Corpus BRASSENS

Présentation générale

Geoffrey Williams, Christophe Ropers
LiCoRN – Université de Bretagne Sud

http://web.univ-ubs.fr/corpus

Financé par ANR

Ce document présente en détail le corpus BRASSENS, un corpus de dialogue entre des enfants et leur institutrice au cours de séances de travail autour du conte. Ce corpus a été réalisé dans la cadre d'un projet ANR (cf. supra) par le laboratoire LiCoRN de l'Université de Bretagne Sud. Il est 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,
- la convention à laquelle elle liée l'utilisation de ce corpus à toutes fins scientifiques ou industrielles,
- les modes de distributions du corpus,

1 Postulat

Nous partons du principe (résultant d'observations préalables) que, dans le genre d'interaction qui a lieu entre enfants, les stimuli de l'interaction ne seront pas tant des questions et des réponses que des points d'entrée dans une histoire partagée par plusieurs interlocuteurs.

2 Constitution du corpus

Partant de ce postulat, le corpus est constitué de parole d'enfant recueillie pendant une activité de création d'histoire en milieu scolaire. L'activité concerne la création d'un conte qui sert de prétexte à la génération de l'interaction.

2.1 Fiche signalétique

Corpus	BRASSENS
Version	1.0 (15 janvier 2009)
Type de dialogue	Dialogue enfant / adulte finalisé (tâche autour du conte)
Locuteurs	Enfants de ? ans + enseignant adulte
Enregistrement	Conditions réelles – micro d'ambiance visible
Contenu	Corpus audio + transcription orthographique
Concepteur(s)	Christophe Ropers, Geoffrey Williams (LiCorRN)
Recueil	Christophe Ropers, Patrice Marquand (LiCorRN)
Transcripteur(s)	Patrice Marquand (LiCorRN)
Diffusion	libre sous réserve du respect d'une convention d'utilisation

2.2 Modélités d'enregistrement

- 2 salles de classe,
- 1 micro d'ambiance par salle
- distance du micro: entre 1m et 3m du locuteur

2.3 Transcription : corpus distribué

L'intégralité du corpus enregistré à été transcrite. Le corpus BRASSENS regroupe ainsi 138 dialogues correspondant à environ 4h10' d'enregistrements audio.

Durée (approximatif)	Nombre de dialogues
< 30 s	3
30 s - 1mn	11
1 mn - 1mn30	27
1 mn 30 - 2 mn	42
2 mn - 2 mn 30	39
2 mn 30 - 3 mn 00	12
> 3 mn	4

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

2.4 Protocole de transcription

Les fichiers de transcription ont été produits par le biais de Transcriber (Barras *et al.* 1998) depuis des extraits des enregistrements bruts. Ces extraits ne sont pas des énoncés échantillonnés; ils n'ont pas été sélectionnés mais ont été produits simplement après élimination des périodes de silence ou de bruit sans parole. L'unité de segmentation retenue est celle du tour de parole. La transcription est orthographique sans trucage orthographique visant à faire ressembler le texte transcrit à la production orale.

Suite à la transcription, les fichiers XML obtenus par Transcriber ont été transformés à l'aide d'un script XSLT afin de les conformer aux recommandations du consortium TEI (www.tei-c.org). Lors de cette transformation, des codes de catégorisation ont été ajoutés dans la balise <catRef> de l'en-tête afin de référencer différemment la parole d'enfant de la parole d'adulte en vue d'une extraction de la parole d'enfant seule (nous ne nous intéressons pas ici à l'interaction avec l'adulte/enseignant).

2.5 Organisation du corpus distribué

La figure 1 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.

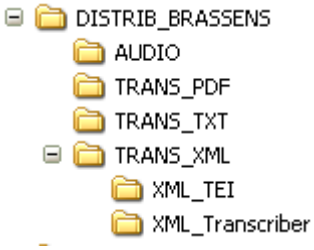

Figure 1 : Organisation des répertoires du corpus ECOLE_MASSY

A l'exception du répertoire `Trans_PDF` qui contient un seul fichier regroupant l'ensemble du corpus au format Acrobat PDF, ces répertoires contiennent autant de fichier qu'il existe de dialogues. Le répertoire `TRANS_MXL` englobe deux sous répertoire :

- `XML_Transcriber` qui correspond au format XML généré par le logiciel *Transcriber*. On y trouvera également le fichier `trans-13.dtd` correspondant à la DTD Transcriber utilisée.
- `XML_TEI` qui correspond au format XML respectant les recommandations de TEI, afin de rendre les transcriptions indexables, entre autres, par le logiciel *Xaira*. Chaque transcription est ainsi séparée entre un fichier regorupant les productions de l'institutrice, et un autre celles des enfants.

3 Utilisation du corpus BRASSENS dans le cadre du projet ANR EmotiRob

L'exploitation du corpus visera à caractériser le type de discours susceptible d'être produit par les enfants qui entreront en interaction avec le robot compagnon. A cette fin, le corpus devra fournir:

-des indications permettant de spécifier le vocabulaire récurrent et les probabilités de co-occurrence lexicale;

-des mesures stylométriques;

-les mesures de la longueur moyenne d'une intervention et de la fenêtre de référence interne potentielle;

-la structure de progression discursive et thématique.

4 Financement

La constitution de ce corpus a été réalisée dans le cadre du projet EmotiRob (projet PSIROB06_174281) financé par l'Agence Nationale pour la Recherche (Programme Systèmes Interactifs et Robotique - PsiRob06).

Toile : http://www-valoria.univ-ubs.fr/emotirob/

5 Distribution du corpus

Le corpus BRASSENS est diffusé suivant deux modes :

- **corpus transcrit seul** — Téléchargement à partir de la page WWW du projet PAROLE PUBLIQUE : http://www.info.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 **15 Euros** correspondant aux frais de constitution et d'envoi du CD.

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

6 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 AT 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.

7 Références bibliographiques

7.1 Publications concernant le corpus BRASSENS

???

7.2 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.