HAL open science

Du bon usage des amorces dans la transcription des corpus

Marie-Noëlle Roubaud

▶ To cite this version:

Marie-Noëlle Roubaud. Du bon usage des amorces dans la transcription des corpus. Recherches sur le français parlé , 2004. hal-01925242

HAL Id: hal-01925242

https://amu.hal.science/hal-01925242v1

Submitted on 16 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Du bon usage des amorces dans la transcription des corpus

Marie-Noëlle ROUBAUD
Université de Provence

Résumé

L'équipe du GARS (Groupe Aixois de Recherches en Syntaxe) relayée par l'équipe DELIC (DEscription Linguistique Informatisée sur Corpus) s'est penchée sur le problème de la transcription des amorces de mots (ou fragments de morphèmes). L'examen de dix-neuf corpus de français parlé montre que cette tâche est ardue. Des questions se posent au transcripteur : Comment identifier les amorces ? Comment les transcrire en restant fidèle à la parole du locuteur ? N'existe-t-il pas plusieurs interprétations possibles ? Cet article tente d'apporter des réponses et veut surtout montrer que la transcription des amorces passe par une analyse syntaxique des corpus.

1. Introduction

Depuis plus de vingt-cinq ans, l'équipe du GARS relayée par l'équipe DELIC collecte des corpus de français parlé en vue de recherches syntaxiques. Ce travail l'a conduite à transcrire les données orales en orthographe standard, sans ponctuation et sans marque intonative et à établir des conventions de transcription. Parmi celles-ci figure la transcription des amorces de mots (en gras dans l'extrait) dont Pallaud (2002 : 79) indique que leur présence se traduit par « une interruption de morphèmes en cours d'énonciation »[1] :

[1] Son étude quantitative révèle que les amorces « constituent plus de la moitié des achoppements de langage identifiés dans un corpus enregistré de français contemporain » (Pallaud, 1999, 2002).

Recherches sur le français parlé, 18, 163-184.
©2004, *Université de Provence*

(1) voilà le jour de l'Ascension il fallait rien faire fallait pas **trava-** faut pas travailler le jour de l'Ascension c'est un jour particulier alors l'herbe de l'Ascension l'œuf de l'Ascension la veilleuse de l'Ascension tout ça ça permettait de se protéger contre les éléments les **élé-** les les éléments naturels et je ne sais pas si ça donnait des bons résultats mais nos vieux ils y croyaient et et le le soleil il revenait et **l'ar-** l'arc-en-ciel on le voyait au loin on disait tu vois et maintenant ils vont rentrer et il n'y aura pas eu de + voilà ça ce sont des des usages qui ont qui ont qui existent encore à l'heure actuelle (COR-PRI001)

L'observation et la correction de dix-neuf corpus de français parlé[2] (soit un peu plus de 46 000 mots) montre que la transcription des amorces n'est pas une tâche facile, contrairement à ce qu'on pourrait penser : elle met en cause une analyse syntaxique.

Cet article veut rendre compte de cette réflexion sur les amorces, reposant sur tout un travail collectif de transcripteurs[3]. Pour cela, nous reviendrons d'abord sur les principes de transcription du GARS pour ensuite considérer le cas des amorces. Puis nous tenterons de dresser une liste d'outils aidant à l'identification des amorces. Et enfin, nous reviendrons sur les problèmes rencontrés dans les corpus observés : nous nous interrogerons sur le bon usage des amorces, examinerons les cas difficiles, proposerons quelques solutions[4].

2. Les principes de transcription

Quand nous lisons, nous avons l'habitude de reconnaître, sous le découpage en mots, des morphèmes, c'est-à-dire des unités de sens. L'orthographe est notre « déchiffreur » courant des morphèmes[5]. C'est pourquoi le GARS puis DELIC ont choisi de transcrire des morphèmes, en orthographe standard, orthographe donnée par les dictionnaires.

---

[2] Ces corpus ont été enregistrés dans différentes régions de France et dans différentes situations (conversations, relations d'évènements, discours de guide, cours de professeur, etc.), dans le cadre d'un projet de constitution d'un *Corpus de référence de français parlé*, conduit par l'équipe DELIC. La dernière phase de ce projet a consisté en la relecture et la correction de ces corpus.

[3] Cette réflexion a été menée par l'équipe du G.A.R.S. et se poursuit au sein de l'équipe DELIC.

[4] Les exemples qui ont été corrigés seront présentés sous deux versions : la version initiale (ex.) et celle corrigée (ex. corrigé).

[5] cf. Blanche-Benveniste et Jeanjean (1986 : 124).

Cette précaution à transcrire tous les morphèmes justifie que dans les conventions de transcription apparaît un *n'*, hésitation orthographique qui rend compte de la présence possible du morphème *ne* de négation, impossible à entendre après une liaison comportant une nasale :

(2) on **(n')** a pas de moyens (crfp, prilar1)

La transcription ne comporte aucun « trucage orthographique »[6] pour rendre compte des réductions de morphèmes absolument courantes, caractéristiques de la langue parlée, quels que soient les locuteurs. Que la voyelle du pronom *tu* ou la consonne /r/ de *parce que* soit élidée, l'orthographe du morphème est conservée :

(3) et **tu** as **tu** as pas mal d'hôtels qui sont aménagés (crfp, prilem1)

et non pas :

et **t'**as **t'**as pas mal d'hôtels qui sont aménagés

(4) mais à part les ruines il te reste vraiment pas grand-chose **parce que** tout a été détruit (crfp, prilem1)

et non pas :

mais à part les ruines il te reste vraiment pas grand-chose **pa'ce que** tout a été détruit

Il en est de même pour les nombreux pronoms *il(s)* ou *elle(s)* parsemant les corpus, dont la prononciation est réduite le plus souvent à l'émission d'une voyelle. La transcription rétablit l'orthographe du morphème :

(5) par exemple **il** tu as remarqué le long de la rivière il y a des petites chapelles (crfp, pricor1)

et non pas :

par exemple **i-** tu as remarqué le long de la rivière

(6) bon la vigne c'est sûr que **elle** + il y a **i-** + elle est plus souvent chez nous du moins elle est plus souvent défeuillée que feuillée (crfp, prigap1)

et non pas :

bon la vigne c'est sûr que **è-** + il y a **il** + elle est plus souvent chez nous ...

La particularité de prononciation peut être mentionnée ailleurs, par exemple sous forme d'une étude phonétique adjointe à la transcription ou en note de bas de page. Il est plus facile de partir de l'identification des

6 cf. Blanche-Benveniste et Jeanjean (1986 : 130).

morphèmes, d'après une transcription orthographique, et de rajouter ensuite les particularités phoniques, que d'aller en sens inverse. En effet si nous considérons l'exemple de *il y a*, plusieurs transcriptions phonétiques sont possibles [ilija], [ilja], [ija], [ja] ou encore chez les enfants [ila], [la]. Toutes les prononciations évoquées représentent la même suite de morphèmes. Si on s'intéresse à la combinaison des morphèmes, et non au niveau inférieur de la prononciation des morphèmes, on ne note pas la prononciation.

C'est au nom de ce principe que nous ne descendons pas en dessous du morphème[7]. Ainsi dans l'exemple qui suit, nous n'indiquerons pas l'enchaînement entre le déterminant *une* et le nom *équipe* :

(7) on peut tomber la première partie contre **une équipe** qui est qui est can- cant- comme nous (crfp, prilil1)

et non pas :

on peut tomber la première partie contre **une /n/ équipe** qui est qui est can- cant- comme nous

3. Le cas des amorces

Les amorces sont des fragments de morphèmes, ou d'hypothèses qu'on fait sur les fragments de morphèmes, elles sont transcrites avec un trait d'union, collé à l'amorce. Ce sont des unités de sens ainsi que des représentants de constituants syntaxiques. En effet, il nous suffit de lire les amorces, dans les trois exemples suivants, pour identifier des unités de sens et les syntagmes correspondants :

(8) j'en connais qui + des élevages porcins la la preuve c'est quand même des + des **polémi-** des polémiques au niveau de environnemental pour le la création de nouveaux élevages et tout ça (crfp, prigap1)

(9) c'est-à-dire que les les **f-** les **états fédérés** ressemblent à l'Allemagne ressemblent à l'Italie ressemblent à l'Angleterre (crfp, prigap2)

(10) et c'est le coq au vin euh est un de ces est un de ces plats + on **l'ap-** au départ on faisait euh le le euh le macérer le le le poulet (crfp, pridij1)

Si pour l'exemple 8, l'amorce est évidente puisque elle est complétée aussitôt après par *des polémiques*, pour l'exemple 9, c'est le contexte qui permet d'interpréter l'amorce : amorce de *fédérations* ou anticipation du lexème *fédérés* produit ensuite. Quant à l'exemple 10, c'est la place syntaxique occupée par l'amorce (en position verbale) qui peut la faire

7 cf. Giovannoni & Savelli (1990 : 21).

interpréter comme un fragment du morphème *apprêtait* (*on l'apprêtait au départ on faisait macérer le poulet*). En effet, comme l'écrit Cappeau (1997 : 121), la reconnaissance des morphèmes dans les corpus se fait au niveau du syntagme :

> « Le travail du transcripteur opère ici au niveau du syntagme. À partir d'un moule phonique, il reconstitue une séquence qui puisse s'insérer dans le contexte. C'est pourquoi il propose des équivalents acceptables à partir de certains indices formels ou sémantiques. »

Comme nous l'indiquent les trois exemples précédents, les amorces nous renseignent sur la composition de la syntaxe et sur la progression du sens dans le discours du locuteur. Cet aller et retour sur l'axe syntagmatique peut être visualisé en prenant comme point d'appui la place du morphème, telle qu'elle est donnée dans la dernière mise au point et en alignant, dans une même « colonne » verticale, les éléments occupant la même place syntaxique. C'est ainsi que le GARS a développé la mise en grille[8] pour visualiser par écrit l'organisation du discours oral. Nous mettrons ici en grille (notée : G), sous les exemples, l'extrait portant sur la recherche du locuteur :

- recherche de l'objet nominal du verbe *avoir* :

(11) les enseignants à l'époque n'avaient peut-être pas euh comme aujourd'hui euh **l'ha- l'ha- l'ha-** l'habitude enfin la la la vocation hein (crfp, procor1)

G : les enseignants à l'époque n'avaient peut-être pas euh comme aujourd'hui
euh
l'ha-
l'ha-
l'ha-
l'habitude
enfin la
la
la vocation hein

Henry & Pallaud (dans ce numéro) montrent d'ailleurs que les amorces répétées (ex. 11) sont complétées, dans 80% des cas.

---

8 Pour l'analyse en grilles, cf. Blanche-Benveniste et Jeanjean (1987 : 167) et Blanche-Benveniste, Bilger, Rouget, Van den Eynde, (1990 : 214). Dans Roubaud & Loufrani (1999), nous avons montré que la mise en grille était un outil commode pour analyser les corpus de type aphasique.

- recherche du nom dans le syntagme nominal *l'intitulé du diplôme* :

  (12) alors que **l'in- l'intitu-** l'intitulé du du diplôme est le même hein (crfp, pubcor1)

  G : alors que **l'in-**
  l'intitu-
  l'intitulé du diplôme est le même hein

- recherche du nom dans le syntagme nominal *un avenir certain* :

  (13) euh hum l'Aligoté aussi a un vin qui + qui aura euh + un un **aven-** un **a-** un avenir certain dans dans des plats euh surtout dans des entrées ou des apéros (crfp, pubdij1)

  G : qui aura un
  un **aven-**
  un **a-**
  un avenir certain

- recherche du nom avec correction du lexème adjectival dans le syntagme nominal :

  (14) quand il y avait une belle **nou-** une bonne nouvelle elle allait elle allait très vite même s'il n'y avait pas de téléphone (crfp, pricor1)

  G : quand il y avait une belle **nou-**
  une bonne nouvelle

- recherche du nom avec correction du nombre du déterminant dans le syntagme nominal :

  (15) on a quand même un laps de de repos moi je suis du matin ma collègue est du soir + euh elle a son samedi moi j'ai mon lundi on a notre **rep-** nos repos + qui sont obligatoires (crfp, pricle3)

  G : on a notre **rep-**
  nos repos

- recherche du verbe dans le syntagme verbal :

  (16) et puis lui il **d-** il désignait deux tireurs deux pointeurs (crfp, prilil1)

  G : et puis lui il **d-**
  il désignait deux tireurs
  deux pointeurs

- recherche du verbe infinitif et abandon de l'amorce (probablement *faire*) pour un lexème plus spécifié *construire* :

  (17) quand on a commencé à détruire le vieux couvent pour **f-** construire le groupe scolaire on a construit en même temps le clocher actuel de l'église (crfp, pricor1)

  G : quand on a commencé à détruire le vieux couvent pour **f-**
  construire ...

Dans les corpus, nous trouvons de nombreux monosyllabes concernés par les amorces, comme les déterminants (ex. 18), les pronoms (ex. 19), les prépositions (ex. 20) :

(18) bon on va repartir sur **c- c-** après **cette** évocation de de de un peu de l'histoire du conflit entre les deux langues (crfp, prigap2)
(19) je **v- vous** citerai pas tout (crfp, pridij1)
(20) je leur souhaite à se donner les moyens + **d- dans** l'Europe c'est peut-être un peu plus difficile (cfp, prigap2)

Ces amorces subissent les déformations phonémiques et leur correction a lieu, toujours en revenant au départ du groupe syntagmatique. Dans l'exemple 21, le locuteur anticipe, à l'intérieur de l'amorce, sur le phonème /r/ qui vient trop tôt dans le mot *patrimoine*[9] et dans l'exemple 22, il anticipe sur une voyelle du mot *appellation* :

(21) il y a des des des sujets qui ont /a-, a/ trait au monde rural et des a- des sujets euh qui ont /a-, a/ trait au au **par-** au au **patrimoine** euh euh naturel (crfp, procor1)
(22) il vient de Mercurey donc c'est en Saône-et-Loire + normalement bon c'est une euh région qui est assez connue euh pour ses blancs + ses blancs de Chardonnay + où les blancs sont souvent élevés en fûts donc on verra si ce vin est élevé en fût euh d- de cette **epp-** de cette **appellation** (crfp, pubdij1)

Comme nous le voyons au travers de ces exemples, ce sont, dans la majorité des cas, des amorces d'anticipation que le locuteur produit, témoignant de sa capacité à effectuer des allées et venues sur l'axe syntagmatique. Comme l'écrit Claire Blanche-Benveniste (1990 : 23) :

> « Ces anticipations montrent que le locuteur tient en réserve, lorsqu'il parle, les syntagmes qu'il vient de dire aussi bien que ceux qu'il projette de dire. »

Dans les exemples précédents, cette mise en mémoire était brève car peu de syllabes séparaient l'amorce du morphème appartenant au même groupe syntagmatique. L'exemple 23 constitue une exception car sept syllabes séparent l'amorce *op-* du lexème *opération*, mais le syntagme *on peut considérer que*, qui s'intercale entre eux, agit comme une sorte de précaution oratoire, dans un discours qui paraît un peu embrouillé. C'est ce que visualise la mise en grille sous l'exemple :

(23) bon maintenant on a quand même une population de près de cent cinquante cerfs bon /dont, donc/ une partie + hein euh **l'op-** on peut

---

[9] Ce phénomène de déplacement de la consonne « r » d'une syllabe à l'autre est fréquent chez les enfants (Blanche-Benveniste & Pallaud, 2001 : 22).

considérer que **l'opération** de réintroduction est en en voie d'aboutissement parce qu'on ne en en général hein une une une opération de réintroduction n'est réussie que lorsque il y a des animaux hein qui sont passés par une phase /d', 0/ élevage qui sont retournés au milieu naturel + (crfp, procor1)

G : l'op-
on peut considérer que **l'opération** de réintroduction est en en voie d'aboutissement parce que ...

Dans les corpus étudiés, on trouve peu d'exemples montrant un locuteur mettant en mémoire le syntagme dans son intégralité puis le reprenant d'une façon fragmentée. Dans l'exemple 24, c'est le syntagme *ça me déplairait* qui est mis en mémoire et dans l'exemple 25, *on avait un effectif* qui sera repris ensuite mais d'une façon incomplète, avec une amorce (signalée en gras) :

(24) mais ça me déplairait pas + ça me **dé-** d'autant plus que je suis assez bon en anglais (crfp, prigap2)
(25) on avait un effectif on a eu jusqu'à un **eff-** euh ici sur Moulin on a été jusqu'à soixante-quatorze (crfp, prilil1)

L'amorce porte sur les premiers éléments du morphème et très rarement sur les derniers éléments[10] ; dans ce cas le trait d'union collé à l'amorce se trouve à gauche de l'élément. Dans l'exemple 26, c'est l'interlocuteur (noté L2) qui corrige la production du locuteur (noté L1) et dans les deux exemples suivants, c'est le locuteur qui se corrige lui-même. En 27, c'est un enfant de 10 ans qui rectifie le temps du verbe et en 28, c'est un adulte qui rétablit la consonne initiale :

(26) L1 ça pousse et ça ça fait des des petites **stalactites**
L2 **-mites -mites** qui montent
L1 c'est ça **-mites** qui montent oui des stalagmites (Corpaix, 1995-Champignons)
(27) bon alors je reviens à la question que tu m'avais posée tout à l'heure Isabelle + tu m'as demandé si je **préférais -rerais** chanter en groupe enfin à deux ou à ou toute seule (Corpaix, 1984-Puget1)
(28) des négociations ont d'ailleurs été entamées avec douze pays Chypre Malte et dix pays de l'Europe de l'Est dont la **-ologne** la **Pologne** pardon et la Hongrie (Corpus Delic, 2002-Poitiers)

En dehors d'une analyse syntaxique, des outils peuvent aider le transcripteur à identifier les amorces.

---

[10] Nous n'en avons trouvé aucun exemple dans les 19 corpus étudiés. Nos exemples, fournis aimablement par Paul Cappeau, proviennent d'autres corpus de français parlé.

4. L'identification des amorces

4.1. Le dictionnaire

Le dictionnaire[11] peut nous renseigner sur l'existence du morphème, comme il le fait pour « ovis », indiquant que ce mot regroupe les moutons et les mouflons :

(29) vous savez que le le mouflon c'est un **ovis** c'est un ovin (crfp, procor1)

Mais il faut être prudent : les mots étrangers, le lexique spécialisé, etc. demandent une recherche complémentaire, non seulement dans les dictionnaires mais également auprès des locuteurs qui les emploient.

4.2. Le contexte

L'identification de l'amorce se fait dans le syntagme et en contexte. C'est ainsi que dans les exemples suivants, c'est le contexte qui suggère l'interprétation.

Dans l'exemple 30, c'est l'idée de *ramasser le bois* qui peut faire interpréter l'amorce comme une anticipation de *cueillette* :

(30) il y avait une période où les où les où les paysans avaient le droit d'aller dans les propriétés des forêts des seigneurs pour faire la **cue-** le bois pour ramasser le bois pour etc. (crfp, pricor1)

Dans l'exemple 31, l'amorce *vi-* occupe la place d'un nom et le contexte de *paysans* peut la faire interpréter comme une amorce de *village* :

(31) et la moitié du **vi-** des **pay-** de nos de de de nos **paysans** eh habitaient tantôt à Baste- à la plaine tantôt à la montagne (crfp, pricor1)

Mais le morphème peut apparaître dans le contexte antérieur et servir d'amorce pour un morphème qui sera corrigé ensuite. Ainsi dans les deux exemples suivants, les locuteurs prononcent la première syllabe d'un morphème produit précédemment, s'en avisent et corrigent le groupe syntagmatique en abandonnant l'amorce pour un autre morphème. Dans l'exemple 32, l'amorce *ci-*, renvoyant au nom *cimetière*, est corrigée par *clocher* et dans l'exemple 33, l'amorce *es-*, renvoyant au nom *escalier*, est corrigée par *chemins-là* :

---

[11] Nous conseillons l'usage du TLF (*Trésor de la Langue Française*) dont il existe une version informatisée.

(32) ce **cimetière** le premier le plus ancien a été inauguré en mille neuf cent trente + un peu à l'époque où on a construit le **ci-** le clo- le clocher de l'église (crfp, pricor1)

(33) on a fait le chemin de l'Inca qui en fait qui est un un **escalier** plus tu descends tu montes qui est un escalier vraiment très tortueux abrupt + et quand on pense que les Incas se se se déplaçaient que par ces **es-** ces chemins-là chargés de de blé de grains de marchandises (crfp, prilem1)

4.3. La connaissance du lexique

L'identification de l'amorce suppose également de la part du transcripteur une connaissance du lexique : des fréquences et de son organisation (champs lexicaux, séries, lexèmes substituables).

4.3.1. Les fréquences

Si les amorces ont une forte probabilité d'apparition dans le syntagme, comme *même* (ex. 34), *plutôt* (ex. 35) ou *déjà* (ex. 36), le transcripteur les identifie facilement :

(34) notre travail est aussi bien fait que si on avait constamment le patron + parce que **m-** en sachant qu'il y a quelqu'un derrière qui regarde ton travail + moi personnellement ça me s- ça m'énerve + c'est pas mon c'est pas mon truc (crfp, pricle3)

(35) il y a euh huit assistants de clientèle qui eux sont **plut-** plus spécialement chargés de du service clientèle aux guichets (crfp, progre1)

(36) d'ailleurs certains qui avaient des exploitations d'arbres fruitiers se sont **déj-** reconvertis vers la vigne (crfp, prigap1)

4.3.2. Les champs lexicaux

Quand l'amorce est corrigée par un autre morphème dans le syntagme, le transcripteur peut établir un rapport de sens étroit entre les éléments syntagmatiques. Ainsi, dans l'exemple 37, le transcripteur peut identifier l'amorce *é-* comme un fragment de *édifices, établissements* ... en rapport de sens avec *monuments* et dans l'exemple 38, l'amorce *en-* comme un fragment de *entreprise* ... en rapport de sens avec *banque* :

(37) bien sûr euh nous avons euh des monuments exceptionnels à voir en Bourgogne des **é-** des des **monuments** qui qui partent bien sûr euh du du roman (crfp, pridij1)

(38) et euh + l'évolution à l'intérieur de la de l'**en-** de de la de la **banque** comment comment elle s'effectue (crfp, progre1)

4.3.3. Les séries

L'identification peut se faire grâce à la connaissance qu'a le transcripteur des séries lexicales, comme ici, celle des jours de la semaine :

(39) et puis le **vendre-** le **samedi** quand je peux parce que avec Thomas à emmener au foot souvent je suis contrainte à jouer au taxi avec lui là je là c'est l'atelier (crfp, prilar1)

4.3.4. Les substitutions

Le transcripteur doit aussi connaître les lexèmes qui sont substituables, comme dans l'exemple suivant où *der-* est l'amorce probable de *dernière* car dans ce contexte, *dernière guerre* et *deuxième guerre* sont interchangeables. C'est au transcripteur que revient la charge d'interroger les locuteurs sur ces substitutions, afin de ne pas passer à côté de l'identification de certains morphèmes :

(40) il produisait encore cinquante mille hectolitres à la fin de la **der-** de la **deuxième guerre mondiale** (crfp, progap1)

4.4. La connaissance des différentes perceptions

La perception du transcripteur n'est pas forcément celle des autres locuteurs. Ainsi un transcripteur du sud de la France qui a l'habitude de prononcer le *e* final, situé après consonne, des verbes du premier groupe, pourrait transcrire l'exemple suivant avec une amorce parce que le *e* n'est pas audible (ex. 41a) :

(41a) et puis on le **veil-** on le veille jusqu'à minuit (crfp, pricor1)

Il est préférable de ne pas amputer la syllabe de son *e* muet et de conserver l'écriture complète du morphème (ex. 41b) :

(41b) et puis on le **veille** on le veille jusqu'à minuit

4.5. La connaissance des prononciations

Le transcripteur doit aussi se méfier de sa prononciation. Nous savons combien il est difficile de savoir comment se prononcent certains noms terminés par *s*, *l*, *z* ou *x*. Doit-on dire pour *cassis* : /kasis/ ou /kasi/ ? Le dictionnaire indique que la consonne finale est prononcée alors que pour *persil*, elle ne l'est pas (/pErsi/). Que faire alors devant l'exemple suivant, transcrit ainsi ?

(42a) euh hum l'Aligoté aussi a un vin qui + qui aura euh + un un aven- un a- un avenir certain dans dans des plats euh surtout dans des entrées ou des apéros euh comme le **cassi-** avec le **cassis** il s'accorde très bien avec le **cassis** (crfp, pubdij1)

Doit-on garder l'amorce (ex. 42a) puisque le locuteur prononce /kasis/ deux fois par la suite ou doit-on rétablir l'écriture du morphème, tout en indiquant la particularité de prononciation à un autre endroit (ex. 42b)[12] ?

(42b) euh hum l'Aligoté aussi a un vin qui + qui aura euh + un un aven- un a- un avenir certain dans dans des plats euh surtout dans des entrées ou des apéros euh comme le **cassis** avec le **cassis** il s'accorde très bien avec le **cassis**

La prise en compte de ces outils a permis de corriger certaines erreurs de transcription, comme nous allons le voir, mais a soulevé d'autres problèmes.

5. Les problèmes rencontrés dans les corpus concernant les amorces

5.1. Quelle orthographe ?

La prise en compte du syntagme dans lequel est inscrite l'amorce nous a amené à corriger l'orthographe de certaines amorces, la mise en grille visualisant l'amorce et le syntagme qui lui correspond :

(43) c'est sûr que il y a un impact sur le terrain bon l'**in-** l'**in-** l'impact ce sont les le le premier impact ce sont les rejets directs (crfp, procor1)

(ex. 43 corrigé) c'est sûr que il y a un impact sur le terrain bon l'**im-** l'**im-** l'impact ce sont les le le premier impact ce sont les rejets directs

G : bon l'**im-**
l'**im-**
l'impact ce sont ...

En 44, la prise en compte du contexte montre que l'amorce n'est qu'un fragment du nom propre qui suit, d'où la majuscule dans l'exemple corrigé :

---

12 Il n'est pas rare qu'un même locuteur ait les deux prononciations : /kasi/ et /kasis/ ou /pErsi/ et /pErsil/.

(44) c'est-à-dire + aient une organisation très régionalisée comme l'**a** l'Allemagne bien entendu depuis longtemps (crfp, prigap2)
(ex. 44 corrigé) comme l'**A-** l'Allemagne bien entendu
G : aient une organisation très régionalisée comme l'**A-**
l'Allemagne ...

Dans l'exemple qui suit, on peut penser que *con-* est l'amorce de *compétences*, occupant la même place syntaxique :

(45) il faut qu'il sorte de l'univer- qu'il ait été formé ici c'est normal + euh euh donc après et après qu'il ait eu une formation corse + sur les **con-** à à compétences égales (crfp, pubcor1)
(ex. 45 corrigé) il faut qu'il sorte de l'univer- qu'il ait été formé ici c'est normal + euh euh donc après et après qu'il ait eu une formation corse + sur les **com-** à à compétences égales
G : et après qu'il ait eu une formation corse sur les **com-**
à
à compétences égales

Mais nous nous rendons compte, au travers de ce dernier exemple, qu'il n'y a pas toujours de certitude mais seulement une probabilité de solution reposant sur des critères formels.

5.2. Amorce ou morphème ?

L'analyse en syntagmes a permis de corriger des amorces qui n'en étaient pas. Ainsi dans l'exemple qui suit, il s'agit d'un cas banal de répétition du déterminant. Pallaud et Henry (dans ce numéro) montrent que les déterminants sont majoritairement concernés par la répétition (dans plus de 80% des cas) :

(46) hein des + enfin parlons pas des grands des des grands crus mais + c'est l'équivalent de + à quoi peut-on le comparer dans l'arrière pays de Montpellier les les Costières les non pas les Costières le + zut le nom m'échappe (...) enfin des + **dé- dé-** des produits sur ces coteaux qu'il y a derrière Montpellier là (crfp, prigap1)

(ex. 46 corrigé) enfin des + **des des** des produits sur ces coteaux qu'il y a derrière Montpellier là

G : hein des
[...]
enfin des
des
des
des produits sur ces coteaux qu'il y a derrière Montpellier là

Le fait que le locuteur revienne, après la production de l'amorce, au départ du groupe syntagmatique aide à l'interprétation et à la correction des exemples qui suivent.

Dans l'exemple 47, le verbe *avoir* est employé comme verbe générique, occupant la place du lexème verbal qui sera produit ensuite[13] (*qui a / qui existe*) :

(47) quand euh quelqu'un mourait euh au village ou à la plaine eh il fallait l'enterrer alors d'abord il y avait déjà la confrérie la confrérie qui **a-** qui existe depuis des temps immémoriaux qui prenait en charge le mort (crfp, pricor1)

(ex. 47 corrigé) il y avait déjà la confrérie la confrérie qui **a** qui existe depuis des temps immémoriaux
G : il y avait déjà la confrérie la confrérie qui **a**
qui existe depuis des temps ...

Dans l'exemple 48, il est probable que le locuteur reprend le syntagme verbal en remplaçant *falloir* par un synonyme *il y a ...+ infinitif* :

(48) ils mettent en place pour les enfants euh bon des activités randonnée(s) des activités escalade(s) bon et puis **fo-** il y a les aménagements à proprement dit hein à à surveiller et à à et à entretenir sur les sites même (crfp, procor1)

(ex. 48 corrigé) bon et puis **faut** il y a les aménagements à proprement dit hein à à surveiller et à à et à entretenir sur les sites même
G :bon et puis **faut**
il y a les aménagements à proprement dit hein à
à surveiller
et à
à
et à entretenir
...

Dans l'exemple 49, le locuteur amorce une question en *qu'est-ce que* puis produit une sorte d'énoncé parenthétique avant de reprendre le syntagme en *qu'est-ce que*, comme le montre la mise en grille :

---

13 Ces verbes génériques ont été appelés « verbes d'approximation » (cf. Roubaud, 1987).

(49) et et dans le domaine du travail **què-** vous qui avez connu disons en tant que fils de vigneron les les la manière de travailler il y a cinquante ans et qui la connaissez maintenant qu'est-ce que vous + donneriez comme différence principale (crfp, prigap1)

(ex. 49 corrigé) et et dans le domaine du travail **qu'est-** vous qui avez connu disons en tant que fils de vigneron les les la manière de travailler il y a cinquante ans et qui la connaissez maintenant qu'est-ce que vous + donneriez comme différence principale

G : dans le domaine du travail **qu'est-**
[vous qui avez connu disons en tant que fils de vigneron les les la manière de travailler il y a cinquante ans et qui la connaissez maintenant]
qu'est-ce que vous donneriez comme ...

5.3. Cas de double-transcription

Mais il nous a aussi parfois été impossible de décider si nous nous trouvions face à un morphème ou une amorce parce que nous hésitions sur le départ du groupe syntagmatique. Dans ce cas nous avons opté pour une double-transcription : l'une avec un morphème complet et l'autre avec seulement l'amorce d'un morphème, les deux éléments étant placés entre barres obliques.

Tout compte de nombreux allomorphes, aussi peut-on se demander si dans l'exemple transcrit (ex. 50a), il s'agit bien de l'amorce de *toute*, lexème produit un peu plus loin ou s'il s'agit du morphème *tous*, réalisé dans le syntagme *tous les occitans* (ex. 50b) :

(50a) au cirque des corbeaux il y avait Marti il y avait **tou-** il y avait **toute** la + tous les occitans qui chantaient là il y avait vraiment la liaison sur le le Larzac (crfp, prigap2)

(50b) au cirque des corbeaux il y avait Marti il y avait **tous** il y avait toute la + **tous** les occitans

La solution, pour le transcripteur, est de présenter une double-transcription (ex. 50c) :

(50c) au cirque des corbeaux il y avait Marti il y avait /**tou-, tous**/ il y avait toute la + tous les occitans

Que décider pour l'exemple 51 ? Le locuteur produit-il le verbe *avoir* au présent, comme l'indique la transcription (ex. 51 a) pour se corriger ensuite en *avait* ou a-t-on une amorce de *avait* : *si on t'a- / si on t'avait* (ex. 51b) ?

(51a) mais pourquoi et si si si on **t'a** et si on **t'a** si on t'avait laissé le choix un jour qu'on **t'avait** dit si tu veux tu peux prendre français euh pour euh pour rehausser un peu tu aurais pris français (crfp, pubcor1)

(51b) mais pourquoi si si si on **t'a-** et si on **t'a-** si on t'avait laissé le choix un jour qu'on t'avait dit ...

Encore une fois, la double-transcription laisse la place aux deux interprétations (ex. 51c) :

(51c) mais pourquoi si si si on **t'/a, a-/** et si on **t'/a, a-/** si on t'avait laissé le choix un jour qu'on t'avait dit ...

Dans l'exemple (52), le transcripteur peut hésiter entre la production d'un adverbe dans la rection du verbe *faire manger* ou l'amorce du syntagme en *il fallait*, les deux mises en grille (G1 et G2) visualiseront ces deux possibilités :

(52) et les hommes sont dans les pièces à côté ou chez les voisins s'il n'y a pas assez de place et alors ceux-là **/i-, il/** on les fait manger comme si c'étaient des gens qui étaient arrivés de très loin et qu'il fallait nourrir pour qu'ils puissent euh passer la nuit (crfp, pricor1)

G1 alors ceux-là **i-** (immédiatement) on les fait manger comme si ...

ou

G2 alors ceux-là **il**
et qu' **il** fallait nourrir pour qu'ils puissent euh passer ...

Dans l'exemple 53, le transcripteur peut hésiter entre la production de l'indéfini *tout* comme élément de la liste des syntagmes infinitifs (G1) ou comme adverbe portant sur le verbe *répondre* (G2) :

(53) les six conseillers commerciaux sont là essentiellement pour provoquer les contacts avec les clients répondre à leurs attentes + **/tout, tou-/** euh leur proposer des placements d'argent (crfp, progrep1)

G1 les six conseillers commerciaux sont là essentiellement
pour provoquer les contacts avec les clients
répondre à leurs attentes
tout
leur proposer des placements d'argent

ou

G2 les six conseillers commerciaux sont là essentiellement
pour provoquer les contacts avec les clients
répondre à leurs attentes **tou-** (toujours)
leur proposer des placements d'argent

Dans l'exemple 54, le transcripteur peut hésiter entre la retouche de la préposition dans le syntagme prépositionnel *à ce moment-là* (G1) ou envisager une mise en mémoire (plus longue que celle que nous avons vue

dans les exemples précédents) du lexème *octobre* produit un peu plus loin (G2) :

(54) les maladies cryptogamiques quoi + bon là c'est environ un traitement tous les quinze jours à peu près + pendant bon + surtout le mois de juin et juillet (...) c'est le plus important le plus important c'est **/au, o-/** à ce moment là que ça craint le plus (...) et puis on arrive aux vendanges vers le mois de fin octobre (crfp, prigap1)

G1 le plus important c'est **au**
à ce moment là que ça craint le plus (...)

ou

G2 le plus important c'est **o-** à ce moment là que ça craint le plus (...)
et puis on arrive aux vendanges vers le mois de fin **octobre**

5.4. Cas où l'identification est plus difficile

Mais il arrive que le transcripteur ait beaucoup de mal à identifier l'amorce pour différentes raisons.

5.4.1. Amorce isolée

Dans certains cas, le syntagme contenant l'amorce n'est pas repris par le locuteur si bien qu'il est difficile d'interpréter quoi que ce soit, c'est le cas de *sans le so-* (ex. 55) :

(55) maintenant pour la langue bon il y a quelques cours par ci par là mais encore une fois + sans le **so-** c'est c'est in- c'est insuffisant (crfp, prigap2)

5.4.2. Discours difficile à percevoir

Si le discours est coupé par les interventions d'un interlocuteur (noté L1), par des chevauchements de paroles (indiquées entre chevrons), la tâche du transcripteur est encore plus difficile, c'est ce qu'illustre la transcription de l'amorce *contè-* dans cet extrait :

(56) L2 j'ai quand mê- j'ai enregistré personnellement des contes j'ai
L1 ah < d'accord
L2 j'ai > au moins deux cassettes
L1 < ah bé oui
L2 de **contè-** > au- conteuses aujourd'hui /deux C.D., décédées/
L1 ah c'est bien
L2 tout à fait d'un d'un bon niveau sur les contes traditionnels en terme de de ça il y en a

L1 ça ça devrait être édité tu penses pas
L2 à ma connaissance il y en a peu de il y en a peu >
L1 l'éditer

Si l'on se réfère au contexte et à la place occupée par cette amorce, il est plus probable d'envisager deux morphèmes à la place de l'amorce :

(ex. 56 corrigé) j'ai enregistré personnellement des contes j'ai (...) j'ai au moins deux cassettes (...) de **contes et** au- conteuses aujourd'hui /deux C.D., décédées/

De la même façon, dans l'exemple qui suit, il nous a semblé possible d'envisager autre chose que ce qui était proposé, l'amorce *almal-* (ex. 57a) suggérant que le locuteur commence le syntagme *les âmes damnées* par une amorce anticipant sur la voyelle « A » et la consonne « L », comme le visualise la mise en grille :

(57a) derrière le vieux couvent il y avait + trois euh nécropoles si tu veux + euh euh il y avait une nef pour euh pour les + centrale pour les religieux + ceux qui étaient très proches de l'église une à l'est c'était pour les chrétiens et puis celle de gauche c'était pour les les **almal-** les âmes damnées les gens qui étaient morts euh en mal de religion etc. (crfp, pricor1)

G : et puis celle de gauche c'était pour les
les **almal-**
les âmes damnées
les gens qui étaient morts ...

Mais ne peut-on pas envisager la production de deux parties de syntagmes (ex. 57b), celui contenant le mot *âmes* avec anticipation d'un /l/ à la place d'un /m/ et celui contenant l'adverbe *mal*, anticipant sur le syntagme à venir *en mal de religion* ? Nous aurions alors, à la place d'une seule amorce, un fragment de morphème et un morphème. Cette possibilité d'interprétation sera visualisée dans la grille sous l'exemple :

(57b) et puis celle de gauche c'était pour les les **al- mal** les âmes damnées les gens qui étaient morts euh en mal de religion etc.

G : c'était pour les
les **al-** **mal**
les **âmes damnées**
les gens qui étaient morts euh en **mal** de religion

5.4.3. Discours spécialisé

Dans les corpus, l'identification des morphèmes et donc des amorces, peut être rendue difficile parce que le transcripteur se trouve face à un discours de spécialiste, qu'il ne connaît pas forcément et dont le lexique

spécialisé sera absent de son dictionnaire. Si le locuteur explique les termes qu'il emploie, le transcripteur n'a plus aucun problème pour identifier les morphèmes ; c'est le cas de *l'ouvrée* dans cet exemple, dont la définition est donnée dans le syntagme en *c'est-à-dire que* :

(58) hein nous on compte à **l'ouvrée** + c'est-à-dire que dans un hectare il y a vingt-quatre ouvrées c'est quatre cent vingt-huit mètres carrés exactement l'ouvrée quand on vend des vignes en Bourgogne eh bien on parle de on parle de l'ouvrée + euh quatre cent vingt-huit mètres carrés donc à l'hectare au je pense que ça c'est facile à à se rappeler (...) + on vend à l'ouvrée alors l'ouvrée eh bien pourquoi parce que c'est ce que faisait dans la journée euh il y a une centaine d'années ou peut-être même moins euh à la main + un un ouvrier qui travaille dans à la tâche dans dans dans la vigne alors voilà c'est pour ça c'est le travail d'une journée hein le l'ouvrée de l- de l- l'ouvrage de la de la de la journée + (crfp, pridij1)

Mais il arrive aussi que le corpus ne renseigne pas le transcripteur. En effet, à quoi correspond l'amorce *éti-* de ce dernier exemple ?

(59a) L2 et puis vers la f- au quinze juin à peu près c'est ce qu'on appelle la floraison c'est-à-dire au ça se passe très rapidement cette floraison ça se passe très rapidement + il f- il faut qu'elle se passe très rapidement pour que chaque chaque grain devienne un raisin quoi si-
L1 hum
L2 sinon bon c'est ce qu'on appelle la coulure s'il pleut ou que c'est /dans, la/
L1 oui
L2 la pollinisation si le pollen est mouillé **éti-** + il vole pas de d'un grain à l'autre
L1 oui c'est
L2 c'est-à-dire qu'il part dans l'air et puis + ça ça pollinise pas quoi (crfp, prigap1)

La transcription (ex.59a) ne livre qu'une interprétation, difficile d'ailleurs pour des non-spécialistes de la vigne, le locuteur essayant d'expliquer que la coulure est un accident qui empêche la fécondation de la fleur et qui fait le plus souvent couler le pollen. Transcrite de cette manière, l'amorce peut s'interpréter comme étant celle d'un adjectif (comme *étiolé*), produit sur la même place syntaxique que *mouillé* dans le syntagme *si le pollen est mouillé*, avec une mise en grille de ce type :

G : si le pollen est mouillé
éti- il vole pas de d'un grain à l'autre

Mais ne peut-on pas envisager une autre transcription, qui ferait apparaître une coordination de deux adjectifs dans la valence de *être* : par

exemple : *si le pollen est mouillé et tiède* (ou *mouillé et tigré* ou *mouillé et tiqueté* ...) ? La mise en grille serait alors différente ainsi que la transcription (ex. 59b) :

(59b) L2 la pollinisation si le pollen est mouillé **et ti-** + il vole pas de d'un grain à l'autre

G si le pollen est mouillé
et **ti-** il vole pas de d'un grain à l'autre

Il aurait pu s'agir également d'un *et qu'il* mal perçu par un locuteur du nord de la France (ex. 59c)[14] et nous n'aurions alors aucune amorce :

(59c) L2 : la pollinisation si le pollen est mouillé **et qu'il** + il vole pas de d'un grain à l'autre

Comme l'écrivent Blanche-Benveniste & Jeanjean (1987 : 112) :

> « Dans toutes les écoutes, aussi méticuleuses qu'elles soient, il y a toujours une part de flottement et d'approximation. »

6. Conclusion

Les amorces ne sont pas seulement des traces d'oralité, elles sont également riches d'informations car elles nous renseignent

- sur le plan syntaxique : sur la composition de la syntaxe avec les allées et venues qu'effectue le locuteur sur l'axe syntagmatique et les recherches paradigmatiques qu'il opère (quels lexèmes ? quelle place ? quelle correction ?) ;
- sur le plan sémantique : sur la progression du sens (le locuteur anticipant le plus souvent sur le morphème à venir), sur ce que veut dire le locuteur ;
- et sur le plan énonciatif : sur les phénomènes de mise en mémoire, de savoirs partagés ou non.

La tâche du transcripteur pour identifier les amorces est complexe. Il doit non seulement se référer à des données disponibles (dictionnaires, contexte) mais aussi se construire ses propres outils (connaissance du lexique, des prononciations, des perceptions) et se documenter dans le cas de morphèmes inconnus ; la connaissance d'un transcripteur isolé ne suffit pas. Transcrire un corpus nécessite d'en effectuer une analyse syntaxique.

---

[14] Je remercie José Deulofeu pour cette remarque.

Une première transcription, même faite par un transcripteur averti, ne donne qu'une première analyse du texte en syntagmes, qu'une première interprétation. Coppieters (1997 : 35) indiquait déjà que l'interprétation d'un corpus demeurait toujours incomplète et donc en attente d'interprétation. Le transcripteur doit ensuite affiner la transcription au niveau des syntagmes et des morphèmes en revenant sur les amorces signalées, en réécoutant les passages difficilement interprétables (ce qui augmente encore le temps qu'il faut pour transcrire une minute de parole)[15].

Une bonne transcription ne peut se faire qu'à ce prix car nettoyer les amorces des textes serait perdre toutes ces précieuses informations et conduirait à produire un autre texte, qui risquerait de ressembler à un texte écrit mais qui, en fait, aurait peu de chose à voir avec le discours du locuteur.

Références

Bilger, M., Blasco, M., Cappeau, P., Pallaud, B., Sabio, F., & Savelli, M.-J. (1997). Transcription de l'oral et interprétation ; illustration de quelques difficultés. *Recherches sur le français parlé, 14*, 57-86.

Blanche-Benveniste, C. (1997). *Approches de la langue parlée en français*. Paris : Ophrys.

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

Blanche-Benveniste, C., Bilger, M., Rouget, C., & van Den Eynde, K. (1990). *Le français parlé. Études grammaticales*. Paris : Édition du CNRS.

Blanche-Benveniste, C., & Pallaud, B. (2001). Le recueil d'énoncés d'enfants : enregistrements et transcriptions. *Recherches sur le français parlé, 16*, 11-37.

Cappeau, P. (1997). Données erronées : quelles erreurs commettent les transcripteurs ? *Recherches sur le français parlé, 14*, 117-126.

Coppieters, R. (1997). Quelques réflexions sur la question des données : corpus et intuitions. *Recherches sur le français parlé, 14*, 21-41.

Giovannoni, D.-C., & Savelli, M.-J. (1990) Transcrire et orthographier le français parlé. De l'impossible copie à la falsification des données orales. *Recherches sur le français parlé, 10*, 19-37.

Jeanjean, C. (1984). Les ratés c'est fa- fabuleux. Étude syntaxique des amorces et répétitions. *Syntaxe et discours LINX, 10*, 171-177.

Pallaud, B. (1999). Lapsus et phénomènes voisins dans la langue parlée : problèmes d'identification. *Recherches sur le français parlé, 15*, 1-33.

Pallaud, B. (2002). Les amorces de mots comme faits autonymiques en langage oral. *Recherches sur le français parlé, 17*, 79-101.

[15] L'écoute du corpus avec un casque facilite la perception des amorces.

Roubaud, M.-N.(1987). *Les mystères de l'approximation lexicale*. Mémoire de DEA. Sciences du Langage, Université de Provence.

Roubaud, M.-N., & Loufrani, C. (1999). Éclairage des corpus de type aphasique par les grilles. *Recherches sur le français parlé*, *15*, 41-57.

Roubaud, M.-N. (2003). Du bon usage des notations phonétiques dans les corpus aphasiques. *Recherches sur le français parlé*, *18*.