À propos de deux bases de données de parole publiées récemment: compte-rendu de “API - Archivio del Parlato Italiano” et de “C-ORAL-ROM”[1]

Antonio Romano

Parmi les nombreuses contributions des dernières décennies, deux ouvrages importants, publiés récemment, ont proposé de nouvelles données linguistiques d’italien parlé (l’une d’entre elles conjointement à celles d’autres langues romanes).

Les deux publications se présentent très différemment dans leurs formats et dans les conditions de diffusion. De même, les *corpus* qu’elles proposent sont issus d’un travail de récolte, d’organisation et d’évaluation qui a été mené dans des centres de recherche d’excellence dans des domaines relativement différents.[2]

Le but de ce compte-rendu est d’offrir un résumé d’informations qui puisse valoriser ces deux publications et en élargir la diffusion, en stimulant l’intérêt d’un public d’utilisateurs aussi vaste que possible. Dans ce sens, je vais présenter les deux bases de données en soulignant les qualités des données et des outils fournis qui les accompagnent. Je profiterai de cette occasion pour signaler au lecteur (et à l’utilisateur final des données) quelques uns des principaux défauts présents dans les produits, en émettant un jugement personnel concernant la confiance que l’on peut accorder aux matériaux disponibles lorsqu’on les soumet aux traitements et aux analyses souhaités.[3]

La première base de données (*BD*), dans l’ordre chronologique de publication, est celle de l’*API – Archivio del Parlato Italiano*, issue d’un projet de recherche financé par le Ministère italien de l’Université et de la Recherche Scientifique et Technologique (*MURST*). La base de données est publiée sous-forme de *DVD*, avec des documents de présentation et des commentaires. Elle est distribuée gratuitement par le centre de recherche *CIRASS* de Naples (v. note 1).

Il s’agit des résultats des recherches menées par 7 équipes de divers instituts universitaires italiens coordonnées par Federico Albano Leoni au Centre de Recherche *CIRASS*. La recherche a été achevée en novembre 2001 et le corpus mis au point (basé sur des données recueillies dans le cadre du projet *AVIP – Archivio delle Varietà d’Italiano Parlato*) ainsi que les outils développés pour sa consultation ont été diffusés peu de temps après (2002).

Rivista di Linguistica 19.2 (2007), pp. 367-390 *(ricevuto nel giugno 2007)*

Cette *BD* concerne la parole spontanée, analysée et étiquetée au niveau segmental aussi bien qu'au niveau suprasegmental. Elle se présente sous forme d'un ensemble organisé de fichiers dans un format.wav ou dans un codage similaire (fichiers.sw, *mono*, *PCM* à 16 bits, *signed*, 22050 Hz) accessibles pour des traitements ultérieurs de la part de l'utilisateur. Aux fichiers sonores d'une sélection de dialogues (voir ci-après) sont associés des fichiers de texte en transcription orthographique conventionnelle et des fichiers d'étiquettes.

Les variétés d'italien prises en compte sont celles de Pise, Naples et Bari (et, dans une moindre mesure, de Florence et de Brindisi), à partir de 18 dialogues semi-spontanés de type *map-task*, chacun avec deux participants adultes, ainsi que 4 dialogues avec des enfants normo-entendants et 3 avec des enfants ayant des légers problèmes auditifs.[4]

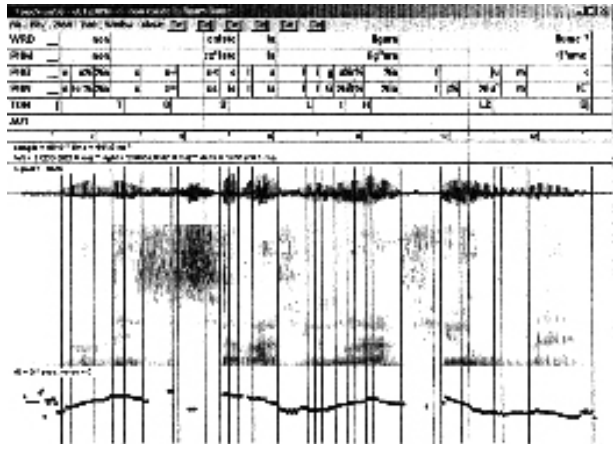
Dans la base de données figurent des dialogues liés à 4 cartes (A, B, C, D) pour les diverses localités. À chaque dialogue sont associées deux séries d'enregistrements de parole lue concernant les mots prévus dans les cartes (*toponymes*).

Au total, pour ce type de données linguistiques, on peut disposer de 18 dialogues pour plus de 4 heures de parole obtenues en conditions d'enregistrement d'excellente qualité (conditions de laboratoire). Une division en tours de parole est disponible (seulement pour une partie de ces dialogues, 14) et des enregistrements supplémentaires sont donnés pour les listes de *toponymes* d'une autre sélection de dialogues (16).[5]

Concernant la qualité linguistique de ces données, l'impression générale qui se dégage est qu'il s'agit de productions assez représentatives des variétés choisies, dans lesquelles on retrouve une grande partie des phénomènes phonétiques typiques (certains très connus d'autres moins) de l'italien parlé dans ces localités. Cependant, la sélection restreinte de locuteurs (et de leurs domaines d'activité) et le nombre limité de contextes et de styles de production ne permettent pas d'avoir une vue d'ensemble satisfaisante de la variation socio- et/ou géo-linguistique dans ces communautés. Ce n'était d'ailleurs pas parmi les objectifs du projet de recherche d'où est issue la *BD* en question, et le souligner ici permet au lecteur d'être mis en garde à ce propos.

Les fichiers d'annotation associés aux fichiers sonores se basent sur des protocoles les plus connus, mais qui ont été adaptés aux contenus spécifiques de cette collection. Les transcriptions phonétiques s'inspirent ainsi de l'ensemble des conventions *SAMPA* (*Speech Assessment Methods Phonetic Alphabet*) et les symboles utilisés sont tirés de cet alphabet.

Parmi les documents et les outils fournis, je citerai surtout les logiciels *QueryGenerator* (© *CIRASS*-Napoli, 2002) et *Segview98* (*Segmented and Annotated Speech Viewer 1.0* © Politecnico di Bari, 2000-2002).[6] Le premier peut être utilisé efficacement pour rechercher les contextes phonétiques, lexicaux et phono-syntaxiques de réalisation des éléments sélectionnés (avec des critères de recherche intuitifs et bien illustrés, voir un exemple ci-après). Le deuxième permet un affichage et une analyse simplifiée de quelques paramètres acoustiques qui caractérisent les signaux de parole choisis (extraits du dialogue, mots des listes lues) représentés dans un format spécifique (mais facile à convertir en format.wav). Il est possible de visualiser, à l'aide de ce logiciel, quatre niveaux parallèles d'étiquetage dans lesquels des annotations de type orthographique et phonologique s'alignent sur le signal acoustique (affiché avec une sélection de courbes). De plus, les transcriptions des principaux phénomènes phonétiques (une plus large et l'autre plus fine) sont détaillées dans les divers segments (voir la figure ci-dessous).



Une attention particulière des auteurs dans l'analyse préalable de ces matériaux est consacrée au codage prosodique, un aspect qui détermine un enjeu actuellement très important.

Dans le document d'accompagnement divers systèmes de transcription des phénomènes suprasegmentaux sont présentés et évalués distinctement (la méthode d'analyse *INTSINT*, la notation *PROLAB* et les principes de transcription *ToBI*). Toutefois, les données acoustiques associées à une transcription ont été analysées et étiquetées (en plus des quatre niveaux cités précédemment) à un niveau prosodique

incluant des éléments descriptifs du développement mélodique de la courbe de F0 et des éléments plus subjectifs d'évaluation de la proéminence.

Au premier niveau, les symboles utilisées (H, L, S, D, U) permettent d'associer à un événement d'évolution de F0 (respectivement pic, creux, stabilité du niveau, progression par paliers vers le bas ou vers le haut) des étiquettes qui peuvent faire l'objet d'une recherche par expressions régulières. À ces étiquettes s'en ajoutent d'autres qui permettent de situer les phénomènes codés par rapport aux paramètres généraux de la courbe analysée (T pour la valeur la plus élevée de l'unité tonale, B pour la valeur la plus basse) et des indicateurs de cibles de frontière prosodique («[» et «]») pouvant être utilisés en combinaison avec les autres étiquettes pour décrire les courbes plus finement.

À ce même niveau, une évaluation est donnée pour les réalisations attendues des accents lexicaux (0 pour les cas de déaccentuation, 1 pour des proéminences issues de variations de F0, durée ou intensité – accents culminatifs –, 2 pour les proéminences principales, 3 pour les accents d'insistance).

Cette notation expérimentale est testée sur une partie du corpus. Elle se révèle peut-être trop subjective, mais elle propose néanmoins un accès sélectif aux positions proéminentes pour une évaluation éventuelle de leurs conditions de réalisation.

L'utilité des fichiers d'étiquetage associés aux enregistrements est cependant surtout liée à la disponibilité des différents types de codage segmental: du fait du caractère biaisé de ses contenus lexicaux (dialogues autours des référents suggérés par les cartes des *maptasks*), la base de données est surtout intéressante pour des analyses de type pragma-linguistique (prosodie et stratégies conversationnelles) ou phonétique (à différents niveaux).[7]

Comme annoncé ci-dessus, le logiciel fourni *QueryGenerator* se révèle particulièrement utile pour la consultation des données à ces niveaux.

L'interrogation de la *BD* repose sur des critères intuitifs et se révèle toujours efficace.[8] Comme démonstration du type de recherches possibles dans cette base de données sonores et de l'utilité de l'outil d'interrogation fourni, je propose ici les résultats d'un exemple d'application.

Dans les sept dialogues des *maptask* (*MT*) entre des locuteurs (hommes, H) de Naples (cartes A, B, 3 de C, 2 de D), sur la base des étiquettes attribuées par les transcripteurs (et si on exclue les quelques rares cas de superposition), 135 occurrences de la mi-occlusive

postalvéolaire (labialisée) sourde /tS/$_{SAMPA}$ présentent des réalisations réellement mi-occlusives de type [tS]$_{SAMPA}$.[9] Le transfert automatique des mesures de durée associées à ces segments dans une feuille de calcul (fourni par le logiciel) nous permet de traiter cet ensemble de données et de calculer par exemple une durée moyenne de 103 ms et un écart-type de 34 ms.[10] De même, on retrouve 48 réalisations géminées du type [ttS] avec une durée moyenne de 129 ms et un écart-type de 25 ms.[11] Cela nous permet d'observer des bonnes conditions de maintien de l'opposition entre les unités phonologiques associées à ces réalisations (à Naples).[12] Si on explore les données de Pise relatives aux mêmes segments (*MT*, H, 7 dialogues: cartes A, 2 de B, 2 de C, 2 de D), on observe que – dans ces productions – le phonème /tS/ présente des réalisations du type [tS] seulement 9 fois, alors qu'il se réalise 126 fois comme [S], avec une durée moyenne de 105 ms et un écart-type de 25 ms, tout à fait comparables aux valeurs-type des dialogues de Naples.[13]

Voici donc un exemple d'utilisation de cette *BD*:[14] il est évident que les possibilités sont nombreuses et varient selon plusieurs axes d'intérêt.

Venons-en maintenant à la deuxième *BD* concernée par ce compte-rendu, celle de C-ORAL-ROM – *Corpus ORAL de langues ROManes*, une base de données linguistiques publiée sous forme de volume + *DVD* (*Integrated Reference Corpora for Spoken Romance Languages*).[15]

C-ORAL-ROM est le résultat d'un projet de recherche financé par l'Union Européenne au sein du programme IST 2000 du 5ème programme-cadre. Il s'agit d'une base de données linguistiques, de plus de 120 heures d'enregistrements, elle aussi conçue en vue d'une étude contrastive des 4 langues romanes considérées (Espagnol, Français, Italien et Portugais).

Pour permettre notamment cette comparaison, le rassemblement des données a demandé la définition d'un paradigme classificatoire des unités linguistiques présentes dans le corpus.

Etant donnée l'existence de traditions d'analyse linguistique légèrement différentes d'une langue à l'autre, cette tâche n'a évidemment pas été facile, et les équipes de recherche qui ont participé au projet ont certainement dû résoudre plus d'un conflit entre les diverses solutions envisagées en fonction d'une remarquable variété d'approches théoriques, de références techniques et de choix terminologiques.

Le volume annexe revêt une importance particulière. Il présente en effet des sections consacrées aux différentes langues traitées, avec

des données quantitatives et des commentaires à propos des contenus et des propriétés linguistiques des diverses parties du corpus. La sélection d'articles proposée offre des informations détaillées en ce qui concerne les points de divergence traditionnelle quant au traitement des phénomènes les plus communs dans les quatre langues. Il présente également un ensemble de paramètres d'évaluation pour surmonter les particularités de chaque langue et les approches spécifiques de chacune des communautés scientifiques des pays concernés. En sont témoins aussi les structures des divers chapitres, où les auteurs se sont efforcés de converger vers une description des matériaux offerts suivant un même schéma de présentation (et d'organisation sous-jacente).

Les chapitres illustrant les divers volets géolinguistiques du corpus proposent de façon plus détaillée une évaluation des contenus de ces quatre sections ainsi que des compte-rendus sur l'état d'avancement de la linguistique de corpus dans les différentes langues, avec des indications sur les autres corpus existants et sur les conditions de rassemblement des matériaux fournis par C-ORAL-ROM, souvent extraits d'autres collections plus vastes et plus variées.[16]

Seul un choix des sections, permettant la comparaison d'une langue à l'autre de par leur contenus et leur taille, est ici proposé: mon objectif étant celui d'une comparaison équilibrée.[17] Dans le premier chapitre de présentation et dans les sections «Corpus Metadata» et «Diagrams»[18] du *DVD*, une sélection de données quantitatives décrivant le corpus est donnée, tout comme est donnée une première évaluation des principales tendances de structuration observables, par exemple, au niveau de la longueur des énoncés ou des unités rythmico-tonales selon les diverses sections.[19] Pour avoir une idée quantitative de l'organisation des données présentes dans la *BD*, on peut également se référer au schéma synoptique proposé par le site *web* italien de C-ORAL-ROM et ici reproduit:[20]

	Nb. de fichiers WAV	Espace mémoire	Durée totale	Nb. d'énoncés	Nb. de mots	Nb. de Locuteurs		
						Total	Hommes	Femmes
Espagnol	210	4,56 GB	31h 06' 00"	35588	333482	410	247	163
Français	206	3,77 GB	26h 21' 43"	21010	295803	305	154	151
Italien	204	5,19 GB	36h 16' 10"	40402	310969	451	276	175
Portugais	152	4,43 GB	29h 43' 42"	38855	317916	261	144	117

On constate au total un bon équilibre quantitatif entre les différentes sections et surtout une quantité impressionnante de données

pour chaque langue. L'un des objectifs des auteurs était en effet de créer une collection suffisamment exhaustive de textes, de styles et de contextes de parole spontanée pour les quatre langues prises en compte.[21]

Les données contenues dans le *DVD*, dans le dossier «Multimedia_Corpus», s'organisent pour chaque langue, en deux sous-sections «Formal» et «Informal».

De la première sous-section, on trouve les enregistrements des trois catégories «media» (divisée en sous-classes: *documentary*, *interviews*, *news*, *scientific_press*, *sport*, *talk_show*, *weather_forecast*),[22] «natural_context» (divisée en sous-classes *business*, *conference*, *law*, *political_debate*, *political_speech*, *preaching*, *professional_explanation*, *teaching*) et «telephone» (divisée en sous-classes *human-machine*, *private_conversations*) avec un nombre légèrement variable d'ensembles cryptés de fichiers sonores et annotations.

Dans la sous-section «Informal», on trouve en revanche les enregistrements des catégories «family_private» et «public» (les deux étant divisées dans les sous-classes *conversations*, *monologues*, *dialogues*) contenant elles aussi un nombre variable, mais assez équilibré, de données (pour les conventions d'attribution des noms des fichiers voir C-ORAL-ROM: 39-40).

Comme on peut le constater par cette organisation, le matériel linguistique présent dans la base de données, bien qu'observé strictement dans le cadre d'une théorie bien définie, s'inscrit dans une vision de la variation linguistique qui repose sur des distinctions diamésiques plutôt intuitives (et universelles).[23]

Pour ne citer que quelques exemples, on pourrait évidemment relever la variété très commune d'*écrit-écrit* qui se manifeste dans des documents administratifs, ou dans des lettres personnelles, destinés à rester écrits et à ne pas être lus à haute voix. Les sermons préparés par écrit et les textes des journaux radiophoniques ou télévisés représentent au contraire des exemples d'*écrit-parlé*, c'est-à-dire des textes rédigés exprès pour être lus. A l'inverse, la production orale improvisée d'un locuteur qui sait que son message sera transcrit est un bon exemple de *parlé-écrit*. Les dialogues spontanés en style informel peuvent être assumés comme un modèle très commun de *parlé-parlé*. Sauf la première variété, qui n'est pas représentée dans la *BD* sonore de C-ORAL-ROM (et la troisième variété qui l'est implicitement dans les productions de locuteurs concernés par le projet – la transposition dans l'écrit étant soignée parfois par eux-mêmes), les autres variétés constituent les principaux objectifs de ce projet et constituent le cœur du corpus recueilli.

Toutefois sans le choix d'un modèle, comme celui défini par Emanuela Cresti, l'analyse du discours et le classement des unités de structuration des productions linguistiques aurait ressenti des «préférences» de chaque langue pour diverses structures syntaxiques et suprasegmentales; la détection de schémas communs aurait échoué ou aurait été au moins perturbée (comme le souligne Claire Blanche-Benveniste dans la présentation de l'ouvrage) par un taux élevé de phénomènes de déviation qui caractérisent la parole spontanée.

La segmentation des matériaux résultant de cette approche spécifique permet cependant de définir des structures homogènes dans lesquelles se retrouvent les stratégies parallèles, liées aux divers choix locaux des langues en question. L'analyse des productions en termes d'actes de parole unitaires recherchés à un niveau prosodique assez large, dans lequel se manifestent des unités pragma-linguistiques compatibles, bénéficie d'une conception de l'énoncé comme unité fondamentale de structuration et d'analyse.[24]

Comme présenté à la fin de l'ouvrage, c'est l'accord entre différents opérateurs dans la notation et dans le positionnement des frontières prosodiques définies à ce niveau (grâce à la présence des indices de rupture prosodique associés à ce genre d'unités) qui prouve l'efficacité et l'utilité de la segmentation proposée.

D'importantes réflexions préliminaires à ce sujet sont résumées par Massimo Moneglia dans la présentation du volume et de la *BD,* et sont ensuite reprises dans les exemples d'application donnés au chapitre 6 par Emanuela Cresti.

Concernant l'utilité plus directe des données brutes qu'offre C-ORAL-ROM, on se heurte aux difficultés qui sont posées par une *BD* sonore dont les données sont en réalité inaccessibles pour un traitement autonome: elles sont audibles et analysables uniquement avec les logiciels propriétaires fournis. Les enregistrements sont associés seulement aux transcriptions orthographiques conventionnelles de leur contenu (avec une fiche complète d'informations sur les données mêmes, *metadata*, cf. C-ORAL-ROM: 28), enrichies de marqueurs d'interruption prosodique, indication de pauses, hésitations etc., mais sans aucun étiquetage segmental: s'agissant de plus de 120 heures d'enregistrements (772 textes d'une dizaine de minutes de parole en moyenne), cela aurait certainement demandé un travail excessif.

Les transcriptions annotées de C-ORAL-ROM proposent l'identification et la délimitation de deux unités prosodiques bien définies (établies à la main): une unité terminale vs. une unité non-terminale (qui, dans d'autres termes, peuvent se retrouver délimitées par une frontière forte vs. une frontière faible).

L'unité prosodique terminale est composée d'une série d'unités prosodiques mineures (les unités prosodiques non-terminales) qui, toutes ensemble, constituent une «allure cohérente» à l'intérieur du *continuum* de la chaîne parlée.[25]

Dans les transcriptions proposées, les diverses productions ont été segmentées en unités intonatives délimitées par une barre double (*double slash* //), insérée chaque fois qu'une unité prosodique a été considérée comme terminale (cela se produit à proximité d'une pause dans seulement 37% des cas).[26]

Les unités non-terminales ont été délimitées par une barre simple (*slash*, /) sur la base de la perception d'un changement intonatif continu (voir ci-dessous).[27]

La consultation est possible grâce à *Macromedia Flash Player* et à un logiciel d'analyse, *WinPitch Corpus*, une version particulière du logiciel *WinPitch* de Philippe Martin.[28] La maîtrise de cet outil formidable n'est pas facile pour un non-spécialiste: dans mon cas, la *BD* n'a pas été accessible aux premiers essais sur un PC avec Windows 2000, mais il a fonctionné du premier coup sur d'autres ordinateurs avec le même système d'exploitation ou avec XP.[29]

Concernant plus spécifiquement la qualité des fichiers sonores (qui sont à l'origine en codage .wav, puis soumis à un traitement similaire à la compression .mp3), il faut aussi signaler dans certains cas la présence fortement indésirable de bruits de fonds et les mauvaises conditions générales d'enregistrements. Cela est tout à fait acceptable lorsqu'il s'agit de parole spontanée en plein air ou – même en conditions de laboratoire – lorsqu'on est en présence de plusieurs locuteurs (on sait que dans ce domaine, on doit réaliser un compromis entre une qualité sonore acceptable et un bon degré de naturel pour une représentation satisfaisante de la parole spontanée). Toutefois, cela est un peu moins justifiable lorsqu'il s'agit par exemple de données de parole radiophonique ou télévisées.[30]

Les auteurs ont prévu trois types pour le classement de la qualité acoustique de chaque enregistrement de la *BD*: A – enregistrements numériques avec microphones monodirectionnels garantissant une excellente qualité audio; B – enregistrements avec une réponse acoustique suffisante en présence de faibles bruits de fond; C – basse qualité audio à cause de sources de bruits d'environnement et/ou superposition entre les productions de différents locuteurs. Malgré cela, dans quelques cas, la mauvaise qualité est liée à la saturation ou à des conditions douteuses de numérisation: même comprimés, les fichiers sont garantis avec la qualité d'un codage PCM à 16 bit et à 22050 Hz, mais en réalité, on peut trouver ça et là, même en utilisant

l'affichage de *WinPitch*, des représentations spectrographiques qui s'arrêtent à 5000 Hz environ (numérisation à 10kHz?) ou bien qui se présentent miroitées au-dessus d'environ 5500 Hz.[31] Même si limité à une petite quantité de données (1% des fichiers ouverts), ceci représente un défaut assez grave pour une *BD* si importante.

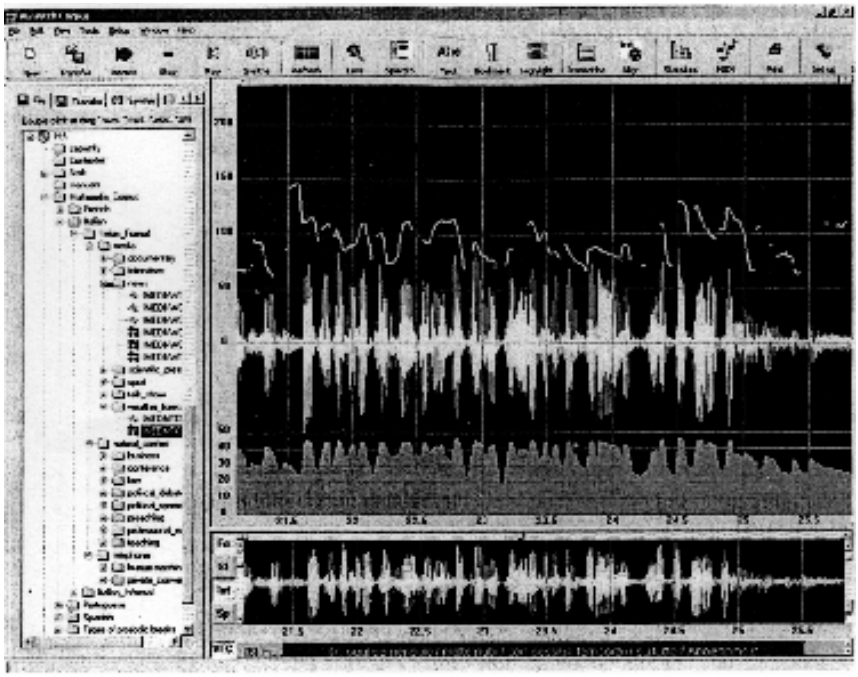
Les transcriptions alignées qui accompagnent les fichiers sonores représentent sans aucun doute un apport considérable à la recherche des mots et d'autres faits linguistiques transcrits: elles facilitent en outre un accès plus immédiat aux données orales dans les fichiers sonores.[32] Au-delà de l'important travail qu'elles ont dû demander, elles représentent aussi une 'mine documentaire' pour les annotations qu'elles contiennent et les autres informations de segmentation prosodique qui se prêtent à des évaluations automatiques (à condition d'avoir accès aux fichiers non cryptés, ce qui n'est pas le cas).

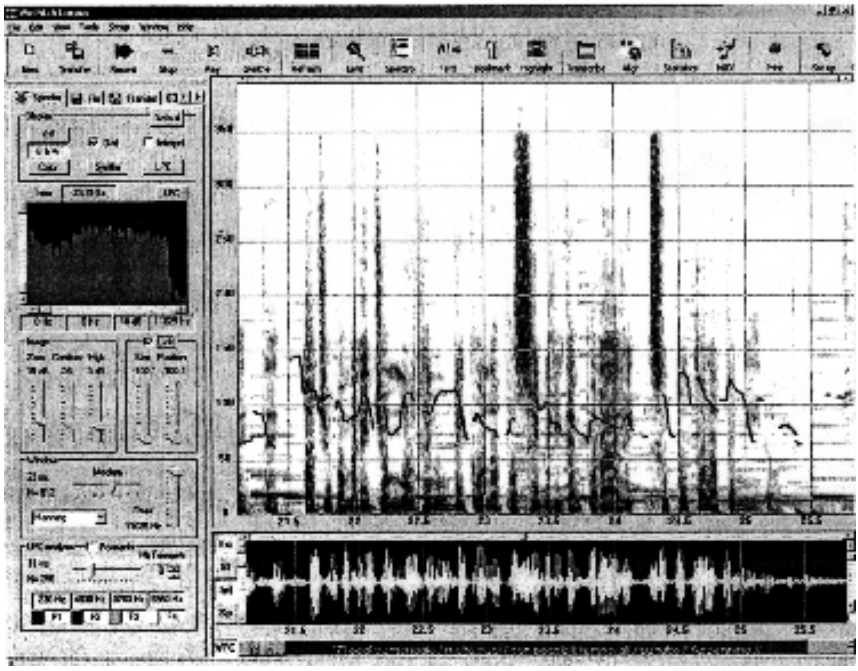
Il est possible de visualiser les informations (textuelles et acoustiques) concernant les divers passages à l'aide du logiciel *WinPitch Corpus*, qui offre plusieurs solutions d'affichage parmi lesquelles celles que je propose dans les figures à la page suivante (j'ai pu les composer moi-même).

Dans la première figure, le signal acoustique, la courbe de l'intensité, une autre version (élargie) du signal acoustique, et la courbe de F0 s'alignent sur la transcription orthographique du contenu linguistique entre deux interruptions terminales. Cette transcription est affichée sous la figure (elle est marquée par les frontières non-terminales: *oggi pomeriggio / molte nubi / con possibili temporali / su tutto l'Appennino*).

Dans la deuxième figure, j'ai choisi de présenter pour le même passage, outre la courbe de F0, un spectrogramme à bande plutôt étroite (sur 512 points) qui permet d'apprécier, au delà des indications segmentales sur la production en question, la présence de traces de musique en arrière plan.[33] Les nombreuses options de visualisation fournies par le logiciel sont affichées dans le cadre à gauche.

Concernant la section italienne de C-ORAL-ROM, il est intéressant d'observer – comme le remarque Cresti (2005) – la quantité d'expressions qui ne satisfont pas à la condition d'efficacité communicative parce qu'elles sont interrompues: le pourcentage atteint 12 à 13% des productions rassemblées.





Pour toutes les expressions qui atteignent un but communicatif, car elles présentent un caractère d'autonomie ou bien parce qu'elles satisfont les conditions structurelles d'auto-suffisance sémantique, on peut observer les chiffres suivants: 61,9% des expressions autonomes contiennent des syntagmes verbaux (par exemple: «non entrò nel partito»); 38,1% des expressions autonomes se présentent comme des "énoncés primitifs nominaux" (constitués par un syntagme nominal, un adjectif, une interjection etc.; par exemple: «eh», «quindici aprile millenovecentonovantasette», «in montagna»); seulement 5% environ des données orales semblent présenter un caractère d'auto-suffisance sémantique et peuvent ainsi satisfaire à la définition de phrase (par exemple: «sono atterrati quattro C130»).[34]

L'autonomie communicative résulte donc de l'interprétabilité pragmatique d'énoncés verbaux ou nominaux qui recouvrent ainsi plus de 90% des productions, alors que les expressions sémantiquement complètes ne dépassent pas 10% du matériel linguistique de ce corpus.

D'autres paramètres de comparaison ont été utilisés pour évaluer la variabilité des matériaux collectés pour ces langues: mesures moyennes de longueur des unités prosodiques, du débit de parole (diverses définitions) et indices de fragmentation.

Les paramètres qui ont été définis à cet effet sont les suivants (tous mesurés en nombre de mots graphiques): *MLU* (*mid-length of utterances*, longueur moyenne de l'énoncé), *MLTone* (*mid-length of the tone unit*, longueur moyenne de l'unité tonale comprise entre deux frontières non terminales), *MLTw* (*mid-length of the dialogic turn*, longueur moyenne du tour de parole dans les dialogues). Pour l'évaluation de la vitesse, on s'est référé à un calcul brut du nombre de mots par seconde, tandis que pour la mesure du degré de fragmentation, on a tenu compte du pourcentage d'énoncés interrompus et des faux départs par rapport au nombre total d'énoncés dans chaque section du corpus. Les graphiques de comparaison et le cadre général qui se dégage sont illustrés dans une section du *DVD* et au premier chapitre du volume (pages 57-62).

Concernant la *MLU*, on observe en général, pour les quatre langues confondues, des valeurs basses dans les conversations téléphoniques privées (la moyenne est de l'ordre de 5 mots), alors que dans la parole spontanée en monologue et en contexte naturel, ces valeurs peuvent augmenter de manière très significative (15 à 22 mots). De manière surprenante, l'italien et le portugais sembleraient globalement privilégier des longueurs d'énoncé réduites, comparés à l'espagnol et surtout au français. Comme on peut le remarquer facile-

ment, une des raisons de cet écart pourrait être une longueur du mot relativement moins importante en français.[35] L'écart se réduit (ou carrément s'inverse entre espagnol et français dans un contexte de dialogue naturel) si on tient compte de la *MLTw*, les nombres moyens de mots par tour de parole pour les quatre langues se concentrent autour des mêmes valeurs (de 7-8 à 36 mots environ). L'influence de l'écrit dans l'évaluation des propriétés de l'oral se manifeste à nouveau dans la comparaison des débits de parole des quatre langues pour les différents genres textuels analysés. Au-delà d'une valeur aberrante (au dessous de 2 mots par seconde) pour le français en contexte de dialogue naturel (alors que les autres langues présentent entre 2,4 et 3,2 mots par seconde et que cette même langue présente d'habitude des valeurs supérieures de 3,5 dans les autres contextes informels), l'effet parasite d'une mesure non phonétique se manifeste dans l'écart systématique entre l'italien et les autres langues. Pour l'italien on observe toujours un débit entre 2 et 2,8 mots par seconde (soit 0,3 à 1,2 mots en moins par rapport aux autres).

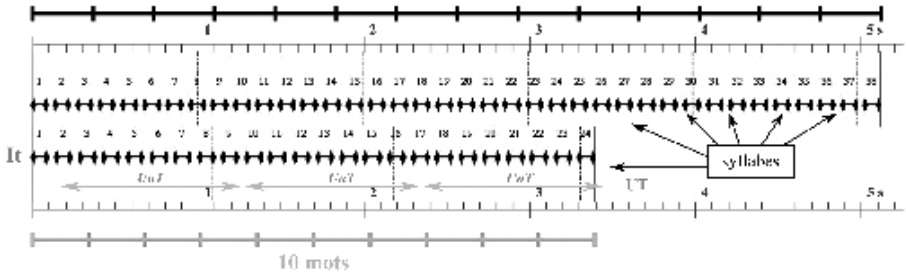
Dans le cadre d'un mémoire de fin d'étude (voir note 31), la vérification des ces paramètres a été effectuée sur la base de mesures phonétiques (cf. Bertinetto & Magno Caldognetto 1993; Zmarich et al. 1996; Giannini & Pettorino 1999; Giannini 2000; Pettorino 2003) pour la section média de l'italien et de l'espagnol. Alors que les mesures de C-ORAL-ROM sont exprimées en mots par seconde (mots/s), dans ces évaluations, le débit de parole a été calculé en termes de syllabes par seconde (syll/s). Cela a permis d'obtenir une réduction de cet écart avec un taux de fluence (*fluency rate*) de 7,3 syll/s pour l'espagnol et de 7,0 syll/s pour l'italien. Une durée moyenne de 5,1 s a été associée à la mesure de 14 mots de *MLU* pour les unités terminales (*UI*) de l'espagnol (constituées de 38 syllabes en moyenne). De la même façon, une durée moyenne de 3,4 s a été déterminée pour ces unités dans les données de l'italien (caractérisé par une mesure de *MLU* de 10 mots et par une quantité moyenne de syllabes égale à 23,6 syllabes par *UI*).

Si on considère le niveau des unités non terminales (*UnT*), les graphiques de la page 60 suggèrent des valeurs de *MLTone* autour de 3,2 mots pour l'espagnol, comme pour l'italien.

Dans notre système de référence les *UnT* de l'espagnol seraient constituées d'un nombre moyen de 7,4 syllabes tandis que celles de l'italien en auraient 7,7. D'après ces mesures, en effet, les *UnT* de l'italien sont moyennement plus longues (1,1 s vs. 1 s) et se présentent à l'intérieur d'*UI* qui sont au contraire plus courtes, comme nous l'avons présenté ci-dessus. Cela détermine un nombre moyen d'*UnT* plus bas par *UI* (5,2 *UnT*/*UI* pour l'espagnol médiatique vs. 3,1 *UnT*/

UI pour l'italien) et confirme ainsi le genre de conclusions des auteurs à ce propos: les *UI* de cet échantillon d'espagnol sont plus 'construites' que celles de l'italien.

Nous proposons dans le schéma suivant une comparaison synoptique de ces mesures (en haut l'espagnol médiatique et en bas l'italien médiatique).



Dans la moitié haute, on remarque la correspondance entre les 14 mots d'une *UI* idéale de l'espagnol (5,1 s) avec ses 38 syllabes et ses 5,2 *UnT* (de 7,4 syllabes et 1 s environ de longueur): le nombre de mots par *UnT* est cependant inférieur à 3. La durée moyenne d'une syllabe est de 137 ms et le débit de parole qui en résulte est de 7,3 syll/s.

Dans la moitié basse, on remarque les 23,6 syllabes de l'*UI*-type de l'italien (3,4 s) divisée en 3,1 *UnT* (de 7,7 syllabes et 1,1 s environ de longueur): la correspondance avec les 10 mots proposés au premier chapitre de C-ORAL-ROM comme mesure de longueur nous mène à en attribuer un peu plus de 3 à l'*UnT*-type. La durée moyenne d'une syllabe est dans ce cas de 143 ms et correspond à un débit de 7,0 syll/s.

Avec ces exemples de lecture, j'espère avoir donné une idée au moins grossière du potentiel de cet énorme corpus (qui présente néanmoins quelques limitations). Les réflexions que cette *BD* a su stimuler sont en effet déjà considérables. Pour ne citer qu'une donnée quantitative, rien qu'en considérant le travail de recherche du LABLITA – si on en juge par les publications disponibles sur le site *web* de ce laboratoire –, 72 publications ont à ce jour déjà fait l'objet d'approfondissements divers, à partir de la première version du corpus de 2002.

Avant de conclure, je voudrais encore rappeler qu'un volet important (du *DVD* et des divers chapitres du livre) est constitué par les tableaux de comparaison entre les rangs d'occurrences de différentes parties du discours étiquetées de manière plus ou moins uniformes pour les quatre langues.

Concernant les critères procéduraux de classification, il est dommage que les auteurs des différentes sections aient fait des choix différents, par exemple dans l'étiquetage des parties du discours (*PoS tagging*, cf. C-ORAL-ROM: 51). Certains éléments linguistiques possédant des fonctions similaires d'une langue à l'autre ont reçu un traitement différent, par exemple ceux qui sont désignés *MD* (marqueurs du discours) en espagnol (*pues*, *mira* etc.) et en portugais (*pois*, *pronto* etc.) mais qui ne correspondent pas à leurs contreparties dans les deux autres langues.[36]

Les divers choix se retrouvent détaillés dans les tableaux des étiquettes (*tagset* ou *tag-set* ou *tag set*) aux différents chapitres et dans la section *Frequency Lists* du *DVD* (où l'on peut aussi retrouver une évaluation des performances du *taggueur* automatique utilisé).

Dans le corpus français, les auteurs ont correctement choisi d'étiqueter comme adverbe le mot multiple *un_peu*. Il est dommage que dans le corpus italien, *un_po'* n'ait pas reçu le même traitement (on retrouve donc «po'» isolé avec une étiquette B=adverbe alors que dans d'autres cas, comme par exemple «per_esempio», «fino_a» etc., le choix a été de garder le lien entre les deux éléments).

Mais on sait que dans ce genre d'entreprise les choix sont parfois forcés et même les auteurs regrettent souvent d'avoir opté pour telle solution sur laquelle ils ont longuement hésité…

Ce même problème affecte, à une échelle microscopique, certains textes qui accompagnent les fichiers sonores. De manière tout à fait occasionnelle et non-uniforme, les transcriptions du contenu des enregistrements – pourtant réalisées par des opérateurs entraînés (et revues par deux autres opérateurs) – présentent des infidélités notables par rapport au contenu des productions effectivement proposées, avec l'omission de passages courts: elles peuvent présenter une interprétation improbable, ou l'omission de l'indication de certains allongements finaux et de certaines formes d'hésitations (dans un cas, ce problème concerne aussi l'échange entre les indications des deux participants à un dialogue).[37]

En général, j'ai observé une négligence particulière dans la transcription de termes et de noms étrangers (cela est d'autant plus surprenant lorsqu'il s'agit d'expressions communes dans le domaine de la linguistique comme *bedring au lieu de *babbling*, ou en général dans le milieu universitaire comme *Prinstone à la place de *Princeton*, *Standford à la place de *Stanford*).[38]

Par rapport à ces faits secondaires, le problème pourrait se résumer au choix de l'anglais comme langue de présentation des données (choix critiqué à de nombreuses reprises lors des premières annonces

de la *BD,* pour des raisons commerciales évidentes); cependant, lorsque ce choix devient un obstacle rendant plus difficile la tâche des analystes et des commentateurs, en générant des sources supplémentaires d'imprécision,[39] on peut se demander s'il n'aurait pas été plus pertinent de choisir une langue romane comme langue 'instrumentale' de présentation.

Ces remarques ne sauraient faire oublier l'apport le plus important de ces deux collections de données: elles nous offrent une grande quantité de matériaux sonores sur lesquels nous pouvons tester nos outils, vérifier un grand nombre d'éléments théoriques, ou encore élaborer de nouvelles méthodes d'analyse de la langue parlée.

Du fait de leur format final, nous avons constaté que: 1) la première *BD* est riche en informations segmentales; elle s'appuie aussi, en partie, sur les éléments expérimentaux d'un codage prosodique mixte (elle propose en plus des données de comparaison entre parole semi-spontanée et parole pathologique); 2) la seconde *BD* est en revanche beaucoup plus riche en termes de variété de registres et de genres textuels: avec un grand choix de formes de communication, elle offre les résultats d'un travail d'analyse (et d'étiquetage) considérable, dévolue à l'évaluation d'aspects pragma-linguistiques et de structuration de l'information, associés à la prosodie de l'énoncé. L'autre nouveauté de cette contribution à la recherche linguistique est la découverte que ces niveaux de structuration peuvent être analysés avec un grand consensus. On peut trouver certainement très utiles à ce propos l'essai de comparaison de l'organisation rythmico-prosodique et la démonstration des techniques et des résultats d'une évaluation croisée de ces données pour les quatre langues considérées.

En conclusion, les deux *BD* représentent un outil formidable, mais souffrent peut-être encore des difficultés à concilier un projet idéal de *ressources* de cet ampleur avec les contraintes pratiques de sa réalisation qui impliquent un travail fastidieux confié trop souvent à du personnel pas assez spécialisé et/ou ayant une formation encore peu interdisciplinaire.

Addresse de l'Auteur

Antonio romano, Università di Torino, Dipartimento di Scienze del Linguaggio, via Sant'Ottavio, 20 10124, Torino
antonio.romano@unito.it

Notes

[1] API – *Archivio del Parlato Italiano* (coord. F. Albano Leoni), sans date (2003), 1 *DVD* distribué gratuitement par le *CIRASS* - Università degli Studi di Napoli "Federico II", via Porta di Massa, 1, I-80133 Napoli (Italie); dvd_api@cirass.unina.it. C-ORAL-ROM – *Corpus ORAL de langues ROManes* (coord. E. Cresti - M. Moneglia), 2005, volume + 1 *DVD*, Amsterdam-Philadelphia, John Benjamins Publishing Company, 120 € (http://www.benjamins.com/cgi-bin/t_bookview.cgi?bookid=SCL%2015); disponible aussi en édition étendue, 9 *DVD*, distribuée par *ELDA*, à des prix différents entre 1500 et 20000 € (http://www.elda.org/catalogue/en/speech/S0172.html).

[2] Dans ce compte-rendu j'appellerai "bases de données" (*BD*) ces deux collections d'enregistrements de productions linguistiques que leurs divers auteurs ont voulu regrouper respectivemenet sous une désignation d'*archive* ou de *corpus*. La notion de "base de données" s'adapte peut-être mal à ce genre de matériaux formatés et recueillis selon un protocole unique et rigoureux, mais l'utilisation de ce terme se revèle efficace pour ses qualités d'hyperonyme (d'autant plus qu'il ne s'agit pas simplement d'un ensemble d'enregistrements: d'important documents d'accompagnement et d'évaluation préalable étant aussi disponibles). Je préfère ne pas évoquer ici les enjeux linguistiques, théoriques et pratiques, industriels et commerciaux des bases des données orales. Je me limite à souligner le tournant récent en linguistique et en phonologie vers ce genre d'approche, des références utiles sur ce thème général étant fournies entre autres par Newmeyer (2003).

[3] Mes remerciements s'adressent à Laurent Girin et aux deux rapporteurs anonymes de ce compte-rendu: leurs suggestions m'ont aidé à améliorer la description (dont je reste le seul responsable) des deux publications concernées. Quant aux auteurs des deux *BD*, j'espère qu'ils voudront bien accueillir cette présentation comme témoignage d'une appréciation générale de l'utilité de leurs corpus: les défauts observés sont ici mentionnés – cela va sans dire – uniquement dans la perspective de l'amélioration future de ce genre d'ouvrages.

[4] Un certain nombre de phrases lues est aussi disponible pour l'analyse de phénomènes particuliers (comme la *gorgia toscana* ou le redoublement phonosyntaxique) pour 5 locuteurs de Pise.

[5] Pour les diverses cartes, un ou plusieurs dialogues sont disponibles pour un choix de localités. Au dialogue de chaque carte sont normalement associés les fichiers sonores du dialogue entier, les fichiers sonores issus de la division en tours de parole (D_T) et les deux fichiers sonores contenant la lecture de la liste des *toponymes* (L_{Top}) de la part des deux participants au dialogue. Pour la carte **A** sont proposés les dialogues: *A01* de Naples et *A03* de Pise. Pour la carte **B** sont proposés les dialogues *B02* de Bari, *B01* de Florence (sans L_{Top}), *B03* de Naples et *B03* de Pise. Pour la carte **C** sont proposés les dialogues *C01* (sans D_T) et *C02* (sans D_T) de Bari, *C02* et *C04* de Naples et *C03* de Pise. Pour la carte **D** on dispose enfin des dialogues *D01* et *D02* (sans D_T) de Bari, *D01* (hors dossier, sans D_T) de Brindisi, *D01* et *D02* de Naples et *D01* (sans L_{Top}) et *D02* de Pise.

[6] Au-delà des manuels qui illustrent les conditions d'usage de ces deux logiciels, enrichis par la disponibilité d'autres outils – tels un extracteur automatique de formants (valeurs brutes) et un syllabateur acoustique – parmi les fichiers annexes au *DVD*, nous retrouvons des documents d'explication des codages utilisés et des chapitres de commentaire sur l'approche syntaxique à la base des annotations morphosyntaxique et pragmatique. Pour ce genre de productions linguistiques, le lecteur pourra trouver également utile la réflexion fournie conjointement à propos du traitement des phénomènes d'hésitation.

[7] Étant donnée l'attention qui a été portée à ce genre de phénomènes, le lecteur sera surpris de remarquer une certaine approximation dans la présentation des symboles phonétiques dans le document des consignes pour la représentation, analyse et codage des données. Cela est d'autant plus étonnant si l'on considère la diffusion actuelle en Italie d'un grand nombre de documents qui ont désormais adopté les conventions API. Certaines d'entre elles ont été ici ignorées malgré le choix heureux de se référer à un codage SAMPA. Ainsi nous retrouvons une terminologie inadéquate dans l'illustration des symboles adoptés pour $[l]_{SAMPA}$ («liquida laterale» au lieu de *laterale alveolare*) et pour $[L]_{SAMPA}$ («liquida palatale» au lieu de *laterale palatale*). De la même façon, alors que le terme approximant est employé pour $[G]_{SAMPA}$ (constrictive vélaire sonore, «allofono approssimante dell'occlusiva velare»), nous observons des termes inappropriés pour [j] («semivocale palatale») et [w] («semivocale velare»). La définition la plus gênante est cependant celle de $[S]_{SAMPA}$ comme «fricativa palatale sorda». Or, nous savons que les sons associés à ce symbole sont tout au plus postalvéolaires et que des constrictives palatales sourdes (allophones de /x/ en allemand ou en grec moderne) auraient dues être représentées par le symbole $[C]_{SAMPA}$.

[8] Une certaine quantité – heureusement limitée – de documents associés aux données sonores produirait des fausses estimations lors d'une interrogation automatique du fait d'un pourcentage élevé d'erreurs et du manque de respect des conventions de transcription. Notamment, dans la section de parole enfantine, j'ai remarqué un manque d'attention non négligeable (par exemple, dans l'interview P01, et d'autres, la transcription orthographique présente la forme *fà pour la voix verbale *fa*; dans l'interview P04 on peut observer la transcription <più parli è meglio è>). Les fréquentes oscillations sont particulièrement néfastes dans la transcription de *c'è* (en P01 on retrouve la graphie traditionnelle <c'è>, tandis que dans d'autres apparaît la graphie non conventionnelle <c'e">). D'autres choix inconstants apparaissent par exemple dans l'interview S03 qui présente l'alternance *Gerri*/*Jerri*. En général, l'étiquette <schiocco di lingua>, malgré la tentative de définir cette notation conventionnelle, peut se retrouver aussi comme <schioccodilingua>, <schiocco_di_lingua> ou <schiocca di lingua>; de même que l'étiquette <inintelligibile> prend alternativement les formes <inintellegibile> et <inintelleggibile>. Le document de transcription de S02 mérite une attention particulière: il n'est pas dans un format .txt et cause des problèmes à l'ouverture du fichier (certains logiciels de protection pour *PC* l'isolent comme porteur d'un virus W97M/Verlor).

[9] Même lorsque ceci n'est pas indiqué explicitement, les transcriptions proposées ci-après se basent sur les normes SAMPA.

[10] Plus en détail, ces réalisations apparaissent: dans 31 cas (23%) dans des occurrences de *c'è* (durée moyenne 109 ms et écart-type 24 ms); dans 10 cas (7%) dans des occurrences de *ce* (durée moyenne 103 ms et écart-type 21 ms); dans 9 cas (7%) dans des occurrences de *centro* (durée moyenne 115 ms et écart-type 17 ms).

[11] Ces réalisations apparaissent: dans 9 cas (19%) comme exemples d'autogémination dans des occurrences de *c'è*, avec une durée moyenne de 137 ms et un écart-type de 19 ms; dans 10 cas (21%) dans des occurrences du contexte phonétique *a_a*/*o* (comme dans les mots *traccia*, *faccia*, *faccio*), avec une durée moyenne de 147 ms et un écart-type de 27 ms.

[12] Même dans les conditions de réduction communes dans la parole spontanée.

[13] Ces réalisations se manifestent: dans 19 cas (15%) pour des occurrences de *c'è* avec une durée moyenne de 87 ms et un écart-type de 17 ms; dans 15 cas (12%) pour des occurrences de *c'hai*/*c'ho* avec une durée moyenne de 103 ms et un écart-type de 17 ms. On peut aussi souligner le fait que /tS/ se réalise mi-occlusif dans un seul cas de *c'è* (sur 20); de la même façon, dans les données relatives aux dialo-

gues enregistrés à Florence, seulement 4 *c'è* sur 50 présentent un /tS/ → [tS]).

[14] Les mesures et les évaluations que je viens de présenter (qui – bien entendu – auraient pu être vérifiées à partir d'une observation directe des données brutes) n'ont demandé pas plus de 10 minutes.

[15] Publié par les soins de Emanuela Cresti et Massimo Moneglia chez John Benjamins Publishing Company (Amsterdam-Philadelphia), C-ORAL-ROM est disponible sous deux formes: une édition en *DVD* avec fichiers sonores compressés et cryptés (l'accès à la base de données est autorisé uniquement à partir des logiciels fournis); une version non-cryptée, conçue pour les centres de recherche en technologies de la parole, distribuée par ELDA (www.elda.fr) dans un format étendu (9 *DVD*) beaucoup plus cher.

[16] Le lecteur intéressé à une ou plusieurs de ces langues pourra retrouver la description des corpus correspondants aux diverses sections du volume. À la suite du premier chapitre de présentation, les chapitres 2, 3, 4 et 5 sont consacrés respectivement aux corpus italien, français, espagnol et portugais.

[17] Pour d'autres corpus d'italien parlé un résumé des principales sources connues est proposé au chapitre 2, avec l'indication de leurs dimensions et de leurs formats (du *LIP* de 1993 à *CLIPS*, en cours de publication).

[18] Voir par exemple les tableaux de données .xls et les diagrammes disponibles dans «Standard measurements». Les autres chapitres de cette section «Lexical Strategies», «Structural Strategies» et «Surface clause indexes» sont également intéressants à ce sujet.

[19] À ce sujet, le lecteur trouvera également intéressant les essais de synthèse dressés par les éditeurs (et, indirectement, par les évaluateurs) qui ont mené une réflexion conjointe sur ces unités vis-à-vis de certaines variables, tels les indices de structuration prosodique et certains aspects pragma-linguistiques. L'observation linguistique menée sur ces niveaux d'analyse, en général très controversés, s'insère évidemment dans le cadre d'une théorie et de méthodes procédurales mises au point au cours d'une expérience pluriennale. La robustesse de ces méthodes a été testée préalablement dans la constitution et l'évaluation du corpus d'italien LABLITA (voir les nombreuses informations disponibles sur le site Internet homonyme). Beaucoup d'outils, de documents et d'autres informations sur les premières phases préparatoires de ce projet (ainsi que des premiers résultats qui ont suivi) sont disponibles sur le site *web* http://lablita.dit.unifi.it/.

[20] Cf. http://lablita.dit.unifi.it/coralrom.

[21] Les variables liées aux facteurs socio- et géo-linguistiques sont peut-être moins équilibrées. A l'exception des échantillons de productions médiatiques, elles conditionnent de manière évidente les enregistrements plus 'sur le terrain'. Une considération déjà souvent soumise aux auteurs de la section italienne est par exemple que la dominance de données d'origine toscane est très forte. Le déséquilibre dans cette direction se justifie par le fait que la variété de cette aire est traditionnellement considérée représentatives de l'italien standard mais – comme l'ont fait remarquer les auteurs eux-mêmes – d'autres accents régionaux sont tout de même bien présents. Néanmoins, le caractère dialectal des données toscanes est souvent très sensible au point que la plupart des italiens ne se sentent pas du tout représentés dans l'échantillon et considèrent la sélection de productions proposée trop connotée sur ce plan (voir la liste de formes orthographiques non-standard présentées dans le texte). Ce déséquilibre est comparable en partie avec la forte connotation géographique qui marque les sections portugaise et espagnole, étant donné le rôle divers que jouent les variétés de portugais de Lisbonne et d'espagnol de Madrid dans les communautés linguistiques correspondantes. La justification que les données recueillies sont uniquement représentative d'une aire restreinte dans laquelle ont opéré les différentes équipes ne s'applique pas de la même mesure à

la section française. Cette dernière présente une meilleure répartition géographique dans l'espace de l'Hexagone, avec des données d'Alsace, d'Auvergne et, bien sûr, de Provence, mais aussi d'enregistrements effectués à Clermont-Ferrand, Limoges, Paris, Poitiers, Rouen etc. Pour ne citer qu'un seul exemple de la forte connotation régionale qui marque le corpus italien, on peut se référer à la liste de fréquence lexicale présentée dans le *DVD*: la dialectalité générale des données recueillies fait remonter au rang 63 (avec 687 occurrences !) la forme *'un*, réalisation dialectale toscane de l'italien *non*.

[22] Cette section n'est pas disponible, cependant, dans le corpus français.

[23] Il est évident que, de façon plus ou moins indépendante de la langue observée, les propriétés phonétiques de la parole lue et de la parole spontanée sont différentes. De la même manière, pour toute langue nationale, on peut prévoir des différences stylistiques plus ou moins importantes entre la variété employée dans les conversations entre amis, par exemple, et la variété employée dans les journaux télévisés. L'étude propose une bibliographie importante qui retrace le progrès des méthodes d'observation des aspects phonologiques des structures phrastiques (de Karcevskij (1931), à nos jours). L'attention portée à la variation spécifique sur ce niveau d'analyse doit beaucoup à des linguistes qui ont stimulé des réflexions importantes dans ce domaine, en proposant des concepts qui sont à la base de nos modèles de la communication et de la variation linguistique. Nous mentionnerons notamment Roman Jakobson et Eugenio Coseriu (Eugeniu Coşeriu), et pour rester plus proches des applications à l'italien, nous rappellerons l'importante contribution de Giovanni Nencioni et de Francesco Sabatini, notamment leurs considérations sur la variation de l'italien *parlato-parlato*, *parlato-scritto*, *parlato-recitato*, auxquels nous pouvons rajouter, comme d'autres ont proposé, l'italien transmis (sans doute des considérations similaires pourraient être étendues *mutatis mutandis* aux autres langues).

[24] Dans ce cadre, l'étude de la structure de tout acte de parole repose sur la notion d'énoncé, une production autonome douée d'une valeur communicative. La définition de l'énoncé comme unité fondamentale de la production linguistique se base sur un critère illocutoire qui permet de distinguer l'énoncé à un niveau perceptif dans le *continuum* de la production linguistique (voir aussi la section *Features-Prosodic tagging* sur le site *web* http://lablita.dit.unifi.it/coralrom). Aussi, sur la base de 't Hart et al. (1990), l'intonation est considérée dans ce cadre comme l'élément fondamental pour la réalisation de l'acte de parole. En effet, l'énoncé ne se concrétise pas sans un profil intonatif terminal. Le concept assume son caractère intuitif par le fait que la perception humaine est en général très sensible à la variation de la fréquence fondamentale volontaire. Sur la base de principes similaires, les énoncés peuvent être analysés ensuite en unités tonales.

[25] Le critère de définition de telles unités linguistiques repose sur la notion d'énoncé présentée ci-dessus. À partir des années 90, le laboratoire LABLITA s'est engagé dans des recherches sur des corpus de parole spontanée pour étudier les propriétés des passages de langue parlée qui présentent de tels pré-acquis. La perception commune qui voit l'énoncé comme la version orale d'une phrase (l'unité élémentaire de l'exécution en opposition à l'unité élémentaire de la compétence), déjà remise en cause dans des ouvrages classiques dans ce domaine, a été ultérieurement démentie par des hypothèses sur les contraintes et les objectifs pragmalinguistiques et par l'observation du caractère autonome assumé par l'expression linguistique grâce à l'intonation. Des énoncés définis sur ces bases mènent aussi à une meilleure définition de la notion de phrase, qui peut ainsi poser sur une autosuffisance sémantique centrée sur le concept de prédication (cf. Cresti 2005).

[26] D'autre part – comme l'a fait remarquer Moneglia lors d'une présentation de la *BD* – 42% des pauses vides ne correspond pas à une frontière prosodique.

[27] D'autres symboles utilisés sont [/], [//] et [///] pour signaler des unités prosodiques non-terminales causées par des faux départs, des bribes, des interruptions… qui néanmoins, d'après les transcripteurs, ne déterminent pas de ruptures prosodiques. La subjectivité translinguistique du positionnement de ces marqueurs a été l'objet d'une validation qui démontre un accord remarquable entre les transcripteurs (v. C-ORAL-ROM: *Appendix*; v. aussi Moneglia et al. 2002): une correspondance biunivoque est alors assumée entre la frontière d'énoncé et la présence d'un marqueur prosodique (mais il s'agit là d'une prosodie perçue qui reste acoustiquement indéfinie, avec des indices objectifs qui ne sont pas vraiment évidents). Il est très surprenant enfin que les désaccords entre transcripteurs (évidemment tous alignés sur les mêmes critères d'analyse), par exemple sur 1400 textes italiens analysés, ait concerné uniquement les frontières non-terminales, et que seulement 28 jugements ait été controversés (Moneglia et al. 2002: 7).

[28] Cette version, expressément conçue pour la gestion et l'analyse de *BD* orales, se présente complexe et efficace mais parfois peu intuitive. Par exemple, pour ouvrir un fichier .wav, il ne faut pas chercher «Open», mais «File» et puis «Sound file»; de même pour ouvrir un fichier sonore aligné avec son texte, fichier .xml, il faut suivre le chemin «File» puis «Alignment file»; pour rouvrir la fenêtre du texte de transcription pendant qu'on écoute le contenu de l'enregistrement, il ne faut pas cliquer sur le bouton «Text» mais sur «Align»; etc.). En effet, la fenêtre du texte disparaît parfois inopinément après le lancement de certaines commandes, et il faut la rouvrir continûment. Pour écouter une unité tonale délimitée dans le texte – une fonction très utile prévue par ce logiciel – il faut cliquer sur la fenêtre de texte, puis cliquer sur la séquence de caractères délimitée par les barres (/) et ensuite cliquer sur le spectrogramme (parfois le spectrogramme n'est pas mis à jour automatiquement, et il faut encore cliquer sur l'oscillogramme pour obtenir cette mise à jour).

[29] Beaucoup de ces problèmes rendent préférable de nos jours le recours à des logiciels *open source* (comme PRAAT ou d'autres), mais il est vrai que la gestion des fichiers proposée par *WinPitch Corpus* se révèle peut-être actuellement la seule adéquate à la consultation d'une *BD* dans ce format. Toutefois, après avoir attendu des mois la disponibilité d'un PC 'compatible', la première utilisation s'est avérée quand même décevante à cause du grand nombre de fonctions désactivées (notamment celle qui permet d'imprimer le texte des transcriptions associées et 'alignées' aux enregistrements). D'autre part, même si on peut comprendre les raisons commerciales qui ont poussé à ce choix, on doit mentionner la déception de l'utilisateur liée à l'impossibilité d'analyser les fichiers sonores avec d'autres logiciels (à cause de leur cryptage dans la version réduite) et les frustrations liées à l'impossibilité d'extraire en format texte les valeurs des paramètres utilisés ou des mesures effectuées.

[30] Dans bien des cas, ces enregistrements sont aussi alourdis par la présence de musique de fond (inévitable), de génériques inutiles et de longs interludes.

[31] Cela peut être la conséquence d'un rééchantillonnage à 22050 Hz de données originairement numérisées à 11025 Hz. Parmi les autres problèmes techniques à signaler, je rappellerai le dysfonctionnement d'autres programmes associés qui n'ont pas donné de résultats sur les PCs sur lesquels ils ont été testés (par exemple, le concordancier *Contextes* fonctionne régulièrement sur d'autres fichiers de texte, mais n'a jamais fonctionné sur les fichiers de la *BD*: s'agit-il de conséquences additionnelles du cryptage?).

[32] La recherche de mots avec la commande «Highlight» aurait surement bénéficié d'une option d'insertion des caractères spéciaux.

[33] Le signal original, venant d'une émission régionale de prévisions météorologiques de la chaîne nationale TV RAI, présente en plus une importante chute d'énergie au delà des 10kHz.

[34] Ces exemples sont extraits des sections “documentari”, “meteo”, “news” du corpus italien de C-ORAL-ROM (et sont discutés dans le mémoire de Licence de Fabrizio Serra, Faculté de Langues de Turin).
[35] Cela peut révéler une influence trop importante de l’écrit (mots graphiques) dans l’estimation d’un paramètre qui aurait dû se référer uniquement à des variables de la langue parlée (syllabes).
[36] Pour l’importance de ces éléments voir aussi le chapitre 6 (par ex. C-ORAL-ROM: 217). À ce propos, on peut signaler les choix discutables de prévoir dans le sous-corpus français un même *tag* pour par exemple *quoi* et *hum* (INT=*interjections and discourse particles*) et dans le sous-corpus italien un *tag* de même type pour *ah*, *mh*, *beh*, *boh*, *mah*, *magari*, *vabbè* et *ciao* (I=*interjections*) alors que par exemple, *cioè* et *ok* sont traités comme adverbes (B). Parmi les autres choix qui rendent difficile la comparaison, on peut remarquer encore que dans la liste des adverbes italiens de C-ORAL-ROM dans le tableau 2.15 ne figurent pas les adverbes en *-mente* alors qu’ils sont pris en compte dans les autres sources comparées et que les correspondants en *-ment* / *-mente* / *-mente* des autres langues ont été retenus pour les classements de fréquence des tableaux 3.13a, 4.7, 5.8.
[37] Une révision supplémentaire aurait peut-être aussi limité les nombreuses ‘coquilles’, surtout celles liées à un usage de l’apostrophe qui se révèle impropre dans certains cas (dans la section italienne on trouve par exemple: *di buon’occhio, *un’aspetto, *qualcun’altro etc.), dans d’autres cas, carrément maladroit (*va’ be’, qui s’alterne avec *vabbè*; *core n’ grato au lieu de *core ’ngrato*, là où il s’agit évidemment d’un cas d’aphérèse; voir aussi l’exemple: «quando lei va via la sera / nell’ascensore ‘un ce più luce», C-ORAL-ROM: 229). Dans la section française, les tableaux 3.11a et 3.11b présentent la forme lemmatisée *metre au lieu de *mettre*. Dans la section portugaise, le tableau 5.12 présente parmi les dix noms les plus fréquents *senior au lieu de *senhor*. Dans les attributions du chapitre 4 (*The Spanish Corpus*) et de la section *Appendix* la plupart des noms des auteurs espagnols sont affectés par des erreurs d’orthographe.
[38] À d’autres occasions, dans le corpus italien, on peut relever des substitutions malheureuses du type *Fürher au lieu de *Führer*, ou pire – puisqu’il s’agit de formes romanes – *Rochê à la place de *Roche*, *Yaoundè à la place de *Yaoundé* etc.
[39] Comme l’usage de formes erronées telles “ridution”, “haedlines”, “interwiee”, “wendsday”, “wich”, “whisles”, “refered “ etc.

Bibliographie

SAMPA (1995-1998). *SAMPA:* computer readable phonetic alphabet, *Speech Assessment Methods Phonetic Alphabet* (a cura di John C. Wells) [http://www.phon.ucl.ac.uk/home/sampa/index.html].

Bertinetto Pier Marco & Emanuela Magno Caldognetto 1993. Ritmo e intonazione. In Sobrero Alberto A. (ed.). *Introduzione all’italiano contemporaneo. Le strutture*, 2. Roma-Bari: Laterza. 141-192.

Coseriu Eugenio 1955. Determinación y entorno. Dos problemas de una lingüística del hablar. *Romanistisches Jahrbuch* 7, 1955. 29-54 (poi in Coseriu Eugenio 1962. *Teoría del lenguaje y lingüística general*. Madrid: Gredos. 282-323).

Coseriu Eugenio 1958. *Sincronía, diacronía e historia. El problema del cambio lingüístico*. Madrid: Gredos (trad. it. 1981. *Sincronia, diacronia e storia*. Torino: Boringhieri).

CRESTI Emanuela 2005. Enunciato e frase: teorie e verifiche empiriche. In BIFFI Marco et al. (a cura di). *Italia linguistica: discorsi di scritto e di parlato, Scritti in onore di Giovanni Nencioni*. Siena: Prolagon [document PDF http://lablita.dit.unifi.it/preprint/preprint-cresti_nencioni.pdf].

GIANNINI Antonella 2000. Range di variabilità della velocità di articolazione in italiano. *Atti del XXVIII Convegno Nazionale dell'Associazione Italiana di Acustica* (Trani, 10-13 Giugno 2000). 253-256.

GIANNINI Antonella & Massimo PETTORINO 1999. I cambiamenti dell'italiano radiofonico negli ultimi 50 anni: aspetti ritmico-prosodici e segmentali. In DELMONTE Rodolfo & Antonella BRISTOT (eds.). Aspetti computazionali in fonetica, linguistica e didattica delle lingue: modelli e algoritmi. *Atti delle IX Giornate di Studio del "Gruppo di Fonetica Sperimentale" dell'Associazione Italiana di Acustica* (Venezia, 17-19 Dicembre 1998). Roma: Esagrafica, 65-81.

JAKOBSON Roman 1958. *Essais de linguistique générale*. Paris: Minuit (trad. it. 1966. *Saggi di linguistica generale*, Milano: Feltrinelli).

KARCEVSKIJ Sergej 1931. Sur la phonologie de la phrase. *Travaux du Cercle de Linguistique de Prague* 4. 188-227.

MESSINA Simona 2004. Il parlato trasmesso. In ALBANO LEONI Federico et al. (ed.). Il Parlato Italiano. *Atti del Convegno Nazionale di Napoli* (13-15 Febbraio 2003). Napoli: D'Auria (CD-ROM).

MONEGLIA Massimo et al. 2002. Validation by expert transcribers of the C-ORAL-ROM prosodic tagging criteria on Italian, Spanish and Portuguese corpora of spontaneous speech. *Paper presented at the public session of the Mid-term C-ORAL-ROM Review Meeting* (Berlin, 25 September 2002) [document PDF http://lablita.dit.unifi.it/preprint/preprint-02coll09.pdf].

NENCIONI Giovanni 1976. Parlato-parlato, parlato-scritto, parlato-recitato. *Strumenti Critici*. X/1. 1-56.

NENCIONI Giovanni 1983. *Di scritto e di parlato*. Bologna: Zanichelli.

NEWMEYER Frederick 2003. Grammar is grammar and usage is usage. *Language*. 79/4. 682-797.

PETTORINO Massimo 2003. La velocità di articolazione. In DE DOMINICIS Amedeo et al. (a cura di). *Costituzione, gestione e restauro di* corpora *vocali. Atti delle XIV Giornate di Studio del "Gruppo di Fonetica Sperimentale" dell'Associazione Italiana di Acustica* (Viterbo, 4-6 Dicembre 2003). Roma: Esagrafica. 227-232.

SABATINI Francesco 1982. La comunicazione orale, scritta e trasmessa: la diversità del mezzo, della lingua e delle funzioni. In BOCCAFURNI Anna Maria & SERROMANI Simonetta (a cura di). *Educazione linguistica nella scuola superiore: sei argomenti per un curricolo*. Roma: Provincia di Roma - Istituto di Psicologia del CNR. 103-127.

SABATINI Francesco 1997. *Prove per l'italiano «trasmesso» (e auspici di un parlato serio semplice)*. In AA.VV. Gli italiani trasmessi. La radio. Firenze: Accademia della Crusca. 11-30.

SORNICOLA Rosanna 1981. *Sul parlato*. Bologna: Il Mulino.

ZMARICH Claudio et al. 1996. Analisi confrontativa di parlato spontaneo e letto: fenomeni macroprosodici e indici di fluenza". In CUTUGNO Francesco (ed.). *Fonetica e fonologia degli stili dell'italiano parlato. Atti delle VII Giornate di Studio del "Gruppo di Fonetica Sperimentale" dell'Associazione Italiana di Acustica* (Napoli, 15-16 Novembre 1996). Roma: Esagrafica. 111-139.