Corpus Accueil_UBS

Présentation générale

Jean-Yves Antoine[1], Judith Muzerelle[2]

[1]LI – Université François Rabelais de Tours
[2]LLL – Université d'Orléans

Université François Rabelais Tours

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

Introduction

Ce document présente en détail le corpus Accueil UBS, un corpus pilote de dialogue oral homme-machine réalisé par le laboratoire VALORIA dans le cadre du projet AGILE-OURAL du programme TECHNOLANGUE du Ministère de la Recherche. Il a été ensuite révisé au sein du laboratoire LI dans le cadre du projet régional ANCOR (région Centre). Ce corpus est diffusé librement par le laboratoire LI de l'Université de Tours, (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,

1 Présentation du corpus : contenu et conditions d'enregistrement

Le corpus Accueil_UBS est un corpus pilote de dialogue oral homme-homme finalisé correspondant à une tâche d'accueil téléphonique par le standard d'une université. Il a été enregistré en conditions réelles au sein de l'Université de Bretagne Sud et regroupe un ensemble de dialogues entre un(e) appelant€ et le personnel d'accueil du standard.

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

1.1 Fiche signalétique

Corpus	Accueil_UBS
Version	1.1 (septembre 2013)
Type de dialogue	Dialogue oral Homme-Homme finalisé (tâche : accueil téléphonique)
Locuteurs	Adultes hommes ou femmes
Enregistrement	Conditions réelles – enregistrement semi-clandestin (appelants non informés avant l'appel).
Contenu	Corpus audio + transcription orthographique
Concepteur(s)	Jean-Yves Antoine (LI, Université de Tours)
Recueil	Julien Foulon (VALORIA, U. Bretagne Sud)
Transcripteur(s)	Julien Foulon (VALORIA, U. Bretagne Sud)
Révision	Judith Muzerelle (LLL, Université d'Orléans), Jean-Yves Antoine (LI, U. Tours)
Diffusion	libre sous réserve du respect d'une convention d'utilisation

1.2 Enregistrement : tâche et conditions d'enregistrement

Le corpus Accueil_UBS a été enregistré par le VALORIA en conditions réelles auprès de l'accueil téléphonique de l'Université de Bretagne Sud (UFR Sciences et UFR Droit), suivant une procédure semi-clandestine : seul le personnel de l'office était préalablement mis au courant de l'enregistrement. Le personnel d'accueil n'a été soumis à aucune consigne particulière. Les enregistrements ont été directement effectués par extraction logicielle sur la ligne téléphonique des standards. Les voix appelant et réceptionnistes n'étaient pas séparées à l'enregistrement. On dispose donc d'un fichier audio par dialogue, de qualité téléphonique.

Un expérimentateur assistait à la prise de son. En fin de dialogue, il s'assurait du respect des règles déontologiques en la matière. En particulier, une fois l'enregistrement effectué, il mettait au courant les clients de cette expérimentation. Il était alors demandé aux clients s'ils acceptaient que l'enregistrement les concernant soit conservés ou non.

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

Au total, 3 heures d'enregistrement ont été effectuées et peuvent être obtenues auprès du laboratoire VALORIA[2]. Le corpus distribué dans le cadre du PAROLE_PUBLIQUE correspond à la transcription d'un tiers du corpus (10 000 mots), reprenant les dialogues jugés les plus intéressants et les plus audibles.

1.3 Transcription orthographique

La transcription orthographique proposée dans cette distribution reprend 40 dialogues, qui correspondent à environ 1 heure d'enregistement pour une taille de 10 000 mots (tableau 1).

durée d'enregistrement	60 minutes environ
nombre de dialogues	40
nombre de locuteurs	2 réceptionnistes / 40 appelants
nombre de mots	10 060

Tableau 1 : Données synthétiques sur le corpus Accueil_UBS.

1.4 Corpus distribué

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.
- formats DOC (Microsoft Word), ODT (Open Office) et PDF regroupant dans trois fichiers la totalité des transcriptions obtenues en format texte.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE Trans SYSTEM "trans-13.dtd">
<Trans scribe="Foulon J" audio_filename="060_0000003d" version="1" version_date="041124">
<Speakers>
<Speaker id="spk1" name="hotesse" check="no" dialect="native" accent="" scope="local"/>
<Speaker id="spk2" name="client" check="no" dialect="native" accent="" scope="local"/>
</Speakers>
<Episode>
<Section type="report" startTime="0" endTime="11.952">
<Turn startTime="0" endTime="1.812" speaker="spk1">
<Sync time="0"/>
U B S bonjour
</Turn>
<Turn speaker="spk2" startTime="1.812" endTime="10.032">
<Sync time="1.812"/>
oui bonjour madame j'aurais voulu avoir des renseignements pour e l' l'inscription en A E
S administration Ã©conomique et sociale
</Turn>
<Turn speaker="spk1" startTime="10.032" endTime="11.952">
<Sync time="10.032"/>
oui [pi] oui conserver je vais vous passer la personne
```

[2] http://www-valoria.univ-ubs.fr/

```
</Turn>
</Section>
</Episode>
</Trans>
```

Figure 1 : Extrait du corpus Accueil_UBS : transcription orthographique (format XML)

```
<01> hotesse
     h: U B S bonjour
<02> client
     c: oui bonjour madame j'aurais voulu avoir des renseignements pour e l'
l'inscription en A E S administration économique et sociale
<03> hotesse
     h: oui [pi] oui conserver je vais vous passer la personne
```

Figure 2 : Extrait du corpus Accueil_UBS : transcription orthographique (format ASCII).

1.5 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 DOC/ODT/PDF (répertoire `Trans_DOC_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.

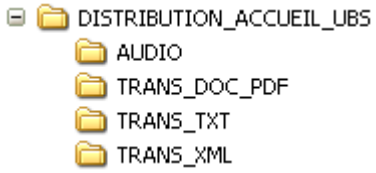

Figure 3 : Organisation des répertoires du corpus Accueil_UBS

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. Etant donné que les transcriptions suivent le format utilisé par le logiciel Transcriber, un renommage des fichiers `.xml` en `.trs` permet l'utilisation directe des transcriptions sous Transcriber.

2 Distribution du corpus et convention d'utilisation

Le corpus Acccueil_UBS est diffusé suivant deux modes :

- **corpus transcrit seul** — Téléchargement à partir de la page WWW du projet 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 le cas d'une distribution par CD, il vous est demandé une participation de **15 Euros** correspondant aux frais de constitution et d'envoi du CD.

Important - Par ailleurs, cette ressource intègre des échanges dont la communication porte atteinte à la protection de la vie privée ou portant appréciation ou jugement de valeur sur une personne physique nommément désignée, ou facilement identifiable, ou qui font apparaître le comportement d'une personne dans des conditions susceptibles de lui porter préjudice. (Code du Patrimoine, art. L. 213-2, I, 3) . A ce titre, ce corpus peut être utilisé à des fins d'analyse, mais en aucun cas ne peut être diffusés publiquement.

La distribution de ces corpus est **libre** quel que soit l'usage de ce corpus.

Par ailleurs, nous vous serions extrêmement reconnaissants de nous signaler toute utilisation du corpus à des fins de recherche ou industrielle, ainsi que de nous communiquer tout article reposant sur des données extraites du corpus. Ceci afin de nous permettre d'identifier les usages faits avec la ressource, pour son amélioration éventuelle à l'avenir.

3 Références bibliographiques

Liste des publications à la date de l'émission de ce rapport technique. Consultez le site Internet du projet Parole Publique pour une bibliographie à jour.

3.1 Publications concernant le projet PAROLE PUBLIQUE

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. pp. 319-324.

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. pp. 649-655.

3.2 Publications sur l'utilisation du corpus Accueil_UBS (brève introduction sur la ressource)

Antoine J.-Y., Goulian J., Villaneau J., Le Tallec M. (2009) Word Order Phenomena in Spoken French : a Study on Four Corpora of Task-Oriented Dialogue and its Consequences on Language Processing. *Corpus Linguistics'2009*, Liverpool, UK, July 2009

3.3 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.

4 Financement

La réalisation de ce corpus a été financée dans le cadre de deux projets distincts :

- projet AGILE-OURAL du programme TECHNOLANGUE du Ministère de la Recherche.
- projet ANCOR de la région Centre.

5 ANNEXE A — Conventions de transcription du corpus Accueil_UBS

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.

5.1 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.

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.

5.2 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[3].

[3] Les linguistes travaillant sur l'oral dénient généralement toute pertinence de la notion de ponctuation dans le langage parlé.

5.2.1 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]

5.2.2 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.

5.2.3 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*

5.2.4 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`

5.2.5 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

5.2.6 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.

5.2.7 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.

5.2.8 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 #

6 ANNEXE B — 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.

6.1 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://trans.sourceforge.net/

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

6.2 Codage ASCII

Ce codage est la traduction simplifiée en ASCII de la transcription XML. Dans ce format (figure 1):

- 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.

```
<01> hotesse
     h: U B S bonjour
<02> client
     c: oui bonjour madame j'aurais voulu avoir des renseignements pour e l'
l'inscription en A E S administration économique et sociale
<03> hotesse
     h: oui [pi] oui conserver je vais vous passer la personne
```

Figure 1 : Extrait du corpus Accueil_UBS : transcription orthographique (format ASCII).

6.3 Formats, DOC, ODT PDF

Ces formats sont la compilation, sous la forme d'un fichier unique, des fichiers ASCII décrits ci-dessus.

Dialogue 060

<01> hotesse
h: U B S bonjour
<02> client
c: oui bonjour madame j'aurais voulu avoir des renseignements pour e l' l'inscription en A E S administration économique et sociale
<03> hotesse
h: oui [pi] oui conserver je vais vous passer la personne

Figure 2 : Extrait du corpus Accueil_UBS : transcription orthographique (format DOC/ODT/PDF).

Le numéro de dialogue précisé dans cette compilation reprend les trois premiers chiffres des fichiers de transcription et audio correspondant. Par exemple, sur la figure 2, la numérotation *Dialogue 060* correspond aux fichier de transcription 060_0000003d.xml.