



**HAL**  
open science

## La transcription comme annotation

Gabriel Bergounioux

► **To cite this version:**

Gabriel Bergounioux. La transcription comme annotation. Ph. Caron, R. Defiolle, M.-H. Lay. L'enjeu des métadonnées dans les corpus textuels : un enjeu pour les sciences humaines, Presses Universitaires de Rennes, 2019. halshs-03058628

**HAL Id: halshs-03058628**

**<https://shs.hal.science/halshs-03058628>**

Submitted on 11 Dec 2020

**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.

## La transcription comme annotation

Gabriel Bergounioux Université d'Orléans / LLL

### Introduction

Quelles transformations se produisent au moment de convertir en textes des enregistrements, en particulier lorsqu'il s'agit, comme en français, d'une langue à tradition écrite ? En adoptant une perspective épistémologique, on a choisi de centrer l'observation sur les étapes initiales du traitement. Ces réflexions font suite à une série de discussions au sein du Laboratoire Ligérien de Linguistique avec Michel Jacobson, à des échanges avec Paola Pietrandra, Iris Eshkol, Pascal Cordereix et Olivier Baude. La question en débat est la suivante : dans la constitution des corpus oraux, comment doit s'effectuer la répartition des tâches entre la transcription, l'annotation et les métadonnées ? De la transcription à l'annotation, de quelle nature est l'enrichissement que réalise *de facto* la transcription – surtout quand le texte obtenu en vient à se substituer à l'écoute de l'enregistrement ? On montrera que c'est en fonction de ce que marque la conversion scripturale que s'effectuera le partage avec ce qui devra être reporté dans les annotations.

Par ailleurs, la distribution des informations entre les annotations et les métadonnées conduit à distinguer deux catégories : (i) des indications fournies pour faciliter l'accès au contenu qui sont directement déductibles des données (langue, durée, thème...) et (ii) des renseignements requis par le catalogage (date et lieu, matériel d'enregistrement utilisé, enregistrements associés...) ou par l'exploitation (qualités sociales du locuteur et de l'enquêteur...).

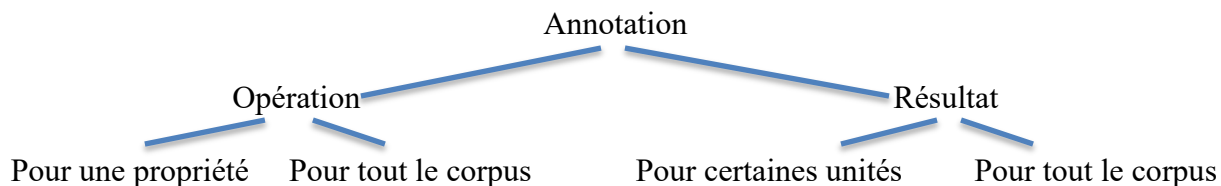
On se propose d'examiner dans quelle mesure la transcription constitue, par elle-même et en tant que telle, une annotation, ce qui suppose un retour sur sa définition afin de restituer à l'écriture sa fonction d'outil. Un usage quotidien, routinisé depuis des siècles, ferait oublier que les effets de la « raison graphique » [Goody, 1979] dans l'élaboration des données ne sont pas si différents de ceux qui résultent de l'utilisation de certains métalangages informatiques sans y être non plus réductibles. En conclusion, on s'interrogera sur ce que signifie la production des métadonnées dans ce dispositif.

#### 1. ESLO : transcription, annotation, métadonnées

Comme beaucoup de noms dérivés d'un verbe au moyen du suffixe *-tion* (du type *construire* > *construction*), *annotation*, tel qu'il est employé en linguistique de corpus, peut recevoir deux interprétations. S'il est compris comme un nom d'action, il correspond à la forme active du verbe et présuppose un agent (un sujet) qui exécute la tâche : l'*annotateur*. Si c'est le résultat qui est envisagé, il s'agit d'un nom d'objet qualifié de façon passive : un document *annoté*. Un énoncé comme « l'annotation du corpus » peut correspondre indifféremment à l'une ou l'autre de ces interprétations.

Chacune des acceptions peut recevoir à son tour une valeur tantôt distributive, tantôt collective. On opposera ainsi, pour l'opération, le fait d'annoter un élément parcellaire et discontinu (par exemple les formes modales [Degand, Cornillie & Pietrandrea, 2013] ou les coréférences [Muzerelle *et al.*, 2012]) et l'annotation de propriétés impliquant tous les éléments du corpus (marquage en POS [Benzitoun, Fort & Sagot, 2012]). On pourrait avancer l'hypothèse que, dans la seconde catégorie, une notation phonétique s'apparenterait à une certaine forme d'annotation : pour chaque séquence, les variations de la production verbale sont consignées en dissociant, selon leur réalisation, les occurrences d'un même mot [Bergounioux, 2008], sans pour autant que soit introduit de commentaire explicite. La distance entre la forme orthographique et la prononciation orthoépique doit être recalculée au fur et à mesure par le lecteur.

Concernant le résultat, on retrouve cette différence d'approche entre l'annotation d'un segment ou d'un type de propriété vs l'annotation globale du corpus.



*Distribution des acceptions d'annotation en linguistique de corpus*

À partir de l'expérience acquise à travers le programme de l'Enquête Sociolinguistique à Orléans [Bergounioux 2016], désormais ESLO, on se propose de revenir sur la chaîne des opérations qui sont constitutives de l'annotation. ESLO représente aujourd'hui le plus grand corpus de français parlé (7.500.000 mots) accessible en ligne. Les données sont établies en conformité avec les préconisations des acteurs majeurs du domaine, – en France, la TGIR Huma-Num (à travers le consortium CORLI), la BnF, l'Equipex Ortolang et CoCoON. La création et la mise en forme de cette ressource, entreprises depuis une dizaine d'années sous la direction d'Olivier Baude, ont mobilisé une trentaine de participants de l'université d'Orléans et du service des documents sonores du département de l'audiovisuel de la BnF. Des centaines de personnes ont été enregistrées et un module spécial, « Langues en Contact à Orléans », est consacré à la diversité linguistique dans l'agglomération. L'analyse sous-jacente est résolument variationniste. Les données sont contrastées en fonction de critères diachronique (une première enquête, dite ESLO1, a été conduite en 1969-1971, la seconde, ESLO2, est en incrémentation continue depuis 2007) et sociaux, avec une diversité des témoins (par âge, par sexe, par niveau d'étude, par situation professionnelle), des langues, des situations d'enquête, des enquêteurs et des transcrip-teurs.

L'une des originalités du corpus est que chaque enregistrement fait l'objet de trois transcriptions codées A, B et C :

- la première, A, se fixe pour objectif de segmenter les tours de parole et de découper des séquences qui correspondent à une proposition, un groupe de souffle, une incise, i.e. à des unités syntagmatiques d'extension plus large que le mot, tout en conciliant brièveté et cohérence grammaticale et discursive.
- la seconde, B, apporte les corrections et compléments nécessaires et finalise l'anonymisation, en balisant les passages qui ne pourront être accessibles au public (l'intégralité du document reste disponible pour des usages de recherche sous réserve d'une demande particulière et de la signature d'une convention spécifique).
- la troisième, C, avant que ne soit décidée la validation de mise en ligne, réalise de nouvelles rectifications en veillant à la qualité orthographique et à la conformité aux normes de présentation (par exemple, la suppression de tout signe de ponctuation à l'exception du point d'interrogation).

Les trois versions sont conservées et accessibles à partir du catalogue. Elles sont définies comme un moyen d'analyser une variation d'écoute dont l'étude est menée en ajoutant aux explications les caractéristiques des transcrip-teurs en plus des incertitudes provoquées par une qualité sonore déficiente ou par les indé-cisions homophoniques. Une thèse exploitant ces données est en train d'être rédigée par Linda Hriba.

## La transcription comme métadonnée

Enregistrement_N	Code_exemples	A	B	Variations_type
78	ESLO1_ENT_078_1		HN 169: qui était sur le sol français en difficulté qui venait pour nous libérer #1 et je tiens	VS
78	ESLO1_ENT_078_2		à #	VS
78	ESLO1_ENT_078_3		JSM: #2 oui #	VS
78	ESLO1_ENT_078_4	HN 169: qui était sur le sol français en difficulté qui venait pour nous libérer et je tiens à leur en rendre au passage un hommage ému et donc à ce moment-là euh j'ai été en contact avec beaucoup d'équipage de la RAF dont j'ai conservé un souvenir vous pensez ému et j'en ai conservé euh des liai- des liaisons avec euh la une société que ils ont fondé après la guerre la société Escaping à Londres	HN 169: leur en rendre au passage euh un hommage ému et donc à ce moment-là euh j'ai été en contact avec beaucoup d'équipage de la RAF dont j'ai conservé un souvenir vous pensez ému et j'en ai conservé euh des liai- des liaisons avec euh la une société que ils ont fondé après la guerre la société Escaping à Londres	VS
78	ESLO1_ENT_078_5	JSM: oui	JSM: oui [rire]	VG
78	ESLO1_ENT_078_6	JSM: mais enfin je voudrais savoir d'abord euh depuis combien de temps vous habitez euh Orléans	JSM: mais enfin je voudrais savoir d'abord euh depuis combien de temps vous habitez euh Orléans ?	VG
78	ESLO1_ENT_078_7	JSM: et est-ce qu'il y a quelque chose en particulier qui vous a amené à Orléans ?	JSM: et est-ce qu'il y a quelque chose en particulier qui vous a amené à Orléans euh ?	VP
		HN 169: et bien c'est surtout euh ma femme puisqu'elle est originaire	HN 169: et bien c'est surtout euh ma femme puisque elle est originaire d'Orléans	

### *Comparaison de transcriptions (d'après L. Hriba)*

Parallèlement, un double jeu de métadonnées est constitué dans des fichiers séparés. L'un, interne au laboratoire et de diffusion très restreinte, réunit l'ensemble des identifiants ; l'autre, public, avec un codage aléatoire du témoin, regroupe les indications accessibles en consultation. La répartition se fait en veillant à fournir un maximum d'informations pour l'exploitation tout en préservant l'anonymat des témoins.

### Référence enregistrement: ESLO2\_ENT\_1055

Fichier son:	ESLO2_ENT_1055.wav
Corpus:	ESLO2
Catégorie:	Entretien
Précisions sur la catégorie:	Discussion en face à face entre un chercheur et un locuteur témoin à partir d'une trame d'entretien
Sujet:	(text_and_corpus_linguistics) Français (Ethnologue: fra)
Sommaire:	
Editeurs:	LLL Université d'Orléans
Créateurs:	LLL Université d'Orléans – ESLOs
Chercheurs:	Baude, Olivier
Chercheurs locuteurs:	Baude, Olivier
Participants:	TQ55 TQ55FEM, TQ55FIL TQ55FIL1
Description des participants:	Nombre d'autres participants : 3 Remarque : femme du témoin et ses deux fils
Descriptions annexes:	
Remarques:	Marantz PMD 661, micros AKG C 417

Fiche modifiée par:	Lkanaan
Date d'enregistrement:	08/06/2010
Droits:	Copyright (c) 2012 Université d'Orléans/LLLFreely available for non-commercial use. This file is licensed under a Creative Commons License.
Format:	(IANA MIME Media Type: audio/x-wav)
Durée:	01:01:00
Acoustique:	Excellente
Précisions acoustiques:	Télé en bruit de fond
Lieu spatial:	Olivet
Lieu TGN:	1034729
Lieu Point:	east=1.899; north=47.873
Locuteurs:	<ul style="list-style-type: none"><li>• <a href="#">ch_OB1</a></li><li>• <a href="#">TQ55</a></li><li>• <a href="#">TQ55FEM</a></li><li>• <a href="#">TQ55FIL</a></li><li>• <a href="#">TQ55FIL1</a></li></ul>
Transcriptions:	<ul style="list-style-type: none"><li>• <a href="#">ESLO2_ENT_1055_A</a></li><li>• <a href="#">ESLO2_ENT_1055_B</a></li><li>• <a href="#">ESLO2_ENT_1055_C</a></li></ul>

*Fichier associé à un enregistrement (ESLO2 locuteur TQ55 Entretien)*  
<http://eslo.huma-num.fr/CorpusEslo/html/fiche/enregistrement?id=1055>

Au nombre des enregistrements inclus dans la ressource, il y a ceux des enquêteurs qui ont été à leur tour soumis à l'interview afin de conserver une trace des acteurs équivalente à celle des témoins.

## 2. Les effets de la transcription dans ESLO

Les conventions de transcription sont détaillées dans le *Guide du transcripteur et du relecteur*, disponible pour téléchargement sur le site ESLO. La mise en forme a été conçue afin d'anticiper une certaine automatisation des tâches, variable selon le type d'objet concerné, en évitant de préjuger d'une interprétation. Ainsi, le relevé des amorces est réalisable d'après la requête : « extraire toutes les séquences de 1 à n lettres consécutives suivies d'un tiret et d'un espace », tout comme l'élosion est repérable par l'utilisation de l'apostrophe.

Obligation est faite d'écrire un mot tel qu'il est orthographié en français standard, ce qui permet de recenser toutes ses occurrences, indépendamment de leur prononciation. Ainsi, bien que le pronom « il » précédant un mot commençant par une consonne soit le plus souvent réalisé /i/, le transcripteur est tenu de le noter avec les deux lettres dans tous les cas. De même, c'est une orthographe d'accord normative qui est suivie, sans préjuger de la conscience que peuvent avoir ou non les locuteurs de leur usage. Par principe, tout ce qui permet de suspendre les jugements de l'auditeur (du transcripteur) dans le texte qu'il produit est préconisé. Si une unité n'est pas identifiable (pour cause de bruit, de chevauchement...) ou figurable de façon conventionnelle (clic, raclement de gorge...), elle est renvoyée à une annotation interpolée entre crochets. Il en va de même, en cas de prononciation non

conforme : la forme perçue est reportée séparément. Quelques formes orales itératives – *ben*, *mh...* – sont figurées de façon homogène. Elles sont répertoriées dans une liste séparée. En revanche, ne sont pas rétablies les unités absentes dans le signal qui seraient obligatoires dans un écrit normatif. On écrira, comme c'est le plus souvent perçu, « faut pas », sans mentionner le pronom ni la première partie de la négation. Ceci permet d'éviter l'artefact d'un décompte a posteriori de « ne » dont on ne retrouverait aucune trace dans les enregistrements.

Sur tous ces points, même si l'on restreignait l'inventaire aux langues qui ont en partage l'alphabet latin, dès lors que les orthographes recèlent des caractéristiques vernaculaires (y compris dans le recours aux diacritiques, la segmentation des unités et l'emploi des majuscules ou de la ponctuation), on s'en tiendra au français mais les résultats et la méthode ne sont limités ni à cette seule langue, ni à un seul corpus dans cette langue.

En même temps que la transcription, un premier jeu d'annotations (entendant par là des éléments qui sont ajoutés par le transcripteur sans correspondre à une séquence sonore distincte dans l'enregistrement) est inclus. D'abord, il y a l'indication des tours de parole avec une notation sténographique ou un code pour l'enquêteur, de même pour les autres locuteurs (tous les témoins ne sont pas anonymes, dans le cas d'une interview de personnalité ou d'une conférence publique par exemple).

HS757: ▶ bonjour

XS952: ▶ bonjour

HS757: ▶ excusez-moi de vous déranger est-ce que vous pourriez nous indiquer le chemin

HS757: ▶ pardon pour aller à la mairie d'Orléans ?

HS757: ▶ s'il vous plaît

*Enregistrement de demande d'itinéraire* ESLO2\_ITI\_1172\_C

On a aussi recours à une extension de la typographie avec des symboles spécifiques (crochets, deux points...) qui permettent de mentionner ce qui est conditionné par l'exécution de la tâche : l'indication que tel passage n'a pas pu être compris [pi:pronounce:instantaneous], ou que le signal ne correspond pas à des unités de langage [pi:rire:instantaneous], ou bien que tel mot a été prononcé d'une façon non conforme – dans le cas d'un lapsus par exemple.

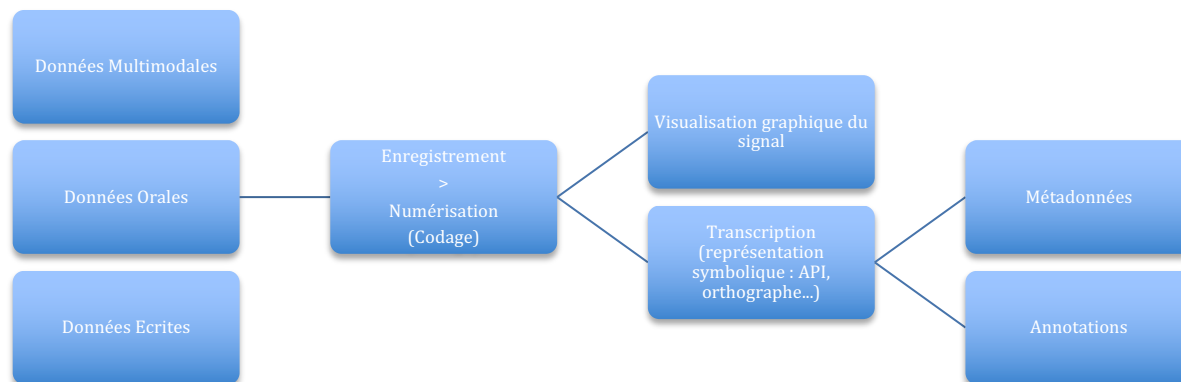
Une tâche supplémentaire est confiée aux transcripteurs : la préparation de l'anonymisation. Il faut pour cela disposer d'une définition de ce qu'est une « entité nommée » afin de pouvoir en relever non seulement les occurrences directes mais également l'ensemble des indications qui permettraient d'identifier la personne (mention de localisation, de profession...) et signaler quels passages devront être bipés dans la version mise en ligne [Maurel, 2011].

Dès ce moment, le choix de rédaction des annotations – en particulier la façon dont elles seront interpolées à l'intérieur de la transcription – et les conventions d'écriture s'ajustent à un format qui sera reproductible dans les traitements ultérieurs afin d'unifier et de limiter les représentations graphiques. Sans exclure d'autres manipulations qui répondraient à des règles et à des attentes différentes, l'annotation et la transcription fixées au départ conditionnent en partie l'enrichissement, l'outillage, l'interopérabilité et le catalogage et les résultats.

Quelle que soit la finalité de l'étude, les annotations sont conçues comme un moyen d'accéder à un certain type d'information en identifiant tous les éléments qui répondent à une caractéristique établie au départ et qui, une fois qu'elles auront été insérées dans le document en suivant un format ad hoc (lui-même déclaré dans les métadonnées), fourniront la réponse attendue à des requêtes ciblées.

### 3. Le signal et la numérisation du corpus

En fonction de ce cadre de réflexion, la question se pose de savoir jusqu'à quel point la transcription constitue effectivement une modalité particulière de l'annotation. On rappellera pour commencer les étapes principales d'une chaîne de traitement avant toute analyse en la figurant de façon simplifiée de la façon suivante.



### *Représentation des phases de constitution d'un corpus oral*

Dans ce schéma ne sont pas prises en considération les opérations effectuées en aval : finalité du corpus, constitution du panel, recrutement et formation des enquêteurs, définition du format d'entretien ou de capture, normes d'enregistrement... De même ne sont pas mentionnées les opérations nécessaires à l'archivage et à la diffusion.

Au départ, il y a le signal sonore. Un signal est une variation temporelle de la pression atmosphérique en un instant donné dans un lieu donné, une succession rapide de surpressions et de dépressions captées par un récepteur (un organe sensoriel, un appareil d'enregistrement). Pour le fixer, depuis un siècle et demi, des techniques variées ont été mises en œuvre. Dès lors que le travail s'effectue à partir d'enregistrements – ce qui est de règle pour tous les corpus oraux et, dès avant, pour les recherches phonétiques –, plusieurs types de supports ont été utilisés depuis les cylindres d'Edison (en faisant abstraction du phonotaugraphe de Scott de Martinville) et les disques de Berliner jusqu'aux serveurs du cloud.

Aujourd'hui, la première manipulation du signal de parole est sa numérisation, i.e. une pérennisation des données sonores sur un support électronique au moyen d'un codage conditionné par le système d'exploitation et les logiciels. A ce stade, la conversion reste contenue dans la réalisation d'un transfert, du signal acoustique à une séquence de symboles en langage machine qui n'est pas directement accessible à l'utilisateur en l'absence d'interface. L'opérateur s'en tient au réglage des appareils pour optimiser le rendu en fonction des attentes de l'utilisateur.

Avec l'irruption de l'informatique, les transformations ont été à la fois qualitatives et quantitatives. En termes de masses de données, les cylindres du phonographe avaient une durée comprise entre deux et trois minutes, les disques ont multiplié cette valeur par vingt-cinq (en format commercial avec l'apparition du microsillon dans les années 1950), les bandes magnétiques ont permis de l'augmenter encore mais c'est avec l'informatique que les données ont connu une extension indéfinie en même temps que l'accès au contenu s'opérait sans manipulation directe d'objets ni intercession d'appareils de lecture. Les indications fournies par des éléments externes (la boîte contenant le cylindre, la pochette du disque, la

fiche associée aux bandes ou aux cassettes...) ont été assurées par la création de fichiers (et de sous-fichiers) distincts mais disponibles sur le même appareil.

Données, annotations et métadonnées se situent désormais au même endroit du point de vue de l'opérateur. Un des avantages est la possibilité d'exploitation de fichiers multimodaux alignant l'enregistrement sonore (voire visuel) et sa transcription, la séparation en plusieurs lignes superposées des annotations et la réalisation d'analyses contrastives entre deux états d'un même document (par exemple un texte et sa traduction ou, comme dans ESLO, les réponses apportées à une même question dans les entretiens ou les différentes transcriptions d'un même enregistrement).

L'accroissement exponentiel des ressources disponibles a requis, entre autres, une harmonisation des enregistrements, des formats et des conventions afin d'assurer, sinon l'homogénéité, du moins l'interopérabilité des dépôts, l'identification des fonds sonores par l'adoption de consignes internationales de catalogage (cf. le programme *Europeana Sounds*) et le repérage des informations dans des dizaines puis des centaines d'heures de parole dont le volume rendait inopérante une recherche manuelle. L'annotation de l'oral, au-delà de la transcription, est une pratique déterminée par la diffusion de l'informatique dont les modalités dépendent des états d'avancement de la technologie. Ces transformations ont des répercussions dans l'organisation de la recherche en linguistique, en particulier l'apparition d'une standardisation et d'une division du travail scientifique dont témoignent la production de guides [Baude, 2006, Mallet-Poujol, 2014, Robin, 2014] et la création de communautés internationales qui se consacrent à la mise au point de normes partagées.

Une conséquence en recherche a été la nécessité de numériser rétroactivement les données sonores consignées sur des supports non informatiques, qui exigent des appareils de lecture obsolètes, difficilement disponibles et pour lesquels les compétences de manipulation disparaissent avec les agents qui en ont eu l'usage [Cordereix, 2012]. Ce travail est conduit par la BnF et les grandes institutions patrimoniales d'archivage sonore à travers le monde avant que ne soit effectué le dépôt sur des plates-formes comme Gallica pour tout public, CoCoON ou Ortolang pour les chercheurs. Ainsi, les bandes magnétiques d'ESLO1 ont été numérisées par la BnF qui en prépare la mise en ligne avec remise d'une copie témoin à l'Université d'Orléans.

Aucun de ces formats, au-delà de leur diversité, n'ajoute d'information à celle qui était déjà présente, selon une autre modalité, dans l'état antérieur. Il ne s'agit que de pérenniser le signal avec tous les aléas qu'implique aujourd'hui l'absence de visibilité, au-delà de la décennie, concernant l'évolution technologique des supports et des systèmes de conservation. Parallèlement, les cadres institutionnels ne sont pas non plus entièrement stabilisés comme le montreraient l'histoire courte mais mouvementée du TGE ADONIS et de l'IR Corpus fusionnés dans Huma-Num ou les difficultés de déclinaison au plan national des infrastructures de recherche européennes CLARIN et DARIAH.

Dans le transfert d'un format analogique au numérique, en dehors des métadonnées, il est possible de filtrer le signal pour améliorer la qualité d'écoute, ou de séparer les propos des différents participants à une discussion en ménageant une sortie sur des canaux séparés, mais il s'agit dans tous ces cas d'enlever du bruit (dans les deux acceptions du terme) et non d'augmenter la quantité de données disponibles comme le font les annotations et, dès avant, la transcription.

#### 4. Les différences entre transcription et annotation

##### 4.1 De la notation à l'orthographe



La transcription est un choix de représentation qui s'efforce de concilier au mieux la fidélité à la source et la compréhension par le lecteur – lui-même orienté par certaines attentes préconstruites – en faisant correspondre à la réalisation sonore initiale une accessibilité visuelle. C'est le résultat d'un arbitrage entre une figuration physique, analogique, du signal (des spectrogrammes de type PRAAT par exemple) et une notation symbolique qui a sa logique propre [Herrenschmidt, 2007]. La projection graphique offre une panoplie de solutions, depuis les notations phonétiques les plus minutieuses jusqu'à l'écriture en logogrammes où le signifiant s'estompe derrière les conventions représentationnelles d'une culture qui a pu choisir une voie où l'exposition reste subordonnée à une *Weltanschauung*, une certaine appréhension à la fois subjective et sociale du monde [Vandermeersch, 2013].

Quelle que soit la restitution optique proposée, elle s'ajuste à des capacités d'interprétation visuelle, autrement dit de lecture. A ce titre, elle effectue un changement de registre sensoriel et fait appel à d'autres fonctionnements cognitifs sous-jacents, eux-mêmes produits par des apprentissages dans un système d'éducation particulier. La transposition des données se fait suivant une mise en forme qui ne leur est pas inhérente mais qui y correspond. On ne traitera en suivant ni de la vocalisation a posteriori du texte (comme pour un air interprété à partir d'une partition) afin de focaliser la réflexion sur la transposition alphabétique, ni des visualisations non linéaires de données, comme la constitution de listes de fréquence, de réseaux de collocations ou la réalisation d'histogrammes.

Dans le travail sur corpus, la *transcription* n'est pas toujours considérée comme une variété d'*annotation*. En suivant le découpage proposé, entre autres, par Leech, Myers & Thomas [1995], on distinguerait deux phases. Une conversion du signal enregistré en un fichier texte est effectuée avant que n'intervienne un retravail sur le texte obtenu afin de l'enrichir, entre autres, au moyen d'annotations (d'autres formes d'enrichissement, par de l'outillage en particulier, n'imposent pas obligatoirement le recours à la transcription).

On se propose d'examiner quelques-uns des arguments qui justifient l'établissement d'une séparation tranchée en restreignant l'analyse au cas du français qui présente les propriétés suivantes :

- la notation se réalise sous une forme alphabétique, avec recours aux caractères spéciaux, aux diacritiques, aux majuscules et à la ponctuation (on n'utilise dans les transcriptions ni chiffres, ni symboles, ni abréviations),
- la figuration est linéaire de gauche à droite et de haut en bas,
- les correspondances phonétiques sont grevées d'adjonctions à valeur morphologique et étymologique.

On fera abstraction de l'investissement social, en particulier de l'imposition scolaire qui donne à l'écriture normée du français une autorité symbolique particulière dont témoigne la virulence des débats que soulève toute velléité de réforme.

On examinera trois arguments qui, sans être exclusifs ni recouvrir l'ensemble des différences possibles, conforteraient l'hypothèse d'une distinction de nature entre la transcription et l'annotation :

- 1°) selon de degré de compétence requis de la part de l'exécutant,
- 2°) selon l'ordre des opérations,
- 3°) selon le type d'action exercée sur les données.

#### 4.2 Argument n° 1 *L'orthographe est une compétence générique, l'annotation est une compétence spécifique.*

Chez des locuteurs natifs massivement scolarisés dans une langue à tradition écrite – ce qui est le cas de la très grande majorité des adultes français –, on distingue les connaissances de tout locuteur-scripteur d'un côté et ce qui relève de l'activité du linguiste de l'autre, de la

même façon qu'ont été séparés ce que savent les locuteurs (l'épilinguistique chez Culioli) et le travail accompli par les grammairiens. Une chose est d'employer le conditionnel, une autre de savoir l'orthographe, une troisième de l'analyser dans des catégories savantes.

Du fait que l'apprentissage de l'écriture et de la lecture se fait dans un cadre scolaire qui se fonde sur la maîtrise de savoirs grammaticaux pour rendre compte de certains des décalages entre ce qui s'entend et ce qui s'écrit, il s'établit une porosité entre l'épilinguistique et le métalinguistique au cours de l'acquisition. L'alphabétisation affine et modèle les compétences des locuteurs en phonologie – par une définition normative des traductions acoustiques – et en morphosyntaxe. Elle les rend conscients d'un certain fonctionnement de la langue jusqu'à ce que, imaginativement, ils en viennent à substituer la représentation écrite à la réalité sonore ou qu'ils projettent des artefacts appris et sanctionnés socialement en les considérant comme appartenant à l'essence de la langue. Jusqu'à quel point le français serait-il ce qu'il est s'il n'était écrit dans l'orthographe qu'on sait ?

Cette aperception, qui affleure dans certains propos, pose une question de méthode. Comment s'assurer que les opinions émises, les retours sur le dire ne constituent pas un artefact des questions de l'enquêteur, que ce qui est répondu aurait trouvé son expression si celle-ci n'avait pas été dûment sollicitée ? Qui parle dans le retour qu'effectue un locuteur sur sa pratique ?

ch\_AC7: et à votre avis

ch\_AC7: l'orthographe la correction de la langue sont-ils enseignés aussi bien ou moins bien

qu'il y a trente-cinq ans

DJ39\_DIA: d'abord

DJ39\_DIA: quant quant à moi quant à moi je je je crois maintenant c'est une opinion qui m'est toute personnelle

DJ39\_DIA: mais j- je crois que c'est vrai

DJ39\_DIA: euh on ne fait pas de fautes d'orthographe on en fait fort peu quand on lit beaucoup

ch\_AC7: mh

DJ39\_DIA: c'est-à-dire que si vous lisez régulièrement ne serait-ce que le journal

DJ39\_DIA: euh bon ben vous voyez

DJ39\_DIA: vous voyez les accords vous voyez les mots

DJ39\_DIA: difficiles et cetera et puis vous les photographiez

ch\_AC7: mh

DJ39\_DIA: la la meilleure preuve est que

DJ39\_DIA: quand vous savez pas comment comment est un mot

ch\_AC7: oui mh mh

DJ39\_DIA: si ça s'écrit euh

DJ39\_DIA: comme ceci ou comme cela vous les écrivez l'un au-dessus de l'autre vous regardez

ch\_AC7: mh mh mh mh

DJ39\_DIA: oh ben oui ça ça passe et ça ça passe pas bon

*Extrait de ESLO\_DIA 1221*

Pour s'en tenir aux propos du témoin – un médecin en retraite – qui tient à signifier son attachement à l'orthographe, on relève la corrélation qui est faite à des pratiques culturelles (la lecture du journal, l'écriture) et le lien exclusif avec la graphie alors même que la question posée mentionnait aussi « la correction de la langue ». La transcription suffit à montrer, à une certaine distance de cette assurance, l'importance des disfluences (amorces et répétitions), les

tournures maladroites (« quand vous savez pas comment comment est un mot »), la scansion par des ponctuations (*euh, ben, oh*). Cet écart confirme le point de vue du locuteur-scripteur : l'apprentissage des règles, leur application et leur vérification s'accomplit de façon automatique en se fiant non à l'oreille mais à l'œil.

L'orthographe dont il est question est l'orthographe d'usage mais la transcription qui a été réalisée ajoute d'autres décisions. Dans le tour de parole « mais j- je crois que c'est vrai », sont assurés :

- la validation des unités lexicales par l'espace ou l'apostrophe (j- est marqué comme un bruit), leur reconnaissance graphique (/mè/= « mais ») et leur coupe,
- la désambiguïsation (c'est ≠ sait),
- les accords grammaticaux (le « s » de « crois » est corrélé à « je » et l'absence de « s » à la fin de « vrai » se conforme au singulier de « est »).

Il y a effectivement une disjonction entre la représentation scripturale de la langue et les annotations, à la fois dans l'explicitation que celles-ci proposent des catégories (POS), des modalités ou des corrélations, qui peuvent avoir une traduction dans le rendu orthographique sans être déclarées comme telles, et dans la maîtrise des conventions. Reste qu'il s'agit plutôt d'une différence de degré que de nature dans un processus qui reste fondamentalement identique : la conversion d'un signal acoustique en une représentation optique.

#### 4.3 Argument n° 2 *L'annotation est seconde ; elle n'est efficiente qu'à partir d'une transcription.*

Outre que l'on peut annoter directement des spectrogrammes, l'alignement sur le signal autorise une disjonction des annotations qui figureront sur une tire séparée. Les analyses peuvent se cantonner à ce niveau sans qu'il soit nécessaire de se fonder sur la transcription, et réciproquement. Néanmoins, la question de l'alignement, qui s'est posée autant pour la traduction ou la détection du plagiat du côté des textes que pour la représentation du signal sous forme iconique ou symbolique, n'est pas construite de la même façon dans le cas de l'annotation. Celle-ci est interpolée ; elle peut apparaître dans la continuité des termes concernés, se différenciant de la pratique philologique des gloses portées en marge ou des notes infrapaginales [Grafton, 1998].

Cette proximité a son envers : étant continue, elle crée de la discontinuité. Il se jouerait là une différence essentielle entre la transcription et l'annotation : celle-ci n'a qu'une portée limitée et, quel que soit le détail de leur contenu, dans un langage de balises par exemple (qui n'est que l'une des façons d'annoter), l'enchaînement, au-delà de la discontinuité, ne fait pas apparaître un document tel qu'une lecture suivie restituerait un texte, un autre texte, ayant sa cohérence globale, intégrant hiérarchiquement tous ses composants. *Membra disjecta* de la source, l'annotation ne permet pas d'en restituer la consistance, la linéarité. D'une certaine façon, c'est une intervention paradigmatique à l'intérieur de la concaténation, la sélection de qualités particulières, partielles, dans un ensemble de possibles. Ce n'est pas une transposition normative du signal mais une décision de celui qui se propose d'en déterminer certaines caractéristiques à des fins de classement ou d'analyse.

Si on a considéré exclusivement le cas d'une transcription orthographique, la réflexion se poserait dans des termes assez comparables pour les transcriptions phonétiques. Dans l'acception technique du terme (généralement entendu comme une notation faite en API), ces notations requièrent un apprentissage supplémentaire. Leur statut est contradictoire : d'un côté, elles livrent moins d'informations sur le plan morphosyntaxique en épousant le signal mais leur pouvoir analogique devrait faciliter la compréhension ; de l'autre, elles se trouvent paradoxalement plus éloignées des compétences du locuteur qui ne peut le plus souvent les déchiffrer faute de maîtriser les correspondances. En se rapprochant au plus près de ce qui a

été énoncé, elles introduisent une distance avec ce qu'on lit par rapport à ce qu'un lecteur s'est habitué à reconnaître par sa pratique de l'écriture, en particulier pour ce qui a trait au découpage en mots.

Un examen des corpus déposés sur les différentes plates-formes montre que cette solution est assez peu utilisée : consommatrice de temps, source d'erreurs (quelle garantie a-t-on d'être en présence de la réalisation ouverte ou fermée, antérieure ou postérieure d'une voyelle moyenne ? [Bergounioux, 2008]), difficile à interpréter même de la part d'un lecteur entraîné, la restitution en API intervient le plus souvent d'après la version orthographique, ne serait-ce que pour des raisons de maniabilité.

#### Transcription en API

BV1 /mmmHpBssédubaRatCpaskE/

BV1

BV1 /spaposibdevifkOmsafCdé/

BV1AMI /sisiséposiJaJattUski/

BV1 /bEEtBlegamClotfwasétBilabitodizVitJèmidésBtUsFllasipRB

lasBsFR/

BV1AMI /hC/

BV1 /tsélgamCla/

#### Transcription orthographique

BV1: je pense c'est du baratin parce que

BV1: [bb:noise:instantaneous]

BV1: c'est pas possible de vivre comme ça enfin

BV1: b- bah attends le gamin l'autre fois

BV1AMI: si c'est possible y a y a de tout hein

BV1: sept ans il habitait au dix-huitième il descend tout seul là s- il prend l'ascenseur

BV1AMI: hein

BV1: tu sais le gamin-là

#### *Comparaison des transcriptions en API et orthographe ESLO2\_ENT\_1001*

Même si la transcription en API ne saurait se substituer au signal, on relève d'emblée plusieurs différences significatives :

- la suppression de certaines émissions sonores : /mmm/
- la notation alphabétique de phonèmes sans réalisation /spa/ pour « c'est pas », /posi/ pour « possible »...),
- la restitution des formes standard (« vivre » pour/vif/, « attends » pour /øtB/),
- le repérage des amorces (« b- », « s- »),
- le figuration de ponctuations par une forme conventionnelle (/bø/ transcrit « bah »).

Si l'annotation intervient le plus souvent après une transcription orthographique, la distance entre celle-ci et une transcription phonétique fonctionne comme un révélateur de la part d'annotation sous-jacente rajoutée par l'orthographe, comme une preuve du travail d'apprentissage requis pour s'en assurer la maîtrise.

#### 4.4 Argument n° 3 *La disjonction tient à la nature des opérations effectuées par les deux types de traitement sur le signal.*

Dans ce cas, en restreignant l'observation à un langage de balises du type XML, quelle distinction peut-on faire entre l'indication du « s » requis par l'orthographe qui ne s'entend

pas à la fin de noms, de pronoms et d'adjectifs et la balise <pluriel> ? Entre la graphie « -er » d'un infinitif et <infinitif> ? Y a-t-il moins d'informations dans un cas que dans l'autre ?

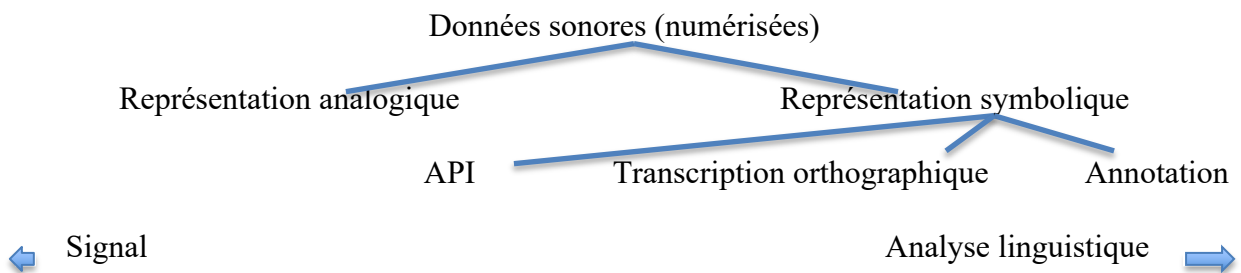
On peut noter d'emblée quatre différences. L'orthographe annote de façon prédictive (*vs a posteriori*), lacunaire (*vs systématique*), polymorphe (l'infinitif peut être marqué par d'autres graphies que « -er ») et homonymique (*les voies* = [det + nom] ou [pronom + verbe : (*que tu les voies*)] ?). Cependant, à sa façon, elle inscrit en continu une partie des propriétés morphosyntaxiques de l'énoncé que coderait une annotation, soit directement (le pluriel), soit indirectement (d'une finale de pluriel en *-ent* se déduit la catégorie « verbe »). En revanche se trouvent surimposées des informations étymologiques rarement exploitées dans une approche synchronique mais efficaces pour discriminer des homonymes. La séquence orthographiée privilégie la reconnaissance optique au détriment d'une reproduction fidèle du signal.

La pétition de principe que ça ne s'entend pas vaut dans l'un et l'autre cas. S'il s'agit du statut des notations (homogènes avec le texte dans l'orthographe, disjointes par des chevrons dans l'annotation), avec intégration continue dans le texte ou interpolation, l'opposition ne résiderait pas tant dans la nature des graphismes que dans le mode de figuration et de sortie d'édition.

```
<annotatedU end="#T175" start="#T174" wh="spk1" xml:id="au72">
  <u>
    <seg xml:id="s343">alors ils faisaient comme ça euh <pause type="short"/>et je
me suis rendue compte que ça n'allait pas </seg>
    <anchor synch="#T183"/>
    <seg xml:id="s344">parce que moi je je lisais et je lisais un rigue </seg>
    <seg xml:id="s345">euh la première ligne </seg>
  </u>
  <spanGrp>
    <span type="com" target="#344">mot italien = ligne</span>
  </spanGrp>
</annotatedU>
```

*Une annotation en XML*

En prenant en compte ces arguments, on proposera une autre distribution que celle présentée en commençant.



*Figuration graphique des données sonores*

Que rajoute l'annotation si l'on introduit une distinction entre ce qui équivaut à la reconnaissance d'une propriété déjà comprise dans la transcription orthographique et ce qui requiert une analyse linguistique complémentaire ? Il faut poser la question différemment selon que sera pris en compte (i) ce qui est contenu dans le cadre d'une unité (POS), (ii) ce qui est itératif entre plusieurs d'entre elles (les accords) et (iii) ce qui est discontinu (les marqueurs de discours).

Entre le signal et l'annotation, la transcription est à considérer comme une modalité de l'annotation, aussi sommaire et élémentaire qu'on voudra, première autant qu'il sera nécessaire dans la chaîne des opérations, mais irréductible à d'autres étapes du travail qui se situent sur d'autres plans. A terme, le développement de systèmes automatiques de reconnaissance vocale pourrait rendre superflue l'intervention du linguiste, la plus coûteuse en temps (et en argent). La conversion graphique se ferait à la demande et autant que de besoin par un traitement effectué directement à partir du signal. Pour l'instant, aucune technique n'est suffisamment fiable pour se substituer au travail manuel.

En conséquence, on posera qu'il existe seulement deux discontinuités dans la chaîne de traitement d'un corpus sonore :

- la première pérennise la labilité du signal en effectuant un enregistrement sur un support qui, d'analogique, est devenu numérique ;
- la seconde annote les données numérisées, en incluant dans ce processus, entre autres, la transcription.

Toute l'histoire de la linguistique, des débuts de la grammaire comparée [Bopp, 1816] à l'avènement du structuralisme, est le reflet de ces deux étapes. Le XIX<sup>e</sup> siècle effectue une remontée qui, partant des textes – ceux qui avaient permis de reconnaître l'identité de la famille indo-européenne – cherche en eux ce qu'ils reflètent du versant sonore des langues jusqu'à en fournir une objectivation non philologique avec l'apparition de la phonétique à partir des années 1870. Le XX<sup>e</sup> siècle substitue à une description constamment affinée du signal la construction d'un système de la langue dont toutes les écoles (Baudouin de Courtenay en Russie, Saussure en Europe occidentale, Sapir aux Etats-Unis), établissent de façon indépendante que l'unité sonore (le signifiant) est à chercher du côté du psychique et non de la phonation ou de l'acoustique.

## 5. Segments et propriétés

Il y a différentes segmentations opérées à l'intérieur des corpus oraux. Sans revenir sur celle qui, dans l'échange dialogique, fixe des bornes au discours, on déclinera, par ordre décroissant :

- la décomposition en séquences,
- la répartition en tours de parole,
- les coupes correspondant à des syntagmes, des propositions, des phrases, des paragraphes,
- les scansion qui s'ajustent aux inflexions prosodiques de l'énoncé, aux groupes de souffle, aux silences (il y a de nombreuses coïncidences avec le point précédent),
- la séparation en mots, congruente ou non avec l'orthographe et l'usage de la ponctuation,
- la séparation en unités élémentaires, morphologiques et phonologiques : détailler le signal en une succession de lettres et d'espaces, c'est aussi segmenter.

La pratique des linguistes en ces matières a largement repris le modèle proposé par Boas [1911] et ses élèves, la distinction entre une première ligne phonétique découpée en séquences morpho-lexicales et une deuxième ligne qui identifie les morphèmes en correspondance, soit par une traduction littérale, soit par l'indication de leur valeur grammaticale, avec une reprise globale du texte *in fine* afin d'en restituer la cohérence.

Dans cette perspective, la transcription orthographique est à concevoir moins comme une anticipation de l'annotation que comme une première annotation, concurremment avec d'autres qui se centrent sur la phonétique (cf. corpus PFC), la prosodie (cf. programme Rhapsodie), la kinésie (cf. corpus COLAJE) ou qui utilisent d'autres systèmes que l'orthographe (sténographie, API...).

En quoi l'écriture du français constitue-t-elle un prétraitement des données aux différents niveaux de l'analyse ? Comme les règles orthographiques interfèrent, au moment du travail

effectué à des fins linguistiques, avec les conventions édictées en amont et les consignes données aux opérateurs, on esquissera la récapitulation de ce qui s'ajoute et se soustrait aux différents niveaux dans la réalisation des transcriptions d'ESLO.

### 5.1 Représentation du phonétisme

\* Ce qui est assuré :

- une meilleure **lisibilité**, les agrégats de symboles en petit nombre permettant une saisie globale des termes et accélérant leur déchiffrement,
- une indication sur la **prononciation** standardisée des mots et des enchaînements (liaisons) représentée par l'orthographe conventionnelle : si la transparence du signal est brouillée par l'étymologie et la morphologie, les incertitudes sur la réalisation (du type *gageure* ou *couvent*) ne sont pas très fréquentes,
- une première **désambiguïsation** (*ce/se, et/est, tes/t(u) es, qu'il/qui le...*),
- la **segmentation** en mots.

\* Ce qui n'est pas pris en compte :

- les réductions (*assiation* pour *association*, *vif* pour *vivre*) et les épenthèses (*omblet* pour *omelette*),
- la réalisation ou non du schwa et des liaisons,
- le mode d'articulation ou de résonance des voyelles – spécialement les voyelles moyennes –, des consonnes palatales (e.g. mouillure) et des liquides,
- les assimilations (*chépas* pour *je ne sais pas*) et les neutralisations,
- l'encombrement des organes phonateurs (bouche pleine, rhume, enrrouement...),
- certains défauts de prononciation (blèsement, nasonnement...),
- la pondération intrinsèque et relative des timbres (hauteur, intensité, durée),
- le contour mélodique – sauf le cas de l'interrogation avec intonation montante notée par un point d'interrogation) :

RN488: oh

RN488: rien

RN488FRE: hé il est déjà dix heures et demie ?

RN488: ben oui

*Désambiguïsation de l'interrogation par « ? » Extrait de ESLO\_REPAS\_1247\_C*

Ce qui est quelquefois consigné (alternativement ou complémentirement avec une annotation) :

- les prononciations erronées (*quand je suis enstré en seconde*) et les lapsus,
- certaines indications prosodiques transcrites par des signes conventionnels (e.g. les pauses par #).

### 5.2 Marques morphologiques

En limitant la recension aux marques quiescentes :

\* Ce qui est pris en compte :

- les **marques nominales** : pluriel en *-s* et en *-x*, féminins en *-e* après voyelle...
- les **marques verbales** : marques de personne et de nombre en *-s (-x)*, *-t*, *-ent* (+ auxiliaire *être* et formes du passé simple), de temps (différence entre *-(r)ai*, *-(r)ais*) et de mode (*-ait*, *-er*, *-é*, *-a /-ât*, *-ut/-ût* etc.),

- en *l'absence de marques*, la disponibilité d'un indice concernant les catégories invariables comme contribution à l'identification des POS (cf. *ils parlent fort*),
- quelques marques de *composition* : adverbes (finale *-ment*), distinction verbe/adjectif (*fatigant/fatiguant*)...

\* Ce qui n'est pas pris en compte :

- les homographies (« chante » = Ind.1Sg, Ind.3Sg, Imp.2Sg, Subj.1Sg, Subj.3Sg ; « fois » = Sg / Pl),
- le découpage interne du mot.

Au-delà des mots, on remarque l'indication d'élisions, de sites de liaisons facultatives et d'accords (non actualisés acoustiquement), certains repères anaphoriques :

DJ39: il est évident que par exemple euh

DJ39: euh des professions comme la pédiatrie la gynéco-obstétrique pourraient être euh euh des spécialités féminines voyez-vous

DJ39: mais elles comportent beaucoup d'urgences qui sont

DJ39: euh

DJ39: peu compatibles avec une vie équilibrée de femme mariée

*Extrait de ESLO1\_ENT\_003C*

Entrent aussi en ligne de compte le regroupement (clusterisation/chunking) et l'organisation sous-jacente en propositions et en phrases.

En résumé, les lacunes sont essentiellement d'ordre phonétique, les additions de type morphologique. Ce qui rend tangentes l'annotation et la transcription, c'est l'ensemble des informations qu'ajoute l'orthographe. Ces adjonctions ne se calculent pas à partir du signal de parole (tout son à prononcer a un reflet dans la graphie) mais par comparaison avec ce que produirait la notation strictement phonétique d'une réalisation orthoépique. Qu'une partie de ce qui a été effectivement prononcé ne soit pas écrit ne signifie pas pour autant que cela ait été entendu, ni par l'auditeur, ni par le locuteur lui-même. Réciproquement, que soient rétablis des éléments qui n'ont pas été dits ne présuppose en rien que le locuteur ne voulait pas les prononcer ni que l'auditeur ne les aurait pas reconstitués.

Reste l'interrogation sur la légitimité de notations imposées par le manuel du transcripteur, telles que « on n'a pas vu » pour un locuteur qui n'utilise à peu près jamais le premier élément de la négation que l'enchaînement fait apparaître ou « les meufs que j'ai vues » quand l'accord du complément d'objet direct antéposé n'est pas sans poser des questions de crédibilité dans de nombreux usages du français. Cette extrapolation n'a pas reçu de réponse satisfaisante. Le principe adopté est que, si rien ne le contredit dans le signal, on retient la forme la plus normative, sans interroger la maîtrise du locuteur en matière d'orthographe. ESLO s'en tient à une application stricte et normée des règles, quitte à rajouter de l'information dans « Il a pas de difficultés », où le pluriel est déjà une interprétation, c'est-à-dire un choix, le singulier en étant un autre, aucun des deux ne pouvant être sans incidence(s).

## 6. Annotations et métadonnées

A partir du moment où annoter, dans l'acception extensive assignée au terme, revient à ajouter de l'information à des données afin de transformer une source sonore en corpus linguistique (dont les exploitations peuvent ne pas relever de la linguistique), il s'opère un partage entre ce qui figure dans les annotations et ce qui est consigné par les métadonnées. Dans un langage de balisage, la mise en forme contribue à estomper la différence entre le



mode de traitement du document source et sa description, offrant un nouvel exemple des transformations radicales introduites par l'informatique en regard des techniques éprouvées dans les domaines de l'archivistique et des bibliothèques.

En matière de catalogage et d'indexation, on ne fouille pas les données orales de la même manière que les textes. Il n'existe pas d'écoute rapide comme on parle de lecture cursive et c'est aux métadonnées d'y suppléer afin d'assurer maniabilité et accessibilité. Leur production répond à une série d'objectifs :

- (i) définir les critères qui permettent de spécifier les corpus selon leur mode d'élaboration, de traitement et d'exploitation,
- (ii) connecter ceux qui sont corrélés en calculant leur interopérabilité,
- (iii) faciliter les opérations d'accès, de gestion et de connexion,
- (iv) préparer les réponses aux requêtes sur le contenu.

Ces prescriptions sont déterminantes dans l'émergence du web sémantique et la construction des ontologies. Elles soulèvent des questions épistémologiques [Rastier, 2004] dès lors qu'on en revient à une conception taxinomique, une définition par des propriétés internes aux objets en lieu et place d'une structure construite à partir d'oppositions.

Concrètement, quelles informations s'attend-on à trouver ? Autrement dit, quels sont les critères à prendre en compte pour sélectionner ce qui doit y figurer ? En théorie, tout ce qui permet d'augmenter la compréhension du corpus et de ses éléments sans faire double emploi avec les renseignements que fournissent la transcription et l'annotation. Sont absents de la notation orthographique comme des annotations :

- l'accessibilité, la localisation (en particulier l'adresse électronique du site),
- l'origine de la production, la source (y compris les implications légales),
- la définition de la nature du corpus (DTD), le type de document, son mode de codage,
- certaines indications de contenu.

Les quinze éléments pris en compte par le Dublin Core (OLAC) afin de couvrir tant les formats que les contenus et les propriétés sont inscrits sous forme de marqueurs (en XML) traités en RDF (Resource Description Framework). Dans le modèle de l'EAD, ils sont détaillés selon un canevas contraint (cf. le site de la BnF) et complétés par une mise en relation avec d'autres documents, l'ajout de références bibliographiques. Ce travail n'a pas seulement pour fin de décrire les ressources, il en prépare le moissonnage notamment en OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting) avec pour objectif d'assurer l'interopérabilité et l'archivage.

Dans ESLO, sont inscrits dans les métadonnées les renseignements pertinents pour affiner l'interprétation, ce qui pose la question de l'anonymisation, toute indication supplémentaire (la voix en est une qui serait déjà par soi seule dirimante) élevant les chances de reconnaissance [Eshkol *et al.*, 2014]. Ce qui correspond pour un livre à la notice d'auteur est développé et complété par des indications qui dépareraient des fiches de bibliothèque comme, par exemple, le niveau scolaire ou le nombre d'enfants. Ces indications sont cruciales pour effectuer certains tris dans une approche sociologique, par un effet en retour des conditions spécifiques des productions orales sur les informations nécessaires à l'exploitation des informations en deçà de la transcription.

Les métadonnées sont séminales. Des chercheurs peuvent à tout moment créer de nouvelles versions de la ressource, ajouter ou supprimer des éléments en fonction de leurs besoins. Il s'agira d'états dérivés avec des éléments qui pourront être ou non inclus dans la présentation générale, introduisant la nécessité d'un versionnage des métadonnées. L'affichage sur site est avant tout conçu pour faciliter l'accessibilité, et pas seulement aux linguistes, avec une réflexion sur la diffusion auprès du grand public – et auprès de ceux qui consulteront, dans quelques décennies ou quelques siècles, ces documents. Les modifications sont elles-mêmes riches d'enseignement.

## Conclusion

On dira pour finir quelques mots de l'annotabilité (en laissant de côté ce que serait une « transcriptibilité »). Pour paraphraser un aphorisme de Jakobson [1984], les transcriptions ne se distinguent pas par ce qu'elles peuvent coder mais par ce qu'elles doivent coder, i.e. par le débord que l'orthographe réalise sur le signal et par les soustractions qu'opère l'écriture. Comme on a distingué deux acceptions d'annotation, on marquera une différence entre deux acceptions d'annotabilité.

Dans un cas, il s'agit de tout ce qui peut faire l'objet d'une addition à des données enregistrées, soit par alignement à partir d'une représentation analogique, soit par interpolation – comme avec un langage de balises. L'annotabilité représente alors la projection d'une théorie, d'une taxinomie ou de requêtes attendues sur des données, liée ou non à un processus d'apprentissage, avec une liste extensible à loisir de ce qui pourra être inclus. Dans l'autre cas, l'annotabilité sera ce que chaque langue, en fonction de son orthographe, de ses conventions, impose via la transcription, depuis les marques à éclipse (liaison) jusqu'à des formes telles « nous croyions » qui ne s'entendent ni, le plus souvent, ne s'imaginent. Il y aurait, à travers ce concept, l'idée que l'annotabilité serait ce qui, dans chaque langue, fera potentiellement ou non double emploi avec telle ou telle modalité de l'annotation. A ce titre, la transcription est conçue comme une annotation, variable en annotabilité selon ce qu'a été l'histoire des langues écrites et l'écriture des langues, inhérente à leur mode de figuration.

Qu'il y ait eu, dans l'histoire des langues, une anamnèse de ce qu'il est apparu comme nécessaire, à certains moments, de faire apparaître dans la transcription – autrement dit l'élaboration d'une orthographe – se retrouve aujourd'hui par différence avec l'ampleur de la tâche qu'il reste à effectuer pour compléter par l'annotation les informations que l'écriture a déjà rendues disponibles. La question se pose en traitement automatique des langues d'une distance entre des conventions, rivées à une forme d'écriture (l'alphabet latin et ses adjuvants), héritières de principes (le poids de l'étymologie) face aux annotations (dont les transcriptions) nécessaires à l'analyse. Une typologie reste à construire des solutions à adopter selon les principes vernaculaires retenus par les langues dans leur(s) écriture(s) et leur(s) orthographe(s).

## Sitographie

ANCOR [http://tln.li.univ-tours.fr/Tln\\_Corpus\\_Ancor.html](http://tln.li.univ-tours.fr/Tln_Corpus_Ancor.html)

BnF Gallica <http://gallica.bnf.fr/>

BnF [EAD] [http://www.bnf.fr/fr/professionnels/formats\\_catalogage/a.f\\_ead.html](http://www.bnf.fr/fr/professionnels/formats_catalogage/a.f_ead.html)

CLARIN <https://www.clarin.eu/>

CoCoON <http://eslo.huma-num.fr/index.php/pagecorpus/pagepresentationcorpus>

COLAJE <http://colaje.scicog.fr/index.php/multimodalite>

CORLI <https://corli.huma-num.fr/>

DARIAH <http://www.dariah.eu/>

ESLO <http://eslo.huma-num.fr/>

[http://eslo.huma-num.fr/images/eslo/pdf/GUIDE\\_TRANSCRIPTEUR\\_V4\\_mai2013.pdf](http://eslo.huma-num.fr/images/eslo/pdf/GUIDE_TRANSCRIPTEUR_V4_mai2013.pdf).

Europeana sounds <https://www.europeanasonsounds.eu/fr/>

OLAC <http://www.language-archives.org/>

Ortolang <https://www.ortolang.fr/>

PFC <http://www.projet-pfc.net/>

Rhapsodie <https://halshs.archives-ouvertes.fr/halshs-01061368/>

TEI <http://www.tei-c.org/index.xml>

XML <https://www.w3.org/XML/>

## Bibliographie

BAUDE, Olivier, *Corpus oraux. Guide des bonnes pratiques*, Paris-Orléans, CNRS Éditions – P.U.O., 2006.

BENZITOUN, Christophe, FORT, Karen, SAGOT, Benoît, « TCOF-POS un corpus libre de français parlé annoté en morphosyntaxe », communication présentée aux JEP-TALN, Grenoble, 2012.

[http://www.shs-conferences.org/articles/shsconf/abs/2012/01/shsconf\\_cmlf12\\_000252/shsconf\\_cmlf12\\_000252.html](http://www.shs-conferences.org/articles/shsconf/abs/2012/01/shsconf_cmlf12_000252/shsconf_cmlf12_000252.html)

BERGOUNIOUX, Gabriel, « Des œufs mués : entendre omelette », *Actes du 1<sup>er</sup> Congrès Mondial de Linguistique Française*, Paris, 2008.

<http://www.linguistiquefrancaise.org/articles/cmlf/abs/2008/01/cmlf08157/cmlf08157.html>

BERGOUNIOUX, Gabriel, *Linguistique de corpus. Une étude de cas*, Paris, Champion, 2016.

BERGOUNIOUX, Gabriel, JACOBSON, Michel & PIETRANDREA, Paola, « Annotating Linguistic Corpora » in Wendy AYRES-BENNETT & Janice CARRUTHERS (dir.), *Romance Sociolinguistics*, Berlin, De Gruyter, à par.

BOAS, Franz, *Handbook of American Indian Languages*, Bureau of American Ethnology, Washington, Government Print Office, 1911.

BOPP, Franz, *Über das Conjugationssystem der Sanskritsprache in Vergleichung mit jenem der griechischen, lateinischen, persischen und germanischen Sprache*, Francfort, Andreae.

CORDEREIX, Pascal, « Les enjeux de la conservation du patrimoine sonore enregistré », in PIERRET, Gilles, *Musique en bibliothèque*, Paris, Cercle de la Librairie, 2012, p. 275-295.

DEGAND Liesbeth, CORNILLIE, Bert, PIETRANDREA, Paola (éd.), *Discourse markers and modal particles: categorization and description*, Amsterdam-Philadelphie, John Benjamins, 2013.

ESHKOL, Iris, BAUDE, Olivier, KANAAN, Layal, MAUREL, Denis, DUGUA, Céline, « Procédure d'anonymisation et traitement automatique : l'expérience d'ESLO », *Journée d'études ATALA, Ethique et TAL*, Paris, 2014 <[halshs-01165957](http://halshs-01165957)>

GOODY, Jack, *La Raison graphique*, Paris, Minuit, 1979.

GRAFTON, Anthony, *Les Origines tragiques de l'érudition. Une histoire de la note en bas de page*, Paris, Seuil, 1998.

HERRENSCHMIDT, Clarisse, *Les Trois Écritures : langue, nombre code*, Paris, Gallimard, 2007.

JAKOBSON, Roman, « Aspects linguistiques de la traduction », in *Essais de linguistique générale*, Paris, Minuit, 1963, p. 71-86.

LEECH, Geoffrey, MYERS, Greg, THOMAS, Jenny (éd.), *Spoken English on Computer: Transcription, Mark-up and Application*, Londres, Longman, 1995.

MALLET-POUJOL, Nathalie, *Protection des données personnelles, de la vie privée et de l'image*, rapport IRCOM/Huma-Num 2014, consultable en ligne [http://ircom.huma-num.fr/wiki/lib/exe/fetch.php?media=wiki:fichiers\\_publicques:corpus\\_droit\\_guide.nmp.pdf](http://ircom.huma-num.fr/wiki/lib/exe/fetch.php?media=wiki:fichiers_publicques:corpus_droit_guide.nmp.pdf)

MAUREL, Denis *et al.*, « Cascades de transducteurs autour de la reconnaissance des entités nommées », *Traitement Automatique des Langues*, 52-1, 2011, p. 69-96.

MUZERELLE, Judith, SCHANG, Emmanuel, ANTOINE, Jean-Yves *et al.*, « Annotations en chaînes de coréférences et anaphores dans un corpus de discours oral spontané en français », *Actes du 3<sup>e</sup> Congrès Mondial de Linguistique Française*, 2012.

[http://www.shs-conferences.org/articles/shsconf/abs/2012/01/shsconf\\_cmlf12\\_000252/shsconf\\_cmlf12\\_000252.html](http://www.shs-conferences.org/articles/shsconf/abs/2012/01/shsconf_cmlf12_000252/shsconf_cmlf12_000252.html)

RASTIER, François, « Ontologie(s) », *Revue des sciences et technologies de l'information*, série : *Revue d'Intelligence artificielle*, vol. 18, n°1, 2004, p. 15-40.

ROBIN, Agnès, *Droit de la propriété intellectuelle*, rapport IRCOM/Huma-Num 2014, consultable en ligne [http://ircom.huma-num.fr/wiki/lib/exe/fetch.php?media=wiki:fichiers\\_publicques:guide\\_corpus\\_dpi.pdf](http://ircom.huma-num.fr/wiki/lib/exe/fetch.php?media=wiki:fichiers_publicques:guide_corpus_dpi.pdf)

VANDERMEERSCH, Léon, *Les Deux Raisons de la pensée chinoise*, Paris, Gallimard, 2013.