, et qui sont du même coup passibles d'une acception / appréhension phonologique. J'illustre cette ambivalence et ses propriétés à partir de transcriptions du mot-vedette *government*.

government ˈgʌv. ən.mənt, - əm.mənt, -və.mənt, US -ɚn-

Sans les théoriser, le dictionnaire indique les variantes imputables à la resyllabification et à l'assimilation. Dans le même ordre d'idée, le dictionnaire compile toute une série d'entrées encyclopédiques de phénomènes phonologiques, la notation des degrés d'accents (primaires ou secondaires) est de type phonologique, même si l'on pourrait argumenter que la hiérarchie accentuelle n'est pas explicitée. Par exemple, au vu de la différence de prononciation du nom et du verbe *delegate*, on ne sait s'il convient de distinguer des degrés de hiérarchie différents pour la dernière syllabe entre syllabe inaccentuée réduite pour le nom *delegate* $_{N}$ /ˈdeləgət/ et inaccentuée pleine pour le verbe *delegate* $_{V}$ /ˈdeləgeɪt/.

La première difficulté distingue les corpus oraux et corpus écrits en ce que la lemmatisation n'est établie de manière certaine que pour l'écrit. J'appelle cette problématique, sans me dissimuler la part d'inexactitude d'une telle formulation, la « tokenisation phonétique du sublexical ». Il s'agit de montrer que l'ensemble des variantes phonétiques des composantes inférieures au lexème (phones, mais aussi syllabes phonétiques et réalisations phonétiques des préfixes) est insuffisamment établie, et qu'elle n'est, en tout état de cause, pas établie pour les corpus oraux, et encore moins automatisée. Pour l'anglais, la variabilité sémantique (et les variations potentielles de choix de focus et d'accent tonal) établit une distinction entre

les affixes : la variabilité phonétique des suffixes est moindre, le système accentuel de l'anglais étant en partie contraint par ces mêmes suffixes. La variation des réalisations est considérable, les variantes accentuelles (Zumstein 2007, Trevian 2007) contribuent fortement au déploiement de ces variantes. Or, non seulement ces variantes sont sous-estimées dans les dictionnaires de référence (Martin 2009, Bauer, communication personnelle), mais leur recensement même est sans doute incomplet : on ne dispose pas d'un répertoire fini de ces formes attestées et surtout pas (ou peu) de leur distribution dans les corpus (et encore moins de probabilités conditionnelles de leur co-occurrence).

2.3. Problèmes de tokenisation phonétique

Le dictionnaire *LPD* donne pour les variantes de *and* des renvois vers les entrées encyclopédiques de la compression, de l'assimilation et de la syllabicité. Rien n'est dit de la réalisation, du spectre allophonique que l'on souhaite étudier. Je donne un aperçu de ce que pourrait être cette problématique à partir du corpus AIX-MARSEC, en examinant successivement les allophones d'un marqueur grammatical (*and*), d'un préfixe (*re-*) et, à défaut de l'ensemble des phones d'un corpus, la distribution des « phonèmes » qui servent à annoter une transcription automatique par alignement. Les données sont tirées de la base extraite du corpus AIX-MARSEC (Auran *et al.* 2004) comprenant 54 432 tokens, 204 696 phones, mais qui ne répertorie pas particulièrement les syllabes ou les variantes.

3. DE QUELQUES DIFFICULTÉS DU TRAITEMENT DES DONNÉES ORALES EN CORPUS

L'analyse d'une forme phonétique suppose que soit connu l'ensemble des réalisations associées à une forme, qu'il s'agisse d'un mot ou d'un préfixe. Il n'y a pas de lemmatiseur pour les corpus oraux qui unifierait l'intégralité des transcriptions (*c'lui, çui, celui*) et la tokenisation phonétique n'est pas encore disponible, les aligneurs forcés ne proposant qu'une transcription de type phonologique sur la base de dictionnaires déjà disponibles (Ballier & Martin 2015), et généralement relevant, pour l'anglais, de la variété américaine.

3.1. Les limites de la tokenisation phonétique des corpus oraux : le cas de and

J'illustre cette problématique de la tokenisation des réalisations des variantes avec l'approximation des réalisations d'un marqueur, en quelque sorte le spectre allophonique de l'ensemble des réalisations du marqueur *and*. Je vais décrire sommairement, c'est-à-dire sans analyser le spectre des voyelles, ni contrôler l'intégralité des formes disponibles, la variation que l'on peut inférer à partir des données extraites d'un corpus oral.

And est dans bien des corpus oraux le type le plus fréquent, on ne dénombre pas moins de 1 505 occurrences de *and* dans le corpus AIX MARSEC de 54 432 tokens. L'analyse de la base de données livre les chiffres suivants : il y a 31 liaisons répertoriées où la consonne finale de *and* est liée au mot qui suit, 1 424 réalisations de voyelles réduites qui sont catégorisées en /ə/ et 79 réalisations avec la voyelle pleine /æ/ dans la transcription du corpus. Le spectre de la variabilité des formes phonétiques est sans doute sous-estimé (Ballier 2013), puisque compressions et syllabicité ne sont pas pris en compte dans la transcription. Pour le moment, les corpus oraux, majoritairement transcrits en alignements forcés, ne proposent que des annotations en phonèmes des données, de sorte que la transcription est potentiellement sujette à caution : elle n'a pas fait l'objet d'une vérification manuelle à partir du signal, et une division syllabique a été reportée à partir du principe de l'attaque maximale. Sans rentrer dans les détails d'une analyse précise que permettrait une extraction des valeurs formantiques, on peut raisonner sur la base des données disponibles dans la base extraite du corpus, le nombre de phones par mot et la durée.

Les diagrammes de dispersion de la durée font apparaître une forte variabilité. Point non-négligeable, la variation en durée est aussi imputable à la tenue de la nasale en cas d'apocope de l'alvéolaire. On peut juger de la dispersion de la durée des *and* en voyelle pleine à partir des boîtes de dispersion de la figure 1.

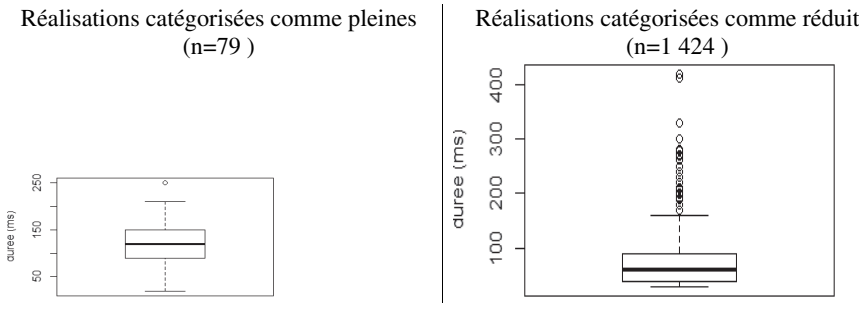

Figure 1: Dispersion de la durée (ms) des *and* en voyelle pleine et en voyelle réduite dans le AIX-MARSEC

Le nombre de valeur aberrantes (petits cercles) est bien plus important pour les réalisations de la voyelle dite réduite, dont la durée moyenne (matérialisée dans le « diagramme à moustaches », par la barre noire en gras) est certes inférieure à la moyenne de la durée des réalisations pleines, mais qui comporte des valeurs de durée bien supérieures aux voyelles pleines. Il est donc vraisemblable que la base de données, fondée sur l'alignement automatique, devrait affiner les réalisations de *and*. Tout un spectre de réalisations se devine derrière cette catégorisation initiale, ce qui donnerait raison à Michel Viel quand il proteste contre l'absence de réalisme des conventions transcriptionnelles retenues dans les dictionnaires de

prononciation et affirme la nécessité d'une palette de symboles API plus étendue qu'il faudrait mobiliser pour rendre compte des réalisations effectivement observées (Viel 2003). Une analyse fondée sur l'extraction automatique des formants confirmerait ce fait.

Les étiquettes des corpus phonétiques, qui ne sont que des étiquettes phonologiques issues de dictionnaires, peinent à rendre compte de la complexité des réalisations observées. Cette problématique des conditions de possibilité de la discrétisation du signal dans son versant segmental questionne la diversité allophonique et sa référenciation parcimonieuse dans les dictionnaires de prononciation. Il est entendu que la question est encore plus complexe encore, lorsqu'il s'agit du suprasegmental (Ballier et Martin 2015). On peut en effet essayer de donner une esquisse de caractérisation des réalisations d'un marqueur, où les propriétés phonétiques de la qualité de la voyelle ainsi que les contours intonatifs sont rattachés à des interprétations sémantiques (voir, pour une esquisse de l'analyse de *so*, Ballier & Filippi 2007). Reste que, comme on pouvait s'y attendre, la granularité des transcriptions tirées des alignements automatiques est insuffisamment fine, on voit bien ainsi les problématiques de tokenisation phonétique à l'œuvre dans les corpus oraux. Elle vaut au niveau lexical, illustré ici par l'ensemble des formes de *and*, mais elle vaut sans doute plus encore pour le sub-lexical, que j'illustre par le cas des préfixes. Les préfixés en *re-* en anglais permettront d'illustrer les problèmes d'itemisation des différentes formes à prendre en compte dans l'analyse et, *a contrario*, de mesurer le chemin parcouru pour l'étude des corpus écrits. Je m'appuierai sur une étude des préfixés de l'italien qui s'est posée la question de l'accès aux formes préfixées en *ri-* dans les corpus.

3.2. Le statut des réalisations des mots en re- en anglais, des dictionnaires aux corpus

La figure 2 reproduit le recensement des variantes du préfixe *re-* (Videau 2013, p. 81) dans le cadre d'une thèse qui inventorie les variantes attestées de certains préfixes dans les dictionnaires avant de chercher à les analyser en corpus.

On peut évidemment se demander si cette matrice, qui répertorie dans les colonnes l'ensemble des variantes attestées pour une prononciation principale, ne révèle pas à elle seule une sous-estimation de la variation. Elle s'apparente à une matrice de confusion, qui compare les formes prédites en colonnes et les formes réelles en ligne, dans une pratique classique de la classification automatique. C'est ici le nombre de variantes théoriquement possibles mais pas attestées dans le dictionnaire qui est frappant. Dans sa thèse, N. Videau ne dénombre pas moins de 32 entrées sans variante pour 51 mots en *re-*. Comme il le souligne, les chiffres de la colonne Total ne sont pas égaux à la somme des chiffres des autres colonnes. Ils

correspondent aux nombres de mots ayant au moins une variante, quelle qu'elle soit, portant sur le schéma accentuel ou la qualité vocalique. Les dictionnaires n'explicitent pas toujours comment est conféré le statut de la « prononciation principale » dans le cas de plusieurs variantes recensées. Wells en est bien conscient, pour avoir lancé une des premières enquêtes mettent à contribution les internautes pour un questionnaire en ligne fondé sur des données écrites (Wells 1999). Figure depuis l'édition de 2000 dans le dictionnaire une indication de la prononciation préférée selon les variables 'âge du locuteur' et 'variété parlée'.

		VARIANTES										
		Total	[rɪ]	['ri:]	[ˌri:]	[(ˌ)ri:]	[ri:]	['re]	[ˌre]	[re]	[ri]	[rə]
PRON. PRINC.	[rɪ]	**418**		14	23	8	385	8	5	1	0	417
	['ri:]	**19**	6		11	3	5	0	0	0	0	6
	[ˌri:]	**32**	14	1			13	0	12	0	0	10
	[(ˌ)ri:]	**3**	2	1				0	0	0	0	1
	[ri:]	**2**	0	0	2			0	0	0	0	0
	['re]	**9**	9	0	0	0	6		0	0	0	9
	[ˌre]	**4**	2	0	3	0	1	0		0	0	1
	[ri]	**11**	0	0	11	0	0	0	0	0		0

Figure 2. Distribution des variantes réalisationnelles recensées dans le *LPD* des mots en *re-* (Videau 2013, p.81).

La thèse donne des exemples d'analyse phonétique en contexte des variations de la réalisation, mais n'explicite pas le point suivant : qu'en est-il de la possibilité d'inventorier en corpus ces réalisations ?

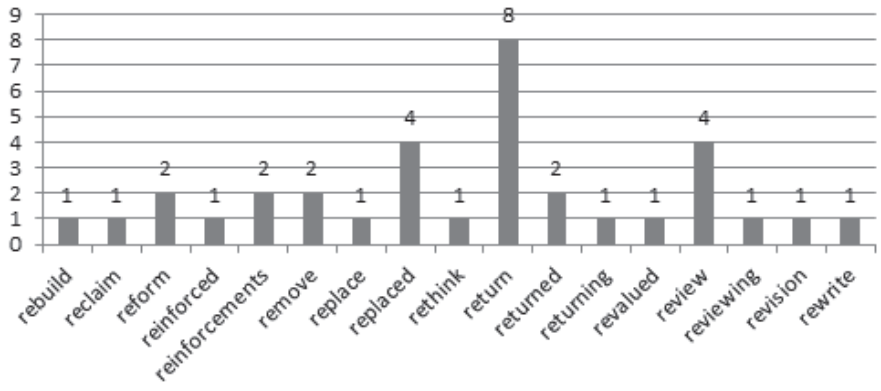

Figure 3. distribution des mots en *re-* dans le corpus AIX-MARSEC

La figure 3 regroupe l'ensemble des occurrences de *re-* dans le AIX-MARSEC, dans une acception maximaliste qui inclut aussi bien les préfixés (*rebuild*) et les

pseudo-préfixés (*revision*). Le premier constat est que la moisson est bien décevante si l'on cherche à analyser la variation : on ne dispose quasiment que d'une seule occurrence à analyser par unité lexicale, ce qui revient à dire que l'objet d'étude a le même statut que la moitié des mots du corpus retenu. Pour apprécier la complexité de la représentativité des occurrences dans un corpus oral, il convient de bien garder à l'esprit que ces occurrences de la forme lemmatisée doivent également prendre en compte les variantes phonétiques et les différents paramètres susceptibles d'en affecter la réalisation. S'agissant de la réalisation d'un préfixe, qu'en est-il du caractère plus ou moins sémantiquement contraint du contexte, commandant à son tour le choix de la syllabe tonique ? Plus généralement, observe-t-on un échange de proéminence pour cause de conflit accentuel (*stress shift*), processus consigné pour certaines unités lexicales dans le dictionnaire ? La recherche n'en est, me semble-t-il, pas encore là dans une approche multiparamétrique des corpus oraux. À l'inverse, les morphologues, sur la base de corpus écrits annotés avec des lemmatiseurs opérationnels, ont un début de réponse pour l'analyse des conditions d'occurrence d'une forme donnée en fonction de sa fréquence et de sa productivité.

Analysant la croissance des préfixés dans un corpus journalistique d'italien, Baroni & Evert 2006 ont montré les propriétés des courbes de croissance du vocabulaire (*vgc, vocabulary growth curve*). On dispose d'une modélisation mathématique qui met en regard l'accroissement du corpus, le nombre d'hapax et l'accroissement des occurrences d'un type donné. Cette courbe de croissance d'un vocabulaire donné (les mots en *ri*-) renvoie le nombre de types (V) et d'hapax (V1) par coupe longitudinale de 1 000 occurrences (N). Il est possible, à partir de la visualisation et de la modélisation de ces courbes de croissance, fondées sur les régularités de la Loi de Zipf, d'extrapoler les occurrences en fonction de la taille du corpus. Cette modélisation des phénomènes repose sur les travaux fondateurs de Baayen (2001) portant sur la productivité lexicale dans une perspective quantitative. En particulier, Evert & Baroni proposent une modélisation des LNRE (*Large Number of Rare Events*) permettant des projections des occurrences des préfixés en *ri*- selon la taille du corpus. La figure 4, à partir de l'une des quatre fonctions définies (fonction Zipf-Mandelbrot finie), représente l'espérance des occurrences des mots en *ri*- en ordonnée, en fonction de l'accroissement de la taille du corpus (en nombre de mots).

Comme le rappellent fort opportunément M. Baroni et S. Evert, « With a bit of creativity in the definition of the target type classes, vocabulary statistics modeling techniques can be applied to a very wide range of linguistic problems » (Baroni et Evert 2006, p. 16). Pour autant, un certain nombre de précautions entourent ce type d'analyse. Ils estiment notamment que la validité des extrapolations est

contestable au-delà du double du corpus initial (la ligne verticale en pointillés sépare le corpus initial de son extrapolation). Ils observent également la tendance de ces modèles à sous-estimer le nombre d'occurrences des types (Baroni et Evert 2006, p. 21).

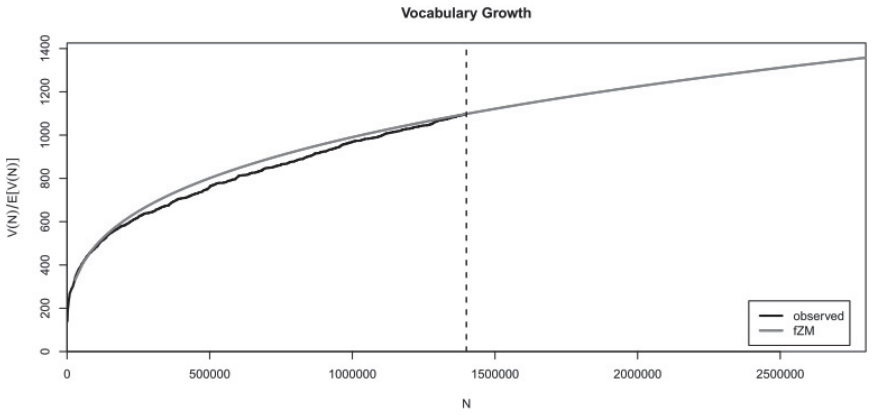

Figure 4. modélisation par extrapolation de l'accroissement des préfixés en *ri-* (Evert and Baroni 2006)

Retenons que les occurrences d'un type donné dans un corpus ne sont pas dans un rapport simple de proportionnalité à la taille du corpus. Nombre d'occurrences et taille du corpus ne s'appréhendent pas par une simple règle de trois, mais dépendent pour les lexèmes d'une distribution LNRE. L'analyse d'un certain nombre de corpus oraux (et de taille singulièrement plus importante que ceux dont nous disposons actuellement) permettra d'établir si les réalisations phonétiques suivent les mêmes lois à partir des spectres de fréquence (Evert et Baroni 2007). Il conviendrait de rechercher non pas tant des occurrences, mais les conditions de possibilité des occurrences. La question qui se pose est de savoir si les réalisations phonétiques sont passibles d'une analyse similaire. A supposer que l'on puisse identifier aussi précisément toutes les variantes, quelles conditions s'imposent sur l'accroissement de la taille du corpus pour les voir apparaître ?

Conclusion

Cet exposé d'une partie des problématiques de Guierre et de ses disciples a défendu une phonologie de corpus qui ne soit pas exclusivement fondée sur les données du signal (c'est un peu l'acception dominante qu'en donne l'*Oxford Handbook of Corpus Phonology*, qui ne mentionne d'ailleurs pas le nom de Guierre). Quand bien même il n'y aurait pas de salut hors du traitement du signal, ce type d'analyse en corpus n'échappe pas à une relative contingence des formes observées. Donner

en pâture à la phonologie l'immensité du divers phonétique n'est pas une opération anodine. On peut tenter de proposer une synthèse de l'entreprise intellectuelle qui prolonge une perspective lexicale en étendant à la langue orale les données du dictionnaire : on est passé du recensement de la langue supposée à la tentative d'analyse empirique de la parole.

La taille actuelle des corpus oraux ne permet pas d'envisager la prévisibilité des occurrences phonétiques à partir de la distribution des types. La base de données mentionnée du corpus AIX-MARSEC compte 204 696 phones, la fréquence des types y varie spectaculairement, ainsi le rapport de /ʒ/ à schwa est-il de un à mille. S'y donne à voir une sorte de loi de Zipf pour les phones, où certes la fréquence n'est pas divisée par deux lorsque le rang décroît, mais où les données disponibles sur les types dissimulent, on l'a vu avec la voyelle de *and*, la variation des occurrences.

Pour bien faire, il faudrait pouvoir sortir du simple relevé des occurrences, car le prélèvement du corpus à la langue ne permet pas un simple rapport de proportionnalité. Les données linguistiques ne sont généralement pas, au sens statistique, normales. Le rapport entre l'échantillon et la population est complexe. Il ne se laisse pas approximer par inférence simple d'une courbe gaussienne. Il relève sans doute également de cette distribution des événements rares extrêmement nombreux (LNRE). Dans le même ordre d'idées, bien des phénomènes phonologiques, notamment la liaison, gagneraient à être appréhendés à partir des conditions de possibilité de leur occurrence (c'est-à-dire, en l'espèce, des catégories syntaxiques) et donc des probabilités conditionnelles. Je ne développe pas ici ce que serait une telle analyse, je renvoie à la remarquable présentation de cette notion qu'en donne (Goldsmith 2007).

Une partie des anglicistes formés dans la tradition de Lionel Guierre a dû batailler avec les limites de la consignation de la variation dans les dictionnaires de la variation, et notamment les effets du prescriptivisme des lexicographes ; la génération suivante est confrontée à des problèmes qui ne sont pas moindres pour l'appréhension de cette variation dans les corpus oraux. J'ai donné deux illustrations du type de difficultés posées par l'examen de ces données : la concurrence des occurrences et la sous-détermination de leur variation, lorsqu'il s'agit d'analyser le mot le plus souvent en tête des fréquences d'occurrence des corpus oraux (*and*), la complexité de l'allophonie sublexicale et au contraire la difficulté à disposer d'attestations de certaines réalisations. Le moins qu'on puisse dire est que cette entreprise nécessite une linguistique singulièrement outillée, qui et encore à bien des égards balbutiante.

Bibliographie

Auran, Cyril, Bouzon, Caroline & Hirst, Daniel, 2004. "The Aix-Marsec Project: an evolutive database of spoken British English", *Proceedings of Speech Prosody 2004*, 561-564.

Baayen, Harald, 2001. *Word Frequency Distributions*, Dordrecht, Kluwer.

Ballier, Nicolas, 2013. « Approche de la cohésion en anglais : le cas de AND en texte lu », Deschamps, Alain, Trevian, Ives et Zumstein, Franck (éd), *11e colloque d'avril sur l'anglais oral*, Université Paris 13, diffusion APLV, 9-22.

Ballier, Nicolas & Fournier, Jean-Michel, 2007. « La compétence phonologique en anglais au-delà des questions à l'agrégation », *Les Langues Modernes*, 101(3), 18-28.

Ballier, Nicolas & Martin, Philippe. 2015. "Speech annotation of learner corpora", Granger, S., Gilquin, G., Meunier, F. (ed.), *The Cambridge Handbook of Learner Corpus Research*, Cambridge, Cambridge University Press, 107-134.

Ballier, Nicolas & Filippi-Deswelle, Catherine, 2007. « Le *though* dit "adverbial" », Celle, Agnès, Gresset, Stéphane et Huart, Ruth (éd.), *Les Connecteurs, jalons du discours*, Berne, Peter Lang, 173-196.

Baroni, Marco & Evert, Stefan, 2006. "The zipfR package for lexical statistics: A tutorial introduction". Disponible sur: http://zipfr.r-forge.r-project.org.

Beal, Joan C., 1999. *English Pronunciation in the Eighteenth Century: Thomas Spence's 'Grand Repository of the English Language',* Oxford: Clarendon Press.

Boersma, Paul, 2001. "Praat, a system for doing phonetics by computer". *Glot International* 5:9/10, 341-345

Bolinger, Dwight, 1972. "Accent is predictable (if you're a mind reader)", *Language* 48, 333-344.

Castanier, Jérémy, 2016. *L'évolution accentuelle du lexique contemporain appréhendée à travers les dictionnaires de prononciation (XVIIe-XXIe siècles)*, thèse non publiée, sous la direction de Jean-Louis Duchet et de Sylvie Hanote, Université de Poitiers.

Chomsky, Noam & Halle, Morris, 1968. *The Sound Pattern of English*, New York, Harper & Row.

Deschamps, Alain & O'Neil, Michael, 2007. "Lionel Guierre", *Language Sciences* 29/ 2-3, 492-495.

Durand, Jacques, Gut, Ulrike & Kristoffersen, Gjert (éd.), 2014. *The Oxford Handbook of Corpus Phonology,* Oxford, Oxford University Press.

Evert, Stefan & Baroni, Marco, 2007. "ZipfR: Word frequency distributions in R", *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Session*, Prague.

Fournier, Jean-Michel, 2007. "From a Latin syllable-driven stress system to a Romance versus Germanic morphology-driven dynamics: in honour of Lionel Guierre", *Language Sciences* 29 (2), 218-236.

Ginésy, Michel, 2000. *Phonétique et phonologie de l'anglais*, Paris, Ellipses.

Goldsmith, John, 2007. "Probability for linguists", *Mathématiques et sciences humaines* 180, 73-98.

Goyvaerts, Didier, & Pullum, Geoff (ed.), 1975. *Essays on the Sound Patterns of English*, Amsterdam, John Benjamins.

Guierre, Lionel, 1966. « Un codage des mots anglais en vue de l'analyse automatique de leur structure phonétique », *Études de Linguistique Appliquée*, Paris, Didier, 48-64.

— 1979. *Essai sur l'accentuation en anglais contemporain*, thèse d'État, Université de Paris VII.

— 1982. « L'isomorphisme vocalique en phonologie de l'anglais », *Recherches en linguistique étrangère,* vol. VII, Paris, Les Belles Lettres, 84-98.

— 1984. *Drills in English Stress-Patterns*, Paris, Armand Colin/Longman.

— 1987. *Règles et Exercices de Prononciation anglaise*, Paris, Armand Colin-Longman.

Halle, Morris & Keyser, Samuel Jay, 1971. *English Stress*, New York, Harper & Row.
Martin, Marjolaine, 2011. *De l'accentuation lexicale en anglais australien standard*, thèse non publiée, sous la direction de Jean-Michel Fournier, Université de Tours.
Milner, Jean-Claude, 1989. *Introduction à une science du langage*, Paris, Seuil (Des Travaux).
Moore Mauroux, Susan, 2002. *Les mots composés : analyse de schémas accentuels de l'anglais britannique standard,* thèse non publiée, sous la direction de Jean-Louis Duchet, Université de Poitiers.
Spence, Thomas, 1775. *The Grand Repository of the English Language*, Newcastle, Thomas Saint.
Trapateau Nicolas, 2015. *Placement de l'accent et voyelles inaccentuées dans la prononciation de l'anglais du XVIII*e *siècle sur la base témoignage des dictionnaires de prononciation, des vers et de la musique vocale*, thèse de doctorat non publiée, sous la direction de Jean-Louis Duchet et de Philippe Caron, Université de Poitiers.
Trevian, Ives, 2007. "Stress-neutral endings in contemporary British English: an updated overview", *Language Sciences* 29 (2), 426-450.
Videau, Nicolas, 2013. *Préfixation et phonologie de l'anglais : analyse lexicographique, phonétique et acoustique*, thèse non publiée, sous la direction de Jean-Louis Duchet et de Sylvie Hanote, Université de Poitiers.
Viel, Michel, 2003. *Manuel de phonologie anglaise*, Paris, Armand Colin / CNED.
Walker, 1791. John. *A Critical Pronouncing Dictionary*, London, G.G.J. and J. Robinson and T. Cadell.
Wells, John, 1990. *Longman Pronouncing Dictionary*, Londres, Longman.
— 1999. "Pronunciation preferences in British English: a new survey", ICPhS-14, 1245-1248.
Zumstein, Franck, 2007. *Variation accentuelle, variation phonétique: étude systématique fondée sur des corpus lexico-phonétiques informatisés anglais*, thèse non publiée, sous la direction de Jean-Louis Duchet, Université de Poitiers.