Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle

Roxane Bertrand, Philippe Blache, Robert Espesser, Gaëlle Ferré, Christine Meunier, Béatrice Priego-Valverde, Stéphane Rauzy

Le CID - *Corpus of Interactional Data* -

Annotation et Exploitation Multimodale de Parole Conversationnelle

R. Bertrand*, P. Blache*, R. Espesser*, G. Ferré**, C. Meunier*, B. Priego-Valverde*, S. Rauzy*

** Aix-Marseille Universités – Laboratoire Parole et Langage*
CNRS/Université de Provence
29 avenue Robert Schuman, 13621 Aix-en-Provence, cedex 1

{roxane.bertrand;philippe.blache;robert.espesser;christine.meunier;beatrice.priego-valverde;stephane.rauzy}@lpl-aix.fr

*** Centre International des Langues, Département d'Etudes Anglaises*
Université de Nantes
Chemin de la Censive du Tertre, 44312 Nantes, cedex 3
Gaelle.Ferre@univ-nantes.fr

RÉSUMÉ. *La compréhension des mécanismes du langage nécessite de prendre en compte très précisément les interactions entre les différents domaines ou modalités linguistiques, ce qui implique la constitution et le développement de ressources. Nous décrivons ici le CID (Corpus of Interactional Data), corpus audio-vidéo de 8 heures, en français, constitué au Laboratoire Parole et Langage (LPL). L'annotation multimodale du CID inclut la phonétique, la prosodie, la morphologie, la syntaxe, le discours et la mimo-gestualité. Les premiers résultats de nos études sur le CID permettent de confirmer l'intérêt d'une annotation multimodale pour mieux comprendre le fonctionnement du discours.*

ABSTRACT. *The understanding of language mechanisms needs to take into account very precisely the interaction between all the different domains or modalities, which implies the constitution and the development of resources. We describe here the CID (Corpus of Interactional Data), an audio-video corpus in French recorded and processed at the Laboratoire Parole & Langage (LPL). The corpus has been annotated in a multimodal perspective including phonetics, prosody, morphology, syntax, discourse and gesture studies. The first results of our studies on the CID lead to confirm the relevance of an analysis which takes into account as many linguistic fields as possible to draw up a more precise knowledge of discourse phenomena.*

MOTS-CLÉS : *schéma d'encodage multimodal, outils et plateforme d'annotation, phonétique, prosodie, morphologie, syntaxe, discours, geste.*

KEYWORDS : *multimodal coding scheme, annotation tools/platforms, phonetics, prosody, morphology, syntax, discourse, gesture.*

1. Introduction

Le CID, corpus audio-vidéo d'interactions, constitué[1] au LPL, est une ressource unique pour l'analyse de la langue parlée en interaction. Il comporte 8 heures de dialogues filmés, transcrits et annotés pour différents domaines : phonétique, prosodique, syntaxique, discursif et mimo-gestuel. L'objectif de ce corpus est de proposer une ressource de haut niveau (comportant une annotation précise de chacun des domaines), de taille suffisante pour permettre une approche quantitative de l'interaction des domaines considérés.

Les projets existants ayant conduit à la production de ressources annotées ont en effet pris en compte un sous-ensemble de ces domaines (la plupart des projets portant sur l'écrit). Parmi ces grands projets, on peut citer : *MATE* (Multilevel Annotation, Tools Engineering; Dybkjaer *et al*. 1998), *ATLAS* (Architecture and Tools for Linguistic Analysis System, Bird *et al*. 2000), *NITE* (Natural Interactivity Tools Engineering), *Map Task* (conduit au HCRC), *DAMSL* (Dialog Act Markup in Several Layers, Core *et al*. 1997) ou encore *Verbmobil*. Ils ont chacun une spécificité : les uns étant plutôt orientés vers la production d'outils, les autres (comme *Map Task*, Anderson *et al*. 1991) construits autour d'un projet scientifique précis. De plus, il n'existe finalement que peu de ressources concernant le français. Les corpus radiophoniques ont longtemps constitué un support d'investigation privilégié pour l'étude de la parole (cf. corpus ESTER, Galliano *et al*. 2005), sans pour autant donner lieu à la constitution de grandes ressources enrichies. Par ailleurs, plusieurs projets d'analyse de la langue parlée ont également donné lieu à des corpus. Il s'agit essentiellement de corpus de parole spontanée transcrits et pouvant comporter des annotations syntaxiques (cf. *CORPAIX*[2], *VALIBEL*[3]) ou plus centrés sur des aspects phonologiques comme le projet *PFC* (Durand *et al*. 2005). Citons enfin la base *CLAPI*[4] qui constitue un projet d'envergure visant à la constitution d'une base de données de parole en situation d'interaction. Cependant, les conditions d'enregistrement de ces corpus en situation naturelle (interactions à la poste, entre amis, etc.) rendent leur exploitation acoustique délicate. En tout état de cause, aucun ne propose un niveau d'annotation exhaustif de l'ensemble des domaines visés.

Il y a donc une véritable lacune dans ce domaine pour qui recherche une ressource comportant à la fois des informations sur le signal acoustique, la structure syntaxique ou les gestes. Ce constat est à l'origine de la constitution du CID qui a été conçu pour répondre aux besoins très spécifiques des chercheurs aux différents

---

[1] Le *CID* a été élaboré par R. Bertrand et B. Priego-Valverde.
[2] Voir Blanche Benveniste C, Rouget C., Sabio F. (2002) Choix de textes de français parlé: 36 extraits. Honoré Champion, Paris.
[3] http://valibel.fltr.ucl.ac.be/val-banque.html
[4] http://lidil.revues.org/document139.html

niveaux linguistiques, qui vont du niveau le plus bas (phonétique) au niveau le plus haut (discursif, interactif) en passant par les niveaux prosodique, syntaxique et mimo-gestuel. Cet objectif a nécessité la mise en place d'un protocole particulier permettant de recueillir des dialogues d'un intérêt suffisant pour le niveau interactionnel (organisation des tours de parole, phénomènes d'écoute, etc.) tout en préservant un enregistrement de qualité optimale permettant les analyses phonétiques et prosodiques.

Dans cet article, après une présentation de la phase de constitution du corpus, nous décrivons précisément l'ensemble des processus permettant son enrichissement. Nous détaillons ainsi pour chaque domaine d'une part le type d'annotation visé et d'autre part les mécanismes et les outils (quand ils existent) utilisés. Nous terminons par une illustration de l'intérêt de ce type de ressource pour l'analyse de phénomènes linguistiques complexes.

2. Dispositif et protocole

Actuellement, le CID compte 8 x 1 heure de dialogues en français. L'objectif final est fixé à 20 dialogues (5 dialogues selon 4 consignes spécifiques).

2.1. *Tâche*

Les dialogues du CID reposent sur une consigne donnée par les expérimentateurs aux participants avant l'enregistrement. Deux séries d'enregistrement ont été menées selon 2 consignes : dans la première les sujets devaient évoquer des conflits professionnels et dans la seconde des situations insolites dans lesquelles ils s'étaient trouvés. Cette consigne a été présentée comme support thématique permettant aux locuteurs de s'engager assez vite dans la conversation mais elle n'était destinée qu'à rester un support puisqu'il a été précisé d'emblée aux locuteurs qu'ils pouvaient, s'ils le souhaitaient, s'en distancer à tout moment. Cependant, un corpus répond toujours à des besoins spécifiques : en l'occurrence, les expérimentatrices ont cherché à éliciter des phénomènes langagiers spécifiques de polyphonie tels que des discours rapportés ou des énoncés humoristiques.

2.2. *Choix des sujets*

10 femmes et 6 hommes sont impliqués dans des dialogues non mixtes en face à face. Les 16 locuteurs, de langue maternelle française, sont pour la moitié d'entre eux natifs de la région Provence-Côte d'Azur (ou y résident depuis plus de 20 ans) et pour l'autre moitié issus d'autres régions françaises. Au moment de l'enregistrement,

ils résidaient tous depuis plusieurs mois à ou aux environs de Marseille. Ce sont des familiers du laboratoire, condition nécessaire pour leur éviter un stress trop important lié à une situation relativement embarrassante en soi (enregistrement filmé), mais qui aurait pu l'être davantage pour des locuteurs peu familiers d'un tel lieu. Tous les participants ont donc été choisis parmi les membres du laboratoire, aucun ne connaissant pour autant les finalités de l'enregistrement. Par ailleurs, ils ont été choisis en fonction de leur degré de familiarité et de leur habitude à converser ensemble. Une telle habitude garantissait qu'ils partageaient une réelle *histoire conversationnelle*, cette dernière favorisant des échanges plus spontanés et fructueux ainsi qu'une certaine facilité à prendre de la distance par rapport à la tâche dans laquelle ils étaient engagés.

2.3. *Conditions d'enregistrement*

Les deux participants ont été enregistrés dans une salle type studio d'enregistrement. Ils étaient assis côte à côte et légèrement orientés l'un vers l'autre, à une distance d'un mètre environ, similaire à celle d'une conversation naturelle. Les sujets portaient un micro-casque afin d'enregistrer chacune des voix sur piste séparée, la qualité optimale des données orales ainsi obtenues les rendant alors exploitables pour l'ensemble des niveaux linguistiques. Le CID présente ainsi l'avantage de permettre l'exploitation acoustique des phases de chevauchement de parole, très fréquemment ignorées en raison non seulement de la difficulté à les transcrire mais aussi de les analyser acoustiquement, les logiciels de traitement du signal peinant encore à séparer les voix[5]. De ce fait, très peu d'études ont été menées sur ces questions en vue de valider leur rôle dans la structuration des discours. Enfin, les sujets ont été également filmés, en plan large et fixe.

2.4. *Caractéristiques conversationnelles du CID*

Si tous les sujets se sont accommodés de la consigne en cherchant à la satisfaire, ils s'en sont également souvent distancés en s'autorisant des séquences parallèles «libres». Le CID apparaît ainsi comme un type intermédiaire entre des données *naturelles authentiques* et des corpus *orientés tâche* (type *Map Task*[6]).

Par ailleurs, les dialogues du CID sont très similaires aux interactions conversationnelles. Parmi les principaux critères retenus pour caractériser ces dernières, seul l'objectif *externe* (lié à la consigne) les distingue d'une réelle conversation dont l'objectif est dit *interne* car centré principalement sur la relation. Excepté ce critère, on constate une totale symétrie des interactants en termes de

5 Voir notamment les travaux du LIA (Laboratoire d'informatique d'Avignon).
6 Il existe de nombreuses versions adaptées à la langue cible (italienne, suédoise, etc.).

statuts et de places, l'absence d'un tiers (comme dans l'interview ou le débat) pour gérer la circulation de la parole. L'organisation des tours de parole obéit aux principes d'alternance établis dans le modèle des tours de parole (Sacks *et al*. 1974): ils ne sont pas pré-déterminés, les locuteurs s'octroient eux-mêmes la parole en se désignant comme locuteur principal et/ou simple allocutaire. Les différents types de transitions coexistent, l'alternance des tours s'opérant de façon plus ou moins coopérative (*smooth/non-smooth transitions*, Koiso *et al*. 1998). Le style est informel. Enfin, si les dialogues du CID sont globalement conversationnels, ils reflètent aussi une hétérogénéité de séquences discursives propre à toute interaction : de nombreuses narrations notamment dues à la consigne, mais aussi des séquences argumentatives, explicatives ou descriptives.

3. Un premier niveau d'enrichissement du CID : la transcription

3.1. *Découpage en unités inter-pausales*

Préalablement à toute annotation, le signal de parole du CID a été pré-découpé automatiquement en unités inter-pausales (Interpausal Unit -désormais IPU-)[7]. Les IPU sont des blocs de parole bornés par des pauses silencieuses d'au moins 200 ms (durée variable selon les langues). Chaque locuteur étant enregistré sur un canal audio distinct, ce découpage en IPU est bien adapté. En raison de sa nature formelle, objective et repérable automatiquement (Koiso *et al*. 1998), l'IPU est souvent utilisée sur des corpus de taille importante. Les auteurs la substituent en l'occurrence à d'autres unités d'analyse dont le découpage nécessite l'intervention manuelle d'experts, telles que le tour de parole ou encore l'unité intonative. A ce stade, le découpage du CID en IPU visait à faciliter la transcription mais aussi, en raison de la longue durée du signal de parole (1 heure), les étapes ultérieures de phonétisation et d'alignement avec le signal audio (cf. la durée des IPU, tableau 1).

Effectif IPU	*durée moyenne (ms)*	*médiane (ms)*	*quantiles [25% 75%](ms)*
13872	1923	1390	[600 , 2770]

Tableau 1 : *Distribution globale des durées des IPU*

[7] La procédure automatique de segmentation en IPU consiste en une détection du voisement et un seuillage sur l'énergie pour distinguer une pause silencieuse d'un temps de silence dans une occlusive par exemple.

3.2. *Transcription Orthographique Enrichie (TOE)*

La transcription du CID, fondée sur celle du Groupe Aixoix de Recherche en Syntaxe (Blanche-Benveniste & Jeanjean 1987), est essentiellement orthographique. Elle spécifie toutefois les phénomènes typiques de l'oral tels que les pauses pleines (« euh », « hum », etc.), les faux-départs, les amorces, les mots tronqués, les répétitions. Elle indique aussi explicitement les noms propres, les noms de lieux et les discours rapportés. Etant donné le découpage en IPU, les transcripteurs n'ont pas dû noter les pauses silencieuses, excepté celles internes aux IPU (donc inférieures à 200 ms mais perceptibles). Les chevauchements de parole (environ 5300 cas), repérés automatiquement grâce aux IPU, n'ont pas été transcrits non plus.

La transcription spécifie par ailleurs certaines réalisations phonétiques particulières observées sur le CID. A priori ces réalisations phonétiques en parole conversationnelle apparaissent plus étendues, en forme et en type, que celles habituellement recensées sur des corpus de parole lue ou contrôlée. Pour ces derniers, une grande partie des phénomènes phonétiques particuliers (type élision ou réduction par exemple) est automatisable par implémentation de règles (Auran *et al*. 2004) puis constitution d'un lexique de variantes systématiques testées lors de la phase d'alignement (« je suis » réalisé [SH][8]; « je sais » réalisé [Se]). En revanche, dans le CID, à côté de ces réalisations phonétiques particulières habituelles, on constate des élisions non standards, des substitutions et/ou ajouts de phonèmes, etc. (« expérience » réalisé [perja~s]; « demande » réalisé [ma~]), qui peuvent rendre la constitution d'un lexique inefficace en raison d'un trop grand nombre d'entrées et de variantes possibles par entrée. De plus, la moitié de nos locuteurs étant d'origine méridionale, les schwas réalisés (en particulier en finale) ou les sons épenthétiques sont aussi spécifiés (« rappelle quand » réalisé [rapel2 ka~t2], extrait de gpd_118, cf. tableau 2). Ces différentes contraintes nous ont donc incités à opérer une désambiguïsation manuelle préalable au stade de la transcription, pour tenter de pallier les limites des outils actuellement disponibles et davantage adaptés à de la parole lue ou contrôlée sur du français standard. C'est en ce sens que nous parlons de transcription orthographique enrichie (désormais TOE), à partir de laquelle deux transcriptions sont automatiquement dérivées :

- une transcription orthographique standard (cf. tableau 2), dont sont dérivés les « tokens orthographiques », est destinée aux modules d'analyse textuelle.

- une transcription orthographique *truquée*, dont sont dérivés les « tokens phonétiques », est destinée au convertisseur graphème-phonème.

La TOE a été effectuée sous le logiciel Praat (Boersma & Weenink 2005) par deux experts, le second corrigeant la version du premier (600 h de travail pour la totalité du CID). Le premier passage a aussi permis d'ajuster les frontières d'IPU.

[8] Les réalisations phonétiques particulières sont codées en SAMPA.

AG gpd_116	mais t par rapport au pyjama tu sais de tout à l'heure dont tu parlais
AG gpd_118	ça me rappelle quand j'étais
AG gpd_119	tu sais j'étais allé j'avais j'ai je suis resté un an à Dumont d'Urville tu sais à Troie
YM gpd_130	ouais ouais
AG gpd_120	en prépa et j'étais interne

Tableau 2 : *Exemple de transcription (orthographe standard) sur un extrait du* CID

3.3. *Annotation phonétique*

L'annotation phonétique manuelle de 8 h de parole n'est guère envisageable. Ce sont dès lors les outils de phonétisation et d'alignement automatique qui déterminent l'annotation phonétique. Ils permettent en outre de traiter des masses de données considérables et donnent une sortie pertinente et exploitable pour une grande partie des études phonétiques.

3.3.1. *Conversion graphème-phonème*

L'étape de conversion est destinée à produire la séquence de phonèmes nécessaire à l'aligneur (cf. liste tableau 3). À partir de la TOE, la transcription orthographique *truquée* seule est fournie au convertisseur utilisé (Di Cristo & Di Cristo 2001) qui s'appuie sur un dictionnaire de formes fléchies et un ensemble de règles modifiables (exception, liaison, etc.). Ici, nous avons implémenté les règles relatives aux seules liaisons obligatoires (un /n/ ami, ces /z/ articles). En sortie, le convertisseur fournit une suite de tokens phonétisés en SAMPA.

SAMPA	A	a	E	e	O	o	2	9	@	i	y	u	a~	o~	e~	9~
LORIA	A		e		o		@			i	y	u	a~	o~	U~	

Tableau 3 : *Liste des phonèmes utilisés par l'aligneur*

3.3.2. *Alignement automatique*

L'aligneur employé, développé au LORIA par D. Fohr et Y. Laprie (Brun *et al*. 2004), est basé sur des modèles de Markov cachés (Hidden Markov Models). L'alignement est effectué IPU par IPU. A partir de la séquence de phonèmes et du signal audio, l'aligneur fournit en sortie la localisation temporelle de chaque phonème sur le signal. L'aligneur connaît 10 macro-classes de voyelles (7 orales et 3 nasales). Il s'agit de modèles acoustiques de phonèmes du français standard, comme

pour la plupart des aligneurs disponibles. L'aligneur fonctionne donc ici dans des conditions difficiles (cf. 3.2). La résolution temporelle de l'aligneur est de 8 ms, et la durée minimale entre deux étiquettes est de 24 ms.

La localisation temporelle des *tokens* phonétiques est ensuite déterminée à partir de leur phonétisation et de la localisation temporelle des phonèmes. Un second module permet, à partir de la phonétisation et de la TOE, de retrouver la correspondance entre *token* phonétique et *token* orthographique (cf. tableau 4).

TOE	t(u)	sais	j'	étais	allé	j'	a(v)ais	+	j'	ai	[je suis,chui]
Tok. phonétique	t	sais	j	étais	allé	j	aais	+	j	ai	chui
Tok. orthographique	tu	sais	j'	étais	allé	j'	avais	+	j'	ai	je_suis

Tableau 4 : *Exemple de tokens phonétique et orthographique sur des phonèmes utilisés par l'aligneur (gpd_119). Les tokens orthographiques composés de plusieurs formes (ex. « je_suis ») sont dissociés lors du traitement morpho-syntaxique*

3.3.3. C*orrection et évaluation de l'alignement*

Afin d'évaluer le taux d'erreur de l'aligneur, nous avons comparé l'alignement automatique avec l'alignement corrigé par deux experts humains. Cette correction consiste à rectifier les marqueurs de début et de fin de voyelles (en fonction du début et de la fin du 2ème formant), à enlever des étiquettes de phonèmes qui ne sont pas réalisés et à insérer des étiquettes de phonèmes absents dans la TOE. Les deux experts ont corrigé chacun un locuteur, la comparaison entre les deux experts n'est donc pas envisageable ici.

L'alignement des voyelles orales de deux locuteurs (un homme, une femme) a été corrigé, soit environ 13000 voyelles. Les résultats portent sur l'ensemble des voyelles, y compris sur les « voyelles » de 24 ms proposées par l'aligneur, et qui comportent une plus forte proportion d'erreurs (Meunier *et al.* 2008). La figure 1 et le tableau 5 illustrent la distribution du désalignement, donné par l'écart (valeur aligneur - valeur corrigée) des frontières gauche et droite des voyelles, et du point milieu.

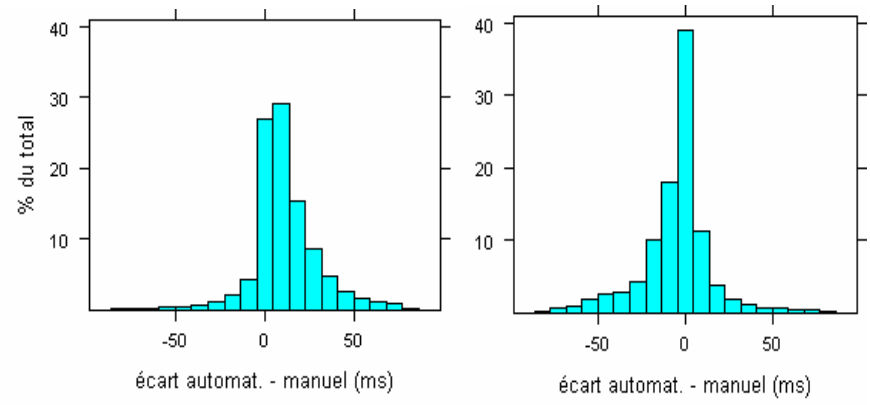


Figure 1 : *Histogramme des écarts en début de voyelle (à gauche) et en fin (à droite)*

L'aligneur tend à centrer les frontières des voyelles. La sous-estimation de la durée des voyelles reste modérée (médiane = 14 ms, et 50% des valeurs comprises entre 2 et 30ms). Les écarts positifs en début de voyelle correspondent à une frontière automatique placée après la frontière manuelle, et inversement pour les frontières en fin de voyelle. 50% des erreurs (écart interquartile) sont comprises entre 0 et 20ms pour la frontière gauche, et entre -13 et 3ms pour la frontière droite; 75% des écarts absolus automatique/manuel sont inférieurs à 20ms en début de voyelle, à 23ms en fin et à 16ms au centre. Ces résultats sont tout à fait comparables à ceux obtenus avec le même aligneur, sur un corpus de mots lus (extrait de PFC), (Nguyen & Espesser 2004). La qualité de l'alignement ne semble pas moindre sur le CID.

écart (ms) (auto. – manuel)	*début voy.*	*fin voy.*	*milieu voy.*
mediane	9	0	3
quantile [25%, 75%]	[0, 20]	[-13, 3]	[-3, 1]
décile [10%, 90%]	[-4, 40]	[-35, 16]	[-15, 24]
écart absolu (ms) |auto – manuel| quantile 75%	20	23	16

Tableau 5 : *Distribution (en ms) de l'écart automatique - manuel*

Nous avons mesuré l'impact du désalignement sur le calcul des 3 premiers formants des 7 macro-classes des voyelles orales de la locutrice, qui présentait le désalignement le plus élevé. Seules les voyelles comprises entre 30 et 300 ms ont été retenues, soit 4378 items automatiques et 5367 items corrigés. Les formants ont été calculés au point milieu de la voyelle avec le logiciel ESPS (Entropic, module *formant*; préemphase de 0.94; le reste des paramètres étant à leur valeur par défaut;

fenêtre d'analyse : 49 ms, LPC à 12 coefficients). Pour chacun des 3 formants, on a estimé l'effet du facteur alignement (manuel vs automatique) en fonction de la voyelle sur la valeur en Bark du formant :

– pour F1, le facteur alignement n'est pas significatif ($p = 0.3$). L'interaction alignement : voyelle est significative pour /i/ ($p = 0.001$), /u/ ($p = 0.016$) et /y/ ($p = 0.035$). Pour ces 3 cas, l'écart entre les valeurs automatiques et manuelles est toujours < 0.2 Bark, et donc inférieur au seuil de discrimination perceptive de 0.28 Bark (Vallabha & Tuller, 2004). La variabilité de F1 sur l'ensemble des voyelles, mesurée par l'écart interquartile (IQR), vaut 1.82 Bark en alignement automatique et 1.93 en corrigé.

– pour F2, le facteur alignement et les interactions sont tous non significatifs ($p > 0.2$). L'IQR en automatique vaut 1.79 Bark, 1.75 en corrigé.

– pour F3, le facteur alignement est légèrement significatif ($p = 0.03$), et correspond à un écart d'environ 0.1 Bark. Aucune interaction n'est significative. L'IQR vaut environ 0.93 Bark pour les deux séries de mesure.

On n'observe donc aucune différence significative de formants entre les mesures faites sur les voyelles alignées automatiquement ou manuellement, ou qui soit supérieure au seuil de discrimination perceptive de 0.28 Bark. La précision des mesures paraît également similaire entre les deux séries, la variabilité des mesures étant très voisine. Ces résultats sont cohérents avec la bonne localisation du point milieu de la voyelle.

3.3.4. *Cas particuliers*

La parole conversationnelle fait apparaître des séquences phonétiques très spécifiques encore peu ou pas abordées. Nous appelons ces phénomènes *conglomérats* (cf. figure 2). Les segments phonétiques ne sont pas identifiables sur le signal (impossible de déterminer ceux qui sont présents ou absents) alors qu'ils sont perçus dans un contexte élargi. Ces séquences sont à distinguer des phénomènes de réduction stéréotypés du type « chais pas » pour « je sais pas ». Il ne s'agit pas non plus de réduction phonologique typique et identifiable comme les élisions classiques (« petit » réalisé /pti/) ou les assimilations (« médecin » réalisé /metse~/). Il s'agit probablement d'une unité de programmation articulatoire non segmentale répondant à certaines contraintes de la parole conversationnelle qu'il convient d'analyser plus finement. Les aligneurs, comme les experts phonéticiens, ne proposent pas de solutions satisfaisantes dans la mesure où les deux approches reposent sur une représentation phonétique discrète du signal de parole. Or, les conglomérats ne peuvent être décrits en ces termes. Ces séquences soulèvent, pour les phonéticiens, des questions théoriques très intéressantes. La figure 2 montre que la sortie de l'alignement donne une succession de phonèmes d'une durée minimale (24 ms) répartis dans le signal sans que cette répartition corresponde à de quelconques événements phonétiques. Un des objectifs des phonéticiens pourrait être de fournir une interprétation linguistique de ces phénomènes.

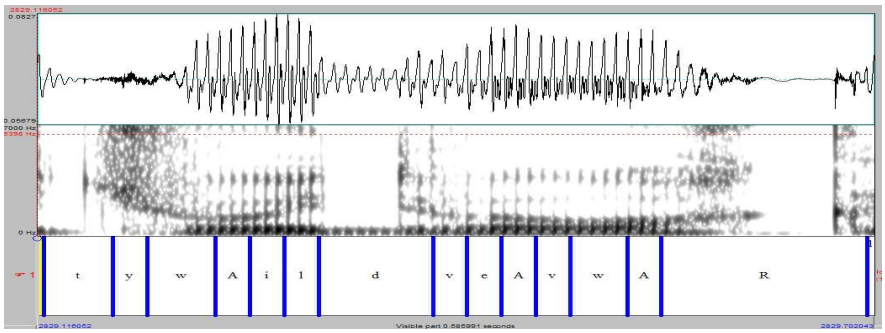


Figure 2 : *exemple de conglomérat "tu vois il devait avoir", transcrit [tywAildveAvwaR]. La réalisation effective est en deçà de la transcription*

La TOE permet de traiter une grande quantité de phénomènes de réduction non prévisibles, mais perceptibles. Or, il est d'autres phénomènes de réduction, eux aussi non prévisibles mais nettement moins perceptibles. Ce fait ne tient pas à la qualité du travail des experts. Le processus de perception de la parole induit une reconstruction partielle de l'information phonétique absente ou déformée. Ce processus échappe à l'introspection des auditeurs. Aussi, est-il normal que ces phénomènes échappent aux transcripteurs, tandis que des réductions plus prototypiques (tel que /pti/), dont la TOE permet de rendre compte, sont parfaitement identifiables.

3.4. *Remarques*

Selon la nature des analyses menées sur de gros corpus, les décalages et omissions sur les étiquettes de phonèmes ne posent pas forcément problème, les erreurs étant noyées dans la masse de données ou même largement éliminées lors du traitement statistique. Ainsi l'approche choisie paraît valide pour certains types de mesures portant sur la partie centrale des voyelles (analyse spectrale, f0 moyen). On observe une sous-estimation des durées, mais son caractère systématique permet justement l'utilisation des durées automatiques de manière relative (analyse inter-voyelles de certaines caractéristiques). Pour d'autres types d'études, portant par exemple sur l'analyse des élisions ou des voyelles très brèves, il ne semble pas possible de se contenter d'une annotation phonétique automatique.
Le taux « d'enrichissement » dans la TOE, défini par le rapport (nombre de réalisations phonétiques particulières) / (nombre de tokens phonétiques) vaut 17%. Ce taux élevé conforte notre choix préalable d'une spécification manuelle. L'analyse et l'évaluation de la TOE, en termes de coût/efficacité, sont désormais possibles.

4. Les autres niveaux d'enrichissement du CID

4.1. *Annotation prosodique*

L'annotation prosodique s'avère une tâche complexe et pour partie seulement, automatisable. D'une part, les primitives et les constructions des systèmes prosodiques se distribuent selon trois axes (métrique, tonal et temporel) qui nécessitent chacun le recours à des représentations plurilinéaires (Di Cristo *et al*. 2004). Ce sont ces travaux s'intéressant à l'interface prosodie/discours, ou ceux décrivant les variations prosodiques liées aux variantes régionales par exemple (Post *et al*. 2006), qui ont proposé les systèmes d'annotation les plus complets pour encoder à la fois les phénomènes globaux (tels que le registre) et locaux (accentuation, proéminences, etc.) aux différents niveaux métrique, intonatif et temporel. D'autre part, l'annotation prosodique reste encore délicate à automatiser dans la mesure où les phénomènes prosodiques sont des objets perceptifs dont les corrélats acoustiques restent encore à mieux identifier. Les outils fondés sur les seuls critères acoustiques (Goldman *et al*. 2007)[9] facilitent l'encodage de gros corpus mais sont inadaptés s'il s'agit d'encoder par ailleurs les catégories phonologiques de la prosodie. Sur ce point précis, des systèmes de transcription ont été développés et sont, pour certains comme *ToBI* (Beckman & Ayers 1997) ou *INTSINT* (Hirst & Di Cristo 1998), devenus des standards. Pour le français, un système tel que *ToBI* ne fait pas encore consensus dans la mesure où son utilisation implique que soit établi l'inventaire phonologique de la langue, ce qui n'est précisément pas le cas du français. À l'opposé, l'un des atouts *d'INTSINT* est qu'il se fonde sur une analyse acoustique qui ne présuppose aucune connaissance du système phonologique de la langue, ce qui le rend utilisable quelle que soit la langue. Il présente en outre l'intérêt d'être automatisé.

Rares sont les systèmes ou les outils permettant de transcrire l'ensemble des phénomènes prosodiques; *ToBI* et *INTSINT* privilégient largement le niveau intonatif. Si le schéma d'encodage du CID vise à terme l'exhaustivité, il concerne pour l'heure, le seul niveau intonatif. Notre approche mêle par ailleurs une annotation à la fois manuelle et automatique : la première est fondée sur une expertise auditive de variations intonatives spécifiques et la seconde utilise les outils *MOMEL-INTSINT* (Hirst *et al*. 2000). Le caractère réversible d'*INTSINT* permet un aller-retour constant entre annotation automatique et manuelle, qui devrait à terme, nourrir le niveau phonologique indispensable au schéma d'encodage du niveau intonatif tel que nous le concevons.

A ce jour, le CID comporte donc un premier niveau d'encodage automatique des cibles tonales. L'algorithme *MOMEL* modélise la courbe de f0 par une séquence de points cibles pertinents linguistiquement, codés ensuite par INTSINT selon un

[9] Certains de ces outils seront testés et évalués sur la parole conversationnelle du CID dans le cadre du projet Rhapsodie (http://rhapsodie.risc.cnrs.fr/fr/index.html).

alphabet de 8 symboles : *T*op, *Mid*dle et *B*ottom, définis globalement par rapport au registre de chaque locuteur; *H*igher, *S*ame et *L*ower définis par rapport aux points précédents ainsi que *D*ownstepped et *U*pstepped (qui sont plutôt liés à des changements moins amples).

Au niveau de l'encodage manuel, un expert a annoté les unités *intonatives (intonational phrase*, désormais *IP*) et *accentuelles* (*accentual phrase*, désormais *AP*)[10] sur 6 heures du CID (soit environ 300 heures de travail). Outre ces unités qui sont les plus communément admises pour le français (pour une revue voir Jun & Fougeron 2002), une autre catégorie (*external phrase EP)* s'est révélée nécessaire pour les cas délicats à classer en IP ou AP. Ces EP peuvent être liés à la présence de marqueurs discursifs tels que *quoi*, *tu vois*, etc. (Post *et al.* 2006), qui sont des éléments typiques des données orales en interaction. Ce point soulève de nouveau (cf. 3.3.4) la question cruciale, valable pour tous les niveaux, du bien-fondé d'utiliser des catégories établies sur de la parole plus « contrainte », mais qui restent néanmoins pour l'heure, les seules catégories disponibles et utilisables. Ceci accroît l'urgence d'analyser des corpus à la fois plus spontanés et variés, pour en faire émerger d'éventuelles nouvelles unités.

Deux experts ont également annoté les *contours intonatifs* sur 6 heures du CID (soit environ 400 heures de travail). Ces contours, définis comme des constructions associant une forme à une fonction (Portes *et al*. 2007), s'ancrent principalement à la fin de l'IP. Leur inventaire, extrait de Bertrand *et al*. (2007), est le suivant : *mineur* (m), *majeur montant continuatif* (RMC), *majeur montant de liste* (RL), majeur descendant (F), *majeur montant terminal* (RT), *majeur montant questionnant* (RQ), *montant-descendant* (RF1), *descendant depuis la pénultième* (RF2), *absence de variation mélodique* (fl). Les annotations manuelle et automatique ne codent pas les mêmes objets mais sont complémentaires. Leur confrontation et l'étude des désaccords entre experts devraient permettre de mieux caractériser formellement les contours intonatifs, favorisant leur implémentation dans de nouveaux outils. Le tableau 6 présente quelques chiffres pour ce niveau d'annotation.

annotations prosodiques	*occurrences*
AP	16.061
IP	21.745
Contours intonatifs	25.997

Tableau 6 : *Nombre d'annotations prosodiques manuelles sur 6 heures du* CID

10 L'*AP* est le domaine de l'accent primaire en français. Des *AP*s peuvent se regrouper pour former une *IP* (marquée par une frontière majeure, une pause, une cohésion mélodique, etc.).

4.2. *Annotation morphosyntaxique*

Le principe de l'annotation morphosyntaxique consiste à associer à chaque mot de l'énoncé la ou les catégories correspondantes. Il existe plusieurs systèmes, ou étiqueteurs, permettant avec un certain succès d'effectuer de façon automatique cette tâche. Pour des raisons de contrôle des formats d'entrée et de sortie, nous avons choisi d'utiliser notre propre étiqueteur.

L'analyse morphosyntaxique repose tout d'abord sur l'utilisation du lexique électronique DicoLPL que nous mettons régulièrement à jour (Vanrullen *et al*. 2005). Il s'agit d'un lexique morphologique très couvrant (environ 500.000 formes), dont la partie verbale (appelée VfrLPL, Rauzy & Blache 2007) a fait l'objet d'une correction très approfondie (disponible sur le site du CRDO : http://www.crdo.fr). La partie nominale est actuellement en cours de correction. Le lexique contient des informations variées morphologiques et syntaxiques.

Notre étiqueteur repose sur une technique probabiliste particulière (Blache & Rauzy 2007, 2008). En l'absence de données fiables concernant l'oral, l'apprentissage de l'étiqueteur a été effectué sur la base de plusieurs corpus étiquetés (notamment le corpus élaboré dans le cadre de la campagne d'évaluation GRACE). Ces données provenant de corpus écrits, le résultat de l'utilisation directe pour l'oral n'est bien entendu pas optimale. Certains items de la TOE (hésitations, amorces, etc.) ne sont pas associés aux tokens fournis en entrée à l'étiqueteur. Le découpage en énoncés est pour l'instant basé sur les IPU (à savoir une pause > 200 ms du locuteur). Une stratégie mieux adaptée au type de données (i.e. dialogues spontanés) est en cours d'élaboration. Une première évaluation effectuée sur les données orales fournies dans le cadre de la campagne d'évaluation EASY montre malgré tout un résultat satisfaisant (un score de F-Mesure de 93% est obtenu pour l'étiquetage des données orales, contre un score de 95% pour l'écrit). Nous sommes en train de corriger manuellement le résultat obtenu pour l'étiquetage du CID de façon à constituer une base de données d'entraînement pour l'étiquetage de l'oral. Un exemple d'étiquetage est proposé dans la figure 3 suivante. A partir de la transcription orthographique enrichie, l'étiqueteur associe à chaque token sa catégorie morphosyntaxique la plus probable codée sous forme d'un jeu de traits au format Multext (ligne CMS) ou sous forme allégée (ligne Tag).

TOE	en	prépa	+	et	j'	étais	interne	#
CMS	Sp-	Nc----		Cc	Pp1-sn-	Vmii1s--	Afpms-	
Tag	Prep	Noun		Conj	Pro	Verb	Adjective	
Chunk	GP				NV		GA	

Figure 3 : *Exemple d'annotations morphosyntaxiques et syntaxiques synchronisées sur la transcription orthographique enrichie (gpd_120)*

Nous avons comparé le résultat de l'étiquetage du CID avec celui d'autres corpus : la partie orale des corpus de la campagne EASY (monologues de parole spontanée, voir par exemple Paroubek *et al*. 2006), un extrait du journal Le Monde et les données fournies dans le cadre de la campagne ESTER (corpus radiophoniques). La figure (4a) montre la répartition des catégories morphosyntaxiques dans chacun des corpus. Le CID présente une proportion significativement plus importante de pronoms et de verbes que les autres corpus, de même qu'un nombre plus faible de noms et de prépositions. Cette différence vient probablement du fait que le CID est constitué de dialogues spontanés qui sont indiscutablement moins contrôlés et moins élaborés que de la parole radiophonique ou des énoncés monologués.

Cette tendance est confirmée par l'étude de la richesse lexicale présentée dans la figure (4b) présentant le nombre de formes différentes par catégorie. Il est intéressant de noter la proximité des résultats entre le CID et ESTER, tandis que la partie orale de EASY est globalement proche de corpus écrits. Ceci peut s'expliquer par le fait que les données de EASY sont des monologues descriptifs professionnels qui entraînent l'usage d'une variété lexicale plus importante qu'en parole spontanée.

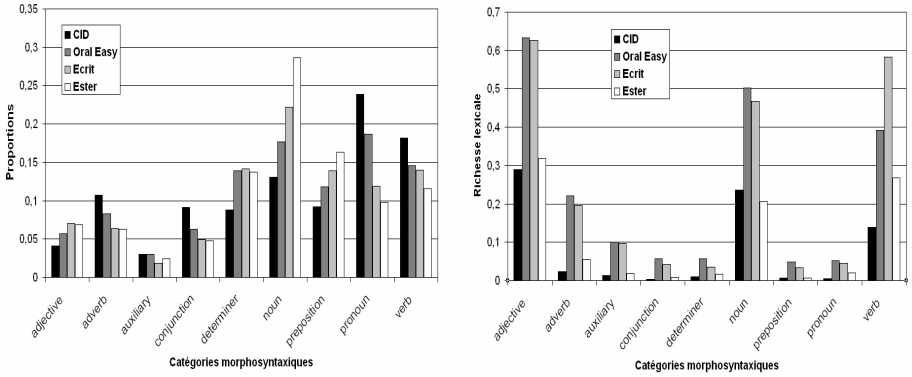


Figure 4a-b : *Proportion et richesse lexicale des catégories morphosyntaxiques*

4.3. *Annotation syntaxique*

L'annotation d'informations syntaxiques reste encore aujourd'hui une tâche complexe et difficilement automatisable. Cette tâche, déjà complexe pour le traitement de l'écrit, l'est encore plus pour l'oral : il n'existe à ce jour aucun système d'analyse syntaxique adapté à la parole spontanée. Nous avons donc décidé, compte tenu de la complexité de la tâche, de nous limiter à une analyse syntaxique superficielle du CID. Nous avons choisi d'utiliser le format d'encodage proposé dans le cadre des campagnes d'évaluation EASY et PASSAGE (cf. de la Clergerie *et al*.

2008). Dans ce type d'analyse, les unités construites ne sont pas récursives. Elles contiennent donc (pour le français) des informations fiables concernant l'identification des types d'unités syntaxiques présents dans la structure ainsi que sur leur frontière gauche. Ce type d'information, même incomplet du point de vue syntaxique, est malgré tout très intéressant dans le cadre de l'étude des interactions avec les autres domaines. Il présente de plus l'avantage de constituer une tâche beaucoup plus simple qu'une véritable analyse syntaxique et donc d'obtenir des résultats plus fiables. Le système utilisé pour annoter le CID, l'analyseur stochastique StP1, a été développé dans le cadre de la campagne PASSAGE (Blache & Rauzy 2008). Il a obtenu de très bons résultats (classé 3ième) : un score de F-Mesure moyen de 93% et, de façon plus intéressante dans le cadre de cet article, une bonne efficacité pour le traitement de l'oral (F-Mesure de 83% en moyenne).

Les annotations produites par ce système correspondent aux unités syntaxiques du formalisme PEAS (cf. Gendner *et al*. 2003). Il s'agit des catégories suivantes : GA (groupe adjectival), GN (groupe nominal), GP (groupe prépositionnel), GR (groupe adverbial), NV (noyau verbal, incluant les clitiques) et PV (verbe non conjugué introduit par une préposition). Un exemple d'annotation syntaxique est présenté figure 3. Les groupes syntaxiques (ligne Chunk) sont constitués d'un ou de plusieurs tokens appartenant à la ligne portant les annotations morphosyntaxiques.

Comme dans le cas de l'analyse morphosyntaxique, nous comparons ici les résultats obtenus pour le CID avec les mêmes corpus (oral de EASY, ESTER et texte écrit extrait du *Monde*). La figure (5a) illustre tout d'abord la taille des groupes formés. S'agissant d'unités non récursives, elles sont donc généralement très courtes en moyenne. On remarque une très grande homogénéité dans les résultats, que ce soit pour du matériel écrit ou oral. Les unités non récursives correspondant en fait à des super-étiquettes (Blache & Rauzy 2008). Cette observation n'est pas surprenante : le chunking appliqué ici consiste essentiellement à déterminer la frontière gauche des unités, chaque nouvelle frontière gauche clôturant l'unité précédente.

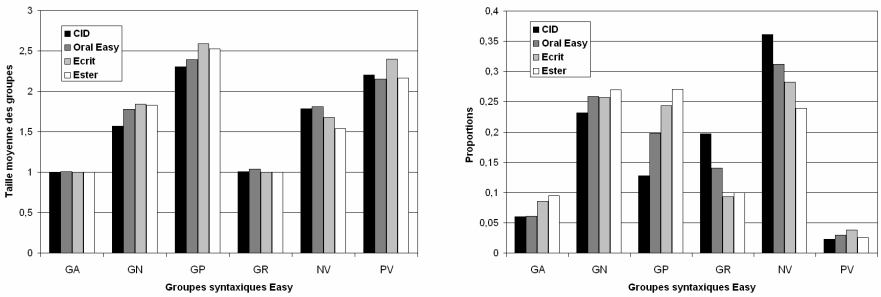


Figure 5a-b : *Taille et proportion des chunks syntaxiques selon le type de corpus*

La figure 5b indique la proportion des types de chunks dans le corpus. On remarque en revanche ici un comportement particulier des données du CID avec d'une part une proportion sensiblement plus forte de noyaux verbaux et de groupes adverbiaux que dans les autres corpus, et d'autre part un nombre significativement plus faible de groupes prépositionnels. Cette observation pourrait refléter une caractéristique de l'oral spontané, s'organisant plutôt autour de constructions verbales (associant des adverbes) que nominales. Cette observation est sans doute également expliquée par la nature de la tâche fixée aux locuteurs, ceux-ci devant raconter et décrire un événement inhabituel.

4.4. *Annotation discursive*

L'annotation discursive a été effectuée de façon totalement manuelle. Ce niveau renvoie ici à des phénomènes proprement discursifs mais concerne également des phénomènes d'ordre conversationnel ainsi qu'énonciatif.

Notre schéma d'encodage envisage un premier niveau d'annotation discursif lié aux événements langagiers affectant les croyances et les savoirs partagés des interlocuteurs. Plus particulièrement, nous avons retenu les unités telles que connecteurs, particules, phatiques/signaux backchannels (produits par l'interlocuteur pour manifester son écoute) qui partagent la fonction de *structurer*, respectivement, l'univers de référence, le discours ou l'interaction (Bertrand & Chanet 2005). Ces unités ne participent pas à l'établissement du contenu propositionnel, elles n'appartiennent pas non plus nécessairement à la référence, mais elles jouent un rôle crucial dans l'interaction en guidant l'interprétation de l'interlocuteur (rôle méta-discursif). Ces unités, dites aussi marqueurs discursifs, du type « quoi, voilà, tu vois, tu sais, enfin », ont été annotées sur 4h du CID. Une étude consacrée aux « enfin » (Bertrand & Chanet 2005), a permis d'illustrer notamment la nature spécifique du corpus. En effet, sur les 4 heures considérées, « enfin » apparaît à 460 reprises dans l'unique fonction de « particule » discursive. A titre de comparaison, sur 80 heures de corpus radiophonique (ESTER) « enfin » apparaît seulement 490 fois. A l'opposé du CID, « enfin » paraît remplir dans ESTER[11] toutes les fonctions (d'organisateur textuel, de connecteur temporel, etc.) recensées dans des corpus de presse écrite. Par ailleurs, les signaux backchannels (du type « mh », « ouais », etc.) ont également été annotés d'un point de vue fonctionnel, sur l'ensemble du CID (cf. 5.3).

Une seconde annotation a permis de caractériser les séquences discursives de narration, établies selon les critères de Küntay & Ervin-Tripp (1997), sur 3 heures du CID. Deux annotateurs ont ensuite annoté les phases formelles des récits, conformément à la typologie de Labov (2007). Il s'agit de l'*Abstract* : début du récit dans lequel est mentionnée de manière anticipée la chute du récit ; l'*Orientation* : présente l'acteur principal, les éléments spatiaux et temporels du récit; la

[11] Etude en cours menée en collaboration avec P. Nocera (LIA)

Complication : série d'événements qui conduisent à un « climax » ('apogée') relatif à l'événement le plus notable du récit; la *Résolution* : dénouement du récit; l'*Evaluation* : commentaire évaluatif de l'un ou l'autre des locuteurs relatif soit à l'apex soit à l'intégralité du récit lui-même; la *Coda* : retour au temps de l'interaction (avec un commentaire de type « et voilà »). Cette structure canonique du récit, construite selon une trajectoire temporelle-séquentielle (Lerner 1992), s'applique plutôt bien à l'ensemble des narrations du CID. Le caractère typiquement conversationnel du CID se traduit cependant par une fréquence élevée de *parenthèses*, grâce auxquelles le locuteur s'assure qu'un point du récit fait bien partie des connaissances partagées. Le Tableau 7 présente quelques données relatives aux séquences narratives du CID :

Phases formelles	*Récit*	*Abstr*	*Orient*	*Compl*	*Resol*	*Eval*	*Coda*	*Parenth*
Effectif	63	0	99	155	28	100	6	145

Tableau 7 : *Annotation manuelle des récits et des phases formelles constitutives pour 3 heures d'enregistrement (6 locuteurs)*

Un autre niveau d'annotation concerne les unités conversationnelles liées à l'organisation des tours de parole. Au stade actuel, 3 heures du CID ont été annotées selon la définition qu'en donne l'analyse conversationnelle : les *unités de construction de tours (turn-constructional units*, désormais *TCU*) sont *les plus petites unités linguistiquement complètes (aux niveaux syntaxique, prosodique et pragmatique) pertinentes au niveau interactionnel* (Selting 2000). Suite au travail de Ford & Thompson (1996) qui ont estimé le poids de chacun des trois niveaux, nous avons considéré les unités intonatives comme étant les plus fiables pour déterminer ces unités. Un TCU doit donc s'achever dans une unité intonative (entre autres). Par ailleurs, nous avons également distingué entre un TCU final (TCU-f) et un TCU non final (TCU-nf). Le TCU-f est constitué d'une seule unité pouvant s'achever dans une place transitionnelle potentielle (TRP), où l'interlocuteur a toute légitimité de prendre son tour. Un TCU non final est l'un des composants incomplets d'un tour complexe défini notamment en termes d'activité discursive (constructions causales, séquence narrative, etc.).

Enfin, un dernier niveau d'annotation concerne les phénomènes énonciatifs qui reflètent l'implication des locuteurs dans leur discours tels que les discours rapportés « repérés » dans la TOE (environ 1000 occurrences) mais sur lesquels aucune annotation supplémentaire, pour l'heure, n'a été faite.

4.5. *Annotation Mimo-Gestuelle*

L'annotation de la gestualité est entièrement manuelle. Elle a été effectuée par un expert et couvre pour l'heure 6 locuteurs pour une durée de 1h15 environ, (soit 100 h de travail). Etant donné le temps dévolu à ce type d'annotation, nous avons choisi d'annoter différents locuteurs plutôt qu'un seul dialogue. Les annotations comprennent les sourires et les rires, les mouvements de tête et des sourcils, ainsi que l'orientation du regard et les gestes manuels des participants.

Pour ce niveau d'annotation, nous avons utilisé le logiciel *ANVIL* Kipp 2003-2006). Notre fichier de spécifications définissant les étiquettes utilisées lors de l'annotation se base sur le standard (McNeill 1992). Allwood *et al.* (2005) ont complété ce standard (*MUMIN*) en y ajoutant les expressions faciales, les mouvements de tête et la direction du regard. Nous avons adapté et complété[12] ce code d'annotation en séparant notamment les différents niveaux d'analyse, à la différence du code MUMIN qui mêle par exemple la description mimo-gestuelle, l'interaction et la modalité du discours en spécifiant par exemple les gestes produits par le locuteur ou l'auditeur. Or, les tours de parole relèvent d'une analyse de l'interaction et non d'une analyse gestuelle. Dans notre annotation gestuelle, nous avons séparé les différentes modalités puisque notre objectif global est de précisément mettre en relation les diverses annotations. ANVIL peut intégrer toutes les annotations, qu'elles soient effectuées sous Praat ou autres, ce qui permet de l'utiliser comme outil de représentation des données. Nous avons donc opté pour l'annotation de la gestualité d'un sujet par fichier d'annotation (que celui-ci soit locuteur ou auditeur) et nous réservons l'information relative aux tours de parole au niveau conversationnel. Nous avons enfin complété le code MUMIN en y ajoutant l'annotation des mouvements du buste. Nous avons également précisé la direction gauche/droite pour les mouvements latéraux, ainsi que la nature de la référence pour les gestes déictiques[13], vers un objet de discours abstrait ou concret.

Les gestes annotés sur le CID sont décrits dans le tableau 8 ci-dessous. Certaines catégories doivent être explicitées pour les non-gestualistes : les phases gestuelles font référence aux travaux de McNeill (1992), dans lesquels le geste est décomposé en différentes phases telles que le temps de préparation (la main par exemple quitte la position de repos et va se mettre dans la configuration pour réaliser le geste), la phase de réalisation du geste à proprement parler, puis la phase de rétraction (où la main, pour reprendre le même type de geste, retourne à sa position de repos). Avant la phase de rétraction, le geste peut également être tenu (*hold*). A ces phases, suite à Loehr (2004), nous avons ajouté une dimension qui note l'apex du geste (point où le geste atteint son déploiement maximal par rapport à la position de repos). Enfin, le

---

[12] Avec la collaboration de Claire Maury-Rouan (LPL).

[13] Nous retenons les gestes déictiques, les plus connus, mais il existe plusieurs catégories de gestes décrits ci-après. La nature de la référence abstraite/concrète ne s'applique en revanche qu'aux gestes déictiques, ainsi qu'à certains mouvements de tête et directions du regard.

point de contact est utilisé pour les adaptateurs, qui impliquent un contact entre la main et une partie du corps, soit du locuteur lui-même, soit de l'interactant.

Tête/Visage	*Mains*	*Corps*
Expression faciale	Symétrie/asymétrie du geste	Mouvement du buste
Mouvements des sourcils	trajectoire de la main	
Ouverture des yeux	Configuration de la main	
Direction du regard	Type sémiotique de geste	
Ouverture de la bouche	Phases gestuelles	
Configuration des lèvres	Apex	
Mouvements de tête	Point de contact	
Emotions/Attitudes exprimées	Hauteur de réalisation du geste	
	Position du geste dans l'espace gestuel du locuteur	

Tableau 8 : *Nomenclature des gestes annotés dans le* CID

L'une des principales difficultés dans l'annotation de la gestualité réside dans la grande variabilité inter et intra-locuteurs. Pour pallier ce problème, nous avons effectué deux types d'annotations :

– Le premier réfère aux gestes dont la variabilité est faible d'un locuteur à l'autre, tels que les mouvements de sourcils qui ne peuvent qu'être levés ou baissés, de manière simultanée ou indépendamment l'un de l'autre. Les mouvements du buste, de la tête et de la bouche relèvent aussi de cette catégorie. Nous avons donc pré-codé dans le fichier de spécifications, un nombre restreint de *valeurs* suffisantes pour décrire l'ensemble des mouvements produits. Leur fonction linguistique est annotée si nécessaire, selon l'étude réalisée. Par exemple, un hochement de tête peut remplir le rôle d'un backchannel ou renforcer le discours.

– Le second réfère aux gestes dont la variabilité intra et inter-locuteurs est forte. Cette variabilité est imputable d'une part au nombre d'articulateurs impliqués dans les gestes manuels (doigts, mains, avant-bras, bras), et d'autre part à l'absence de conventions dans la réalisation de ces gestes. A l'exception des emblèmes, qui renvoient aux gestes manuels codifiés qui remplacent la parole, il n'existe pas en effet de correspondance unilatérale entre un geste manuel et le discours qu'il accompagne. La description de la forme de ces gestes donnera une glose différente pour chaque geste ou presque, ce qui rend le corpus difficilement interrogeable, sans pour autant rendre compte des points communs entre ces gestes. L'annotation des gestes manuels rend compte de leur relation avec le discours, selon la terminologie de McNeill. Il s'agit des gestes iconiques (figurant une action ou un objet concret), des gestes métaphoriques (figurant une idée abstraite), des battements (gestes réalisés en deux temps et qui scandent le discours d'un point de vue rythmique), des emblèmes (gestes conventionnels), des adaptateurs (gestes d'auto-contact), des déictiques (gestes de pointage) et enfin, des gestes désordonnés (*butterworths*). Ces

premières catégories rendent compte de la relation du geste au discours sans toutefois préjuger de leur fonction dans le discours. Un geste iconique peut ainsi renforcer le discours, le compléter ou y suppléer. Il peut également jouer un rôle dans l'organisation discursive sous la forme d'un *catchment* (McNeill *et al*. 2001). Un *catchment* (cf. figure 6) consiste en la récurrence d'un ou plusieurs traits gestuels dans au moins deux gestes dans une même séquence discursive. Au sein d'une même séquence narrative du CID, le trait de verticalité s'est révélé récurrent : tous les gestes iconiques réalisés s'effectuaient dans un plan vertical, trait qui s'opposait à celui d'horizontalité d'une autre séquence discursive (Ferré 2008).

(a) (b) (c)

Figure 6 : *Catchment vertical pour trois gestes iconiques produits dans une même séquence narrative du* CID *accompagnant le discours suivant en (a)* **:** *« on appuie sur l'interrupteur », en (b)* **:** *« ils avaient aussi enlevé les ampoules », en (c)* **:** *« y avait plus que les fils qui pendaient »*

Le tableau 9 présente les statistiques du niveau d'annotation gestuel pour ¼ d'heure d'enregistrement sur 2 locuteurs (représentant environ 25 heures de travail).

Mouvement/geste	effectif
Regard	2390
Tête	2089
Sourcils	506
Bouche	396
Mains	281

Tableau 9 : *Occurrences de gestes pour 2 locuteurs sur ¼ d'heure de dialogue*

5. Exploitation du corpus

Cette partie est consacrée à la présentation de quelques résultats d'exploitation préliminaires sur le CID.

5.1. *Données segmentales*

L'exploitation des données de l'alignement du corpus nous a permis d'extraire quelques données chiffrées de l'occurrence des phonèmes dans un corpus de parole conversationnelle. Le CID est constitué de 272166 phonèmes répartis en 144841 consonnes (53,22%) et 127325 voyelles (46,78%) (cf. tableau 10). Le ratio consonnes/voyelles est conforme aux données issues des analyses d'autres corpus. Les voyelles orales représentent environ 40% des 272166 phonèmes du CID. Ensemble les voyelles e, A et @ représentent 70% des voyelles orales du corpus.

API	*SAMPA*	*nombre*	*%*
e ɛ	e	35055	12,88
a ɑ	A	25106	9,22
ø œ ə	@	15742	5,78
ʀ	R	15637	5,75
s	s	15477	5,69
t	t	14736	5,41
l	l	13737	5,05
i	i	13376	4,91
k	k	12394	4,55
m	m	10841	3,98
p	p	10353	3,80
d	d	9806	3,60
ã	a~	7988	2,93
w	w	7634	2,80
o ɔ	o	7193	2,64
y	y	7054	2,59
n	n	6548	2,41
õ	o~	5997	2,20
v	v	5793	2,13
ɛ̃ œ̃	U~	5047	1,85
f	f	4782	1,76
u	u	4767	1,75
ʒ	Z	4260	1,57
b	b	3495	1,28
z	z	3172	1,17
j	j	2975	1,09
ʃ	S	1969	0,72
ɡ	g	1232	0,45
total		*272166*	*100*

voy.	*nombre*	*%*
e	35055	27,53
A	25106	19,72
@	15742	12,36
i	13376	10,51
a~	7988	6,27
o	7193	5,65
y	7054	5,54
o~	5997	4,71
U~	5047	3,96
u	4767	3,74
total	*127325*	*100*

cons.	*nombre*	*%*
R	15637	10,80
s	15477	10,69
t	14736	10,17
l	13737	9,48
k	12394	8,56
m	10841	7,48
p	10353	7,15
d	9806	6,77
w	7634	5,27
n	6548	4,52
v	5793	4,00
f	4782	3,30
Z	4260	2,94
b	3495	2,41
z	3172	2,19
j	2975	2,05
S	1969	1,36
g	1232	0,85
total	*144841*	*100*

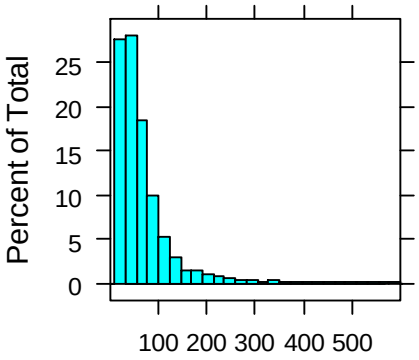
Tableau 10 : *Nombre d'occurrences et pourcentage des phonèmes dans le* CID *(et dans chacune des catégories) d'après la TOE et la conversion graphème-phonème*

La fréquence des phonèmes est globalement conforme aux distributions que l'on observe dans les bases de données (New & Pallier, www.lexique.org). Toutefois, on note une sur-représentation de la voyelle /e/, due d'une part, au fait que /e/ et /E/ ne sont pas distinguées par l'aligneur et, d'autre part, aux réalisations fréquentes de « ouais » (2ème token le plus fréquent dans le CID). Ceci est confirmé par la sur-

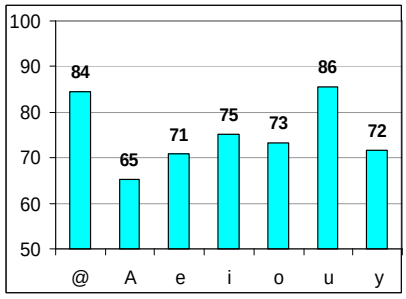
représentation de la consonne /w/. Les voyelles fermées et les voyelles arrondies sont les plus rares. Le /r/ est en revanche sous-représenté dans le CID (en 4ème position alors qu'il arrive en 1ère position dans plusieurs études). La TOE a permis de rendre compte des élisions fréquentes de ce phonème dans la langue parlée (par exemple « l'autre jour » réalisé [lotZur]).

Les durées des voyelles sont très diminuées par rapport aux durées habituelles relevées en parole plus contrôlée (Bartkova & Sorin 1987), cette diminution étant largement supérieure à celle due à l'aligneur (cf. 3.3.3.). La distribution des durées est concentrée vers des valeurs très brèves (figure 7a), ce qui est conforme aux observations faites sur des corpus de parole spontanée (Adda-Decker 2006) : les valeurs des durées des voyelles vont de 24 à 2000 ms, la médiane est à 48 ms. L'importante proportion de voyelles très brèves peut également être due à l'alignement de voyelles non réalisées (cf. 3.3.4) tout comme les valeurs très longues sont dues, soit à des erreurs, soit à des pauses remplies. Pour cette raison, nos analyses acoustiques portent sur les voyelles dont la durée est comprise entre 30 et 300 ms, la durée médiane des voyelles passant alors à 56 ms.

Les durées moyennes des voyelles ne sont pas conformes aux durées standards (Bartkova & Sorin 1987) selon lesquelles les voyelles ouvertes sont plus longues et les voyelles fermées plus courtes. Dans le CID, la voyelle /a/ est la plus courte et le /u/ la plus longue (figure 7b). Il est probable que les durées intrinsèques des voyelles se trouvent ici modifiées en raison des multiples facteurs (fréquence, caractéristique lexicale, fonction discursive, etc.) qui peuvent affecter les paramètres temporels.



7a : *distribution des durées des voyelles dans le* CID *(ms)*



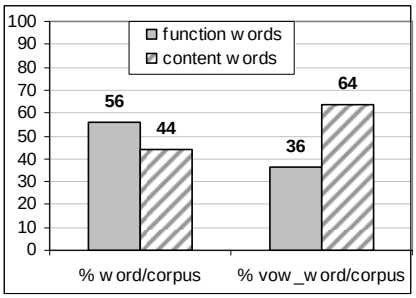
7b : *durées moyennes des voyelles du* CID *(filtre 30-300 ms)*

Figure 7a-b : *Distribution de la durée des voyelles*

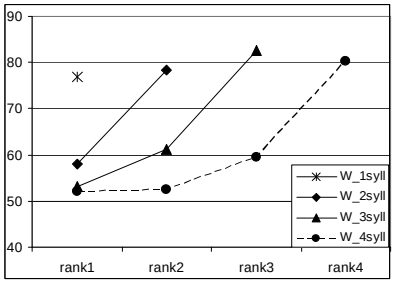
5.2. *Interaction phonologie/lexique*

L'analyse conjointe des niveaux d'annotation phonétique et morphosyntaxique apporte un éclairage sur la réalisation des voyelles du français en parole conversationnelle. 50% des voyelles orales du CID sont réalisées dans seulement 70 mots différents. Une très grande partie des occurrences du corpus concerne très peu de mots (cf. 4.2.). En moyenne, pour chaque voyelle 50% des réalisations concerne une douzaine de mots. On peut ainsi estimer que la production des voyelles se fait dans très peu de mots apparaissant très souvent, ce qui n'est probablement pas sans incidence sur leur réalisation.

Le CID est majoritairement constitué de mots *fonction* (déterminants, prépositions, pronoms, auxiliaires), tandis que les mots de *contenu* (noms, verbes, adjectifs, adverbes) sont proportionnellement moins nombreux. Toutefois, les mots *fonction* sont majoritairement monosyllabiques à l'inverse des mots de *contenu*. Il en résulte que les voyelles se trouvent plus souvent dans des mots de *contenu*, alors que les mots *fonction* sont les plus fréquents dans le corpus (figure 8a).



8a : *à gauche, pourcentage de mots fonction et de mots de contenu dans le* CID*; à droite, pourcentage de voyelles dans deux catégories de mots*



8b : *durées moyennes (ms) des voyelles en fonction de la place dans le mot (rankX) et de la taille du mot (W_Xsyll).*

Figure 8a-b : *Caractéristiques des voyelles selon le type et la taille du mot*

Certains travaux (Johnson 2002 ; Ernestus *et al.* 2006) ont montré que les mots *fonction* les plus fréquents, sont généralement plus affectés par les phénomènes de réduction (élision, assimilation, etc.) que les mots de *contenu*. Nous avons donc étudié l'effet du type de mot sur la réduction temporelle des voyelles : nos résultats n'ont pas confirmé de lien entre la durée des voyelles et le type de mot (fonction vs contenu). On note cependant un net allongement de la voyelle en syllabe finale de mot (figure 8b) qui a pu brouiller l'éventuel lien durée vocalique vs type de mot. Pour neutraliser l'effet de la position syllabique, seuls les mots de structure CV ont donc été retenus. Les mots fonction sont alors plus courts (-10 ms) que les mots de

contenu, ce qui corrobore les travaux antérieurs. Mais conserver uniquement les mots de structure CV présente toutefois l'inconvénient de réduire considérablement la variété des mots restants : en effet 4 mots fonction et 6 mots contenu représentent à eux seuls 50% des occurrences de chaque type. Par exemple, pour la voyelle /A/ : deux mots représentent 94% des mots fonction et deux autres 85% des mots de contenu. La comparaison porte donc plus sur quelques mots sur-représentés que sur deux catégories de mots. S'il peut être dû au choix opéré (structure CV), ce constat reflète en outre une caractéristique majeure de la parole conversationnelle qui comporte des éléments sur-représentés au détriment d'autres (cf. richesse lexicale). L'annotation aux différents niveaux du CID revêt alors tout son sens. Prenons le cas de « ouais », l'un des mots les plus fréquents, qui peut fonctionner comme simple réponse ou comme signal backchannel (cf. 4.4). Plus que les catégories lexicales, les catégories discursives fonctionnelles associées à ces mots très récurrents comme « ouais », pourraient avoir un effet sur la variabilité phonétique constatée (étude en cours).

5.3. *Interaction Prosodie/Morpho-syntaxe/Discours/Geste*

Une autre analyse conjuguant les 4 niveaux prosodique, morpho-syntaxique, discursif et gestuel a été effectuée sur les signaux *backchannels* (désormais BC). Au niveau vocal, les BCs sont largement représentés par le marqueur « ouais » et ses dérivés (« eh ouais », « ah ouais », etc.) ainsi que « mh », « ok », « d'accord ». Au niveau gestuel, les plus fréquents sont les hochements de tête ou les sourires.

Dans l'interaction, les BCs signalent l'attention de l'interlocuteur au discours produit, en vue de préserver la relation interactionnelle tout en régulant l'échange. Les BCs informent à la fois sur le processus d'interprétation mais aussi d'élaboration des discours (Fox Tree 1999), en ponctuant certaines étapes de ces derniers. Nous nous sommes donc interrogés 1/ sur l'existence éventuelle d'indices linguistiques, voire d'une combinaison d'indices, susceptibles de favoriser l'apparition d'un BC; 2/ si ces indices existent, varient-ils en fonction de la modalité vocale ou gestuelle des BCs?

La nature et le niveau d'enrichissement du CID nous ont permis de répondre à ces questions, bien que notre étude ait été limitée à 2 x 15 minutes de dialogues. Les indices linguistiques retenus sont les unités prosodiques (IP et AP), les contours intonatifs, les catégories morphosyntaxiques, les marqueurs discursifs et l'orientation du regard. L'obtention des séquences d'événements suivies de BC repose principalement sur des procédures de tri et d'interclassement des indices retenus. La distribution de ces séquences a été analysée à l'aide de tests de proportion. L'originalité de l'approche réside notamment dans le fait que l'analyse ne portait pas sur le dernier élément adjacent précédant le BC mais sur une fenêtre plus large pouvant comporter plusieurs événements successifs possibles.

Les principaux résultats ont montré un comportement similaire entre les BCs gestuels et vocaux quant aux facteurs prosodiques et discursifs. Comme attendu et à l'opposé des AP, les IP favorisent l'apparition des BC, quels qu'ils soient. Concernant les contours intonatifs, deux d'entre eux, à savoir les montants majeurs de continuation (RMC) et les montants terminaux (RT), privilégient très nettement les BCs, quels qu'ils soient. A l'inverse, les marqueurs discursifs se caractérisent par une absence significative de BC. Quant aux différences liées à leur modalité (vocal vs gestuel), elles apparaissent aux niveaux morphosyntaxique et gestuel : les noms, verbes et adverbes, ainsi que l'orientation du regard favorisent l'apparition des seuls BC gestuels (Bertrand *et al*. 2007). Il s'agit de confirmer ces résultats sur plus de données, en tenant compte notamment du facteur « fonction discursive » des BCs.

6. Conclusion

La création de corpus enrichis est un enjeu essentiel de la linguistique moderne. Ce type de ressource, pour être efficace, doit d'une part être de grande taille et d'autre part comporter des annotations précises pour un grand nombre de domaines linguistiques. Le CID répond à ces critères : il s'agit à ce jour du corpus annoté de parole conversationnelle le plus important pour le français.

Les résultats obtenus à l'aide de ce type de ressource de haut niveau sont extrêmement intéressants par leur variété, mais également par la possibilité d'aborder des questions impliquant de façon systématique plusieurs domaines. Nous avons dans cet article insisté plus particulièrement sur les aspects phonétiques et prosodiques, en présentant différents travaux conduits au sein du LPL. D'autres études exploitant le CID sont en cours, portant plus précisément sur des phénomènes syntaxiques (les détachements) ou encore sur la structure de l'information. Là encore, l'intérêt de ce type de corpus est la possibilité de mettre en relation ces phénomènes avec l'ensemble des domaines et en particulier d'aborder de façon précise la question de la multimodalité[14] (cf. le projet AMI, Carletta *et al.* 2005).

Du point de vue pratique, le processus d'annotation que nous avons présenté dans cet article reste extrêmement coûteux. Si une partie a pu être automatisée, les annotations ou corrections manuelles restent extrêmement importantes. Un des enjeux pour le futur consistera à utiliser le CID comme corpus d'apprentissage permettant de créer de nouveaux outils d'aide à l'annotation. De plus la question de l'accès aux données par un système de requêtage adapté à nos besoins est un enjeu essentiel. Cette question est critique, il ne s'agit pas de développer un énième langage de requêtes propriétaire, mais au contraire de tirer parti de la standardisation de l'annotation pour développer cette fonctionnalité. Toutes ces questions

[14] Nous réutilisons actuellement l'expérience du CID dans le cadre d'un projet portant sur l'étude des interactions multimodales en situation de réalité virtuelle (projet CNRS PEPS « *IHM multimodale en réalité virtuelle* »).

(développement d'outils, stabilisation du standard d'encodage, accès aux données) sont désormais traitées dans le cadre du projet ANR « OTIM »[15].

7. Bibliographie

Adda-Decker M. «De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux», *Actes des XXVI Journées d'Étude sur la Parole*, Dinard, 2006, p. 389-400.

Allwood J., Cerrato L., Jokinen K., Navarretta C. & Paggio, P. «The MUMIN Annotation Scheme for Feedback, Turn Management and Sequencing», in Allwood J., Dorriots B. & Nicholson S. (eds), *Proceedings of the 2nd Nordic Symposium on Multimodal Communication.* Gothenburg, Sweden, p. 91-109.

Anderson A., Bader M., Bard E., Boyle E., Doherty G., Garrod S., Isard S., Kowtko J., Mc Allister J., Miller J., Sotillo C., Thompson H., & Weinert R. «The HCRC map task corpus». *Language and Speech* 34, 1991, p. 351-366.

Auran C., Bouzon C., Hirst D., Lévy C. & Nocera P. «Algorithme de prédiction d'élisions de phonèmes et influence sur l'alignement automatique dans le cadre du projet Aix-MARSEC», *Actes des XXV Journées d'Etudes sur la Parole*, Fès-Maroc, 2004 p. 57-60.

Bartkova K. & Sorin C. «A model of segmental duration for speech synthesis in French», *Speech Communication*, 6 (3), 1987, p. 245-260

Beckman M. & Ayers G. *Guidelines for ToBI Labelling*, 1997. http://ling.ohio-state.edu/phonetics/E_ToBI

Bertrand R. & Chanet C. «Fonctions pragmatiques et prosodie de *enfin* en français spontané», *Revue de Sémantique et Pragmatique* 17, 2005, p. 41-68.

Bertrand R., Ferré G., Blache P., Espesser R. & Rauzy S. «Backchannels revisited from a multimodal perspective», *Proceedings of Auditory-visual Speech Processing* Hilvarenbeek, The Netherlands, 2007, Cederom.

Bertrand R., Portes C. & Sabio F. «Distribution syntaxique, discursive et interactionnelle des contours intonatifs du français dans un corpus de conversation», *TRANEL*, 47, 2007, p. 59-77.

Bird S., Day D., Garofolo J., Henderson J., Laprun C. & Liberman M. «ATLAS : A Flexible and Extensible Architecture for Linguistic Annotation», *Proceedings of the Second International Conference on Language Resources and Evaluation*, Greece, 2000, p 1699-1706.

Blache P. & Rauzy S. «Le moteur de prédiction de mots de la Plateforme de Communication Alternative», *Traitement Automatique des Langues*, 48, 2, 2007, p. 47-70.

[15] «OTIM : Outils de traitement d'information multimodale», est le projet ANR BLAN08-2_349062, coordonné par Philippe Blache, Laboratoire Parole et Langage.

Blache P. & Rauzy S. «Influence de la qualité de l'étiquetage sur le chunking : une corrélation dépendant de la taille des chunks», *Actes de TALN*, Avignon, 2008, p. 290-299.

Blanche-Benveniste C. & Jeanjean C. *Le français parlé, Transcription et édition.* Paris. Didier-Erudition/ InaLF, 2e éd., 1987.

Boersma P. & Weenink D. *Praat : doing phonetics by computer*, http://www.praat.org/

Brun A., Cerisara C., Fohr D., Illina I., Langlois D., Mella O. & Smaïli K. «Ants : le système de transcription automatique du Loria», *Actes des XXV Journées d'Etudes sur la Parole*, Fès, 2004, p. 101-104.

Carletta J., Ashby S., Bourban S., Flynn M., Guillemot M., Hain T., Kadlec J., Karaiskos V., Kraaij W., Kronenthal M., Lathoud G., Lincoln M., Lisowska A., McCowan I., Post W., Reidsma D & Wellner P. «*The AMI Meetings Corpus*», *Proceedings of the Measuring Behavior 2005 symposium on Annotating and measuring Meeting Behavior*, 2005, 12 p.

Core M. G. & Allen J. F. «Coding dialogs with the DAMSL annotation scheme», In *Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*, Boston, MA, 1997, p. 28-35.

de la Clergerie E., Ayache C., de Chalandar G., Francopoulo G., Gardent C. & Paroubek P. «Large Scale Production of Syntactic Annotations for French», *Proceedings of the International Workshop on Automated Syntactic Annotations for Interoperable Language Ressources*, Hong Kong, 2008, p. 45-52.

Di Cristo A. & Di Cristo P. «Syntaix, une approche métrique-autosegmentale de la prosodie», *Traitement Automatique des Langues*, 42, 1, 2001, p. 69-111.

Di Cristo A., Auran C., Bertrand R.**,** Chanet C., Portes C. & Regnier A. «Outils prosodiques et analyse du discours», in A.C. Simon, A. Auchlin et A. Grobet (eds), *Cahiers de Linguistique de Louvain 30/1-3*, Louvain-la-neuve : Peeters, 28, 2004, p. 27-84.

Durand J., Laks B. & Lyche C. «Un corpus numérisé pour la phonologie du français», In G. Williams (ed.) *La linguistique de corpus.* Rennes : Presses Universitaires de Rennes, 2005, p. 205-217.

Dybkjaer L., Bernsen N., Dynkjaerand H., Mckelvie & Mengel A., *The MATE Markup Framework*. Rapport interne, MATE Deliverable D1.2, 1998, http://mate.nis.sdu.dk/

Ernestus M., Lahey M., Verhees F. & Baayen R. H. «Lexical frequency and voice assimilation», *JASA*, 120(2), 2006, p.1040-1051.

Farnetani E. «Coarticulation and connected speech», *The Handbook of Phonetic Sciences*, Hardcastle W.J., Laver J. (eds), Blackwell, Oxford, GB, 1997, p. 371-404.

Ferré G. «Récits de femmes -Analyse multimodale du récit conversationnel en français : une étude de cas-», *Actes du Congrès Mondial de Linguistique Française*, Paris, 2008, p. 715-730.

Ford C. E. & Thompson S. A. «Interactional Units in Conversation : syntactic, intonational and pragmatic resources for the management of turns», In *Interaction and Grammar*, E. Ochs, E. A. Schegloff & S. A. Thompson (eds), 1996, p. 134-184, Cambridge UP.

Fox Tree J. E. «Listening in on Monologues and Dialogues», *Discourse Processes* 27(1), 1999, p.35-53.

Galliano S., Geofftois E., Mostefa D., Choukri K., Bonastre J.-F. & Gravier G. «The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News», *Proceedings of Interspeech*, Lisboa, 2005, p. 1149-1152.

Gendner V., Illouz G., Jardino M., Monceaux L., Paroubek P., Robba I. & Vilnat A. «PEAS, the first instantiation of a comparative framework for evaluating parsers of French», *Research Notes of EACL 2003*, Hongrie, p. 95-98.

Goldman J.P, Avanzi M., Simon A.C., Lacheret A. & Auchlin A. «A methodology for the automatic detection of perceived prominent syllables in spoken French», *Proceedings of Interspeech*, Antwerp, 2007, p. 98-101.

Hirst D. & Di Cristo A., *Intonation Systems*, Cambridge University Press, 1998.

Hirst D., Di Cristo A. & Espesser R. «Levels of description and levels of representation in the analysis of intonation», in M. Horne (ed) *Prosody* : *Theory and Experiment*, Kluwer : Dordrecht, Pays-Bas, 2000, p. 51-87.

Johnson, K. «Massive reduction in conversational American English», *Proceedings of the Workshop on Spontaneous Speech: Data and Analysis*, 2002, Tokyo; http://vic.ling.ohio-state.edu/massive_reduction.pdf

Jun S.-A. & Fougeron C. «Realizations of accentual phrase in French intonation», *Probus* 14, 2002, p.147-172.

Kipp M. 2003-2006. *Anvil 4.0. Annotation of Video and Spoken Language.* http://www.dfki.de/~kipp/anvil

Koiso H., Horiuchi Y., Ichikawa A. & Den Y. «An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs», *Language and Speech*, 41, 1998, p. 295-321.

Küntay A. & Ervin-Tripp S. «Conversational Narratives of Children : Occasions and Structures», *Journal of Narrative and Life History*, 7, 1997, p. 113-120.

Labov W. «Narrative pre-construction», In M. Bamberg (Ed.), *Narrative -State of the Art*. Amsterdam : John Benjamins, 2007, p. 47-56.

Lerner G. H. «Assisted Storytelling : Deploying Shared Knowledge as a Practical Matter», *Qualitative Sociology 15 (3)*, 1992, p. 247-271.

Lindblom B. «Spectrographic study of vowel reduction», *JASA*, 35, 1963, p. 1773-1781.

Loehr D. Gesture and Intonation. *Doctoral dissertation*, Georgetown University, 2004.

McNeill D. *Hand and Mind. What Gestures Reveal about Thought*. Chicago : The University of Chicago Press, 1992.

McNeill D., Quek F., McCullough K.E., Duncan S., Furuyama N., Bryll R., Ma X.F. & Ansari R. «Catchments, prosody and discourse», *Gesture 1(1)*, 2001, 9-33.

Meunier C., Meynadier Y. & Espesser R. «Voyelles brèves en parole conversationnelle», *Actes des XXVII Journées d'Etude sur la Parole*, Avignon, 2008, p. 97-100.

MUMIN : A Nordic Network for MUltiModal INterfaces. http://www.cst.dk/mumin/

Nguyen N. & Espesser R. «Méthodes et outils pour l'analyse acoustique des systèmes vocaliques», *Bulletin PFC*, 3, 2004, p. 77-85.

NITE : Natural Interactivity Tools Engineering. http://nite.nis.sdu.dk/

Paroubek P., Robba I., Vilnat A. & Ayache C. «Data Annotations and Measures in EASY the Evaluation Campaign for Parsers in French», *Proceedings of the 5th International Conference on Language Resources and Evaluation*, 2006, p. 314-320.

Portes C., Bertrand R. & Espesser R. «Contribution to a grammar of intonation in French. Form and function of three rising patterns», *Nouveaux Cahiers de Linguistique Française*, 28, 2007, p. 155-162.

Post B., Delais-Roussarie E. & Simon A.C. «IVTS, un système de transcription pour la variation prosodique», *Bulletin PFC*, 6, 2006, p. 51-68.

Rauzy S. & Blache P. «Un lexique syntaxique des verbes du français : VfrLPL», *Rapport de recherche RAU-3055*, 2007, Laboratoire Parole et Langage.

Sacks H., Schegloff E.A. & Jefferson G. «A simplest systematics for the organization of turn-taking for conversation», *Language* 50, 1974, p. 696-735.

Selting M. «The construction of 'units' in conversational talk», *Language in Society*, 29, 2000, p. 477-517.

Vallabah G.K. & Tuller B. «Perceptuomotor bias in the imitation of steady-state vowels», *JASA*, 116(2), 2004, p. 1184-1197.

Vanrullen T., Blache P., Portes C., Rauzy S., Maeyhieux J.-F., Guénot M.-L., Balfourier J.-M. & Bellengier E. «Une plateforme pour l'acquisition, la maintenance et la validation de ressources lexicales», *Actes de TALN*, Dourdan, 2005, p. 41-48.

VERBMOBIL : Alexandersson J., Buschbeck-Wolf B., Fujinami T., Kipp M., Koch S., Maier E., Reithinger N., Schmitz B. & Siegel M. *Dialogue Acts in VERBMOBIL-2* - Second Edition. Rapport interne, 1998, DFKI GmbH.