

1 Le projet "Phonologie du français contemporain"

Le projet "Phonologie du français contemporain" (ci-après *PFC*) vise à constituer une vaste base de données de phonologie française. *PFC* se propose de documenter et de décrire la prononciation du français saisie dans sa variation et dans la réalité de ses usages attestés. Une attention particulière est portée à la diversité géographique, sociale et stylistique du français contemporain. A partir d'un protocole d'enquête uniforme, un groupe international d'une trentaine de chercheurs est impliqué dans la constitution d'un vaste corpus de français parlé à travers le monde. En prenant appui sur des méthodes d'analyse et des outils développés en commun, le projet a pour ambition d'offrir une vision globale et unitaire de la phonologie du français contemporain.

Lorsque qu'on considère les travaux de phonologie française, on ne peut qu'être surpris par l'étroitesse des données sur lesquelles ils s'appuient. Depuis (Dell, 1973), qui propose essentiellement une synthèse des données standard en s'appuyant sur (Fouché, 1959), les données prises en compte apparaissent étonnamment stables, limitées et aseptisées. Toute dimension de diversité sociale, de variation inhérente, stylistique ou géographique en semble exclue. On note pourtant, au cours de la même période, la publication de nombreux travaux descriptifs concernant des usages ou des sociolectes particuliers. Ces travaux, menés dans des perspectives théoriques différentes, fonctionnalistes ((Walter, 1977; Walter, 1982)), laboviennes ((Durand et al., 1987; Laks, 1983), ou plus descriptives ((Hansen, 1997; Wachs, 1998)) sont très difficiles à unifier et restent largement extérieurs au corpus reçu de la phonologie du français. Parallèlement on note également la publication de travaux purement descriptifs, souvent de bonne qualité, à des fins didactiques (Wioland, 1991), (Léon, 1992). Mais les contraintes propres à l'enseignement du Français Langue Etrangère y limitent évidemment l'importance accordée à la variation des usages. Enfin, on sait que les travaux de phonétique expérimentale dans le domaine de l'analyse automatique de la parole (analyse et synthèse vocale, reconnaissance de locuteurs, dictée vocale, interaction vocale homme-machine etc.) sont de gros consommateurs de données fiables, transcrites, alignées sur le signal et proprement étiquetées, mais que des corpus de qualité présentant ces caractéristiques font encore cruellement défaut.

C'est à partir de ces constats que le projet *PFC* s'est construit et défini. Notre premier objectif consiste à définir un nouveau corpus de référence pour la phonologie du français. Il s'agit de recueillir, à partir d'un même protocole, des données fiables dont la comparabilité soit garantie. L'application d'un même protocole, avec les mêmes méthodes d'enquête, les mêmes tests phonologiques, les mêmes tâches de lecture de mots et de lecture d'un texte unique, à une très vaste population de locuteurs du français garantit cette comparabilité interne (*Cf. infra*). De plus chacun de ces outils a été construit pour assurer une comparabilité externe avec les analyses et corpus précédents. Les hypothèses, exemples et contextes phonologiques pertinents de Fouché, Martinet, Walter et Dell ont ainsi été systématiquement pris compte lors de la construction du protocole *PFC*. Outre qu'elle permet une comparabilité externe, et une cumulation inter théorique, une telle stratégie nous fournit également une profondeur historique appréciable. Les données *PFC* pourront ainsi très utilement être comparées à des descriptions partielles d'états de langue plus anciens (Martinet, 1945; Martinet and Walter, 1973; Walter, 1982; Warnant, 1962).

Pour constituer ce corpus de référence, le projet *PFC* se propose de couvrir 40 points d'enquête en France métropolitaine et environ une vingtaine de points dans le reste de la francophonie. A chaque point d'enquête 10 locuteurs en moyenne sont interviewés selon un protocole unique et constant. Deux niveaux stylistiques sont pris en compte, un niveau d'entretien guidé en face à face permettant de documenter un style relativement formel, et un niveau de conversation libre en interaction avec plusieurs enquêtés permettant d'approximer un style relativement informel. Chaque enregistrement dure environ une heure trente et est entièrement digitalisé. A terme, le corpus contiendra donc entre 800 et 1000 heures d'enregistrements numérisés (60 à 80 GO de données). Ce qui constitue la plus grosse base de données orales portant sur le français et l'une des plus grosses base toutes langues confondues. A la fin 2002, deux tiers des enquêtes auront été achevées, la phase d'enquête proprement dite du projet devant prendre fin à l'automne 2003. Pour choisir les points d'enquête, on a tenu compte des informations dialectologiques et sociogéographiques disponibles. Sans rechercher une représentativité construite ou une exhaustivité quelconque, le corpus *PFC*, intègre donc une forte dimension de variation sociale et sociogéographique. Dans chaque région, on a, autant que faire se peut, documenté les zones urbaines, périurbaines et rurales. Pour chaque point d'enquête on a de même cherché à équilibrer les variables de sexe, d'âge, de catégorie socioprofessionnelle et de niveau scolaire. Le corpus *PFC*, sans être construit autour d'une enquête sociolinguistique totalement représentative, objectif hors de portée compte tenu de la taille du corpus visé et des moyens disponibles, est donc un corpus équilibré du point de vue de chacun des axes de variation

repérés par la sociolinguistique. Notre deuxième objectif est ainsi de fournir une meilleure image du français parlé dans son unité et sa diversité sociale, stylistique et géographique.

Nous avons rappelé ci-dessus l'étroitesse du corpus de données récurrentes que les travaux de phonologie théorique et de phonologie française prennent pour objet. Nous avons également souligné leur caractère systématiquement biaisé par l'absence d'enquêtes précises et fiables et par le recours conscient ou inconscient à une norme orthoépique cachée telle qu'articulée notamment par Fouché. Le troisième objectif du projet PFC, est ainsi de critiquer empiriquement cette cristallisation normative fictive et hétérogène que (Morin, 2000) a fort justement baptisé *Français de Référence*. En reprenant systématiquement les hypothèses et contextes clés des diverses phonologies du français, mais en fournissant une empirie nouvelle qui prenne au sérieux la variabilité et la diversité des usages réels de la langue, notre objectif consiste donc également à soumettre à une épreuve critique les analyses et les modèles phonologiques et phonétiques, sur les plans synchroniques comme diachroniques.

Une telle base de données rigoureusement construite au plan des hypothèses phonologiques, intégrant les dimensions de variation et documentant les usages réels est également importante pour tout ce qui concerne le traitement automatique de la parole. L'un des objectifs du projet PFC est donc de favoriser les échanges croisés entre les analyses phonologiques, données empiriques nouvelles et développement d'outils et de modèles spécifiques pour de traitement automatique de la parole; tant en analyse qu'en synthèse. Pour ce faire, une attention particulière est portée à la normalisation des transcriptions, à l'étiquetage, au balisage et à l'alignement texte/son.

Nous avons signalé ci-dessus l'étroitesse des liens existant entre description phonologique du français et didactique du français langue seconde. C'est également une préoccupation du projet PFC que de fournir une base empirique fiable pour l'enseignement. Un formatage particulier permettant un usage pédagogique de la base est ainsi prévu.

Enfin, par son ampleur, par sa volonté de prendre en compte les usages réels dans toute leur variabilité et par la quantité de documents phonologiques réunis, le projet PFC s'inscrit dans une dynamique patrimoniale nettement affirmée et revendiquée. Cette dimension patrimoniale, illustrée dans le passé par les enquêtes de Coquebert de Montbret, Gilliéron, Ferdinand Brunot et plus récemment par l'enquête ESLO, est aujourd'hui assumée par les Archives de la Parole, le dépôt sonore légal à la Bibliothèque Nationale de France et le travail d'archivage de l'Institut National de l'Audiovisuel. Un partenariat avec ces institutions, notamment pour ce qui concerne le dépôt, l'archivage la communication à tiers et la diffusion dans la communauté scientifique des linguistes est ainsi en cours de construction.

2 Méthodologie et protocoles

2.1 Enquête

L'originalité du projet PFC réside dans la combinaison de plusieurs facteurs très rarement réunis : une grande couverture sociogéographique avec plus de 60 points d'enquête au total, la prise en compte de deux styles nettement distincts et définis (le style guidé et le style libre), une diversité sociologique permise par la construction d'un corpus de plus de soixante locuteurs différents, le tout conduit dans un laps de temps relativement bref (3 ans) permettant de maintenir une réelle cohérence et une stabilité des méthodologies. C'est surtout l'administration en tous ces points d'un même protocole d'enquête rigoureusement construit et contrôlé qui constitue l'originalité du projet. Le projet est ainsi à la fois très centralisé au niveau scientifique et fortement décentralisé au niveau des enquêtes. Comme nous l'avons dit, un groupe d'une trentaine de chercheurs participe à l'heure actuelle au projet et conduit des enquêtes locales de façon autonome et décentralisée. Ces chercheurs sont reliés par le web (un site dédié au projet sera opérationnel fin 2002), un Bulletin de Liaison qui paraît deux à trois fois par an, ainsi que deux ateliers de travail et journées d'étude annuels. La centralisation est assurée par la direction scientifique du projet qui contrôle et précise la méthodologie et le protocole commun, tout en pilotant le développement d'outils spécifiques de traitement (*cf. infra*).

D'inspiration labovienne la méthodologie d'enquête fait appel au réseau personnel local de chaque enquêteur. La technique des réseaux denses (Milroy, 1987) permet à l'enquêteur de construire progressivement son panel. Sans prétendre atteindre à tout coup le vernaculaire, la méthodologie d'enquête garantit un type d'interaction linguistique qui ne soit pas trop inégalitaire. « On a ainsi pris le parti de laisser aux enquêteurs la liberté de choisir les enquêtés parmi les *gens de connaissance* ou des gens auprès de qui ils pouvaient être introduits par des gens de connaissance. La proximité sociale et la familiarité assurent en effet deux des conditions principales d'une communication « non violente ». (Bourdieu and alii, 1993), 1395).

Lors de l'interview en face à face, 3 registres principaux peuvent être élicités pour ce qui concerne le style guidé. L'enquêteur enclenche tout d'abord une conversation libre qui permet de recueillir un grand nombre d'informations biographiques et sociologiques qui viendront alimenter le talon sociologique de l'enquête.

permettront de rédiger la fiche de profil sociologique de chaque enquêté. La seconde phase comporte la lecture d'un texte et permet d'accéder à un style un peu plus formel et surveillé. La troisième phase consiste dans la lecture d'une liste de 94 mots, dont dix paires minimales classiques de la phonologie du français. Cette dernière lecture particulièrement formelle permet d'accéder au registre le plus soutenu du locuteur. Un quatrième registre, nettement moins formel et plus proche du vernaculaire est atteint dans une seconde session, séparée, où l'enquêteur enregistre une conversation libre entre plusieurs enquêtés liés par des relations personnelles fortes. Cette situation permet de lever, dans une certaine limite, le paradoxe de l'observateur (Labov, 1976), et de s'approcher du parler ordinaire des enquêtés.

2.2 Hypothèses

Les 4 registres considérés permettent de construire de façon sûre le profil phonologique de chaque enquêté. La lecture du texte, de la liste de mots et des paires minimales permet de construire la charte phonémique moyenne de chaque locuteur en testant systématiquement les oppositions phonémiques qu'il pratique selon les divers registres. On obtient ainsi un profil phonologique qui peut très utilement être comparé à ceux, un peu plus sommaires, dressés il y a trente ans par (Walter, 1977). Les exercices de lecture ont été construits en incorporant systématiquement tous les phénomènes connus affectant le système phonémique français : opposition de hauteur et de longueur, loi de position et ajustement en syllabe fermée, neutralisation en syllabe ouverte, allègement des groupes obstruente/liquide en position finale, assimilation de sonorité progressive et régressive, inventaire des voyelles nasales, nasalisation et dénasalisation des groupes orale/nasale, diérèse et synérèse, palatalisation, réalisation de /r/ etc. Une attention particulière a été portée aux phénomènes les plus massifs et les plus problématiques de la phonologie du français : liaison et élision, e muet, h aspiré. En particulier, tous les contextes pertinents cités par les analyses classiques ((Dell, 1973; Fouché, 1959) ont été incorporés. Seuls les phénomènes accentuels, rythmiques et prosodiques n'ont pas été directement pris en compte.

L'analyse phonologique permise par le corpus *PFC* est encore affinée par ce que nous nommons le principe d'asymétrie des lectures. En effet, puisque nous disposons de 4 registres différents appartenant à deux styles nettement distincts, nous pouvons rendre compte des usages réels avec une finesse jamais acquise auparavant. Tel locuteur qui maintient encore des oppositions phonémiques apparemment stables dans le test le plus formel de lecture des paires minimales a un comportement beaucoup plus fluctuant dans la lecture des mots séparés, et ne maintient plus ces oppositions dans la lecture du texte. Tel locuteur qui ne réalise que très peu de liaisons facultatives en conversation libre, en maintient un nombre significatif en lecture. Enfin le comportement de E muet tant en syncope qu'en épenthèse peut être approché sur un nombre de contextes quantitativement considérables et avec une variabilité stylistique interne et géographique externe jamais atteintes à ce jour. On documente ainsi en moyenne 4000 contextes de schwa et 2000 contextes de liaison possibles par point d'enquête.

2.3 Protocoles et transcriptions

Au terme du projet, la base *PFC* comportera environ 60 GO octets de données stockées sous la forme de bases de données relationnelles interrogeables selon un protocole de type SQL. L'ensemble des données recueillies pour les 2 styles, les 60 points d'enquête et les 600 locuteurs est archivé, versé au dépôt légal et accessible à la communauté des linguistes. Pour chaque locuteur on dispose d'une fiche sociologique et d'un ensemble de fichiers dont la dénomination est normalisée. Les quatre enregistrements, correspondant aux quatre styles, de chaque locuteur sont entièrement numérisés (mono, 16 bits, 44 à 66 KHZ selon la qualité originelle du signal), seules les répétitions, incises et commentaires lors des lectures sont expurgés.

On obtient alors un fichier par registre considéré. Ces fichiers sont utilisés dans le travail de transcription et d'alignement du texte sur le signal. On utilise le logiciel Praat développé par Boersma et Weenink Pour chaque style, conversation guidée et conversation libre on génère sous Praat un fichier de balisage Textgrid dans lequel on fournit à chaque fois dix minutes de transcription orthographique alignée sur le signal. Les principes de transcription s'inspirent des travaux et des expériences antérieures dans la transcription de gros corpus de français parlé (GARS à Aix-Marseille, VALIBEL à Louvain-la-neuve), notamment pour ce qui concerne la gestion des tours de parole, la transcription des pauses, des reprises, des incises et des erreurs. Ils visent à alléger la tâche tout en permettant un balisage complet et un alignement précis sur le signal.

Les deux transcriptions de dix minutes alignées sont soumises à une vérification croisée par un transcripateur différent. Sur la base de cette transcription on génère alors, pour chaque style, une analyse exhaustive des contextes de schwa et des contextes de liaison. Chacune de ces analyses prend la forme d'un codage complet de 5 et 3 minutes respectivement, inscrit sur une tire séparée, de ces contextes alignés. On fournit d'autre part un codage complet pour schwa et la liaison du registre 2 du style guidé (lecture du texte). Un ensemble de fiches de commentaires phonétiques et phonologiques complètent la livraison.

Ainsi, en même temps qu'elle offre à chaque chercheur la possibilité de suivre ses propres hypothèses et de construire ses propres objets phonologiques sur la base sonore digitalisée, *PFC* offre un ensemble de préanalyses

particulièrement utiles en fournissant des transcriptions stables et vérifiées alignés sur le signal. S'agissant de schwa et de la liaison, PFC offre un ensemble d'analyses finales quantifiées avec un détail et précision jamais atteints : codage complet et exhaustif de 3 et 5 minutes de parole selon deux styles, ainsi que la lecture du texte, aligné sur le signal.

Pour ce qui concerne la liaison, le système de codage rend compte de la présence ou de l'absence de liaison, de son type de réalisation (enchaîné ou non), de la nature phonique de la consonne liaisonante (avec une attention particulière portée aux nasales) ainsi que de la longueur syllabique du mot liaisonant. Le codage liaison du premier paragraphe du texte lu pourra se présenter ainsi :

Le village de Beaulieu est11t en grand12h émoi. Le Premier Ministre a en11nVO effet décidé de faire étape dans cette commune au cours de sa tournée de la région en fin d'année. Jusqu'ici les seuls titres de gloire de Beaulieu étaient son vin blanc sec, ses chemises20 en soie, un champion local de course à pied (Louis Garret), quatrième aux jeux11z olympiques de Berlin en 1936, et plus récemment, son11nVN usine de pâtes10h italiennes. Qu'est-ce qui a donc valu à Beaulieu ce grand11t honneur Le hasard, tout bêtement, car le Premier Ministre, lassé des circuits20 habituels qui tournaient toujours21z autour des mêmes villes, veut découvrir ce qu'il appelle "la campagne profonde".

On obtient ainsi à terme environ 200 codes quadrilitères par locuteur soit 120 000 contextes de liaison codés dans PFC.

Pour ce qui concerne le schwa, le codage prend en compte la présence ou l'absence d'une réalisation, la position, le contexte droit et gauche. Le digit de position sous code la longueur syllabique et le rang, dans le cas des polysyllabes. Les deux digits de contexte sous codent la qualité vocalique ou consonantique du contexte, la présence de groupes consonantiques lourds ou simplifiés, la force des frontières intonatives et des structures rythmiques. Le codage schwa du deuxième paragraphe du texte pourra se présenter ainsi :

Le112 maire0412 de1122 Beaulieu - Marc0422 Blanc - est en re1212vanche0412 très inquiet. La cote0412 du Premier Ministre2422 ne1142 cesse0412 de1122 baisser de1212puis les élections. Comment, en plus0413, éviter les manifestations qui ont eu tendance0411 à se1112 multiplier lor0412s des visite1412s officielle0413s ? La côte0411 escarpée du Mont Saint Pierre0413 qui mène0411 au village0413 connaît des barrage0413s chaque0412 fois que1112 les opposants de1112 tous les bor0412ds manifeste1422nt leur0412 colère0413. D'un autre1422 côté, à chaque0412 voyage0412 du Premier Ministre1423, le1132 gouverne1322ment prend contact0421 avec0412 la préfecture0412 la plus proche0413 et s'assure2412 que1142 tout est fait pour0412 le1122 protéger. Or0413, un gros détache0312ment de1112 police0413, comme0411 on en a vu à Jonquière0413, et des vérifications d'identité risque1422nt de1112 provoquer une0411 explosion. Un jeune0412 membre1422 de1112 l'opposition aurait déclaré : "Dans le1112 coin, on est jaloux de1112 notre1422 liberté. S'il0412 faut montrer patte0412 blanche0412 pour0412 circuler, nous ne1112 répondons pas de1112 la réaction des gens du pays

On obtient ainsi à terme environ 400 codes quadrilitères de schwa par locuteur soit 240000 contextes de schwa codés dans PFC.

3 Outils et traitements

Engagés dans la construction et le traitement d'un corpus de français parlé à grande échelle, nous n'avons pas trouvé sur le marché tous les outils d'analyse qui nous étaient nécessaires. En effet, comme nous l'avons rappelé ci-dessus, notre opération est exceptionnelle à plus d'un titre, par l'ampleur, la variété géographique, sociale et stylistique des interviews conduits. Il nous fallait donc disposer d'outils spécifiques de gestion du corpus et de comparaison systématique des productions des différents locuteurs. De plus, le projet PFC incorpore des hypothèses phonologiques très précises quant aux processus actifs en français et aux contextes pertinents dans lesquels ils sont susceptibles d'avoir lieu. Pour assurer la comparabilité externe et garantir la profondeur historique critique de nos traitements nous avons d'autre part cumulé les hypothèses et les résultats de nos prédécesseurs. Il nous fallait donc disposer d'outils de fouille, d'analyse contextuelle et d'inspection systématique du corpus tout à fait particulier. Pour l'ensemble de ces raisons, nous nous sommes engagés dans la construction d'outils PFC spécifiques.

3.1 Transpraat

La base PFC propose des transcriptions larges alignées sur le signal ainsi qu'un sous-ensemble de transcriptions et de codages, beaucoup plus fins eux, également balisés sous Praat et alignés sur le signal. La grande souplesse de Praat pour la segmentation, le balisage et l'alignement texte/signal ainsi que les souplesses qu'il offre dans la gestion de tires multiples synchronisées se paye d'un inconvénient majeur concernant la lisibilité des fichiers TextGrid qu'il génère. Voici par exemple un écran Praat et le TextGrid correspondant :

intervals [26] :
xmin = 126.92362249533851
xmax = 128.64820506304423
text = "E : Vous êtes à Toulouse depuis combien de temps ?"

intervals [27] :
xmin = 128.64820506304423
xmax = 131.31852645820149
text = "D.P : Alors ça va faire la cinquième année, que je suis à Toulouse."

intervals [28] :
xmin = 131.31852645820149
xmax = 133.26563580883698
text = "D.P : Voilà < E : et qu'est-ce que vous faites comme études ? >"

intervals [29] :
xmin = 133.26563580883698
xmax = 136.60353755278356
text = "D.P : Alors là, je suis en licence et en maîtrise de lettres modernes,"

intervals [30] :
xmin = 136.60353755278356
xmax = 141.10970490711142
text = "D.P : euh je vais terminer ma licence et puis ma maîtrise en même temps."

Pour assurer la lisibilité des fichiers et faciliter la consultation de base, notre outil Transpraat génère des fichiers textes standard lisibles sous tout logiciel. Pour l'exemple ci-dessus on obtient :

E : Vous êtes à Toulouse depuis combien de temps ?

D.P : Alors ça va faire la cinquième année, que je suis à Toulouse. Voilà < E : et qu'est-ce que vous faites comme études ? > Alors là, je suis en licence et en maîtrise de lettres modernes, euh je vais terminer ma licence et puis ma maîtrise en même temps.

On remarquera notamment que les quatre derniers intervalles ont été fusionnés dans la mesure où il s'agit d'un seul tour de parole produit par le locuteur DP.

3.2 Extracteur automatique de formants et constructeur de charte formantique.

L'une des dimensions de la variation sociogéographique parmi les plus importantes de la phonologie du français concerne le système phonémique moyen de chaque locuteur. Il s'agit d'un problème particulièrement bien documenté en phonologie structurale pour lequel Martinet puis Walter ont proposé des analyses détaillées, tant pour la couche synchronique qu'ils considéraient que pour la diachronie d'ensemble du système. Pour la synchronie considérée par *PFC*, il était donc important de pouvoir disposer pour chaque locuteur d'une répartition de ses voyelles moyennes dans un espace F1/F2 et de pouvoir ainsi documenter la variabilité sociogéographique au plan systémique. De plus, de telles analyses permettaient une comparabilité externe et rencontraient notre souci de profondeur historique déjà souligné. Si l'enquête de (Walter, 1982) qui ne comportait que 111 locuteurs et était bien plus limitée que la notre en termes d'ampleur du corpus pouvait encore se contenter d'analyses manuelles faites à l'oreille, il était clair que *PFC* ne pouvait adopter un fonctionnement identique. Nous avons donc engagé la construction d'un outil spécifique qui nous permette de disposer pour chaque locuteur du corpus de son système phonémique moyen, de le comparer à celui d'autres locuteurs situés en divers points de l'espace sociogéographique couvert par *PFC* et de le comparer aux descriptions plus anciennes de Walter et de Martinet.

La construction de cet outil, assurée par l'équipe *PFC* d'Aix en Provence est en cours. Pour chaque locuteur, on dispose d'enregistrements absolument comparables : la lecture du texte, de la liste de mots et des paires minimales. De plus, la lecture de chaque mot de la liste est précédée de la lecture de son numéro d'ordre qui fournit un référentiel simple, stable et particulièrement utile. Le balisage temporel et l'alignement sur le signal étant acquis, on peut pour chaque classe phonémique extraire automatiquement la structure formantique correspondante. Pour l'ensemble des structures formantiques correspondant à une classe phonémique donnée on analyse ensuite sa dispersion statistique autour d'une cible centrale afin de décider s'il s'agit bien d'une cible unique comme c'est le cas dans les dialectes qui neutralisent par exemple la différence entre voyelles moyennes et basses (piqué/piquet) ou si la classe est scindée comme c'est le cas pour les dialectes qui maintiennent cette différence. On obtient finalement par extraction automatique des formants une charte moyenne pour chaque locuteur, indiquant pour chaque phonème l'ensemble de ses réalisations dans l'espace F1/F2 et définissant des cibles statistiques moyennes. Ce résultat correspond au profil phonémique moyen du locuteur, il permet une comparaison sociogéographique et historique simple et très éclairante.

3.3 Comparateur

L'équipe de Toulouse a développé un outil *PFC* qui permet une comparaison simple des performances de lecture des mots de la liste pour un nombre quelconque de locuteurs. Le logiciel intègre un outil de recherche permettant une comparaison en fonction de divers paramètres tels que l'âge, la localité ou la profession des locuteurs. Dans sa version actuelle, l'outil est limité à la lecture de la liste (autrement dit celle des 94 mots précédés du numéro d'ordre). Une nouvelle version permettra notamment de comparer les mots de la liste et ceux du texte lu. Par la suite, le comparateur sera intégré à l'interface globale de visualisation de la base de données *PFC*. L'outil comprend un séquenceur qui permet d'aligner très précisément le mot sur sa réalisation et un moteur de comparaison. Il permet au phonologue de pouvoir produire et tester rapidement des hypothèses phonémiques.

3.4 Le classeur Schwa

Cet outil, également développé par l'équipe *PFC* de Toulouse prend en entrée aussi bien des TextGrids que des sources reformatées par Transpraat. Les textes d'entrée sont codés pour schwa. L'outil permet La fonction de cet utilitaire est de classer et de comptabiliser les occurrences des codes mis au point pour l'analyse du schwa.

4 Conclusion

La description et l'analyse du français oral constituent un terrain scientifique ancien et très fréquenté. La description des processus phonologiques remonte ainsi aussi loin que l'institutionnalisation de la langue elle-même puisque dès 1530 (Palsgrave, 1530) propose une analyse assez précise de la liaison, de l'élision, de l'accentuation et du rythme. Les descriptions et les analyses, souvent basées sur l'intuition du linguiste ou sur sa reconnaissance de la norme académique, se sont depuis lors multipliées. A la fin du 20^{ème} siècle, les développements de la phonétique instrumentale et descriptive comme celle des premiers dispositifs d'enregistrement de la parole font naître un grand espoir dont Paul Passy en France et Daniel Jones en Grande-Bretagne se sont fait les chantres et les artisans. Espoir vite déçu puisque aussi bien la phonologie du français poursuivra imperturbablement son cours d'exégèse normative ou de contemplation théorique et que la phonétique du français continuera quant à elle d'accumuler des descriptions éparses et ponctuelles, mais sans systématique et sans construction d'un cadre hypothético-descriptif précis.

Des atlas linguistiques de (Gilliéron and Edmont, 1902-1912), aux grandes enquêtes dialectologiques ou sociolinguistiques, l'histoire de la phonologie de corpus en France est ainsi celle d'un descriptivisme marginal sans grande influence sur le cours des phonologies analytiques ou prescriptives plus nobles parce que plus académiques. Pourtant, ce que nous apprennent l'histoire et l'épistémologie des sciences du langage c'est qu'il ne suffit pas de décrire et d'observer pour rendre compte d'un objet linguistique; il faut encore le construire et l'interroger dans une problématisation scientifique qui seule permet de le saisir comme objet. A l'inverse, toute l'histoire de la phonologie du français montre qu'il ne suffit pas de s'interroger sur les usages, voire d'exercer son introspection sur le sien propre, pour rendre compte de la langue dans sa diversité et sa variation. Sans protocole empirique et sans recollection systématique de données, la phonologie se trouve sans objet et c'est, comme l'on sait, la norme et le standard le plus reçu, et donc le moins fondé, qui viennent alors, *volens nolens*, occuper le vide empirique ainsi créé.

L'histoire des rapports entre phonologie et phonétique du français est ainsi celle d'un mariage annoncé mais sans cesse repoussé faute de premier rendez-vous. L'objectif premier du projet *PFC* est d'abord d'organiser la rencontre en rabattant sur l'enquête les problématizations théoriques de phonologie française dont elle a besoin pour construire son interrogation comme systématique et informative; c'est aussi de renvoyer à la phonologie du français le *datum* qui lui fait défaut, un *datum* mis en forme et problématisé, qui entre directement en résonance avec les questionnements théoriques qui sont les siens. Au-delà, l'objectif du projet *PFC*, en construisant une base de donnée représentative de la variété sociale, géographique et stylistique des usages phonologiques du français vise à constituer le référentiel contemporain dont la linguistique française a besoin. C'est assurément le rôle de la linguistique de corpus que de construire ce référentiel empirique, de développer les bases de données, de définir leur organisation, de développer les outils de fouille dont la linguistique a besoin. Mais ceci ne peut se faire hors des problématizations théoriques qui structurent les sciences du langage. On sait qu'il n'existe pas de corpus omnibus et qu'une base de données n'est pas un couteau suisse. C'est la raison pour laquelle le projet *PFC*, projet de phonologie de corpus se veut à la fois phonétique et phonologique, empirique et théorique, de terrain et d'élaboration conceptuelle. C'est à ce prix nous semble-t-il que nous comprenons mieux ce qu'il en est à l'oral des usages du français.

Bibliographie

BERGOUNIOUX, G. « Les enquêtes de terrain en France », *Langue Française* 93, 1992, p. 3-23.

BOURDIEU, P. et alii, *La misère du monde*, Paris, Editions du Seuil, 1993.

DELL, F., *Les règles et les sons : Introduction à la phonologie générative*, Paris, Hermann, 1973.

DURAND, J., SLATER, C. et WISE, H., « Observations on schwa in southern French », *Linguistics* 25, 1987, p. 993-1004.

ENCREVE, P., *La liaison avec et sans enchaînement : phonologie tridimensionnelle et usages du français*, Paris, Seuil, 1988.

FOUCHE, P., *Traité de prononciation française*, Paris, Klincksieck, 1959.

GILLIERON, J. et EDMONT, E., *Atlas linguistique de la France*, Paris, Champion, 1902-1919.

GRAMMONT, M., *Traité pratique de prononciation française*, Paris, Delagrave, 1914.

HANSEN, A.B., « Le nouveau [] prépausal dans le français parlé à Paris », PERROT, J. (ed.) *Polyphonie pour Ivan Fonagy*, Paris, Editions l'Harmattan, 1997, p. 173-198.

LABOV, W., *Sociolinguistique*, Paris, Editions de Minuit, 1976.

LAKS, B., « Langage et pratiques sociales : étude sociolinguistique d'un groupe d'adolescents, Actes de la recherche en sciences sociales, 46, 73-97, 1983.

LAKS, B., « Description de l'oral et variation : la phonologie et la norme », *L'information grammaticale*, 2002, 5-11.

LEON, P., *Phonétisme et prononciation du français : avec des travaux pratiques d'application et leurs corrigés*, Paris, Nathan, 1992.

MARTINET, A., *La prononciation du français contemporain : témoignages recueillis en 1941 dans un camp d'officiers prisonniers*, Genève, Droz, 1945.

MARTINET, A. et WALTER, H., *Dictionnaire de la prononciation française dans son usage réel*, Paris, France-Expansion, 1973.

MILROY, J., *Observing and analysing natural language. A critical account of sociolinguistic method*, Oxford, Blackwell, 1987.

MORIN, Y.-C., « Le français de référence et les normes de prononciation », *Cahiers de l'Institut de linguistique de Louvain*, 26-1, 2000, p. 91-135.

PALSGRAVE, J., *L'esclarcissement de la langue françoise, composé par maistre Jehan Palsgrave, Angloys, natyf de Londres et gradué de Paris*, Londres, 1530.

THUROT, C., *De la prononciation française depuis le commencement du XIV^e siècle, d'après les témoignages des grammairiens*, Paris, Bibliothèque Nationale, 1881-1883.

TRANEL, B., *Concreteness in Generative Phonology : Evidence from French*, Berkeley, University of California Press, 1981.

TRANEL, B., *The Sounds of French : an Introduction*, Cambridge, Cambridge University Press, 1987.

WACHS, S., « Le relâchement de la prononciation en français parlé de l'Ile de France. Analyse linguistique et sociolinguistique par générations », Thèse de doctorat non publiée, Paris X-Nanterre, 1998.

WALTER, H., *La phonologie du français*, Paris, Presses Universitaires de France, 1977.

WALTER, H., *Enquêtes phonologiques et variétés régionales du français*, Paris, Presses Universitaires de France, 1982.

WARNANT, L., *Dictionnaire de la prononciation française*, Munich, Gembloux, 1962.

