Corpus ECOLE_MASSY

Présentation générale

Jean-Yves Antoine

LI – Université François Rabelais de Tours

http://www.info.blois.univ-tours.fr/~antoine/parole_publique/

Rapport technique : VALORIA-CORAIL-2002-01, VALORIA, Université de Bretagne Sud

1 Introduction

Ce document présente en détail le corpus ECOLE_MASSY, un corpus pilote de dialogue orale homme-machine réalisé par le laboratoire VALORIA et diffusé librement (sous réserve de respect d'une convention d'utilisation) sur Internet dans le cadre du projet PAROLE_PUBLIQUE[1].

Plus précisément, ce rapport présente :

- le contenu du corpus distribué ainsi que les conditions dans lesquelles il a été recueilli,
- les modes de distributions du corpus,
- la convention à laquelle elle liée l'utilisation de ce corpus à toutes fins scientifiques ou industrielles,
- les références bibliographiques associées à ce corpus.
- les conventions de transcription et d'encodage suivies lors de la réalisation du corpus,

2 Présentation du corpus : contenu et conditions d'enregistrement

Le corpus ECOLE_MASSY est un corpus pilote de dialogue oral homme-homme finalisé. Relevant du cadre applicatif du renseignement touristique sur une tâche précise (planification d'activités de loisirs), il s'agit d'un corpus particulier recueilli auprès d'un public de jeunes enfants d'une classe de CE1 interagissant avec un compère (leur enseignante) jouant le rôle de hôtesse d'accueil d'un office de tourisme.

Le corpus distribué comprend les fichiers audio enregistrés ainsi qu'une transcription orthographique des dialogues ainsi recueillis.

2.1 Fiche signalétique

Corpus	ECOLE_MASSY
Version	1.0 (3 mai 2002)
Type de dialogue	Dialogue oral Homme-Homme finalisé (tâche de renseignement touristique)
Locuteurs	Enfants de 7 ans + enseignant adulte
Enregistrement	Conditions réelles – micro visible
Contenu	Corpus audio + transcription orthographique
Concepteur(s)	Sabine Letellier-Zarshenas (VALORIA), Igor Schadle (LI), Jean-Yves Antoine (LI)
Recueil	Sabine Lettelier-Zarshenas (VALORIA)
Transcripteur(s)	Sabine Letellier-Zarshenas (VALORIA)
Diffusion	libre sous réserve du respect d'une convention d'utilisation

2.2 Enregistrement : tâche et conditions d'enregistrement

Le corpus ECOLE_MASSY a été enregistré en conditions réelles dans une classe de CE1 d'une école primaire de Massy. Les consignes fournies aux enfants concernaient uniquement l'objectif de la transaction : recherche d'une séance de cinéma, puis planification libre de loisirs sur la région parisienne dans un second temps. A l'opposé, l'enseignant, qui jouait le rôle de l'agent, avait pour consigne de simuler un dialogue relativement directif. Afin de garantir une certaine naturalité, les transactions se sont faites sur les possibilités réelles de loisirs offertes au moment de l'enregistrement. A la demande de l'enseignant, les enregistrements ont été réalisés en l'absence d'opérateur de notre laboratoire. Aussi :

- la prise de son a été effectuée sur un magnétophone sans enregistrement sur pistes séparées. On aura donc un fichier audio par dialogue,
- les productions des enfants, qui faisaient face à leur enseignant, se sont traduites par une certaine perte de spontanéité. Elles reflètent une adaptation langagière qui peut être rapprochée d'une forme de dialogue homme-machine.

[1] Le projet PAROLE PUBLIQUE est désormais développé au sein de l'Université de Tours : http://www.sir.blois.univ-tours.fr/~antoine/parole_publique

La couverture sémantique du domaine par les enfants est par contre restée très libre, sans jamais sortir du périmètre défini par les consignes. Elle est donc représentative de la tâche et pourrait être comparée à celle d'usagers adultes. Le corpus comprend 45 minutes d'enregistrement et regroupe 31 dialogues.

2.3 Transcription : corpus distribué

L'intégralité du corpus enregistré à été transcrite. Le corpus ECOLE_MASSY regroupe ainsi 31 dialogues (tableau 1) correspondant à environ 5 300 mots transcrits.

Durée	Tâche : cinéma	Tâche : planification loisirs
< 30 s	2	0
30 s - 1mn	6	0
1 mn - 2mn	6	10
2 mn - 3 mn	0	7
> 3 mn	0	0

Tableau 2 — Distribution par type de tâche et par durée des dialogues du corpus ECOLE_MASSY

Chaque dialogue donne lieu à un fichier audio au format wav et un fichier de transcription orthographique. Les conventions de transcription et de codage suivies reprennent les normes les plus utilisées au sein de la communauté, à savoir :

- conventions de transcription du français parlé utilisées par le laboratoire DELIC (Blanche-Benveniste et Jeanjean 1987) et légèrement enrichies par certaines recommandations issues du projet SPEECHDAT (Gibbon, Moore et Winski 1997). Ces conventions sont détaillées en annexe de ce document,
- codage au format structuré XML avec utilisation de l'alphabet Unicode codé sur 8 bit.

La transcription a été réalisée à l'aide du logiciel libre Transcriber (Barras *et al*. 1998) dont nous reprenons la DTD XML en format de sortie.

Au final, les transcriptions sont distribuées suivant trois formats de sortie correspondant à des usages potentiels différents :

- codage XML (figure 1),
- codage en format texte (ASCII) reprenant une structuration en tours de parole (figure 2). Les chevauchements éventuels restent représentés dans ce format. L'information d'alignement temporel des tours de parole n'est par contre par reprise ici.
- format PDF regroupant dans un seul fichier l'ensemble des transcriptions obtenues en format texte.

```
<?xml version="1.0" encoding="UTF-8"?><!DOCTYPE Trans SYSTEM "trans-13.dtd">
<Trans scribe="Letellier" audio_filename="c1" version="10" version_date="020502">
<Topics> <Topic id="to1" desc="cinéma 1"/><Topic id="to2" desc="c1"/> </Topics>
<Speakers>
<Speaker id="spk1" name="institutrice" check="no" dialect="native" accent="" scope="local"/>
<Speaker id="spk3" name="élève" check="no" dialect="native" accent="" scope="local"/>
</Speakers>
<Episode>
<Section type="report" startTime="0" endTime="11.500" topic="to2">
<Turn startTime="0" endTime="3.055" speaker="spk1">
<Sync time="0"/>
qu'as tu choisi comme activité
</Turn>
<Turn speaker="spk3" startTime="3.055" endTime="7.374">
<Sync time="3.055"/>
# un film
</Turn>
<Turn speaker="spk1" startTime="7.374" endTime="9.587">
<Sync time="7.374"/>
quel film veux tu voir
</Turn>
<Turn speaker="spk3" startTime="9.587" endTime="11.500">
<Sync time="9.587"/>
```

[tx] les Razmoket
</Turn> </Section> </Episode>

Figure 1 : Extrait du corpus ECOLE_MASSY : transcription sans annotation (format XML)

fichier audio : c1
<01> institutrice
i: qu'as tu choisi comme activité
<02> élève
é: # un film
<03> institutrice
i: quel film veux tu voir
<04> élève
é: [tx] les Razmoket

Figure 2 : Extrait du corpus ECOLE_MASSY : transcription sans annotation (format texte)

2.4 Organisation du corpus distribué

La figure 3 décrit l'arborescence des fichiers du corpus distribué. A un premier niveau, on trouve le fichier de présentation du corpus ainsi que 3 répertoires regroupant les transcriptions aux formats XML (répertoire `Trans_XML`), ASCII (répertoire `Trans_TXT`) et PDF (répertoire `Trans_PDF`). Dans le cas d'une distribution avec fichiers sonores (cf. § 3 ci-dessous), un quatrième répertoire `Audio` regroupe les fichiers sons correspondant aux dialogues.

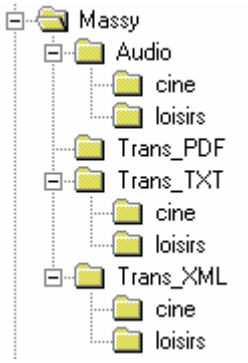

Figure 3 : Organisation des répertoires du corpus ECOLE_MASSY

A l'exception du répertoire `Trans_PDF` qui contient un seul fichier regroupant l'ensemble du corpus au format Acrobat PDF, ces répertoires sont eux-mêmes divisés en deux répertoires correspondant aux deux tâches de recherche de séance de cinéma (répertoire `cine`) et de planification de loisirs (répertoire `loisirs`). Dans ces répertoires terminaux se trouvent les fichiers audio ou de transcription, à raison d'un fichier par dialogue. Dans le cas des transcriptions XML, on trouvera également le fichier `trans-13.dtd` correspondant à la DTD Transcriber utilisée.

3 Distribution du corpus

Le corpus ECOLE_MASSY est diffusé suivant deux modes :

- **corpus transcrit seul** — Téléchargement à partir de la page WWW du projet PAROLE PUBLIQUE : http://www.sir.blois.univ-tours.fr/~antoine/parole_publique

- **corpus transcrit + corpus audio** — Compte tenu de la taille des fichiers audio, le corpus (fichiers son + transcription au divers formats) est distribué sur CD adressé par courrier postal.

Dans les deux cas, il vous est demandé de respecter une convention d'enregistrement peu contraignante détaillé dans le paragraphe suivant.

La distribution de ces corpus est **libre** quel que soit l'usage de ce corpus. Cependant, dans le cas d'une distribution par CD, il vous est demandé une participation de **10 Euros** correspondant aux frais de constitution et d'envoi du CD.

Pour plus de renseignements, contactez : Jean-Yves.Antoine@univ-tours.fr

4 Convention d'utilisation du corpus

Afin de favoriser les recherches francophones en linguistique de corpus et en ingénierie des langues, ce corpus est diffusé librement. Afin de préserver les droits intellectuels des concepteurs du corpus, il vous est cependant demandé de respecter la convention d'utilisation suivante :

- signaler auprès de Jean-Yves.Antoine@univ-tours.fr toute utilisation de ce corpus, que ce soit à des fins scientifiques ou industrielles,
- mentionner toute utilisation de ce corpus (nom + laboratoire) dans toute publication scientifique ou tout produit (logiciel ou autre) commercial concernés.
- donner dans tout article faisant mention de ce corpus une référence bibliographique concernant ce dernier. Une sélection de ces références sont donnés à la fin de ce document.

5 Références bibliographiques

5.1 Publications concernant le corpus ECOLE_MASSY

J.-Y. Antoine, S. Letellier-Zarshenas, P. Nicolas, I. Schadle (2002). Corpus OTG et ECOLE_MASSY : vers la constitution d'un collection de corpus francophones de dialogue oral diffusés librement. Actes TALN'2002. Nancy, France. Juin 2002.

P. Nicolas, S. Letellier-Zarshenas, I. Schadle, J.-Y. Antoine, J. Caelen (2002). Towards a large corpus of spoken dialogue in French that will be freely available: the "*Parole Publique*" project and its first realisations. Actes *LREC'2002*. Las Palmas de Gran Canaria, Espagne. Mai 2002.

5.2 Publications citées dans ce document

C. Barras *et al*. (1998). Transcriber : a free tool for segmenting, labeling and transcribing speech, Actes *LREC'1998*, Grenade, Espagne, pp. 1373-1376.

C. Blanche-Benveniste, C. Jeanjean (1987), Le français parlé, Paris, Didier Erudition.

D. Gibbon, R. Moore, R. Winski (Eds.) (1997) Handbook of standards and ressources for spoken language systems, Berlin, Mouton de Gruyter, pp. 825-834.

6 ANNEXE A — Conventions de transcription du corpus ECOLE_MASSY

La transcription est strictement orthographique, avec mention minimale des événements acoustiques connexes (voir ci-après). D'une manière générale, les conventions de transcription s'inspirent des fortement des recommandations utilisées dans le projet SPEECHDAT (Gibbon *et al*., 1997), ainsi que des conventions définies par la laboratoire DELIC pour le français.

6.1 Structuration de la transcription : tours de parole

Chaque dialogue est segmenté en tours de parole. La définition du tour de parole varie dans la littérature d'un auteur à l'autre. Dans le cadre de ce corpus, nous avons utilisé la définition opérative suivante : un nouveau de parole apparaît lorsqu'un nouveau locuteur se met à parler. Deux situations peuvent alors survenir :

Tour de parole sans chevauchement — Le tour de parole est délimité par (début) la prise de parole d'un locuteur et (fin) par la fin de sa production. Ce tour de parole ne concerne donc qu'un seul locuteur. Exemple de tour de parole sans chevauchement transcrit au format ASCII :

```
<03> institutrice
   i: quel film veux tu voir
```

Tour de parole avec chevauchement — Le tour de parole est délimité par le début et la fin du chevauchement. Ce tour de parole regroupe alors deux (voire plus) locuteurs. Leurs productions orales sont représentées simultanément dans ce tour de parole, en distinguant chaque locuteur. Exemple de tour de parole avec chevauchement transcrit au format ASCII :

```
<04> client + hôtesse
   c: d'accord
   h : on a simplement
```

Dans les dialogues, les périodes sans chevauchement succèdent bien entendu sans arrêt à des périodes avec chevauchement. Dans le cas du corpus ECOLE_MASSY, les chevauchements sont cependant très rares.

A titre d'exemple, supposons qu'un locuteur prononce un certains énoncé (par exemple « Tiens j'ai vu Paul hier ») tandis que le second locuteur se contente d'une marque d'étonnement (« ah ouais ») en milieu d'énoncé. Cette « tranche » de dialogue sera alors segmentée en 3 tours de parole :

- début d'énoncé sans chevauchement du locuteur 1,
- partie chevauchée avec prononciations des locuteurs 1 et 2,
- fin d'énoncé sans chevauchement du locuteur 2.

6.2 Conventions de transcription

La transcription est strictement orthographique, avec mention minimale des événements acoustiques connexes (voir ci-après). Elle suit les normes orthographiques standards du français. Notons cependant que tout mot sera séparé par un espace (blanc), le tiret entre deux mots n'étant conservé que si ceux-ci constituent un lemme insécable. Ainsi :

puis-je	sera transcrit	puis je	(2 mots)
plate-forme	sera transcrit	plate-forme	(1 mot)

La description des événements acoustiques ou prosodiques est limitée au minimum et est non exhaustive.

On se contente ainsi de marquer seulement les pauses longues, sans distinction de type. De même, la transcription ne comprendra aucune marque de ponctuation[2].

[2] Les linguistes travaillant sur l'oral, tels les chercheurs du GARS/DELIC, dénient généralement toute pertinence de la notion de ponctuation dans le langage parlé.

6.2.1 Bruits

Ce corpus a été enregistré en conditions réelles avec un médiocre rapport signal sur bruit. Les bruits non humains n'ont pas été transcrits. Nous avons par contre opéré réalisé une annotation minimale de certains bruits de l'appareil phonatoire :

rire	annoté	[rire]
bruits de bouche	annoté	[bb]
toux	annoté	[tx]
souffle	annoté	[pf]

6.2.2 Majuscules / minuscules

De manière générale, les transcriptions ne comportent que des caractères minuscules. L'emploi de majuscules est néanmoins pertinent pour marquer les noms propres de la langue ainsi que les caractères épelés. D'une manière plus précise :

- les énoncés transcrits ne débutent pas par une majuscule (on retrouve ici l'absence de ponctuations),
- Les acronymes et les caractères épelés (ou sigles) sont transcrits en majuscule. Ils ne sont pas séparés par des points :

S N C F	et non	*S.N.C.F.*

- les noms propres commencent par une majuscule (par exemple : `Jospin`, `Grenoble`). L'application de cette règle est stricte afin d'éviter d'englober autant que possible des noms communs. Ainsi, on transcrit :

monsieur Lionel Jospin	et non	*Monsieur Lionel Jospin*
mairie de Grenoble	et non	*Mairie de Grenoble*

  A l'opposé, les noms propres correspondant à des sigles sont mentionnés à l'aide de majuscules. L'existence d'un acronyme correspondant à ce sigle est un bon indice de "capitalisation". Par exemple :

Société Nationale des Chemins de Fer	(SNCF)
Transports de l'Agglomération Grenobloise	(TAG)

- les noms communs ayant fonction de nom propre (par exemple : titre de film) ne correspondant pas à un sigle sont transcrits entre guillemet et restent en minuscule. Lorsqu'on relève un nom propre dans ce type de nom commun, il prend bien entendu une majuscule. Par exemple :

  le bureau "info montagne"

  "l'amicale laïque de la ville de Massy"

  Remarque — Cette règle de transcription était optionnelle, la délimitation des situations sigle / nom commun ayant fonction de nom propre / nom commun étant relativement floue.

6.2.3 Nombres

A l'exception du nombre *un* qui peut être confondu avec l'article indéfini, les nombres ont été codés en chiffre lorsque leur prononciation suivait celle du français standard. Par exemple :

128	et non	*cent vingt huit*

Dans le cas contraire, les nombres ou séquences de nombres sont transcrites en caractères afin de refléter la prononciation exacte du locuteur. Par exemple :

septante deux	et non	*72*

6.2.4 Acronymes et sigles

La transcription des sigles, déjà évoquée, suit bien entendu la prononciation du locuteur :

- Intégralement s'il est prononcé mot à mot : `Société Nationale des Chemins de Fer`

- Sous forme de caractères épelés si son acronyme est prononcé lettre à lettre : `S N C F`
- Sous forme d'un nom propre particulier si son acronyme n'est pas épelé : `Tag` et non `T A G`

6.2.5 Prononciations incomplètes

Sont considérées ici les prononciations incomplètes de mots dues au caractère spontané de la parole : phénomènes de reprises ou répétitions, ou interruptions par l'autre locuteur. Elles seront marquées à l'aide des parenthèses placées en fin du fragment prononcé. Ce fragment sera transcrit sous forme orthographique en suivant les règles standard de prononciation. Lorsqu'il y a difficulté d'interprétation du fragment, la transcription complète du mot attendu est précisée entre les parenthèses. Par exemple :

donne moi une po() une poire ou encore

donne moi une po(pomme) une poire

6.2.6 Délétions, contractions

Le français parlé présente de nombreuses occurrences de contractions ou de délétions de syllabes qui concernent en particulier les locutions fréquentes ou les petits mots outils. Ces délétions ne peuvent être considérées comme des prononciations incomplètes, puisqu'elles relèvent de la stratégie d'élocution et non du caractère spontané de la production.

Certaines transcription rivalisent de conventions particulières destinées à rendre compte le plus précisément possible de la prononciation réalisée (par exemple : `y'a ka` pour *il n'y a qu'à*). Au contraire, on s'est limité ici — à l'instar des recommandations du DELIC (ex-GARS) — à une transcription aussi proche que possible de l'écriture standard. Par exemple :

je vais pour *j'vais* (en phonétique : /jve/)

il y a pour *y'a*

Dans le cas d'une délétion complète de mot (cas de la chute du discordantiel *ne*, par exemple), le mot ne sera pas transcrit.

6.2.7 Erreurs de prononciations, prononciations idiomatiques

Les formes correspondant à une erreur manifeste de prononciation (lapsus, par exemple), ou à une prononciation idiomatique, sont transcrites sous leur forme régulière, précédée d'un astérisque. La forme réellement prononcée est alors transcrite sous forme orthographique, en respectant les règles standard de prononciation du français, entre crochets après la forme corrigée. Exemple :

*je *répète{récapépète} depuis le *début{bédut}*

Si la forme inattendue ne peut se traduire fidèlement sous forme orthographique, on adopte la notation phonétique ajoutée en signes "/". On utilise pour cela la convention de notation SAMPA.

6.2.8 Evénements acoustiques : pauses

Deux types de pause ont été distinguées :

- pauses remplies (hésitations du type *euh*, *mmh* etc...) notées par le sigle e
- pauses silencieuses notées par le sigle #

7 ANNEXE A — Codage : formats de transcription en sortie

Trois formats de sortie ont été définis pour les fichiers de transcription

- codage XML,
- codage en format texte (ASCII),
- format PDF regroupant dans un seul fichier l'ensemble des transcriptions obtenues en format texte.

7.1 Codage XML

La transcription a été réalisée à l'aide du logiciel libre Transcriber. Le format XML de sortie suit donc la DTD définie par ce logiciel. Nous ne détaillerons pas ici cette DTD : le lecteur intéressé se référera à (Barras *et al*. 1998) ou consultera le site Internet consacré à Transcriber :

http://www.etca.fr/CTA/gip/Projets/Transcriber/IndexFr.html.

On notera simplement que ce format de sortie permet de décrire les chevauchements ainsi que l'alignement temporel des débuts et fin de tours de parole.

Précisons enfin que la version de Transcriber utilisée (version Windows) présentait un bug quant au codage du « à » en Unicode. Dans le corpus distribué, ce codage erroné a été corrigé.

7.2 Codage ASCII

Ce codage est la traduction simplifiée en ASCII de la transcription XML précédente. Dans ce format :

- ne sont conservés que les informations concernant le dialogue par lui-même (pas d'entête à l'exception de l'étiquette du dialogue concerné),
- ne sont pas conservées les informations d'alignement temporel
- est par contre conservée la segmentation en tours de parole. Chaque tout de parole se voit accorder un numéro spécifique par incrément. Pour un tour de parole donné, on précise ensuite à la ligne l'identité du locuteur ainsi que l'énoncé prononcé. Ce format permet toujours une représentation des chevauchements : dans ce cas, deux énoncés sont donnés dans un tour de parole particulier, avec toujours en tête d'énoncé la mention de l'identité du locuteur correspondant.

La figure 1 donne un exemple de sortie dans ce format.

```
                    fichier audio : c1
<01> institutrice
    i: qu'as tu choisi comme activité
<02> élève
    é: # un film
<03> institutrice
    i: quel film veux tu voir
<04> élève
    é: [tx] les Razmoket
```

Figure 1 : Extrait du corpus ECOLE_MASSY : transcription sans annotation (format ASCII)

7.3 Format PDF

Ce format de sortie est la simple compilation, sous la forme d'un fichier Acrobat PDF unique, des fichiers ASCII de transcription décrits ci-dessus.