La détection des proéminences syllabiques. Un aller-retour entre l'annotation manuelle et le traitement automatique

Anne-Catherine Simon, Mathieu Avanzi, Jean-Philippe Goldman

La détection des proéminences syllabiques. Un aller-retour entre l'annotation manuelle et le traitement automatique

1 L'annotation semi-automatique de la prosodie dans les corpus oraux

Notre objectif est de développer un système d'annotation semi-automatique des proéminences syllabiques en français, sur la base d'une annotation manuelle experte qui résout les difficultés rencontrées par les tentatives antérieures[1]. La première partie de cet article résume les résultats que nous avons obtenus précédemment [Avanzi *et al.* 2007 ; Goldman & Avanzi 2007 ; Goldman *et al.* 2007]. Nous renvoyons le lecteur intéressé à ces travaux et aux références bibliographiques qu'ils contiennent, et nous synthétisons ci-dessous les points importants qui s'en dégagent.

L'outil que nous présentons, ainsi que la méthodologie qui a conduit à sa création, concernent le français. Notre ambition est que l'outil soit robuste aux différentes variétés de français, et principalement à la variation stylistique[2].

Nous visons la mise au point d'un outil d'annotation implémenté sous Praat [Boersma & Weenink 2008][3] qui permettrait d'annoter de façon semi-automatique la structure prosodique dans les corpus oraux[4]. Par rapport à une annotation manuelle, l'annotation semi-automatique permet de (liste non exhaustive) :

- limiter la variation intra- et inter-transcripteurs, et donc de minimiser le problème de la subjectivité ;
- gagner un temps non négligeable, et donc faciliter le traitement de grosses bases de données ;
- donner un caractère reproductible à la procédure d'annotation, la rendre contrôlable et testable par d'autres ;
- mettre à disposition de la communauté scientifique un outil libre et convivial, qui permettra à tous (entre autres aux non-experts en phonétique/phonologie) d'identifier facilement les proéminences syllabiques dans les corpus de français parlé, d'une manière fiable.

1.1 Spécificités de l'oral non préparé

Une difficulté dans la détection des proéminences réside dans l'amalgame qui peut être fait entre « proéminence » (notion strictement perceptive, qui caractérise un élément, en l'occurrence une syllabe, qui se détache de son environnement par un paramètre acoustique donné) et « accent » (notion grammaticale et phonologique, dépendante d'une langue particulière et d'un modèle accentuel) [Post *et al.* 2006 ; Avanzi *et al.* 2007]. Par exemple, quel statut donner à un « euh » d'hésitation, vraisemblablement proéminent par sa durée ? Afin d'identifier les syllabes qui peuvent poser problème sans devoir d'emblée trancher sur leur statut, une tire d'annotation (nommée « delivery ») est réservée au codage des phénomènes suivants : allongements liés à une hésitation, réalisation de schwas post-toniques, appendices (voir Tableau 1). L'intérêt majeur de cette façon de faire est qu'elle permet de mettre sur deux plans distincts, lors de l'analyse, les processus réguliers d'encodage prosodique structural et les phénomènes de planification discursive ou « marques du travail de formulation » [Morel & Danon-Boileau 1998].

1.2 Procédure d'annotation manuelle des proéminences

Pour mettre au point notre logiciel de détection automatique de syllabes proéminentes, nous avons procédé de la façon suivante. Deux experts ont annoté, sur la base de critères perceptifs uniquement, les proéminences[5] dans un corpus aligné préalablement en syllabes[6]. Les transcripteurs ont suivi un protocole

strict (durée des segments soumis à l'écoute restreinte entre 3 et 5 sec., nombre d'écoutes limité, recours à l'affichage du signal réservé aux cas « retors », etc. ; voir, pour les détails de la procédure, Avanzi *et al.* 2007), sur lequel ils se sont entendus et entraînés au préalable. L'annotation a été effectuée par chaque codeur de manière autonome, sur la base des symboles présentés dans le Tableau 1.

Tableau 1. Catégories de symboles pour le codage manuel des phénomènes prosodiques

1. Codage des intervalles vides	
%	segment inaudible ou inexploitable (rire, toux, chevauchement, bruit)
*	prises de souffle
_	silence (issu de la détection de l'alignement automatique)
2. Codage des proéminences	
P	proéminence forte
p	proéminence faible
0	syllabe non proéminente
3. Codage des phénomènes de *delivery*	
z	hésitation (allongement, contour plat, *creaky voice*)
@	schwa post-tonique (p. ex. *c'est dingu**e*** [sɛdɛ̃gə])
$	appendice (p. ex. *c'est dingue **quoi***)

Les symboles utilisés pour l'annotation sont rangés en trois grandes classes :

1. classe des symboles d'intervalles vides (pauses, prises de souffle, « junk »[7]) ;
2. classe des symboles des syllabes exclues (pour ne pas fausser la détection des syllabes proéminentes, *cf.* le cas de la durée les conclusions de Morel *et al.* [2006]) ;
3. classe des symboles des proéminences proprement dites (réparties en deux degrés[8]).

Ensuite, les codages ont été comparés, les divergences listées et les codeurs ont réglé les divergences en vue de s'accorder sur une annotation dite « de référence ». Ce codage de référence ne prétend pas être exempt d'erreurs ou de mésinterprétations, mais il est important, car c'est à cette annotation consensuelle que le logiciel va se mesurer.

Nous avons comparé, dans le corpus présenté *infra* (Tableau 2), les codages des deux experts, puis comptabilisé le nombre d'intervalles pour lesquels il y avait un conflit quant à la présence d'une proéminence ou d'un autre symbole (syllabe 0 ou symbole de la tire « delivery »). Sur les 12688 syllabes composant le corpus (un sous-corpus de 128 syllabes a servi de corpus d'entraînement), 1185 intervalles syllabiques font l'objet d'un désaccord entre les deux codeurs. Cela représente donc un taux d'accord d'environ 90.67%. Comme Buhman *et al.* [2002], nous pensons que ces bonnes performances s'expliquent en grande partie par les étapes préliminaires au codage (entraînement conjoint ; apprentissage d'un protocole).

1.3 Résultats des études antérieures

Dans des études antérieures, nous avons travaillé sur un corpus de 18 minutes [Avanzi *et al.* 2007 ; Goldman *et al.* 2007], puis de 20 minutes [Goldman & Avanzi 2007]. Ces données représentaient deux types de situation de parole (des interviews radiophoniques et des prescriptions d'itinéraires en milieu urbain[9]), avec des locuteurs francophones natifs de Belgique et de France.

Nous avions retenu pour la détection automatique deux paramètres acoustiques : la **durée** et la **hauteur de f0**. Ces deux paramètres étaient calculés de façon relative (*cf.* section 3.2.1 ci-dessous). Les résultats obtenus étaient relativement encourageants : nous étions parvenus à un taux de convergence de 84.4% entre l'annotation experte et la détection automatique, score qui figure parmi les meilleurs pour ce type d'étude[10]. Dans cet article, nous reconduisons l'expérience sur un corpus élargi, incluant des francophones natifs de Suisse, de Belgique et de France et des conditions de production de parole plus variées. La composition précise de notre corpus est fournie dans le Tableau 2.

2 Le corpus d'étude

Mises à part les interviews radiophoniques et les prescriptions d'itinéraire, dont le contenu est détaillé dans [Goldman & Avanzi 2007 ; Goldman *et al.* 2007], les autres styles de parole (discours politiques, journaux radiophoniques et récits de vie) sont représentés par trois enregistrements monologiques continus de 3'30'' en moyenne (soit un par zone géographique retenue, Belgique, Suisse et France). Le corpus est échantillonné minimalement, et contient des enregistrements de style plus ou moins formel (le discours politique lu *vs* le récit monologique « libre »).

Ce corpus, qui a servi à la mise au point du script **Prosoprom** (cf. section 3), est baptisé **C-PROM**. Il sera prochainement publié et accessible librement sur Internet. Il comprendra, outre les cinq genres de parole présentés ici, des conférences universitaires et des extraits de conversation libre (cf. [Avanzi, Goldman & Simon en prép.]).

Tableau 2. Statistiques de l'annotation manuelle sur le corpus d'étude. De haut en bas : durée des enregistrements et nombre total de syllabes ; syllabes exclues pour le codage des proéminences (marquées par un signe de la tire « delivery ») – avec pourcentage par rapport à la totalité des syllabes ; syllabes proéminentes (P ou p), syllabes non proéminentes (0) – avec pourcentage par rapport à la totalité des syllabes ; total de syllabes dites « valides » et prises en compte pour l'annotation, *i.e.* ensemble auquel doit se mesurer l'automate.

	+ formel				**- formel**	**Corpus complet**
	Discours Politiques (cote DP)	**Journaux radiophoniques (cote JP)**	**Interviews radiophoniques (cote IRT)**	**Prescriptions d'itinéraires (cote ITI)**	**Récits de vie (cote RCV)**	
Durée (sec)	633	619	626	587	621	**3086**
Nb total syll.	2175	3164	2593	2229	2655	**12816**
« Delivery »	18	97	184	186	206	**691**
%	0.82	3.06	7.09	8.34	7.75	**5.39**
P/p	625	808	637	531	643	**3276**
%	28.73	25.53	24.56	23.82	24.21	**25.56**
0	1532	2257	1762	1508	1790	**8849**
%	70.43	71.33	67.95	67.65	67.41	**69.04**
Total syllabes « valides »	**2157**	**3067**	**2409**	**2043**	**2449**	**12125**

Sur un total de **12816** intervalles syllabiques, 691 syllabes ont été exclues via la tire « delivery » (5.39%), 3276 syllabes ont été codées « p » ou « P » (soit 25.56%). Restent 8849 syllabes non proéminentes (69.04%). Le total des syllabes valides (c'est-à-dire codées « 0 » ou « p/P » et non exclues par un symbole de la tire « delivery ») s'élève à **12125**.

3 L'annotation automatique

La procédure automatique de détection de proéminences se décompose en deux temps. Le système procède d'abord à une identification des noyaux vocaliques via une version enrichie du script de stylisation mélodique **Prosogram** de Mertens [2004]. Nous choisissons et calculons ensuite les paramètres acoustiques (et les seuils) à utiliser pour la détection des proéminences syllabiques.

3.1 Identification des noyaux vocaliques

La première étape de la procédure d'identification des proéminences accentuelles consiste en une **stylisation** du signal mélodique. Pour ce faire, une version modifiée du script **Prosogram** a été développée. Elle permet de repérer les noyaux vocaliques sur la base d'une segmentation phonétique et du paramètre d'intensité. Le but de cette opération est double : d'une part cette stylisation permet d'éliminer un maximum de risque d'erreur dans la détection de f0 (seules les parties les plus stables des noyaux de syllabes sont stylisées). D'autre part, cette simplification de la courbe mélodique ne retient que les variations mélodiques perçues / fonctionnelles [Mertens & d'Alessandro 1995 ; Hermes 2006] et réduit l'impact des contraintes articulatoires microprosodiques, supposées non pertinentes.

Le traitement automatique comporte deux étapes.

1. La segmentation en noyaux consiste à repérer les portions voisées et intenses de chaque syllabe. Plus précisément, le noyau est délimité de part et d'autre du maximum d'intensité de chaque syllabe pour autant que la parole soit toujours voisée et que l'intensité ne soit pas inférieure à un seuil défini par rapport au maximum local d'intensité[11]. Il résulte de cette segmentation que certaines syllabes non voisées ne donnent pas lieu à la détection d'un noyau[12].
2. La courbe mélodique de cette portion de parole voisée est stylisée en un ou plusieurs segments de droite tout en respectant des seuils perceptifs de glissando fixés au préalable[13]. Afin de travailler sur une représentation plus proche de la substance, nous avons fixé ce seuil à 0.16 (au lieu de 0.32, qui est la valeur par défaut du script **Prosogram**).

3.2 Paramètres acoustiques retenus pour identifier les syllabes proéminentes

Une fois les noyaux détectés, nous traitons pour chacune des syllabes les mesures acoustiques suivantes :

- la **durée de la syllabe** (en millisecondes), que l'on préfère à la durée du noyau syllabique stylisé, laquelle est dépendante des traits de voisement des consonnes contenues dans l'attaque et la coda de la syllabe ;
- le **maximum de f0** (en semi-tons, désormais ST) atteint sur le noyau.

Ces mesures acoustiques sont ensuite « relativisées », c'est-à-dire recalculées par rapport au contexte syllabique immédiat (plus précisément, relativement aux deux syllabes précédentes et à la syllabe suivante[14]), pour obtenir des durées relatives (en pourcentage) et des mesures de différence relative de f0 (en ST). A noter que les pauses silencieuses de plus de 250 ms ont été utilisées pour contraindre ce calcul : des syllabes adjacentes, normalement utilisées pour le calcul des paramètres relatifs, seront ignorées si elles sont séparées de la syllabe en cours d'analyse par une pause de plus de 250 ms. Cette mesure est inspirée des travaux de Lacheret-Dujour [2003], qui considère que des pauses de cette durée, couplées avec d'autres indices acoustiques (contour postposé d'une certaine amplitude, présence d'une réinitialisation, pas de *euh* d'hésitation dans l'environnement immédiat), constituent des indices fiables de fin d'unités prosodiques maximales (ou *périodes*). Partant, il n'y aurait pas lieu de comparer les syllabes qui se succèdent si elles appartiennent à deux macro-unités prosodiques distinctes.

Bien qu'elle ait donné de bons résultats, la prise en compte de ces deux paramètres demeure insuffisante (voir pour une critique [Obin, Rodet & Lacheret 2008]). C'est pourquoi, en vue d'améliorer les premiers scores, nous avons ajouté deux paramètres acoustiques supplémentaires.

- **La montée mélodique (ou mouvement mélodique intrasyllabique)** : il s'agit de la partie montante de la fréquence fondamentale stylisée dans le noyau syllabique. Autrement dit, cette mesure, qui correspond au mouvement mélodique d'un ton montant, est considérée nulle pour les noyaux plats et pour les noyaux descendants. Dans le cas des glissandos complexes (montant-descendant ou descendant-montant), seule la partie montante est comptabilisée. Ce choix provient de l'idée que les mouvements mélodiques, bien qu'en partie expliqués comme un phénomène transitoire reliant une cible à une autre, peuvent également par eux-mêmes véhiculer la perception de proéminence.
- **La pause subséquente** est une marque évidente de fin de groupe en français [Lacheret-Dujour & Beaugendre 1999]. S'appuyer sur la pause pour repérer les proéminences signifie qu'on éloigne la notion de proéminence de critères purement perceptifs pour la rapprocher de la notion d'accentuation. En effet, on considère qu'une syllabe se détache ou « ressort » du contexte environnant en vertu d'une pause subséquente, et non pas d'une variation intrinsèque (allongement, montée mélodique) à la syllabe elle-même. Ce choix se fonde sur une connaissance du système accentuel français, qui est largement oxytonique.

Sur la base de ces étapes de repérage des noyaux syllabiques, de stylisation de la courbe mélodique et finalement de calcul des paramètres acoustiques (dont certains relativement aux syllabes adjacentes), il s'agit de prendre une **décision** sur le caractère proéminent d'une syllabe donnée.

La stratégie la plus simple consiste à considérer chacun des paramètres comme un critère de décision autonome, c'est-à-dire de considérer une syllabe proéminente si un seul des paramètres dépasse un seuil choisi. Autrement dit, une syllabe sera dite proéminente si son maximum relatif de f0 est supérieur à un seuil à déterminer, ou si sa durée relative est supérieure à un seuil à déterminer, etc. Ces conditions sont suffisantes mais ne sont pas exclusives : par exemple, il se peut qu'une syllabe proéminente soit à la fois aiguë, longue et suivie d'une pause. En effet l'interdépendance des paramètres, c'est-à-dire leur non-orthogonalité, est connue.

3.3 Résultats de l'analyse sur le corpus élargi

Pour mémoire, le logiciel de détection automatique **ProsoProm** doit se mesurer aux 12125 syllabes valides que contient notre corpus (total qui exclut les syllabes annotées avec un symbole de la tire « delivery », cf. Tableau 2).

Pour fixer les valeurs des quatre seuils retenus pour la détection automatique, nous nous sommes basés sur les résultats d'une précédente étude [Obin, Goldman *et al.* 2008]. Nous avions dans ce travail mené une étude exploratoire dichotomique, qui consistait à comparer systématiquement les résultats de la détection automatique avec le corpus de référence annoté par les experts. Nous avions ainsi observé que les valeurs adéquates (sur une partie plus restreinte du corpus) étaient les suivantes[15] :

- durée syllabique relative = 2 (sans unité) ;
- maximum relatif de f0 = 2 demi-tons ;
- montée mélodique (mouvement intrasyllabique) = 3.5 demi-tons ;
- pause subséquente = 300 millisecondes.

L'**évaluation** des résultats de l'automate se fait en comparant pour chaque syllabe l'annotation manuelle et la détection automatique. Quatre cas sont envisageables (ce sont les quatre cases grisées du Tableau 3, de gauche à droite et de haut en bas).

1. Une syllabe notée manuellement non proéminente est détectée comme non proéminente (on parle de **rejet concordant**).
2. Une syllabe annotée comme proéminente par les experts est détectée automatiquement comme non proéminente. Il pourrait s'agir d'une sur-détection des experts ou d'une omission de l'automate.

3. Une syllabe est annotée comme non proéminente par les codeurs et est détectée comme proéminente par l'automate.
4. Une syllabe est détectée comme proéminente par les experts et par l'automate (**proéminence concordante**).

Seuls les cas 1 et 4 sont évalués comme corrects, les cas 2 et 3 renvoient à des discordances entre le codage manuel et la détection automatique. Dans le Tableau 3 ci-dessous, la matrice de confusion résume les scores obtenus pour ces quatre cas de figure.

Tableau 3. Matrice de confusion entre détection automatique et codage manuel par les experts sur la totalité du corpus d'étude (en %)

		Codage manuel par les experts	
		Non proéminent	Proéminent
Détection automatique	Non proéminent	64.7	6.7
	Proéminent	10.3	18.2

La somme des cases en gris clair donne un score de bonne détection par l'automate dans **82.9%** des cas (64.7% + 18.2%).

4 Analyse des divergences entre détections manuelle et automatique

Les proéminences concordantes et les rejets concordants résultent d'un accord entre l'annotation manuelle de référence et la détection automatique. Ils ne seront pas commentés ici.

4.1 Proéminences détectées par l'automate et omises par les experts

Les proéminences détectées uniquement par l'automate constituent 6.7% de cas. Elles peuvent faire l'objet de trois interprétations.

- La machine détecte une syllabe proéminente sur la base des paramètres locaux alors que les codeurs perçoivent et interprètent cette syllabe dans le cadre d'une modification mélodique globale, comme une réinitialisation mélodique [Grobet & Simon 2002].
- Il s'agit d'une omission des codeurs, que la détection automatique permet de récupérer après-coup ; et la démarche d'aller-retour entre annotation manuelle et détection automatique prend tout son sens.
- Le calcul qui génère la détection de proéminence se base sur une information erronée due à une erreur dans le pré-traitement de segmentation et de stylisation des noyaux syllabiques.

Quantitativement, voici dans quelles proportions les paramètres de détection sont mis en cause pour les 812 détections d'une proéminence selon l'automate uniquement, dans notre corpus.

Tableau 4. Fréquence de recours aux paramètres de détection pour les 812 proéminences détectées par l'automate

f0 relative	610
Montée de f0	114
Durée relative	56
Pause	32
Total	812

Dans 75% des cas, une hauteur relative supérieure à 2 ST de la syllabe analysée est responsable de la détection par l'automate de cette syllabe comme proéminente. Comment expliquer qu'une syllabe d'au moins 2 ST plus haute que les syllabes environnantes n'ait pas été perçue comme proéminente par les codeurs experts ? La coexistence de variations prosodiques locales (utilisées par la détection automatique) et de variations prosodiques globales explique certaines de ces discordances.

Par exemple, le phénomène de la déclinaison tonale, qui s'observe dans certains styles de parole comme la lecture, produit une diminution progressive de la f0 plus ou moins corrélée au déroulement d'un groupe de souffle [Lacheret-Dujour & Beaugendre 1999 : 243]. Cette déclinaison peut être suivie d'une réinitialisation mélodique qui affecte les premières syllabes du groupe suivant. Les deux proéminences détectées par l'automate uniquement sont portées par les premières syllabes de l'énoncé « et avec le terrorisme ».

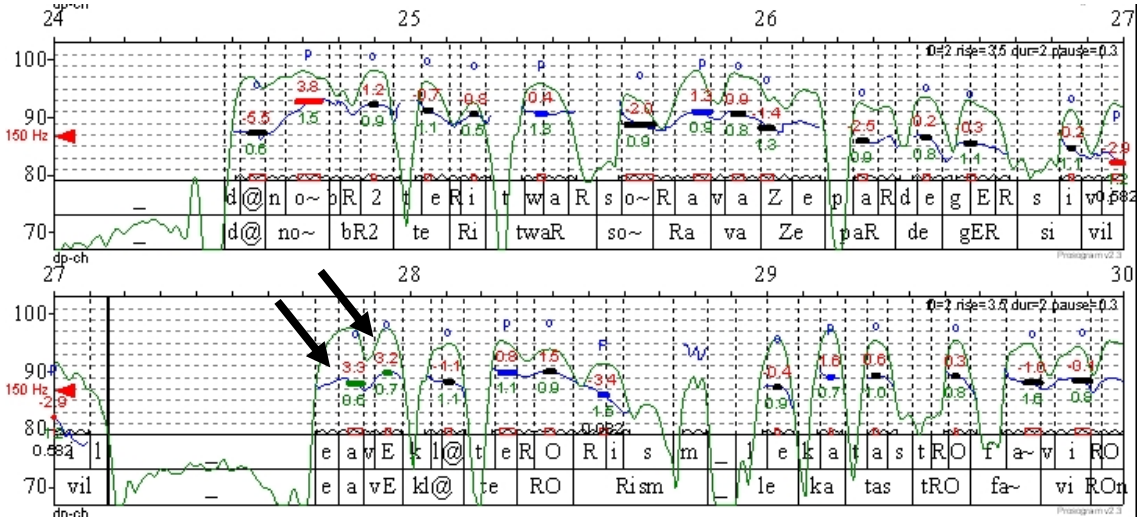

Figure 1. Extrait d'un discours politique : *de nombreux territoires sont ravagés par des guerres civiles / et avec le terrorisme … (DP-CH).* Interaction entre phénomènes globaux (ligne de déclinaison observable sec. 25 à 27, suivie d'un reset mélodique) et locaux (proéminences). Légende du prosogramme. Un trait rouge représente une proéminence concordante, détectée par les experts et par l'automate, tandis qu'un trait noir représente une syllabe non proéminente (« rejet concordant »). Un trait vert représente une proéminence détectée uniquement par l'automate et un trait bleu une proéminence détectée uniquement par les experts. Un trait gris (absent de ce schéma) représente une syllabe « non valide » pour la détection automatique (hésitation, schwa, etc.) et exclue de la comparaison.

La détection automatique identifie ces deux syllabes comme proéminentes à cause de leur hauteur relative locale (respectivement mesurée à +3.3 et +3.2 ST par rapport aux syllabes environnantes), tandis que les annotateurs ont considéré qu'il s'agissait d'une réinitialisation mélodique globale, et non pas de syllabes proéminentes localement. Normalement, ce type d'erreur est évité par la prise en compte de la pause

(lorsqu'une syllabe est séparée de la précédente par une pause supérieure à 250 ms, elle n'est pas comparée aux syllabes avant la pause)[16].

D'autre part, certaines proéminences détectées uniquement par l'automate constituent vraisemblablement des syllabes que les codeurs, pour différentes raisons, ont omis d'annoter comme proéminentes. L'exemple à la Figure 2 illustre une telle omission, où les codeurs n'ont pas retenu la syllabe finale du groupe « n'a pas empêché » comme proéminente, alors qu'elle présente une hauteur mélodique relative plus haute que les syllabes environnantes (+3.4 ST). Cette omission s'explique par le débit rapide de ce passage : quand le débit s'accélère, on a tendance à ne retenir que les proéminences les plus marquées, à la fois par la hauteur et la durée (la syllabe en question est brève).

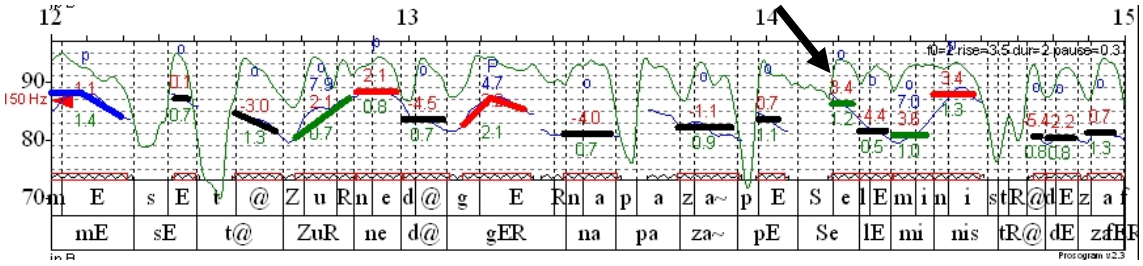

Figure 2. Extrait d'un journal radiophonique : *mais cette journée de guerre n'a pas empêché les ministres des affaires étrangères…* (JP-B). La dernière syllabe de « empêché » (sec. 14.2), proéminente par sa hauteur mélodique, a été manquée par les codeurs.

Dans ces cas, la détection automatique, de par son caractère cohérent et régulier, peut constituer un outil d'aide à l'annotation, voire augmenter la fiabilité d'une annotation manuelle qui résulte du consensus entre deux codeurs.

4.2 Proéminences détectées par les experts et omises par l'automate

Il arrive que l'analyse automatique ne détecte pas une syllabe annotée proéminente par les codeurs parce que la syllabe analysée est juste inférieure aux valeurs fixées comme seuils. À l'inverse, il arrive que les codeurs aient identifié comme proéminente une syllabe qui ne présente aucune saillance acoustique. Ces deux cas de figure vont être analysés successivement.

Lors de l'application de la détection automatique sur le corpus élargi (voir section 3.3 *supra*), différents paramètres acoustiques et différentes valeurs de seuils ont été testés afin d'aboutir à la meilleure convergence entre détections manuelle et automatique. Il arrive que des syllabes annotées manuellement comme proéminentes ne soient pas identifiées par la machine, en raison de leur **infériorité aux seuils** (durée relative < 2 ; maximum relatif de f0 < 2 ST ; montée mélodique interne < 3.5 ST et pause subséquente < 300 ms). Dans l'énoncé « donc on est restés euh six heures dans la gare », les syllabes détectées comme proéminentes par les codeurs sont légèrement en dessous des seuils (par ex. la syllabe finale de « restés » est allongée mais seulement de 1.9 alors que le seuil de durée est fixé à 2).

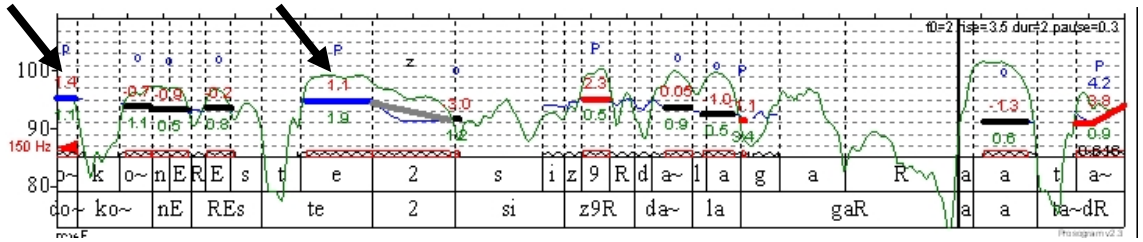

Figure 3. Extrait d'un récit de vie : *donc on est restés euh six heures dans la gare / à attendre…* (RCV-F).

La solution qui consisterait à abaisser systématiquement les seuils produirait une sur-détection de la part de l'automate et augmenterait le nombre de proéminences détectées uniquement par l'automate.

Plusieurs solutions sont à explorer pour améliorer le score de la détection automatique.

- La procédure de décision utilisée dans notre étude impose que le seuil de proéminence soit dépassé pour au moins un des paramètres retenus (durée relative / f0 max. relative / mouvement de f0 / pause). Une autre procédure consisterait à accepter comme proéminente une syllabe pour laquelle les seuils sont approchés (mais pas atteints) pour plusieurs paramètres. De cette manière, on attribuerait un score à chaque paramètre (selon que le seuil est largement manqué, presque atteint, atteint ou largement dépassé) et on calculerait un **degré de proéminence** pour chaque syllabe. Cela reviendrait à parler de contribution graduelle des paramètres.

- On peut aussi chercher à améliorer le score de détection en cherchant des seuils plus spécifiques. On pourrait par conséquent analyser séparément chaque sous-corpus (récits de vie *vs*. discours politique *vs*. etc.) afin d'optimiser les seuils de détection en fonction des « genres » de parole (vitesse d'articulation, amplitude du registre utilisé, etc.). Par exemple, dans l'extrait de récit de vie illustré à la Figure 3, les mouvements mélodiques et les allongements sont moins amples que dans les discours politiques formels, de sorte que les seuils pourraient être abaissés. Cette solution repose sur l'hypothèse qu'il pourrait y avoir des paramètres de marquage des proéminences lié à des « phonostyles » différents [voir Auchlin *et al*. soumis ; Obin, Veaux *et al.* soumis]. On risque toutefois de s'égarer si on ne tient pas compte du fait que les paramètres sont également spécifiques au locuteur.

- Enfin, on peut chercher à détecter les proéminences en faisant varier le domaine prosodique au sein duquel le caractère proéminent d'une syllabe est calculé. L'empan de traitement, actuellement fixé à 1 + 3 syllabes, pourrait être défini de manière dynamique sur la base d'une segmentation indépendante en macro-unités prosodiques, comme les *périodes* intonatives [Lacheret-Dujour & Victorri 2002].

Un cas particulier de non-détection d'une proéminence par la machine concerne les **syllabes initiales** de mots polysyllabiques (que l'on peut interpréter comme porteuses d'un accent dit « initial » ou « secondaire »). On sait qu'en français [Mertens 1987 ; Lacheret-Dujour & Beaugendre 1999], l'accent initial se caractérise souvent par une proéminence de hauteur mais sans allongement (et, par conséquent, sans mouvement mélodique intrasyllabique). Un accent initial de mot n'est par définition jamais suivi d'une pause. En outre, c'est parfois un pic d'intensité ou une modification de la qualité vocale (hyperarticulation, etc.) qui est utilisé comme marqueur de proéminence. Cette configuration particulière de l'accent initial fait qu'un seul paramètre (la hauteur relative de f0) va être utilisé par la machine, ce qui explique les fréquentes non-détections de l'automate[17]. C'est le cas sur la première syllabe de « ravagés » dans l'exemple suivant.

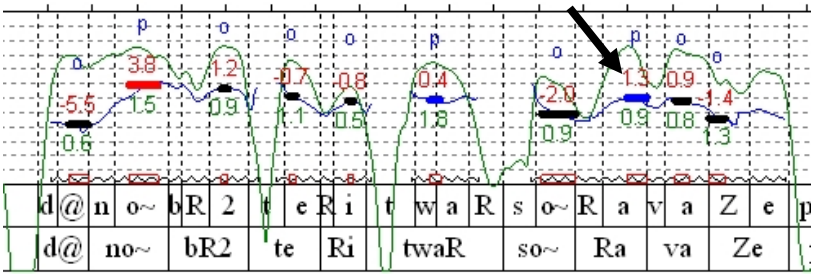

Figure 4. Extrait d'un discours politique : *de nombreux territoires sont ravagés… (DP-CH)*. La syllabe initiale de « ravagés » n'a pas été détectée à cause d'une hauteur relative de 1.3 ST (inférieure au seuil de 2 ST).

Une solution pour récupérer ces proéminences initiales de mots consisterait à abaisser le seuil de hauteur relative de f0 dans cette position syllabique particulière (initiale de polysyllabe), en injectant de l'information morphosyntaxique à partir d'un étiquetage grammatical. Nous n'avons pas implémenté ce type de solution dans le cadre de cette étude, par souhait de conserver une définition perceptive et acoustique de la proéminence[18].

Que dire, enfin, des cas où les annotateurs ont conjointement désigné une syllabe comme proéminente alors qu'elle n'approche aucun des seuils de proéminence ?

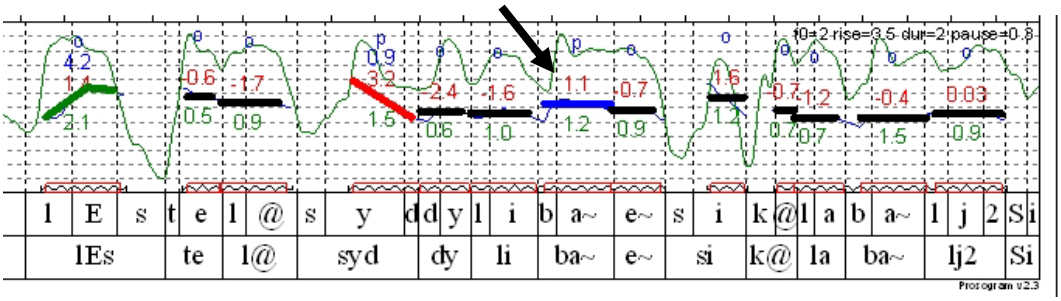

Figure 5. Énoncé extrait d'un journal parlé : *l'Est et le Sud du Liban… (JP-B)*. La deuxième syllabe de « Liban » est annotée proéminente par les codeurs indépendamment d'une saillance acoustique mesurable.

Dans l'énoncé « l'Est et le Sud du Liban ainsi que la banlieue shiite de Beyrouth ont encore été pilonnés », les annotateurs ont codé la deuxième syllabe de « Liban » comme proéminente alors qu'elle ne ressort objectivement de son contexte ni par la durée (1.1), ni par la hauteur (1.2 ST). Il est probable que la frontière de constituant syntaxique ait pu influencer le codage. Ces cas de figure, assez rares, révèlent une limite de la procédure d'annotation qui se fonde sur les enregistrements non filtrés : le codeur perçoit le contenu segmental, ou il est influencé par d'autres aspects comme la position de la syllabe ou des aspects de dynamique (attaque).

5 Conclusion

Cette étude comporte deux volets : une procédure de détection automatique des syllabes proéminentes dans un corpus avec différentes situations de parole ; une analyse des divergences entre le codage manuel et la détection automatique.

L'étude se base de l'annotation par deux experts, selon une méthodologie rigoureuse, de 3276 syllabes proéminentes dans un corpus varié de plus de 50 minutes de parole. Cette annotation de référence a servi à entraîner un automate pour réaliser une détection des syllabes proéminentes. Par rapport au « codage de référence » élaboré par les codeurs humains, 82.9% des syllabes sont détectées correctement par l'automate. Ce score peut se comparer à l'accord entre les codeurs humains, qui était de 90.67% avant résolution des cas problématiques.

L'analyse qualitative des cas de divergences entre l'annotation manuelle et la détection automatique nous a permis d'envisager une série de modifications à apporter à notre procédure de détection automatique.

- L'empan de relativisation des syllabes : le « contexte » par rapport auquel la durée, la hauteur, le mouvement d'une syllabe sont relativisés pourrait être modifié. Nous avons privilégié un empan d'analyse très local (quatre syllabes) mais l'existence de macro-unités prosodiques, démarquées entre autres par des mouvements mélodiques globaux, pourraient nous inciter à détecter les proéminences au sein d'unités de taille plus importante.
- La procédure de décision : nous avons privilégié une procédure qui repose sur le dépassement, pour une syllabe donnée, d'au moins un des quatre seuils parmi les critères acoustiques retenus. Une stratégie qui tiendrait compte de l'interaction entre ces paramètres ou d'une pondération de ceux-ci pourrait être testée également.
- L'ajout d'une annotation linguistique : la question d'utiliser ou non de l'information morpho-syntaxique (la syllabe analysée est-elle initiale ou finale de mot ?) a été posée. Jusqu'à présent, nous nous sommes limités aux critères acoustiques[19], avec les conséquences que certains types de proéminences, comme les accents initiaux (secondaires) qui présentent des caractéristiques spécifiques, sont plus difficilement identifiés.

- La spécification de paramètres et de seuils en fonction des « styles de parole » : pour calculer les taux de bons résultats et d'erreurs, les différentes variétés (régionales et stylistiques) ont été regroupées. Dans des études en cours, nous essayons de voir si, dans certaines variétés, les taux de concordance dépendent de causes de nature différente. Ainsi, dans certaines conditions de production (plus ou moins formelles), les proéminences peuvent être davantage marquées par la durée [Goldman, Auchlin *et al.* 2007 et 2008 ; Obin, Veaux *et al.* soumis ; Auchlin *et al.* soumis].

Quoi qu'il en soit, il reste à décider si proéminences détectées uniquement par les experts, ou celles détectées uniquement par l'automate, doivent, en fin de compte, être considérées ou non comme des syllabes proéminentes. Il arrive que la détection automatique incite les experts à revoir leur annotation ; ou que l'annotation des experts soit considérée comme la référence à laquelle se mesurent les performances de l'automate. Des critères univoques et explicites doivent encore être trouvés pour trancher dans les cas de discordance, si tant est que cela soit possible.

Notons pour finir que nous avons su maintenir le score global de réussite de la détection automatique, alors qu'on a augmenté l'hétérogénéité du corpus d'étude par rapport aux études précédentes. Les régularités qui émergent nous permettent de mieux comprendre comment améliorer notre outil en même temps qu'elles nous amènent à souligner que la proéminence syllabique est un phénomène complexe, difficile à modéliser.

6 Remerciements

Ce travail a 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 ; ainsi que du soutien du Fonds National Suisse de la Recherche Scientifique, qui finance le projet *La structuration interne des périodes* (dirigé par M.-J. Béguelin à l'Université de Neuchâtel, subside n°100012-113726/1). Les auteurs tiennent à remercier vivement les deux relecteurs anonymes pour leurs remarques pertinentes. Toutes les imperfections qui subsistent sont de notre responsabilité.

Références bibliographiques

Auchlin, A., Avanzi, M., Goldman, J.-P. & Simon, A.-C. (soumis). Les *phonostyles* : une description prosodique des styles de parole en français. Communication au colloque *Les voix du français : usages et représentations*, Taylorian Institution, The University of Oxford, 3-5 septembre 2008.

Avanzi, M., Goldman, J.-P., Lacheret-Dujour, A. Simon, A.-C. & Auchlin, A. (2007). Méthodologie et algorithmes pour la détection automatique des syllabes proéminentes dans les corpus de français parlé. *Cahiers of French Language Studies*, 13/2, 2-30.

Avanzi, M. Goldman, J.-P. & Simon, A.C. (en prép.). C-PROM. An Annotated Corpus for French Prosodic Studies.

Beckman, M., Hirschberg, J. & 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.

Boersma, P. & Weenink, D. (2008). Praat: doing phonetics by computer (Version 5.0). www.praat.org.

Crystal, D. (2003). *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell Publishing.

Goldman, J.-P. (2008). EasyAlign: a semi-automatic phonetic alignment tool under Praat, http://latcui.unige.ch/phonetique.

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

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

Goldman, J.-P., Auchlin, A., Simon, A.C. & Avanzi, M. (2007). Phonostylographe, un outil de description des phonostyles prosodiques. Chroniques radiophoniques et style lu. *Nouveaux Cahiers de Linguistique Française*, 28, 219-237.

Goldman, J.-P., Auchlin, A., Avanzi, M. & Simon, A.C. (2008). ProsoReport. An Automatic Tool for Prosodic Description. *Proceedings of Speech Prosody 08*, Campinas, May 6-9, 2008.

Grabe, E., Post, B. & Nolan, F. (2001). Modelling Intonative Variation in English: The IViE System. In S. Puppel & G. Demenko (eds), *Prosody 2000*. Poznan: Adam Mickiewicz University, 51-58.

Grobet A. & Simon, A.C. (2002). Différents critères de définition des unités prosodiques maximales. *Cahiers de linguistique française* 23, 143-163.

Gussenhoven, C. (2002). Intonation and Interpretation. *Proceedings of Speech Prosody'02*, Aix-en-Provence, April 11-13, 2002, 47-57.

Hermes, D.J. (2006), Stylization of Pitch Contours, in Sudhoff S. *et al.* (eds). *Methods in Empirical Prosody Research*, Berlin-New York, Walter de Gruyter, 29-61.

Hirschberg, J. (2002). The Pragmatics of Intonational Meaning. *Proceedings of Speech Prosody'02*, Aix-en-Provence, April 11-13, 2002, 65-68.

Jun, S. A. (ed.) (2005). *Prosodic Typology – The Phonology of Intonation and Phrasing*. Oxford University Press.

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.

Lacheret-Dujour, A. (à par.). Séquençage et mouvements intonodiscursifs en français parlé. *Cahiers de praxématique*.

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

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. (2004). Le prosogramme : une transcription semi-automatique de la prosodie. *Cahiers de l'Institut de Linguistique de Louvain*, 30/1-3, 7-25.

Mertens, P. & d'Alessandro, Ch. (1995). Pitch contour stylization using a tonal perception model. *Proc. 13th International Congress of Phonetic Sciences*, Vol. 4, 228-231.

Morel, M.-A. & Danon-Boileau, L. (1998). *Grammaire de l'intonation. L'exemple du français oral*. Paris-Gap : Ophrys.

Morel, M., Lacheret-Dujour, A., Lyche, C. & Poiré, F. (2006). Vous avez dit proéminences ? *Actes des 26èmes journées d'étude sur la parole (JEP'06)*, Dinar, 12-16 juin 2006, 183-186.

Obin, N., Rodet, X. & Lacheret-Dujour, A. (2008). Prominence model: a probabilistic framework. *Proceedings of the 33rd International Conference on Acoustics, Speech, and Signal Processing (ICASSP'08)*, Las Vegas, April 01-04, 2008.

Obin, N., Goldman, J.-P., Avanzi, M. & Lacheret-Dujour, A. (2008). Comparaison de trois outils de détection automatique des proéminences en français parlé. *Actes des 27èmes journées d'étude sur la parole (JEP'08)*, Avignon, 8-13 juin 2008.

Obin, N., Veaux, C., Rodet, X. & Lacheret-Dujour, A. & Simon, A.C. (soumis). A Method for Automatic and Dynamic Estimation of Discourse Genre Typology with Prosodic Features. *Interspeech'08*, Brisbane, September 22-26, 2008.

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

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

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

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

---

[1] Un workshop organisé en 2004 dans le cadre du projet international Phonologie du Français Contemporain (www.projet-pfc.net) a suscité une première étude visant à évaluer le degré d'accord entre experts annotant les proéminences prosodiques dans un extrait de parole spontanée [Poiré 2006]. Cette étude a été complétée ultérieurement [Morel *et al.* 2006].

[2] Par ailleurs, nous utilisons notre corpus d'étude pour essayer de dégager automatiquement les phénomènes prosodiques permettant de discriminer automatiquement différentes variétés de français – principalement des variétés stylistiques (degré de formalité et conditions de production plus ou moins spontanée de la parole). Voir [Auchlin *et al.* (soumis)] et [Goldman, Auchlin, Simon & Avanzi 2007].

[3] Ce script, élaboré par J.-P. Goldman, est dénommé **ProsoProm**. Il repose sur le script **Prosogram** de P. Mertens [2004], dont il constitue un prolongement (*cf.* section 3). Voir http://bach.arts.kuleuven.be/pmertens/prosogram/

[4] La plupart des systèmes de transcription prosodique, qu'ils engendrent une interprétation phonologique (comme ToBI [Beckman *et al.* 2006]) ou non (*cf.* p. ex. Le système IViE [Grabe *et al.* 2001] ou IVTS [Grabe *et al.* 2001 ; Post *et al.* 2006]), partagent le point de vue selon lequel l'identification des proéminences est une étape indispensable pour la mise au jour de la structure prosodique. Reste que, si cela est vrai pour le français et les autres langues à intonation, il semblerait que pour les langues fonctionnant différemment, la question mérite d'être posée et discutée [Jun 2005].

[5] Notre définition des proéminences recoupe celle de 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 LENGTH, PITCH, STRESS and inherent SONORITY are all factors which contribute to the relative prominence of a UNIT ».

[6] L'annotation des proéminences est réalisée directement dans Praat, et la segmentation en syllabes est obtenue grâce au script **EasyAlign**, mis au point par [Goldman 2008]. L'alignement phonétique et la syllabation ont a été vérifiés manuellement.

[7] Cette catégorie correspond aux chevauchements (audibles pour un codeur humain mais inexploitables par un système automatique) et aux réalisations paraverbales (rires, toux, etc.) ou non verbales (bruit) jugées non pertinentes à coder du point de vue des proéminences syllabiques.

[8] A noter que la distinction entre « proéminence forte » et « proéminence faible » a une fonction heuristique : elle incite les codeurs à développer une écoute plus fine en ne se limitant pas à l'annotation des proéminences les plus marquées. Cependant, nous ne prétendons pas qu'il existerait deux degrés de proéminences (ni un, ni trois ou quatre…). Lors de l'analyse, les syllabes « P » et « p » sont regroupées en une seule catégorie, qui s'oppose à « 0 ».

[9] Cotes IRT (interviews) et ITI (itinéraires), *cf.* Tableau 2.

[10] Si l'on en croit les recensions de Tamburini & Caini [2005], et bien que de telles comparaisons soient parfois hasardeuses.

[11] La durée minimale du noyau a été mise à 9 ms.

[12] Voir cependant section 3.2 : ces syllabes « sans noyau (stylisé) » sont néanmoins traitées pour la détection de proéminences sur la base des paramètres de durée et de pause subséquente.

[13] Les seuils de glissando ont été établis à partir des mesures de Rossi [1978] et de [Mertens & d'Alessandro 1995].

[14] Notre empan de traitement, qui couvre 3 syllabes autour de la syllabe analysée, est fixé *a priori* : chaque syllabe soumise à l'analyse est comparée aux syllabes qui l'entourent. L'empan de traitement diffère en cela du « domaine d'implémentation » utilisé par exemple par [Post *et al.* 2006] dans le système d'annotation prosodique IVTS. Dans ce dernier cas, le domaine d'implémentation est construit *a posteriori*, sur la base d'une identification préalable des syllabes proéminentes. L'outil **ProsoProm** pourrait servir à définir automatiquement les domaines d'implémentation pour un système d'annotation comme IVTS.

[15] Le logiciel ProsoProm dispose d'une interface conviviale qui permet à l'utilisateur de régler lui-même les paramètres et les seuils des paramètres. Cette flexibilité est utile si l'on veut tester des hypothèses ou adapter la procédure de détection à différentes variétés de français, par exemple.

[16] Cette « sécurité » n'a pas fonctionné dans le cas présent à cause d'une erreur de stylisation sur la syllabe « et », dont le noyau est très bref.

[17] Actuellement, l'automate ne tient pas compte d'une pause particulièrement longue (> 300 ms) qui précèderait la syllabe initiale de mot, pas plus qu'il ne tient compte d'un éventuel allongement affectant l'attaque de la syllabe (VOT allongé).

[18] Un autre cas de proéminence difficile à identifier automatiquement concerne les proéminences finales de groupe intonatif caractérisées par un contour mélodique grave. Le pré-traitement de ces syllabes finales est déjà problématique, car la voix est parfois dévoisée ou « creaky », ce qui rend difficiles la segmentation et la stylisation (la partie intense et stable de la syllabe est souvent très réduite). Si ces proéminences sont détectées, c'est grâce à la pause subséquente (quand elle dépasse 300 ms) ou à leur allongement relatif. Les deux autres paramètres, concernant la f0, ne sont jamais saillants. En fait, les syllabes finales de niveau bas posent la question même de la « proéminence », puisqu'elles sont souvent réalisées avec une faible énergie et une chute de f0.

[19] Le seul recours à de l'information linguistique se trouve dans la tire « delivery » qui identifie des marques du travail de formulation, comme les hésitations.