Corpus OTG et ECOLE_MASSY : vers la constitution d'une collection de corpus francophones de dialogue oral diffusés librement

Jean-Yves Antoine(1), Sabine Letellier-Zarshenas(1), Pascale Nicolas(1), Igor Schadle(1), Jean Caelen(2)

(1) VALORIA , Université de Bretagne Sud,
rue Yves Mainguy, F-56 000, France
Mel = Jean-Yves.Antoine@univ-ubs.fr
(2) CLIPS-IMAG , Université Joseph Fourier,
BP 53, F-38041 Grenoble Cedex 9, France

Résumé – Abstract

Cet article présente deux corpus francophones de dialogue oral (OTG et ECOLE_MASSY) mis librement à la disposition de la communauté scientifique. Ces deux corpus constituent la première livraison du projet *Parole Publique* initié par le laboratoire VALORIA. Ce projet vise la constitution d'une collection de corpus de dialogue oral enrichis par annotation morpho-syntaxique. Ces corpus de dialogue finalisé sont essentiellement destinés à une utilisation en communication homme-machine..

This paper presents two corpora (*OTG* et *ECOLE_MASSY*) of French spoken dialogue which are the first delivery of the *Parole_Publique* (in English : *Public Speech*) project held by the VALORIA laboratory. This project aims at the achievement of a collection of spoken dialogue corpora that is freely distributed on the WWW. It is primarily intended for researches on man-machine communication.

Mots Clés — Keywords

ressources linguistiques francophones ; dialogue oral ; communication homme-machine

French speaking linguistic ressources ; spoken dialogue ; man-machine communication.

1 Introduction

Le dialogue oral homme-machine (DOHM) a atteint au cours de la dernière décennie un degré de maturité qui s'est traduite par le développement d'applications commercialisées. Cette réussite doit beaucoup à la généralisation de méthodes empiriques basées sur les données (estimation de modèles stochastiques sur de grands corpus). Les ressources linguistiques ont ainsi acquis un rôle central en DOHM comme dans l'ingénierie des langues en général. La représentativité de ces corpus étant pour partie fonction de leur taille, on observe une tendance générale à la constitution de ressources linguistiques de plus en plus vastes.
Le développement de larges ressources linguistiques orales constitue donc un enjeu majeur du DOHM. Or, le retard du français, et plus particulièrement du français oral, est considérable (Véronis 2000). Là où le *British National Corpus (*Leech 1994) comprend plus de 10 millions de mots de parole transcrite, on ne peut citer au mieux qu'un corpus francophone d'un million de mots collecté par le laboratoire DELIC (ex-GARS). Ce corpus n'est que partiellement informatisé et ne concerne guère le dialogue oral. De même, le corpus réalisé dans le cadre du projet ELICOP (un million de mots) ne semble pas comprendre de dialogues oraux. La plupart des corpus francophones de dialogue oral ne sont d'ailleurs pas distribués.
Dans cet article, nous présentons deux corpus francophones de dialogue oral (OTG et ECOLE_MASSY) distribués librement. Ces corpus constituent la première livraison du projet *Parole Publique* du laboratoire VALORIA. Ce projet vise la constitution d'une collection de corpus de dialogue oral enrichis par annotation morpho-syntaxique destinés au DOHM.

2 Présentation générale des corpus

Les deux corpus ont été constitués suivant une méthodologie commune de transcription et de codage définie dans le cadre du projet *Parole Publique*. Dans cette section, nous décrivons les caractéristiques communes de ces corpus, pour revenir ultérieurement sur leurs spécificités.

2.1 Corpus pilotes

Les corpus OTG et ECOLE_MASSY sont des corpus pilotes de dialogue oral finalisé. Un corpus pilote permet le recueil d'interactions réelles (dialogue homme-homme) mettant en évidence les phénomènes linguistiques et dialogiques inhérents naturellement à la tâche considérée (Caelen *et al*. 1997). L'intérêt des corpus pilotes en DOHM réside dans cette analyse des usages et des besoins réels — voire « idéaux ». Dans la perspective d'une mise en place de bonnes pratiques ingénieriques, ces analyses devraient être utiles à la conception de systèmes de dialogue oral mais aussi à la préparation de leur évaluation (critères de test).

2.2 Contenu des corpus distribués

Chaque corpus regroupe un ensemble de dialogues. Tout dialogue donne lieu à l'enregistrement d'un ou plusieurs (enregistrements sur pistes séparées) fichiers audio au format `wav`. Chaque dialogue est décrit par un fichier de transcription et d'annotation à raison. La méthodologie de transcription et d'annotation que nous avons suivi reprend les normes les plus utilisées au sein de la communauté francophone, 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),

- système d'annotation morphosyntaxique défini par l'action GRACE (Adda *et al*. 1999),
- 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. L'annotation morpho-syntaxique de nos corpus n'a pas encore été réalisée. Elle sera effectuée à l'aide du logiciel Cordial Analyseur de la société Synapse avec post-validation par un expert. Les études de (Valli et Véronis 1999) ont en effet montré que ce système, utilisé en étiqueteur morpho-syntaxique, conservait une bonne robustesse sur de l'oral spontané. Au final, les transcriptions sont distribuées suivant deux formats de sortie correspondant à des usages potentiels différents :

- codage XML avec ou sans (figure 1) annotation morpho-syntaxique .
- codage en format texte (ASCII) reprenant une structuration en tours de parole avec ou sans (figure 2) annotation morpho-syntaxique. On remarquera sur la figure 2 que les chevauchements restent représentés dans ce format. L'information d'alignement temporel des tours de parole n'est par contre par reprise ici.

Ces corpus peuvent être librement récupérés sur la Toile après signature d'une convention d'utilisation peu contraignante : http://www.univ-ubs.fr/valoria/antoine/parole_publique.

```
<?xml version="1.0" encoding="UTF-8"?> <!DOCTYPE Trans SYSTEM "trans-13.dtd">
<Trans scribe="Nicolas" audio_filename="1ag0365" version="1" version_date="011008">
<Speakers>
<Speaker id="spk1" name="hôtesse" check="no" type="female" dialect="native" accent="" scope="local"/>
<Speaker id="spk2" name="client" check="no" type="female" dialect="native" accent="" scope="local"/>
</Speakers>
<Topics><Topic id="to1" desc="1ag0365"/></Topics>
<Episode>
<Section type="report" startTime="0" endTime="5.980" topic="to1">
<Turn startTime="0" endTime="0.629" speaker="spk1">
<Sync time="0"/>
bonjour madame
</Turn>
<Turn speaker="spk2" startTime="0.629" endTime="3.420">
<Sync time="0.629"/>
bonjour est ce que vous avez le programme de oui e e je
</Turn>
<Turn speaker="spk1 spk2" startTime="3.420" endTime="3.856">
<Sync time="3.420"/>
<Who nb="1"/>
oui
<Who nb="2"/>
connaissances
</Turn>
<Turn speaker="spk2" startTime="3.856" endTime="4.24">
<Sync time="3.856"/>
du monde
</Turn>
```

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

```
                    fichier audio : 1ag0365
<001> hôtesse
```

```
    h: bonjour madame
<002> client
    c: bonjour est ce que vous avez le programme de oui e e je
<003> hôtesse+client
    h: oui
    c: connaissances
<004> client
    c: du monde
```

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

3 Corpus OTG (*Office du Tourisme de Grenoble*)

Le corpus OTG (*Office du Tourisme de Grenoble*) a été constitué dans le cadre de l'ARC « Dialogue Oral » de l'AUF. Il a été enregistré par le laboratoire CLIPS-IMAG et transcrit par le VALORIA. Le cadre d'application étudié était le renseignement touristique.

3.1 Enregistrement

Le corpus OTG a été enregistré à la Maison du Tourisme de Grenoble. Les clients et l'agent n'ont été soumis à aucune consigne. La prise de son s'est effectuée en conditions réelles par deux microphones directifs orientés l'un vers le client (masqué[1]) et l'autre vers l'agent. Les enregistrements ont été recueillis sur deux pistes séparées par un enregistreur DAT. On dispose donc de deux fichiers audio par dialogue. Au total, une sélection de 5 heures d'enregistrements a été conservée pour la constitution du corpus audio.

3.2 Transcription : corpus distribué

Enregistré en conditions réelles, ce corpus présente un nombre important de transactions de médiocre qualité sonore. La transcription de ces dialogues s'est avérée difficile voire impossible, les transcripteurs ne parvenant pas à s'accorder sur de nombreux passages. Les conventions de transcription du DELIC permettent la représentation de transcriptions alternatives. Compte tenu du nombre important de passages conflictuels dans certains dialogues, nous n'avons pas utilisé cette possibilité. La transcription n'a ainsi été réalisée que sur des dialogues ne présentant aucune ambiguïté d'écoute. Certaines transactions retenues correspondaient à des trilogues. Il s'est alors avéré difficile de faire une distinction sûre entre les productions des deux clients concernés. Ces dialogues n'ont donc pas été transcrits.
Au total, 315 dialogues ont été transcrits, qui correspondent à 2 heures d'enregistrement. Ce corpus a une taille globale à 26 000 mots transcrits (tableau 1). A terme, le corpus atteindra une taille critique de 40 000 mots.

Durée	< 30s	30s – 1 mn	1 mn – 2 mn	2 mn-3 mn	> 3 mn
Nbre de dialogues	294	77	36	2	0

Tableau 1 : Distribution des dialogues du corpus OTG suivant leur durée.

[1] Ce n'est qu'à la fin du dialogue que le client était informé de l'expérience.

4 Corpus ECOLE_MASSY

Le corpus ECOLE_MASSY a été enregistré et transcrit par le laboratoire VALORIA. Le cadre d'application relevait également du renseignement touristique sur une tâche plus précise : la planification d'activités de loisirs. Ce corpus répond à une motivation scientifique spécifique : l'étude différentielle des usages linguistique suivant le type d'utilisateur. La population étudiée ici était de jeunes enfants de sept ans. De part ses motivations, ce corpus n'est pas directement utilisable pour la conception d'un système classique de dialogue adulte-machine. Il intéressera surtout linguistes et chercheurs en sciences de l'éducation. L'adaptation des systèmes de dialogue oral à des publics spécifiques (personnes âgées, handicapés, enfants…) représentera cependant une problématique importante dans les années à venir. Elle nécessitera alors le recours à des observations sur des corpus tels que celui-ci.

4.1 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 (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 (Spérandio et Létang-Fogeac 1986).

La couverture sémantique du domaine par les enfants est 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.

4.2 Transcription : corpus distribué

Contrairement au corpus OTG, l'intégralité du corpus ECOLE_MASSY a été transcrite. Elle regroupe 31 dialogues (tableau 2) correspondant à environ 5 300 mots transcrits.

Durée	< 30s	30s – 1 mn	1 mn – 2 mn	2 mn-3 mn	> 3 mn
Nbre de dialogues	2	6	16	7	0

Tableau 2 : Distribution des dialogues du corpus ECOLE_MASSY suivant leur durée.

Une première analyse du corpus montre un taux très faible de chevauchements. Nous sommes donc en présence d'une parole plus contrôlée que spontanée, qu'il serait intéressant de comparer aux genres oraux tels que ceux définis par (Biber 1988).

5 Conclusion : le projet PAROLE PUBLIQUE

A eux seuls, ces deux corpus représentent une ressource linguistique relativement limitée. Ce travail ne prend tout son sens que dans le cadre du projet *Parole Publique* de collection d'un ensemble de corpus oraux restreints au seul champ du DOHM. Ce projet concernera :

- différents types de dialogue (corpus pilotes, magiciens d'Oz, dialogue H-M réel) pouvant être utilisés à différents stades de développement des systèmes de dialogue oral.

- différents domaines d'application (renseignement touristique ou administratif, réservation hôtelière, portail vocal) afin de répondre à l'importante question de la généricité des recherches en dialogue oral homme-machine.

- différents types d'usagers : personnes âgées, enfants ou adolescents, handicapés…

La constitution de ce corpus dépendra des ressources budgétaires de notre laboratoire. Cet effort demande donc à être poursuivi et relayé dans d'autres laboratoires francophones. Aussi, un des objectifs de ce projet est-il de favoriser, par cette première diffusion libre de corpus, la constitution de larges ressources linguistiques francophones accessibles librement. La mise en place du projet ASILA (www.loria.fr/projets/asila/) — auquel contribue notre laboratoire par l'intermédiaire de la distribution de ces deux corpus — constitue une initiative prometteuse de ce point de vue.

Remerciements

Le travail présenté dans cet article a été partiellement financé par l'AUF dans le cadre de l'ARC « Dialogue Oral », ainsi que par une bourse doctorale de la Région Bretagne.

Références

Adda G. *et al*. (1999), L'action GRACE d'évaluation de l'assignation des parties du discours pour le français, *Langues* 2(1), pp. 119-129.
Barras C. *et al*. (1998), Transcriber : a free tool for segmenting, labeling and transcribing speech, Actes de *LREC'98*, Grenade, Espagne, pp. 1373-1376.
Biber D. (1988), Variation across speech and writing, Cambridge, Cambridge Univ. Press.
C. Blanche-Benveniste, C. Jeanjean (1987), Le français parlé, Paris, Didier Erudition.
Caelen J. *et al.* (1997), Les corpus pour l'évaluation du dialogue homme-machine. Actes de *JST'97 FRANCIL*, Avignon, France, pp. 215-222.
Gibbon D., Moore R., Winski R. (Eds.) (1997) Handbook of standards and ressources for spoken language systems, Berlin, Mouton de Gruyter, pp. 825-834.
Leech G. (1994), 100 million words of English : the British National Corpus, *English Today*, 9(1). pp. 9-15.
Spérandio J.-C., Létang-Figeac C. (1986), Simulation expérimentale de dialogues oraux en communication homme-machine, rapport technique, GRECO Communication Parlée, CNRS.
Valli A., Véronis J. (1999), Etiquetage grammatical des corpus de parole, *Revue Française de de Linguistique Appliquée*, *RFLA*, 4(2), pp. 113 :134.
Véronis J. (2000), Annotation automatique de corpus, In Pierrel J.-M. (2000*), Ingénierie des langues*, Paris, Hermès Sciences., pp. 111-130.