Méthodologie et algorithmiques pour la détection automatique des syllabes proéminentes dans les corpus de français parlé

Anne Lacheret, Mathieu Avanzi, Jean-Philippe Goldman, Anne Catherine Simon, Antoine Auchlin

Méthodologie et algorithmes pour la détection automatique des syllabes proéminentes dans les corpus de français parlé

Mathieu Avanzi[1], Jean-Philippe Goldman[2], Anne Lacheret-Dujour[3], Anne Catherine Simon[4], Antoine Auchlin[5]

Universités de Neuchâtel de Paris X Nanterre[1], Université de Genève[2], Université de Paris X Nanterre et IUF[3], UCLouvain[4], Université de Genève[5]

mathieu.avanzi@unine.ch; goldman@lettres.unige.ch; anne@lacheret.com; anne.catherine.simon@uclouvain.be; auchlin@lettres.unige.ch

1. Avant-propos

La ponctuation orthographique de l'écrit n'est pas opératoire pour transcrire l'oral (Blanche-Benveniste, 1998 ; Béguelin, 2002), aussi tout le monde est d'accord pour dire qu'il faut trouver d'autres façons pour segmenter les corpus de langue parlée en 'unités' de différents rangs, utiles pour la description des multiples niveaux de l'analyse linguistique (grammaire, structure informationnelle, analyse des interactions, etc.). Un des moyens les plus couramment convoqués pour ce faire est la prosodie :

> La transcription des corpus oraux à l'aide de la ponctuation de l'écrit est loin d'être satisfaisante, et il serait largement préférable d'en transcrire la prosodie. (Campione, 2003 : 103)

Toute transcription de la prosodie repose au minimum sur le repérage de *syllabes proéminentes* et sur l'appréciation de *degrés de frontière*. Le système ToBI (acronyme pour *Tones and Break Index*), développé dans la mouvance des travaux de Pierrehumbert (1980) sur la phonologie supra-segmentale de l'anglais américain standard, constitue le système de transcription de ce type le plus célèbre et le plus répandu à l'heure actuelle[1] :

[1] Cf. Wightman (2002) et Beckman *et al.* (2006) pour des mises au point récentes. ToBI dispose également d'une page web : http://www.ling.ohio-state.edu/~tobi/.

> Par sa notation des événements prosodiques sur une couche tonale en tons Haut et Bas, événements liés aux syllabes accentuées (accent mélodique ou *pitch accent*), aux frontières de constituants (tons de frontière ou *boundary tones* et accents de syntagme, *phrase accents*), le système ToBI se veut à la fois proche de la réalité phonétique et apparaître comme une notation phonétique à vocation universelle, tout en étant lié par sa définition même à des entités phonologiques (Martin, 2003 : 109)

Même si la transcription de ces phénomènes prosodiques donne lieu à une interprétation phonologique (consensuelle ou non), elle est effectuée manuellement par des annotateurs humains. Elle exige donc un temps de traitement considérable. De ce fait, elle demeure difficilement envisageable pour le traitement de gros corpus. En outre, par son caractère manuel, elle reste empirique, aléatoire et subjective. La variation inter-juges dans le repérage d'objets prosodiques pertinents constitue un problème majeur. D'autre part, restreindre les points prosodiques remarquables à des saillances de F0 pose un problème de fond, surtout quand on sait que dans beaucoup de langues, et tel est le cas du français, le marquage des phénomènes prosodiques est multiparamétrique par essence[2]. D'où la nécessité de développer et de mettre en œuvre des algorithmes robustes pour leur identification (semi-)automatique dans les corpus de langue parlée.

Dans cet article, nous exposons les premiers résultats d'un algorithme implémenté sous Praat (Boersma and Weenink, 2007) pour effectuer une telle tâche. À partir d'une transcription orthographique standard, notre système procède à une phonétisation, à un alignement phonétique et à un étiquetage automatique du signal en syllabes. Ensuite, une détection des *proéminences syllabiques* est réalisée sur la base d'une analyse acoustique. Cette détection automatique a été systématiquement comparée à une détection auditive et les divergences entre la détection automatique et celle des annotateurs experts (humains) a permis d'ajuster progressivement les seuils de l'algorithme.

2. L'annotation prosodique des grands corpus

L'idée de mettre au point une procédure pour faciliter l'annotation de la prosodie dans les corpus de langue orale n'est pas inédite. Plusieurs propositions ont déjà

[2] Cf. en particulier le cas de la durée, qui joue un rôle fondamental (Beller *et al.*, 2006).

été faites pour transcrire la prosodie dans les corpus de langue parlée spontanée. Nous en signalerons trois, en renvoyant également le lecteur intéressé aux états de la question dressés par Williams (1996), Delais-Roussarie (2003) et plus récemment Delais-Roussarie *et al.* (2006) ou Martin (à par.).

2.1. Corpus Gesproken Nederlands

Pour le néerlandais, une méthodologie a été imaginée par Martens (2002). Cette procédure a été mise en place dans le but de parvenir à un étiquetage minimal de la prosodie dans les textes oraux du *Corpus Gesproken Nederlands* (corpus de néerlandais parlé)[3]. Compte tenu de l'ampleur de la tâche (transcrire la prosodie dans une sous-partie du corpus équivalente à 250.000 mots) et des contraintes matérielles, techniques et temporelles, les responsables du projet ont décidé de faire appel à des annotateurs non experts pour traiter les informations de nature suprasegmentale. Ils ont ainsi été amenés à bâtir un protocole simple mais strict, que nous résumons très brièvement ci-dessous.

Après transcription et découpage du corpus en énoncés majeurs (des segments d'une durée approximative de 10 secondes au maximum, se situant entre deux grandes pauses silencieuses, *cf.* Martens *et al.* (2002) pour l'origine et la justification de ce traitement), des transcripteurs non experts annotent directement sur la transcription en mots graphiques, alignée dans le logiciel d'analyse acoustique Praat. Les annotateurs s'entraînent conjointement avant de s'atteler à la tâche. Quatre types de phénomènes sont notés :

1. Les ruptures prosodiques fortes (notation : ||) sont signalées par une prise de souffle plus ou moins audible, et la perception d'une pause d'une certaine longueur.
2. Les ruptures prosodiques faibles (symbole : |) signalent des ruptures discursives de force inférieure aux précédentes, elles signalent surtout un non-enchaînement entre deux mots.
3. Les syllabes porteuses d'un accent interne (dit de 'focalisation') sont signalées entre deux ^.

[3] http://lands.let.kun.nl/cgn/ehome.htm

4. Les allongements liés à une hésitation sont aussi mis en exergue, la syllabe allongée étant encadrée par des %.

Dans la phase d'évaluation de l'étude pilote (Buhman *et al.*, 2002), il a été montré que la qualité de ces annotations – qu'elle soit mesurée à l'aide du coefficient kappa (Cohen, 1960), en fonction du pourcentage d'accord inter-annotateur ou encore par rapport à une annotation de référence – était relativement satisfaisante, par rapport à d'autres études du même type (§ 4.1.3. *infra*).

2.2. EUROM1 ET FREF[4]

Concernant le français, une proposition avait été faite par Campione et Véronis (2001). Leur idée était que l'annotation des phénomènes prosodiques ne devait pas reposer sur des principes théoriques particuliers (comme ToBI p. ex., qui présuppose que l'inventaire des contours phonologiques de la langue étudiée ait été fait) et qu'elle ne devait pas concerner les niveaux d'analyse de bas niveau, mais seulement les informations utiles pour l'analyse de la structure syntaxique et discursive. Le traitement envisagé par Campione et Véronis comprend plusieurs étapes automatiques, qui doivent faire l'objet de vérifications manuelles intermédiaires :

1. stylisation de la F0 avec l'algorithme MOMEL, qui permet de lisser la courbe mélodique en supprimant les variations micromélodiques supposées non pertinentes ;
2. détection et typage des pauses silencieuses selon leur durée ;
3. discrétisation des mouvements mélodiques (direction du geste, niveau de hauteur du contour, ton statique/dynamique, simple ou complexe) à partir d'un modèle de codage discret, qui repose sur une simplification du système INTSINT (Hirst and Espesser, 1993) ;
4. transcription, alignement et étiquetage des proéminences et des frontières (codage phonologique, voir ci-dessous pour les symboles utilisés). A ce moment-là, sont également notées les hésitations (allongements syllabiques et 'euh') et les accents internes de mot (accents focalisateurs).

[4] Noms des deux corpus annotés par Campione dans sa thèse (2001).

Au final, le corpus transcrit en mots graphiques est aligné manuellement dans Transcriber (Barras *et al.*, 2001), selon le découpage en unités intonatives majeures qui émerge. Ces unités sont ponctuées par un type de flèche qui indique leur fonction discursive. Elles sont respectivement notées ↗, ↘ et →, selon qu'elles symbolisent une frontière prosodique non terminale, terminale ou neutre du point de vue de l'opposition +/-conclusif (parenthétique). La direction de la flèche n'indique donc pas la forme de la pente perçue. Cela veut dire, pour le formuler autrement, que ces contours phonologiques recouvrent un ensemble de réalisations phonétiques variables. Par exemple un mouvement mélodique descendant non suivi d'une pause pourra très bien être interprété comme une fin de segment continuatif (dans le cas des inversions de pente décrites par Martin (1975)) ; une intonation montante, si elle est associée à l'expression de la modalité interrogative, sera en revanche symbolisée ↘.

Les problèmes que pose cette approche, qui constituent sans doute une des raisons pour lesquelles elle n'a jamais été utilisée par d'autres, résident dans le fait que (i) elle tient compte de la dimension perceptive des phénomènes prosodiques, puisque c'est l'interprétation fonctionnelle de l'annotateur qui est en définitive conservée (voir étape 4 ci-dessus), mais sans rendre explicite le lien entre la perception et la catégorisation phonologique (Post *et al.*, 2006) et (ii) elle s'inscrit dans un cadre théorique particulier (alphabet phonologique INTSINT), ce qui est en désaccord de principe avec le fait que la transcription prosodique doit faire abstraction d'un cadre théorique particulier.

2.3. C-ORAL-ROM

Partant de l'idée selon laquelle les sujets parlants natifs identifient facilement la fin des unités prosodiques majeures dans les discours de tous les jours, les concepteurs du corpus C-ORAL-ROM ont cherché à identifier ces séquences sur la base de critères perceptifs uniquement (Cresti and Moneglia, 2005). Ce corpus constitue, à notre connaissance, le seul corpus de français parlé annoté prosodiquement jamais publié. Les ruptures prosodiques qui scandent le flux discursif sont catégorisées en deux types de frontières.

1. Les ruptures prosodiques terminales marquent des frontières d'*énoncés*. Elles indiquent le moment où peut se faire le passage du tour de parole et sont signalées par des barres obliques doubles.
2. Les ruptures non terminales actualisent des unités non autonomes contextuellement, c'est-à-dire des *constituants d'énoncés*. Elles sont signalées par une barre oblique simple.

L'ensemble du C-ORAL-ROM a été annoté sur ces bases. La détection de ces ruptures n'est pas automatique. Elle ne résulte pas d'un traitement instrumental du couplage de différents indices prosodiques, mais repose sur la seule interprétation subjective du signal auditif par des locuteurs natifs[5]. L'hypothèse selon laquelle les natifs discrimineraient aisément entre les signaux terminaux et non terminaux dans les discours spontanés a d'ailleurs été confirmée par l'évaluation d'une entreprise externe (Loquendo, Turin). Les protocoles et les résultats des expérimentations effectuées pour vérifier la pertinence du découpage perceptuel ont été consignés dans le chapitre *Appendice* de l'ouvrage, dont il ressort au final que :

> given the high scores on agreement, it is safe to say that the prosodically annotated data of the C-ORAL-ROM corpus are very trustworthy[6]

Cela dit, on peut douter de la pertinence de ce découpage, surtout quand on sait que :

1. L'annotation manuelle des phénomènes prosodiques – et la catégorisation des différents types de frontières ne constitue pas une exception – est extrêmement variable d'un auteur à l'autre : elle demeure fortement subjective. Ainsi, Pickering *et al.* (1996 : 67) signalent-ils que, entre les deux experts qui ont annoté manuellement le même sous-ensemble du

[5] The labelling is based only on perceptual judgments and in principle does not require any specific knowledge, although the notion of speech act is always familiar to the experts transcribers (comprising PhDs and PhD students) who annotated the corpus (Cresti and Moneglia, 2005 : 24).
[6] Commentaires de Marc Swerts (Université de Tilburg) à propos du « Final Report of the C-ORAL-ROM Prosodic Tagging Evaluation ». Texte accessible depuis http://lablita.dit.unifi.it/coralrom/reports.html.

Spoken English Corpus, le taux de désaccord quant à la présence d'une frontière est de 27%.

2. Dans le C-ORAL-ROM, la notion de 'rupture prosodique terminale' (*terminal prosodic break*), comme son nom ne l'indique pas, n'est pas définie selon des propriétés prosodiques *stricto sensu*, mais mêle des considérations résultant de la prise en considération d'informations relevant de plusieurs niveaux d'analyse. Selon la théorie de la *lingua in atto* développée par Cresti (2000), un acte de langage = un énoncé = une frontière prosodique terminale[7]. Or, on peut douter du bien-fondé d'un tel isomorphisme. En effet, l'accomplissement d'un 'acte de langage' ne va pas de pair avec une rupture prosodique terminale, de même, inversement, la présence d'une rupture prosodique terminale ne veut pas dire qu'un acte de langage a été performé, comme le montre bien Berrendonner (à par.).
3. L'opposition +/- terminal et ses avatars (+/- conclusif, +/- final, etc.) n'est pas opératoire pour définir les unités prosodiques d'intégration maximales. Le fait qu'un segment soit considéré comme terminal ou non terminal, outre qu'il s'agisse d'une variable extrêmement subjective, n'est pas déterminant pour repérer les ruptures prosodiques pertinentes, dans la mesure où une frontière prosodique majeure, *i. e.* d'intégration maximale, peut être indifféremment continuative ou conclusive. *Cf.* Avanzi et Martin (2007) pour les détails de la démonstration.

2.4. Notre parti pris méthodologique

La critique de la méthodologie des autres essais dans le même domaine sert non seulement de mise en relief de la propre façon de faire. Elle sert aussi à ouvrir ou continuer le dialogue, toujours prometteur dans notre travail. Dans ce contexte, notre démarche s'articule autour des points suivants :

- d'abord, elle s'inscrit dans une perspective phénoménologique en s'appuyant sur des données perceptives et en cherchant dans un second

[7] The rough equivalence between utterance and textual string ending with a terminal break is based on the idea that the performance of language actions is necessarily correlated to prosody, which constitutes the interface between the accomplishment of illocutionary and locutionary acts. (…) The perception of the property 'terminal' in a prosodic break seems to be at least in correlation with this: we perceive the break as terminal because an act has been accomplished (Cresti and Moneglia, 2005: 16).

temps à automatiser la détection des phénomènes tels qu'ils sont été annotés par les experts ;

- ensuite, elle refuse tout a priori théorique (théorie de l'accentuation, symboles intonatifs, etc.) : il s'agit de la clé de voûte méthodologique de notre démarche ;
- enfin l'annotation fournit les primitives prosodiques nécessaires et suffisantes pour permettre une analyse qualitative et quantitative exhaustive des corrélats prosodiques associés aux structures et processus linguistiques analysés (fréquence fondamentale, intensité, durée).

3. Matériel

Le corpus sur lequel se base cette étude est constitué de deux types d'enregistrements : des prescriptions d'itinéraires (cote Iti) *in situ*, recueillis à micro ouvert dans la région grenobloise (Avanzi, 2004) ; des interviews radiophoniques des radios publiques française et belge (cote Irt), captées analogiquement puis numérisées (Simon, 2004)), à dominante monologale. Les locuteurs sont aussi bien des hommes que des femmes. Les transcriptions en orthographe standard ont été phonétisées puis alignées semi-automatiquement sur les sons à l'aide du software *EasyAlign* (Goldman, 2007), qui fonctionne sous Praat.

Cet outil construit une annotation multi-niveaux (phonèmes, syllabes, mots et pseudo-phrases) à partir d'un son et de sa transcription orthographique (ou phonétique) par une succession d'étapes automatiques et manuelles. Les avantages de cette approche semi-automatique sont multiples : en premier lieu le temps de traitement est largement inférieur à celui d'une approche entièrement manuelle (on estime le temps nécessaire à 5 fois la durée du corpus, c'est-à-dire 50 minutes avec EasyAlign pour 10 minutes de corpus, alors que certaines études estiment le temps d'alignement manuel à plusieurs dizaines de fois la durée du corpus) ; de plus cette approche apporte une plus grande consistance (car l'automate applique la même technique à l'ensemble du corpus alors que la précision de l'annotateur humain qui peut varier au cours du corpus, et on notera d'autant plus de différences si le corpus doit être réparti, du fait de sa taille, entre plusieurs annotateurs), finalement on obtient un résultat reproductible. Au final, la

précision des positions de frontières segmentales est comparable à celle des annotateurs humains : plus de 80% des frontières sont distantes de moins de 20ms par rapport à un corpus aligné manuellement, c'est aussi la précision que l'on note entre deux alignements humains indépendants. Autrement dit, l'imprécision de notre outil par rapport à un alignement humain est comparable aux différences notées entre deux annotateurs humains.

La durée totale du corpus est d'environ 20 minutes (1213 secondes). Le contenu détaillé est donné dans le tableau 1 :

Tableau 1 **: Contenu détaillé du corpus d'étude avec, pour chacun des échantillons des deux sous-corpus, la durée en secondes et le nombre de phonèmes, de syllabes et de mots graphiques.**

Situation de parole	Sexe	Nom du corpus	Durée (sec.)	Nb. Phonèmes	Nb. Syllabes	Nb. Mots
Demandes d'itinéraires	M	Iti-10	50	396	181	137
	M	Iti-12	46	321	148	110
	F	Iti-14	100	948	430	342
	F	Iti-22	203	1737	820	658
	M	Iti-B	27	298	128	100
	M	Iti-D	128	983	436	375
	M	Iti-S	33	312	140	114
		sous-total	*587*	*4995*	*2283*	*1836*
Interview radio	M	Irt-LF1r	331	3151	1403	1062
	F	Irt-WL1r	295	2723	1195	996
		sous-total	*626*	*5874*	*2598*	*2058*
		Total	**1213**	**10869**	**4881**	**3894**

Le traitement du corpus s'est opéré en deux temps :

- Deux experts ont détecté auditivement les syllabes proéminentes ainsi que certaines marques du travail de formulation (hésitations, faux-départ…) selon une méthodologie contraignante (§ 4) et à l'aide d'une liste fermée de symboles ; les annotations discordantes ont été identifiées et les experts

ont dû prendre une décision commune pour chaque désaccord, afin de constituer un corpus de référence annoté pour les proéminences syllabiques.

- Une détection automatique des syllabes proéminentes a été effectuée sur le corpus aligné phonétiquement (§ 5). La suite du traitement consistera en une comparaison systématique de la détection manuelle et de la détection automatique.

4. Détection auditive des proéminences

L'identification, auditive et automatique, des syllabes proéminentes constitue le second volet de notre étude. L'identification des proéminences est à la base de n'importe quelle analyse ou modélisation prosodique, quelle que soit l'interprétation donnée aux phénomènes identifiés comme proéminents[8].

Or, la détection auditive des proéminences, même lorsqu'elle est réalisée par des phonéticiens experts, est loin de donner des résultats fiables, comme l'a encore démontré la récente étude de Poiré (2006). En outre, si la proéminence est par définition un phénomène perceptif, on sait que la perception est largement informée par les connaissances, linguistiques en l'occurrence, de celui qui perçoit (Martin, 2006).

En développant un système semi-automatique de détection des proéminences syllabiques, nous avons voulu éviter de nous inscrire dans une approche trop théorique et/ou trop naïve :

- une approche trop théorique ferait dépendre l'identification des proéminences d'un modèle phonologique de l'accentuation ou de l'intonation du français fondé sur une pré-identification des syllabes accentuables, et donc potentiellement proéminentes ;

[8] Par exemple, le système de codage IViE pour décrire la variation prosodique régionale dans différentes variétés d'anglais (Grabe and Post, 2002 ; adapté au français par Post, Delais-Roussarie and Simon, 2006) établit le domaine de codage à partir des syllabes détectées auditivement comme proéminentes. N'importe quel système de transcription symbolique (ToBI ; le modèle tonal de Mertens (1987) ; l'alphabet INTSINT de l'équipe d'Aix-en-Provence (Hirst and Espesser, *loc. cit.*)) constitue la mise en relation de symboles avec des phénomènes proéminents, que ceux-ci soient détectés auditivement (ToBI, Mertens) ou automatiquement (INTSINT), avec ou sans information linguistique de type morphosyntaxique.

- une approche naïve sous-estimerait l'importance des spécificités du français parlé spontané (hésitations, interruptions…) et l'influence que ces marques du travail de formulation ont sur l'identification des proéminences[9].

Nous avons cherché à élaborer la meilleure méthodologie possible pour la détection auditive des proéminences par des experts. Sur la base de l'expertise de cette annotation, nous avons développé un algorithme de détection semi-automatique (voir § 5) destiné à faciliter l'annotation de grands corpus et à la rendre plus fiable, pour servir de point de départ à d'autres études. Nous avons essayé de garder l'apport d'informations linguistiques au degré minimal.

4.1. Méthodologie pour le codage manuel

Des études antérieures sur le codage prosodique dans le projet PFC[10], qui sont en partie à l'origine de ce travail, il faut tirer un certain nombre de leçons, afin d'éviter d'être confronté aux mêmes difficultés que celles rencontrées par Poiré (2006).

4.1.1. Principes

Avant toute chose, il faut définir le plus précisément possible ce qu'on entend par *proéminence*. Les ouvrages de référence ne fournissent pas toujours des définitions opérationnalisables :

> PROEMINENCE : mise en valeur perceptive d'une syllabe, qui se manifeste par la perception d'un accent
>
> ACCENT : sur le plan perceptif, il s'agit d'un élément qui correspond à une proéminence ; sur le plan acoustique, trois paramètres phonétiques peuvent le manifester : une montée mélodique marquée et/ou un allongement de durée pour l'accent primaire, une montée mélodique, voir une augmentation de l'intensité pour les accents secondaires (Lacheret et Beaugendre, 1999 : glossaire)

La circularité des définitions met en évidence le danger de définir *a priori* la localisation d'une proéminence (sur certaines syllabes et pas sur d'autres) ou son

[9] Par exemple, la manière dont un codeur perçoit une syllabe allongée de type *euh* varie très fortement selon le statut théorique (accentuable ou non…) qu'il donne à une particule d'hésitation.
[10] Cf. Poiré (2005, 2006), (Martin, 2006), Morel *et al.* (2006) et Goldman *et al.* (2006).

statut théorique (un certain type d'accent[11]). Dans l'étude de Poiré, une telle définition *a priori* restreignait le nombre de syllabes susceptibles d'être codées:

> (i) il existe un consensus assez général quant à la distribution des syllabes normalement accentuées en français (et donc proéminentes), à savoir la dernière syllabe d'un mot plein ; en conséquence, toutes les syllabes toniques doivent être codées ; (ii) une syllabe non tonique mais proéminente (emphase, accent secondaire, allongement associé à l'hésitation) devrait facilement être remarquée et donc codée (Poiré, 2006 : 70)

Selon nous, identifier les proéminences ne présuppose pas que l'on sache quelles fonctions elles ont, ni quel est leur type, et encore moins quelles sont leurs localisations possibles dans la chaîne parlée.

La seule hypothèse externe que nous ayons utilisée a été la **syllabe**. Le phénomène de proéminence peut en théorie affecter n'importe quel 'empan' du signal de parole, puisque 'prominence is the property by which linguistic units are perceived as standing out from their environment' (Terken, 1991). En nous basant sur les travaux de perception, nous avons considéré que chaque syllabe était susceptible de recevoir une proéminence :

> Les travaux de House (1990) sur la perception des variations mélodiques dans la parole montrent qu'une même variation du fondamental sera perçue différemment selon sa place par rapport aux frontières syllabiques. Si elle apparaît au cours de la voyelle, la variation sera audible compte tenu du seuil de glissando. Si elle est située en partie pendant la transition à la frontière syllabique, seule la partie sur la voyelle sera bien intégrée auditivement. Tout semble indiquer que les changements simultanés d'intensité, de spectre et de voisement entravent l'intégration perceptive des variations mélodiques. Le phénomène est d'autant plus prononcé que les changements acoustiques sont importants. Ceci donne lieu à la *segmentation du continuum mélodique* en chaînons correspondant aux noyaux syllabiques. (Mertens, 2004 : 117)

La deuxième leçon à tirer des expériences passées est qu'il faut définir un 'domaine' temporel au sein duquel on identifie les proéminences. La proéminence est hautement relative puisqu'elle concerne la perception d'un élément comme se détachant des éléments qui l'entourent. Si l'on écoute trois syllabes successives, il

[11] D'autant qu'il existe autant de types d'accents que de modèle de l'accentuation. Lacheret et Beaugendre (1999) en recensent 22 types (avec des regroupements) : 'Accent affectif, d'insistance, de focalisation, de groupe, de mot, de phrase, didactique, émotionnel, emphatique, énonciatif, externe, intellectif, interne, lexical, logique, objectif, oratoire, primaire, rhétorique, secondaire, rythmique, tonique'.

y en a forcément une qui se détachera des autres ; en revanche, si l'on écoute un segment long contenant plusieurs dizaines de syllabes, la perception sera adaptée à l'empan d'écoute et un segment devra ressortir avec plus de force pour être perçu comme se détachant de son environnement élargi. La consigne 'écoutez et codez ce que vous percevez comme proéminent' (Poiré, 2006 : 71) n'est donc pas suffisamment précise. Il faut d'une part définir la longueur moyenne des segments à écouter, et d'autre part limiter le nombre d'écoutes – car plus on écoute un extrait, plus sa structure interne est perçue avec précision. Cette double restriction (domaine et nombre d'écoutes, voir § 4.1.2) devrait permettre d'éviter que 'certaines propositions de codage contiennent des énoncés complets sans aucune proéminence notée' (Poiré, 2006 : 78).

Enfin, dernier point, il faut faire intervenir un minimum d'information linguistique. Nous avons tenté d'établir une position raisonnable entre la posture qui refuse toute forme d'information symbolique et cherche à faire remonter la modélisation de la pure substance (variations de F0, variation d'intensité ou silence[12]) et la posture qui cherche à établir les corrélats acoustiques d'objets symboliques définis *a priori* (p. ex., la durée moyenne des syllabes en position finale de lexèmes dits 'pleins').

L'identification préalable de syllabes réalisant un *euh* d'hésitation, p. ex., évite de grandes disparités dans le traitement qui en est fait par les codeurs (voir encore sur ce point Poiré (2006 : 70-71) qui remarque qu'un seul expert parmi 7 codeurs a systématiquement identifié les *euh* très allongés comme des proéminences). Ce type d'information est également crucial pour l'exploitation ultérieure du codage en vue d'analyses acoustiques : il est impossible de corréler les durées aux proéminences si on n'analyse pas certaines syllabes comme des hésitations (Morel *et al.*, 2006 : 185). Nous suivons donc Poiré lorsqu'il remarque que 'le fait qu'il s'agisse de langue spontanée mérite aussi que certains phénomènes reçoivent une attention particulière' (2006 : 79).

[12] Si l'on va jusqu'à refuser l'alignement syllabique ou phonématique pour analyser la prosodie, on s'empêche de faire intervenir le paramètre de la durée puisque l'allongement concerne une syllabe, ou une voyelle…

4.1.2. Consignes de codage

La notion de proéminence ne doit pas être confondue avec celle d'accent, trop polysémique et trop empreinte de considérations théoriques. Elle doit être neutre, et se rapprocher de la définition qu'en donne Mertens[13] :

> A syllable is prominent when it stands out from its context due to a local difference for some prosodic parameter. Prominence is continuous (not categorical) and contributions of multiple parameters can interact (1991: 218).

La procédure d'annotation manuelle des proéminences prosodiques dans les corpus oraux [14] doit circonscrire un domaine temporel limité, et spécifier le nombre d'écoutes.

- À partir de notre expérience, nous avons retenu des segments d'une durée moyenne de 3,5 secondes que le codeur écoute à 3 reprises. Cette durée permet au codeur de répéter mentalement ou à voix haute le fragment de discours et, par cette répétition, d'identifier les syllabes proéminentes.

L'identification des proéminences, si on veut l'automatiser, doit tenir compte de certaines spécificités de la langue parlée informelle (hésitations, faux départs, changements de locuteurs, schwas post-toniques). Deux ensembles de symboles sont utilisés pour le codage :

Tableau 2 : Catégories de symboles pour le codage manuel des phénomènes prosodiques

1. Codage des proéminences	
P	proéminence forte
p	proéminence faible
0	syllabe non proéminente

[13] Voir aussi Crystal (2003 : 375), cité par Poiré (2006: 70) : '**prominent** (adj.) A term used in AUDITORY PHONETICS to refer to the degree to which a sound or SYLLABLE stands out from others in its ENVIRONMENT. Variation in LENGHT, PITCH, STRESS and inherent SONORITY are all factors which contribute to the relative **prominence** of a UNIT'.
[14] L'annotation peut prendre 5 fois le temps réel du corpus à coder, une fois que celui-ci est aligné en syllabes.

2. Codage des phénomènes de *delivery*	
z	hésitation (allongement, creaky voice)
@	schwa post-tonique (comme dans *c'est dingue* (sE de~g@))
$	appendice (p. ex. *c'est dingue* quoi)
!	interruption de mot ou de syntagme
%	partie inaudible ou inexploitable (rire, toux, chevauchement, bruit…)
*	prises de souffle
_	silence (issu de la détection de l'alignement automatique)
3. Combinaison des symboles	
z!	hésitation dans un mot interrompu
$!	appendice interrompu
p!	proéminence dans un mot ou syntagme interrompu

- L'étiquetage des **proéminences** proprement dites se fait avec les symboles 'P', 'p' et '0'. La distinction entre 'proéminence forte' et 'proéminence faible' a une fonction heuristique : elle force à développer une écoute plus fine. Lors de l'analyse, les syllabes 'P' et 'p' sont regroupées en une catégorie, qui s'oppose à '0'.
    - Nous avons proscrit l'utilisation d'un symbole de type '?' destiné à coder les proéminences incertaines. En effet, le traitement de ce type de catégorie pose toujours problème, qu'on décide finalement de le regrouper avec les syllabes non proéminentes, ou avec les syllabes dites faiblement proéminentes. Les cas de désaccord entre codeurs ('p' vs '0', par exemple) étaient tranchés lors d'une discussion (voir ci-dessous).
- L'étiquetage de phénomènes typiques de la production d'oral spontané fait l'objet d'une tire d'exception (***delivery***) : elle permet d'isoler des syllabes que chaque codeur, selon ses a priori théoriques, pourrait coder d'une manière propre, engendrant par là des divergences ou des incohérences quant à la perception des proéminences. Ces syllabes (hésitations,

interruptions, schwas post-toniques) peuvent faire l'objet d'un traitement spécifique ultérieur.

- Le symbole 'z' est attribué à chaque syllabe allongée et prononcée sur un ton plat, éventuellement avec une *creaky voice*, et qui correspond à ce qui est perçu comme une hésitation (qu'il s'agisse d'un *euh* ou d'une autre syllabe).
- Le symbole '@' identifie les schwas réalisés en position post-tonique, à cause de leur statut particulier par rapport au placement des frontières intonatives en français: une syllabe proéminente finale de mot produit un effet de frontière; le schwa post-tonique est inclus dans une telle unité intonative.
- Le même type de raison justifie l'annotation des syllabes atones en appendice par le symbole '$' (souvent des *ponctuants* comme *quoi*, *hein*, etc.)
- Les syllabes des mots ou de syntagmes interrompus (par exemple *il sav/ euh il savait la vérité*) sont identifiées par '!'. Ceci afin de permettre ultérieurement de calculer la longueur des unités intonatives délimitées par les proéminences en les incluant ou en les excluant. Une syllabe qui fait partie d'une interruption peut être proéminente ou non proéminente.
- Les parties impossibles à exploiter acoustiquement, et parfois même à segmenter en syllabes (rires, toux, chevauchements, etc.) sont notées '%' et exclues.
- Les silences '_' sont détectés automatiquement lors de l'alignement ; les prises de souffle '*' sont notées manuellement.
- Certains symboles peuvent se combiner (voir tableau 2).

- Enfin, une tire permet d'indiquer l'identité du **locuteur**, en vue de pouvoir séparer les données de chaque locuteur d'une conversation. Les parties en chevauchement sont également identifiées dans cette couche.

4.1.3. Résultats de la détection auditive des proéminences

Partant de là, les deux experts codeurs (H1 et H2) se sont entraînés conjointement au cours de sessions communes sur les corpus iti-B et iti-D. Ils ont ensuite annoté

le reste du corpus chacun de leur côté. Les résultats de cette première série d'annotations se présentent comme suit :

Tableau 3 : Taux d'accord entre les deux experts annotateurs

				Agrément H1 – H2		
Corpus	Durée	Total syll.	Total syll. 0p/P	Total désaccord syllabes	Moyenne brute	Moyenne pondérée
Irt-LF1r	331	1403	1279	136	89,37	
Irt-WL1r	295	1195	1054	167	84,16	
Iti-10	50	181	161	11	93,17	
Iti-12	46	148	124	6	95,17	
Iti-14	100	430	365	39	89,32	
Iti-22	203	820	747	44	94,11	
Iti-S	33	140	126	8	93,66	
Total	**1058**	**4317**	**3856**	**411**	**91,28**	**89,35**

Ici encore, nous avons rappelé pour chaque corpus sa durée et le nombre total de syllabes qui le compose. Le taux d'accord entre H1 et H2 a été calculé par rapport à l'ensemble des syllabes codées p/P ou 0 – *i.e.* par rapport à l'ensemble des syllabes auxquelles aucun marqueur de *delivery* n'a été assigné – après discussion des désaccords (*cf.* tableau 4). Les meilleurs résultats sont pour iti-12 (95,17%), les moins bons pour irt-WL1r (84,16%). Après pondération, le pourcentage d'agrément inter-annotateur atteint 89,35 %. On peut considérer que ce score comme vraiment satisfaisant, puisque d'après les recensions faites par (Tamburini and Caini, 2005 : 43), le taux d'accord entre deux experts dans l'identification des proéminences prosodiques oscille généralement entre 80 et 84 %, dans le meilleur des cas. Comme Buhman *et al.* (2002) (*cf.* § 2.1), nous pensons que ces bonnes performances s'expliquent en grande partie grâce aux étapes préliminaires au codage des proéminences à proprement parler (entre autres du fait de la mise en place et de l'apprentissage du protocole, aussi de l'entraînement conjoint).

4.2. Traitement des désaccords : les règles qui émergent

Nous aurions pu en rester là mais notre objectif – l'élaboration d'un corpus de référence annoté pour les proéminences en vue d'automatiser cette détection – nous a conduit à régler les cas de désaccord d'une manière cohérente et fiable, c'est-à-dire en formulant des règles explicites.

4.2.1. Désaccords sur la perception d'une proéminence

Les désaccords peuvent porter sur la perception d'une syllabe comme proéminente (p ou P) ou non proéminente (0). Par exemple l'enregistrement Irt-WL1, qui contenait le pourcentage de désaccord le plus élevé (voir tableau 3), présentait 1195 syllabes, dont 1054 syllabes à annoter pour la proéminence (ne contenant ni hésitation, ni faux départ, etc.) :

- 20% des syllabes ont été identifiées comme proéminentes par les deux codeurs;
- 72,2% ont été identifiées comme non proéminentes par les deux codeurs;
- 7,8% (soit 83 syllabes sur 1054) ont fait l'objet d'un codage divergent.

C'est sur les divergences qu'a porté la discussion pour arriver à un codage unifié. S'il est rare qu'un codeur identifie une syllabe fortement proéminente quand l'autre n'identifie aucune proéminence, il est en revanche fréquent qu'une syllabe soit perçue comme **faiblement proéminente** par l'un et comme **non proéminente** par l'autre. Dans ces cas, une réécoute conjointe est effectuée pour trancher.

Dans l'énoncé *donc là vous partez dans <u>cette</u> direction* un désaccord a été constaté pour le codage de la syllabe *cette*, perçue comme faiblement vs non proéminente. La réécoute a permis de prendre la décision en faveur du symbole 'p' (faiblement proéminente).

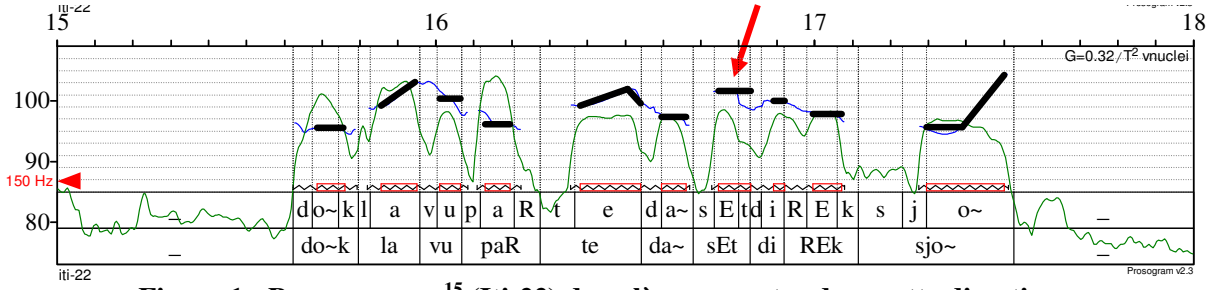

Figure 1 : Prosogramme[15] (Iti-22) donc là vous partez dans cette direction

Un type de désaccord intéressant se produit quand chaque codeur identifie comme proéminente une **syllabe différente d'un même mot di- ou polysyllabique**. Dans *on est loin du minimalisme de la poésie en général*, la première syllabe de *minimalisme* a été codée comme proéminente par un codeur, tandis que l'autre a codé la deuxième syllabe comme proéminente. Le Prosogramme montre un mouvement global de montée de F0 réparti sur plusieurs syllabes, comme si 'l'énergie' était diffusée sur plusieurs syllabes successives.

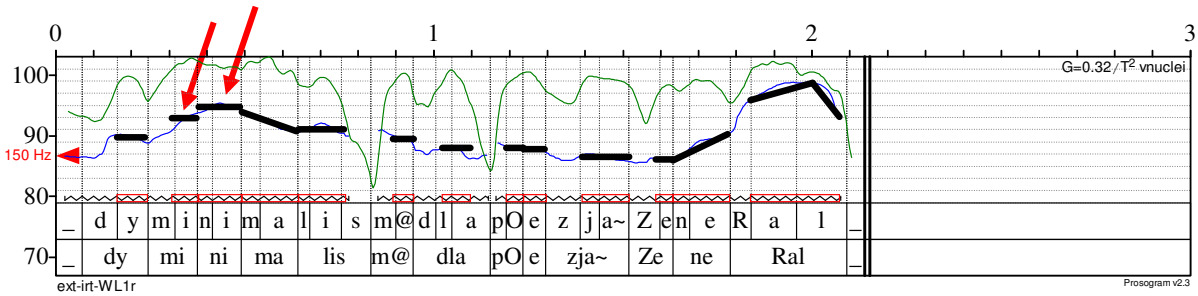

Figure 2 : Prosogramme (IrtWL1r) du minimalisme de la poésie en général

[15] Le Prosogramme est un script Praat élaboré par P. Mertens. Il est présenté en détail dans la section 5.1.

Un autre cas de **bascule** dans la perception de proéminence par les deux codeurs se produit sur les deux syllabes de l'adjectif *premier* dans l'énoncé suivant : *le premier poème du recueil est un long poème*, issu du même enregistrement.

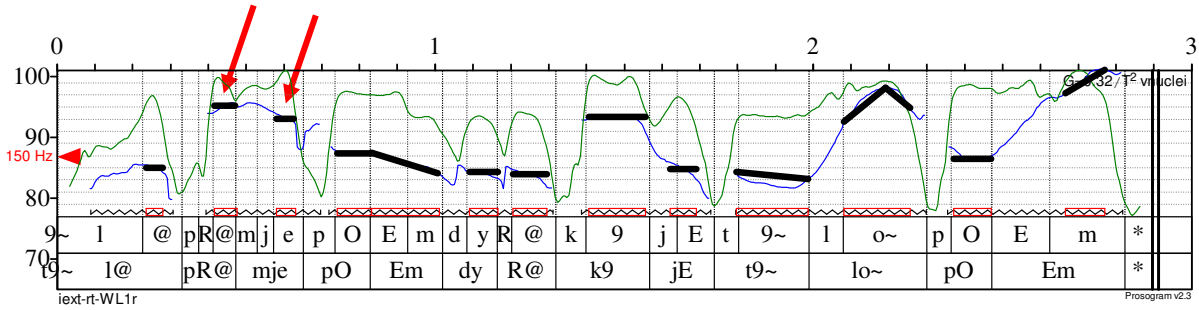

Figure 3: Prosogramme (IrtWL1r) le premier poème du recueil est un long poème

Dans *alors là c'est pas compliqué*, le désaccord a porté sur deux syllabes contiguës mais appartenant à deux mots graphiques successifs (l'adverbe *pas* et la première syllabe de l'adjectif *compliqué*). À la réécoute, les deux codeurs se sont accordés pour coder les deux syllabes successives comme proéminentes.

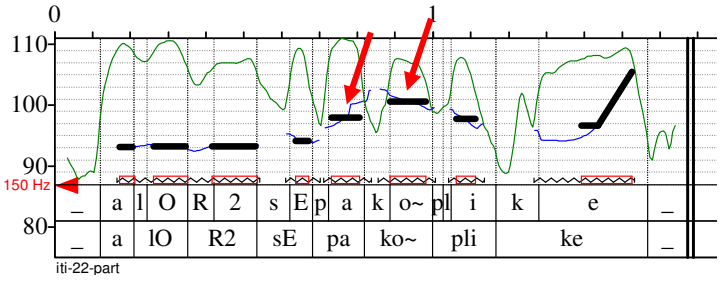

Figure 4 : Prosogramme (Iti-22) alors euh c'est pas compliqué

La réécoute conjointe permet de trancher. Aucun *a priori* théorique (comme par exemple une règle interdisant le clash accentuel) n'a empêché les experts à coder deux syllabes successives comme proéminentes, si telle était leur perception.

4.2.2. Désaccords concernant l'usage des autres symboles

Une autre série de désaccords provient de ce qu'un codeur perçoit une syllabe comme proéminente, tandis que l'autre y identifie une marque typique de formulation de l'oral (comme une hésitation)[16]. Ce cas fut surtout observé dans deux contextes. Les connecteurs (de type *et* ou *mais*) réalisés avec un certain allongement mais sans la *creaky voice* typique de l'hésitation, comme à l'initiale de l'énoncé : *et si vous écrivez ce qui s'est passé.*

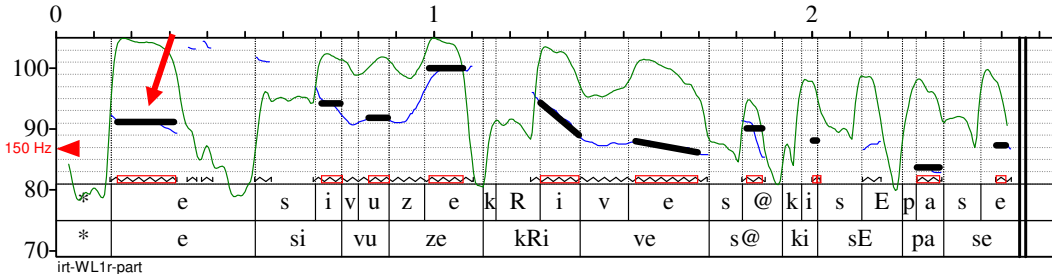

Figure 5: Prosogramme (IrtWL1r) et si vous écrivez

Soit pour une syllabe fortement allongée en fin de groupe, l'allongement pouvant être interprété comme un corrélat acoustique de la proéminence ou comme la marque d'une hésitation quant à la suite du discours. Les codeurs ont divergé sur le statut à accorder à la dernière syllabe du verbe *passer* dans l'énoncé *vous allez passer devant la poste*.

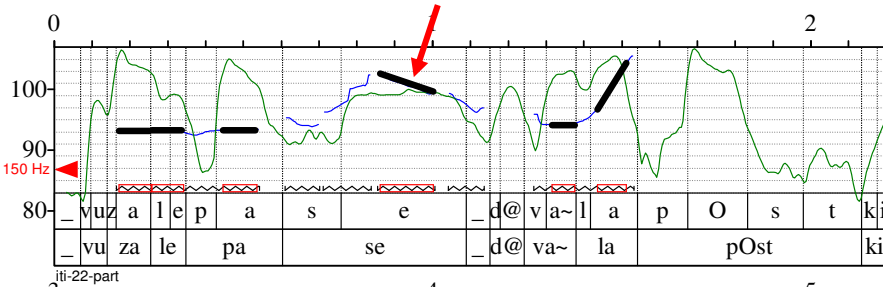

Figure 6 : Prosogramme (Iti-22) vous allez passer devant la poste

Chaque divergence a été résolue de manière à obtenir un codage unifié, que nous considérons comme notre codage de référence pour les proéminences. Il présente les caractéristiques décrites à la section suivante.

[16] Rappelons qu'une syllabe peut être codée à la fois comme proéminente et appartenant à une amorce de constituant syntaxique, mais qu'une syllabe ne peut pas être codée comme à la fois proéminente et faisant partie d'une hésitation. Ceci constitue une des rares, sinon la seule, contraintes 'théoriques' de notre procédure.

4.3. Proportion de syllabes proéminentes sur le corpus final

Comme il a été précisé plus haut, les experts ont discuté et réglé les divergences de codage au cours de discussions communes. Il en ressort une annotation consensuelle, dont le tableau 4 donne les résultats :

Tableau 4 : statistiques finales de l'annotation manuelle sur le corpus d'étude. De bas en haut : durée des corpus et nombre de syllabes total ; syllabes exclues (marquées par un signe de delivery) – avec pourcentage par rapport à la totalité des syllabes ; syllabes proéminentes, syllabes non proéminentes – avec pourcentage par rapport à la totalité des syllabes ; total de syllabes 'valides', i.e. ensemble par rapport auquel doit se mesurer l'automate

	Irt-LF1r	Irt-WL1r	Iti-10	Iti-12	Iti-14	Iti-22	Iti-B	Iti-D	Iti-S	**Total**
Durée (sec)	331	295	50	46	100	203	27	128	33	**1213**
Total syll.	1403	1195	181	148	430	820	128	436	140	**4881**
Delivery	124	141	20	24	65	73	12	37	14	**509**
%	8,83	11,79	11,04	16,21	15,11	8,90	9,37	8,48	10	**10,42**
P/p	333	314	35	42	104	192	27	106	30	**1182**
%	23,73	26,27	18,33	28,37	24,18	23,41	21,09	24,31	21,42	**24,21**
Non P/p	946	740	126	82	261	555	89	293	96	**3190**
%	67,42	61,92	69,61	55,40	60,69	67,68	69,53	67,20	68,57	**65,35**
Total syllabes valides	**1279**	**1054**	**161**	**124**	**365**	**747**	**116**	**399**	**126**	**4372**

Au final, sur un total de 4881 intervalles syllabiques, 509 syllabes ont été exclues via la tire d'exception (10,42%), 1182 syllabes ont été codées p ou P (soit 24,21%). Restent 3190 syllabes non proéminentes (65,35 %). L'outil de détection

automatique mis en place prendra comme mesure de référence le nombre total de syllabes non exclues, qui s'élève à 4372 unités.

5. Détection automatique des proéminences

La méthode de détection auditive des proéminences a donné lieu à la constitution d'un corpus annoté à partir duquel une méthode de détection semi-automatique a été entreprise. Notre proposition pour combiner l'approche auditive et les outils d'analyse acoustique ne consiste donc pas à 'afficher la courbe de F0' pendant l'expertise manuelle[17]. En fait, les résultats d'une identification auditive experte servent de référence pour l'entraînement et la validation d'un système automatique de détection de syllabes proéminentes, basé sur des paramètres prosodiques 'classiques' comme la mélodie, l'intensité et la durée. La procédure se décompose en deux moments : (i) identification des noyaux vocaliques ; (ii) computation des paramètres acoustiques prosodiques pour la détection de proéminences.

5.1. Détection de noyaux vocaliques

La première étape prend pour point de départ un script élaboré par P. Mertens (2004) fonctionnant avec le logiciel Praat : le Prosogramme[18]. A l'origine, cet outil avait été développé pour faciliter la transcription semi-automatique de la prosodie, en opérant à une stylisation de la mélodie. Cette stylisation peut être faite à partir du signal seul, mais elle est plus robuste si un alignement phonétique est fourni. Pour chaque syllabe, le noyau vocalique est délimité comme la partie voisée qui présente une intensité suffisante (en utilisant des seuils relatifs au pic d'intensité local). Puis, pour chaque noyau, la F0 est stylisée en un ou plusieurs segments de droite. Ces segments peuvent être stylisés comme plat ou avec une pente mélodique, selon des seuils perceptuels de glissando qui sont réglables[19].

---

[17] Cette suggestion est régulièrement mentionnée dans les publications sur la détection de proéminences. Outre qu'elle suggère erronément au codeur de s'aider de ses yeux pour repéré un phénomène supposément auditif, elle présente l'inconvénient de corréler prioritairement la proéminence au mouvement mélodique, alors que la durée, l'intensité ou l'énergie articulatoire sont également en jeu.

[18] http://bach.arts.kuleuven.be/pmertens/prosogram/

[19] L'idée sur laquelle repose ce script est qu'un changement de F0 peut être perçu comme un ton statique ou dynamique selon la vitesse de variation mélodique au cours du temps. Le seuil de glissando qui permet de déterminer s'il s'agit d'une cible statique ou d'un mouvement de F0 a été établi à partir des mesures de Rossi (1978a et b).

Malheureusement, dans nos corpus, un nombre non négligeable de noyaux n'a pas été détecté ou mal stylisé. Les raisons en sont diverses.

- En premier lieu, la version originale du Prosogramme a été développée pour styliser de la parole avec ou sans l'aide d'une segmentation phonétique préalable. De ce fait, certains seuils non réglables par l'utilisateur ont été ajustés de la même manière pour les deux modes de fonctionnement.
- La seconde source d'erreurs concerne les frontières de phonèmes. Celles-ci sont détectées par la segmentation semi-automatique avec EasyAlign (*cf.* supra, § 3.) mais il est possible qu'elles puissent ne pas être tout à fait exactes.
- Enfin, le paramètre d'intensité utilisé pour la segmentation des noyaux vocaliques n'est pas entièrement fiable, parce qu'il est à la fois instable et dépendant de la nature des segments. En effet dans la version originale du Prosogramme, si le pic d'intensité a lieu dans une partie non voisée de la consonne, aucun noyau n'est détecté.

Pour pallier ces problèmes, et, afin qu'un maximum de noyaux soient stylisés, nous avons apporté de légères modifications à la version originale du Prosogramme.

- Si le pic d'intensité, utilisé comme 'base de construction du noyau syllabique', est recherché dans la voyelle (délimitée par la segmentation phonétique) et qu'il n'y est pas trouvé, on autorise la stylisation à s'étendre à l'attaque et/ou à la coda de la syllabe complète, tout en maintenant la contrainte de voisement.
- Des systèmes de routines de *back-off* ont été implémentés pour forcer la détection du noyau.
- Nous avons rendu réglables certains paramètres comme les seuils d'intensité pour la segmentation du noyau[20].

[20] A l'origine, les mesures sont de -3 dB à gauche et -9 dB à droite par rapport au pic d'intensité local (Mertens 2004).

Ces modifications se sont avérées fort utiles, puisqu'au final, la proportion de noyaux stylisés est passée de 85% à 95%.

5.2. Mesures et paramètres acoustiques retenus pour la détection de proéminences

Une fois les noyaux détectés, nous avons comptabilisé pour chacun des intervalles syllabiques les mesures acoustiques suivantes :

- La durée de la syllabe (en millisecondes), que l'on préfère intuitivement à la durée du noyau syllabique stylisé, laquelle est dépendante des traits de voisement des consonnes contenues dans l'onset et la coda ;
- Le maximum de F0 atteint sur le noyau (en semi-tons)

Ces premières mesures acoustiques sont ensuite 'relativisées', c'est-à-dire recalculées par rapport au contexte syllabique immédiat (par rapport au deux syllabes précédentes et à la syllabe suivante), pour obtenir des durées relatives (sans unités), et des mesures de F0 relatives (en ST). Les pauses silencieuses de plus de 250 ms ont été utilisées pour contraindre ce calcul : certaines de ces syllabes adjacentes, normalement utilisées pour le calcul des paramètres relatifs, peuvent être ignorées si elles sont au-delà d'une pause par rapport à la syllabe en cours de calcul.

5.3. Résultats de la détection automatique de proéminences

Sur la base de ces paramètres, une stratégie de décision basée sur des seuils acoustiques a été testée sur les 4372 syllabes valides (c'est-à-dire non exclues par la tire delivery). Il s'agissait de déterminer la pertinence des deux paramètres prosodiques relatifs retenus (durée et hauteur relatives en vue d'estimer le caractère proéminent ou non des syllabes du corpus. Dans cette optique, une syllabe sera considérée comme proéminente si l'un ou l'autre des deux paramètres (F0 max ou durée syllabique) est supérieur à un certain seuil. Seront donc considérées comme proéminentes les syllabes hautes ou/et longues relativement aux syllabes adjacentes. Les seuils optimaux pour ce corpus sont :

- 3 ST pour le paramètre de hauteur
- 2 comme durée syllabique relative

Les résultats obtenus par cette méthode de classification par seuils et présentés dans le tableau 5 montrent un taux de bonne classification totale de **84.54%** (= 66.7% + 17.8%). Parmi ces syllabes 'bien classifiées', 66.7% sont des syllabes identifiées comme non proéminentes par les humains et détectées comme telles par l'automate, et 17.8 % sont des syllabes identifiées proéminentes par l'annotation manuelle et l'automate. Le reste sont des fausses alertes (5.5 % de syllabes saillantes selon l'algorithme mais pas d'après l'annotation manuelle) ou des détections manquées (10% des syllabes marquées 'P' ou 'p' par les experts ne sont pas identifiées comme telles par l'automate).

Tableau 5 : Matrice de confusion entre détection automatique et codage manuel sur la totalité du corpus, en %

		Codage manuel	
		Non proéminent	Proéminent
Détection automatique	Non proéminent	66.7	10.0
	Proéminent	5.5	17.8

Les deux *scatter plots* de la figure 7 offrent une autre illustration de la répartition des syllabes proéminentes dans le corpus d'étude. Les 4372 syllabes y sont représentées par leur F0 maximum relative (en abscisse) et leur durée relative (en ordonnée). L'annotation manuelle préalable a permis de tracer la distribution de deux types de syllabes (proéminente à gauche / non proéminente à droite) ainsi que des ellipses de confiance (à 1 fois la variance, c'est-à-dire représentant 68% des points de chaque catégories). En traits plus clairs, les ellipses pour chaque sous-corpus ; en traits foncés, les ellipses de l'ensemble des corpus. Les catégories proéminente/non proéminente y sont assez distinctes conformément aux scores de classification automatique choisis (3 ST pour la hauteur relative, 2 pour la durée relative).

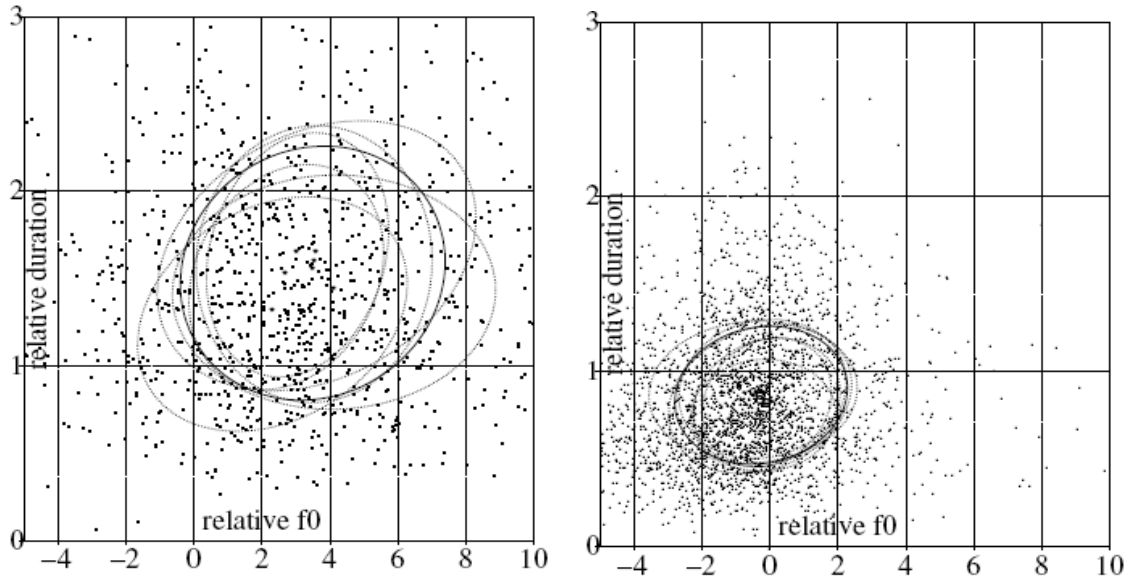

Figure 7. Distribution des syllabes dans le plan 'F0 relative /durée syllabique relative' des 3190 syllabes non-proéminentes (à gauche) et des 1182 syllabes proéminentes (à droite)

5.4. Représentation graphique de la détection automatique

Le Prosogramme original permet de tracer des graphiques représentant la mélodie stylisée en regard de la segmentation phonétique. La version modifiée présentée ici y ajoute quelques informations, comme :

- La valeur des paramètres prosodiques relatifs pour chaque noyau (en vert et souscrit, la durée relative, en rouge et suscrit la hauteur mélodique)
- La tire d'annotation manuelle, comprenant à la fois les proéminences mais aussi les symboles de la tire *delivery* comme les hésitations, les faux départs… (en bleu au-dessus de chaque segment stylisé)

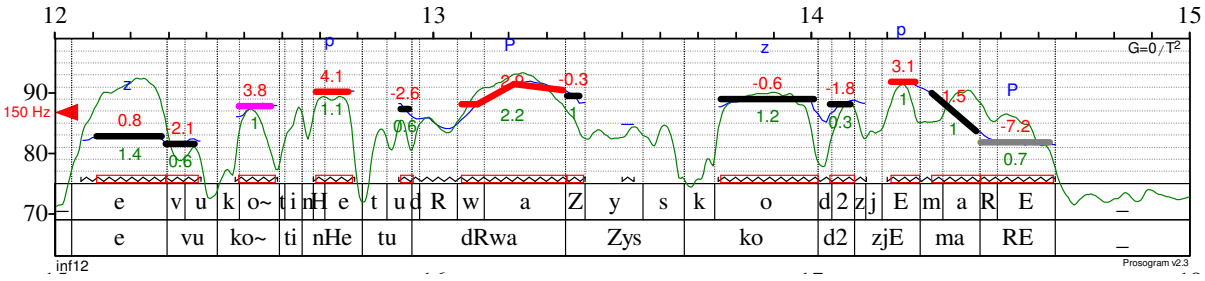

Figure 8: Prosogramme enrichi (Iti-12) et vous continuez tout droit jusqu'au deuxième arrêt

Les syllabes annotées proéminentes par les experts et identifiées comme telles par l'algorithme sont colorées en rouge, les syllabes annotées proéminentes par les experts mais non reconnues comme telles par l'algorithme sont en gris (détection manquée) ; enfin, dernier cas de figure, les fausses alertes (syllabes proéminentes selon l'automate mais non codées P/p par les experts) sont en rose.

Cette façon de faire permet à l'utilisateur de travailler sur une version *ad hoc* du corpus, dans laquelle le discours est réduit à ses seules informations pertinentes du point de vue des proéminences, à savoir, selon notre hypothèse : les pics conjoints d'intensité et de F0, la stylisation de la fondamentale, la segmentation phonétique et syllabique et les valeurs des paramètres acoustiques retenus. Chaque erreur de détection peut ainsi être diagnostiquée *a posteriori* (mauvaise segmentation, codage incertain, exceptions, non pertinence des paramètres acoustiques choisis, etc.).

6. Conclusion et perspectives

Dans cet article, nous avons présenté une approche à deux volets devant aboutir à l'annotation semi-automatique des proéminences syllabiques dans des corpus oraux non lus. Cette entreprise se justifie par le fait que les proéminences sont des phénomènes acoustiques basiques, préliminaires obligés à toute étude s'intéressant aux rapports entre prosodie et discours.

D'abord, nous avons présenté une méthodologie pour l'annotation manuelle des proéminences syllabiques. Nous avons justifié le fait que ce genre de transcription, pour être opératoire et exploitable par d'autres, ne devait pas faire intervenir trop de considérations théoriques. Nous avons aussi insisté sur le fait qu'elle devait suivre des règles strictes.

Dans la deuxième partie de ce travail, nous avons montré de quelle façon nous nous y sommes pris afin d'automatiser la procédure de détection des proéminences. En nous basant sur un corpus aligné de 20 minutes, composé de deux genres discursifs différents (des prescriptions d'itinéraires *in situ* ainsi que des interviews radiophoniques), nous avons montré que notre algorithme de détection semi-automatique offrait des résultats très prometteurs. En effet, la marge de flottement entre les deux experts humains qui ont annoté le corpus est de 89%. Or la marge de flottement entre l'annotation humaine consensuelle et la détection semi-automatique n'est pas si éloignée, puisque les paramètres acoustiques choisis dans cette étude ont permis

que l'on atteigne 84,5% de consensus. Contrairement à d'autres procédures du même type (Tamburini, 2003, 2005), l'outil que nous avons présenté permet de traiter des corpus de parole spontanée, et pas seulement de la parole lue, issue des laboratoires où elle a été enregistrée. Il ne remplace pas le codage d'experts humains, mais permet de le faciliter grandement (il est possible de corriger les erreurs de détection semi-automatique après-coup, mais aussi de s'appuyer sur la détection pour vérifier les intuitions des codeurs humains).

Ce travail, encore exploratoire, offre d'intéressantes perspectives d'amélioration pour la suite.

- On aimerait tester de nouveaux paramètres acoustiques pour le repérage semi-automatique des proéminences, et : (i) utiliser des valeurs globales et non plus locales, en cherchant à déterminer les valeurs moyennes de F0 et de durée dans des segments discursifs restreints, de l'ordre de la période intonative (Lacheret-Dujour et Victorri, 2002 ; Lacheret, 2003) p. ex. ; (ii) injecter de la connaissance linguistique, ce qui permettra sans doute de reconnaître plus aisément les tons bas extrêmes, que l'algorithme a encore du mal à détecter à l'heure actuelle).
- Prendre en compte d'autres corpus, de genres discursifs différents de ceux-ci, avec des locuteurs parlant des variétés de français moins 'standard'. Ainsi cela nous permettra de voir si notre outil – encore en cours de développement – est robuste à la variation phonostylistique, que celle-ci soit liée à la situation de communication (les journalistes qui commentent un événement sportif ne parlent pas de la même façon que quelqu'un qui prescrit une recette de cuisine) ou à l'origine géographique (p. ex. les francophones de Suisse Romande n'utilisent pas les mêmes paramètres acoustiques que les francophones de la région parisienne ou de Marseille)[21].

En résumé, l'outillage présenté et discuté ici se montre à la fois souple et robuste. Il permet d'envisager de traiter automatiquement ou semi automatiquement de grandes masses de données orales, et d'extraire des résultats statistiques dont la validité est fonction de l'ampleur du corpus, et de la fiabilité de l'outil automatique. Sa valeur, l'utilité réelle de l'outil, est fonction des tâches qui lui seront confiées et auxquelles il sera confronté. Outre une meilleure connaissance des prosodies, de la parole lue/préparée ***et*** de la parole spontanée, il peut servir à

[21] Cf. dans cette optique les analyses phonostylistiques du parler des chroniques radiophoniques de France Inter (Goldman et Auchlin, 2006 ; Burger et Auchlin, 2007 ; Goldman *et al.,* 2007).

caractériser, en interaction avec une analyse linguistique et discursive, des stratégies prosodiques associées à des buts communicatifs, de façon plus ou moins locale, aussi bien que des profils phonostylistiques, récurrents ou occasionnels, associés à des sociolectes, dialectes, etc. – c'est-à-dire aussi bien à repérer ou détecter des points communs et des différences entres variantes sélectionnées qu'à décrire avec précision les caractéristiques de telle ou telle variante. Une fois les prolongements annoncés effectués, nous espérons que cet outil pourra être utilisé pour l'annotation des grands corpus de langue parlée.

7. Remerciements

Mathieu Avanzi tient à remercier le Fonds National Suisse de la Recherche Scientifique, qui a financé le projet *La structuration interne des périodes* (subside n°100012-113726/1), dans le cadre duquel s'inscrit cette recherche. Ce travail a également bénéficié du support financier du projet FRFC n° 2.4523.07 *Établissement d'une procédure de segmentation du discours oral en unités minimales (MDU) sur la base de critères syntaxiques et prosodiques* du Fonds national de la Recherche scientifique belge. Les auteurs remercient également Piet Mertens (KUL) et Bernard Victorri (Lattice, ENS, Paris) pour leur précieuse collaboration, ainsi que les deux relecteurs anonymes pour leurs suggestions.

8. Références citées

Avanzi, M. (2004) *L'indication d'itinéraire en français parlé. Schématisation cognitive et organisation macro-syntaxique*. Mémoire de maîtrise : Grenoble 3.

Avanzi, M. et Martin, Ph. (2007) L'intonème conclusif : une fin (de phrase) en soi ? *Cahiers de linguistique française* 28 : 244-258

Barras, C., Geoffrois, E., Wu, Z. and Liberman, M. (2001) *Transcriber*: development and use of tool for assigning speech corpora production *Speech Communication* 33/2: 5-22.

Béguelin, M.-J. (2002) Clause, période ou autre ? La phrase graphique et les niveaux de l'analyse *Verbum,* 24/1-2 : 85-108.

Beckman, M., Hirschberg, J. and Shattuck-Hufnagel, S. (2006) The Original ToBI System and the Evolution of the ToBI Framework In: J. Sun-Ah (ed.) *Prosodic models and transcription: Towards prosodic typology*. Oxford: University Press: 9-54.

Beller, G., Hueber, T., Schwarz, D. and Rodet, X. (2006) Speech Rates in French Expressive Speech. *Proceedings of Speech Prosody* 2006, Dresden.

Berrendonner, A. (à par.) Qu'est-ce qu'une période ? In: Groupe de Fribourg *Grammaire de la période*.

Blanche-Benveniste, C. (1998) Ponctuation et langue parlée In *Le Discours Psychanalytique*, Revue de l'Association Freudienne, numéro spécial 18, *La Ponctuation*, octobre 1997, Actes des Journées de l'Association Freudienne Internationale, 14-15 juin 1997, au Centre Hospitalier Sainte Anne : 73-109.

Boersma, P. and Weenink, D. (2007) *Praat: doing phonetics by computer (Version 4.5)*. www.praat.org

Buhmann, J., Caspers, J., van Heuven, V., Hoekstra, H., Martens, J.-P. and Swerts, M. (2002) Annotation of Prominent Words, Prosodic Boundaries and Segmental Lengthening by Non Expert Transcribers in the Spoken Dutch Corpus In M.C. Rodriguez and Suarez Araujo (eds) *Proceedings of the 2nd International Conference on Language Resources and evaluation (LREC 2002)*. Paris: ELRA, 779-785

Burger, M. et Auchlin, A. (2007) Quand le parler radio dérange : remarques sur le phono-style de France Info In: M. Broth, M. Forsgren, C. Norén, and F. Sullet-Nylander (eds) *Le Francais parlé des médias. Actes du colloque de Stockholm 8-12 juin 2005*. Stockholm: Acta Universitatis Stockholmiensis, 97-111.

Campione, E. (2001) *Étiquetage prosodique semi-automatique de corpus oraux : algorithmes et méthodologie.* Thèse de doctorat : Université de Provence.

Campione, E. (2003) Quelques outils pour l'étiquetage prosodique des corpus oraux. In V. Aubergé et A. Lacheret-Dujour (eds.) *Actes du colloque international Journées Prosodie 2001*. Université de Grenoble : 103-107.

Campione, E. & Véronis, J. (2001) Etiquetage prosodique semi-automatique des corpus oraux In *Actes de la conférence Traitement Automatique des Langues (TALN'2001)*, Tours, ATALA: 123-132.

Cohen, J. (1960) A Coefficient of Agreement for Nominal Scales *Educational and Psychological Measurement* 20: 37-46.

Cresti, E. (2000) *Corpus di italiano parlato*. *Vol. 1: Introduzione*. Firenze : Academia de la Crusca.

Cresti, E. and Moneglia, M. (2005) *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam : Benjamins.

Delais-Roussarie, E. (2003) Constitution et annotation de corpus : Méthodes et recommandations' In E. Delais-Roussarie et J. Durand (eds) *Corpus et Variation en phonologie du français*. Presses Universitaires du Mirail : 91-125.

Delais-Roussarie, E., Post, B. and Portes, C. (2006) Annotation prosodique et typologie *Travaux Interdisciplinaires du Laboratoire Parole et Langage* 25 : 61-95.

Eriksson, A., Grabe, E. and Traunmüller, H. (2002) Perception of syllable prominence by listeners with and without competence in the tested language In: *Proceedings of the Speech Prosody 2002 Conference*, 11-13 April 2002, Aix-en-Provence: 275-278.

Goldman, J.-P. (2007). *EasyAligner: a semi-automatic phonetic alignment tool under Praat*. Available at http://latcui.unige.ch/phonetique.

Goldman, J.-P. and Auchlin, A. (2006) Quelques observations intuitives et mesurées sur le phonostyle de France Info. Communication au Colloque international Phonologie du Français Contemporain 2006, *Approches phonologiques et prosodiques de la variation sociolinguistique: le cas du français*, Louvain-la-Neuve, 6-8 juillet 2006.

Goldman, J-P., Avanzi, M., Lacheret-Dujour, A., Mertens, P. et Simon, A.-C. (2006) Pour un codage prosodique outillé : outils de segmentation automatique sous Praat et détection de proéminences. Communication présentée aux *Journées Phonologie du Français Contemporain : du social au cognitif*, Paris, 12-13 décembre 2006.

Goldman, J.-P., Avanzi, M., Lacheret-Dujour, A., Simon, A.-C. et Auchlin, A. (2007). A Methodology for the Automatic Detection of Perceived Prominent Syllables in Spoken French In Proceedings of Interspeech'07, Antwerp, Belgium, August 27-31: 98-101.

Goldman, J-P. et Avanzi, M. (2007) Vers un algorithme de détection (semi-)automatique des proéminences en français parlé In *Actes des 7èmes Rencontres des Jeunes Chercheurs sur la Parole (RJCP07)*, Paris, 05-06 juillet 2007 : 84-87.

Goldman, J.-P., Auchlin, A., Avanzi, M. et Simon, A-.C. (2007) Phonostylographe: un outil de description prosodique. Comparaison du style radiophonique et lu *Cahiers de linguistique française* 28 : 219-237.

Grabe, E. and Post, B. (2002) Intonational Variation in English In: B. Bel and I. Marlin (eds) *Proceedings of the Speech Prosody 2002 Conference*, 11-13 April 2002, Aix-en-Provence: 343-346.

Groupe de Fribourg (à par.) *Grammaire de la période*.

Hirst, D. and Espesser, R. (1993) Automatic Modelling of Fundamental Frequency Using a Quadratic Spline Function *Travaux de l'Institut de Phonétique d'Aix-en-Provence* 15 : 75-85

House, D. (1990) *Tonal Perception in Speech*. Lund: University Press.

Jenkin, K.L. and Scordilis, M.S. (1996) Development and Comparison of Three Syllable Stress Classifiers In: *Proceedings ICSLP'96*, Philadelphia: 733-736.

Lacheret-Dujour, A. et Beaugendre, F. (1999) *La prosodie du français*. Paris : CNRS.

Lacheret-Dujour, A. et Victorri, B. (2002) La période intonative comme unité d'analyse pour l'étude du français parlé : modélisation prosodique et enjeux linguistiques *Verbum* 24/1-2 : 55-73.

Lacheret-Dujour, A. (2003) *La prosodie des circonstants en français parlé*. Leuven/Paris : Peeters.

Martens, J.-P. (2002) *Protocol voor prosodische annotatie*. Unpublished document.

Martens, J.-P., Binnenpoorte, D., Demuynck, K., Van Parys, R., Laureys, T., Goedertier, W. and Duchateau, J. (2002) Word Segmentation in the Spoken Dutch Corpus. In: M.C. Rodriguez and A. Suarez (eds) *Proceedings of the 2nd International Conference on Language Resources and evaluation (LREC 2002)*. Paris: ELRA.

Martin, Ph. (1975) Analyse phonologique de la phrase française *Linguistics* 146 : 35-68.

Martin, Ph. (2003) ToBI : l'illusion scientifique ?'. In: V. Aubergé et A. Lacheret-Dujour (eds) Actes du colloque international *Journées Prosodie 2001*. Université de Grenoble : 109-113.

Martin, Ph. (2006) La transcription des proéminences accentuelles : mission impossible ? *Bulletin PFC* 6 : 81-87.

Martin, Ph. (à par.) A propos de la perception et la transcription des unités prosodiques In : *Actes du colloque Transcription de la langue parlée. Aspects théoriques, pratiques et technologiques*, Perpignan, 27-30 juin 2005.

Mertens, P. 1987. *L'intonation du français. De la description linguistique à la reconnaissance automatique*. Unpublished Ph.D. dissertation: University of Leuven, Belgium.

Mertens, P. (1991) Local prominence of acoustic and psychoacoustic functions and perceived stress in French In: *Proc. 12th Int. Cong. Phon. Sc.*, vol. 3 : 218-221.

Mertens, P. (2004) Un outil pour la transcription de la prosodie dans les corpus oraux', *Traitement Automatique des langues* 45 /2 : 109-130.

Moneglia, M. and Cresti, E. (2006) C-ORAL-ROM – Prosodic Boundaries for Spontaneous Speech Analysis In: Y., Kawaguchi, S., Zaima and T. Takagaki (eds) *Spoken Language Corpus and Linguistic Informatics*. Amsterdam: Benjamins, 89-113.

Morel, M., Lacheret-Dujour, A., Lyche, C. et Poiré, F. (2006) Vous avez dit proéminences ? In: *Actes JEP 06*: 183-186.

Pickering, B., Williams, B. and Knowles, G. (1996) Analysis of transcribers differences in the SEC In: G. Knowles, A. Wichmann, and P. Alderson (eds) *Working with Speech:*

perspectives and research into the Lancaster/IBM Spoken English Corpus. London/New-York: Longman, 59-105.

PierreHumbert, J. (1980) *The Phonetic and Phonology of English Intonation*. PhD Thesis: MIT.

Poiré, P. (2006) La perception des proéminences et le codage prosodique *Bulletin PFC* 6 : 69-79.

Poiré, F. et Simon, A-.C. (2006) Constitution d'une base de données pour l'analyse prosodique en langage spontané. Communication au colloque *Les français d'ici - Acadie, Québec, Ontario, Ouest canadien*, 5-8 juin 2006, Queens University, Kingston.

Post, B., Delais-Roussarie, E. et Simon, A-.C. (2006) IVTS, un système de transcription pour la variation prosodique *Bulletin PFC* 6 : 51-68.

Rossi, M. (1978a) La perception des glissandos descendants dans les contours prosodiques *Phonetica* 35/1 : 11- 40.

Rossi, M. (1978b) Interactions of intensity glides and frequency glissandos *Language & Speech* 21: 384-396.

Simon, A.-C. (2004) *La structuration prosodique du discours en français. Une approche multidimensionnelle et expérientielle*. Berne : Peter Lang.

t'Hart, J. Collier, R. and A. Cohen, (1991) *A Perceptual Study of Intonation. An Experimental-Phonetic Approach to Speech Melody*. Cambridge: University Press.

Tamburini, F. (2003) Automatic Prosodic Prominence Detection in Speech using Acoustic Features: an Unsupervised System In: *Proc. 8th European Conference on Speech Communication and Technology - Eurospeech 2003*, Geneva: 129-132.

Tamburini, F. (2005) *Fenomeni prosodici e prominenza : un approccio acustico*. Bologna: BUP.

Tamburini, F. and Caini, C. (2005) An Automatic System for Detecting Prosodic Prominence in American English Continuous Speech *International Journal of Speech Technology* 8: 33-44.

Terken, J. (1991) Fundamental frequency and perceived prominence *Journal of the Acoustical Society of America* 89: 1768-1776.

Wightman, C. (2002) ToBI or not ToBI? In: *Proc. of the Speech Prosody 2002 Conference*, 11-13 April 2002, Aix-en-Provence.

Williams, B. (1996) The formulation of an intonation transcription system for British English In: G. Knowles, A. Wichmann, and P. Alderson (eds) *Working with Speech:*

perspectives and research into the Lancaster/IBM Spoken English Corpus. London/New-York: Longman, 38-58.

Young, S. *et al.* (2000) *The HTK book*. Cambridge: CUP.