Conventions de transcription

Toute transcription est un compromis forcément boiteux entre le respect des particularités orales et la lisibilité. Afin de faciliter la lecture[1] et de simplifier les traitements automatisés du corpus, nous avons adopté le code orthographique avec quelques aménagements, suivant en cela les pratiques initiées par les chercheurs québécois, par ESLO, par le DELIC et VALIBEL[2]. L'alignement sur le son permettant de toute façon de revenir facilement à la version orale, nous avons décidé de ne pas essayer de rendre compte de l'intonation ou des variantes phonétiques.

1 Identification des corpus et des locuteurs

Chaque fichier est identifié au moyen d'un code individuel (nom du corpus CFPP2000 suivi du numéro de l'arrondissement ou des initiales de la banlieue où a été réalisé l'enregistrement, pseudonyme du locuteur principal et d'un chiffre correspondant à l'ordre d'enregistrement dans la base de donnée ; sexe (F ou M), âge, numéro de l'arrondissement. Le pseudonyme est choisi autant que faire se peut pour évoquer les caractéristiques du nom de départ (origine géographique, notoriété générationnelle des prénoms, etc. Dans l'identifiant `CFPP2000 [03-01] Ozgur_Kilic_H_32_alii_3e`, le locuteur principal est, par convention Özgur Kilic et ce pseudonyme évoque son origine turque ; le prénom Kilian, d'origine celtique est un prénom que des Français commencent à donner à leurs enfants à partir des années 90. Les autres locuteurs reçoivent aussi des pseudonymes qui permettent de les identifier et qui sont récapitulés dans la fiche descriptive, par exemple Michel Chevrier.

2 Une transcription orthographique avec quelques aménagements

Nous transcrivons les mots en orthographe sans correction des écarts à la norme qui correspondent à un morphème attesté en français. Ainsi, bien que ce soit une femme qui s'exprime, nous écrivons *mis* et non *mise* lorsqu'à l'oral la locutrice n'a pas réalisé l'accord du participe :

1 c'est pas pour des raisons euh catholiques qu'ils m'ont **mis** dans une école de bonnes sœurs
2 CFPP2000 [11-01] Anita Musso-F-46-11e

[1] Il s'agit de ne pas rebuter des utilisateurs non linguistes, mais aussi de permettre une lecture cursive à des linguistes. On sait que même des spécialistes ont une lecture extrêmement ralentie lorsque les conventions s'écartent trop de leurs habitudes.

[2] http://sites.univ-provence.fr/delic/corpus/conventions.html; http://www.uclouvain.be/81836.html

En effet, *mis* est un morphème du français. Selon le même principe, nous pouvons « respecter » un futur comme « cousera » (au lieu de *coudra*) qui se laisse ramener à une succession de morphèmes français : radical du verbe + *-er-*, morphème du futur présente pour de nombreux verbes + morphème de troisième personne, *-a*.

En revanche, les variantes de prononciation qui n'ont pas de correspondance orthographique reçue dans les dictionnaires ne sont pas notées : nous n'écrivons pas « f'nêt », « f'nêtr », pour *fenêtre*, ce qui rendrait très difficile les recherches assistées par l'ordinateur dans les corpus. Nous pouvons cependant noter la variation entre « oui » et « ouais », puisque « ouais » est un morphème recensé par les dictionnaires.

Nous ne notons pas non plus les allongements de syllabes, ou les *e* qui apparaissent en fin de mot (la frontière avec des « euh » d'hésitation ou de remplissage est incertaine).

Nous avons cependant – et c'est un point de divergence avec d'autres projets – élidé les clitiques quand cette élision s'entendait à l'oral. Nous écrivons alors : *j'* devant consonne, *t'* devant voyelle *: j'sais ; t'arrives.* Ces graphies se sont bien répandues, gagnant peu à peu la bande dessinée, le roman, les blogs. Les adopter amène à produire une description plus proche de la réalité morphologique. Il est faux par exemple que la forme sujet et la forme objet direct du pronom de deuxième personne soient en distribution complémentaire, puisque *tu* et *te* sont parfois neutralisés sous la forme *t'*. Cette décision ne devrait pas poser de problèmes puisque les morphèmes « t', j', qu', c', s', d', n' » existent et que la liste en est fermée :

> 3 Lies : en fait *tu t'rends* compte que finalement tous les vendeurs […] parlent arabe et donc en plus avec leur client *t'as* l'impresssion qu- qu'ils sont vachement potes et toi *t'arrives* bon j'suis sûre qu'on paie plus que les autres
> 4 CFPP2000 [18-01] Paul Simo… Lees, F-30 néerlandaise

En revanche nous nous refusons à « i » ou « iz » pour *ils* qui s'écartent des habitudes orthographiques du français.

Nous avons respecté les apocopes qui constituent selon nous des néologismes (cela conduira à revoir les règles de lemmatisation des concordanciers et à poser des équivalences entre : « appart- ; prof ; aspi-, etc. ; » et « appartement, professeur, aspirateur, etc. »).

Les emprunts sont autant que possible transcrits selon l'orthographe usuelle dans leur langue d'origine. Lorsque le mot est inconnu, ou qu'il n'a pas été répertorié par les dictionnaires parce qu'il est non standard, il est transcrit selon une graphie qui se rapproche des séquences usuelles du français. Pour les mots dits « des

banlieues », nous avons eu recours à des dictionnaires comme Le Dictionnaire de la Zone[3].

Nous employons des majuscules à l'initiale des noms propres ; cependant nous avons renoncé à les isoler par des signes spéciaux, leur délimitation posant des problèmes qui nécessitent une étude spéciale

Les passages incompréhensibles sont représentés par un (pour une syllabe) ou plusieurs X majuscules. Par convention, nous limitons, comme le font Delic et Valibel, nous nous limitons à 3 X.

Nous suivrons les conventions de Valibel pour opposer les sigles transcrits en capitales sans points : *SNCF* et les acronymes transcrits par une majuscule au début du mot, le reste restant en bas de casse comme un nom propre ordinaire : *Fnac*.

3 Amorce, multi-transcription et alternances orthographiques

L'amorce d'un mot est notée par un tiret accolé au mot :

5 un mi-

Pour faciliter la lecture, nous avons sévèrement restreint la notation des alternances de transcription. Nous nous contentons de noter l'interprétation la plus probable. Lorsqu'une multi transcription apparaît, nous suivons les conventions de DELIC en mettant entre parenthèses les deux possibilités séparées par une virgule :

6 (d'accord, d'abord).

Les alternances orthographiques n'ont pas été notées systématiquement. Lorsqu'elles le sont, c'est entre-parenthèses, conformément à l'usage du DELIC :

7 on (n') a pas

4 Liaisons

Les liaisons fautives ont été indiquées par un « z » ou un « t » entre tirets. Ce principe s'étend à des cas où le statut du « z » n'est pas très clair et où il pourrait s'apparenter à un morphème flottant ou à une variante morphologique :

8 Mille-z-assiettes

[3] http://www.dictionnairedelazone.fr/

9 donne moi-z-en
10 des chefs de gare-z-en retraite

5 Pauses et ponctuation

Nous n'utilisons ni les points ni les virgules. Il aurait peut-être été intéressant d'indiquer par une ponctuation forte ou demi-forte toutes les segmentations en unités de discours que le transcripteur perçoit, – que son impression soit due à une pause, par un allongement vocalique ou par un contour intonatif descendant. C. Blanche Benveniste (2007) est d'ailleurs revenue sur la proscription qu'elle avait largement contribué à établir. Nous avons finalement maintenu pour cette première tranche les conventions suivantes :

+	pause brève ;
++	pause longue ;
///	interruption du discours.

Le point d'interrogation, le point d'exclamation et les guillemets ont été utilisés lorsque le transcripteur entendait nettement l'intonation :

?	interrogations avec montée de la voix
!	Exclamation
« »	Les décrochages liés au discours direct sont signalés. Les transcripteurs ont décidé de ne pas noter la frontière droite lorsqu'ils hésitaient.

6 Chevauchements et tours de parole

Le système Transcriber est mal adapté à la notation des chevauchements, en particulier il ne permet pas de noter le cas où plus de deux personnes prennent la parole simultanément. Ces chevauchements multiples sont rares dans notre base. Ils ne sont pas pris en compte. Par ailleurs, nous avons décidé pour des raisons d'exploitation informatique de ne pas couper les mots, même lorsqu'une partie seulement est concernée par le chevauchement. C'est pourquoi, nous n'avons pas ajouté de symboles permettant de délimiter le début et la fin des paroles prononcées. Ces indications resteraient de toute façon approximatives. Un retour au son s'impose donc si l'on veut travailler sur les interactions.

Nous n'avons pas interrompu le tour par un retour à la ligne, lorsque le locuteur principal poursuit sa prise de parole et qu'un second locuteur intervient en arrière-plan en se bornant à des « mm » approbatifs ou à des interjections (hum)

sans interrompre le tour : nous faisons figurer entre parenthèses ces régulateurs. Ces interventions n'ont pas été notées dans les transcriptions réalisées jusqu'en avril 2009. (Les « mm » du locuteur principal ne sont pas mis entre parenthèses).

Nous notons également entre crochets des bruits tels que les rires ou la toux. Dans Transcriber, ils apparaissent sous cette forme :

11 j'l'ai fait non j'ai pas arrêté ils m'ont renvoyé j'vais vous dire comment ils ont fait [rire]

Dans la version en ligne via Real Player, ils apparaissent avec le codage qui permet leur non prise en compte par les outils informatiques, concordanciers, étiqueteurs, etc. :

12
13 j'l'ai fait non j'ai pas arrêté ils m'ont renvoyé j'vais vous dire comment ils ont fait[rire|noise|instantaneous]
14 CFPP2000 [SO-02] Youcef-Zerari, 29 ans