LE PROJET PAROLE PUBLIQUE DE CONSTITUTION D'UN LARGE CORPUS FRANCOPHONE DE DIALOGUE ORALE : RÉALISATIONS ET PERSPECTIVES

Jean-Yves ANTOINE, Sabine LETELLIER-ZARSHENAS, Igor SCHADLE

RESSOURCES LINGUISTIQUES INFORMATISÉES : LE RETARD DU FRANÇAIS

Corpus et ingénierie des langues

À l'heure de la généralisation de l'outil informatique et de l'accès à de grandes banques de connaissances par l'intermédiaire d'Internet, la place d'une langue dans la société mondiale dépendra de plus en plus de ses possibilités d'utilisation dans les nouvelles technologies d'information et de communication (NTIC). L'ingénierie des langues se doit donc de proposer pour chaque idiome un ensemble de technologies permettant la recherche et la restitution d'informations sous forme écrite, orale ou multimédia. Or, la mise en œuvre de ces technologies dépend de plus en plus de la mise à disposition de vastes ressources linguistiques dans la langue concernée.

Les ressources linguistiques ont en effet acquis un rôle central en ingénierie des langues du fait la généralisation de méthodes empiriques basées sur les données (estimation de modèles de langage stochastiques sur de grands corpus). Parallèlement, la linguistique accorde une attention de plus en plus manifeste aux études sur corpus. La représentativité de ces corpus étant pour partie fonction du nombre d'observations qu'ils contiennent, on observe une tendance générale à la constitution de ressources linguistiques de plus en plus importantes. Un seul exemple suffira à montrer l'importance de cette question. Il s'agit du développement récent de l'*American National Corpus,* (Ide & Macleod, 2001) financé par le *Linguistic Data Consortium* (LDC) américain. Alors que la communauté scientifique disposait déjà avec le *British National Corpus* (BNC) d'une ressource linguistique de taille considérable concernant l'anglais britannique (Leech, Garside Bryant, 1994) le LDC a jugé impérieux le développement d'un corpus équivalent consacré exclusivement à l'anglais américain. Lorsqu'on connaît la proximité de ces deux idiomes, on saisira alors la nécessité de la mise en place de ressources linguistiques équivalentes pour le français.

Le retard du français

Avec des ressources pionnières telles que le *Trésor de la Langue Française*[1] pour l'écrit, ou BDSONS (Carré et al. 1984) pour la parole, le français était certainement une des langues les mieux représentées en corpus au début des années 1980. Malheureusement, le développement de l'ingénierie des langues n'a pas été accompagné d'un accroissement comparable des ressources linguistiques francophones. Au contraire, le retard pris par le français dans ce domaine est considérable. Quelques chiffres pour en donner la mesure :

– on dispose à l'heure actuelle pour l'anglais écrit d'un corpus annoté grammaticalement de 100 millions de mots (*BNC*) alors que la communauté francophone ne dispose même pas d'un corpus équivalent d'un million de mots (Véronis 2000),

– dans le domaine des corpus arborés (*treebanks*) correspondant à une annotation des structures syntaxiques des énoncés écrits, l'anglais dispose depuis une dizaine d'années de plusieurs corpus de l'ordre de 3 millions de mots tels que le *Penn Treebank* (Marcus, Santorini et Marcinkiewicz 1993) ou le corpus *IBM/Lancaster* (Leech et Garside 1991). Ce type de ressource n'est toujours par disponible en français. La constitution d'un corpus arboré francophone de taille sensiblement inférieure a été lancée par le laboratoire TALANA (Abeillé, Clément et Kinyon 2000). À ma connaissance, cette ressource n'est cependant toujours pas distribuée.

– En ce qui concerne le langage parlé, la situation est tout aussi inquiétante. On dispose pour l'anglais de corpus de parole transcrite de plus de 10 millions de mots en anglais (partie orale du *BNC*). Pour le français, seuls trois corpus oraux (CORPAIX, ELILAP et VALIBEL) présentent une taille atteignant ou dépassant le million de mots. Leur statut est en outre relativement précaire. Le corpus CORPAIX (Blanche-Benveniste, Rouget et Sabio 2002), collecté depuis de nombreuses années par le GARS à Aix-en-Provence, ne semble pas pouvoir être utilisé facilement en ingénierie des langues. Il s'agit en effet d'un corpus de plus d'un million de mots recueilli et transcrit par des linguistes à une époque où leur utilisation par des moyens informatiques n'était pas envisagée. Il n'est de toute manière pas diffusé. À ma connaissance, le corpus VALIBEL[2] (Dister 2002) de l'Université de Louvain-la-Neuve (Belgique), qui forme avec ses 4 millions de mots le plus grand corpus oral en francophonie, n'est lui non plus pas distribué. Le corpus ELILAP[3], recueilli par l'université de Leuven (Belgique), est lui disponible librement sous format électronique. Regroupant environ un million de mots pour 500 heures d'enregistrement, il reprend pour partie le corpus d'Orléans enregistré entre 1968 et 1971 par des sociolinguistes. Comme CORPAIX, ces deux corpus ne concernent malheureusement que des monologues ou des interviews peu interactifs. Leur utilisation en dialogue oral homme - machine n'est donc pas envisageable. Défendant une « *approche interactive différente de celle qu'adoptent les grammairiens de l'oral* », Catherine Kerbrat-Orecchioni (1999) rappelle à ce sujet combien le degré d'interactivité influe sur les productions orales. Dans cette perspective, les seules ressources actuellement disponibles semblent être quelques corpus pilotes ou de magicien d'Oz (corpus Air France[4], CIO, SNCF) collectés au cours des années 1980 dans le cadre du GDR-PRC Communication Homme - machine. Ils regroupent chacun moins de 50 000 mots. Cette taille de devrait pas être dépassée, pour ce qui concerne le dialogue oral proprement dit, par les corpus *Français de Référence* (DELIC 2002) et C-ORAL-ROM

(Cresti *et al.* 2002) sur lesquels travaille actuellement le laboratoire DELIC. Ne comportant qu'une petite partie de dialogues oraux réellement interactifs, ceux-ci ne seront en outre pas diffusés librement au terme de ces projets.

– Enfin, si l'on considère les corpus de parole annotés grammaticalement, on peut citer la disponibilité de donnée de tailles conséquente pour l'anglais, toujours dans le cadre du *BNC*. À l'opposé, Valli et Véronis (1999) notent qu'aucun corpus oral annoté significatif n'était pour l'instant disponible en français. À mon sens, le plus grand corpus de français (écrit) étiquetté grammaticalement et distribué est constitué par la partie française du corpus JOC réalisé dans le cadre du projet MULTEXT. La partie francophone distribuée dans le cadre de ELDA regroupe 200 000 mots uniquement.

Ce petit état de l'art pourrait laisser croire à une domination écrasante de l'anglais sur les autres langues. Si cette prédominance est réelle, la situation du français est cependant nettement plus préoccupante que celle d'autres langues. Pour ne citer que quelques exemples, on observe ainsi que l'allemand, le néerlandais, le suédois, le japonais ont bénéficié d'efforts qui sont sans aucune mesure avec la taille des corpus francophones actuellement disponibles.

Pour une libre diffusion des corpus francophones

On peut donner deux causes à cette situation inquiétante du français. La première provient de l'absence, jusqu'à une date récente, d'une politique cohérente de développement de ressources francophones de la part du Ministère français de la Recherche, mais aussi de celui chargé de la Francophonie. Quoique salutaires, les actions menées par la francophonie dans le cadre de l'AUF (*Agence Universitaire de la Francophonie*) furent ainsi loin d'être à la mesure des besoins des chercheurs francophones. L'AUF a d'ailleurs abandonné depuis toute politique de soutien à la recherche en ingénierie des langues. Nous sommes ainsi loin des efforts consentis par la (D)ARPA américaine pour l'anglais, et il est symptomatique que les ressources francophones les plus significatives proviennent bien souvent de recherches menées au Canada (corpus bilingue aligné HANSARD par exemple), en Belgique (corpus oraux ELILAP et VALIBEL déjà cités) ou par des départements de français langue étrangère (on citera par exemple le corpus d'Orléans, qui a été repris et saisi par l'Université Libre d'Amsterdam). Le Ministère de la Recherche semble avoir enfin pris conscience de l'importance de ce problème. Le programme *Technolangue* comprend ainsi un volet ressources linguistiques qui devrait favoriser non seulement la constitution de corpus francophones, mais surtout leur diffusion.

En effet, si les ressources francophones actuellement diffusées sont relativement rares, de nombreux corpus existent pourtant au sein des laboratoires français. En l'absence d'une politique favorisant la diffusion de ces ressources, ces laboratoires ont adopté une attitude de repli qui a consisté à conserver les données qu'elles étaient parvenues à constituer par un travail de longue haleine et sur financement propre. Comme le fait remarquer Laurent Romary (2000 : 194): *« [ces équipes] ont l'impression de se faire piller ou, tout simplement, de perdre une partie de ce qui fait leur connaissance scientifique. »*

Rares sont ainsi les laboratoires français qui se sont lancés, à l'image par exemple des universités de Mons ou de Louvain en Belgique, dans une politique d'ouverture consistant

à diffuser librement les ressources (corpus ou logiciels) qu'ils ont constitué. Cette politique peut pourtant s'avérer très intéressante scientifiquement. Laurent Romary (2000 : 194-195) poursuit ainsi : *« Ce dernier argument s'avère dans les faits trompeur : les équipes qui se sont engagées dans la voie de la diffusion large de leurs ressources et des méthodologies associées ont bénéficié d'un regain de renommée non négligeable. »*

Notre équipe a précisément choisi de diffuser librement l'ensemble des ressources qu'elle produit. En adoptant cette politique, notre objectif est tout d'abord de participer à l'effort de constitution de larges ressources linguistiques francophones. Mais nous sommes également soucieux de faire « vivre » au maximum les données que nous avons recueillies. Celles-ci peuvent en effet rencontrer un intérêt en dehors des problématiques qui ont présidé à leur constitution.

Comme je l'ai rappelé précédemment, les corpus de dialogue oral ou écrit sont très peu représentés en français. L'action ASILA [5] sur l'étude du dialogue, qui a été récemment lancée par le CNRS, devrait favoriser la diffusion de ressources conservées jusqu'à présent dans certains laboratoires et compléter les maigres ressources constituées il y a plus de 15 ans dans le cadre du GDR-PRC Communication Homme-Machine. En réponse à cette absence de données, notre équipe développe depuis deux ans dans le cadre du programme *Parole Publique* une activité de constitution de corpus de dialogue oral diffusés librement sur Internet [*http://www.univ-ubs.fr/valoria/antoine/parole_publique*]. Le développement *stricto sensu* de ressources linguistiques, qui se traduit rarement par de grandes avancées théoriques, est peu valorisé scientifiquement. Il n'en est pas moins essentiel au développement de l'ingénierie des langues. Il me semble donc important de présenter ici cette initiative.

PAROLE PUBLIQUE :
UNE INITIATIVE POUR LA LIBRE DIFFUSION DE CORPUS DE DIALOGUE ORAL

Objectifs

Le corpus *Français de Référence* du DELIC vise une description linguistique globale de ce que l'on pourrait appeler abusivement le « français parlé standard ». À l'opposé, le programme *Parole Publique* répond aux besoins spécifiques de la Communication Homme – machine et se focalise sur un genre ou registre (Biber 1988) particulier : le dialogue oral spontané et finalisé. Ce « genre » pourrait être caractérisé par sa spontanéité ainsi que par une grande interactivité. Favorisée par le caractère finalisé de l'échange et l'engagement des interlocuteurs, celle-ci se manifeste entre autres par de fréquents chevauchements et interruptions. Il en résulte souvent – mais pas toujours, loin de là – une assez grande brièveté des échanges. Enfin, le caractère finalisé de l'interaction se traduit par une couverture lexicale relativement restreinte qui limite l'ambiguïté sémantique et favorise *a priori* les procédés de raccourci (ellipses, co-référence…).

Compte tenu de la carence en corpus d'études amont, le programme *Parole Publique* concerne principalement le recueil de corpus pilotes, voire éventuellement de magicien d'Oz (simulation de la machine par un expert humain). Il vise la réalisation d'une ressource échan-

tillonnée qui soit représentative de la diversité du dialogue oral finalisé. C'est-à-dire que notre propos est de regrouper des données correspondant à des contextes interactifs variés.

Cette variabilité concerne en premier lieu la tâche et le domaine applicatif au sein desquels l'interaction est circonscrite. On sait en effet que la généricité des systèmes interactifs vis-à-vis de l'application est une des questions les plus débattues actuellement en Communication Homme –Machine (Hirschmann 1998). Pour l'heure, les domaines d'applications suivants sont étudiés :

– renseignement touristique,
– réservation hôtelière,
– renseignement administratif,
– portail vocal entreprise,
– accueil standard téléphonique.

Cette couverture thématique devrait cependant être étendue au fur et à mesure du développement du corpus. En particulier, nous comptons faire appel et répondre aux suggestions de la communauté scientifique, afin de « faire vivre » au maximum les corpus de nous réaliserons.

Par ailleurs, si notre objectif n'est pas d'arriver à un échantillonnage sociolinguistique équilibré des locuteurs enregistrés, l'adaptation des systèmes au type d'usager est une forme de généricité qui sera de plus en plus étudiée à l'avenir. Ici, cette problématique est uniquement abordée du point de vue de l'âge du locuteur (adolescents, adultes et personnes âgées). Nous espérons recueillir une collection de corpus suffisamment diversifiée pour autoriser une analyse différentielle des usages langagiers vis à vis de cette variable.

Enfin, précisons qu'un mode d'interaction orale sera privilégié tout au long du programme. Il s'agit du dialogue téléphonique, qui répond à des besoins socio-économiques bien identifiés. Les autres modes d'interaction orale ne sont néanmoins pas oubliés, comme en témoignent les corpus que nous distribuons dès à présent (cf. *infra*). De même, la constitution de corpus multimédia pourra être envisagée.

Présentation technique : contenu des corpus distribués

L'ingénierie des langues, comme les sciences du langage, ayant par le passé souffert de pratiques individuelles trop disparates, il était impératif que les corpus réalisés dans le programme *Parole Publique* répondent à un soucis de normalisation permettant leur réutilisabilité. C'est pourquoi nous avons défini une méthodologie de transcription et de codage qui sera reprise tout au long du projet *Parole Publique*. Celle-ci s'inscrit dans les pratiques les plus communément admises au sein de la communauté scientifique francophone. En particulier :

– La transcription suit les conventions définies pour le français parlé par le GARS (Blanche-Benveniste et Jeanjean 1987) et désormais reprises par l'équipe DELIC. Ces conventions ont été légèrement enrichies par certaines recommandations EAGLES issues du projet SPEECHDAT (Gibbon, Moore, Winski, 1997). Au total, nos transcriptions restent totalement en accord avec les principes d'objectivités défendus par Claire Blanche-Benveniste. À savoir que la transcription ne doit pas être corrompue par des sur-interprétations de prononciation ou par l'indication de faits paralinguistiques. Notre approche peut

même être considérée comme encore plus orthodoxe que celle du GARS. Nous écartons en effet le recours aux multi-transcriptions en cas d'ambiguïté irrépressible de perception. Dans de telles situations, le dialogue n'est tout simplement pas transcrit.

– Les corpus distribués dans le cadre de Parole Publique seront enrichis par une annotation morpho – syntaxique (parties du discours associées à chaque entité lexicale du corpus). Celle-ci se base le jeu d'étiquettes définies au sein de l'action GRACE (Adda et al. 1999), qui avait réuni la plupart des assigneurs de parties du discours du français.

Cette annotation représente une valeur ajoutée qui est essentielle à la mise en œuvre de traitements linguistiques puissants dépassant la simple observation de co-ocurrence de mots réalisée sur des corpus bruts. D'une manière générale, l'utilisation de corpus annotés linguistiquement a montré son intérêt (Véronis 2000) pour l'extraction de ressources terminologiques, le développement d'outils multilingues mais aussi, par exemple, l'apprentissage de modèles de langage avancés pour les systèmes de reconnaissance de parole (Chelba et Jelinek, 2000).

Les fichiers de transcription annotés sont encodés dans le format structuré XML (figure 1). Nous reprenons pour cela la DTD [6] définie par le logiciel libre *Transcriber* (Barras *et al.* 1998) que nous utilisons pour la transcription. L'annotation morphosyntaxique se greffe sur la transcription par l'ajout d'un jeu très limité de balises XML à cette DTD.

```
< ?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 sans annotation morpho – syntaxique en format XML

Ce format XML est également traduit, pour des usages spécifiques, dans un format texte (ASCII) qui conserve une structuration en tours de parole avec ou sans annotation morpho – syntaxique (figure 2). On remarquera que les chevauchements sont toujours représentés dans ce format. L'information d'alignement temporel des tours de parole n'est par contre pas reprise ici.

```
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 sans annotation morpho – syntaxique en format texte (extrait du identique à celui de la figure 1).

Les corpus distribués peuvent être librement récupérés sur le site du programme *Parole Publique*, sous réserve d'acceptation d'une convention d'utilisation peu contraignante. Ils sont également distribués librement dans le cadre de l'action spécifique ASILA et du projet ANANAS[7] du CNRS . Le paragraphe qui suit présente très brièvement les corpus déjà disponibles à l'heure actuelle.

Premiers résultats : corpus OTG et ÉCOLE MASSY

Pour le moment, deux corpus pilotes de dialogue oral (*OTG* et *École Massy* : voir les tableaux 1 et 2 page suivante) sont d'ores et déjà disponibles. Ces deux corpus correspondent à un domaine applicatif, sur lequel portent nos travaux en compréhension de parole : le renseignement touristique .

CORPUS	OTG	ÉCOLE MASSY
Durée d'enregistrement	117 minutes	45 minutes
Nombre de dialogues	315	31
Nombre de locuteurs	5 réceptionnistes / 315 touristes	1 enseignant/19 élèves
Nombre de mots	25 695	5 300

Tableau 1. Description synthétique des corpus OTG et École Massy

Durée	< 30s	30s – 1 mn	1 mn – 2 mn	2 mn-3 mn	> 3 mn
OTG	294	77	36	2	0
École Massy	2	6	16	7	0

Tableau 2. Distribution des dialogues des corpus OTG et École Massy suivant leur durée

OTG

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.

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 et l'autre vers l'agent. Les micros étaient masqués à la vue du client, qui n'était informé de l'expérience qu'au terme de l'échange (procédure semi-clandestine). 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.

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 (ex-GARS) 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). À terme, le corpus atteindra une taille critique de 40 000 mots annotés morpho-syntaxiquement.

École Massy

Sensiblement plus petit (5 300 mots pour 31 dialogues), le corpus *École Massy* a été enregistré et transcrit par notre équipe. Tout en relevant également du renseignement touristique, les dialogues recueillis concernent une tâche plus précise : la planification d'activités de loisirs. Ce corpus répond à une motivation scientifique spécifique : l'étude différentielle des usages langagiers suivant le type d'utilisateur. La population étudiée était constituée de jeunes enfants de sept ans enregistrés dans leur classe. Le corpus *École Massy* regroupe en fait des dialogues homme-homme simulés portant sur la tâche étudiée. Il 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. À 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. À la demande de l'enseignant et à notre grand regret, les enregistrements ont été réalisés en notre absence

De part ses motivations, ce corpus n'est pas directement utilisable pour la conception de systèmes de dialogue oral. L'adaptation des systèmes de dialogue oral à des publics spécifiques – personnes âgées (Privat 2000), handicapés, adolescents, 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.. Par ailleurs, ce corpus est également susceptible d'intéresser psycholinguistes et chercheurs en sciences de l'éducation (Le Cunff 2002).

Vers un large corpus de dialogue oral mis à disposition de la communauté scientifique

Ces deux corpus sont un aperçu de la recherche de diversité (domaine applicatif, type de locuteurs, contexte dialogique) qui préside au programme *Parole Publique*. Ce souci se retrouvera dans nos futures réalisations dans le cadre du programme *Technolangue* du ministère de la Recherche. Plus précisément, nous intervenons dans le projet AGILE-OURaL en qualité de fournisseurs de corpus de dialogue oraux. Ce projet, porté par l'entreprise *Sinequa*, vise à terme la distribution de composants de base pour le traitement automatique des langues (segmenteur, étiqueteur grammatical, outil de repérage d'entités, outil de segmentation thématique). Ces outils, qui sont développés en premier lieu pour le traitement de l'écrit, seront de plus en plus utilisés à l'avenir sur du langage parlé. C'est pourquoi les outils mis en œuvre dans le cadre du projet OURaL seront validés sur des corpus de dialogue oral recueillis et transcrits à cet effet. Ces corpus proviendront des dialogue oraux finalisés recueillis par notre laboratoire, tandis qu'une autre ressource développée par le laboratoire SILEX (Université de Lille) sera constituée de dialogues radiodiffusés (Gasiglia 2002).

Au terme du projet, le laboratoire VALORIA disposera d'un corpus de dialogue oral qui comportera 200 000 mots transcrits et annotés et concernera les différents domaines d'application cités précédemment (§ 2.1). Au vu de la taille des corpus oraux déjà existants, cette ressource peut sembler de faible envergure. C'est oublier que la transcription de dialogue oraux très interactifs, comportant de nombreux chevauchements et interruptions, constitue une activité lourde et coûteuse. Aussi, cette ressource devrait représenter, à la fin de l'année 2004, le plus grand corpus francophone de dialogue oral distribué librement.

Par sa variabilité, cette ressource devrait intéresser un nombre important de chercheurs en linguistique de corpus. C'est en tout cas notre souhait le plus cher.

Notes

1. Constitué à partir des années 1960-1970, le *Trésor de la Langue Française* est un corpus d'envergure à forte connotation littéraire (80 % des données) est également structuré sous forme de dictionnaire. Il est désormais informatisé (TLFi) comme d'autres ressources de l'ATILF (ex-INALF). Contrairement à la base textuelle FRANTEXT, il est d'accès libre (Bernard et al. 2002).
2. Corpus VALIBEL ; [*http://valibel.fltr.ucl.ac.be/val-banque.html*].

3. Corpus ELILAP ; [*http://bach.arts.kuleuven.ac.be/elicop/projetELILAP.htm*].
4. Corpus Air France ; [*http://www.inalf.fr/ananas/site/htm/AirFrance.html*].
5. Projet ASILA ; [*http://www.loria.fr/projets/asila/*].
6. DTD = *Document Type Definition*. Il s'agit d'une notion définie avec le langage de balisage structuré SGML. La DTD attachée à un document SGML décrit formellement l'organisation de l'information au sein de ce dernier. XML constituant un sous-ensemble de SGML, cette notion se retrouve dans ce langage. Pour une présentation rapide de SGML, XML et de leur utilisation en ingénierie des langues, on consultera utilement (Bonhomme, 2000).
7. Projet ANANAS ; [*http://www.inalf.fr/ananas/site/htm/*].

BIBLIOGRAPHIE

ABEILLÉ A. , CLÉMENT L., KINYON A., « Building a treebank for French », Actes *2nd Conference on Linguistic Ressources and Evaluation*, *LREC'2000*, Athènes, 2000, p. 87-94.

ADDA G., MARIANI J., PAROUBEK P., RAJMAN M., LECOMTE J., « L'action GRACE d'évaluation de l'assignation des parties du discours pour le français », *Langues*, 2(2), 1999, p. 119-129.

BARRAS C. *et al*., « Transcriber : a free tool for segmenting, labeling and transcribing speech », Actes *1st Conference on Language Ressources and Evaluation*, *LREC'98*, Grenade, 1998, p. 1373-1376.

BERNARD P., DENDIEN J., LECOMTE J., PIERREL J.-M., « Un ensemble de ressources informatisées et intégrées pour l'étude du français : FRANTEXT, TLFi, Dictionnaires de l'Académie et logiciel Stella », *Actes TALN'2002*, Nancy, 2002, p. 3-36.

BIBER D., *Variety across speech and writing*, Cambridge, Cambridge University Press, 1998.

BLANCHE-BENVÉNISTE C., JEANJEAN C., *Le français parlé : transcription et édition*, Paris, Didier Erudition, 1987.

BLANCHE-BENVÉNISTE C., ROUGET C., SABIO F., *Choix de textes de français parlé : 36 extraits*, Paris, Honoré Champion, 2002.

BONHOMME P., « Codage et normalisation de ressources textuelles », dans Pierrel J.-M. (dir.), *Ingénierie des langues*, Paris, Hermès, coll. « I^{2}C », 2000, p. 173-192.

CARRÉ R., DESCOUT R., ESKÉNAZI M., MARIANI J., Rossi M., « The French language database : defining, planning and recording a large database », *Actes 1984 International Conference on Acoustics, Speech and Signal Processing*, *ICASSP'1984*, San Diego, vol. 3 42-10.1 – 42.10-4, 1984.

CHELBA C., JELINEK F., « Structured Language Modeling », *Computer Speech and Language*, 14(4), 2000, p. 283-332.

CRESTI E. *et al.*, « The C-ORAL-ROM project. New methods for spoken language archives in a multilingual romance corpus », *Actes 3rd International Conference on Language Resources and Evaluation*. *LREC'2002*, Las Palmas de Gran Canaria, vol. I, 2002, p. 2-9.

DELIC, « Le corpus de référence de français parlé », *Actes 2e journées de la linguistique de corpus*, Lorient, 2002, p. 41.

DISTER A., « Normalisation de corpus oraux retranscrits : jusqu'à quel point ? », *Actes des 2es journées de la Linguistique de Corpus*, Lorient, 2002, p. 15.

Gasglia N., « Vers un corpus thématisé de dialogues radiodiffusés : défense et illustration » *Actes 2es journées de la Linguistique de Corpus*, Lorient, 2002, p. 19.

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

Hirschman L., « Language understanding evaluations : lessons learned from MUC and ATIS », *Actes 2nd Conference on Language Ressources and Evaluation*, *LREC'98*, Grenade, 1998, p. 117-122.

Ide N., Macleod C., The American National Corpus : a standardized resource for American English. *Actes Corpus Linguistics'2001*, Lancaster, 2001, p. 274-280

Kerbrat-Orecchioni C., L'oral dans l'interaction : une liberté surveillée. *Revue Française de Linguistique Appliquée*, *RFLA*, 4(2), 1999, p. 41-55.

Le Cunff C., « De l'usage des corpus en didactique de l'oral : recherche et formation », *Actes des 2es journées de la Linguistique de Corpus*, Lorient, 2002, p. 25.

Leech G., Garside R., « Running a grammar factory : the production of syntactically analysed corpora or " treebanks " », Johansson S., Strenström A.-B. (Eds.), *English computer corpora : selected papers and research guide*, Mouton de Gruyter, Berlin, 1991, p. 15-32.

Leech G., Garside R., Bryant M., « CLAWS4 : The tagging of the British National Corpus », *Actes 14th International Conference on Computational Linguistics*, *COLING'1994*, Kyoto, 1994, p. 622-624.

Marcus M., Santorini B., Marcinkiewicz M., « Building a large annotated corpus of English : the Penn Treebank », *Computational Linguistics*, 19(2), 1993, p. 313-330.

Privat R., « Interrogation multimodale de consultation de serveurs d'informations : application aux personnes âgées », *Actes des 1res Rencontres Jeunes Chercheurs en IHM*, *RJC-IHM'2000*, Île de Berder, 2000, p. 127-130.

Romary L., « Outils d'accès à des ressources linguistiques », dans Pierrel J.-M. (dir.), *Ingénierie des langues*, Paris, Hermès, coll. « I²C », 2000, 193-212.

Valli A. et Véronis J., « Étiquetage grammatical de corpus de parole : problèmes et perspectives » *Revue Française de Linguistique Appliquée*, 4(2), 1999, p. 113-133.

Véronis J., « Annotation automatique de corpus : panorama et état de la technique », dans Pierrel J.-M. (dir.), *Ingénierie des langues*, Paris, Hermès, coll. « I²C », 2000, p. 235-250.