Erreurs d'écoute dans la transcription de données orales

Berthille Pallaud

Berthille Pallaud

Université de Provence Laboratoire Parole et Langage (UMR 6057)

Erreurs d'écoute dans la transcription de données orales

Mots-clefs : erreurs d'écoute, transcription, corpus de français parlé

1. INTRODUCTION

Affirmer, à partir d'énoncés enregistrés, la présence de dysfonctionnements langagiers soulève de nombreuses difficultés et ne peut se faire qu'avec une extrême prudence. En effet, les textes écrits ne peuvent constituer des éléments témoins fiables. De nombreux travaux du Groupe Aixois de Recherche en Syntaxe ont décrit les spécificités des énoncés oraux. Ces derniers semblent constituer les corpus deréférence lorsqu'un énoncé oral en vient à être soupçonné de variation. Les textes oraux présentent des caractéristiques, véritables marques de l'oral, qui sont depuis mieux connues. On a montré en particulier pour les enfants (Blanche-Benveniste & Pallaud, 2001; Pallaud & Savelli, 2001), combien il était important de tenir compte des descriptions de ces marques de l'oralité afin de ne pas attribuer à tort, à ces enfants des défauts d'acquisition ou des dysfonctionnements langagiers.

L'établissement du texte transcrit sur la base de l'énoncé oral requiert des conventions et des précautions à différentes étapes. Outre les difficultés d'écoute qui peuvent avoir lieu en dépit de la qualité de l'enregistrement, les problèmes de transcription et d'interprétation ne sont pas rares (Giovannoni & Savelli, 1990; Pallaud, 2001). Même si les caractéristiques de la transcription sont précisées (orthographiques, prosodiques, de prononciation), les hésitations interprétatives en dépit du contexte ne se produisent pas à n'importe quel moment de l'énoncé. "Les écarts entre la version primitive et la version finale ne se font pas de façon anarchique" (Cappeau, 1997, p. 119). En fait la transcription recèle parfois, en ces points contestés, l'indice d'une deuxième voix qui s'introduit dans l 'énoncé du locuteur; celle du transcripteur. C'est ce que révèle le travail de vérification des documents linguistiques obtenus même après une transcription soigneuse. Le texte obtenu contient en quelques endroits des erreurs d'écoute qui, n'étant pas la représentation écrite de ce que le locuteur a dit, sont la transcription de ce que le transcripteur a cru percevoir. En ces endroits, le texte n'appartient donc pas au locuteur mais bien au transcripteur. En ce sens, on peut dire que le discours du locuteur se trouve "ajouré" par un autre discours, celui du transcripteur.

Le propos de cette étude est de recenser les erreurs d'écoute commises lors de la transcription d'énoncés de français oral. L'écoute réfère à la complexité de l'activité de perception d'un énoncé oral quand s'élabore le texte transcrit. Nous verrons que les erreurs que font les transcripteurs obéissent à certaines règles et dépendent du secteur auquel elles se rattachent. L'objectif ici est de fournir des données quantitatives, ce qui vient poursuivre les réflexions déjà conduites dans ce domaine par l'équipe du GARS à Aix (Bilger, Blasco, Cappeau, Pallaud, Sabio & Savelli, 1996; Cappeau, 1997; Giovannoni & Savelli, 1990; Blanche-Benveniste, & Jeanjean, 1987).

2. LE MATÉRIEL ANALYSÉ

Notre étude a retenu douze corpus de français parlé enregistrés en septembre 1999 dans le cadre du Projet de Corpus de Référence du Français Parlé[1] conduit à Aix par l'équipe du CNRS ESA 6060 que dirigeait Claire Blanche-Benveniste. Trois villes du Nord-Est (Belfort, Nancy et Strasbourg), une ville du Centre (Troyes) et une ville du Sud-Ouest de la France (Bordeaux) avaient été, à cet égard, sélectionnées (cf. tableau 1).

Ville des locuteurs	Transcripteurs	Nombre de mots
Nancy 1	N. R.	4537
Nancy 2	S. H.	4901
Strasbourg 1	L. R.	3256
Strasbourg 2	L. V.	3555
Strasbourg 3	N. R.	4578
Belfort 1	L. V.	3957
Belfort 2	I. A.	3808
Belfort 3	L. R.	3309
Bordeaux 1	S. H.	4280
Troyes 1	C. V.	3862
Troyes 2	C. V.	3385
Troyes 3	C. V.	3221

Tableau 1: présentation des douze corpus retenus pour cette étude.

L'ensemble du Projet prévoyait de réunir des énoncés oraux (d'une durée, chacun de vingt à trente minutes) recueillis dans une quarantaine de villes françaises réparties sur toute la France (132 enregistrements). Les situations de parole étaient échantillonnées sur chacune des villes: conversation, situation de parole publique, énoncé professionnels. L'échantillonnage portait également sur

[1] Le rapport sur ce Projet est prévu pour publication dans le numéro 18 de la revue *Recherche Sur le Français Parlé* en 2003.

l'âge des locuteurs (trois catégories d'âge: 18 à 30, 30 à 65 ans; plus de 65 ans) et sur leur niveau de scolarité (trois catégories: niveau collège et professionnel, niveau Bac; niveau Bac + 3 et plus). Ces corpus vont donc permettre des comparaisons essentielles entre les mécanismes utilisés oralement par les locuteurs et ceux qui sont développés, par exemple, dans les productions écrites. La division en sous-corpus permet d'apercevoir les répartitions qui sont faites selon les situations, les âges et les niveaux de scolarisation des locuteurs.

Au total, le nombre de mots dans l'ensemble des douze corpus retenus est de 47 000 soit 4 heures d'enregistrement si on se base sur un débit moyen de parole fixé à 200 mots par minute. La moyenne du nombre de mots par corpus est de 3900 mots avec comme limites inférieure et supérieure: 3221 et 4901 mots.

3. LA TRANSCRIPTION DE CES DOUZE CORPUS

À l'exception des trois corpus de Troyes, tous ces corpus ont été recueillis par mes soins et ont été transcrits par six étudiants du Département de Linguistique française d'Aix (niveau maîtrise et DEA). Les trois corpus de Troyes ont, eux, été recueillis et transcrits par une étudiante en thèse de sociolinguistique à Paris-Nanterre.

Les conventions du GARS ont été appliquées pour la totalité de ces enregistrements. Ces conventions prévoient que la transcription se fait en orthographe standard, sans "trucage" orthographique pour rendre compte de faits de prononciation. Les dictionnaires servent de référence pour les mots de la langue, les noms propres, les interjections et les onomatopées. Elles ne comportent pas non plus de ponctuation. L'évaluation objective des pauses dans le cours de l'énoncé nécessiterait une approche acoustique que les équipements disponibles dans l'équipe ne permettent pas. Elles ne sont donc évaluées que subjectivement par le transcripteur.

pause brève - *des clients qui sont particulièrement - désagréables*
pause longue - - *c'était une - - une assurance*

Les répétitions comme les amorces de mots sont relevées soigneusement.
amorce de mot *il faut les* rem- *les remplacer.*

Si les passages incompréhensibles sont notés par un artifice typographique (X ou XXX),
syllabe inaudible: X *c'est X c'est là*
suite de syllabes inaudibles: XXX *et elle XXX encore un petit peu plus loin.*

Les séquences dans lesquelles on ne peut identifier des morphèmes sont transcrites phonétiquement
il a acquis [akeri] la voiture.

Lorsque le transcripteur hésite entre plusieurs solutions pour un même énoncé, il propose une multi-transcription qui rend compte de son hésitation.

multi-transcription: / , / *elle /a acheté, achetait/ des meubles*
il y a /des, les/ clients désagréables
bien euh /je vais, il faut/ pas regarder la différence

quelque chose ou zéro: / , Ø/ *et /alors, Ø / dans ce cas on s'en débarrasse*

Il peut y avoir également des hésitations orthographiques:
on (n')en parle plus
il(s) travaille(nt)
il est venu /à accepter, a accepté/ ce qu'on lui offrait (Lafage, 95).

4. LA VÉRIFICATION DES CORPUS

Lorsque les corpus ont été transcrits, leur transcription est vérifiée par une autre personne que le transcripteur[2]. Cette vérification comporte deux étapes dont seule la première concerne cette étude. Dans un premier temps, il s'agit en effet d'écouter à nouveau l'enregistrement et de proposer ou non des modifications. La deuxième étape concerne la vérification de l'édition du texte c'est-à-dire le respect des règles typographiques et de la mise en page. Elle ne sera pas abordée ici.

Les modifications apportées peuvent rectifier des oublis, déceler des éléments ajoutés ou même proposer des solutions de remplacement[3]:

*rectifier un oubli: *c'est à dire que* le *euh la proie* (Nancy5, 3, 6)
*ôter un ajout: *elle* elle *est arrivée* (Strasbourg2, 5, 36)
une catégorie particulière: une multitranscription abusive
*changement (le 2ème élément remplace le 1er):
elle a réussi à passer son CAP (de, euh*) de coiffeuse* (Strasbourg2, 5, 23)
*proposition d'une multitranscription:
/comme, quand*/ je suis pas directement au front* (Belfort5a, 3, 11)
*élucidation de passages incompris par le transcripteur (le 2ème):
j'ai fait des choses pour moi des petites sculptures (XXX, des petits nus*) des choses* (Belfort5B, 3, 7).

On peut objecter que la vérification, elle-même n'est pas exempte d'erreurs. C'est sans doute vrai et Blanche-Benveniste et Jeanjean (1987, p 112) le souligne : *Si j'entends aussi bien "a" que "b" dans une séquence de français parlé, je ne saurai donc jamais ce que le locuteur a "dit en vrai". Il semble bien que la réponse soit "non", même si on améliore les appareils d'enregistrements, et l'oreille de celui qui écoute, même si l'on met les gestes et les mimiques et même*

[2] Rappelons que l'auteur de l'article est la personne qui vérifie ces transcriptions. Elle est familiarisée depuis plusieurs année avec la transcription d'énoncés oraux.
[3] L'élément oublié ou ajouté n'est pas en italique dans le texte.

si l'on questionne le locuteur sur ce qu'il a voulu dire. Le texte authentique fuit, du moins par certains bouts.

Il est évident que cette étude se situe dans un cadre de recherche où l'objectivation des faits a des limites fondées théoriquement. En ce qui concerne le langage humain, pour qu'elle soit transcrite, la source orale doit être interprétée. Cette tâche d'interprétation peut parfois donner l'illusion d'une simple copie écrite de la version orale. Elle s'apparenterait alors au fonctionnement des logiciels de reconnaissance vocale dont on sait bien que malgré l'habituation indispensable à la voix du locuteur, ils sont loin de fournir une bonne représentation écrite de l'énoncé oral qui leur est soumis. Les outils informatiques permettent de mieux en mieux, grâce à leur puissance de calcul et à leurs ressources langagières, de prévoir les bonnes séquences et donc de proposer des solutions écrites plus approchées. Il reste que ces méthodes fondées sur la description statistique des éléments langagiers s'appuient sur des fréquences et non comme le transcripteur sur une compréhension d'un énoncé. À n'en pas douter, les erreurs de transcription de ces logiciels sont différentes de celles que font les transcripteurs (mais ceci reste à vérifier).

Le souci de rigueur scientifique, on l'a vu, conduit à reconnaître les limites des certitudes sur ce qui a été prononcé. Pourtant rares sont les consignes de transcription qui prévoient la possibilité de noter les diverses solutions d'interprétation (les multi-transcriptions) quand il n'est pas possible d'en privilégier une plutôt qu'une autre[4]. La plupart des conventions ne comportent pas de multi-transcription ce qui veut dire que le transcripteur ne peut que trancher, c'est à dire choisir. À ces endroits, l'objectivité de la méthode fait place à la subjectivité du transcripteur. Il est sans doute plus rigoureux de ne pas ignorer ces difficultés interprétatives et de ne pas masquer ces hésitations.

[4] Voir les conventions de transcription du GARS et celles de Kristol (1997) en Suisse.

5. ANALYSE DES MODIFICATIONS APPORTÉES AUX TRANSCRIPTIONS DE CES DOUZE CORPUS (TABLEAUX 2 ET 3 EN ANNEXE)

Six personnes différentes sont intervenues pour la transcription de ces douze corpus, chacune transcrivant, selon le cas, un ou plusieurs enregistrements. La vérification des transcriptions a été faite par une seule personne (l'auteur de l'article). Les résultats montrent une variabilité inter-transcripteurs mais aussi intra-transcripteurs (du moins pour les transcripteurs qui ont fait deux ou trois transcriptions.

Les propositions de modification analysées concernent trois catégories d'éléments:

- les tours de parole;

- les pauses courtes ou longues;

- les éléments du lexique parmi lesquels ont été distingués:

*les onomatopées et interjections du type:

heu (les plus fréquentes), *hein, ben, mh quoi, bon.*

*les éléments répétés (à tort ou oubliés); que la modification concerne la totalité des éléments répétés ou seulement une partie:

- répétition totale oubliée: *de* de *la possibilité (StgC2, 1, 14)*
- répétition partielle oubliée:

si il y a que des que *des adolescents* (StgC2, 3, 20).

Le nombre total de modifications proposées lors de la vérification s'élève à 581 soit en moyenne 48.1 propositions par corpus. L'écart type de 35.51 témoigne de la grande variabilité d'un corpus à l'autre. Les modifications proposées ne sont ni les mêmes ni aussi nombreuses d'un corpus à l'autre.

5.1. Les tours de parole

Des modifications sur les tours de parole n'ont été proposées que pour deux corpus (chacun ayant eu un transcripteur différent). Un certain nombre de relances émises par le locuteur qui conduit l'entretien ont été oubliées. Les modifications proposées se traduisent, donc, toutes par des ajouts d'onomatopées du type "mh" prononcées faiblement mais régulièrement par le locuteur enquêteur. Ces émissions sonores de l'auditeur qui selon les termes employés en analyse conversationnelle "régulent l'entretien" (Cosnier & Kerbrat-Orechionni, 1987). Elles n'ont pas été comptabilisées dans les oublis d'onomatopées, ces derniers nous paraissant d'un autre type. Si ces marques de relance (ou plutôt d'accompagnement de l'énoncé du locuteur par l'auditeur) sont isolées et constituent dans ces cas la seule insertion langagière de l'auditeur, les autres onomatopées apparaissent au sein même des énoncés et ne sont pas isolées.

Les modifications que nous avons proposées sont au nombre de 15 sur 169 relances et 8 sur 155 relances. Ces oublis pourraient être liés aux conditions d'enregistrement puisque le locuteur enquêteur est toujours le même et que les transcripteurs ne sont pas les mêmes. Ces transcripteurs ont transcrit d'autres corpus dans ce projet: leurs transcriptions n'ont pas fait l'objet de modifications en ce qui concerne les relances.

5.2. Les pauses

La pause est le terme couramment employé pour nommer un silence dans l'énoncé. Elle correspond à une cessation apparente de l'activité verbale du locuteur. Il n'est pas question ici de distinguer si les pauses identifiées dans ces enregistrements sont respiratoires, d'hésitation, de segmentation discursive ou rhétorique. Il n'est pas non plus question de confronter la réalité perçue et la réalité acoustique (pour vérifier dans quelle mesure il y a adéquation entre les deux niveaux). Cette distinction requiert des techniques d'enregistrement et des analyses que nous ne pouvions mener. Dans ce domaine, des travaux comme ceux qu'a

conduits notamment Danielle Duez (1991, 1993, 1997) ont montré le rôle fondamental de l'information prosodique.

Comme le souligne Claire Blanche-Benveniste *et al.* (1990, p. 37), *Nous notons sans grande rigueur, trois degrés de pauses: courtes, moyennes et longues. Nous pensons qu'une notation plus minutieuse ne serait pas nécessairement utile pour l'étude syntaxique.* En fait, les conventions de transcription ont même évolué jusqu'à ne distinguer que deux sortes de pauses: les brèves et les longues.

La pause est d'abord une notion contrastive en ce sens qu'elle ne peut être détachée de son contexte acoustique. La durée objective d'une pause ne peut être la même lorsque le débit est lent ou rapide. De plus, une "impression" de pause ne correspond pas toujours au niveau acoustique à une interruption du signal sonore. *L'existence de pauses subjectives...confirme le lien étroit qui lie la perception de la pause à son environnement acoustique* (Duez, 1993).

La perspective adoptée dans notre étude est celle de la perception des pauses et les modifications apportées lors de la vérification des transcriptions. Comme on pouvait s'y attendre, le nombre des pauses courtes varie considérablement d'un corpus à l'autre. Le nombre moyen de mots entre deux de ces pauses varie de 14 à 267 mots selon les corpus (Moyenne: 81.38; écart-type: 82.02). Rien de tel concernant les pauses longues (en moyenne: 4 par corpus soit une tous les 1000 mots).

Il n'est donc pas étonnant que, dans les corpus que nous avons étudiés, les pauses courtes soient plus sujettes à contestation que les pauses longues. Elles sont l'objet de nombreuses modifications (12% du total des modifications proposées) mais leur nombre varie d'un corpus à l'autre (Moyenne: 5.75; écart-type: 29.61; avec pour extrêmes une à 17 modifications). Ce qui est remarquable, c'est que les modifications apportées concernent exclusivement des oublis. Aucune ne signale une pause qui serait notée en trop. Ces résultats confirment ce qu'a rappelé Duez (1997): la transcription tend à omettre les pauses.

5.3. Élucidation d'éléments du texte incompréhensibles pour le premier transcripteur

Les éléments restés incompréhensibles dans ces corpus, après vérification, varient selon les textes. Ils sont en moyenne au nombre de 6 avec des limites extrêmes de 0 à 15 (Moyenne: 6.16; écart-type: 5.22). Un seul corpus se présente comme "entièrement" élucidé.

L'élucidation de passages incompréhensibles constitue 5% des modifications apportées, lors de la phase de vérification. Seuls quatre corpus sur 12 n'ont pu être améliorés. Le nombre de passages incompréhensibles décryptés dans ce deuxième temps varie aussi beaucoup d'un texte à l'autre: de 1 à 6 (Moyenne: 2.41; écart-type: 2.36).

La taille de l'énoncé ainsi élucidé varie aussi beaucoup: de 1 à 24 phonèmes (Moyenne: 5. 87; écart-type: 4.86).

Les catégories grammaticales concernées sont surtout le lexique nominal (15 cas sur 31) dont la moitié sont des noms propres ou du lexique spécialisé:
une disposition de crochets dite *orthognate* (Nancy24a, 3, 25)
les araignées d'ici *les Epeires* (Nancy24a, 3, 26).

On trouve aussi des cas de lexique verbal (5)
auparavant dans votre euh *c'est de* /sEd/ votre premier emploi (Strasbourg41c, 5, 5)
des amorces de lexique (1)
pourquoi pas *ap-* après tout (Bordeaux C7b, 2, 25)
des numérateurs (3)
je pense pas que *quelqu'un* ça je sais même pas comment ça fonctionne (Strasbourg41c, 12, 20)
une négation (*ne*) et un connecteur (*et*)
c'étaient les familles en difficultés financières qui *ne* venaient euh moi (Belfort5a, 2, 1)

non euh elle avait des lapins des lapins - un chien *et* des lapins (Belfort5a, 20, 6)
des séquences de plusieurs mots (en italique)
elle a vite bouclé le cycle (Nancy24a,1, 20)
donc on faisait des choses comme ça, (Nancy24b, 2, 12)
ouais non *ça serait trop difficile à expliquer* (Starsbourg 41c, 5, 13)
c'est elle qui me fait le dossier *et moi donc* (Belfort5a, 7, 3)
j'ai fait des choses pour moi des petites sculptures *des peits nus* des choses (Belfort5b, 3, 7).

Ce ne sont pas seulement les noms propres ou les termes techniques qui ont été, dans un premier temps et pour un même locuteur, incompréhensibles. Des énoncés au lexique tout à fait courant n'ont été élucidés que lors de la vérification. La perception des énoncés oraux est donc une activité complexe. Même après cette vérification, les passages restés incompréhensibles ne peuvent être taxés "d'inaudibles". On ne peut exclure que d'autres écoutes permettent d'identifier des morphèmes supplémentaires dans cette "gangue sonore" restante.

5.4. Les éléments oubliés

Les éléments oubliés sont beaucoup plus nombreux que ceux qui sont en trop. Ils constituent la plus grande partie des modifications; le tiers très exactement (203/581). Ce n'est pas négligeable mais on est loin de l'hypothèse de Cappeau (1997) selon laquelle les oublis représenteraient la quasi-totalité des erreurs de transcription.

Tous les corpus sauf un ont nécessité des modifications liées à l'oubli d'éléments. Ces modifications varient beaucoup selon les corpus mais aussi selon les catégories concernées:
- Onomatopées Interjections
- Éléments répétés
- Autres éléments.

Les "onomatopées et interjections" constituent les deux tiers des oublis. En particulier, les pauses remplies (*euh*) constituent l'essentiel de ces modifications (53/70). Le nombre de ces pauses varie selon les corpus entre 52 et 155. Le nombre moyen sur la totalité des 12 corpus est de 94,7 (écart-type: 39,5), soit 2,4 *euh* tous les 100 mots. Il reste que mis à part quatre corpus, le pourcentage de corrections sur les *euh,* ne dépasse pas 3% (des modifications). Quatre textes n'ont reçu aucune modification.

Les autres onomatopées qui ont dû être corrigées sont *hein* et *mh* et *ben*. La présence de ces éléments varie beaucoup d'un corpus à l'autre. Ils sont environs quatre fois moins nombreux que les pauses remplies. Les oublis sont infimes (en tout: 7 pour les *hein*, 4 pour le *mh* et 1 pour les *ben* et 4 pour les *bon*).

Contrairement à ce qu'on aurait pu penser, ce n'est pas à l'occasion de la répétition d'éléments que se produisent le plus fréquemment des oublis. Ces derniers sont deux à trois fois moins nombreux que dans les autres catégories. Quatre corpus ne nécessitent aucune correction d'oubli de répétition. Pour les corpus restant, ces corrections varient de 1 à 10 (Moyenne: 2.7; écart-type: 3.6). Elles concernent plusieurs catégories morphosyntaxiques mais rarement le lexique nominal:

- les déterminants (10): *un, de, les, le, la*;
- les pronoms clitiques (10): *on, vous, je, elle, leur*;
- des prépositions (3): *en, à, dans*;
- une amorce de mot (1) :

  à des influences *différen-* différentes (Belfort5c, 7, 15);
- des séquences verbales (4):

  ça a *ça a* induit (StrasbourgC2, 3, 20);

  oui moi je pense que c'est *c'est* toute une vie (BelfortC5c, 5, 36);

  s'il ya que des *que* des adolescentes (StrasbourgC2, 1, 8);

  et oui et vous l'avez *vous l'avez* trouvée l'astuce (Srtas41c, 5, 18);

  euh vous aviez *euh vous aviez* dit que (Belfort5a, 11, 7).

Les modifications qui ne concernent ni les éléments répétés ni les onomatopées et pauses silencieuses sont aussi nombreuses que ces deux catégories réunies. Là aussi la situation varie beaucoup d'un corpus à l'autre ; dans quatre cas, le nombre de modifications ne dépasse pas 2; dans d'autres cela peut aller jusqu'à 28 corrections (Moyenne: 8.16; écart-type: 7.45). Comme pour les oublis d'éléments répétés, les éléments oubliés appartiennent rarement au lexique nominal (7 en comptant les adjectifs et noms y compris sous forme d'amorce). On retrouve les catégories "pronoms clitiques et déterminants" mais la proportion des oublis dans ce domaine diffère beaucoup de ce qui a été trouvé avec les répétitions: alors que ces oublis constituaient les deux tiers (20/33) du total dans les répétitions, ils ne représentent plus que 23%. Ces modifications se répartissent ainsi :

- des pronoms clitiques (14): *le, il, vous, je, ce, ça, moi*;
- des déterminants (11): *un, une, des, le les, ce*;
- des connecteurs (5): *et, ou*;
- des prépositions (9): *de, pendant, à, à l', en, entre*;
- des adverbes (7): *là, trop, encore, beaucoup*;
- formes de négation (3): *ne, n', pas*;
- "que" ou "qu'" relatif ou non (5);
- amorces de mots (6): *b- beaucoup, à m- à mener, div- diverses, d'es- d'esprit, la der- la dernière, vous li- vous lisez*;
- adjectif et noms (4): *petit, de caution;*
- des phatiques (10): *oui, quoi, voilà, au fond, alors, pardon,* en *fait*;
- des verbes et des séquences verbales (25):

  bon euh j'ai *eu* une période assez marginale (Belfort 5b, 1, 21);
  oui *c'est ça oui* (BelfortC5a,16, 7);
  c'est pour le faire *et pour voir si* c'est fonctionnel (Strasbourg41c, 6, 11).

L'analyse de ces deux catégories d'oublis montre que ne se trouve pas vérifiée l'hypothèse selon laquelle les transcripteurs tendraient à occulter majoritairement "certains phénomènes tels que bribes et amorces" (Cappeau, 1997, p. 117). Les éléments répétés qui sont oubliés ne constituent que le tiers des oublis. Quant aux amorces de mots oubliées, elles sont rares (sept au total).

En particulier avec des éléments monosyllabiques, le contexte phonétique peut provoquer une mauvaise perception, voire même un oubli:

c'est à dire que *le* euh la proie (Nancy24a, 3, 6)

lorsqu'elle est revenue *là* il y a [ja] deux trois mois (StrasbourgC2, 6, 30).

Les bribes où l'énoncé piétine sur la même place syntaxique ne sont pas toujours bien perçues:

le logement c'est quelque chose *que* qui m'a toujours paru important (BelfortC5a, 4, 7)

elles passent des périodes d- d- *elles ont* des périodes difficiles (BelfortC5a, 6, 7)

j'ai eu *div-* diverses réactions (BelfortC5c, 6, 13).

Le transcripteur tend également à "écarter" ce qui éloigne de la phrase canonique comme dans les exemples suivants où le contexte phonétique favorise ce glissement:

au fur et à mesure *le* que la mygale liquéfie sa proie (Nancy24a, 3, 20)

l'action des pierres vivantes sont *ce* sont des pierres qui sont colonisées par des bactéries (Nancy24a, 7, 37).

On trouve aussi des exemples où, au contraire, l'oubli rend l'énoncé étrange mais ces cas sont les moins fréquents (14/100):

ils nous demandent si on *est* là pour du tourisme (NancyC6, 9, 10)

on allait mon- *on* montrait (NancyC6, 7, 7).

Dans d'autres cas, ce qui est oublié a une valeur de modalisation (adverbe phatique, négation, adjectif):

quand je dis on c'est *au fond* aussi les éducateurs (StrasbourgC2, 3 30)

lorsqu'elle est revenue *là* il y a deux trois mois (StrasbourgC2, 6, 30)

ah je sais *pas* comment préciser (Strasbourg41c, 7, 10)

j'ai un *petit* peu fait le tour, (Belfort5a, 4, 13).

5.5. Les ajouts d'éléments

L'écoute plurielle permet de constater la présence d'éléments transcrits qui sont en fait absents dans la source orale. Ces cas sont peu fréquents (4,5%; 26/581) mais existent tout de même. Ils varient d'un corpus à l'autre (cinq n'en présentent aucun). Quelques ajouts ont lieu qui concernent les onomatopées (3: *ben, mh euh*) et des répétitions erronées (6):

elle *elle* est arrivée (StarsbourgC2, 5,36)
par rapport aux techniques de la *de* la puce électronique (Strasbourg41c, 8, 29).

La plus grande partie des ajouts (17/26) n'appartiennent pas à ces catégories:
ça dure un *petit* peu plus longtemps StrasbourgC2, 1, 13
c'est vrai que *moi* je suis à l'aise (BelfortC5a, 14, 5).

De la même façon que certains oublis tendent à ôter du texte ce qui l'éloigne de la norme écrite, de même certains ajouts tendent à rétablir ce qui pourrait manquer à l'oral, en particulier pour les usages de la négation:
Je *ne* parle pas de l'indépendance (StrasbourgC2, 1, 21)
quelque chose s'est passé pour elle qui *n'*est pas du côté du (StrasbourgC2, 7, 2)
je *ne* veux plus rien en savoir (StrasbourgC2, 7, 2).

5.6. Les éléments remplacés[5]

La deuxième grande catégorie de modifications concerne les passages des corpus pour lesquels une autre transcription a été proposée en remplacement de celle qui avait été préalablement choisie. Elles constituent 21,4% (123/581) du total des propositions avec pour valeurs extrêmes 3 et 18 (Moyenne: 10. 25; écart-type: 5.44). Tous les corpus ont eu des passages remplacés.

[5] Dans les exemples, la correction de l'erreur est le 2ème terme dans la parenthèse, le premier étant l'erreur elle-même.

Une seule erreur de référence a été trouvée: elle est attestée par une divergence dans la cohérence textuelle. Le locuteur, enregistré dans un muséum, décrit les besoins des poissons observés:
pour *qu'ils puissent* construire *leur* squelette - et après donc le dernier point important c'est le traitement de l'eau - donc c'est-à- parce qu'il faut (il; *ils*) (demande, *demandent*) une eau de très grande qualité Nancy5, 7, 34.

Un certain nombre d'erreurs (18) se traduisent par l'omission d'un phonème:
- d'une voyelle (8):
  on allait mon- (montrer, *on* montrait) (Nancy6, 7, 7)
- d'une consonne (6)
  et (quel était, quels étaient) l'organisme (Nancy6, 4, 3)
- d'une syllabe ou d'un mot (4)
  il y a déjà trois (muci-, munici-) municipalités (Bordeaux7b, 7, 4)
  puisqu'il y a eu la (distribution, redistribution) des terres (Troyes102, 17, 5)
  puis après elle a eu euh elle (travaillait, avait travaillé) dans le temps donc
  et elle a eu droit a une petite retraite (Belfort5a, 18, 9)
  c'est euh un comment dire un (un, une) interaction (Strasbourg41c, 11, 10).

Une autre catégorie d'erreurs, tout aussi nombreuses (17) et qui se traduisent au contraire par l'ajout d'un phonème:
- d'une voyelle (5)
  après donc (que; euh) le dernier point important c'est le traitement de l'eau (Nancy5, 7, 25)
- d'une consonne (7)
  le fait qu'il y ait niveau (bac, bas) et que peut-être ça pose aussi un problème (StrasbourgC2, 4, 7)
- d'un mot (5)
  une fonction qui m'a i- été rajoutée (d'une part, par) après quoi Strasbourg41c, 2, 27)
  en fait il est (grand en; en) comparaison avec d'autres bacs (Nancy5, 3, 29).

Les erreurs les plus fréquentes (30) se traduisent par un phonème erroné, concernant:

- une voyelle (12)

c'est la (défense; défonce) pour la proie (Nancy5, 3; 16)

est-ce que j'ai bien à être là (un, en) F3 (Belfort5b, 16, 10)

- une consonne(16)

d'autres bacs (de haute mer, d'eau de mer) Nancy5, 3, 29)

- une interversion de phonèmes (2)

vous aviez déjà (euh fait, fait euh) des prototypes (Strasbourg41c, 5, 4)

parce qu'il y a un petit duplex pas cher et puis (enfin euh, euh enfin) (Belfort5a, 16, 15).

On trouve aussi de nombreuses erreurs (22) qui se traduisent par plusieurs phonèmes erronés. En général, le terme erroné est un mot:

l'exposition dont vous parlez là (du Wissendorf, à Wissenbourg) (Belfort5b, 3 12)

je vais faire une autre toile (qui, quoi) donc le cadre de la de l'histoire bien bien cernée m'embêtait un peu (Belfort5b, 3, 29)

J'ai pris (le, un) minitel (Belfort5b, 16, 5).

Plusieurs exemples (14) révèlent qu'une séquence de mots peut être erronée:

sinon (ce- ce qu'est; c'est que) les coraux ont besoin (Nancy5, 5, 28)

et là (dessus X, c'était) pendant un stage (Strasbourg41c, 6, 15)

je leur ai fait un proto (avec la fonctionna-, et il a fonctionné) (Strasbourg41c, 6, 21).

La description de ces erreurs en terme d'omission ou d'ajout de phonèmes permet de relier ce qui est observé à ce qui a été décrit à propos des oublis ou ajouts de pauses et de lexique. On constate que, contrairement aux résultats obtenus alors, les omissions ne sont pas plus fréquentes que les ajouts. Ce constat suggère que l'essentiel dans l'activité de transcription ne se résume pas à une perception phonologique précise.

Certains de ces exemples peuvent suggérer que le transcripteur soit "contraint" par les règles de la phrase canonique ce qui suspendrait sa vigilance perceptive:

et (quel était, quels étaient) l'organisme (Nancy6, 4, 3).

Dans d'autres exemples au contraire, la complexité de l'élaboration orale transcrite ne s'en trouve pas simplifiée:

donc (la , le) plus souv- la plupart du temps (Nancy6, 6, 20)

gravure (sous, sur) bois euh pyrogravue enfin (Belfort5b, 1, 10).

5.7. Multitranscriptions insérées[6]

Les multitranscriptions déjà proposées par le transcripteur ne sont pas rares (102). Tous les corpus en comportent mais de façon variable (Moyenne: 8.5; écart-type: 4.23) avec des valeurs extrêmes: 2 et 15. Certaines de ces multitranscriptions proposent des solutions homophones. Ce type d'incertitudes est relativement peu fréquent (Bilger., Blasco, Cappeau, Pallaud, Sabio & Savelli, 1997) et correspond aux alternances orthographiques. Ces hésitations peuvent être dues soit à une difficulté irréductible d'interprétation (quelle qu'en soit la raison) soit au fait que l'usage orthographique n'est pas stable:

c'étaient les familles en difficulté(s) financière(s) (Belfort5a, 1, 16)

ce genre de difficulté(s) (Belfort5a, 2, 8)

ce d- signalement d'enfant(s) (Belfort5a, 2, 8).

Lorsqu'en certains passages, la vérification conclut à une incertitude sur ce qui est "écouté", la solution adoptée est de proposer une multitranscription. Cet artifice typographique permet de signaler et transcrire la solution du transcripteur et celle obtenue lors de la vérification. Ces modifications-là ne sont pas très nombreuses: 32/581 soit 5,5% (Moyenne: 2.6; écart-type: 2.95). Trois corpus ne sont pas concernés, les extrêmes des valeurs se situant entre 0 et 9.

---

[6] La solution obtenue lors de la vérification est le 2ème terme dans la multitranscription.

Environ le tiers (9) de ces modifications conduisent à proposer des solutions homophones. Les segments concernés sont courts (1 à 2 phonèmes) et les amorces de mots se trouvent impliquées dans ces hésitations:

et /j'ai, j'ai-/ /j'ai, jai-/ j'aimais pas quoi (Belfort5b, 2, 16)

Je suis née à Belfort et bon alors dans /ma, la/ famille on /s'est, sait/ évidemment ça dépend beaucoup de la famille (Belfort5c, 4, 21).

La modification a révélé une hésitation référentielle dans un seul cas (le locuteur explique le comportement alimentaire des mygales exposées):

est-ce que celle(s)-là enfin bon la vulgarisation le dit - est-ce que celle(s)-là c'est identique elle(s) euh comment elle(s) euh consomme(nt). (Nancy5, 3, 2).

Parmi les multitranscriptions proposées non homophones (23) dix traduisent une variation sur un seul phonème. Comme le souligne (Cappeau, 1997) cette variation infime sur le plan perceptif peut avoir des conséquences non négligeables sur le plan syntaxique:

je suis dans une recherche d'emploi est-ce qu'il vaut peut-être pas mieux prendre un appartement /qu'il, qui/ soit plus proche (Belfort5a, 16, 13).

La mise en grille de cet énoncé (Blanche-Benveniste *et al.*, 1990) n'est pas du tout la même selon la solution choisie:

1- je suis dans une recherche d'emploi
est-ce qu'il vaut peut-être pas mieux prendre un appartement
qu'il soit plus proche

2- je suis dans une recherche d'emploi
est-ce qu'il vaut peut-être pas mieux prendre un appartement *qui* soit plus proche.

Les 13 autres propositions de multitransciption traduisent une variation sur plusieurs phonèmes. Là aussi les conséquences morphosyntaxiques sont plus ou moins importantes.

5.8. Multitranscriptions supprimées[7]

Dans un certain nombre de corpus (5 sur 12), la vérification a conduit à proposer la suppression de quelques multitranscriptions puisqu'un des termes pouvait être reconnu. Leur nombre n'est pas négligeable (deux fois plus que des insertions de multitranscriptions): 72 sur les 581 modifications (soit 12,5%).

La moitié de ces "réductions" concernent les alternances- zéro /X, Ø/. Dans ce cas, le transcripteur avait hésité sur la présence (X) ou non (Ø) d'un élément. L'élément incriminé peut comporter un seul phonème (10), deux phonèmes (23) ou même plus (2):

parce que c'est très différent la vie en Italie /et, Ø/ en Espagne (TroyesPr101, 4, 9)

c'est ce qui fait /que, Ø/ généralement il saute (Troyes102, 10, 3)

ça /ne, Ø/ peut être qu'un sicilien (Troyes102, 18, 4)

je vous en avais déjà parlé /Ø, ici/ du droit (TroyesPu101, 13, 7).

Dans trois cas, la "réduction" a au contraire conclut à l'absence de tout élément. L'erreur a consisté alors à ajouter de l'énoncé là où il n'y en avait pas:

le patronat a toujours dit que il /n', Ø/ aurait pas euh il /n', Ø/ aurait pas d'histoire (TroyesPu101, 12, 4).

Comme pour les changements proposés, les noms propres et le lexique spécialisé ne constituent pas l'essentiel de ces modifications:

elle travaille sur la restauration des /lésions, *liaisons*/ (Strasbourg41c, 3, 11

il y avait pas eu de (/réduction, *révision*/, du temps de travail (Strasbourg41c, 17, 2).

Dans 19 cas, la variation ne porte que sur un phonème. Même si cette variation est faible ses conséquences sur le plan morpho-syntaxique peuvent être importantes.

---

[7] La solution erronée est en italique.

que chacun puisse loger le plus confortablement possible selon */si,* ses/ moyens et donc (Belfort5a, 5, 11).

Sur les 35 cas relevés, 10 correspondent à un changement de catégorie grammaticale:

vous en vous euh /au fond, *vous f-/* dans une journée (Belfort5a, 7, 10)
c'est l'aventure totale quand ils vont /la voir, *là-bas/* (Belfort5a, 19, 13).

Pour les 25 autres cas, la catégorie grammaticale ne change pas. Ce qui varie concerne par exemple:

- le temps du verbe:
  /je fais, *j'ai fait/* la démarche d'accéder à un logement (Belfort5a, 16, 10)
- le type de verbe:
  puisqu'elle y arrive plus là on /aide, *est/* on a une réelle marge de manoeuvre (Belfort5a, 8, 8)
- le nombre dans le nom ou l'adjectif:
  pour présenter */votre*, vos/ dossiers (Belfort5a, 14, 10)
- la construction syntaxique (double marquage ou non) et type de verbe
  donc le projet /il doit, *va/* se mettre en place (Belfort5a, 15, 2).

6. CONCLUSION

La vérification de la transcription de 12 corpus de français parlé révèle un certain nombre d'erreurs très variables en quantité et en types d'un corpus à l'autre. Les pauses silencieuses courtes comme les éléments phatiques de relance sont *oubliées* de façon très diverse selon les corpus mais de façon non négligeable. Les ajouts en revanche sont assez rares. La même suprématie des oublis sur les ajouts est observée dans les modifications d'éléments lexicaux onomatopées et interjections (qui constituent le tiers des modifications). Parmi ces dernières, la catégorie "pauses remplies" (euh) constitue le quart. Les oublis de répétition sont encore moins nombreux. Si toutes les catégories morpho-syntaxiques semblent

touchées, celle des verbes l'est de façon prépondérante. Toutes les transcriptions sauf une conservent des éléments non élucidés; la vérification permet d'améliorer les textes dans 5% des cas pour l'interprétation de passages pouvant varier de 1 à 24 phonèmes. Le remplacement de certains passages constitue la deuxième grande catégorie des modifications. Ces erreurs de reconnaissance morphologique concernent tous les corpus. La solution morphologique correspond le plus souvent à un ou plusieurs phonèmes erronés; dans les cas restants, on remarque que ces solutions se traduisent aussi bien par un oubli que par un ajout de phonèmes. Les multitranscriptions supprimées sont deux fois plus nombreuses que celles qui sont proposées; la moitié concerne des "alternances -zéro" où le doute était émis sur la présence ou l'absence d'un élément.

La transcription requiert de quitter le terrain des phonèmes pour entrer dans la reconnaissance des morphèmes. Des transcripteurs expérimentés savent qu'en certains points il s'agit moins d'être fidèle à ce qui est entendu qu'à ce qui est écouté. On l'a vu les erreurs constatées sont rarement des erreurs de cohérence textuelle ou des glissements vers la traduction d'une prononciation. Elles semblent plutôt constituer des "mal-entendus", erreurs d'écoute chez le transcripteur (Cutler, 1981).

Berthille PALLAUD
Université de Provence
UMR 6057 Parole et Langage
29, avenue Robert Schuman
F-13621 Aix-en-Provence cedex.
France
☎ (33-4) 42953644
Fax (33-4) 42595096
Courrier électronique: pallaud@newsup.univ-mrs.fr

ANNEXE

	Onomatopées Interjections	**Répétitions**	**Autres**	**Total**	**%**
Éléments oubliés	70 *5,75 (7,07)*	33 *2,75 (3, 36)*	100 *8,16 (7,45)*	203	34,9 %
Éléments en trop	3	6	17	26	4,5 %
Éléments remplacés				123	21,2%
Multitranscriptions proposées				32	5,5%
Multitranscriptions réduites				72	12,5%
Éléments élucidés				31	5,0%
Oubli de relances				23	3,9%
Relances en trop				0	0%
Pauses courtes Oubliées				69	11,9%
Pauses courtes en trop				0	0%
Pauses longues Oubliées				2	0,5%
Pauses longues en trop				0	0%
TOTAL				581	100%

Tableau 2 : modifications proposées lors de la vérification des 12 corpus.

	col.A Total	Moyenne *(Écart-type)*	Nombre moyen de mots entre deux éléments[8]	Nombre de modifications
Ligne 1 Taille (mots)	**46 649**	3887 *(574,64)*		
Relances	1397	87,3 *(50,36)*	33,4	3 tous les 100 mots
Pauses courtes	1123	93,6 *(75,33)*	41,5	2,3 tous les 100 mots
Pauses longues	48	4 *(4,53)*	971,8	1 tous les 1000 mots
Amorces lexicales	197	16,42 *(10,06)*	236, 8	4 tous les 1000 mots
Phatiques *euh*	1151	95,9 *(39,59)*	40,5	2,4 tous les 100 mots
Multitranscriptions	102	8,5 *(4,23)*	457,3	2 tous les 1000 mots
Passages non interprétés	74	6,6 *(5,22)*	630,4	1,6 tous les 1000 mots
Modifications apportées	581	48,41 *(35,51)*	80,3	1,24 tous les 100 mots

Tableau 3: statistiques sur certains paramètres dans les 12 corpus.

[8] Ce chiffre est obtenu à partir des données du tableau 2 par la formule suivante: pour les corpus "Français de Référence": total des mots /total de l'élément (colonne A).

RÉFÉRENCES BIBLIOGRAPHIQUES

ARRIVÉ, M., NORMAND, C. (éds), *Linguistique et psychanalyse*, Cerisy-la-Salle, In Presse, 2001.

BILGER, M., BLASCO, M., CAPPEAU, P., PALLAUD, B., SABIO, F., SAVELLI, M., Transcription de l'oral et interprétation; illustration de quelques difficultés, *Recherches sur le français parlé*, **14**, pp. 57-86, 1996.

BLANCHE-BENVENISTE, C., JEANJEAN, C., *Le français parlé. Transcription et édition*, Didier Érudition, Paris, 1987.

BLANCHE-BENVENISTE, C., BILGER, M., ROUGET, C., VAN DEN EYNDE, K., *Le français parlé. Études grammaticales*, Édition du C.N.R.S., Paris, 1990.

BLANCHE-BENVENISTE, C., PALLAUD, B., Le recueil d'énoncés d'enfants. Enregistrement et transcription, *Recherches en Syntaxe du Français Parlé*, **16**, pp. 11-37, 2001.

CAPPEAU, C., Données erronées: quelles erreurs commettent les transcripteurs?, *Recherches en Syntaxe du Français Parlé*, **14**, pp.117-126, 1997.

COSNIER, J., KERBRAT-ORECCHIONI, C. (eds), *Décrire la conversation*, Presses Universitaires de Lyon, 1987.

CUTLER, A. (ed.), *Slips of the Tongue and Language Production*, Moton publishers, Amsterdam, 1981.

DUEZ, D., *La pause dans la parole de l'homme politique*, Éditions du CNRS, Collections sons et parole, Paris, 1991.

DUEZ, D., Acoustic Correlates of Subjective Pauses, *Journal of Psycholinguistic Research*, **22 (1)**, pp. 21-39, 1993.

DUEZ, D., La signification des pauses dans la production et perception de la parole, *Revue PArole*, **3/4**, pp.275-299, 1997.

GIOVANNONI, D., SAVELLI, M., Transcrire et orthographier le français parlé. De l'impossible copie à la falsification des données orales, *Recherches Sur le Français Parlé*, **10**, pp. 19-39, 1990.

KRISTOL, A. M., *Atlas linguistique audiovisuel du Valais romand*, Centre de dialectologie et d'étude du français régional de Suisse romande, Université de Neuchâtel, 1997.

PALLAUD, B., Les lapsus: des pierres dans le champ linguistique, in ARRIVÉ, M., NORMAND, C. (éds), *Linguistique et psychanalyse*, Cerisy-la-Salle, In Presse, pp. 47-66, 2001.

PALLAUD, B., SAVELLI, M., L'oral enfantin. Quelques précautions pour l'évaluer, *Revue Française de Linguistique Appliquée*, **VI, I**, pp. 121-136, 2001.

RÉSUMÉ

Lors de la vérification des transcriptions de douze corpus de français parlé, un certain nombre d'erreurs très variables en quantité et en types ont pu être observées. On retrouve la très grande prédominance des oublis dans les pauses silencieuses et remplies ainsi que dans les tours de parole. De même, les éléments lexicaux, onomatopées et interjections, beaucoup plus souvent oubliés qu'ajoutés, constituent le tiers des modifications apportées aux textes. La taille des passages qui ne sont élucidés que lors de la vérification peut varier d'une vingtaine de phonèmes. Le quart des modifications concernent des propositions de remplacement d'énoncés. Si des multitranscriptions sont supprimées, celles qui sont proposées après vérification sont deux fois moins importantes. Ces deux catégories sont presque aussi nombreuses que les éléments remplacés. Ces erreurs d'écoute nous semblent constituer les "mal-entendus", erreurs d'écoute chez le transcripteur.

SUMMARY

The transcription of French oral corpora was checked and transcription errors were quantified and analysed. Errors differ in number and in type according to the corpora. We observed the great predominance of omission over addition in silent and filled pauses as well as turn taking. Similarly, lexical units and interjections are far more omitted than added. These changes form a third of suggested modifications in the texts. The size of utterances which are deciphered only during the verification phase may vary from one to 24 phonemes. A fourth of the changes are composed of substituted sequences. If some multitranscriptions are withdrawn, those which are inserted are twice less numerous. Both categories are nearly as important as substituted units. These listening errors may be considered as "slips of the ear" encountered during the transcribing activity.