L'annotation des corpus oraux

Gabriel Bergounioux, Michel Jacobson, Paola Pietrandrea

L’annotation des corpus oraux

Gabriel Bergounioux, Michel Jacobson, Paola Pietrandrea

0. Introduction

Deux approches opposées canalisent l’étude de l’oral :

- l’analyse de la parole comme signal : la phonétique expérimentale est intrinsèquement liée à la dialectologie et à la géographie linguistique ;
- la recherche anthropologique qui s’attache au contenu (ce qui est dit) plutôt qu’à la forme (comment c’est dit).

L’observation de la variation linguistique, à tous les niveaux, requiert des données en masse et une comparaison interne aux données. A ce titre, pour les langues à tradition scripturale, la sociolinguistique a été le lieu de constitution d’une technique qui réunissait la pratique philologique des collections de textes et la constitution des données en linguistique de terrain. Le corpus est apparu comme la meilleure présentation des ressources collectées de façon contrôlée. La diversité des réalisations exigeait que soient définies des principes de transcription, d'identification et de catégorisation, ce qui sera réalisé par l’annotation comme technique de quantification et taxinomie.

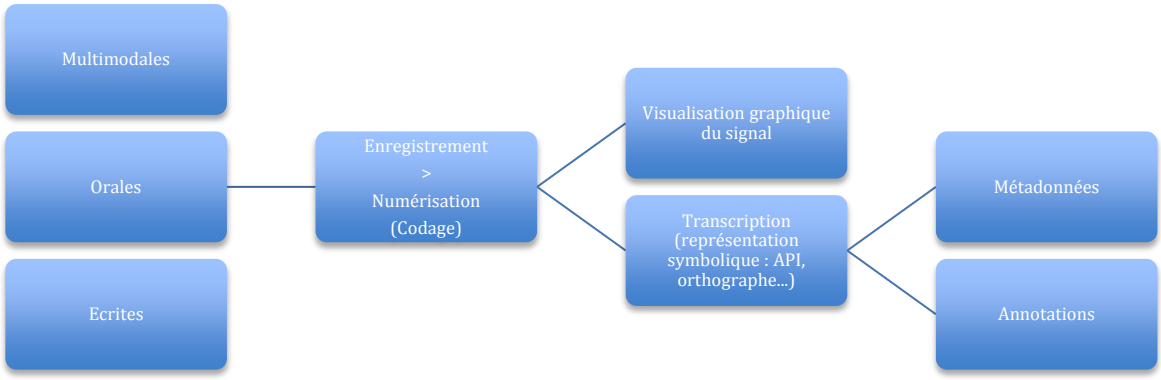
Tout en choisissant pour l’annotation une définition extensive qui inclut la transcription, ce chapitre se restreint à une acception restreinte : l’insertion d’informations linguistiques dans un corpus dans le respect de règles d’écriture codifiées et explicites.

On peut distinguer pour commencer trois grands types de ressources :

- écrites,
- multimodales,
- orales.

Ce sont les objectifs de la recherche qui prescrivent la forme des données, lesquelles conditionnent en retour les modalités de représentation. Le multimodal est indispensable pour les études sur l’acquisition du langage par l’enfant ; l’écrit est au principe des études philologiques. Ce chapitre traite des corpus oraux.

Tableau 1 Chaîne opératoire



Le schéma peut être complété : d'un côté en tenant compte de la diversité des formes de transcription en relation avec ce qui a été prononcé, de l'API aux logogrammes ; de l'autre, en figurant la suite des opérations d'outillage et d'archivage, de diffusion et d'analyse.

Ainsi conçue, à côté de la transcription, l'annotation se présente à la fois comme une technique (plus ou moins automatisée) et une méthode (à dominante empirique), au point de convergence des sciences humaines (l'enquête, le terrain) et de l'informatique, en particulier du TAL (Traitement Automatique des Langues) pour l'outillage. A ce double titre, elle requiert une explicitation de ses principes, l'adoption d'une démarche procédurale, ne serait-ce que pour distinguer les emplois qui apparaissent sous une même figuration.

Exemple 1 : Trois annotations pour un énoncé
Transcription : « Y a pas de bug »
Annotations : (i) <Prononcé : /byg/>
(ii) <Alternative : « bugs »>
(iii) <POS : N / Emprunt / Informatique>

On étudiera successivement :
- la confection et le traitement des corpus
- l'annotation : rôle du chercheur, process, outils, typologie
- l'état des lieux et les perspectives.

1. Principes d'annotation

1.1 Qu'est-ce qu'annoter ?

Dans son acception ordinaire, l'annotation se présente comme une suite de caractères interpolée dans un texte, qu'il ait été transmis sous forme écrite ou qu'il résulte de la transcription de données orales. Dans le premier cas, l'annotation s'inscrit dans la tradition philologique des marginalia, des gloses et des scholies, c'est-à-dire de commentaires dont la fonction est variable : correction d'un mot, traduction, expression d'une opinion… En ce sens, l'annotation constitue une intervention postérieure au document et lui demeure extérieure.

On peut considérer que la transcription des corpus oraux constitue de fait la première forme d'annotation de la source. Dans les langues à tradition écrite, au moins pour les langues romanes, un très grand nombre de locuteurs ont un accès direct à cette représentation.

L'usage ordinaire que les linguistes ont de l'*annotation* tend à restreindre sa définition à un langage de balises, en dehors du texte proprement dit mais dans la continuité de son écriture, suivant deux approches :

- la première, héritée de l'édition de texte, inscrit les balises dans le fil du texte : leur suppression ne compromet pas la possibilité d'un retour au document initial,
- l'autre, utilisée par exemple dans les bases de données relationnelles, structure intégralement le document de telle sorte que seule l'interprétation du balisage permettra de distinguer les contenus d'origine des notes ajoutées.

1.2 Pourquoi annoter ?

Le recours à l'annotation s'est imposé dès lors que l'éditeur d'un document écrit ou le producteur d'une transcription a voulu relever un phénomène particulier ou attirer l'attention du lecteur sur un phénomène donné. L'appel de notes (ou le renvoi) est le procédé par excellence de l'apparat scientifique des philologues, jouant sur une répartition visuelle qui dissocie le texte et les commentaires. Dans les corpus oraux, la préférence est donnée à une notation multilinéaire, qui sépare différents plans – comme [Boas 1911] l'a mise en œuvre pour des traductions alignées –, distinguant la représentation scripturale de l'énoncé et les indications fournies par le rédacteur.

Cette façon de faire, obligatoire pour des langues à tradition orale, s'imposait également pour la forme parlée des langues écrites. La confection de corpus sociolinguistiques ou dialectologiques, la transcription de données d'acquisition (en langue maternelle ou en tant qu'apprenant) ou de troubles du langage impliquaient de compenser, dans la restitution scripturale, certaines informations perdues au moment d'une conversion graphique, à commencer par les indications que fournit le signal.

1.3 Comment annoter ?

Le développement de l'informatique,

d'abord avec des propositions propriétaires comme le langage GML (Generalized Markup Language) d'IBM,

puis avec la normalisation en 1986 au sein de l'ISO (*l'Organisation Internationale de Normalisation*) d'un format générique SGML (Standard Generalized Markup Language),

enfin en 1996 avec l'apparition des premiers travaux de spécification de XML (eXtensible Markup Language)

a permis de séparer :

- la structure logique du document qui peut ou doit être défini dans un schéma tel que DTD (Document Type Definition) pour SGML,
- la représentation (ou structure physique) du document dont les règles de dérivation peuvent être précisées dans des feuilles de style.

Cette distinction des deux plans a été prévue pour distinguer, dans la chaîne d'édition, le domaine d'intervention des différents métiers. L'élaboration des schémas et des feuilles de style relève du travail de l'éditeur afin d'assurer une meilleure maîtrise des traitements, pour la validation et la mise en forme. L'utilisation de ces outils par les rédacteurs correspond à une aide à la saisie.

SGML a donné naissance à différentes applications. HTML (*Hypertext Markup Language*), EAD (*Encoded Archival Description*) et TEI (*Text Encoding Initiative*) sont les plus employées dans le domaine des humanités numériques. XML est plus facilement intégrable à Internet.

1.4 Métadonnées et annotations

Annoter un texte, c'est ajouter des informations dont on conjecture qu'elles feraient défaut aux destinataires alors qu'elles leur seraient potentiellement utiles, voire nécessaires. Les pratiques des linguistes ont été transformées par le transfert numérique des données. Il y a deux stratégies complémentaires :

- soit les informations sont fournies dans les métadonnées, à l'intérieur d'un fichier séparé ;
- soit elles sont insérées dans le texte lui-même.

La répartition effectuée entre métadonnées et annotations est variable : plus les renseignements portent sur l'ensemble du corpus ou du texte (identifiants des locuteurs, situation d'enquête, genre de texte…), plus souvent ils seront portés dans les métadonnées. Plus au contraire ils concernent un mot (ou une séquence sonore courte : clic, *filler*, amorce, locution…), plus ils tendront à être assignés aux annotations.

La nécessité d'ajouter certaines indications afin d'assurer leur exploitation (par exemple, la lemmatisation ou le calcul statistique des occurrences) a introduit en linguistique la notion d'*étiquetage*, à l'écrit pour le corpus Brown par exemple [Greene/Rubin 1971], à l'oral pour le LLC (London-Lund Corpus) issu du Survey of English Usage Corpus [Crystal/Quirk, 1964]. Dans le domaine roman, la première expérience d'étiquetage d'un corpus oral a été réalisée en italien sur le Corpus LIP (De Mauro et al. 1993).

L'exemple ci-dessous, extrait du « London-Lund Corpus » dans sa version numérique, fait

apparaître dans son codage des informations de différentes natures que l'on pourrait assimiler à des étiquettes. Celles-ci sont, en suivant l'ordre d'apparition sur chaque ligne : Text category, Text within category, Identifier, Tone unit number […], Speaker identity […], Text. La transcription (Text) utilise des conventions d'écriture pour noter des éléments non verbaux (rires, sonneries de téléphones, etc.), des pauses, l'intonation… (Pour une liste des conventions, cf. http://clu.uni.no/icame/london-lund/index.htm)

```
1 3   9 1420 1 1 b    20  *[mhm]* /
1 3   9 1410 1 1(A     11  ^th\en they _said# /
1 3   9 1430 1 1 A     11  well "^now that you`ve done th/ese# /
1 3   9 1440 1 1 A     11  and they`ve been "^s\o succ/essful# /
1 3   9 1450 1 1 A     11  we`d ^like you to do our s\uper# . /
1 3   9 1460 1 1 A     11  ^alpha:m\atic# /
1 3   9 1470 1 1 A     11  or ^s/omething# /
1 3   9 1480 1 1 A     11  and ^this is one of th/ese# /
1 3   9 1490 1 1 A     11  that ^goes s/ideways# /
1 3   9 1500 1 1 A     11  and ^fr/ontwards# /
1 3   9 1510 1 1 A     11  and em^br/oiders# /
1 3   9 1520 1 1 A     11  and *^d/arns# /
1 3   9 1530 1 1 A     11  and sews* ^b\uttons on# /
1 3   9 1540 1 1 b    20  *( - laughs) yes* /
1 3   9 1550 1 1(A     11  - - and I ^s=aid# /
```

Dans l'exemple, ci-dessous, extrait du corpus « Vienna-Oxford International Corpus of English » (VOICE : cf. http://ota.ox.ac.uk/desc/2542), l'expression du découpage d'un texte en unités lexicales ou le marquage sur chaque unité de sa forme, de son lemme et de sa partie du discours, utilisent la syntaxe XML avec les conventions de la TEI (~~*Text Encoding Initiative*~~), un programme international qui fixe les directives permettant la mise en forme et l'échange de textes électroniques sous norme ISO).

```
<u who="#LEcon351_S5" xml:id="LEcon351_u_17">
    <w ana="#PPfPP" lemma="it">it</w>
    <w ana="#VBSfVBS" lemma="be">'s</w>
    <w ana="#NPfNP" lemma="austria">austria</w>
    <w ana="#PPfPP" lemma="it">it</w>
    <w ana="#VBSfVBS" lemma="be">'s</w>
    <w ana="#RBfRB" lemma="very">very</w>
    <w ana="#JJfJJ" lemma="cold">cold</w>
    <w ana="#CCfCC" lemma="and">and</w>
    <w ana="#PPfPP" lemma="it">it</w>
    <w ana="#VBSfVBS" lemma="be">'s</w>
    <w ana="#PAfPA">_0</w>
    <w ana="#JJfJJ" lemma="hot">hot</w>
    <w ana="#RBfRB" lemma="enough">enough</w>
    <w ana="#PAfPA">_0</w>
</u>
```

Comme on peut l'observer dans ces deux exemples, l'annotation d'un texte ou d'une transcription consiste à enrichir l'information de départ par de nouvelles rubriques d'information qui, selon les époques, utilisent diverses conventions de codage (texte structuré, XML).

1.5 Diversité des pratiques

Les propositions des informaticiens et les attentes des linguistes n'étaient pas si faciles à concilier, ce qui explique certains retards [Léon 2015]. Faire collaborer des chercheurs issus des mathématiques (de la logique ou du calcul) ou de l'électronique avec des philologues ou des anthropologues n'allait pas de soi.
En linguistique, le recours aux corpus ne revêtait pas la même signification selon qu'ils étaient conçus comme le témoignage d'une culture ou comme des ressources pour l'analyse linguistique. Qu'on les appréhende comme un réservoir d'exemples ou un ensemble de données dont il s'agit de rendre compte, qu'on en fasse un élément de critique à l'encontre de théories ou qu'on privilégie une démarche empirique telle que le « modèle fondé sur l'usage » des grammaires de construction, les exigences en matière d'annotation diffèrent sensiblement.

Une différence supplémentaire tient à la façon de produire les annotations. Elle peut être manuelle, semi automatique (pré et post-édition) ou automatique. Celle-ci, qu'elle privilégie une méthode symbolique ou une méthode par apprentissage, suppose une théorie sous-jacente qui fournit à l'annotation (et à la pré-annotation) les outils dont elle a besoin et permet un retour sur l'évaluation des résultats [Fort 2012]. La méthode manuelle est considérée comme plus fiable mais elle ne peut maîtriser une grande quantité de données et, si elle est collective, pose la question de l'accord entre annotateurs. Les méthodes automatiques s'évaluent avant tout en fonction du nombre de corrections à apporter au moment de réviser les résultats.

Par exemple, pour une transcription de l'« Enquête Sociolinguistique à Orléans » entreprise quarante ans après la collecte, une comparaison a été effectuée sur un échantillon du corpus pour savoir si le coût d'une transcription manuelle serait moindre qu'une transcription générée par logiciel avec une relecture pour correction. Finalement, le choix a été fait, pour ce corpus, de ne pas recourir à la transcription automatique, car la surcharge cognitive en relecture lors de la correction faisait perdre en temps ce que la reconnaissance avait permis de gagner.

Sur ESLO [Baude/Dugua 2011], une organisation rationnelle du travail a conduit à la définition de plusieurs niveaux d'annotation :
- Le niveau zéro (T0) s'appuie sur des conventions minimales et a pour objectifs de faciliter la navigation dans le signal (synchronisation pour réécoute), de proposer une transcription pour tous les mots, y compris les amorces et les disfluences, avec un codage proche des usages de la langue écrite afin de faciliter lecture et édition. Une règle du *Guide de*

transcription spécifie qu'un transcripteur ne doit pas écouter plus de deux fois un segment.
- Le second niveau (T1) s'attache à produire une transcription d'exploitation destinée à une analyse linguistique finalisée : affinement de la restitution écrite (corrections, choix théoriques), ajouts de codages spécifiques (prosodie, multitranscription…) et d'annotations cumulées.
Pour l'annotation en T0, trois transcriptions ont été systématiquement réalisées :
- une version A : transcription brute réalisée le plus rapidement possible (coût : 10 fois le temps d'écoute);
- une version B : relecture de la version A par un autre transcripteur (coût : 5 fois le temps d'écoute) ;
- une version C : correction de la version B par un relecteur confirmé (coût : 5 fois le temps d'écoute).
L'ensemble des versions est conservé afin de pourvoir étudier les différences inter-individuelles et inter-groupes.

La difficulté d'harmonisation entre différents chercheurs travaillant sur un même corpus se retrouve à une échelle bien plus large dès qu'il s'il s'agit d'homogénéiser les annotations indépendamment des langues, des pays, des types de données. Un débat oppose les tenants d'une normalisation effectuée selon des principes universels a priori (comment définir un standard en fonction des contraintes de l'objet et des principes de la science ?) et les partisans des bonnes pratiques. Pour un exemple de standardisation sous norme ISO en cours sous l'égide de la TEI voir [Stührenberg 2012].

2. Qualification des corpus

Les annotations se situent au départ d'une chaîne d'opérations qui permet l'enrichissement des contenus et leur exploitation (décompte, analyse, structuration, repérage des collocations, concordances…). L'exemplification et la réalisation de tests d'entraînement en ~~Traitement Automatique des Langues (~~TAL) posent la question de la granularité des propriétés et de l'échantillonnage des données.

2.1 Taille

A partir de quel moment peut-on considérer qu'un échantillon de langue est représentatif de l'ensemble de ses usages ? La question, qui s'est posée pour le lexique (cf. la création du BASIC English ou du Français Fondamental) et pour la syntaxe, notamment en grammaire générative, avait pour repères la fréquence des emplois, la structure de la langue et leur corrélation (loi de Zipf). La variation sociale, devant intégrer l'ensemble des paramètres – dialectaux, diaphasiques et diastratiques – s'avérait moins assurée dans ses critères de classement et plus exigeante quantitativement. L'aspect quantitatif est le premier pris en compte. Il conditionne tous les autres. Un accroissement des mémoires des ordinateurs et la diminution du coût des équipements ont permis d'élever progressivement les exigences en ce domaine (cf. tableau).

Tableau 2 Corpus Ecrits

Brown Corpus	1.000.000	1961
Frantext (français)	300.000.000	1975
British National Corpus	100.000.000	1995
CORIS-CODIS (italien)	130.000.000	2001
Corpus de Referencia del Español Actual	160.000.000	2008
Reference Corpus of Contemporary Portuguese	300.000.000	2012
French frTenTen	10.000.000.000	2012

Tableau 3 Corpus Oraux

ESLO 1 et 2 (français)	7.000.000	1969
London Lund Corpus of Spoken English	500.000	1990
LIP (italien)	500.000	1993
CLAPI (français)	2.500.000	2005
Corpus de Referencia del Español Actual	9.000.000	2008
Reference Corpus of Contemporary Portuguese,	1.600.000	2012

On mentionnera également le Corpus de Catalanà Contemporani de l'Université de Barcelone, les corpus occitans Symila de l'Université de Toulouse Jean-Jaurès et le Thesoc de l'Université de Nice Sophia Antipolis.

2.2 Pondération

Un second critère concerne l'équilibre entre différentes pratiques, différents contextes. En règle générale, la collation de documents écrits reflète moins un état de langue dans l'absolu que certains de ses usages, littéraire (Frantext) ou journalistique (corpus du *Monde*, corpus français de l'Université de Leipzig). La facilité d'accès aux ressources conditionne les choix effectués. Il était plus simple d'engager le travail en commençant par des textes typographiés et non des manuscrits, par des documents directement lisibles pour ceux qui en assuraient le traitement plutôt que sur des textes médicaux par exemple, sans que ces données soient pour autant exclues de l'investigation.

La majorité des documents sonores conservés dans les phonothèques, en dehors de la musique, concernaient des langues exotiques – où ils suppléaient à l'absence d'écrit – ou une parole publique officielle. La création de corpus oraux a transposé les pratiques des linguistes de terrain et des dialectologues dans les échanges ordinaires, urbains, en se rapprochant des méthodes de la sociologie.

La situation diffère selon les langues et les pays. Les corpus du français semblent moins variés que ceux d'autres langues romanes. L'exclusion d'une partie des usages reflète la diversité des rapports de domination linguistiques. Le catalan et l'espagnol, le portugais européen et le brésilien, l'italien et ses dialectes, le français confronté à la distance entre langue écrite et langue parlée, autant de cas de figure qui nourrissent les discussions sur la définition d'un « corpus de référence ». Les réponses ne sont pas les mêmes d'un pays à l'autre. Elles aboutissent parfois à des résultats paradoxaux : LIP (Corpus *Lessico Italiano Parlato*), qui cherchait à inventorier les différents usages de l'italien parlé, a mis en évidence une propension à la standardisation.

2.3 Conditions d'exploitation

La disponibilité est un sujet d'achoppement majeur, soit que les chercheurs (ou les institutions) n'en permettent pas la diffusion, soit que la conservation n'ait pas été assurée (cf. les enquêtes du *Groupe Aixois de Recherche en Syntaxe* – GARS). Les autorisations – qu'elles relèvent du droit d'auteur ou de la protection des témoins – constituent un obstacle d'autant plus difficile à résoudre dans le cas des corpus oraux que l'anonymisation se heurte aux possibilités de reconnaissance vocale et que les enquêtes les plus anciennes n'ont pas sollicité le consentement des personnes enregistrées. Il en va de même pour les métadonnées. La possibilité d'associer les enregistrements à des locuteurs, des situations, une date, etc. exploitables sous forme de base de données confronte la collecte des informations nécessitées par la recherche à des prescriptions légales contraignantes.

Une autre difficulté est inhérente aux formats et aux outils. Qu'il s'agisse de matériel ou de support, de logiciel ou de langage, l'obsolescence des équipements et des systèmes représente un défi. Le choix de formats et d'outils disponibles et pérennes apparaît encore plus complexe dans le cas de documents sonores où la source première est elle-même soumise aux risques de disparition ou d'inaccessibilité.

3. Usages de l'annotation

Dans l'acception courante du terme, le caractère second de l'annotation et son interpolation la distinguent de la transcription d'une part, des métadonnées d'autre part. [Leech 2004] a proposé de la situer en opposant {transcription vs annotation} et {représentation vs interprétation}.

En linguistique, l'annotation a pour finalité d'intégrer au corpus des informations au moyen de descripteurs qui permettent de rendre plus efficientes, voire simplement possibles, les requêtes effectuées pour une recherche donnée (exemple 1).

Exemple 2 : Image d'annotation

```
<anchor id="u-trigger-3-start" type="AnalecDelimiter" subtype="UnitStart"/>
il m'a dit
<anchor xml:id="u-trigger-3-end" type="AnalecDelimiter" subtype="UnitEnd"/>
<anchor xml:id="u-target_portion-3-start" type="AnalecDelimiter" subtype="UnitStart"/>
il travaillait pas
<anchor xml:id="u-target_portion-3-end" type="AnalecDelimiter" subtype="UnitEnd"/>
```

3.1 Format de codage et consignes

La première étape de l'annotation consiste à spécifier les données mobilisées par l'hypothèse de travail. Certains contenus sont plus souvent utilisés que d'autres (identification des *Parts of speech* (POS), des morphèmes, de traits sémantiques…). Aucun n'est obligatoire ou exclu dès lors que des propriétés de langue sont consignées.

Une fois circonscrit le type de données, on fixe les conventions qui les figureront dans un format qui doit être à la fois :

- distinctif (définir autant de classes que la recherche le nécessite),
- extensif (développer l'annotation au degré de granularité requis),
- unitaire (un même phénomène doit être décrit de la même façon),
- économique (aucun élément de l'annotation ne doit être redondant ou superflu),
- explicite (chaque élément de l'annotation doit être identifié dans un document associé).

Le codage apparaît d'autant mieux adapté qu'il est :

- limité en nombre de caractères,
- hiérarchisé dans la série des informations qu'il livre,
- accessible (il représente une économie de temps en apprentissage et mémorisation).

En pratique, cette dernière condition incite à l'utilisation d'abréviations bien répertoriées (par exemple /N = nom/ dans un étiquetage en POS).

L'annotation est *time consuming*. Elle peut être effectuée par d'autres personnes que le chercheur du fait de son caractère répétitif. L'investissement en temps doit être amorti au moment de l'exploitation dont, tendanciellement, la vitesse d'exécution est inversement proportionnelle au temps de préparation comme le montrent *Computational Analysis of Present-Day American English* [Kučera/Nelson 1967] sur le Brown Corpus qui a exploré une variété d'aspects linguistiques, psychologiques, statistiques et sociologiques du corpus ou les analyses sociologiques sur le corpus ESLO [Bergounioux 2016]. Le gain est d'autant plus appréciable que le corpus est étendu et que l'enrichissement par les annotations peut être exploité dans d'autres études et par d'autres chercheurs, linguistes ou non (statisticiens, informaticiens, sociologues…). Il s'y ajoute une objectivation des critères qui permet un examen critique argumenté et, rétroactivement, un retour sur les choix initiaux.

Les objectifs de la recherche, la sélection des données et la caractérisation des propriétés qui leur sont associées, le choix des chaînes de caractère qui les indicient en recourant au balisage imposent de définir le mode d'exécution des tâches. Les annotations en corpus

impliquent la rédaction d'un manuel ou d'un guide qui explicite, voire commente, les règles appliquées à leur confection, par exemple pour ESLO :
http://eslo.huma-num.fr/images/eslo/pdf/GUIDE_TRANSCRIPTEUR_V4_mai2013.pdf

3.2 Critiques et usages sociolinguistiques

Les critiques formulées concernent d'abord l'accès aux documents. L'annotation alourdit le fichier et le produit obtenu peut décourager par l'encombrement des indications et la distension conséquente de la linéarité du texte initial. Surtout, on a relevé une grande variation des annotations d'un programme à l'autre – résultant souvent d'orientations et d'approches opposées – et, à l'intérieur d'un même programme, entre les annotateurs. De plus, les erreurs commises à chacune des phases affaiblissent la fiabilité de l'ensemble, demandant, pour accroître la qualité des résultats, un travail de relecture et de correction qui compromet l'avantage comparatif attendu en terme de temps.

Les annotations sociolinguistiques étiquettent des caractéristiques liées au changement (de langue ou de dialecte, d'usage ou de registre), des emplois non standard ou innovants par rapport à la norme et des formes typiques d'une culture ou d'une subculture, concernant :
- les traits dialectaux, le code switching…
- les niveaux de langue – comme les dictionnaires en ont répandu la pratique – ou sur des emplois caractéristiques d'un milieu social, d'une tranche d'âge,
- le marquage des erreurs (constructions erronées, hypercorrection…),
- les modalités de catégorisation subjectives (dénominations appréciatives et dépréciatives, opérations de classement liées à un groupe social défini…) et les projections d'identité,
- les marques d'adresse et les reformulations etc.

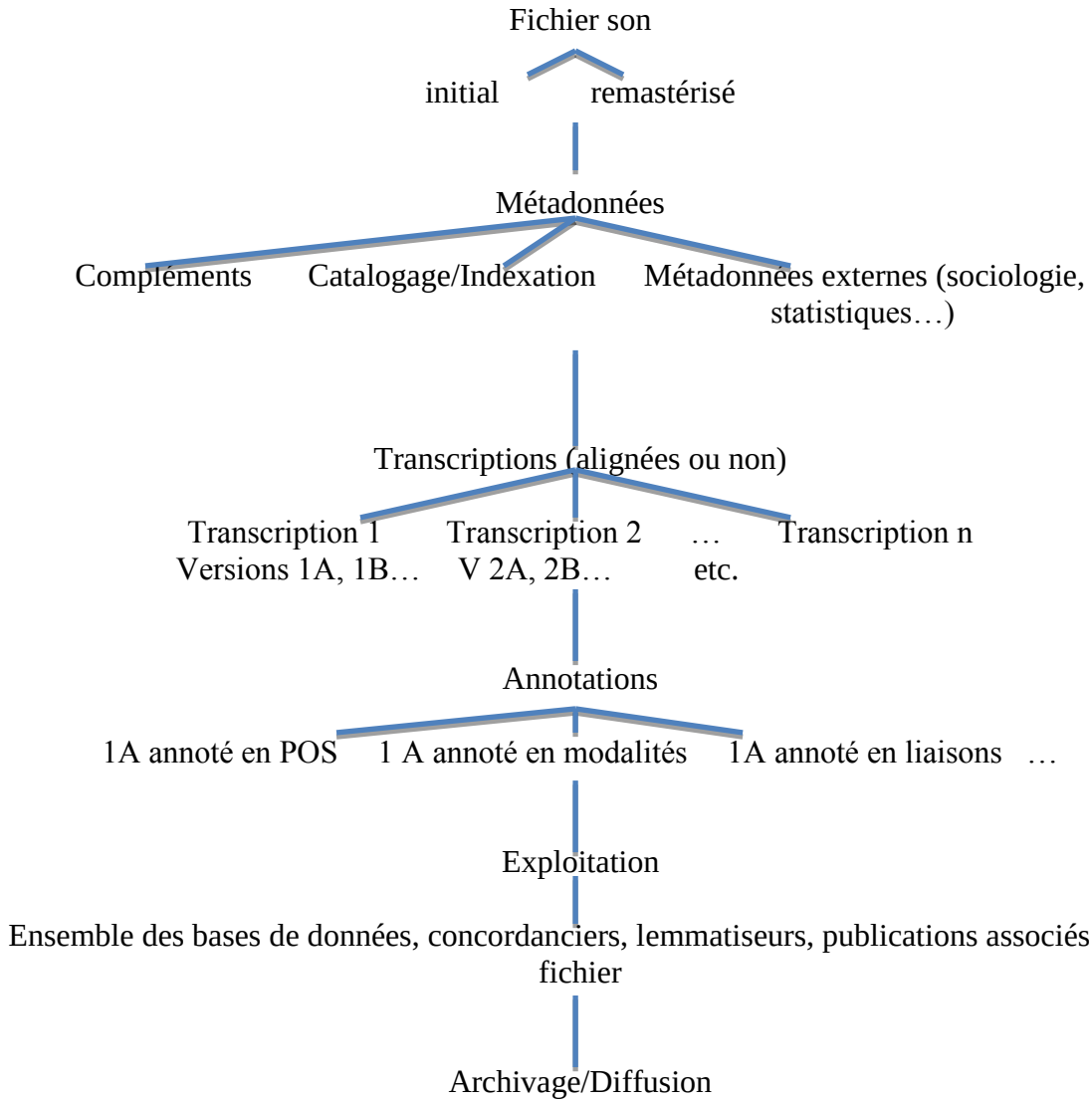
Dans tous ces cas, deux paramètres entrent en jeu. Le premier concerne la façon dont des propriétés de langue s'actualisent en discours. Le second se focalise sur les modalités de verbalisation d'une formation sociale, de ses traits culturels. Quel que soit le contenu d'une annotation, tout codage s'inscrit à l'intersection de ces deux domaines.

4. Consignes d'annotation

Certaines annotations sont spécifiques à l'écrit (orthographe), d'autres à l'oral (e.g. la réalisation de la liaison), d'autres communes aux deux (POS, modalités, niveau de langue…). Malgré leurs différences, elles sont homogènes quel que soit le type de collecte des données – de l'aspiration de la Toile sur une langue à large diffusion jusqu'à une enquête courte auprès d'un groupe linguistique de taille restreinte sans tradition écrite.

4.1 Génération des fichiers

Tableau 4 : La génération des fichiers et la place de l'annotation



En règle générale, l'annotation est conçue comme une opération intervenant a posteriori sur un texte ou une transcription, après qu'ont été renseignées les métadonnées, et avant les exploitations scientifiques. Cette conception situe l'annotation en aval de la mise en forme des données et en amont des différentes exploitations. Celles-ci, tributaires de l'instrumentation, sont réalisées à partir d'une version de transcription qui peut continuer à évoluer parallèlement (versionnage). Le partage proposé par [Habert 2005] entre *instruments* et *outils* est arbitré par les possibilités d'automatisation de l'annotation et le système de gestion des bases de données (SGBD).

La correspondance des produits peut être complétée de la manière suivante (les différents fichiers sont codés par le premier chiffre, les versions d'un même fichier par l'ajout d'une lettre, les enrichissements par un chiffre après la lettre).

Tableau 5 : Opération et production

Opération	Produit
collecte/numérisation du signal	Fichier audio numérique

catalogage/indexation/métadonnées	Fichier texte 1
transcription/codage/alignement	Fichier texte 2 (versionnage 2A, 2B, 2C…)
annotation/instrumentation	Fichier texte 2 enrichi (ex. 2B1, 2C1, 2C2…)
outillage/analyse	Travaux de recherche
conservation/diffusion	Fichier textes 1 et 2 (avec hyperlien aux analyses)

4.2 Trois types d'annotation

S'il est possible de procéder aux analyses, à l'archivage ou à la mise à disposition d'un corpus sans réaliser d'annotations, peut-on envisager d'anticiper certaines annotations au moment de rédiger les métadonnées ou de transcrire les enregistrements ? Il y a une intersection entre ce qui relève des métadonnées et des annotations. Le découpage thématique (*topics*) peut apparaître dans un fichier distinct ou être inséré en tant qu'annotation. En général, les métadonnées incluent des informations « externes » globales (identité des locuteurs, date de l'enregistrement, formats numériques, titulature…) ; les annotations privilégient des informations internes portant sur des éléments courts inclus dans le signal (un bruit, un clic), et concernent des unités qui vont de l'amorce et du mot (avec de l'étiquetage et de la lemmatisation) à la phrase (coréférence, arborescence) ou au tour de parole.

Selon le degré d'intrication avec le fichier de base, on distingue trois types d'annotations.

1. EMBARQUE (*embedded/online*)

Exemple 3

```
<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>
```

(extrait de http://ircom.huma-num.fr/wiki/lib/exe/fetch.php?media=myautolinks:exemples_codage_teiml.pdf

2. DEBARQUE (*standoff/standalone*)

Exemple 4

```
<texte>alors ils faisaient comme ça euh et je me suis
rendue compte que ça n'allait pas parce que moi je je lisais et je lisais un rigue
euh la première ligne</texte>
```

```
<annotatedU end="#T175" start="#T174" wh="spk1" xml:id="au72"
xmlns:xi="http://www.w3.org/2001/XInclude">
   <u>
      <seg xml:id="s343">
         <xi:include href="texte.xml" xpointer="xpointer(substring(., 1, 32))"/>
         <pause type="short"/>
         <xi:include href="texte.xml" xpointer="xpointer(substring(., 33, 48))"/>
      </seg>
      <anchor synch="#T183"/>
      <seg xml:id="s344">
         <xi:include href="texte.xml" xpointer="xpointer(substring(., 82, 48))"/>
      </seg>
      <seg xml:id="s345">
         <xi:include href="texte.xml" xpointer="xpointer(substring(., 130, 30))"/>
      </seg>
      <spanGrp>
         <span type="com" target="#344">mot italien = ligne</span>
      </spanGrp>
   </u>
</annotatedU>
```

3. INTERLINEAIRE (*muti-tiered/interlinear*),

Exemple 5

Insérer ici une image tirée de PRAAT ou d'ELAN

L'annotation embarquée permet d'obtenir un document autosuffisant car l'ensemble des annotations sont regroupées au sein du même document, ce qui peut contrevenir à la lisibilité. À l'opposé, l'annotation débarquée assure une meilleure lisibilité du document « maître » mais rend les annotations débarquées dépendantes du document qui les contient, ce qui exige une coordination entre les deux et rend la maintenance plus complexe. En effet, la dépendance repose sur des liens logiques (positions ou identifiants) qui doivent être préservés à travers toutes les modifications qui peuvent advenir dans les données. Quant aux annotations multi-tires ou interlinéaires, elles seront utilisées dés lors que que l'on veut disposer conjointement, dans une annotation, de plusieurs points de vues ou de plusieurs niveaux d'analyse pour un même phénomène.

Ne serait-ce que par la segmentation en mots ou l'application de règles qui ne sont pas représentatives des usages orthographiques de la personne interviewée, une transcription non phonétique anticipe l'annotation. En quoi est-il justifié de noter l'accord pluriel sur le participe passé dans « les difficultés que ça m'a faits » pour un locuteur de faible niveau scolaire ? La réponse dépend des principes énoncés dans le manuel de transcription correspondant au corpus.

4.3 Consignes

L'annotation d'un corpus doit répondre à un certain nombre d'exigences interdépendantes :

1°) Séparer le fichier sur lequel s'effectue le travail (fichier de transcription et fichier audio) et les fichiers annotés. La *séparabilité* implique la *réversibilité* : d'un fichier annoté, on doit pouvoir revenir à la version précédente.
2°) Assurer la *reproductibilité* des fichiers et leur conservation dans un état qui en garantit la *validité*, c'est-à-dire l'identité et la lisibilité dans les formats évolutifs de l'informatique.
3°) Garantir l'*accessibilité* des données qui doivent être retrouvées rapidement dans l'état souhaité (*traçabilité*) et sous une forme qui en rende aisée la manipulation en mettant à profit l'intuitivité et l'affordance.
4°) Veiller à la pérennité de la ressource et à son interopérabilité.
5°) Faciliter les tâches d'outillage et d'*instrumentation* et permettre l'*incrémentation*.

Ainsi conçue, l'annotation se présente comme un processus qui, à chaque palier, étend la ressource en la transformant sans jamais effacer un état précédent (versionnage).

5. Formats d'annotation

L'annotation est une opération au service d'une fin. Elle ne représente pas un état premier de la donnée : elle est établie en fonction d'une ressource qui a déjà fait l'objet d'une préparation. Elle se visualise de différentes manières, souvent – mais non exclusivement – sous forme écrite, *linéaire* et *segmentée*. Le fichier texte offre une mise en page, un volume global (nombre de caractères ou espaces), des métadonnées etc. Ainsi, l'annotation constitue à la fois une méthode d'analyse et un instrument du TAL (Traitement Automatique des Langues). Le spectre de représentation va d'une figuration analogique (pour les analyses phonétiques) à des notations symboliques (pour des études syntaxiques ou stylistiques), les études sociolinguistiques privilégiant un alignement de la transcription sur le signal et une extension des commentaires à l'intérieur des balises. Ce compromis permet d'obtenir un maximum de lisibilité en distinguant un palier de restitution orthographique et des indications récupérables par requête pour conduire les analyses.

5.1 Les principes

[Leech, 2005] a énoncé quatre principes sur lesquels devrait se régler l'instrumentation d'un corpus :

- donner accès à la documentation sur les données et les traitements ;
- expliciter les choix qui ont décidé des opérations ;
- assurer la reproductibilité des résultats (l'instrument doit permettre, dans des conditions identiques, d'aboutir aux mêmes conclusions) ;
- pouvoir être vérifié par des procédures indépendantes de l'observateur (une difficulté récurrente destinée, entres autres, à surmonter les désaccords entre annotateurs).

L'annotation est exécutée en suivant des conventions définies au sein d'un format de représentation qui spécifie :

- la segmentation en unités élémentaires,
- leur organisation à l'intérieur du document,
- la manière d'intercaler de façon réversible des informations (métalinguistiques) exploitables automatiquement.

Sans que la distribution soit absolue, on parle plutôt de « balise » quand on se réfère à la partie technique des opérations de traitement et d'« annotation » quand on traite des orientations scientifiques du chercheur.

5.2 Codage et contenu

L'annotation peut être réalisée, entre autres, en XML/TEI, une pratique répandue aujourd'hui, aussi bien pour le *document* (fichier) que pour le *schéma* (cf. infra « validation ») avec une *entête* (header) qui introduit soit le corpus, soit un de ses fichiers.

On hiérarchise à différents niveaux :
- le fichier pris dans sa globalité,
- sa structure d'ensemble,
- sa décomposition en paragraphes (et sous-paragraphes),

en renseignant quatre items :
- la description du fichier qui équivaut à la référence bibliographique,
- l'indication d'origine qui stipule la relation au texte source,
- la caractérisation du texte contenant les indications pertinentes pour sa manipulation (à commencer par la langue dans laquelle le texte est rédigé),
- le versionnage.

Exemple 6 : Métadonnées
Métadonnées TEI extrait de CLAPI (http://clapi.ish-lyon.cnrs.fr/V3_TEI.php)

```
<teiHeader xml:lang="fr">
  <fileDesc>
    <titleStmt>
      <title> Réunion de conception en architecture - mosaic ~ Mosaic - architecture ~ Mosaic
        - architecture - xml </title>
      <principal>Detienne Françoise</principal>
      <principal>Traverso Véronique</principal>  […]
      <respStmt>
        <resp>collecté par</resp>
        <name>Detienne Françoise</name>
        <name>Visser Willemien</name>
      </respStmt>                                  […]
      <respStmt>
        <resp>préparé et balisé par</resp>
        <name>CLAPI - Equipe Médiathèque</name>
      </respStmt>
    </titleStmt>
    <publicationStmt>
      <publisher>Groupe ICOR/ Plateforme CLAPI</publisher>
      <pubPlace>http://clapi.univ-lyon2.fr</pubPlace>
```

```
<availability status="restricted">
  <licence target="http://clapi.univ-lyon2.fr/V3_CGU.php">
    <p>Conditions générales d'accès pour ce document</p>
    <p>Copyright © ICAR. Tous droits réservés.</p>
    <p> Enregistrement vidéo d'une durée de 1h18m45s téléchargeable sous convention
      de recherche </p>
    <p> Transcription mosaic - architecture - adaptée CLAPI au format doc en
      téléchargement libre </p>
    <p> Transcription mosaic - architecture - clan au format clan - ca ou cha en
      téléchargement libre </p>
    <p>Transcription requêtable par les outils librement</p>
    <p>Agrément CNIL de Clapi numéro : 2-12064</p>
  </licence>
</availability>
</publicationStmt>
```

5.3 Formatage, balisage et validation

Le formatage inclut deux opérations :

- le prétraitement qui vise à limiter au maximum les risques de bruit et de silence – dans l'acception qu'ont ces termes en théorie de l'information – en réalisant le nettoyage des données et leur normalisation ;
- la formulation de l'ensemble des prescriptions qui régissent l'annotation.

A côté du balisage utilisé dans le traitement de texte, le balisage destiné à étiqueter des chaînes de caractères ajoute de l'information (phonétique, grammaticale, sémantique, sociolinguistique…) et fournit sur ces propriétés linguistiques des indications récupérables automatiquement.

La transcription du corpus et le balisage sont linéairement solidaires. La réalisation d'une annotation arborescente, hiérarchique, ne peut être insérée à cette étape, ce qui en reporte l'exécution à une étape ultérieure du traitement.

Si l'annotation est encodée en XML, un premier niveau de validation syntaxique peut s'effectuer à l'aide d'outils de validation standard qui permettent d'établir la conformité d'un document à un modèle définissant les règles de sa grammaire et de son vocabulaire. Au nombre des langages de définition de ces modèles, on citera XML-Schema, Schema Relaxng, Schematron, DTD.

5.4 Annotation et instrumentation

Les annotations, en respectant la linéarité dans la distribution de l'information, présentent l'avantage de permettre une exploration systématique des données sur tout ou partie d'un ou de plusieurs fichiers et en fonction de chacune des requêtes ajustées à son/leur contenu.

Pour prolonger l'investigation, il faut recourir à d'autres formes d'exploitation tels qu'on peut en produire par des treebanks, des concordanciers ou encore par inférence ou

unification (Web sémantique).

Exemple 7 Concordancier

Concordance du mot « poêle » – CoCoON http://cocoon.huma-num.fr/exist/crdo/

... vous faites cuire ça dans la dans la	poêle	ou alors vous avez euh l'omelette ...
... n on on verse tout ça dans la dans la	poêle	et puis on tourne jusqu'à temps que ...
... ir de manière uniforme dans la dans la	poêle	et voilà soit tu la retournes si t ...
et on les verse dans la dans la	poêle	une fois que on laisse cuire et après on ret ...
... urre dans euh dans euh dans la dans la	poêle	on fait chauffer le beurre on on ...
... es battre et les verser dans la dans la	poêle	
... avez versé le les oeufs dans la dans la	poêle	euh vous euh vous quand ça prend c ...
... erse mon ome- mes oeufs dans la dans la	poêle	
... euh progressivement dans la dans la	poêle	alors j'aime pas trop qu'elle soit ...
... hop après je je verse sur ma dans la	poêle	qui a bien qui est bien chaude
et je verse ça dans la	poêle	et je remue avec la fourchette
... reposer puis après on verse ça dans la	poêle	avec du beurre ou de l'huile puis ...
... rchette et puis nous jetons ça dans la	poêle	euh après avoir fait fondre notre be ...
... bats puis bah ma foi on met ça dans la	poêle	et je remue la poêle avec euh je m'a ...
et on met tout ça dans la	poêle	on bat d'abord évidemment et on met tout ça dans la ...
... d évidemment et on met tout ça dans la	poêle	
... néral sans matière grasse donc dans la	poêle	pour euh je me sers donc de la mat ...
... rais ensuite le le le liquide dans la	poêle	une fois qu'elle est un petit peu c ...
... ttue je mets du beurre dans le dans la	poêle	je fais dorer mon beurre pas bru ...
... elette on met les oeufs dans le dans la	poêle	on soulève le l'omelette de temps e ...
... e vous les faites assez cuire dans la	poêle	et puis vous mettez euh les oeufs ...
... des morceaux de pomme de terre dans la	poêle	pour faire une omelette euh voire d ...
... du poivre un morceau de beurre dans la	poêle	et c'est tout
... je mettais un morceau de beurre dans la	poêle	hm
... on fait fondre un peu de beurre dans la	poêle	on fait chauffer bon auparavant on ...

Pour le traitement des corpus, un recours à l'instrumentation s'impose. Une fois défini le type de document guidant le choix d'un modèle (EAD, TEI, CES ou autre), et en fonction de la transcription réalisée (que soit privilégiée une restitution phonétique avec PRAAT ou une transcription alignée sur le signal avec Transcriber, par exemple), différents outils d'annotation sont disponibles. On mentionnera, parmi les plus utilisés :

- ANVIL dédié à l'annotation des vidéos,
- ELAN (*Eudico Linguitic Annotator*) pour des fichiers multimédias,
- EXMARaLDA (*Extensible Markup Language for Discourse Annotation*) pour les corpus oraux. EXMARaLDA est à la fois un outil de transcription, d'annotation, de gestion, de requête et d'analyse.

Exemple 8 : Représentation sous EXMARaLDA

Insérer une image d'écran

La diversité des formats et leur production par des équipes plus souvent que par des institutions fait naître, au-delà des questions d'accessibilité, de disponibilité et de libre utilisation, des interrogations essentielles sur leur interopérabilité et leur pérennité. Une question récurrente concerne les procédures de codification du métalangage afin que soient repérés et exploités des éléments dont la pertinence s'étage sur différents niveaux.

6. La construction des annotateurs

6.1 Annotation, annotabilité, méta-annotation

L'annotation d'un fichier est conditionnée par les analyses qu'elle rend efficientes – ou qu'elle invalide. On identifie les occurrences correspondant aux critères définis, phonologiques, morphologiques, syntaxiques, sociolinguistiques ou autres. En partant des données, et dans le respect des conventions préétablies, les unités sélectionnées sont inscrites sous une forme qui permette l'interrogation par requêtes. Deux opérations sont effectuées simultanément :

- la démarcation : segmentation des éléments à annoter (une chaîne linéaire de symboles),
- la sélection d'une propriété de cet élément, un de ses *attributs*, i.e. une paire nom/valeur.

Il existe deux méthodes : par système de règles ou bien par système d'apprentissage supervisé.

On appelle *annotabilité* l'ensemble ouvert des propriétés éligibles à ce type d'opération, en intégrant dans ce concept les traitements qui en facilitent la mise en œuvre, par exemple le découpage en syntagmes dans un corpus oral. L'annotabilité serait, dans une perspective informatique, l'équivalent de l'*observable* d'un point de vue linguistique. Dans le cas d'annotation des annotations, on parlera de *méta-annotation*.

6.2 Opérations

Les éléments peuvent être segmentés à différents paliers, depuis les unités constitutives (lettres, chiffres, symboles, ponctuation, espace) jusqu'aux unités thématiques (*topics*) en passant par les morphèmes, les mots, les syntagmes, les propositions et les phrases. Néanmoins, c'est le mot qui reste central, au moins dans les langues non agglutinantes.

On distingue trois étapes :

- la décomposition en chaînes de caractères (*tokenization*). Dans le cas des corpus oraux, il s'agit d'une reprise de la tâche réalisée par le transcripteur ;
- la *lemmatisation* : *avions* est-il la 1re personne du pluriel de l'imparfait de l'indicatif du verbe *avoir* ou le pluriel du nom *avion* ? L'unification des formes (le lemme *aller* pour *vont*, *aille* ou *irions*) et leur classement en parties du discours (POS) dessine en creux

la structure morphologique. Particulièrement en français, la morphologie des langues romanes, héritières des graphies latines et limitées dans l'exploitation des oppositions prosodiques, est plus pauvre à l'oral qu'à l'écrit.

- le *traitement* : à partir de la caractérisation linguistique des unités, il est possible de déterminer les coréférences et d'identifier les entités nommées [Eshkol et al. 2012], d'exécuter des analyses syntaxiques et sémantiques (modalités), de structurer en thèmes…

6.3 Difficultés

Au nombre des difficultés, on mentionnera la compatibilité des outils informatiques, des théories linguistiques – en particulier de celles qui ont faiblement recours aux formalismes, comme la sociolinguistique – et de données non scripturales, un problème récurrent en TAL. De plus, si l'exécution n'est pas effectuée par le chercheur lui-même, il peut en résulter des dysfonctionnements supplémentaires qui requièrent une supervision et un contrôle.

Les annotations sociolinguistiques n'ont pas dégagé de consensus en raison du caractère éristique des concepts sociologiques. Leur fonction de critique ou de sociodicée [Bourdieu 1983], ne permet pas d'établir un accord au-delà des classifications statistiques par âge, genre, niveau de revenu ou de diplôme. Les divergences d'un pays à l'autre dans la définition des professions et catégories socio-professionnelles est emblématique et rend difficile l'adoption d'une nomenclature unifiée dans les métadonnées.

Les jugements sociaux contreviennent aux exigences d'objectivation. Ainsi, l'assignation d'une forme d'expression comme « avoir le seum » à « jeunes des cités » va au-delà d'une caractérisation par l'âge ou le lieu de résidence. Si, en termes d'instrumentation, la façon d'inscrire un attribut est une, une convention définissant de façon consensuelle les attributs sociaux ne peut être posée, sauf à reproduire des classifications bureaucratiques ou à entériner une vision unilatérale du monde social.

7. Types d'annotation et niveaux d'analyse

Trois éléments concourent au conditionnement des annotations :
- les logiciels utilisés,
- les données d'entrée,
- l'appareil conceptuel qui, selon le type d'annotation retenu, détermine la qualité des réponses aux requêtes.

En linguistique, ces requêtes s'étagent selon le niveau de l'analyse choisi [Benveniste 1966] en fonction de l'état initial des données (textes, paroles, images, multimodal). Après un exemple sur des données écrites, on étudie l'annotation de différentes unités.

7.1 Graphies

Le travail sur manuscrit avait attiré l'attention sur l'importance des graphies et sur les variantes, les abréviations, les ligatures, les omissions… Initialement, il s'agissait de parvenir à un texte nettoyé de ses scories, de ses interpolations, de suivre les différentes versions d'un état primitif, connu ou non, qu'on s'efforçait de restituer en deçà de ses avatars. Pour des raisons économiques et sociales – la prévalence d'une conception religieuse, la revendication d'une filiation gréco-latine –, la transmission médiévale s'est cantonnée à la reproduction d'un petit nombre d'ouvrages avec une hétérogénéité des pratiques scripturales (les ateliers de copistes).

La diffusion de l'alphabétisation et la réduction du prix du papier, la transformation des échanges (l'apparition de la poste) et des comportements ont transformé les usages. L'étude s'est déplacée des codex à des corpus de « peu lettrés » comme les cahiers de doléance de 1789 [Branca-Rosoff/Schneider 1994] ou la correspondance des poilus [Steuckardt 2015]. Leur traitement exige une restitution orthographique normée, voire des restitutions en français standard.

7.2 Phonétique, prosodie, phonologie

La notation en API est souvent utilisée à petite échelle. Peu de corpus de grande taille y ont recours pour l'ensemble des transcriptions. La technicité de l'opération est consommatrice de temps, chaque choix prête à discussion (dans « meuf », est-ce un /œ/ ou un /ø/ qui a été réalisé ?) et le résultat n'est pas aussi lisible qu'un fichier orthographié. Après l'utilisation de notations semi-conventionnelles (e.g. « ch'crois » pour « je crois »), on privilégie aujourd'hui des transcriptions normalisées alignées sur le signal qui permettent de retrouver à coup sûr toutes les formes d'un mot. Alors que, dans l'exemple précédent, le lemme « je » ne connaît qu'une variation, « j' », une troisième (« ch' ») devrait être ajoutée pour éviter du silence à une requête portant sur les pronoms sujets de première personne.

Les indications phonétiques, qu'elles concernent les prononciations (e.g. la réalisation du schwa) ou la prosodie (Rhapsodie), figureront dans l'annotation dès lors qu'elles ne sont pas prédictibles. L'intégralité du corpus Perseval [Gomez Molina et al. 2007] est segmenté en groupes prosodiques et, à côté de C-Oral Rom, on mentionnera [Prieto/Roseano 2007].

A la différence de la phonétique qui, à partir du signal, est directement accessible et mesurable objectivement par l'instrumentation, à la différence de la morphologie qui est d'ordre métalinguistique, la phonologie est une compétence mentale des locuteurs. Aussi, les utilisations qui en sont faites concernent en priorité des traitements de surface corrélés à la morphologie, soit dans l'actualisation d'un indice phonétique (les liaisons dans PFC, le placement de l'accent tonique), soit dans la recension des différentes prononciations correspondant à la FSJ (forme sous-jacente) des unités lexicales [Bergounioux 2016].

L'annotation effectuée sur une transcription part des conventions qui ont été appliquées pour la réaliser :

- usage ou non de marques de ponctuation, de majuscules, d'italique,
- indication des bruits vocaux (rires, toux…),
- notation des pauses, des disfluences etc.

Quelques exemples d'utilisation de l'annotation des phénomènes phonétiques ou phonologiques en sociolinguistique :

- l'identification des phénomènes de réduction consonantique dans les variétés régionales d'italien [Vallone et al 2002],
- la caractérisation prosodique des genres discursifs en italien [Giordano/Savy 2003] et en français [Beliao et al. 2014],
- l'analyse de la variation diatopique de la prosodie de l'espagnol européen et latino-américain [Prieto/Roseano 2010],
- l'analyse de la variation diastratique et diatopique de la réalisation de la liaison en français [Durand et al. 2002 ; 2009].

7.3 Morphologie, lexique

Le TAL a privilégié le mot comme unité de traitement. L'une des premières applications concrètes de l'informatique aux langues a été la constitution d'une linguistique quantitative qui reprenait les intuitions de [Zipf 1935] afin de produire des statistiques lexicales (listes de fréquence :

- TLF [Imbs/Quemada 1971-1974], [Guiraud 1954],
- LIP [De Mauro et al. 1993],
- Colfis [Bertinetto et al. 2005],
- NVDB [Chiari/De Mauro 2014],
- CLUC [Do Nascimiento 2001],
- Frecuencias del español [Almela et al. 2005].

Au fur et à mesure du développement des ressources, les annotations morphologiques et lexicales se sont enrichies, en particulier dans la confection automatique de lexiques et la production de concordanciers. Au nombre des premières exploitations des données, on peut citer la réalisation de listes lemmatisées organisées par ordre décroissant de fréquence et utilisées pour la recherche lexicale statistique, pour le développement de ressources didactiques et pour la production de dictionnaires à destination du TAL comme de l'analyse sociolinguistique.

Cet accès au corpus pose des questions aussi bien sur la valeur des unités considérées comme pertinentes que sur les éléments d'analyse. Le statut du mot en tant que concept scientifique reste problématique. Sa définition est essentiellement scripturale, peu compatible avec la nature même de la langue qui est perçue comme un flux verbal sans découpage tranché. La littérature scientifique mentionne :

- les morphèmes flottants (le préfixe dans « repolir » ou « non-agréer », le suffixe dans « ordinatouille » attesté sur Internet)
- les regroupements (chunking) en syntagmes figés (« pour autant que », « condition nécessaire et suffisante »), en locutions, en proverbes,

- les phénomènes de grammaticalisation (« je sais pas » équivalant à « à peu près » dans « y avait je sais pas moi sept ou huit personnes »).

Par ailleurs, l'adaptation au TAL de catégories linguistiques constituées sur des textes écrits en alphabet dans une tradition logiciste fondée sur les langues de la famille indo-européenne occidentale limite les capacités de généralisation des annotations et leur pertinence.

La reconnaissance des entités nommées, à la frontière du lexique et de la syntaxe, est un exemple de tâche fréquente en fouille de données et en documentation automatique pour la construction d'ontologies et l'établissement du Web sémantique. Cette application revêt une importance particulière en sociolinguistique où certaines désignations complexifient l'identification en fournissant des indications d'autant plus significatives sur la relation du locuteur au contenu de son discours (« le Président des riches », « ma fifille à moi »).

L'annotation des phénomènes morphologiques et lexicaux a permis de conduire des analyses sociolinguistiques qui ont conduit :

- à relativiser les préjugés sur la présence de dialectalismes et régionalismes en italien parlé – beaucoup moins fréquents qu'on ne le croyait [De Mauro et al. 1993],
- à mesurer l'étendue de la variation grammaticale en espagnol populaire [Fernández-Ordóñez 2011],
- à explorer les phénomènes de code-switching lexical entre espagnol et catalan en Catalogne espagnole [Diaz 2009],
- à étudier la diffusion des mots familiers, du tutoiement et des formules de politesse en français contemporain [Beeching 2012].

D'autres exemples de phénomènes morphologiques et lexicaux annotés dans une perspective sociolinguistique dans une perspective diachronique courte :

- la diffusion des mots familiers, du tutoiement et des formules de politesse en français contemporain [Beeching 2012],
- la composition lexicale avec différenciation diatopique et diaphasique en italien contemporain [De Mauro 2014].

7.4 Syntaxe

Une détermination des POS tenant compte des relations morphologiques rend possible une annotation syntaxique des propositions suivant une représentation linéaire, un procédé similaire aux parenthétisations de Hockett.

Quand les données incluent des unités supérieures au mot (par exemple en tours de parole), les séquences découpées figurent avant tout des projections de l'analyse syntaxique (et sémantique). Les treebanks sont la représentation la plus courante aujourd'hui.

Quelques exemples de Treebank :

Catalan	Cat3LB
Espagnol	Cast3LB

Français	French Treebank, Rhapsodie
Italien	ISST (Italian Syntactic-Semantic Treebank)
Latin	Latin Dependency Treebank
Portugais	Projecto Floresta Sintáctica
Roumain	RDT (Romanian Dependency Treebank)

Ces ressources et leur site sont identifiés sur Internet, entre autres dans le catalogue ELRA : http://catalog.elra.info/index.php?language=fr
On peut ajouter à ces treebanks syntaxiques, les corpus C-Oral Rom (Cresti & Moneglia 2005) et le corpus Rhapsodie (Lacheret et al. forthcoming) qui intègrent une annotation des structures « macrosyntaxiques » orales du français, de l'italien, de l'espagnol et du portugais (C-Oral Rom) et du français (Rhapsodie).

Autant le mot peut être considéré comme l'unité de base manipulée par les informaticiens, autant les analyseurs syntaxiques seraient les outils privilégiés des linguistes. Les exploitations linguistiques et sociolinguistiques des annotations syntaxiques, en plus de l'annotation des structures macrosyntaxiques dans C-Oral Rom et Rhapsodie déjà mentionnés, se retrouvent dans :

- l'étude de la variation syntaxique dans les dialectes des langues romanes [Sauzet/ Dagnac/Sportiche 2015]
- l'analyse de la variation diaphasique des structures de dépendance en français parlé [Pietrandrea forthcoming ; Kahane/Gerdes forthcoming],
- l'étude de la variation diaphasique des structures macrosyntaxiques en français, anglais, espagnol et portugais [Cresti/Moneglia 2005 ; Pietrandrea forthcoming],
- l'analyse de la distribution diastratique de certaines structures syntaxiques – e.g. le *déqueismo* en espagnol valencien de Valence [Gomez Molina 1995].

7.5 Sémantique

Au-delà de ce qui est de facto résolu, volontairement ou pas, par la transcription (le choix entre « en dix ans tout ça va mieux » vs « en disant tout ça va mieux »), la première intervention sémantique effectuée par les annotations concerne la désambiguïsation en POS pour des termes homographes relevant de la même catégorie. « Vers » est-il une préposition ou un nom et, si c'est un nom, s'agit-il d'une unité métrique ou du pluriel du lemme *ver* ? Une partie de la tâche est facilitée par une indication, en métadonnée, du domaine de spécialité concerné ou par un lien établi entre les données et un dictionnaire de terminologie.

Les annotations sémantiques sont également utilisées pour :
- l'analyse des rôles thématiques, des classes verbales et nominales,
- l'organisation de la dimension temporelle,
- le traitement des modalités ou des métaphores,
- l'étude de la structure argumentale,
- l'exploitation de la structure informationnelle.

On citera pour la modalité en italien et en français MODAL [Pietrandrea forthcoming], en portugais [Avila et al. 2015], SenSem Corpus en espagnol et catalan [Fernández/Vásquez 2014]. On mentionnera également pour une annotation sémantique en sociolinguistique concernant l'analyse de la pratique de construction d'un savoir épistémique partagé dans des conversations spontanées [Pietrandrea forthcoming].

7.6 Discours

Au niveau du discours, une réorganisation hiérarchisée apparaît dans l'annotation des coréférences et des anaphores associatives. La discontinuité de la chaîne sonore qui les caractérise et les indices qui en permettent le rappel ont constitué un obstacle à l'annotation automatique. Le regroupement d'unités adjacentes n'est plus opératoire, requérant d'autres conventions, par exemple pour la coréférence dans ANCOR-Centre [Schang et al. 2014].

Qu'il s'agisse de marquer les tours de parole ou d'établir une typologie des actes de langage, l'annotation intervient pour caractériser :

- les formes du dialogue et de la conversation (CID – Corpus of Interactional Data en français, PraTid en italien),
- la structure de l'information (IPIC – Information Structure Database pour l'italien et le portugais brésilien),
- les relations discursives (Annodis, français) et les marqueurs discursifs (pour une application aux corpus de français parlé Valibel, Clapi et Corpage [Bolly et al. 2015]).

La possibilité de transposer d'une langue romane à l'autre les formats d'annotation pose la question d'un palier intermédiaire entre des langues où prévalent d'autres modes de fonctionnement morpho-phonologiques (langues agglutinantes, langues à tons…) ou sémantiques (langues à classes, langues apophoniques…) et les langues indo-européennes et, au sein même des langues indo-européennes, entre les langues du groupe occidental. A l'intérieur de la strate que constituent les langues romanes, certaines différences sont manifestes, par exemple l'emploi du neutre en roumain ou les tableaux de conjugaison.

8. Critiques et perspectives

L'annotation sociolinguistique pâtit d'une absence de consensus, inhérente à la discipline et à sa fonction critique, quant aux catégories qui doivent figurer entre balises. Se pose la question de leur nécessité (qu'apportent-elles aux requêtes ?), de leur standardisation (les désignations varient d'une théorie à l'autre) et de leur pertinence (qu'ajoutent-elles à la description linguistique ?). Les caractérisations génériques (âge, sexe, profession etc.) sont en général portées dans les métadonnées en sorte qu'une annotation n'est requise qu'à la condition de concerner un phénomène potentiellement répétitif, une variation dans les occurrences dont l'explication serait liée à des usages différenciés, « distinctifs ».

8.1 Annotations sociolinguistiques

L'observation des phénomènes relevant de cette catégorie se concentre sur :

- les unités lexicales (terminologie professionnelle, langage « jeune » ou archaïsmes, argot et verlan…),
- certaines tournures syntaxiques, en particulier celles qui, fragilisées dans le système, sont utilisées de façon contrastée [Blanche-Benveniste 2010], par exemple sur les constructions relatives,
- les effets de l'interaction discursive (marques d'adresse, euphémisation…),
- les représentations (et autoreprésentations) collectives des agents et de leur environnement.

Inversement, peu d'annotations sociolinguistiques concernant la phonologie et la prosodie, sauf à ce qu'interviennent des considérations d'un autre ordre comme dans le cas de la liaison [Encrevé 1988] où se conjoignent la morphologie et la variation sociale. Quoiqu'il en soit, l'annotation sociolinguistique apparaît comme une donnée supplémentaire qui se greffe sur l'annotation grammaticale. L'alignement de la transcription sur le signal, conçu comme un moyen de fiabiliser l'analyse, a permis de surmonter l'homogénéisation créée par l'orthographe mais l'auditeur ne peut manquer d'ajouter ses jugements sur les voix et sur les agents.

8.2 Effet des champs disciplinaires et annotation de la variation

Au nombre des obstacles qui se dressent à l'encontre d'une approche intégrative, il faut tenir compte des attentes en matière d'applications et des conditions de constitution des communautés de chercheurs. Le TAL n'a pas été finalisé pour traiter les variations internes aux langues mais les variations entre langues en se fondant sur des constantes (décompte lexical, structures de phrases). Même quand la dispersion des réalisations rendait manifeste la disparité des productions, notamment en reconnaissance vocale, celles-ci étaient assignées préférentiellement à une différence inter-individuelle, secondairement à une origine géographique [Boula de Mareüil/Woehrling/Adda-Decker 2013]. Le cursus universitaire des informaticiens ne les prédisposait pas à s'adresser aux sociologues, et celui des sociologues aux informaticiens. Il en résulte qu'un très petit nombre de travaux ont été effectués dans ce domaine jusqu'à ce que la disponibilité de grandes quantités de données dans les corpus oraux ne rende nécessaire la prise en compte des variations.

L'instabilité des ressources, des théories et aussi des pratiques des transcripteurs tend à ce que la variation des données se retrouve, transposée, dans les annotations. Il en résulte des divergences qui se répercutent sur le traitement et entraînent des difficultés d'exploitation et d'interopérabilité, une compétition entre solutions. L'élaboration de critères d'évaluation est alors rendue nécessaire.

A un second niveau, les annotations sociolinguistiques peuvent porter sur les difficultés rencontrées. Il peut s'agir de mentions portées par les transcripteurs, plus souvent encore de discordances entre annotateurs. A leur tour, ces différentes compréhensions sont exploitables comme des indicateurs concernant des compétences d'auditeurs.

Exemple 9 : Variation d'écoute (locuteur ESLO1 / 109)

Transcription A
euh en euh j'ai **dû la regagner oui** cette année c'est impressionnant les progrès quand même
Transcription B
euh en euh j'ai **vu l'an dernier puis** cette année c'est impressionnant les progrès quand même
Transcription C
euh en euh j'ai vu l'an dernier **j'ai vu** cette année c'est impressionnant les progrès **qu'on a faits**

Si les travaux fondateurs du TAL voyaient dans la variation une difficulté plutôt qu'un surplus d'information, certains domaines de la linguistique, dont les données se fondent sur la comparaison, étaient en attente de solutions sur ce point. C'est le cas des études sur l'acquisition du langage (cf. CHAT, CHILDES), sur les changements liés à l'âge [Gerstenberg/Voeste], sur les corpus d'apprenants ou de la linguistique clinique. D'autres problèmes concernent les corpus multimodaux, pour l'annotation de corpus vidéo indispensables pour l'étude des langues des signes par exemple.

9. Conclusion

L'annotation résulte de la transformation des pratiques du linguiste lorsqu'il s'est trouvé confronté, par le développement de l'informatique, à une augmentation exponentielle des ressources à sa disposition. Pour en assurer l'exploitation, il lui a fallu maîtriser des outils qui, sans être spécifiques (la plupart sont en partage dans toutes les humanités numériques), demandaient à être adaptés, aussi bien pour l'acquisition des données (en particulier l'oral), que pour le traitement ou l'archivage.

Il s'est formé un consensus sur le périmètre du domaine, les méthodes et certaines normes. Une démarche par étapes (*process*), les modalités d'insertion dans un fichier d'éléments d'analyse et le partage des tâches entre la transcription, l'annotation et les métadonnées se retrouvent dans tous les corpus. Les propositions divergentes des écoles théoriques, la nature des données et la taille des unités (du phonème au discours) impliquent certains partages. L'orientation du travail (selon que la démarche privilégie une approche à partir de l'informatique ou de la linguistique) et les différences d'approche entre sous-disciplines limitent la polyvalence des corpus.

A ces partitions s'ajoutent les conditions d'émergence en fonction de la structure du champ académique et la spécificité des conditions du marché linguistique. La distance entre les pratiques orales et l'écrit, entre le standard et les dialectes, mais aussi les formes de domination (diglossie) ou la concurrence avec les variétés ultramarines ont des répercussions sur les conceptions et les usages de l'annotation. Ces conséquences sont plus lisibles dans le cas des études sur l'oral qui est, par nature, moins homogène que l'écrit.

10. Bibliographie

Gabriel BERGOUNIOUX, Michel JACOBSON, Paola PIETRANDREA