PFC, corpus et systèmes de transcription

Jacques Durand et Jean-Michel Tarrier*

La constitution de grands corpus oraux soulève la question de l'indexation des données. Dans cet article, nous présentons la solution adoptée dans le projet 'Phonologie du français contemporain (PFC) : usages, variétés et structure', à savoir la construction initiale d'une tire de transcription proche de la norme orthographique. Nous présentons ici les conventions de transcription orthographique utilisées au sein du projet PFC et examinons quelques-unes des difficultés auxquelles on se heurte devant la diversité des usages. En dépit de désavantages possibles, nous défendons ici l'idée que les conventions orthographiques usuelles très légèrement modifiées restent le meilleur système d'accès aux données orales.

The constitution of large oral corpora raises the question of how one should index one's data. In this article, we present the solution adopted within the PFC project ('Phonologie du français contemporain : usages, variétés et structures'), namely the setting up of an orthographic tier that is close to standard spelling. We introduce here the orthographic conventions used within the PFC project and examine in some detail the difficulties one faces given the diversity of usage. Despite possible drawbacks, we argue that standard orthographic transcription with minor modifications remains the best entry point for a close study of large oral corpora.

* ERSS (UMR 5610) & Université de Toulouse-Le Mirail.

Cahiers de Grammaire 30 (2006)
« Spécial Anniversaire », pp. 139-158

1. Introduction[1]

Si on se fie au contenu des revues scientifiques, aux collections qui apparaissent, aux colloques internationaux, aux projets déposés auprès des grands organismes de recherche, la linguistique est entrée dans une ère nouvelle. Au cœur des préoccupations actuelles se situe le traitement de grandes masses de données à travers la constitution de « corpus »[2]. La plupart des défenseurs de l'introspection (au sens de « jugement de bonne formation ») comme technique incontournable d'appréhension du langage s'accordent désormais à reconnaître que la construction et l'exploitation de grands corpus est indispensable si l'on veut éviter les pièges de l'intuition et confronter les modèles théoriques aux observables. Le projet PFC (*Phonologie du français contemporain : usages, variétés et structure*), coordonné par Jacques Durand, Bernard Laks et Chantal Lyche, place au centre de sa démarche l'élaboration d'un corpus de référence du français parlé[3]. Ce dernier ne consiste cependant pas en un simple enregistrement de données orales mais dans la construction d'une véritable base de données formatée et annotée de façon à assurer son exploitation par la communauté et sa pérennité dans le temps. Dans cet article, nous abordons un seul problème qui est celui de la transcription orthographique adoptée dans PFC comme « couche zéro » du système d'annotation. Nous présentons et défendons les choix méthodologiques qui ont été faits au sein du projet et essayons d'en examiner les limites possibles.

2. Pourquoi une transcription orthographique ?

Nous ne décrirons pas ici le protocole d'enquête de PFC. Nous rappellerons qu'il vise à constituer une base d'enregistrements sonores dans un format rigide donc strictement comparable qui comprend quatre tâches : lecture (à haute voix) d'une liste de mots puis d'un texte, conversation guidée et conversation libre. Une fois les enregistrements effectués, ils sont numérisés et transférés sur des supports informatiques[4]. Dans le projet PFC, nous avons

1 Nous tenons à remercier Béatrice Akissi-Boutin, Julien Eychenne, Chantal Lyche, François Poiré, Gisèle Prignitz et Gabor Turcsan de leur aide lors de la préparation de cet article. Leurs réflexions et leurs exemples se sont révélés très précieux même s'ils ne sont pas responsables de l'usage qui en été fait ici.

2 Voir par exemple Sampson (2005).

3 Durand et Lyche (2003), Durand, Laks, Lyche (2005), Durand (sous presse).

4 Nous utilisons ici le terme 'numériser' dans un sens non technique puisque les enregistrements à partir de DAT ou d'enregistreurs de type minidisc sont déjà, par définition, numériques. Quel que soit le support sur lequel les enregistrements ont été faits, il faut les rendre interprétables par un ordinateur. C'est le sens dans lequel nous prenons donc le terme 'numériser'. Sur ces questions, voir Tarrier (2003).

opté pour des fichiers son au format standard .WAV. Nous avons également mis au point une structuration commune de tous les fichiers de données et d'analyse, ce qui est indispensable au travail scientifique dans un projet décentralisé[5].

La numérisation effectuée, se pose alors la question de la transcription ou de l'annotation des bases de données sonores. Dans le projet PFC, nous adoptons comme « couche zéro » une transcription orthographique qui sert de point de départ à des indexations, des codages et des investigations phonétiques. Cette transcription se fait à partir d'un outil informatique, PRAAT développé par P. Boersma et D. Weenink, qui permet d'aligner les représentations graphiques sur le signal (http://www.fon.hum.uva.nl/praat/).

Lors du passage d'un corpus oral à un texte écrit, le transcripteur est confronté d'emblée à la question suivante : comment refléter le caractère oral du corpus ? Une première réaction, qui a souvent tenté phonologues et phonéticiens, consiste à proposer une transcription phonétique large ou étroite à partir des symboles que fournit un système comme l'Alphabet Phonétique International. Cette stratégie, qui est tout à fait motivée pour un projet à petite échelle, pose des problèmes épineux dans un projet décentralisé où l'on doit analyser une masse de données importante. En effet, il est rare que les transcripteurs s'accordent sur la valeur des symboles et des diacritiques dès qu'on veut noter des phénomènes relativement fins. Mais au-delà du manque d'accord inter-subjectif, il est important de signaler qu'une transcription dite phonétique, comme premier niveau d'abstraction à partir du signal, pose de réelles difficultés. Si on adopte une transcription large, ou plus précisément phonémique, on met la charrue avant les bœufs : on suppose qu'on a déjà découvert le système qu'on cherche à établir à travers l'enquête. Si on privilégie au contraire des transcriptions étroites (de type allophonique), on se heurte à une difficulté de taille. Quel degré de finesse phonétique doit-on adopter ? En effet, au niveau phonétique, de nombreuses réalisations ne correspondent pas à des choix binaires mais à des valeurs sur des échelles continues. Ainsi la longueur des voyelles au sein des énoncés n'est pas simplement une opposition entre long (:) et bref, ni même entre long, demi long et bref, mais entre divers degrés de longueur que seule une étude acoustique précise peut résoudre. De même, l'assimilation de voisement n'est pas toujours un simple changement de valeur (voisé *vs* non voisé) mais un phénomène scalaire et qui, selon divers spécialistes, entre en interaction avec d'autres paramètres phonétiques comme la dimension tendu/lâche. Ce travail ne peut être fait de manière fiable qu'en utilisant des outils de mesure acoustique. Mais s'engager dans une étude acoustique demande beaucoup de temps et d'énergie même pour un corpus de petite taille et exige de toute façon un déblayage du terrain. Pour ces diverses

5 Voir Durand, Laks & Lyche (2002) sur le format des rendus PFC.

raisons et pour d'autres (cf. §3.3), nous avons privilégié un premier niveau de transcription orthographique qui constitue la base du travail que nous effectuons sur le corpus. Cette couche zéro, du fait même qu'elle ne prétend pas refléter fidèlement la structure phonique, permet une exploration beaucoup plus rapide des enregistrements et en retour un travail plus précis sur les données sonores chaque fois que cela s'avère nécessaire.

3. Le système de transcription PFC

Les conventions employées dans PFC ont déjà été présentées dans d'autres travaux et seront reprises ici pour permettre au lecteur de mieux comprendre nos questionnements en §4. Ces conventions doivent beaucoup au travail effectué autour du corpus du GARS à Aix (Université de Provence) et de VALIBEL (UCL, Louvain-la-Neuve)[6]. Nous avons aussi tenu compte des recommandations du groupe européen EAGLES (1996a,b) et des instructions de Gjert Kristoffersen (Université de Bergen) pour la transcription d'une banque de données sur les dialectes norvégiens. Nous avons opté dans le projet pour une transcription orthographique standard (TOS ci-après), y compris une ponctuation standard (mais plus réduite que la norme orthographique française)[7]. Nous commencerons par un exemple concret en 3.1. avant d'expliquer les principales conventions en §3.2. En §3.3, nous revenons sur quelques-uns des avantages du type de transcription orthographique proposée.

3.1. Exemple de transcription (conversation guidée)

Contexte: L'enquêtrice (E) qui est originaire de Belgique interviewe une jeune locutrice (AB) de Pézenas (Hérault, 34120 France) qui lui parle de l'enseignement à l'université en France et des qualifications requises.

AB: Par rapport à des amis que j'ai, qui, <E: C'est vrai?> qui ont beaucoup de mal maintenant à enseigner sans l'agrégation à l'université, <E: A l'université?> ouais, <E: Ah bon?> alors qu'il y a cinq six ans en, en arrière, c'était plus facile, apparemment, <E: Ah bon?> ouais. Peut-être, (rire) peut-être en histoire, c'est peut-être différent pour d'autres domaines. <E: Ça dépend peut-être des

[6] Pour les conventions du GARS, cf. Blanche-Benveniste & Jeanjean (1987), Blanche-Benveniste et alii (1991), Bilger (2000). Voir également le site suivant: http://www.up.univ-mrs.fr/delic/corpus/index.html. Pour VALIBEL, Centre de recherche sur les Variétés linguistiques du français en Belgique (dir. M. Francard), voir le site suivant http://valibel.fltr.ucl.ac.be/. Pour une discussion plus générique des questions de transcription, voir Delais-Roussarie (2003).

[7] Les conventions de transcription PFC et leur intégration à PRAAT sont expliquées dans Delais-Roussarie, Durand, Lyche, Meqqori et Tarrier (2002).

domaines ?> Hein, voilà, c'est ça, hein, peut-être que des profs d'histoire, il y en a tellement que, bon, euh, le nombre de jeunes gens qui ont des doctorats en histoire, que, bon, maintenant on peut pas être en fac sans l'agrégation.
E: Peut-être est-ce, peut-être est-ce, voilà une autre contrainte qu'ils imposent. Pour ça, je ne saurais pas le dire. Je ne saurais pas le dire, je sais pas du tout.
AB: Vous avez le même cursus, le même système universitaire en Belgique qu'en France, oui, c'est pareil?
E: Euh, ça s'appelle pas <E: C'est la même chose?> mais enfin c'est la même chose.

3.2. Conventions de transcription PFC
3.2.1. Formes de mots

Au niveau lexical, les diverses réalisations d'un mot sont ramenées à la forme standard de ce mot en contexte. Dans la séquence parlée 'Je suis plus petite', que le locuteur ait dit [ptit] ou [ptitə] ou [pœtit] ou [pt^{s}ɪt], etc., nous transcrivons 'petite'. Que le locuteur ait dit [ply] ou [py], nous transcrivons 'plus'. Qu'il ait dit [ʒə sɥi] ou [ʒsɥi] ou [ʃsɥi] ou [ʃɥi] etc., nous transcrivons 'je suis'. En revanche, nous ne réintroduisons pas des éléments lexicaux absents. Ainsi [ivwapasakɔmsa] sera transcrit 'Il voit pas ça comme ça' et non pas 'Il ne voit pas ça comme ça' avec ré-insertion normative de 'ne'.

3.2.2. Tours de parole et ponctuation

Chaque locuteur est désigné par ses initiales, l'enquêteur par 'E' ou bien par E1, E2, etc., s'il y a plusieurs enquêteurs. Les initiales sont suivies de deux points (:), et sont reprises à chaque intervalle même s'il n'y a pas de changement de locuteur. Il n'y a aucun paragraphe, ni retour à la ligne. Une pause brève (ou un mouvement mélodique indiquant une continuation) est indiquée par une virgule, une pause plus longue (ou une fin d'énoncé marquée mélodiquement) par un point. Une question est signalée par un point d'interrogation. Nous n'utilisons ni le point d'exclamation ni le point virgule.

3.2.3. Chevauchements

Lorsqu'un locuteur L1 parle et qu'un autre locuteur L2 se manifeste uniquement par des appréciations en arrière-plan comme *oui, non, hum, pff*, ces remarques discursives sont ignorées et ne sont pas transcrites. Par contre, si L2 interrompt véritablement son interlocuteur en produisant un énoncé, son intervention sera indiquée entre chevrons à l'intérieur du discours de L1, si ce dernier continue à parler. S'il y a véritable interruption et que cela entraîne un changement de locuteur, nous l'indiquons par un changement d'intervalle. Dans les deux cas, le locuteur qui cause le chevauchement est indiqué de

façon univoque : <L : sans espace entre les différents symboles, mais avec un espace après les deux points.

- *Exemple 1* : L1 conserve la parole, malgré l'intervention de L2
L1 : Tu as vu Pierre ces derniers <L2 : Pierre, non.> temps ? Il est parti à Paris.
L2 interrompt L1 qui garde cependant la maîtrise du tour de parole.

- *Exemple 2* : L1 conserve la parole, même si L2 intervient plus longuement.
L1 : Tu as vu Pierre ces derniers < L2 : Pierre, non. Je suis plus en contact avec lui.> temps ? Il est parti à Paris.

- *Exemple 3* : L2 interrompt L1
Ici nous adopterons la convention suivante avec minuscules pour la prise de parole de L2.
L1 : Tu as vu Pierre ces derniers <L2 : Pierre, non.> temps ? Il est parti à Paris et il devait rentrer la semaine passée <L2 : Pierre.>
L2 : est vraiment un chic type. Et d'ailleurs, tout le monde l'apprécie.
L2 interrompt L1, 'Pierre' est articulé en même temps que 'passée' et L2 garde la parole.

3.2.4. Mots tronqués

La troncation d'un mot sera indiquée par une barre oblique suivie d'un espace.

- *Exemple 1* :
L1 : Il m'a pro/ pro/ mis de revenir vite.
L1 : Oui, et main/ maintenant il a enfin compris.

S'il y a une pause à l'intérieur d'un mot, mais que le locuteur, après cette pause, termine le mot sans répéter de syllabe, cela sera indiqué par une barre oblique immédiatement suivie d'un trait d'union.

- *Exemple 2* :
L1 : Il m'a pro/-mis de revenir vite
Les mots répétés, eux, sont repris et séparés par une virgule.
L1 : Il lui, lui, lui a dit de revenir vite.

3.2.5. Sigles

Les sigles épelés sont transcrits avec chaque lettre séparée par un point, alors que les sigles lus sont transcrits comme un mot. La prononciation inattendue d'un sigle lu peut être indiquée entre parenthèses en SAMPA[8].

- Exemple 1 :
L1 : La S.N.C.F. est à nouveau en grève.
NB : ici le sigle est prononcé [laɛsɛnseɛf] (soit [laEsEnseEf] en SAMPA), ce qui est attendu et n'est donc pas noté.
L1 : Il a été admis à l'I.U.T. de Caen.

De même, le sigle est prononcé [liyte] (également [liyte] en SAMPA) et n'est pas noté.

- Exemple 2 :
L1 : Le CNET (snET) embauche du personnel.

Dans ce dernier exemple, la personne a prononcé ce sigle d'une façon inhabituelle (à savoir, [snɛt]), d'où l'utilisation de la transcription SAMPA (snET) entre parenthèses.

3.2.6. Hésitations, bruits, onomatopées

Les hésitations seront transcrites *euh*. A la finale des mots il est parfois difficile de décider s'il y a hésitation ou prononciation d'un schwa. Dans TOUS les cas, la transcription sera *euh*.

Certaines productions du locuteur peuvent être transcrites de façon naturelle par des graphies conventionnelles admises dans les dictionnaires usuels : *pfft, psitt, tss-tss, tsoin-tsoin*. Dans tous les autres cas, on note les bruits entre parenthèses, exemple (bruit *chchchch*).

- Exemple :
L1 : Ben, je veux dire euh, que, euh, je fais plus de ski maintenant.

3.2.7. Le discours rapporté

Le discours rapporté sera signalé par des guillemets simples (',') en début et en fin de discours.

[8] Rappelons que SAMPA est un système permettant de coder les symboles API standard à partir de symboles ASCII.

- *Exemple :*
L1 : Il m'a dit 'tu es complètement idiot d'avoir accepté ce boulot', et je crois bien qu'au fond il avait raison.

3.2.8. L'utilisation des parenthèses

Les syllabes incompréhensibles seront indiquées entre parenthèses et un 'X' correspond à une syllabe.

- *Exemple 1 :*
L1 : Paul s'en va et juste à ce moment, (XXX) et il tombe sur lui.

Les parenthèses sont aussi utilisées pour tout commentaire :

- *Exemple 2 :*
L1 : Paul s'en va et juste à ce moment, (bruit) il tombe sur lui.

3.2.9. Résumé

- Les seules marques de ponctuation dans le texte sont la virgule, le point et le point d'interrogation.
- Les chevrons signalent un chevauchement et commencent toujours par identifier le locuteur.
- Les guillemets simples signalent un discours rapporté.
- Les parenthèses ont pour fonction d'incorporer des commentaires.

4. Avantages d'une transcription orthographique standard (TOS)

Dans l'introduction de cette section, nous avons souligné les inconvénients d'une transcription phonémique ou phonétique si elle se donne comme principal point d'accès aux enregistrements. Au-delà des problèmes théoriques et pratique déjà soulevés, on remarquera qu'un des grands avantages d'une TOS est de permettre d'indexer rapidement toutes les occurrences d'une unité lexicale ou de séquences de mots données pour en examiner la prononciation. Si, pour rester concret, on s'intéresse de près aux prononciations de 'il y a', il est plus efficace de repérer tous les 'il y a' dans les transcriptions orthographiques et d'en examiner attentivement la prononciation que d'essayer de retrouver cette forme à partir de séquences phonétiques du type [ilia, ilja, ija, ja]. Cette tâche inverse est infiniment plus complexe puisqu'elle présuppose qu'on ait établi un inventaire fini des séquences phonétiques pertinentes et qu'on puisse les séparer d'autres fragments de transcription auxquelles elles peuvent être accolées. C'est précisément pour cette raison que nous ne modifions pas la transcription orthographique de départ avec des conventions pseudo-phonétiques comme les suivantes : *J' crois qu'y a Paul qu'est v'nu pa(r)c' qu'i peut la supporter*.

Dans PFC, nous respectons scrupuleusement l'orthographe standard. Nous écrivons, par exemple, 'je' que le locuteur ait dit [ʒə, ʒø, ʒœ, ʒ, ʒ, ʃ, etc.], sauf évidemment devant voyelle (*j'ai acheté*). Contrairement aux apparences, ramener toutes les variantes attestées à *j'*, comme on le fait souvent, revient à court-circuiter une analyse plus précise des données. Sans parler des ambiguïtés qu'on multiplie et du manque de cohérence dans l'adaptation des conventions de l'écrit aux méandres de l'oral, il devrait être clair que la diversité des réalisations phonétiques n'est pas directement répertoriables par des remaniements ad hoc de l'orthographe usuelle. De plus, les transcriptions orthographiques aménagées fournissent en général une image déformée de l'oral, le stigmatisent et le confinent dans un rôle de parent pauvre (voir §4.5).

5. Limites possibles de la transcription orthographique standard (TOS)

Aussi réels que soient les possibilités et avantages d'une TOS alignée sur le signal de la parole, il serait erroné de croire que celle-ci puisse répondre à toutes les situations que peut rencontrer le transcripteur. Il existe en effet différentes circonstances où une TOS ne permet pas de satisfaire pleinement aux critères de non-altération et de récupérabilité de l'information. Ces situations, que nous tenterons ici de circonscrire, ont quasiment toutes en commun de comporter des réalisations dont la transcription nécessite crucialement une référence explicite à l'oralité et à ce qu'on a pu appeler sa nature polyphonique.

Les lignes qui suivent ne sauraient prétendre résoudre les difficultés de transcription soulevées. Leur objectif sera, tout en précisant certaines de ces difficultés, d'envisager les problématiques sous-jacentes à chacun des choix susceptibles d'être envisagés par le transcripteur. Cinq types de situations seront ici examinés : la présence de formes linguistiques étrangères à la langue transcrite (§4.1), l'éventuelle nécessité de retranscrire la variation linguistique lorsque le discours porte précisément sur cette question de la « variation » (§4.2), l'éventualité d'ambiguïtés du signal phonétique (§4.3), la présence de mots ou d'expressions dont les réalisations s'écartent de l'orthographe standardisée (§4.4). Enfin nous aborderons brièvement la transcription de variétés linguistiques qui apparaissent si différentes du français « standard » que l'on s'interroge souvent sur la légitimité de la TOS (§4.5).

5.1. Irruption de formes linguistiques étrangères à la langue transcrite

De ce que le plurilinguisme est la situation la plus répandue à travers les langues du monde, il ressort que le discours d'un locuteur peut être couramment caractérisé par des productions relevant de langues différentes. Le cas des francophones n'échappe pas à cette règle, ce qu'attestent fort bien

les enquêtes du projet PFC. Ce type de production, bien que naturel pour tout locuteur maîtrisant plusieurs langues, n'en est pas moins délicat dès lors qu'il s'agit d'en établir la transcription. En effet, pour ce qui est ici de la TOS, si l'on considère que celle-ci fait a priori référence à une orthographe standard du français, qu'advient-il lorsque interviennent des réalisations « non françaises », réalisations qui échappent par conséquent au domaine de la graphie considérée?

Lorsque la langue étrangère possède elle-même une orthographe standard compatible avec celle de la langue française, la difficulté peut être assez aisément contournée. Les deux exemples figurant en (1) montrent deux locuteurs utilisant un terme « emprunté » au basque : « amatxi » pour 'grand-mère'.

(1) Basque

Locuteur 1: Amatxi, vous aviez un lavoir là-bas dans la maison euh, tu te rappelles?

Locuteur 2: Mais, Amatxi toi, tes, tes parents ils étaient commerçants.

Dans ces exemples, le transcripteur a choisi de transcrire le terme emprunté en utilisant l'orthographe basque habituelle, laquelle ne soulève pas ici de difficultés particulières. Notons que l'alignement de la transcription sur le signal sonore permet à l'utilisateur qui ne serait pas familier avec certaines des conventions de cette orthographe d'être informé sur la réalité phonético-phonologique du mot transcrit (ainsi, dans nos exemples, l'utilisation de la lettre « x » pour le son [ʃ]). Il est toutefois légitime de considérer la nécessité de fournir une explicitation plus détaillée. De fait, il est primordial de préciser le choix opéré par le transcripteur et ici, pour notre exemple, de préciser le choix d'utiliser l'orthographe standard basque. En outre, la réalité phonético-phonologique évoquée ci-dessus pourrait être dans certains cas de nature plus complexe, de sorte que la simple correspondance avec le signal audio pourrait parfois se révéler insuffisante pour l'utilisateur « profane ». De manière générale, toute précision déterminante dans la compréhension de la transcription devrait pouvoir être intégrée au sein d'un fichier de commentaires annexe prévu à cet effet. Il reste qu'il peut y avoir débat sur l'étendue de ces précisions. En effet, si l'explicitation du choix opéré s'avère de fait indispensable, certains ajouts de précisions phonético-phonologiques, aussi pertinentes soient ces dernières, ne devraient a priori pas être intégrés dès lors que ces informations ne présenteraient pas d'utilité directe pour la description et l'analyse du système considéré (ici la phonologie du français).

Les situations analogues à celle présentée en (1) et l'utilisation d'une autre orthographe elle même standardisée pourraient donc ne pas présenter de réelles difficultés, mais cela, à condition que la typographie de l'orthographe

considérée soit elle-même compatible avec celle retenue pour la TOS. Dans le cas d'interférences de langues comme l'arabe qui utilisent un système d'écriture différent de celui utilisé pour une langue comme le français, il devient plus difficile d'envisager la possibilité d'une « mixité » et il est alors impératif d'envisager un autre système afin de transcrire ces interférences. Se présentera alors la question du choix entre différents standards de translittération et de leur possible adéquation à une utilisation informatique. Ce recours à un autre système sera d'autant plus indispensable lorsque aucune orthographe standard, voire simple écriture, n'est disponible pour la transcription. Pour l'ensemble de ces situations, l'utilisation d'un alphabet phonétique (IPA ou SAMPA) apparaît alors indispensable[9]. Nous abordons à nouveau ces questions à la section 5.3.

5.2. Lorsque la discussion transcrite porte sur la variation linguistique

Lorsque le discours du ou des locuteurs porte lui-même sur les réalisations linguistiques ainsi que leurs variations (et est donc de type métalinguistique), une utilisation exclusive de la TOS peut dans certains cas s'avérer par trop réductrice et ne pas toujours permettre une explicitation intelligible des contenus transcrits.

Prenons l'exemple en (2) où le locuteur commente la manière de dire « épée ». Ce locuteur s'interroge ici sur le fait (imaginaire !) qu'il était peut-être attendu qu'une réalisation de type [ə] soit produite à la finale de la lecture du mot « épée ».

(2) Aude (Douzens)

Locuteur3: Alors là, c'est là qu'on réfléchit de dire il faut dire (epe@)[10], peut-être ou alors, ou alors parce qu'on, on a envie de faire mieux.

La simple utilisation de la TOS avec la transcription « épée » ne laisserait en aucun cas transparaître ce que cherche à exprimer le locuteur.

De la même manière, en (3)

9 La présence d'interférences de la langue arabe dans la transcription devrait être envisagée avec attention dès lors que l'on considère la présence non négligeable de cette langue dans le champ francophone.

10 La transcription phonétique a été établie en SAMPA ici ainsi que dans tous les exemples où nous illustrons une transcription qui serait intégrable sous PRAAT.

(3) Canada (Québec)

Locuteur4: ce que j'entends moi dans mon milieu c'est (s t^{s} i)

la transcription orthographique « ostie » ne permettrait en aucun cas de comprendre que le locuteur discute ici de l'aphérèse de ce mot dans son parler (mot qui a déjà été évoqué et « identifié » dans les énoncés qui précèdent).

Les occurrences occitanes qu'il est possible d'observer en (4) sont ici transcrites phonétiquement afin qu'elles restent fidèles à la variation des productions, variation qui est précisément l'objet de discussion des locuteurs en présence.

(4) Aude (Douzens)

E: nous, par exemple à Pézenas, pour dire 'il est fou', ma grand-mère disait (ɛskalyk). <Loc: (ɛskalyk)> (kalyk). <Loc: Nous, (ɛskaburt).> Et à Pézenas, on disait (ɛskalœk).

Certes, l'audition simultanée du signal aligné permet d'éclairer les propos de ces locuteurs, mais il peut sembler légitime que la transcription apporte elle-même un certain nombre d'informations de manière plus immédiate (*i.e.* à la simple lecture). Une fois encore, il peut y avoir débat autour de la richesse et de la granularité de ces informations. Dans le cas de (3) il pourrait toutefois être avancé que la simple transcription orthographique, bien que ne permettant pas de comprendre le propos précis du locuteur, permettrait en revanche de compléter très utilement la liste des unités susceptibles d'être utilisées, non seulement dans l'étude de l'affrication, mais aussi dans la recherche systématique d'autres aphérèses sur l'ensemble du corpus et d'en apprécier ainsi la variabilité[11]. La situation n'est par contre pas la même dans (4) puisque les réalisations variables appartiennent non pas au français mais à l'occitan, or ce dernier n'est pas la cible de l'enquête impliquée. Pourtant, dès lors que la réalisation considérée ne correspond plus à un mot identifiable orthographiquement, mais renvoie elle-même à une autre réalisation effective (ou plusieurs autres!) opérée par le locuteur lui-même, il n'est alors plus possible d'envisager une représentation de type orthographique[12]. Ainsi, en (5) :

[11] Voir également en ce sens ce qui est dit ci-dessous à propos des « réalisations 'non-standard' ».

[12] De telles situations sont en outre assez largement favorisées par la nature même de l'enquête ici menée par PFC.

(5) Aude (Douzens)

Locuteur3: moi j'ai dit deux fois [epe]

toute représentation orthographique (*i.e.* « épée » ou « épais ») à la place de la représentation phonétique [epe] est ici impossible dans la mesure où le locuteur fait, dans ce contexte, à la fois référence à une réalisation du mot « épais » et à une réalisation du mot « épée », mots qu'il prononce tous deux de manière identique et ce, conformément à son accent conservateur du sud-ouest[13].

5.3. Ambiguïtés et représentations orthographiques multiples

Dans le même temps, on comprendra aussi qu'une représentation orthographique est de nature à forcer une interprétation là où la seule écoute phonétique est de nature à en suggérer plusieurs. Aussi, particulièrement problématique s'avère l'exemple de (6) où dans ce chevauchement de parole, réunissant notre locuteur3 de Douzens et l'enquêtrice, le transcripteur a choisi de trancher pour une transcription orthographique « épée » :

(6) Aude (Douzens)

Enquêtrice: (epE) et nous on dit (epe) <Locuteur3: épée> un gâteau (epe) un matelas (epe)

Une telle transcription est en effet singulièrement déconcertante. Outre qu'il est ici difficile de comprendre pourquoi le transcripteur a choisi pour l'occurrence du locuteur 3 d'utiliser la TOS (« épée ») alors qu'il utilise les transcriptions phonétiques ([epE], [epe]) pour l'enquêtrice, il est encore plus troublant de constater qu'il a choisi la représentation orthographique qui semble précisément la moins correspondre à celle qui est pourtant clairement attendue. En effet, l'ensemble du signal aligné laisse aisément ici entendre que notre locuteur répète ce que dit l'enquêtrice à propos des différentes réalisations du mot « épais » et non qu'il énonce le mot pour lui homophone « épée ».

Vu le contexte, notre solution serait :

Enquêtrice: épais (epE) et nous on dit (epe) <Locuteur3: (epe)> un gâteau épais (epe) un matelas épais (epe)

Rappelons cependant que le principe est toutefois d'éviter autant que possible toute surcharge métalinguistique.

[13] Ce renvoi « polysémique » opéré dans la réalisation [epe] n'est bien entendu compréhensible ici que par le contexte et l'ensemble de l'échange discursif.

Dès lors, nous voyons concrètement se dégager ici deux types de difficultés que le transcripteur doit surmonter :

– maîtriser et conserver la cohérence des choix opérés,
– identifier de manière correcte, dans les situations d'ambiguïté, l'énoncé effectivement réalisé.

Dans un certain nombre de cas, il n'est pas rare que l'identification en question soit extrêmement délicate, voire impossible. Les cas d'occurrences comme [ɔ̃nɛpa] sont à cet égard particulièrement exemplaires. Doit-on noter ici la réalisation d'une double négation (*i.e.* « ne…pas ») et donc une transcription « on n'est pas » ? ou bien doit-on considérer que le locuteur n'a pas réalisé de « ne », que la consonne nasale ne correspond ici qu'à une liaison enchaînée et que, par conséquent, la transcription sera « on est pas » ?

Dans ce type de situation, lorsque aucun indice fiable ne permet de trancher, la prudence pourrait conduire à opter pour l'option qui stigmatisera le moins le locuteur et donc, dans le cas de notre exemple, pour la représentation « on n'est pas ». Lorsqu'un tel choix arbitraire est opéré, il est crucial qu'il soit observé de manière uniforme dans l'ensemble de la transcription et expliqué dans le fichier de commentaires de la transcription.

5.4. Les réalisations « non standard »

Il arrive que l'on rencontre dans certains corpus des mots dont les réalisations sont sensiblement éloignées de ce que peut évoquer leur représentation sous une forme orthographique standard. De telles « divergences » ont d'ailleurs pu parfois conduire les usagers à adopter pour de telles réalisations une nouvelle orthographe a priori « non standard » ou, disons, « moins standard » que celle habituellement reconnue comme « légitime ».

Ainsi, dans le parler français du Québec, le terme « tabernacle » est fréquemment utilisé comme juron[14] et ce particulièrement sous les formes [tabarnak] et [barnak]. En outre, il semble qu'une orthographe « tabarnak » puisse être utilisée. Confronté à ces réalisations, le transcripteur voit s'offrir à lui plusieurs possibilités :

a) Rester fidèle à l'orthographe standard « tabernacle »,
b) Rester fidèle à la réalisation phonétique,
c) Utiliser l'orthographe « particulière ».

Le choix n'est guère aisé car chaque option souligne une dimension spécifique au détriment des autres. En aucune façon elles ne sont neutres pour l'analyse et chacune infère une problématique d'étude différente. Alors que c) et b) insistent sur une certaine particularité, voire autonomie (par

[14] Un « sacre » dans la terminologie canadienne.

exemple une autonomie de ces réalisations au regard de la réalisation standard), a) met en revanche l'accent sur une communauté d'appartenance pour l'ensemble de ces réalisations différentes (leurs différences pouvant être alors éventuellement expliquées par des phénomènes généraux propre au système étudié, par exemple ici la simplification des groupes consonantiques finaux). Il est clair que telle ou telle analyse est ici par trop tranchée et peut être prématurée pour l'enquête en question (« lexicalisation » ou « phonologisation » hâtives). Ce qu'il est important de souligner ici, c'est la nécessaire lucidité avec laquelle le choix devra être opéré, et la vigilance qui devra être observée au regard des analyses préconçues que ce choix pourrait implicitement véhiculer.

5.5. Limites du système ?

A plusieurs reprises au cours des paragraphes précédents, le lecteur aura constaté que la question de la circonscription et de l'interaction des langues se posait dans un projet comme PFC. Pour nombre de locuteurs dans l'espace hexagonal, la question de l'identité de la langue qu'ils parlent dans nos enquêtes ne se pose pas. Ils considèrent qu'ils parlent français (même s'ils déclarent le « massacrer ») et la TOS, même si elle peut soulever de réelles difficultés ici et là, s'avère efficace. Des difficultés plus grandes surgissent dès qu'on élargit le champ d'observation. Il se peut que les réponses données jusqu'ici suffisent à traiter ces difficultés, il nous paraît néanmoins nécessaire d'examiner de façon plus concrète divers exemples rencontrés au sein du projet, même si cela entraîne des répétitions éventuelles. Il existe tout d'abord de nombreuses communautés bilingues où le « code-switching » (alternance codique) est monnaie courante. Pour certaines enquêtes au Canada, par exemple, il devient nécessaire d'adopter une transcription qui autorise l'utilisation systématique de la graphie anglaise à côté de l'orthographe française tant la mixité des deux systèmes est grande. Walker (2003) note des exemples comme (a) et (b) tirés de l'enquête PFC en Alberta :

(a) yeah c'est comme OK mais ça devient boring.
(b) OK. Ben. Une fois à l'école j'ai assis sur une chaise pis ça a brisé. And everybody laughed so I was totally embarrassed.

Tout aussi épineux est le cas de communautés parlant des variétés qu'elles identifient comme n'étant pas du français mais qui partagent de nombreux traits avec le français à tous les niveaux d'analyse (par exemple des créoles ou des langues régionales d'oïl). Nous ne traiterons pas cette question en détail ici mais nous illustrerons un fort écart par rapport à la norme hexagonale à partir d'un exemple concret emprunté aux usages acadiens. Nous avons précisément choisi de partir d'une transcription orthographique aménagée, celle qui est adoptée dans *La Sagouine* d'Antonine Maillet. Antonine Maillet s'est illustrée en recevant plusieurs récompenses dont le

prestigieux prix Goncourt en 1979 pour Pélagie-la-Charrette. Dans *La Sagouine*, le lecteur est confrontée au monologue d'une illettrée de 72 ans qui à travers sa vie de laveuse de plancher explore la réalité qui l'entoure. L'auteur utilise avec succès dans ce roman une transcription qui conserve l'orthographe standard comme toile de fond mais qui la modifie systématiquement pour toute une partie du lexique. Nous prendrons comme exemple le premier paragraphe de cette œuvre :

> « J'ai peut-être ben la face nouère pis la peau craquée, ben j'ai les mains blanches. Monsieur ! J'ai les mains blanches parce que j'ai eu les mains dans l'eau toute ma vie. J'ai passé ma vie à forbir. Je suis pas moins guénillouse pour ça... J'ai forbi sus les autres. Je pouvons ben passer pour crasseux : je passons notre vie à décrasser les autres. Frotte, pis gratte, pis décolle des tchas d'encens... ils pouvont ben aouère leux maisons propres. Nous autres, parsoune s'en vient frotter chus nous. »

Il nous semble raisonnable de considérer ce passage comme un exemple intéressant de transcription orthographique aménagée (TOA) et nous notons d'ailleurs que de nombreux projets sur de telles variétés adoptent des conventions similaires. Supposons maintenant qu'une enquête PFC nous livre une intervention de témoin correspondant à ce passage. Un examen rapide de la transcription d'Antonine Maillet pourrait suggérer qu'il faut inévitablement recourir à une orthographe modifiée si l'on veut serrer cette variété acadienne de près. Nous ne croyons pas cette conclusion nécessaire pour les raisons suivantes.

Considérons tout d'abord les exemples suivants, 'aouère' pour 'avoir', 'parsoune' pour 'personne', 'chus' pour 'chez', 'ben' pour 'bien', 'pis' pour 'puis'. Même au sein des usages hexagonaux on rencontre des écarts aussi grands entre variantes régionales et français de référence. Ainsi, la prononciation méridionale [entak] pour *intact* semble aussi loin de la norme [ɛ̃takt] que 'parsoune' pour 'personne'. Pourtant les transcripteurs de ces variétés méridionales n'éprouvent aucune gêne dans le recours à la TOS 'intact'. Si la forme 'aouère' est censée dénoter une prononciation du type [awɛr], elle illustre un écart par rapport à la norme /avwar/ qui est loin d'être isolé. On entend de nombreux locuteurs hexagonaux qui effacent le /v/ de mots comme *voir* et qui réalisent la voyelle <a> avec un [æ] ou un [ɛ] devant le phonème /r/. La transcription 'pis' pour 'puis' est également discutable. Il faut établir si 'puis' est utilisé en parallèle et, si la réponse est positive, s'assurer que ces deux formes ne sont pas en distribution complémentaire.

Les formes '(je) pouvons', '(je) passons', '(ils) pouvont' peuvent susciter des interrogations. Les formes 'pouvons' et 'passons' sont attestées en français

écrit standard, bien que normalement régies par 'nous', et seraient donc employées sans problème dans nos transcriptions. En revanche 'pouvont' n'appartient pas à l'orthographe standard. Cependant, cette forme se compose de deux allomorphes parfaitement attestés qui sont phonologiquement éloignés de peuvent et qui s'intègrent à un système de conjugaison qui n'est pas nécessairement isomorphe au système standard. Pour toutes ces raisons, il serait anormal d'adopter la forme 'peuvent' pour ce type d'exemple.

On constate aussi dans ce passage qu'il y a des lexèmes ('forbir', 'guenilloux') n'appartenant pas au français hexagonal standard. Cette difficulté n'est pas propre à cette variété et dans de nombreux cas des dictionnaires existent qui prêtent renfort au transcripteur. Le Canada est particulièrement riche en dictionnaires représentant la diversité des usages et où de tels mots sont répertoriés ; mais ce n'est pas le seul pays dans ce cas. Béatrice Akissi-Boutin (en préparation) note que pour la Côte d'Ivoire les difficultés lexicales qu'on rencontre sont le plus souvent réglées à partir de sources lexicographiques comme l'*Inventaire des particularités lexicales de français en Afrique noire* ou encore Lafage (2003, 2004). Puisque les conventions qu'adoptent ces dictionnaires s'appuient sur les graphèmes du français on peut dans certains cas utiliser des translittérations. Ainsi, alors que le plat cuisiné 'djoumblé' est attesté dans Lafage (2004), on n'y trouve pas le plat [k͡ple] (avec double articulation du [k] et du [p]) attesté dans l'enquête PFC en Côte d'Ivoire. On pourrait également proposer la translittération 'kplé' avec la prononciation en sampa placée entre parenthèses (kple). Nous pensons néanmoins qu'il faut éviter l'emploi simultané d'une translittération et de parenthèses fournissant la prononciation. Le danger est de créer des transcriptions fourmillant d'informations entre parenthèses qui constituent des pré-analyses massives des données. Il faut également se souvenir que nous visons à des traitements automatiques des données et qu'il faut minimiser les informations qu'on sera obligé d'éliminer dans cette perspective.

Nous ne voulons pas suggérer ici que la TOS de certaines variétés éloignées de la norme hexagonale ne pose pas de difficultés mais nous pensons qu'elle est tout à fait utilisable dans la plupart des enquêtes recensées à ce jour. En fait, comme nous le fait remarquer François Poiré, la principale difficulté ne surgit pas nécessairement au moment de la transcription orthographique initiale mais dans les codages ultérieurs qu'on peut vouloir opérer. Si l'on transcrivait la forme fréquente au Canada 'icit' par 'ici' alors on s'interdirait par là-même de coder la consonne finale pour un schwa potentiel et on s'écarte donc des codages systématiquement adoptés dans PFC pour l'étude de la réalisation des consonnes finales. On notera par ailleurs que la forme 'icit' est parfaitement recensée au Canada et que, sans la présence de la

consonne finale, on ne pourrait pas non plus comprendre la réalisation de la seconde voyelle. Une standardisation sur le modèle du français hexagonal a des limites ! Le travail doit donc se poursuivre pour répertorier les cas les plus problématiques et converger vers des solutions.

6. Conclusion

Dans cet article, nous avons examiné la transcription orthographique standard (TOS) adoptée dans PFC comme point de départ essentiel des analyses. Il faut en effet rappeler que l'orthographe française, dans la mesure où elle est en partie morphophonologique, est une voie d'accès utile à de nombreuses alternances de surface du français. Il ne s'agit pas de s'engager ici dans le débat qui a pu opposer les linguistes qui ont considéré l'orthographe du français comme relativement motivée dans son rapport à la phonologie (Schane, 1968, Dell 1973) et ceux qui au contraire l'ont vu comme fondamentalement arbitraire (M. Durand 1936, Blanche-Benveniste & Chervel 1969). Nous affirmons sans le démontrer ici qu'une annotation de l'orthographe standard est utile pour repérer et explorer divers phénomènes phonologiques en français. C'est en tout cas notre thèse en ce qui concerne le schwa et la liaison pour lesquels des codages explicites ont été établis[15]. Nous avons présenté plusieurs arguments pour écarter soit une transcription 'phonétique', soit une transcription orthographique aménagée (TOA) en dehors de cas très restreints. Les TOA ont de nombreux défauts. Il est tout d'abord difficile de garantir la cohérence des solutions comme on le constate en lisant des romanciers dont les œuvres se veulent fidèles à l'oralité. En particulier, on s'aperçoit que les TOA gomment souvent la variation en généralisant abusivement des formes non standard. Ce faisant on finit par marginaliser l'oral et le stigmatiser. Deuxièmement, en adoptant des translittérations de type phonémique on s'interdit de saisir les morphèmes dans la diversité de leurs réalisations allomorphiques et donc de permettre des recherches systématiques et rapides sur les variantes effectives d'un morphème donné. Enfin, il faut se souvenir qu'une TOS est un point de départ idéal pour l'utilisateur de phonétiseurs / phonémiseurs dans des études qui visent à des alignements fins entre le signal et son codage. Cette condition n'est évidemment pas remplie par une TOA. Cela dit, un certain nombre de problèmes demeurent. Nous avons commencé à les recenser et à proposer des solutions partielles. La voie est en partie ouverte pour des recherches plus approfondies.

[15] Voir références.

Références bibliographiques

Akissi-Boutin, B. (en préparation) Réflexions pratiques, théoriques et éthiques à propos de l'encodage du français de Côte d'Ivoire dans le cadre de PFC.

Bilger, M. (2000), *Petite typologie des conventions de transcription de l'oral : aspects pratiques et théoriques.* Manuscrit non publié. Université de Perpignan.

Blanche-Benveniste, C., Bilger, M., Rouget, Ch. & Van den Eynde, K. (1991), *Le français parlé : Études grammaticales*, Paris, Éditions du Centre National de la Recherche Scientifique.

Blanche-Benveniste, C. & Chervel A. (1969), *L'orthographe*, Paris, Maspero.

Blanche-Benveniste, C. & Jeanjean C. (1987), *Le français parlé : transcription et édition.* Paris, Didier Erudition.

Delais-Roussarie, E. (2003a), « Constitution et annotation de corpus : méthodes et recommandations », *in* E. Delais-Roussarie & J. Durand (éd.), *Corpus et variation en phonologie du français : méthodes et analyses*, Toulouse, Presses Universitaires du Mirail, pp. 89-125.

Delais-Roussarie, E. (2003b), « Quelques outils d'aide à la transcription et à l'annotation de données audio pour constituer des corpus oraux », *in* E. Delais-Roussarie & J. Durand (éd.), *Corpus et variation en phonologie du français : méthodes et analyses*, Toulouse, Presses Universitaires du Mirail, pp. 127-157.

Delais-Roussarie, E. & J. Durand (éd.) (2003), *Corpus et variation en phonologie du français : méthodes et analyses*, Toulouse, Presses Universitaires du Mirail.

Delais-Roussarie, E., Meqqori, A. & Tarrier, J.-M. (2003), « Annoter et segmenter des données de parole sous PRAAT », *in* E. Delais-Roussarie & J. Durand (éd.), *Corpus et variation en phonologie du français : méthodes et analyses*, Toulouse, Presses Universitaires du Mirail, pp. 159-185.

Delais-Roussarie, E., Durand, J., Lyche, C., Meqqori, A. & Tarrier, J.-M., (2002), « Transcriptions des données : outils et conventions », *Bulletin PFC* 1, pp. 21-34. CNRS ERSS-UMR5610 et Université de Toulouse-Le Mirail.

Durand, J. (sous presse), « Mapping French Pronunciation. The PFC project », in J.-P. Montreuil (ed.) (à paraître 2006), *New Analyses in Romance Linguistics*, Current Issues in inguistic Theory, Amsterdam, John Benjamins.

Durand, J. & Lyche, C. (2003), « Le projet 'Phonologie du Français Contemporain' (PFC) et sa méthodologie », *in* E. Delais-Roussarie & J.

Durand (éd.), *Corpus et variation en phonologie du français : méthodes et analyses*, Toulouse, Presses Universitaires du Mirail, pp. 213-276.

Durand, J., Laks, B. & Lyche, C. (2002), « (PFC) Protocole d'enquête », *Bulletin PFC* 1, pp. 7-20. CNRS ERSS-UMR5610 et Université de Toulouse-Le Mirail.

Durand, J., Laks, B. & Lyche, C. (2005), « PFC : Un corpus numérisé pour la phonologie du français », *in* G. Williams (éd.), *Les linguistiques de corpus*, Rennes, Presses Universitaires de Rennes, pp. 205-217.

EAGLES (1996a). *Recommendations on corpus encoding. EAG-TCWG-CES/R-F*. Pisa, Consiglio Nazionale delle Ricerche, Istituto di Linguistica Computazionale.

EAGLES (1996b). *Preliminary recommendations on spoken texts. EAG-TCWG-SPT/P*. Pisa, Consiglio Nazionale delle Ricerche, Istituto di Linguistica Computazionale.

Poiré, F. (en préparation), *Transcription et orthographe standard en français canadien*.

Prignitz, G. (en préparation), *Observations sur un corpus récent recueilli à Ouagadougou.*

Sampson, G. (2005), « Quantifying the shift towards empirical methods », *International Journal of Corpus Linguistics* 10(1), pp. 15-36.

Tarrier, J.-M. (2003), « L'enregistrement et la prise de sons », *in* E. Delais-Roussarie & J. Durand (éd.), *Corpus et variation en phonologie du français : méthodes et analyses*, Toulouse, Presses Universitaires du Mirail, pp. 187-212.