

Le CID -Corpus of Interactional Data- : **protocoles, conventions, annotations**

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

Introduction

Le *CID*, corpus audio-vidéo d'interactions, recueilli¹ et transcrit au LPL, constitue une ressource unique en son genre pour l'analyse de la langue parlée en interaction, aux différents niveaux linguistiques (phonétique, prosodique, syntaxique, sémantique, pragmatique et mimo-gestuel).

Depuis plusieurs années, tant en sciences du langage qu'en traitement automatique des langues, il existe pourtant de nombreux projets liés à la constitution et à l'enrichissement de grandes bases de données (cf les projets tels que MATE -Multilevel Annotation, Tools Engineering- ; ATLAS -Architecture and Tools for Linguistic Analysis System- ; NITE -Natural Interactivity Tools Engineering-, Map Task -HCRC- ; DAMSL -Dialog Act Markup in Several Layers- ; Verbmobil). Or, très peu de ressources concernent le français. Et lorsqu'elles existent (cf par exemple CORPAIX², VALIBEL³), ces ressources ne sont pas toujours très facilement accessibles. De plus, les ressources constituées le sont toujours dans un but spécifique, pour répondre aux besoins d'un domaine ou d'une communauté particulière, sans que jamais ne soient pris en compte les besoins d'autres utilisateurs potentiels.

Ce constat est à l'origine de la constitution d'une ressource telle que le *CID*. En effet, les corpus existants référencés ne sont pas « suffisants » pour mener à bien une étude multilinéaire de la parole. Ce point peut sembler paradoxal lorsque l'on sait la taille des corpus utilisés par le TAL notamment. Mais cette notion de « taille », notion toute relative de fait, explique précisément l'une des différences et l'une des difficultés de l'entreprise dans laquelle nous sommes engagés ici. Un corpus TAL doit comporter plusieurs centaines d'heures de parole pour être pertinent. Or, il reste encore très souvent limité à l'écrit bien que des tentatives récentes aient été menées sur de la parole radiophonique (corpus *ESTER*). Dans d'autres domaines tels que la linguistique interactionnelle, des corpus de plus en plus comparables, en taille, à ceux du TAL (voir la base *CLAPI*, Lyon 2) ont été constitués. Or, si ceux-là concernent bien des données orales, leurs conditions d'enregistrement en situation naturelle (interactions chez les commerçants, à la poste, entre amis, etc.) rendent leur exploitation acoustique délicate. Concernant enfin précisément l'analyse des niveaux phonético-prosodique, des corpus d'une trentaine de minutes peuvent être considérés comme de « gros corpus » tant le niveau d'observation requis (à l'échelle de la syllabe, voire du phonème) est fin.

Le *CID* a donc été conçu pour répondre aux besoins très spécifiques relatifs aux différents niveaux linguistiques, qui vont du niveau le plus bas (phonétique) au niveau le plus haut (discursif, interactif) en passant par les niveaux prosodique, syntaxique, sémantique pour aller jusqu'au mimo-gestuel. Par répondre aux besoins, on entend à terme exploiter, c'est-à-dire manipuler et interroger simultanément, les différents niveaux.

Cet objectif général a donc impliqué une réflexion commune préalable à l'enregistrement des données lui-même qui permette d'allier par exemple l'analyse du niveau prosodique -qui nécessite un enregistrement de qualité optimale- tout en conservant des interactions dignes d'intérêt pour le niveau interactionniste (organisation des tours de parole, phénomènes d'écoute, etc.). Les étapes d'annotation, et principalement celle de transcription, se sont accompagnées d'un travail collectif (en termes de recensement des besoins,

¹ Le *CID* est constitué par R. Bertrand et B. Priego-Valverde.

² Corpus non diffusé. Voir Blanche Benveniste C, Rouget C., Sabio F. (2002) Choix de textes de français parlé : 36 extraits. Honoré Champion, Paris.

³ Corpus de l'Université de Louvain-la-Neuve, non diffusé. <http://valibel.fltr.ucl.ac.be/val-banque.html>

des protocoles et des outils nécessaires), avec toujours présente à l'idée la démarche générale sous-tendant l'enrichissement de ces ressources, à savoir que les divers niveaux d'annotation avaient vocation à être intégrés et mis en relation.

1. Dispositif et protocole

À l'heure actuelle, le *CID* compte 8 heures de dialogue en français : chaque interaction (non mixte) entre 2 participants dure au total environ 1 heure. Les 16 locuteurs (10 femmes et 6 hommes)⁴ sont d'origine régionale différente, mais ils résident pour la majorité d'entre eux depuis plusieurs années dans la région du Sud-Est de la France.

Le *CID* se présente comme un type intermédiaire entre des corpus de données dites « naturelles authentiques » (type *CLAPI*) et des corpus de type Map Task [Anderson&al.91] qualifiés de « corpus orientés tâche »⁵.

1.1. Tâche

Les dialogues du *CID* reposent sur une consigne que l'expérimentateur donne aux participants avant l'enregistrement. Cette consigne⁶ est présentée comme support thématique permettant aux locuteurs de pouvoir s'engager assez vite dans la conversation et de ne pas être « à court » dans une situation imposée de discours et de mise en présence d'un autre. Mais elle ne reste qu'un support et il est précisé aux locuteurs qu'ils peuvent à tout moment s'en distancer s'ils le désirent.

Cependant, comme nous l'avons rappelé en introduction, un corpus répond toujours à des attentes et/ou à des besoins spécifiques. Le choix d'une consigne est directement dépendant de l'intention des expérimentateurs qui souhaitent, par le biais de celle-ci, voir émerger des phénomènes langagiers spécifiques⁷.

1.2. Conditions d'enregistrement

Les deux participants sont enregistrés dans une salle type studio d'enregistrement. Ils sont assis côte à côte et légèrement orientés l'un vers l'autre, à une distance d'un mètre environ qui est celle adoptée lors d'une conversation naturelle. Les sujets portent un micro-casque qui permet d'enregistrer chacune des voix sur piste séparée. La qualité optimale des données orales ainsi obtenues les rend alors exploitables par l'ensemble des niveaux linguistiques. À titre d'exemple, le *CID* présente l'avantage de permettre l'exploitation acoustique des phases de chevauchement de parole qui sont très fréquemment évacuées des analyses en raison de la difficulté à les transcrire (problème d'audibilité), mais surtout de les analyser dans des logiciels de traitement du signal de parole qui n'ont pas encore cette capacité de séparer les voix⁸. De ce fait, très peu d'études ont été menées sur ces questions alors qu'on prétend souvent (sans pouvoir le valider, faute de matériau adapté) qu'elles jouent un rôle fondamental dans la structuration des discours et l'organisation en tours de parole.

Enfin, les sujets sont également filmés (en plan large et fixe).

1.3. Choix des sujets

Les sujets sont tous des familiers du lieu. Ce critère est une condition nécessaire présidant au choix des participants. Il permet d'éviter un stress trop important lié à une situation relativement embarrassante en soi (enregistrement filmé), mais qui pourrait l'être davantage pour des locuteurs peu familiers d'un tel lieu.

⁴ Le *CID* est actuellement toujours en cours de constitution.

⁵ Ce type de corpus a fait l'objet de nombreuses versions adaptées à la langue cible (Map Task italienne, suédoise, française, etc.) et aux besoins ponctuels des auteurs.

⁶ Consigne série 1 : évoquez des conflits professionnels. Consigne série 2 : évoquez des situations insolites dans lesquelles vous vous êtes trouvés.

⁷ En l'occurrence dans le *CID*, les expérimentateurs ont cherché, par le biais de cette consigne, à faire émerger entre autres des discours rapportés directs (environ 1000 occurrences).

⁸ Le LIA (Avignon) notamment a développé des outils qui permettent de distinguer par exemple les séquences musicales des séquences parlées (corpus radio).

Tous les participants sont donc des membres du laboratoire (personnel permanent ou doctorants). Par ailleurs, ils sont choisis en fonction de leur degré de familiarité et de leur habitude à converser ensemble. Une telle habitude garantit qu'ils ont une réelle histoire conversationnelle, ce qui favorise une plus grande spontanéité, des échanges plus fructueux et une facilité à prendre de la distance par rapport à la consigne si nécessaire, tout en cherchant à la satisfaire.

1.4. Caractéristiques du corpus

Comme nous l'avons dit, tous les sujets se sont accommodés de la consigne en cherchant à la satisfaire, mais ils s'en sont également largement distancés en s'autorisant des séquences parallèles « libres ». Le *CID* comporte donc des « séquences » variées, dont de nombreuses séquences de narration (essentiellement dues à la consigne), mais aussi d'argumentation, d'explication ou de description. Les dialogues obtenus sont très semblables à des conversations « plus » spontanées : la parole peut être fluente et parfois très hésitante (pauses remplies, amorces, faux-départs, etc.). L'organisation en tours de parole semble obéir aux principes d'alternance connus depuis le célèbre article fondateur de [Sacks et al.74]). On observe à la fois des transitions dites « douces » dans lesquelles l'alternance de locuteur s'effectue sans heurt, c'est-à-dire sans pause trop longue ou sans chevauchement, ainsi que de très nombreuses phases de chevauchements (*smooth vs non-smooth transitions*) [Koiso&al.98].

2. Niveau de transcription du *CID*

2.1. Une Transcription Orthographique Enrichie (TOE)

La transcription de *CID* est essentiellement orthographique, fondée sur celles du GARS [Blanche-Benveniste&Jeanjean87] (voir tableau des conventions, fin de document). Mais elle comporte également tous les phénomènes typiques de l'oral tels que les pauses pleines (euh, mhm, hum, etc.), les faux-départs, les répétitions, les mots tronqués. Elle indique aussi explicitement les noms propres, les noms de lieux, les discours rapportés, etc. Enfin, elle contient quelques « détails » de type phonétique (schwa, particularité régionale, prononciation spécifique, etc.) nécessaires aux étapes suivantes de phonétisation et d'alignement avec le signal audio. Nous parlons de *TOE* (transcription orthographique enrichie) pour caractériser ce premier niveau d'annotation.

La transcription est effectuée sous le logiciel Praat [Boersma&Weenink05] par 2 experts. On conserve la version 2 revue et corrigée. Avant l'étape de phonétisation, un troisième expert effectue un dernier contrôle afin de réduire encore le nombre d'erreurs résiduelles.

2.1.1. Le découpage en *Interpausal-Units* (IPU)

Les transcriptions sont effectuées à partir d'un pré-découpage automatique du signal de parole en *Interpausal-Units* (IPU)⁹. Les IPU sont des blocs de parole bornés par des pauses silencieuses d'au moins 200 ms (durée pouvant varier selon les langues). L'IPU est souvent utilisée sur des corpus de taille importante. Par sa nature formelle et objective, elle se distingue d'autres unités « prosodiques » telles que les unités intonatives¹⁰ par exemple, dont le découpage nécessite l'intervention manuelle d'experts, pouvant de surcroît afficher un désaccord [Koiso&al.98].

Ce découpage en IPU facilite non seulement la transcription mais s'avère indispensable pour les étapes de phonétisation et d'alignement avec le signal audio.

Etant donné le découpage en IPU, il n'est pas nécessaire pour les transcrip-teurs de noter les pauses silencieuses, excepté celles qui sont internes aux IPU, les pauses perçues pouvant être largement inférieures à 200 ms. Les transcrip-teurs corrigent en outre d'éventuelles erreurs de segmentation et

⁹ La procédure automatique de segmentation en IPU consiste en une détection du non voisé/voisé et d'un seuillage pour distinguer une réelle pause silencieuse d'un temps de silence dans une occlusive par exemple.

¹⁰ L'annotation en unités intonatives est effectuée au niveau prosodique.

réajustent alors le découpage. Enfin, comme pour les pauses silencieuses, les phases de chevauchement de parole ne sont pas notées par les transcrip-teurs puisqu'elles sont repérées automatiquement¹¹.

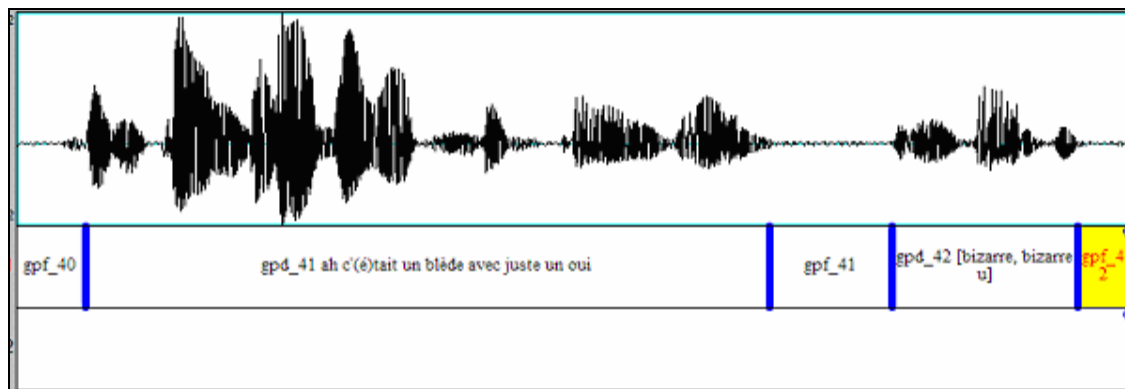


Figure 1: exemple de TOE sur l'extrait "AB_nesterenko_1_clairet_2_nesterenko_espesser_3":
Speaker_AB : ah c'(é)tait un bled avec juste un oui [bizarre, bizarreu]

2.1.2. Illustrations

Le même passage que celui de la figure 1, extrait du dialogue « AB_clairet_1_nesterenko_2_espesser_3.TextGrid », est présenté ci-dessous en (a) dans le format 'textgrid' généré par Praat, qui concerne un seul des 2 locuteurs, et en (b) sous la forme obtenue une fois que les deux textgrids de chacun des deux locuteurs ont été fusionnés. En (a) sont notés les temps de début et de fin (xmin et xmax en secondes) de chaque IPU qui se voit de surcoût attribuer un rang. En (b) apparaissent d'abord les initiales de chaque locuteur ('sp=speaker'), puis le rang de l'IPU, son temps de début en secondes et enfin son contenu.

Pour le détail des conventions de transcriptions, se référer au tableau de conventions situé à la fin du document.

a/ Speaker_AB

```

intervals [60]:
  xmin = 182.41999999999999
  xmax = 184.27000000000001
  text = «gpd_41 ah c'(é)tait un bled avec juste un oui»
intervals [61]:
  xmin = 184.27000000000001
  xmax = 184.59999999999999
  text = «gpf_41»
intervals [62]:
  xmin = 184.59999999999999
  xmax = 185.09999999999999
  text = «gpd_42 [bizarre, bizarreu]»
intervals [63]:
  xmin = 185.09999999999999
  xmax = 188.69999999999999
  text = «gpf_42»
intervals [64]:
  xmin = 188.69999999999999
  xmax = 191.59400351663046
  text = «gpd_43 le bizarre voilà l'insolite ç(a) a un côté bizarre»
intervals [65]:
  xmin = 191.59400351663046

```

¹¹ La procédure d'extraction des phases de chevauchements de parole consiste en une mise en relation temporelle des IPU des deux locuteurs.

xmax = 199.23642504130461
text = «gpf_43»

b/ dialog AB-CM

Sp_AB gpd_41 182.42 ah c'(é)tait un bled avec juste un oui
Sp_CM gpd_70 183.592 complèt(e)ment paumé ouais ouais ouais
Sp_AB gpd_42 184.6 [bizarre, bizarreu]
Sp_CM gpd_71 185.21 ouais
Sp_CM gpd_72 186.05 et euh
Sp_CM gpd_73 187.81 et voilà
Sp_AB gpd_43 188.7 le bizarre voilà l'insolite ç(a) a un côté bizarre

2.2. La phonétisation

L'étape de phonétisation est destinée à produire la suite de phonèmes nécessaire à l'aligneur. À partir de la TOE, la transcription pseudo-orthographique seule est fournie au phonétiseur. Le phonétiseur utilisé [DiCristo&DiCristo01], qui s'appuie sur un ensemble de règles, fournit une suite de *tokens* et leur phonétisation codée en *SAMPA*.

Une bonne qualité de la phonétisation améliore les résultats de l'aligneur, d'où l'importance de la TOE mentionnée plus haut.

2.3. L'alignement

L'aligneur employé, développé au LORIA par D. Fohr et Y. Laprie [Brun&al.04], est fondé sur une technique HMM¹² (<http://www.loria.fr/equipes/parole/>). Il utilise en entrée la liste des phonèmes et le signal audio. L'aligneur fournit en sortie la localisation temporelle de chaque phonème sur le signal.

L'aligneur s'appuie sur un modèle de phonèmes du français standard et fonctionne donc ici dans des conditions difficiles : voyelles dévoisées, réalisation atypique de phonème, etc... La TOE tend donc aussi à faciliter le fonctionnement de l'aligneur (*via* le phonétiseur), en décrivant, voire en simulant dans la mesure du possible, ces particularités : élision, prononciation particulière (ex : « je sais » prononcé « chai »). La localisation temporelle des *tokens* phonétiques est ensuite déterminée à partir de leur phonétisation et de la valeur 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.

exemple :

token orthographique *token* phonétique

je_suis	Sui
allé	Ale
heu	@
c'est-à-dire	stAdiR
nourrir	nuRiR@ (prononciation méridionale)

On obtient ainsi un alignement des *tokens* orthographiques sur le signal. Cette synchronisation entraîne d'ailleurs des vérifications diverses ; après une vérification orthographique (uniquement lexicale), la transcription initiale est rectifiée, et la phonétisation et l'alignement refaits.

Ces deux niveaux d'annotation servent de référence aux autres niveaux, et permettront leur mise en relation, notamment temporelle.

¹² (<http://www.loria.fr/equipes/parole/>)

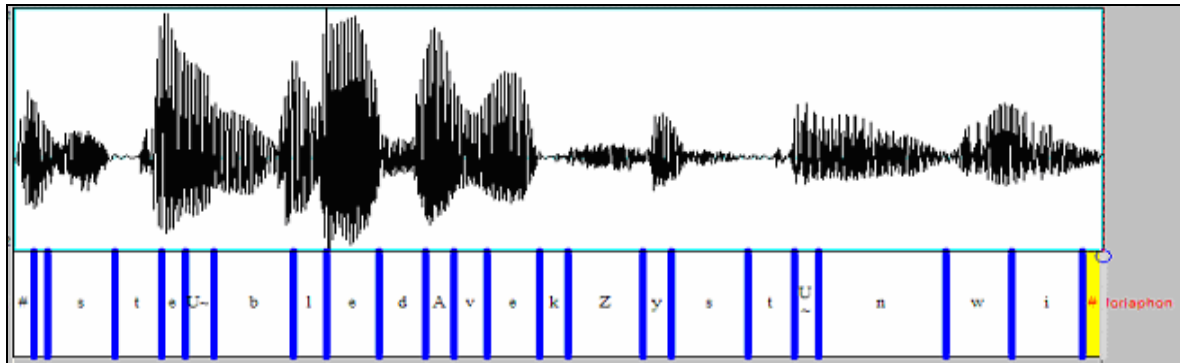


Figure 2: exemple d'alignement sur l'extrait "AB_nesterenko_1_clairet_2_nesterenko_espesser_3":
Speaker_AB : ah c'(é)tait un bled avec juste un oui

3. Annotations multimodales

Les parties qui suivent sont consacrées à l'exposé de chacun des niveaux d'annotation retenus pour le CID. Le contenu relatif à chacun reflète indéniablement une certaine hétérogénéité qui peut s'expliquer entre autres par la volonté de conserver une totale autonomie entre eux et corollairement par le fait que tous, sur cette question de l'annotation, n'en sont pas au même stade de développement. Par exemple, si pour le premier niveau phonétique, nous discutons plutôt des difficultés propres à l'étape de segmentation des unités concernées ainsi que des questions sous-jacentes que soulève précisément l'annotation d'un corpus d'oral spontané, le niveau de l'annotation syntaxique porte lui davantage par exemple sur les outils et le type d'analyseurs développés.

3.1. Annotation du niveau phonétique

3.1.1. Caractéristiques et problèmes généraux de l'annotation phonétique

Le niveau de l'annotation phonétique segmentale a ceci de particulier, avec les annotations prosodique et mimo-gestuelle, qu'il est question de faire correspondre une réalité physique avec un code abstrait. Cette particularité ne va pas sans poser problème et suppose une proposition de choix théoriques et méthodologiques appropriés. À l'inverse des niveaux prosodique et mimo-gestuel, le niveau phonétique ne pose pas de problème quant au choix du code d'annotation. Il n'y a pas plusieurs modèles théoriques possibles dans la nomenclature des unités phonétiques. Ce qui peut être débattu, en revanche, c'est le degré de finesse de l'annotation (phonèmes, phases articulatoires stables et transitoires, etc.) et l'emplacement des marqueurs de frontières.

Le problème majeur de l'annotation phonétique concerne la définition temporelle des unités segmentales de la parole. Il peut parfois être délicat de mettre en correspondance un code phonétique abstrait et discret avec un signal de parole plus ou moins continu. Le problème temporel des unités phonétiques soulève deux questions différentes. La première concerne *l'étendue* des unités segmentales, et donc leur limite, la deuxième concerne *l'existence* d'une frontière entre deux unités.

La coarticulation¹³ est au centre de la première question. Chaque phonème est caractérisé par un ensemble de traits articulatoires. Ainsi, la voyelle /u/ est marquée, entre autres, par le trait d'arrondissement. Ce trait se manifeste d'un point de vue articulatoire par une projection des lèvres en avant. Or, ce geste d'arrondissement est habituellement anticipé, c'est-à-dire qu'il va commencer bien avant la production de la voyelle : dans la syllabe /su/, l'arrondissement des lèvres commence dès le début de la production du /s/. Si l'exemple donné présente deux phonèmes en contact, il a été montré que des traces articulatoires caractéristiques de certains phonèmes pouvaient être identifiées à de plus grande distance du phonème lui-même [Nguyen&al.04] [Hawkins&Nguyen04]. La conséquence de la coarticulation, ou co-production, est

¹³ Chevauchement et interaction des différents articulateurs au cours de la production de segments phonétiques successifs [Farnetani97]

que l'étendue temporelle d'une unité phonétique est assez incertaine et dépasse, le plus souvent, le segment que l'on pense avoir identifié sur le signal de parole.

La deuxième question concerne le problème de la segmentation et donc, de la pose de *marqueurs de frontière*. Il n'y a pas de correspondance bi-univoque entre les discontinuités du signal de parole et les frontières de segments phonétiques [Meunier94]. Cela signifie que les frontières de phonèmes ne correspondent pas à des frontières naturelles dans le signal de parole. De nombreuses discontinuités sont identifiables dans le signal de parole, mais ces discontinuités ne sont pas en rapport avec les frontières d'unités phonétiques. Ainsi, une discontinuité majeure peut être interne à un segment phonétique (l'explosion d'une occlusive, par exemple), tandis que le passage d'un segment au suivant peut n'être marqué par aucune discontinuité identifiable. Un problème important est alors posé lorsque ce que l'on suppose être une frontière est caractérisée par un changement continu d'un état stable vers un autre. Évidemment, dans ce cas, la pose d'un marqueur de début et/ou de fin est arbitraire et dépend des choix théoriques et méthodologiques de l'annotateur.

Si l'on résume les deux questions énoncées ci-dessus, l'annotation phonétique est donc une opération consistant à placer des bornes temporelles sur des segments dont 1/ la durée est incertaine, 2/ la frontière est souvent indéterminable. La segmentation du signal de parole serait-elle alors une erreur ? D'une certaine façon, la réponse est affirmative. Ce débat a longtemps animé la communauté scientifique [Abry&al.85]. Il faut admettre que marquer des frontières d'unités phonétiques est une opération reposant sur des critères arbitraires et répondant à une nécessité d'analyse de ces unités. Il a souvent été question de ne marquer que le centre de la partie stable de chaque unité, ce qui règle le problème des frontières. En revanche, ce choix ne permet pas de travailler sur les durées des unités phonétiques et pose le problème de la synchronisation d'une annotation multi-niveaux. Notre choix est donc ici de poser des frontières de début et de fin d'unité segmentale. Cette délimitation est arbitraire. Il faut admettre que ce que l'on annoté, ce n'est pas le « phonème », mais une structure acoustique qui reflète une partie du phonème. Il a été montré que la coarticulation est plus particulièrement marquée par des indices de lieu d'articulation, tandis que le mode d'articulation pourrait servir de fondement à un marquage temporel plus délimité [Rossi90]. D'une certaine façon, il importe peu que certaines propriétés acoustiques se situent en dehors des étiquettes de frontières. Le principal est de pouvoir, d'une façon ou d'une autre, récupérer cette information dans nos analyses. L'annotation et la délimitation des unités phonétiques sont, en fait, un balisage et un repérage temporel dans le signal qui nous permettent de faire des analyses phonétiques reliées avec d'autres secteurs de l'analyse linguistique.

3.1.2. Spécificité de l'annotation phonétique de la parole spontanée

En parole spontanée, le problème est moins un problème de frontière, bien qu'il le soit encore, qu'un problème d'identification des unités phonétiques prononcées réellement. L'annotation phonétique du corpus du *CID* est une correction d'une annotation préalable réalisée automatiquement (voir 2.3.). Pourquoi est-il nécessaire de corriger cet alignement ? La première raison est que, même si cet alignement semble très performant, on peut constater quelques décalages qui pourraient rendre les analyses futures erronées. Notamment, les marqueurs de début et fin de voyelles empiètent le plus souvent à l'intérieur des voyelles. De façon générale, les marqueurs manquent souvent de précision (voir figure 3). Cela n'est pas forcément un problème dans l'analyse de très gros corpus. Les éventuelles imprécisions sont noyées dans la masse de données. Le corpus *CID*, bien que considérable pour une étude phonétique, représente « seulement », 8 heures de parole.

Une deuxième raison fondamentale de corriger l'alignement concerne le décalage éventuel existant entre le travail des transcripateurs et les sons réellement prononcés. Il est fréquent que les experts transcrivent des segments qui sont absents dans le signal (voir figure 3) ou, plus rarement, que des sons apparaissant dans le signal de parole n'aient pas été transcrits par ces experts. Ce phénomène 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. Il s'agit d'une propriété fondamentale de ce processus. L'efficacité du traitement de la parole passe par cette reconstruction inconsciente. Ce processus échappe ainsi à l'introspection des auditeurs. Aussi, est-il normal que les transcripateurs soient *sourds* à ces phénomènes. La correction de l'alignement doit donc rétablir l'annotation phonétique de façon à rendre compte de ce qui est réellement produit.

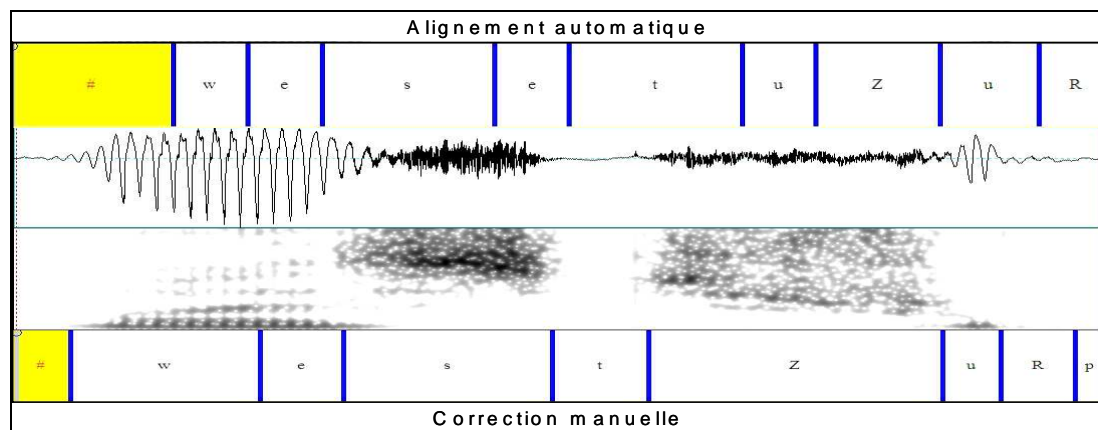


Figure 3: un exemple de correction de l'alignement automatique. En haut, l'alignement automatique de la phonétisation en SAMPA de la séquence transcrite « Ouais c'est toujours... » → /wetuZur/. En bas, la version corrigée de l'expert. Le signal de parole montre 1/ les décalages entre l'alignement automatique et la version corrigée; 2/ la non réalisation de certains phonèmes non perçue par le transcripteur (/wetuZur/ devient /westZur/, le dévoisement n'est pas signalé à cette étape).

Concernant la tâche de segmentation proprement dite, c'est-à-dire, l'identification des frontières, les problèmes spécifiques de la parole spontanée tiennent dans la réalisation « approximante » de nombreux segments. Cela signifie que, pour une forte proportion de fricatives ou d'occlusives réalisées dans un contexte VCV, la constriction est relâchée et leur réalisation devient alors plus proche d'un segment vocalique. Or, les enchaînements phonétiques les plus difficiles à segmenter sont les suites Voyelles+Consonnes Vocaliques [Meunier94]. Les réalisations approximantes augmentent donc, de façon considérable, la proportion d'enchaînements phonétiques pour lesquels la segmentation est délicate. D'une certaine façon, la correction de l'alignement phonétique est déjà un travail de recherche. Nous avons actuellement peu, voire pas, de connaissance sur la réalisation sonore et la phonotactique du français en situation de parole naturelle. Quels sons sont réellement produits ? Lesquels sont omis ? Comment les omissions réorganisent-elles la phonotactique (syntaxe des sons) du français ? La correction de l'alignement nous permettra d'apporter des éléments de réponses à ces questions. La correction nous invite, en outre, à un véritable travail méthodologique. L'observation du signal de parole spontanée nous oblige à reconsidérer les réalisations phonétiques et les critères de segmentation obtenus à partir de parole contrôlée. Il est par ailleurs extrêmement rare de trouver des corpus de parole spontanée de cette dimension dont l'étiquetage phonétique va au-delà d'un alignement automatique d'une transcription. Nous considérons donc ce travail comme un enrichissement majeur concernant l'analyse phonétique de la parole non contrôlée.

3.2. Annotation prosodique

L'annotation prosodique est une tâche complexe car la prosodie est plurisystémique, plurilinéaire et pluriparamétrique [DiCristo&al.04]. Selon les auteurs, on peut l'envisager selon une conception tripartite dans laquelle les primitives et les constructions des systèmes prosodiques se distribuent selon trois axes : l'organisation métrique, l'organisation tonale et l'organisation temporelle, chacune nécessitant le recours à des représentations plurilinéaires.

Les travaux de plus en plus nombreux menés sur corpus durant la dernière décennie ont suscité une large réflexion autour du type de phénomènes prosodiques à encoder, le développement d'outils facilitant cet encodage ainsi que des systèmes de transcription qui sont, pour certains, devenus des standards comme ToBI [Beckman&Ayers97] [Beckman&al.05] ou INTSINT [Hirst&DiCristo98]. Sur ce dernier point cependant, rares sont les systèmes permettant de transcrire l'ensemble des phénomènes prosodiques, ToBI et INTSINT privilégiant largement les phénomènes intonatifs. Par ailleurs, l'adaptation au français d'un système tel que ToBI s'avère encore délicat dans la mesure où son utilisation implique que soit établi l'inventaire phonologique de la langue étudiée, ce qui n'est pas encore le cas pour le français. À l'opposé, l'un des principaux atouts d'INTSINT réside dans le fait qu'il se fonde sur une analyse acoustique qui ne présuppose aucune connaissance *a priori* du système phonologique de la langue. Comme le revendiquent ses concepteurs, INTSINT est donc utilisable sur n'importe quelle langue.

Plus particulièrement pour le français, ce sont les travaux plus récents qui s'intéressent aux liens qu'entretiennent les phénomènes prosodiques avec les phénomènes discursifs [DiCristo&al.04] ou ceux qui cherchent à décrire les variations prosodiques liées aux variantes régionales par exemple [Post&al.06] qui proposent des systèmes d'annotation plus complets, sous la forme d'une *grille multilinéaire* par exemple pour les premiers, où sont encodés à la fois les phénomènes globaux (registre, *span*, *downstep*) et locaux (accentuation, proéminences, etc) aux différents niveaux temporel, métrique et intonatif. Dans une démarche similaire, Post et ses collègues ont développé le système IVTS¹⁴ (adapté du système IViE [Grabe&al.01] qui permet d'encoder différents aspects de la variation prosodique.

Un système d'annotation prenant en compte l'ensemble des phénomènes prosodiques est non seulement souhaitable mais indispensable si l'on cherche à comprendre les relations qu'entretiennent les différents éléments du seul niveau prosodique, avant même celles qu'entretiennent les différents niveaux linguistiques. Or, l'entreprise s'avère extrêmement lourde et coûteuse pour un corpus de plusieurs heures si l'on ne dispose pas d'un ensemble d'outils automatisant le maximum d'étapes dans cet encodage. Aussi avons-nous choisi un système mêlant une approche à la fois manuelle et automatique, fondée pour la première sur l'identification auditive, par des experts, de phénomènes prosodiques particuliers (approche similaire à ToBI) et l'approche MOMEL-INTSINT [Hirst&al.00] pour la seconde.

Notre annotation utilise donc plusieurs tires d'annotation : INTSINT, pour la première, qui permet de repérer et d'encoder automatiquement les cibles tonales, dont le caractère réversible n'interdira pas, si nécessaire, d'éventuelles modifications. Plus précisément, INTSINT utilise l'algorithme MOMEL qui permet de modéliser la courbe de f0 et fournit en sortie une séquence de points cibles pertinents linguistiquement. Le système de transcription INTSINT permet ensuite de coder symboliquement ces points cibles. INTSINT comprend un alphabet de 8 symboles : *Top*, *Middle* et *Bottom* sont définis globalement par rapport au registre de chaque locuteur, *Higher*, *Same* et *Lower* sont définis par rapport aux points précédents tout comme *Downstepped* et *Upstepped* lesquels concernent des changements de plus faible ampleur.

Les autres tires d'annotation sont le fruit de l'annotation manuelle : l'une d'entre elles concerne le phrasé prosodique des énoncés, c'est-à-dire la détermination de domaines ou d'unités prosodiques. Nous avons annoté les *unités intonatives/intonational phrase (IP)* et les *unités accentuelles/accental phrase (AP)*, les deux unités les plus communément admises pour le français¹⁵ (pour une revue voir [Jun&Fougeron02]). Nous avons ajouté une troisième catégorie (*EP, external phrase*) pour les cas ambigus ou impossibles à classer dans l'une ou l'autre des deux catégories. Ces cas peuvent être liés à la présence de marqueurs discursifs (tels que « quoi », « tu vois », etc.) comme l'avaient déjà souligné [Post&al.06].

Une autre tire d'annotation concerne les « contours intonatifs ». Chacune des catégories retenues est symbolisée par la forme du contour associée pour certains à sa fonction. Pour l'heure, le codage ayant été élaboré dans le cadre d'un travail parallèle relatif au contour de continuation [Portes&Bertrand05], seuls les contours montants ont été caractérisés fonctionnellement. Les symboles retenus sont les suivants : plat/*flat (f)*, montée mineure/*minor rising (mr)*; autres contours mineurs/*other minors (m0)*; descendant/*falling (F)*; montant-descendant/*rising-falling (RF1)*; montant-descendant depuis l'avant dernière syllabe/*rising-falling from penultimate (RF2)*; montée de continuation majeure/*rising major continuation (RMC)*; montée terminale/*rising terminal (RT)*; montée de question/*rising question (RQ)*; montée d'énumération/*rising enumerative (RENUM)*; descente d'énumération/*falling enumerative (FENUM)*.

3.3. Annotation morphosyntaxique

Le principe de l'annotation morphosyntaxique consiste à associer aux mots 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 le français, pour ne faire référence qu'à des systèmes facilement accessibles, on peut citer WinBrill, Cordial, ou encore l'étiqueteur développé par le LPL (intégré à la chaîne de traitement LPLSuite, [VanRullen05]). La technologie utilise les informations stochastiques apprises sur un corpus d'apprentissage pour proposer les étiquettes morphosyntaxiques les plus probables de l'énoncé. L'adaptation de l'étiqueteur du LPL pour enrichir un corpus oral avec des annotations

¹⁴ voir pour le détail Delais-Roussarie et coll., dans ce même numéro.

¹⁵ L'AP est le domaine de l'accent primaire en français. L'IP regroupe plusieurs AP et est définie par sa cohésion mélodique.

morphosyntaxiques dans un format donné (dans notre cas, sous la forme d'arcs représentés par des éléments XML) est en cours.

3.4. Annotation syntaxique

L'annotation d'informations syntaxiques reste une tâche complexe et difficilement automatisable. Il existe malgré tout un certain nombre d'analyseurs automatiques pouvant être utilisés au moins en tant que base pour la construction d'une annotation. On distingue pour cela deux types d'approche selon le niveau de description désiré. L'annotation la plus simple (le parenthésage) peut s'appuyer seulement sur des techniques d'analyse superficielle (ou encore *shallow parsing*). Mais il est également possible d'envisager des annotations plus fines, générées à l'aide d'analyseurs approfondis. Ces derniers, à la différence des analyseurs superficiels, permettent d'identifier non seulement les unités et leurs structures, mais également les relations syntaxiques qui les lient, en même temps que les fonctions grammaticales de ces différentes unités. Dans tous les cas, et cette remarque s'applique à tous les niveaux d'analyse présentés ici, la technique d'annotation et le formalisme utilisé sont indépendants du cadre théorique choisi pour la description syntaxique (par exemple HPSG, GP ou les grammaires de dépendance).

Les analyseurs superficiels, issus des travaux d'Abney [Abney91] ne fournissent que des informations au niveau des frontières des constituants, et ont prouvé leur efficacité pour l'analyse de corpus. Ils sont en particulier robustes et permettent de fournir des approximations de constituants syntaxiques intéressantes et pouvant servir de base à une description. Ce type d'analyse est généralement utilisé en tant que composant d'applications plus générales comme la recherche d'informations, les systèmes de dialogue ou les systèmes de synthèse de la parole. Plusieurs systèmes sont disponibles et notamment ceux qui ont été développés dans le cadre du LPL [VanRullen05] [VanRullen&al.05]. Si l'on se situe dans une perspective de représentation et de traitement de l'information sur la base de techniques symboliques, ce type de programme s'appuie sur un ensemble de règles qui déterminent les frontières gauches et droites des constituants en fonction du constituant courant et de diverses propriétés du mot lu. Il prend en entrée un texte étiqueté et désambiguïsé et se sert des catégories fonctionnelles comme frontières ouvrantes et fermantes d'un syntagme.

Quelques outils d'interrogation de corpus annotés syntaxiquement existent depuis peu suite aux efforts fournis pour la constitution de ressources françaises syntaxiquement annotées. Nous ne citons ici que les systèmes les plus évolués à savoir ceux qui dépassent les simples langages de requête basés sur des expressions régulières. [Christ94] a développé un outil d'interrogation appelé XKWIC qui permet d'extraire des passages d'un corpus en fonction d'expressions régulières exprimées sur les formes et les catégories. [Kallmeyer00] décrit un outil de requête des corpus syntaxiquement annotés qui permet de tenir compte des relations de dominance et de précédence.

3.5. Annotation sémantique

L'annotation se situe pour le moment au niveau du lexique et des relations entre les unités lexicales. Il s'agit de mettre à jour les éléments pertinents pour la construction du sens d'un discours en marquant à la fois les unités lexicales et les relations que ces unités entretiennent. La constitution d'une telle ressource est un premier pas vers la possible validation de formalisations de phénomènes sémantiques et discursifs proposés par Kamp et Asher par exemple (cf. [Kamp&Reyle93] et [Asher&Pustejovsky00]). Elle représente de plus une ressource non négligeable pour la création d'outils d'annotations sémantiques basés sur des méthodes numériques.

Une annotation sémantique qui se donne pour but la mise à jour de phénomènes lexicaux et interlexicaux devra comprendre au moins trois niveaux informationnels :

- un premier niveau d'annotation marquera les fonctions sémantiques telles que Agent, Patient, autour des prédicats verbaux ou nominaux. On pourra aussi marquer des informations lexicales plus complexes comme celles qui sont marquées par le lexique génératif initié par James Pustejovsky (cf. [Pustejovsky95]). Nous entendons par là le marquage précis de la contribution de la lexie (i.e. état, processus et transition) à la mise à jour de la structure événementielle de la phrase, ainsi que le marquage de la structure « qualia ». Cette structure permet de représenter un niveau d'information particulier en sémantique lexicale. Elle est plus précisément destinée à catégoriser les objets du monde. La structure qualia contient quatre rôles précisant les propriétés sémantiques de l'item. Ces rôles (*constitutif, formel, téléique et agentif*) précisent

respectivement la relation de l'objet à ses composants, les attributs caractéristiques de l'objet, la fonction de l'objet et les acteurs associés à l'objet.

- un deuxième niveau marquera des informations de type ontologique comme les notions de verbes de mouvement, ces annotations motivées linguistiquement permettent de rendre compte de phénomènes intéressants. Cet ensemble de traits sémantiques sera sélectionné sur la base des études linguistiques préalables qui ont prouvé leur pertinence dans plusieurs langues, à l'instar par exemple du *Generalised Upper Model* de Bateman 1992 et des travaux de Beth Levin (cf. [Levin93])

- enfin un troisième niveau marquera des relations entre les unités lexicales. Il s'agit donc de marquer à la fois les relations d'ordre hiérarchique comme les relations anaphoriques par exemple mais il s'agit aussi et surtout d'indiquer comment le sens de chaque unité dans l'énoncé est obtenu. Ceci pourra se faire en prenant en compte les interactions d'une unité avec les autres unités polysémiques de l'énoncé. Une unité sémantique ne se construisant qu'en contexte, il est nécessaire de marquer chaque contribution à la construction du sens.

3.6. Annotation pragmatique

3.6.1 Quel niveau pragmatique ?

Parler de niveau pragmatique nécessite de préciser ce que le terme recouvre. En effet, dans la littérature, il renvoie à des courants théoriques, des méthodologies, voire même des objets d'étude fort divers. Le terme recouvre en l'occurrence ici trois perspectives qui déterminent des niveaux de nature différente, à savoir par exemple les actes de langage, mais aussi les phénomènes d'ordre conversationnels liés à la construction des tours de parole et d'autres enfin relevant de la dimension énonciative des discours.

Un premier niveau d'annotation concerne donc ce qui a été construit sur le socle des avancées descriptives effectuées dans le domaine du dialogue orienté tâche. Bien que résolument restreinte, cette conception du dialogue est cependant un terrain d'investigation qui a su produire de véritables procédures d'analyse. Une piste marque les relations de discours selon Wolf et Gibson (2005) et Hobbs (1979). Une autre piste marque les événements langagiers (réalisés par les marques d'acceptation et les assertions) qui affectent les croyances des interlocuteurs. Ceci permet de noter le stock d'informations présent dans le contexte du dialogue qui se construit. Cette proposition est reprise des schémas d'annotation proposés dans le cadre des projets DAMSL [Core&Allen97] et TRINDI [Traum&al.99].

La figure 4 fournit un exemple détaillé de notre annotation, illustré par la présentation des références indexées entre les pistes annotées. Le premier élément, qui a une référence temporelle, sert d'index pour les autres éléments qui eux mêmes prennent l'index de leur position dans l'élément structurel qui les englobe.

```
<el index=«32» start=«5.8588» end=«6.0908»>
<attribute name=«SpellSp1»>graphemes</attribute> </el> ...
<el index=«26» start=«32» end=«32»>
<attribute name=«CommonNoun»>Common</attribute>
<attribute name=«Agreement»>4</attribute> </el> ...
<el index=«15» start=«31» end=«32»>
<attribute name=«NounPhrase»>Standard</attribute>
<attribute name=«Agreement»>4</attribute> </el> ...
```

Figure 4 : exemple d'annotation pragmatique

D'autres éléments typiques de l'oral ont également été annotés. Une première tire d'annotation concerne les marqueurs discursifs du type « quoi, voilà, tu vois, tu sais, enfin », sur lequel un premier travail a été réalisé par [Bertrand&Chanet05]. Deux autres tires d'annotation concernent les phénomènes d'écoute. Suite à [Laforest92], nous distinguons entre les signaux d'écoute simples communément appelés signaux *backchannels* et les signaux complexes tels que certaines répétitions, reformulations, métaquestions, compléments, etc. Une tire est consacrée aux catégories formelles (mhm, ouais, ah bon, ah ouais, d'accord, etc), une autre aux catégories fonctionnelles des seuls *backchannels*, étant entendu que la catégorisation en

complexes comporte déjà un aspect fonctionnel. Les *backchannels* sont donc classés en *continuer* (écoute minimale, prendre note) ou *assessment* (implique une évaluation, un jugement), selon la distinction de [Schegloff82], et une troisième catégorie pour les cas ambigus¹⁶.

La dernière tire d'annotation concerne le type d'unités utilisée dans le cadre de la CA pour rendre compte des tours de parole : les *unités de constructions de tours/turn-constructional units* (TCU). Les TCU sont des unités considérées comme potentiellement « complètes » du point de vue syntaxique, prosodique et pragmatique [Ford&Thompson96]. La fin d'un TCU constitue un lieu de complétion potentiel qui renvoie à ce qu'on appelle une *place transitionnelle/transition-relevance place* (TRP). Selon [Selting98 : 40], "les TCU sont donc les plus petites unités linguistiquement complètes pertinentes au niveau interactionnel". Mais ils peuvent être final (complet des trois points de vue) ou non final. Le TCU final représente donc un tour constitué d'une seule unité tandis que le TCU non final est l'un des composants incomplets (des points de vue sémantique ou pragmatique notamment) d'un tour complexe défini en termes d'activité discursive (constructions causales, séquence narrative, etc.). Un TCU final s'achève dans une TRP. Le *CID* est annoté en TCU final et non final.

3.6.2. Les difficultés de l'annotation pragmatique

Si, contrairement à l'annotation phonétique, l'alignement entre l'activité discursive/énonciative notée et le signal n'est pas d'une nécessité absolue, l'annotation de tels phénomènes langagiers pose cependant un certain nombre de problèmes.

Le premier concerne le terme même « d'annotation ». Ainsi, si un connecteur ou un signal *backchannel* peuvent facilement être repérés et annotés en tant que tels, d'autres activités plus floues comme les « modulations du discours » (atténuation) nécessitent une analyse préalable pour être reconnues comme telles et posent alors partiellement la question de la légitimité d'une annotation *après coup*. Plus globalement, et pour tous les niveaux, ce point renvoie à la problématique de l'annotation elle-même et pose la question de la relation entre interprétation et annotation. Selon nous, il n'y a pas d'annotation sans interprétation préalable mais on ne peut nier également l'importance d'effectuer un constant aller-retour entre les deux, et l'influence qu'aura donc l'annotation elle-même sur l'interprétation.

Le deuxième problème concerne la prise en compte de la temporalité, donc de frontière du phénomène observé. Nous savons par exemple que l'humour peut revêtir, dans une conversation, de nombreuses facettes allant du simple jeu de mots (facilement délimitable) à une longue séquence co-construite par les interactants. Dès lors, où commence vraiment cet humour et où finit-il ? Comment arriver à le délimiter ? La question se pose d'autant plus que parfois, un seul locuteur produit de l'humour sans pour autant qu'il soit suivi dans ce mode de communication ludique par l'interlocuteur. Deux modes de communication (sérieux / non sérieux) se confrontent alors et rendent les frontières très floues.

Enfin, et de manière plus générale, la fluidité même des phénomènes discursifs et énonciatifs et leur dimension nécessairement contextuelle (prise en compte de la situation d'interlocution, des implicites partagés par les locuteurs, des réactions de l'interlocuteur...) rendent la question de l'annotation problématique. Il s'avère donc indispensable de travailler sur une source la plus complète possible (audio-vidéo). Une analyse multimodale pour ce niveau s'avère donc d'une importance capitale pour au moins deux raisons : la première concerne l'interprétation des phénomènes étudiés qui ne peut qu'être améliorée par la prise en compte de tous les autres niveaux participant à la construction du sens, et la seconde pour affiner et enrichir nos propres annotations.

3.7. Annotation Mimo-Gestuelle

L'annotation mimo-gestuelle du corpus est en cours de développement. Nous utilisons le logiciel *ANVIL* (développé au DFKI par M. Kipp, 2003-2006). Pour fonctionner, *ANVIL* utilise un fichier XML de spécifications (spec), dans lequel figurent toutes les étiquettes qui seront utilisées lors de l'annotation. Le fichier *spec*, fourni avec le logiciel et utilisé par défaut, se base sur le standard [McNeill92], étendu par [Kita03] pour la notation des gestes manuels. Ces deux standards ont eux-mêmes été complétés par [Allwood&al.05] qui ont pris en compte non seulement les gestes manuels tels qu'ils ont été décrits par

¹⁶ Cette annotation a été mise au point et réalisée sur de nombreux extraits dans le cadre d'un mémoire de M2, d'Amandine Tenet, sous la direction de R. Bertrand.

McNeill (*op. cit.*) et Kita (*op. cit.*), mais également les expressions faciales et les mouvements de tête ainsi que la direction du regard. Nous avons donc repris ce code d'annotation et l'avons également complété¹⁷ et modifié de la façon suivante afin qu'il réponde à nos besoins :

a) Le code MUMIN mêle des descriptions qui ne relèvent pas de la même modalité : la description mimogestuelle est construite sur l'interaction et la modalité du discours. Par exemple, les auteurs notent les gestes produits par le locuteur ou l'auditeur, or, les tours de parole relèvent d'une analyse de l'interaction, et non pas d'une analyse gestuelle. De même, comme ils travaillent en particulier sur le *feedback* de l'interlocuteur, ils notent les gestes produits par le locuteur ou l'auditeur selon que celui-ci donne un *feedback* ou le sollicite. Dans notre propre transcription de la gestualité, nous préférons séparer les différentes modalités. Les informations ne seront pas perdues cependant puisqu'il sera possible par la suite non seulement d'afficher les transcriptions de telle et telle modalité simultanément dans le logiciel pour un meilleur visuel de la transcription, mais aussi de faire des requêtes informatiques à travers plusieurs fichiers d'annotation. Ainsi, nous avons préféré transcrire la gestualité d'un seul sujet dans un fichier d'annotation (que celui-ci soit locuteur ou auditeur) puisque l'information sur les tours de parole sera rendue dans une autre modalité (pragmatique), qu'il sera donc toujours possible de la retrouver, et qu'il est parfaitement possible d'adapter plus tard l'affichage des différentes partitions d'*ANVIL* pour des études ponctuelles.

b) Nous avons complété le code MUMIN, déjà relativement élaboré, en y ajoutant l'annotation des mouvements du buste qui n'étaient pas du tout pris en compte dans l'annotation d'Allwood [Allwood&al.05]. Nous avons également précisé la direction gauche/droite pour les mouvements latéraux, ainsi que la nature de la référence pour les gestes déictiques¹⁸, vers un objet de discours abstrait ou concret.

En résumé, les gestes que nous avons décidé d'annoter sont donnés dans le tableau suivant, sans entrer dans le détail des différentes valeurs attribuées aux gestes (comme par exemple main gauche/main droite/deux mains qui vient définir plus précisément le type sémiotique de geste – les différents types sémiotiques étant donnés dans la note 13) :

Tête/Visage	Mains	Corps
Expression faciale	Symétrie/asymétrie du geste	Mouvements 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 1 : nomenclature des gestes annotés dans le CID

Dans le tableau 1 ci-dessus, certaines catégories sont claires pour les non-gestualistes, d'autres doivent être décrites plus clairement : les phases gestuelles font référence aux travaux de Kendon, repris par McNeill [92], 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, nous avons ajouté, en suivant ici le travail de Loehr [04], 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 point de contact est utilisé pour les adapteurs, qui impliquent un contact entre la main et une partie du corps, soit du locuteur lui-même, soit de l'interactant.

¹⁷ Avec la collaboration de Claire Maury-Rouan (LPL)

¹⁸ Nous prenons ici l'exemple des gestes déictiques car ce sont les plus connus, mais la nomenclature compte aussi des gestes iconiques, métaphoriques, des adapteurs, des battements (*beats*) et des gestes désordonnés (*butterworths*), cf. [McNeill92]. 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.

À l'heure actuelle, nous avons créé le fichier de spécification selon ces standards. Nous procédons maintenant à la description détaillée de nos étiquettes d'annotation afin de pouvoir établir un nouveau code d'annotation de la gestualité. Nous allons également procéder à l'annotation à proprement parler de la gestualité à partir des fichiers vidéo.

À titre d'exemple, voici un extrait de notre fichier de spécification. Il indique que la tire *Eyebrows* (mouvements des sourcils) fait partie du groupe d'annotation des mouvements faciaux (*group name=Face*) et que les valeurs que l'on peut rencontrer pour les mouvements des sourcils sont *frowning* (froncement des sourcils) ou *raising* (haussement des sourcils). Nous avons également introduit une valeur *other* (autre) pour le cas rare où un locuteur hausserait un seul sourcil par exemple. Les autres tires d'annotations sont définies de la même manière.

<pre><?xml version="1.0" encoding="ISO-8859-1"?> <annotation-spec> - <head> - <valuetype-def> - <valueset name="EyebrowsType"> <value-el color="#9df4a9">Frowning</value-el> <value-el color="#f1f07a">Raising</value-el> <value-el color="#f5ce16">Other</value-el> </valueset> </valuetype-def> </head> - <body> - <group name="Face"> - <track-spec name="Eyebrows" type="primary"> <attribute name="Eyebrows" valuetype="EyebrowsType" display="true" /> </track-spec> </group> </body> </annotation-spec></pre>	<p>Cette première ligne décrit le fichier comme un fichier xml Nom de l'annotation La première partie du fichier donne toutes les valeurs possibles dans la tire d'annotation, ici <i>frowning</i>, <i>raising</i>, <i>other</i></p> <p>La deuxième partie du fichier décrit la tire d'annotation <i>eyebrows</i>, et indique entre autres que cette tire est primaire (elle ne dépend pas d'une autre tire), mais fait partie du groupe <i>face</i>, d'annotation des mouvements faciaux. La syntaxe XML nécessite que toutes les parenthèses soient refermées.</p>
----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

La création du fichier nous a permis notamment de repenser différemment la structure de l'annotation gestuelle : ainsi, plutôt que de créer plusieurs étiquettes spécifiques comme :

<i>Gaze</i> >	<i>Sideways</i> >	<i>left / right</i>
<i>Head</i> >	<i>Side turn</i> >	<i>left / right</i>
	<i>Single tilt</i> >	<i>left / right</i>
<i>Trunk</i> >	<i>Sideways</i> >	<i>left / right</i>
<i>Hands</i> >	<i>Single hand</i> >	<i>left / right</i>

Nous entrons une seule fois la valeur *left/right* qui sera applicable aux mouvements de toutes les parties du corps. Si le mouvement ne fait pas intervenir la dimension *left/right*, comme par exemple *eyebrow raising / frowning*, c'est la valeur par défaut *none* qui est appliquée à l'étiquette *left/right*. De la même façon, nous avons créé une étiquette *concrete/abstract* qui ne s'applique qu'aux gestes déictiques et à la direction du regard. La dimension *contact point* quant à elle ne concerne que les adaptateurs, mais est néanmoins présente et spécifiée *none* pour tous les autres gestes.

Une fois les annotations entrées dans ANVIL, le logiciel génère un fichier XML qui indique pour chaque annotation le rang, le temps de début et le temps de fin. Les valeurs par défaut ne sont pas comptabilisées dans le fichier final d'annotation et ne viennent donc pas alourdir ce fichier en y ajoutant des informations non pertinentes pour tel ou tel type de geste.

4. Schéma d'annotation multimodal

Au-delà de l'annotation de la gestualité, *ANVIL* nous a permis de regrouper les annotations des autres niveaux linguistiques décrits plus haut. Ce logiciel nous permet non seulement d'importer des annotations issues de Praat par exemple mais également de les modifier à la condition d'avoir préalablement spécifié les étiquettes utilisées dans le fichier de spécification. Ceci nous a obligés à penser la hiérarchie des étiquettes au sein de chaque niveau d'annotation : les annotations réalisées par exemple de manière linéaire sous Praat ont été hiérarchisées pour répondre à la structure du format XML -format d'entrée et de sortie d'*ANVIL*-. En ce qui concerne les autres niveaux dont les annotations n'ont pas été réalisées sous Praat, elles sont au format XML mais nous avons néanmoins également spécifié les étiquettes dans un fichier *spec*.

Précisons enfin que parmi les fonctionnalités offertes par *ANVIL*, il est possible de créer autant de fichiers *spec* d'entrée que l'on souhaite, utilisant tel ou tel niveau d'annotation plutôt que tel autre ainsi que tel ou tel jeu d'étiquettes, de même qu'il est possible de créer autant de fichiers d'annotation de sortie que l'on souhaite. Il est également possible d'intervenir directement sur ces fichiers de sortie puisqu'ils sont au format XML.

5. Synchronisation et exploitation des données

L'intérêt d'adopter une approche intégrée pour l'annotation de corpus (en particulier multimodaux) réside dans la possibilité d'envisager des utilisations très variées de ces informations. Il devient en particulier possible d'envisager des requêtes de haut niveau, faisant intervenir différents domaines d'annotation et permettant d'étudier de façon plus systématique les interactions pouvant exister entre eux. Il y a cependant une condition préliminaire à ce type d'utilisation : la nécessité de relier entre eux chacun des domaines.

La synchronisation temporelle constitue une réponse naturelle dès lors que les objets annotés sont clairement identifiés et surtout sont délimités dans un intervalle. Il est ainsi possible d'identifier clairement dans le déroulement temporel du signal des frontières de début et de fin. C'est le cas des événements phonétiques (e.g. les phonèmes), prosodiques (e.g. les pauses) ou mimo-gestuels. Les niveaux non connectés directement au signal acoustique ou vidéo peuvent eux aussi être mis en relation avec les autres domaines, de façon indirecte. Le niveau morpho-syntaxique par exemple est connecté au niveau phonétique grâce à un alignement graphème-phonème qui permet pour un mot d'indiquer par transitivité sa position dans le signal.

Dans certains cas cependant, les frontières peuvent être difficiles à marquer : il peut être délicat d'identifier précisément les frontières d'un contour intonatif. De façon encore plus problématique, certains objets peuvent ne pas être véritablement réalisés physiquement. C'est tout d'abord le cas d'objets dont la production est difficile à identifier. Certains phonèmes peuvent par exemple par assimilation n'être quasiment pas détectables dans le signal et pourtant présents dans la transcription. Ce problème s'accroît au niveau morpho-syntaxique ou des mots peuvent pour la même raison n'être pas véritablement présents dans le signal (la préposition « de » peut quasiment disparaître si elle est suivie d'un mot commençant par une occlusive dentale comme dans « armée de terre »). Enfin, il existe des ensembles d'objets relevant de domaines particuliers qui ne sont pas reliés au signal. C'est par exemple le cas du domaine sémantique qui est notamment constitué d'un ensemble de référents constituant « l'univers du discours » sans que ceux-ci ne correspondent nécessairement à des objets présents dans le signal acoustique. Cette situation est typique des corpus multimodaux dans lesquels certains des référents peuvent être constitués par des objets présents dans la scène. Il est dans ce cas impossible de relier un tel objet avec une quelconque position dans le signal.

La question de la synchronisation est donc un problème en soi : comment mettre en relation des objets appartenant à des domaines différents sans base de référence unique ? Nous avons proposé dans [Blache03] une solution reposant sur le positionnement de chaque objet à l'aide d'un ancrage complexe. Le principe consiste à indiquer pour chaque objet plusieurs repérages lorsque c'est possible. Bien entendu, la plupart des objets directement liés au signal sont ancrés par une position temporelle. De même, comme indiqué plus haut, certains objets d'autres domaines peuvent, par transitivité, être également ancrés sur le signal temporel. C'est par exemple le cas des mots, des catégories morpho-syntaxiques. Mais dans ce cas,

un autre type d'ancrage peut être proposé ; la position de l'objet dans la chaîne (i.e. le rang du mot dans la séquence ou la phrase). Ce type d'ancrage relève plus traditionnellement du traitement de matériel écrit, mais nous proposons de l'intégrer dans notre système. Enfin, de façon à intégrer à l'ensemble les informations sémantiques, nous proposons l'utilisation d'une indexation des éléments de l'univers du discours. Au total, le repérage de chaque élément peut se faire *via* une ancre complexe, renseignant tout ou partie des trois composants (situation temporelle dans le signal, position dans la chaîne, indice du contexte). Cette ancre est représentée par la matrice suivante :

$$\text{Ancre} \begin{bmatrix} \text{Temporelle} <i, j> \\ \text{Position} <k, l> \\ \text{Contexte} c \end{bmatrix}$$

Un objet référentiel spécifié par un mot ayant été prononcé verra son ancre renseignée pour les trois attributs. En revanche, un objet présent dans une scène ne sera ancré que par l'indice contextuel. La synchronisation entre les différents domaines se fait par clôture transitive de toutes les ancres.

Cet ancrage réalisé, il devient alors possible d'interroger les corpus en exploitant les différents niveaux d'annotation. Il s'agit dans ce cas d'effectuer des requêtes parcourant simultanément plusieurs domaines tout en synchronisant les recherches *via* le système d'ancrage complexe. Les exemples de requêtes suivants illustrent le fonctionnement du système :

- **Q0 Rechercher les pauses intervenant au milieu d'une unité syntaxique**
 - ✓ Les domaines concernés sont la syntaxe et la prosodie. La recherche consistera à identifier des objets d'un type prosodique particulier dont les frontières sont incluses dans celles d'un objet appartenant à un autre domaine, la syntaxe. Dans ce cas, les ancrages temporels et de position permettent l'identification des objets.
- **Q1 Rechercher les gestes déictiques associés à un pronom**
 - ✓ Cette requête porte simultanément sur les domaines gestuel et morpho-syntaxique. Elle consiste à identifier dans un domaine (gestuel) tous les objets d'un type particulier (déictique) et voir le recouvrement avec les objets de type pronom ayant un recouvrement de portée (dont les intervalles sont inclus dans ceux du geste).
- **Q3 Rechercher les SN agents extraits avec contour montant**
 - ✓ Dans ce cas, les objets sélectionnés doivent avoir une caractéristique syntaxique (ils sont extraits), une propriété sémantique (ce sont des agents) et appartenir à un contour intonatif particulier (montant). Les ancrages de position et temporels seront ici utiles.
- **Q4 Rechercher les backchannels de L1 chevauchant un référentiel sujet de L2**
 - ✓ Cette requête vise l'identification de marqueurs discursifs particuliers, les *backchannels*, produits par un locuteur, et qui apparaissent en chevauchement avec des objets d'un certain type (SN référentiel sujet) produits en même temps par l'interlocuteur.

La possibilité d'effectuer de telles requêtes permet donc d'envisager une exploitation très complète des corpus multimodaux. Nous aurons en effet la possibilité d'analyser les phénomènes d'interaction entre les différents domaines de façon systématique. Ce type d'investigation sur des corpus de grande taille n'a à ce jour pas d'équivalent, aucun corpus proposant des annotations complètes des différents domaines linguistiques n'étant réellement disponible. Notre approche offre une solution efficace à la représentation et la synchronisation d'annotations provenant de différents domaines. Il s'agit donc d'une contribution importante à l'analyse de données multimodales.

Conventions de transcription (inspirées du GARS)

orthographique	transcription orthographique «du dictionnaire», sont donc incluses les transcriptions canoniques des <i>beu, hum, ah</i> , etc.
pseudo-orthographique	approximation (ortho)graphique de la réalisation acoustique ex : /pneu/ prononcé [p 2 n 2] est transcrit [peuneu] /s'il vous plaît/ est transcrit [s'i l vous ,siou] plaît
variante graphique	{variante 1, ..., variante n} {il chante, ils chantent}
«phonèmes(s)» non réalisés	(ortho) je t(e) le dis
prononciation particulière	[ortho, pseudo_ortho] [je sais, chai]
pause silencieuse	+
liaison non standard	= pseudo-ortho = =z=
liaison standard absente	deux # échecs
plusieurs transcriptions possibles	<ortho1,ortho2...> <oui, ouais, ouah>
amorces pseudo_ortho	-
séquence incompréhensible	*
rires	@
nom propre	\$ortho, PTS /\$ \$Senderens, P /\$
titre	«...»
discours rapporté	§ ... §
ex	il dit § moi + j'ai dit ça § le menteur

Références

[Abney91] Abney, S. (1991) "Parsing by chunks", in Berwick, R., Abney, S., Tenny, C. (eds) *Principle-based parsing* (pp. 257-278), Kluwer Academic Publishers, Dordrecht.

[Abry&al.85] Abry, C., Benoît, C., Boë, L.J., Sock, R. (1985) "Un choix d'événement pour l'organisation temporelle du signal de parole", *Actes des 14èmes Journées d'Études sur la Parole (JEP)*, Paris, 133-137.

[Allwood&al.05] Allwood, J., Cerrato, L., Dybkjaer, L., Jokinen, K., Navarretta, C. & Paggio, P. (2005) "The MUMIN Multimodal Coding Scheme", *NorFA yearbook 2005*.
<http://www.ling.gu.se/~jens/publications/B%20files/B70.pdf>

[Allwood&al.06] Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. & Paggio, P. (2006) "A Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing Phenomena", Accepted paper to the LREC 2006 Workshop on «Multimodal Corpora», Genoa, Italy, 27th May.
http://www.speech.kth.se/~loce/papers/lrec_annot_mumin_v5.pdf

[Anderson&al.91] 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. (1991) "The HCRC map task corpus". *Language and Speech* 34, 351-366.

[Asher&Pustejovsky00] Asher N. & Pustejovsky J. (2000) "The metaphysics of words in context", ms., submitted to *Journal of Logic, Language and Information*.

ATLAS: Laprun C., Fiscus J. G., Garofolo J. & Pajot S. (2002) "A practical introduction to Atlas". In *actes de LREC 2002*.

[Beckman&Ayers97] Beckman M. & Ayers, G. (1997) Guidelines for ToBI Labelling.
http://ling.ohio-state.edu/phonetics/E_ToBI

[Beckman&al.05] Beckman, M. E.; Hirschberg, J.; Shattuck-Hufnagel, S. (2005) "The original ToBI system and the evolution of the ToBI framework". In *Prosodic Typology: The phonology of Intonation and Phrasing*, S.-A. Jun (ed.). Oxford: Oxford University Press.

- [Bertrand&Chanet05] Bertrand, R.; Chanet, C. (2005) "Fonctions pragmatiques et prosodie de *enfin* en français spontané". *Revue de Sémantique et Pragmatique*, no. 17, p. 41-68.
- [Blache03] Philippe Blache (2003) "Meta-level constraints for linguistic domain interaction", in proceedings of *International Workshop on Parsing Technologies (IWPT-03)*.
- [Blanche-Benveniste&Jeanjean87] Blanche-Benveniste C & Jeanjean C. (1987) *Le français parlé, Transcription et édition*. Paris : Didier-Erudition/ InaLF, 2^e éd.
- [Boersma&Weenink05] Boersma P. & Weenink D. (2005) Praat : doing phonetics by computer (version 4.3.14). Logiciel téléchargé le 26 mai 2005; <http://www.praat.org/>
- [Brun&al.04] Brun A., Cerisara C., Fohr D., Illina I., Langlois D., Mella O., Smaili K. (2004) "Ants : le système de transcription automatique du Loria". In *Actes des Journées d'Etudes sur la Parole*.
- [Core&Allen97] Core M. G. & Allen J. F. (1997) "Coding dialogs with the damsl annotation scheme", In *Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*, Boston, MA.
- DAMSL: voir [Core&Allen97]
- [DiCristo&DiCristo01] Di Cristo A. & Di Cristo P. (2001) "Syntaix, une approche métrique-autosegmentale de la prosodie", *TAL*, 42(1), 69–111.
- [DiCristo&al.04] Di Cristo, A. ; Auran, C. ; Bertrand, R. ; Chanet, C. ; Portes, C., Regnier, A. (2004) "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, vol. 28, p. 27-84.
- [Farnetani97] Farnetani E. (1997) "Coarticulation and connected speech", *The Handbook of Phonetic Sciences*, Hardcastle W.J., Laver J., eds., Blackwell, Oxford, GB, 371-404.
- [Ford&Thompson96] Ford C. E. & Thompson S. A. (1996) "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), 134-184, Cambridge UP.
- [Grabe&al.01] Grabe E., Post B., Nolan F. (2001) *Intonational Variation in English. The IViE Corpus on CD-Rom*. Linguistics, Cambridge.
- [Hawkins&Nguyen04] Hawkins, S.; Nguyen, N. (2004) "Influence of syllable-coda voicing on the acoustic properties of syllable-onset /l/ in English", *Journal of Phonetics*, vol. 32, no. 2, 199-231.
- [Hirst&DiCristo98] Hirst, D. & Di Cristo, A. (1998) *Intonation System*, Cambridge University Press.
- [Hirst&al.00] Hirst D., Di Cristo A. & Espesser R. (2000) *Prosody : Theory and Experiment*, chapter Levels of description and levels of representation in the analysis of intonation, p. 51-87. Kluwer : Dordrecht, Pays-Bas.
- [Hobbs79]. Hobbs J.R. (1979) Coherence and coreference. *Cognitive Science*, 3(1), 67-90.
- [Jun&Fougeron02] Jun, S.-A. and Fougeron, C. (2002). "Realizations of accentual phrase in French intonation", *Probus* 14. 147-172.
- [Kallmeyer00] Kallmeyer L. (2000) "A query tool for syntactically annotated corpora", in proceedings of ACL 2000.

[Kamp&Reyle93] Kamp H. and U. Reyle (1993) *From Discourse to Logic: Introduction to Model theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, Kluwer, Dordrecht.

[Kipp03-06] Kipp, M., 2003-2006. *Anvil 4.0. Annotation of Video and Spoken Language*. <http://www.dfki.de/~kipp/anvil>

[Kita03] Kita, S. (2003) "Interplay of gaze, hand, torso orientation and language in pointing", In S. Kita (Ed.), *Pointing: where language, culture, and cognition meet* (pp.307-328). Mahwah, NJ: Lawrence Erlbaum.

[Koiso&al.98] Koiso H., Horiuchi Y., Ichikawa A. & Den Y. (1998) "An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs", *Language and Speech*, 41, 295–321.

[Laforest92] Laforest, M. (1992) "Le back-channel en situation d'entrevue", In *Recherches Sociolinguistiques*, 2. Québec : Université Laval.

[Levin93] Levin B. (1993) *English Verb Classes and Verb Alternations: A Preliminary Investigation*. University of Chicago Press.

MATE: (<http://mate.nis.sdu.dk/>, cf. [Dybkjær98])
(<http://www.nist.gov/speech/atlas/overview.html>, cf. [Laprun02])

Dybkjaer L., Bernsen N., Dynkjaerand H., Mckelvie & Mengel A. (1998) *The MATE Markup Framework*. Rapport interne, MATE Deliverable D1.2.

[Loehr04] Loehr, D. (2004) "Gesture and Intonation." Doctoral dissertation (Thesis advisor: E. C. Zsiga). Georgetown University.

[McClave01] McClave E. Z. (2001) "The relationship between spontaneous gestures of the hearing and American Sign Language", *Gesture* 1.1, 51–72.

[McNeill92] McNeill, D. (1992) *Hand and Mind. What Gestures Reveal about Thought*. Chicago and London: The University of Chicago Press.

[Meunier94] Meunier, C. (1994) "Les groupes de consonnes: problématique de la segmentation et variabilité acoustique", *Thèse de l'Université de Provence (Aix-Marseille I)*, Présentée le 7 mars 1994.

MUMIN: A Nordic Network for MultiModal INterfaces. <http://www.cst.dk/mumin/>
Voir [Allwood&al.05] et [Allwood&al.06]

[Nguyen&al.04] Nguyen, N., Fagyal, Z., Cole, J. (2004) "Perceptual relevance of long-domain phonetic dependencies", in *Actes des Journées d'Études Linguistiques (JEL)*, 4, mai 2004, Nantes, France.

NITE: Natural Interactivity Tools Engineering. <http://nite.nis.sdu.dk/>
Carletta J., Heid U. & Kilgour J. (à paraître). "The nite xml toolkit : data model and query", *Language Resources and Evaluation Journal*.

[Portes&Bertrand05] Portes, C.; Bertrand, R. (2005) "Some cues about the interactional value of the 'continuation' contour in French". *Actes Discours et Prosodie comme Interface Complexe (IDP) Cederom* (14 pages).

[Post&al.06] Post B., Delais Roussarie E., Simon A.C. (2006) "IVTS, un système de transcription pour la variation prosodique", *Bulletin n° 6 PFC*, coordonné par A.C. Simon, G. Caelen-Haumont et C. Pagliano, Edition Nicole Serna, 51-68.

[Pustejovsky95] Pustejovsky J. (1995) *The Generative Lexicon*, MIT Press.

[Rossi90] Rossi, M. (1990) "Segmentation automatique de la parole: pourquoi? Quels segments? ", *Traitement du Signal*, vol. 7, n° 4, 315-326.

[Sacks&al.74] Sacks H., Schegloff E.A. & Jefferson G. (1974) "A simplest systematics for the organization of turn-taking for conversation", *Language* 50, 696-735.

[Schegloff82] Schegloff, E.A. (1982). "Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences", In D. Tannen (Ed.), *Analyzing discourse: Text and talk* (pp. 71-93). Washington, DC: Georgetown University Press.

[Selting98] Selting M. (1998), "TCUs and TRPs: the construction of 'units' in conversational talk", In *InLiSt (Interaction and Linguistic Structures)*,4, 1-48.

[Tenet06] Tenet A. (2006) "Phénomènes d'écoute en français : Une étude prosodique du contexte d'apparition des régulateurs à travers l'analyse des contours intonatifs", *Mémoire de Master II*, Mention Sciences du Langage, Parcours phonétique, sous la direction de R. Bertrand, Université de Provence.

[Vanrullen&al.05] Vanrullen T., Blache P. , Portes C. , Rauzy S. , Maeyhieux J.-F., Guenot M.-L., Balfourier J.-M. , Bellengier E. (2005) "Une plateforme pour l'acquisition, la maintenance et la validation de ressources lexicales", in *Actes de TALN'05*.

[VanRullen05] Van Rullen, T. (2005) "Vers une analyse syntaxique à granularité variable", *Thèse de Doctorat*, Université Aix-Marseille I, Décembre 2005.

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

[Wolf&Gibson05] Wolf F. & Gibson T. (2005) "Representing discourse coherence: A corpus-based analysis", *Computational Linguistics*, 31(2), 249-287.