

# Corpus OTG

## Présentation générale

**Jean-Yves Antoine**  
**LI – Université François Rabelais de Tours**

Rapport technique : VALORIA-CORAIL-2002-02 — Université de Bretagne Sud



[http://www.info.univ-tours.fr/~antoine/parole\\_publicue/](http://www.info.univ-tours.fr/~antoine/parole_publicue/)



## Introduction

Ce document présente en détail le corpus OTG (Office du Tourisme de Grenoble), un corpus pilote de dialogue oral homme-machine réalisé par les laboratoires VALORIA et CLIPS-IMAG (désormais équipe GETALP du LIG) avec le soutien de l'AUF<sup>1</sup> dans le cadre de l' Action de Recherche Concertée « Dialogue Oral » (ARC-ILOR B2) ainsi que dans le cadre du projet AGILE-OURAL du programme TECHNOLOGUE du Ministère de la Recherche. 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<sup>2</sup>.

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 OTG est un corpus pilote de dialogue oral homme-homme finalisé relevant du cadre applicatif du renseignement touristique. Il a été enregistré en conditions réelles au sein de l'Office du Tourisme de Grenoble et regroupe un ensemble de dialogue entre un (ou plusieurs) touriste(s) et le personnel d'accueil de l'office.

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

### 1.1 Fiche signalétique

|                         |   |
|-------------------------|---|
| <b>Corpus</b>           | OTG   |
| <b>Version</b>          | 1.0 (3 juin 2002)   |
| <b>Type de dialogue</b> | Dialogue oral Homme-Homme finalisé (tâche de renseignement touristique)     |
| <b>Locuteurs</b>        | Adultes (touristes francophones + réceptionnistes) hommes ou femmes         |
| <b>Enregistrement</b>   | Conditions réelles – enregistrement semi-clandestin (micro touriste caché). |
| <b>Contenu</b>          | Corpus audio + transcription orthographique                                 |
| <b>Concepteur(s)</b>    | Jean-Yves Antoine (LI, Université de Tours)                                 |
| <b>Recueil</b>          | Mariette Bessac (CLIPS-IMAG désormais LIG GETALP)                           |
| <b>Transcripteur(s)</b> | Pascale Nicolas (VALORIA), Julien Foulon (VALORIA)                          |
| <b>Diffusion</b>        | libre sous réserve du respect d'une convention d'utilisation                |

### 1.2 Enregistrement : tâche et conditions d'enregistrement

Le corpus OTG a été enregistré par le CLIPS-IMAG (désormais LIG GETALP) en conditions réelles au sein de l'Office du Tourisme de Grenoble 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 conditions d'enregistrement sont celles d'un office très fréquenté, d'où un rapport signal sur bruit assez médiocre. Les enregistrements ont été effectués sur deux pistes séparés à l'aide d'un enregistreur DAT. Deux microphones directifs étaient orientés l'un vers le client (caché) et l'autre vers l'agent. On dispose donc de deux fichiers audio par dialogue.

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.

Au total, 7 heures d'enregistrement ont été conservées. Ce corpus oral a fait l'objet d'une première distribution sur CD-ROM par le CLIPS-IMAG. Cette distribution est restée limitée aux membres de l'ARC

<sup>1</sup> AUF : ex AUPELF-UREF.

<sup>2</sup> [http://www.info.univ-tours.fr/~antoine/parole\\_publicue](http://www.info.univ-tours.fr/~antoine/parole_publicue)

« Dialogue Oral ». Ce corpus comprenait pour chaque dialogue deux fichiers audio au format wav ainsi qu'un fichier d'annotation décrivant brièvement la transaction, ses buts et sa réalisation.

### 1.3 Transcription orthographique

Enregistré en conditions réelles, ce corpus présente un nombre important de transactions de qualité sonore passable ou médiocre. La transcription des dialogues fortement bruités s'est avérée difficile voire impossible : bien souvent, les transpositeurs ne sont pas parvenus à s'accorder sur de nombreux passages. Dans une telle situation, le laboratoire DELIC suggère de représenter les différentes transcriptions alternatives. Compte tenu du nombre important de passages conflictuels dans certains dialogues, nous avons au contraire choisi de ne pas intégrer d'hypothèses alternatives et de privilégier les dialogues ne présentant aucune ambiguïté d'écoute pour le transpositeur. C'est pourquoi la transcription n'a été réalisée que sur des dialogues de qualité sonore jugée "excellente" ou "bonne" (tableau 1). Notons toutefois que certains énoncés de bonne qualité sonore présentaient encore des parties inaudibles et n'ont pas été transcrits dans cette première phase. Il en va de même pour une trentaine de transactions qui correspondaient à des trilogues. Dans ce cas, il s'est avéré difficile de faire une distinction sûre entre les productions des deux clients concernés.

| Durée       | Qualité sonore : Excellente | Qualité sonore : Bonne |
|-------------|-----------------------------|------------------------|
| < 30 s      | 159                         | 135                    |
| 30 s - 1 mn | 35                          | 42                     |
| 1 mn - 2mn  | 12                          | 24                     |
| 2 mn - 3 mn | 0                           | 2                      |
| > 3 mn      | 0                           | 0                      |

**Tableau 1** : Répartition par durée des dialogues du corpus OTG (qualité excellente ou bonne).

Au final, 315 dialogues ont été transcrits, qui correspondent à environ 2 heures d'enregistrement (tableau 2). Le corpus distribué a une taille de 26 000 mots transcrits.

|                        |                                   |
|------------------------|-----------------------------------|
| durée d'enregistrement | 117 minutes                       |
| nombre de dialogues    | 315                               |
| nombre de locuteurs    | 5 réceptionnistes / 315 touristes |
| nombre de mots         | 25 695                            |

**Tableau 2** : Répartition par durée des dialogues du corpus OTG (qualité excellente ou bonne).

### 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.
- 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="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"/>
Connaissance
</Turn>
<Turn speaker="spk2" startTime="3.856" endTime="4.24">
<Sync time="3.856"/>
du monde
</Turn> </Section> </Episode>

```

Figure 1 : Extrait du corpus OTG : transcription sans annotation morpho-syntaxique (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: Connaissance
<004> client
c: du monde

```

Figure 2: Extrait du corpus OTG : transcription sans annotation morpho-syntaxique (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 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 OTG

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.

## 2 Distribution du corpus et convention d'utilisation

Le corpus OTG 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.



Hormis les frais d'envois susmentionnés, le corpus OTG est distribué gratuitement sous licence *Creative Commons* CC-BY-SA. Cela signifie que vous devez respecter le contrat d'utilisation suivant :

- **BY : paternité** - Vous devez citer les auteurs de ce corpus pour toute utilisation du corpus. Dans le cas d'une publication s'appuyant sur ces travaux, nous vous demandons ainsi de citer les articles référencés dans la description de la ressource jointe à la distribution ou dans la liste ci-dessous.
- **SA : partage des conditions initiales à l'identique** - Vous ne pouvez créer une nouvelle ressource à partir de la ressource existante et en faire ensuite un usage différent de celui imposé par ce contrat. Là encore, nous sommes ouverts à toute utilisation du corpus pour création de nouvelles ressources, mais nous vous demandons de nous contacter pour discuter de ces nouveaux usages.

**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 corpus OTG

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 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 ANNEXE A — Conventions de transcription du corpus OTG

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.

### 4.1 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.

### 4.2 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 :

|                    |                |                    |          |
|--------------------|----------------|--------------------|----------|
| <i>puis-je</i>     | sera transcrit | <i>puis je</i>     | (2 mots) |
| <i>plate-forme</i> | sera transcrit | <i>plate-forme</i> | (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<sup>3</sup>.

<sup>3</sup> Les linguistes travaillant sur l'oral, tels les chercheurs du GARS/DELIC, dénie généralement toute pertinence de la notion de ponctuation dans le langage parlé.

#### 4.2.1 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 :

|                         |        |        |
|-------------------------|--------|--------|
| <i>rire</i>             | annoté | [rire] |
| <i>bruits de bouche</i> | annoté | [bb]   |
| <i>toux</i>             | annoté | [tx]   |
| <i>souffle</i>          | annoté | [pf]   |

#### 4.2.2 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 :

|                               |        |                               |
|-------------------------------|--------|-------------------------------|
| <i>monsieur Lionel Jospin</i> | et non | <i>Monsieur Lionel Jospin</i> |
| <i>mairie de Grenoble</i>     | et non | <i>Mairie de Grenoble</i>     |

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 :

|  |        |
|--|--------|
| <i>Société Nationale des Chemins de Fer</i>      | (SNCF) |
| <i>Transports de l'Agglomération Grenobloise</i> | (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.

#### 4.2.3 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*

#### 4.2.4 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

#### 4.2.5 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

#### 4.2.6 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'veis (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.

#### 4.2.7 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{béduť}

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.

#### 4.2.8 Événements acoustiques : pauses

Deux types de pause ont été distinguées :

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



## 5 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.

### 5.1 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.

### 5.2 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 tour 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 : 1ag0365   |
| <001> hôtesse<br>h: <b>bonjour madame</b>   |
| <002> client<br>c: <b>bonjour est ce que vous avez le programme de oui e e je</b> |
| <003> hôtesse+client<br>h: <b>oui</b><br>c: <b>Connaissance</b>                   |
| <004> client<br>c: <b>du monde</b>  |

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

### 5.3 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.

### 1. Présentation

Les enregistrements ont été effectués sur deux pistes séparés à l'aide d'un enregistreur DAT. Deux microphones directifs étaient orientés l'un vers le client (caché) et l'autre vers l'agent. On dispose donc de deux fichiers audio par dialogue. Leur extension est respectivement .afs et .cfs

Ces enregistrements sonores ont été numérisés sous un format brut (raw format) à la fréquence d'échantillonnage de 16000 Hz (16 KHz). Cette information est essentielle pour une bonne écoute des enregistrements.

### 2. Ecoute des fichiers sonores distribués

Les fichiers sons distribués peuvent être écoutés avec n'importe quel éditeur de signal. Nous recommandons cependant l'utilisation de deux utilitaires bien connus dans la communauté scientifique :

- ✓ le logiciel de transcription *Transcriber*
- ✓ l'éditeur de signal *SfSWin*

#### 2.1 Ecoute avec Transcriber

Transcriber est un outils d'aide à la transcription de corpus oraux développé par la DGA (Claude Barras, Direction Générale de l'Armement) et le LDC américain. Ce gratuiciel (*freeware*) permet d'éditer la plupart des formats de signaux de parole et offre une interface interactive très bien conçue pour écouter et transcrire en parallèle ces corpus oraux.

Dans le cas du corpus OTG, l'intérêt de cet utilitaire est précisément de permettre une écoute séparée ou simultanée des deux pistes audio. L'installation de ce logiciel peut nécessiter quelques efforts pour des utilisateurs non informaticiens.

Une fois le logiciel complètement installé, l'écoute de nos fichiers sons se réalise comme celle de tout format supporté par Transcriber. Nous vous recommandons de consulter le manuel d'utilisation du logiciel, qui peut être consulté sur la page de téléchargement du programme : <http://trans.sourceforge.net/>

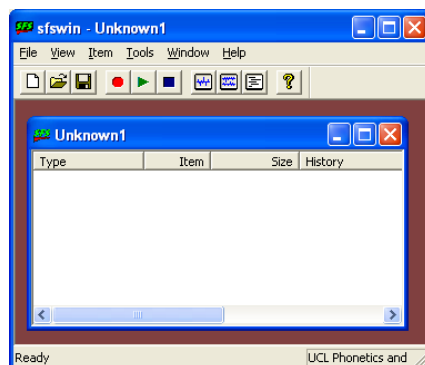
#### 2.2 Ecoute avec SfSWin

SFSWin (*Speech Filling System for Windows*) est un éditeur de signal développé par Mark Huckvale (University College, London) qui est dédié au traitement du signal de parole. Ce gratuiciel (*freeware*) permet d'écoute très simplement les fichiers audio du corpus OTG, ainsi que réaliser des traitements de base (calcul de spectrogramme, suivi de formants, détection de fréquence fondamentale...) sur ces fichiers.

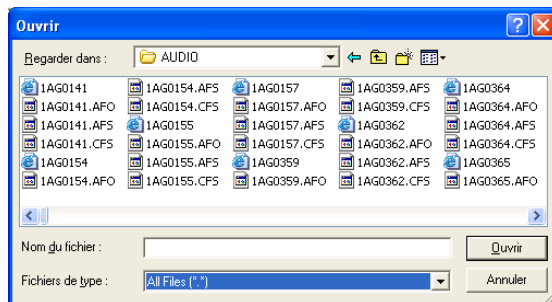
L'installation de ce logiciel sur votre ordinateur est triviale. SFSWin peut être récupéré à l'URL suivante : <http://www.phon.ucl.ac.uk/resource/sfs>.

La lecture des fichiers audio distribués sous SfSWin n'est par contre pas immédiate. SfSWin n'est en effet pas capable de détecter automatique la fréquence d'échantillonnage des fichiers du corpus OTG. Pour écouter correctement ces fichiers, vous devez suivre la procédure suivante.

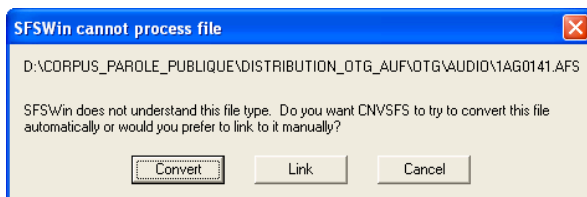
1. Lancer l'utilitaire. La fenêtre de travail représentée à droite apparaît à l'écran.



- Ouvrir le fichier considéré en allant dans le menu File / Open. Choisissez le fichier recherché dans l'explorateur qui apparaît à l'écran, en sélectionnant le type de fichier All Files (\*.\*) dans le menu déroulant du bas.



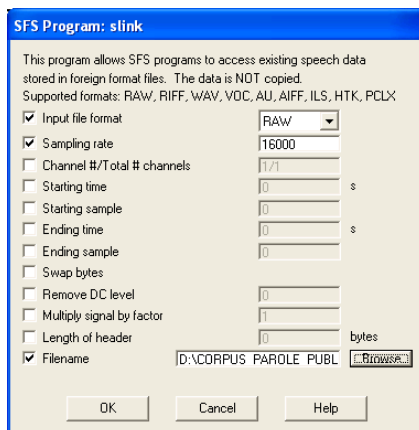
- SfSWin n'arrive pas reconnaître le format des fichiers OTG. Nous allons donc devoir lui préciser certaines informations à la main. Choisissez pour cela l'option Link dans le menu qui apparaît :



- Définissez à la main les informations manquantes dans le formulaire qui apparaît :

- ✓ Type de données (input file format) : RAW
- ✓ Fréquence d'échantillonnage (sampling rate) : 16 000 Hz
- ✓ Nom du fichier (filename) : choisissez celui-ci dans l'explorateur de fichiers qui apparaît après clic sur le bouton Browse.

Cliquez sur OK une fois ces informations saisies. Elles seront conservées pour les prochaines écoutes, pour lesquelles vous n'aurez qu'à modifier le nom du fichier concerné.



- Après validation sur OK, le signal de parole est correctement reconnu : l'item correspondant s'affiche dans la fenêtre de travail de SfSWin. Pour écouter le signal, vous procédez alors comme pour tout signal de parole reconnu par l'utilitaire. Pour savoir comment procéder à partir de ce point, lisez le manuel d'utilisation joint au logiciel.