Journal of Speech Sciences 7(2): 79-91. 2018.
Available at: http://revistas.iel.unicamp.br/joss

GENERATION AUTOMATIQUE DE LA STRUCTURE PROSODIQUE EN FRANÇAIS

Automatic prosodic structure from transcribed speech recordings in French

MARTIN, Philippe[1]*

[1]Université Paris Diderot

Abstract: *An automated process for building a prosodic structure form transcribed speech recordings in French is presented, based on the incremental prosodic model [1, 2, 3]. In this model, the prosodic structure is defined incrementally by dependency relations instantiated by melodic contours located on the last syllable of the last word of stress groups, subject to a rhythmic constrain limiting the gap between successive stressed syllable to a 250-1250 ms range. Although they frequently contain lexical words (noun, verb, adverb, adjective), stress groups in French can also include only grammatical words (pronoun, conjunction, preposition). Melodic contours are phonologically defined from their melodic rise or fall and their glissando value ensuring their function as dependency markers between stress groups.*
The algorithm proceeds from an orthographic transcription as follows:
1. Automatic segmentation of the orthographic text into IPA and word tiers
2. Automatic annotation of stressed vowels in three classes (followed by 250 ms silence, above the glissando threshold and lexical category based)
3. Assignment of melodic contours from fundamental frequency values at stressed vowels boundaries.
Comparisons with automatic and manual stressed syllable annotation on existing corpora are given, showing the validity of the phonological rules implemented in the algorithm.

Keywords: Prosodic structure; French; stressed syllables; phrasing; WinPitch.

1 L'annotation des syllabes accentuées : mission impossible ?

Le français est une langue dépourvue d'accent lexical. Alors que les autres langues romanes comme l'italien ou l'espagnol possèdent un accent lexical généralement interne au mot et marquant une frontière morphologique (par exemple entre racine et suffixe ou entre suffixes [4]), le français ne possède, en plus des marques d'emphase ou d'insistance, qu'un accent dit « tonique » placé sur la syllabe finale de certains mots. Une croyance répandue veut que les groupes accentuels ainsi délimités par l'accent tonique constituent des groupes de sens, mais en réalité, c'est seulement un critère rythmique qui en détermine la distribution. Ce critère contraint les groupes accentuels selon leur durée d'énonciation, qui doit être comprise entre 250 ms et 1250 ms environ [5]. Il résulte de cette contrainte de durée que le nombre de syllabes, et donc de mots, contenu dans un groupe accentuel dépend du débit de parole du locuteur, ou du lecteur, que ce soit en production orale ou silencieuse (i.e. quand on lit ou qu'on se parle « dans sa tête »). De par la durée maximale de 1250 ms des groupes accentuels, plus on lit ou on parle vite, plus le nombre de mots contenu dans les groupes accentuels est grand, et inversement, plus on parle ou lit lentement, moins les groupes accentuels contiennent de syllabes et donc de mots, jusqu'à la prononciation très lente en syllabes détachées, impliquant des groupes accentuels d'une seule syllabe. Ainsi la phrase *la banlieue de Paris*, prononcée avec un débit moyen de l'ordre de 4 syllabes / seconde, sera réalisée avec deux syllabes accentuées terminant les mots *ban**lieu**e* et *Pa**ris***, alors qu'un débit plus rapide, par exemple de 7 syllabes / seconde, n'en créera qu'une seule : *la banlieue de Pa**ris***. La correspondance des groupes accentuels définis par les syllabes accentuées successives avec les unités syntaxiques sera donc différente, *la banlieue* et *de Paris* dans le premier cas et *la banlieue de Paris* dans le second. Dans ces deux réalisations, les groupes accentuels sont alignés avec des « groupes de sens », dont le contenu dépend de la vitesse d'élocution ou de lecture. Cette variation de débit modifie aussi la durée moyenne des syllabes

*Corresponding author: philippe.martin@utoronto.ca

contenues dans chaque groupe [5], permettant de comprimer jusqu'à neuf ou dix syllabes prononcées (ou lues) en moins de 1250 ms, leur durée moyenne se rapprochant alors de la limite théorique de perception, soit environ 100 ms [6].

Le problème pour un annotateur de l'accent tonique en français est donc de s'adapter au débit de parole de l'enregistrement. Ainsi, habitué à un débit moyen de par exemple 4 syllabes/seconde, un annotateur aura tendance à percevoir comme accentuée une syllabe qui serait dépourvue des indices acoustiques attendus dans une phrase prononcée par un locuteur à débit rapide. Inversement, habitué à parler très vite, le même annotateur ne considèrera pas comme accentuée une syllabe qui le serait du point de vue acoustique, par un allongement par exemple. La perception de l'annotateur sera influencée par un processus de prédiction, tendant à détecter les syllabes accentuées aux endroits où il les aurait placées en lisant ou en parlant non pas avec le débit du locuteur mais avec le sien. Ainsi, M. Avanzi [7], confronté à l'incertitude dans l'annotation des syllabes accentuées, décrit en détail une procédure complexe impliquant deux experts, éventuellement aidés d'un troisième en cas de désaccord entre les deux premiers. Même avec ce protocole, l'accord entre les annotateurs varie entre 60% et 80%.

Pour affranchir l'annotation de la dépendance au débit de parole, on a élaboré plusieurs systèmes de détection basés sur les seuls paramètres acoustiques des syllabes accentuées en français. Le plus souvent, ces détections automatiques opèrent de bas en haut (bottom-up), recherchant dans les enregistrements des variations acoustiques syllabiques significatives en durée, fréquence fondamentale et intensité (pour des exemples récents, voir [8, 9]). Cette approche bottom-up notent fréquemment comme non accentuées des syllabes pourtant perçues comme accentuées par une annotation manuelle.

Dans la même veine, Christodoulides et Avanzi [10] ont mis en œuvre un détecteur automatique de proéminence (intégrant également l'accent d'emphase) par des méthodes d'apprentissage automatique appliquées à un grand corpus de 11 heures de durée. Les auteurs utilisent un ensemble élargi de paramètres acoustiques censés être appropriés pour différencier les syllabes proéminentes des autres syllabes (durée syllabique minimale et maximale, fréquence fondamentale moyenne, minimale et maximale, intensité maximale, équilibre spectral, partie de l'étiquette vocale, présence et durée des pauses, structure syllabique, position de la syllabe dans le mot, etc.). Leurs meilleurs résultats atteignent un niveau d'identification de 90%, ce taux étant basé sur une annotation manuelle d'experts admise comme référence (mais dont les jugements pourraient être sujets à caution comme on l'a vu).

Compte tenu de ces difficultés, il semble que la détection des syllabes accentuées devrait procéder par une approche différente qui ne soit pas uniquement basée sur l'analyse acoustique, surtout si leur validité est évaluée par une analyse perceptive. Un examen plus attentif des caractéristiques de l'accent tonique en français devrait mener à une solution plus satisfaisante pour une annotation des syllabes accentuées, étape essentielle à la génération automatique de la structure prosodique de la phrase.

2 Accent tonique et structure prosodique

Loin d'être la cerise sur le gâteau syntaxique, l'intonation de la phrase constitue un élément essentiel du système linguistique, permettant au locuteur de structurer rapidement la parole prononcée, et à l'auditeur de hiérarchiser et de structurer efficacement les segments de parole reconnus et identifiés, ce qui ne serait pas possible en temps réel dans la parole continue en se basant uniquement sur les relations syntaxiques existant entre les mots [3]. On rappellera ici quelques points théoriques qui font le plus souvent l'objet d'une incompréhension récurrente, et parfois agressive, de la part de spécialistes de la prosodie, frileux à l'idée de se distancier de la doxa autosegmentale-métrique dans ce domaine. Ces principes sont à la base

de l'algorithme de segmentation automatique en groupes accentuels dont l'organisation hiérarchique constitue la structure prosodique.

Le premier point a trait aux relations entre l'intonation et la syntaxe. Alors qu'une approche plus traditionnelle, influencée en cela par la linguistique de l'écrit, envisage la structure prosodique comme dérivée de la syntaxe, la vision dont il s'agit ici adopte un point de vue diamétralement opposé, selon lequel les constructions syntaxiques produites dynamiquement au cours du temps sont précédées et non pas suivies des constructions prosodiques. Selon ce processus, le locuteur, pour pouvoir mettre en place une construction syntaxique, doit au préalable sélectionner une construction prosodique. De nombreuses analyses de l'oral spontané et en particulier des « erreurs » relevées dans le spontané [11, 12], mais aussi de neurolinguistique (Arnal et Giraud, [13]) permettent d'étayer cette hypothèse, qui reste surprenante voire iconoclaste pour certains. On en trouvera une discussion détaillée dans [3].

Le second point porte sur la notion d'accent lexical, pour lequel il semble parfois difficile (surtout pour les linguistes anglophones) d'admettre qu'il n'existe pas en tant que tel dans certaines langues comme le français ou le coréen. Le français, à la différence du latin et des autres langues romanes comme l'espagnol ou l'italien, a progressivement perdu au cours de son évolution à partir du latin [14] les syllabes posttoniques des mots lexicaux, c'est-à-dire des mots de classe ouverte que sont les verbes, les noms, les adjectifs ou les adverbes (par opposition aux mots grammaticaux de classe fermée, conjonctions, pronoms, prépositions). Il en reste toutefois des traces dans certains parlers régionaux, ou dans des enregistrements anciens où on entend encore des mots avec un suffise ***-a**tion* accentué sur la pénultième comme dans *l'organis**a**tion, la n**a**tion* [15].

On pourrait alors se demander quelle est la fonction des syllabes accentuées, en dehors d'indiquer éventuellement le caractère emphatique ou d'insistance de certains mots du discours (comme dans ***ex**traordinaire* ou ***in**croyable* accentué sur la première syllabe). Or, une simple observation intuitive permet de constater que les syllabes accentuées en français sont toujours réalisées sur la syllabe finale de certains mots. Quoique le plus souvent ces mots accentués sur leur syllabe finale appartiennent au groupe de classe ouverte (les mots lexicaux), ils peuvent tout aussi bien terminer des mots grammaticaux, de sorte que des groupes accentuels en français peuvent ne pas contenir du tout de mots lexicaux. Cette question a fait l'objet de recherches empiriques récentes qui ont apparemment laissé la question ouverte [8]. Or, le simple fait que tout locuteur du français restitue les syllabes accentuées selon les principes énoncés plus haut aussi bien en parole oralisée que silencieuse suggère que la perception des syllabes accentuées ne dérive pas directement du traitement des caractéristiques acoustiques spécifiques de la parole, telles que la durée de la voyelle, le changement de fréquence fondamentale ou la modulation d'intensité, les paramètres prosodiques classiques souvent mentionnés dans la littérature comme paramètres de l'accent. Dès lors, la perception des syllabes accentuées peut être considérée comme le résultat d'un mécanisme d'identification comparant les caractéristiques acoustiques réelles des syllabes avec une position prédite dérivée de la connaissance de la langue [13].

Des praticiens de l'enseignement du français langue étrangère (FLE), attentifs et peu influencés par des a priori théoriques, ont fait depuis longtemps des observations conduisant à considérer que l'accent en français, et donc le découpage en unités accentuelles (le phrasé), est de nature rythmique et tend à équilibrer les durées séparant des groupes accentuels successifs [16]. Cet effet d'eurythmie affecte donc les durées moyennes des syllabes, de sorte que plus le nombre des syllabes entre deux syllabes accentuées successives est grand, plus ces syllabes sont courtes et inversement [5].

Ce qui peut paraitre encore plus surprenant, c'est qu'il semble impossible (sauf pour une machine tel un synthétiseur texte-parole) d'énoncer une séquence de syllabes, donc de mots, sans en accentuer au

moins une si la durée d'énonciation de la séquence dépasse une certaine valeur, de l'ordre de 1250 à 1450 ms. Si l'intervalle entre deux syllabes accentuées successives dépasse ce seuil, tout locuteur du français insérera au moins une syllabe accentuée supplémentaire dans cet intervalle. Ainsi, la phrase *la ville de Paris est agréable à vivre en été*, semble très difficile à lire sans réaliser au moins une autre syllabe accentuée en plus de la syllabe finale sur ***été***, par exemple sur *Pa**ris***, alors qu'aucune contrainte linguistique apparente ne le force à la faire.

Une autre observation souvent évoquée concerne les syllabes finales des mots, de quelque classe que ce soit, qui sont perçus comme accentuées en français lorsqu'elles sont suivies d'une pause silencieuse (voir par exemple [9]). En fait, l'analyse instrumentale [3, 5] montre que cette pause doit avoir une durée d'au moins 250 ms pour rendre la syllabe perceptivement accentuée, et ce indépendamment de ses caractéristiques acoustiques mélodique, de durée ou d'intensité. Il est du reste facile de tester ce mécanisme dans la réalisation détachée des syllabes d'un mot à des fins stylistiques d'insistance : ***la ban lieue de Pa ris***, chacune des syllabes de détachée paraissant alors accentuée. Cette propriété ne s'observe pas dans des langues à accent lexical (sauf bien sûr dans des mots accentués sur leur dernière syllabe), puisqu'une autre position phonologique non finale est attendue par l'auditeur, alors qu'en français, la syllabe finale du groupe accentuel est phonologiquement accentuée et est nécessairement attendue comme telle.

Il est une autre observation empirique aisée à vérifier. Lorsque deux syllabes successives et accentuées sont séparées de moins de 250 ms, la première n'est plus perçue comme accentuée, quand bien même elle porterait les mêmes caractéristiques acoustiques que la syllabe accentuée suivante. Une phrase comme *le travail de nuit nuit* ou *elle aime le café chaud* voit la compréhension impossible ou modifiée selon qu'il y a ou non une pause suffisante entre les deux syllabes accentuées finales : ce qui oppose *le travail de nuit **nuit*** et *le travail de **nuit** # **nuit*** d'une part, et *elle aime le café **chaud*** ("c'est le café chaud qu'elle aime") et *elle aime le ca**fé** # **chaud*** ("c'est chaud qu'elle aime le café") d'autre part.

Un dernier point concerne la notion même de structure prosodique. Alors que le modèle autosegmental-métrique considère essentiellement l'organisation hiérarchique des unités prosodiques comme résultant d'une interaction entre des évènements mélodiques avec des unités syntaxiques, tels que révélés en fin de groupe syntaxique [17], l'approche incrémentale-dépendantielle retenue ici considère la structure prosodique comme une construction a priori autonome, indiquée par des événements mélodiques à l'endroit des syllabes accentuées (et des syllabes finales de groupes pour les langues pourvues accent lexical). Ces évènements sont instanciés par des contours mélodiques, placés sur les voyelles des syllabes accentuées, et dont la fonction est d'indiquer des relations de dépendance entre les unités minimales prosodiques, les *mots* prosodiques, que sont les groupes accentuels. Cette relation de dépendance opère « vers la droite », c'est-à-dire vers un mot prosodique situé plus loin dans la phrase. Il peut cependant y avoir une dépendance « à gauche », i.e. envers une unité qui précède, dans des constructions appelées *thème-propos*, comme dans des exemples prototypiques tels que *à la **caiss**e ils se **pès**ent* [17] où la syllabe accentuée de ***pès**ent* est pourvue d'un contour mélodique plat, placé après le contour terminal conclusif porté par la syllabe accentuée de ***caiss**e*. Cette notion de dépendance apparait en filigrane dans les dénominations de *continuation majeure* et de *continuation mineure* de Delattre, dans un article souvent cité [19].

Formellement, les contours mélodiques phonologiques sont définis comme suit :

C0 contour terminal conclusif déclaratif (Ci étant la version interrogative), descendant et bas

C1 contour de continuation majeure, montant et de variation mélodique supérieure au seuil de glissando [19]

C2 contour de continuation mineure, descendant et de variation mélodique supérieure au seuil de glissando

Cn contour neutralisé, montant ou descendant mais variation mélodique inférieure au seuil de glissando

Ces contours indiquent des relations de dépendance définissant les regroupements successifs des groupes accentuels de manière à former la structure prosodique :

Cn -> {C2, C1, C0} Le contour Cn marque une dépendance « à droite » envers un contour quelconque C2, C1 ou C0

C 2 -> {C1} Le contour C2 descendant, marque une dépendance « à droite » envers un contour C1 montant (principe du contraste de pente)

C 1 -> {C0} Le contour C1 montant, marque une dépendance « à droite » envers un contour terminal C0 descendant et bas (principe du contraste de pente)

On trouvera des détails sur les principes de la structure prosodique incrémentale-dépendantielle par exemple dans [3]. Le choix du glissando comme paramètre acoustique dans la définition des contours résulte tout naturellement de principe de contraste de pente mélodique propre à l'intonation du français. Pour que ce principe puisse être mis en œuvre, il faut évidemment que les variations mélodiques montantes et descendantes soient effectivement perçues.

3 Un algorithme montant et descendant (bottom-up et top-down)

Pour innover par rapport aux approches *bottom-up* opérant à partir des seules données acoustiques, on propose un algorithme à la fois *top-down* et *bottom-up* basé sur les propriétés accentuelles citées plus haut, en retenant les règles suivantes :

1. Toute syllabe suivie d'un silence de plus de 250 ms est accentuée (propriété de l'accent tonique)
2. Toute syllabe accentuable avec changement de F0 au-dessus du seuil de glissando est accentuée (définition des contours mélodiques définissant la structure prosodique)
3. Toute syllabe finale d'un nom, adjectif, verbe ou adverbe est accentuable (définition classique du groupe accentuel contenant un seul mot lexical)
4. Si deux syllabes accentuables ou accentuées successives sont séparées par moins de 250 ms, la première n'est pas accentuée (durée minimale du groupe accentuel)
5. Si deux syllabes accentuées consécutives sont séparées de plus de 1250 ms en parole continue, au moins une syllabe accentuable dans cet intervalle est accentuée (durée maximale du groupe accentuel). Celle ayant avec la plus grande durée est accentuée.
6. Une syllabe accentuable doit être réalisée dans n'importe quelle durée de fenêtre a) égale à la durée moyenne de l'accentuation (eurythmie) ou b) supérieure à 1250 ms (durée maximale des groupes accentuels). L'eurythmie est mise en œuvre en évaluant la moyenne des durées des groupes accentuels dans une fenêtre glissante incluant les trois derniers groupes accentuels. Une syllabe accentuée supplémentaire sélectionnée selon sa durée est ajoutée en cas d'absence d'accent dans une fenêtre eurythmique donnée.

4 L'algorithme

L'algorithme de détection des syllabes accentuées en français procède de manière dynamique, de « gauche à droite », c'est-à-dire selon une fenêtre temporelle glissante. Opérant à la fois en bottom-up et en top-down, à partir des informations suivantes :

1. Une transcription orthographique

2. Une phonétisation segmentée (qui s'obtient dans WinPitch [23] par un programme basé sur un alignement Viterbi des tronçons à segmenter avec une parole de synthèse dans la langue désirée, dans ce cas le français standard ou le français québécois)

3. Un étiquetage des mots en catégories syntaxiques, obtenues par consultation d'un lexique de quelque 142700 formes (dérivée de Lexique3, [21]) suivi d'une désambiguïsation par trigrammes. Les catégories syntaxiques sont NOM, VER, ADV, AUX, ADJ, ART, CON, POS, DET, PRO, XXX (XXX pour les groupes accentuels inachevés). Les 5 premières catégories ont une syllabe finale accentuable.

4. Une courbe mélodique (ou une annotation mélodique dérivée)

Les différentes classes de voyelles des syllabes accentuées sont obtenues par :

1. Classe 1, segments vocaliques dont la syllabe est suivie par un silence de plus de 250 ms

2. Classe 2, segments vocaliques de glissando supérieur au seuil (0.32 est le paramètre adopté dans le calcul du glissando)

3. Classe 3, segments vocaliques des syllabes finales des mots de catégorie syntaxique VER, ADV, ADJ, NOM, AUX.

Chaque classe d'accentuation est indiquée sur la courbe mélodique correspondante par un surlignage de couleur différente (classe 1 rouge, classe 2 bleu et classe 3 vert). Les contours mélodiques liés aux syllabes accentuées supplémentaires déterminées par application de la règle eurythmique ou la règle de 1250 ms sont surlignées en marron. Un fichier Excel détaillant les différents paramètres peut être obtenu en un seul clic de souris (Fig. 2).

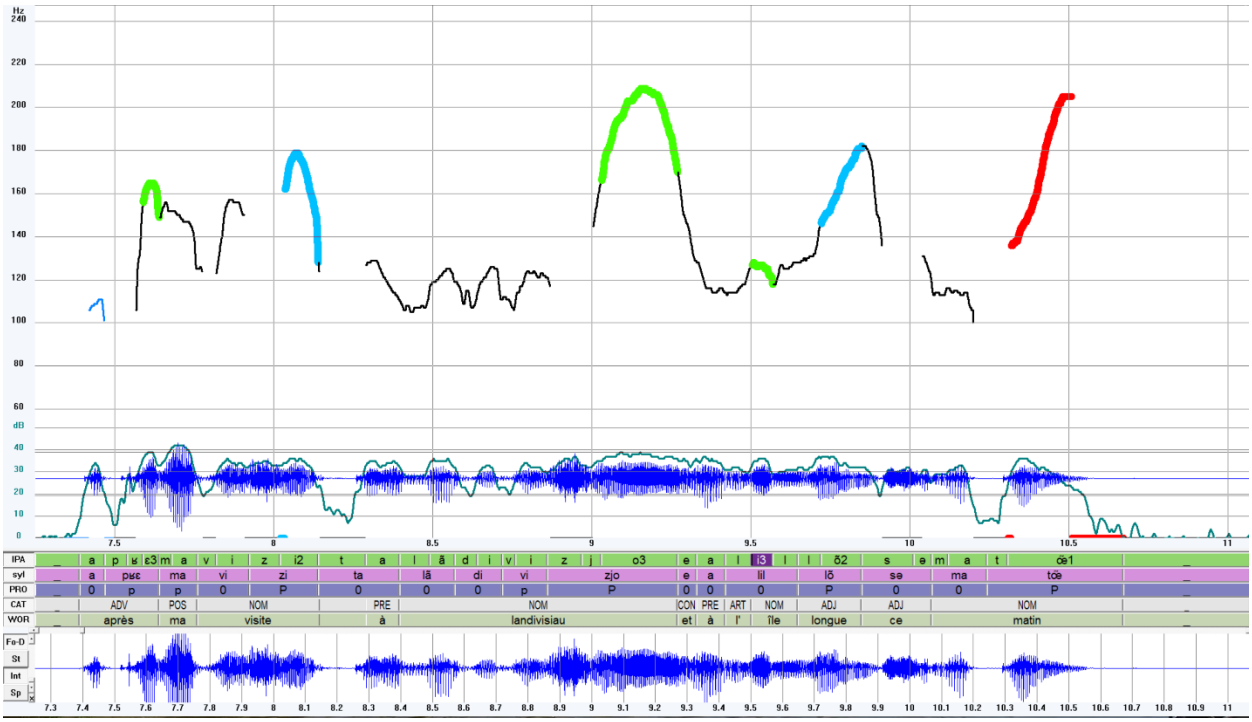

Figure 1 : Surlignage de couleur des différentes classes de syllabes accentuées de l'extrait *aprè**s*** [classe 3] *ma visi**te*** [classe 2] *à Landivisi**au*** [classe 3] *et à l'**î**le* [classe 3] *l**on**gue* [classe 2] *ce mat**in*** [classe 1] (Corpus M201, Rhapsodie [22]).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
156		k	20.540	0.062												
157		d	20.602	0.082	dy	20.602	0.184	0	20.602	0.184	ART	20.602	0.184	du	20.602	0.184
158		y	20.685	0.102												
159		l	20.787	0.051	ljê	20.787	0.248	p	20.787	0.248	NOM	20.787	0.248	lien	20.787	0.248
160		j	20.839	0.047												
161		ê2	20.886	0.149												
162		d	21.035	0.062	di	21.035	0.158	0	21.035	0.158	ADJ	21.035	0.657	direct	21.035	0.657
163		i	21.098	0.096												
164		ʁ	21.194	0.057	ʁɛkt	21.194	0.498	P	21.194	0.498						
165		ɛ2	21.251	0.169												
166		k	21.420	0.109												
167		t	21.530	0.161												
168		k	21.692	0.082	ki	21.692	0.152	0	21.692	0.152	PRO	21.692	0.152	qui	21.692	0.152
169		i	21.775	0.069												
170		y	21.844	0.089	y	21.844	0.089	0	21.844	0.089	VER	21.844	0.256	unit	21.844	0.256
171		n	21.934	0.065	nil	21.934	0.259	p	21.934	0.259						
172		i3	22.000	0.100												
173		l	22.101	0.092							ART	22.101	0.092	le	22.101	0.092
174		p	22.194	0.050	pʁe	22.194	0.170	0	22.194	0.170	NOM	22.194	0.429	président	22.194	0.429
175		ʁ	22.244	0.039												
176		e	22.284	0.080												
177		z	22.364	0.079	zi	22.364	0.129	0	22.364	0.129						
178		i	22.444	0.050												
179		d	22.494	0.050	dã	22.494	0.129	p	22.494	0.129						
180		ã3	22.544	0.079												
181		d	22.624	0.030	dla	22.624	0.130	0	22.624	0.130	ART	22.624	0.030	de	22.624	0.030
182		l	22.654	0.050							ART	22.654	0.100	la	22.654	0.100
183		a	22.704	0.050												

Figure 2 : Extrait du fichier Excel après positionnement automatique des syllabe accentuées, détaillant les paramètres des différentes classes d'annotation accentuelle.

5 Un outil de recherche

Plus qu'un programme de segmentation en phrasé et de génération automatique de la structure prosodique, l'algorithme implémenté dans WinPitch [23] se veut comme un outil d'évaluation des hypothèses présentées plus haut, et en particulier celles portant sur 1) la perception d'une syllabe accentuée finale suivie d'au moins 250 ms de silence, 2) la desaccentuation perçue sur la première syllabe accentuable de classe 2 ou 3 suivie à moins de 250 ms d'une autre syllabe accentuée, 3) la perception d'au moins une syllabe accentuée dans un intervalle sans accentuation détectée de plus de 1250 ms et 4) les segments vocaliques de moins de 50 ms sont notés comme non accentuables. Un algorithme dynamique applique ces différentes contraintes et le signale par un codage couleur aux fins d'examen des paramètres impliqués.

Par des commandes ergonomiques, l'utilisateur peut modifier différents paramètres et observer immédiatement l'effet de ces changements sur l'écran de WinPitch. Il s'agit donc d'un système interactif, les modifications peuvent porter sur la segmentation sur les différentes couches de transcription (transcription phonétique API, segmentation en syllabes, mots, étiquetage syntaxique de chaque mot). Les frontières temporelles de ces segments peuvent être modifiées à la souris, avec affichage simultané des courbes oscillographique, d'intensité et de fréquence fondamentale, ainsi que d'un spectrogramme. Les erreurs fines ou grossières d'une segmentation automatique sont ainsi facilement corrigées si nécessaire. De même, un clic de souris permet d'afficher les catégories syntaxiques alternatives pour un mot donné, suppléant ainsi à une déficience de la correction par trigramme. Les trigrammes relatifs à un fichier donné peuvent être sauvegardés et réinstallés pour une utilisation ultérieure.

Des tests ont été effectués sur des exemples de parole lue et spontanée. La plupart des erreurs évidentes, c'est-à-dire celles qui ne correspondent pas à l'intuition d'un locuteur natif, sont dues soit à une catégorisation syntaxique incorrecte, soit à une erreur de segmentation syllabique. Une autre source

d'inexactitude est liée aux erreurs de suivi de fréquence fondamentale pour la détection de l'accent de classe 2. Pour en minimiser les effets, les valeurs de glissando sont obtenues à partir d'une approximation par moindres carrés de la courbe de fréquence fondamentale, ce qui permettant des mesures relativement fiables pour des enregistrements bruités.

6 Exemples d'annotation automatique

Les figures suivantes donnent quelques exemples d'annotation obtenues par l'algorithme, illustrant l'application de différentes règles acoustiques et rythmiques (glissando Fig. 3, durée minimale des groupes accentuels Fig. 4, durée maximale des groupes accentuels Fig. 5 et 6, et structure prosodique Fig. 7).

6.1 Glissando

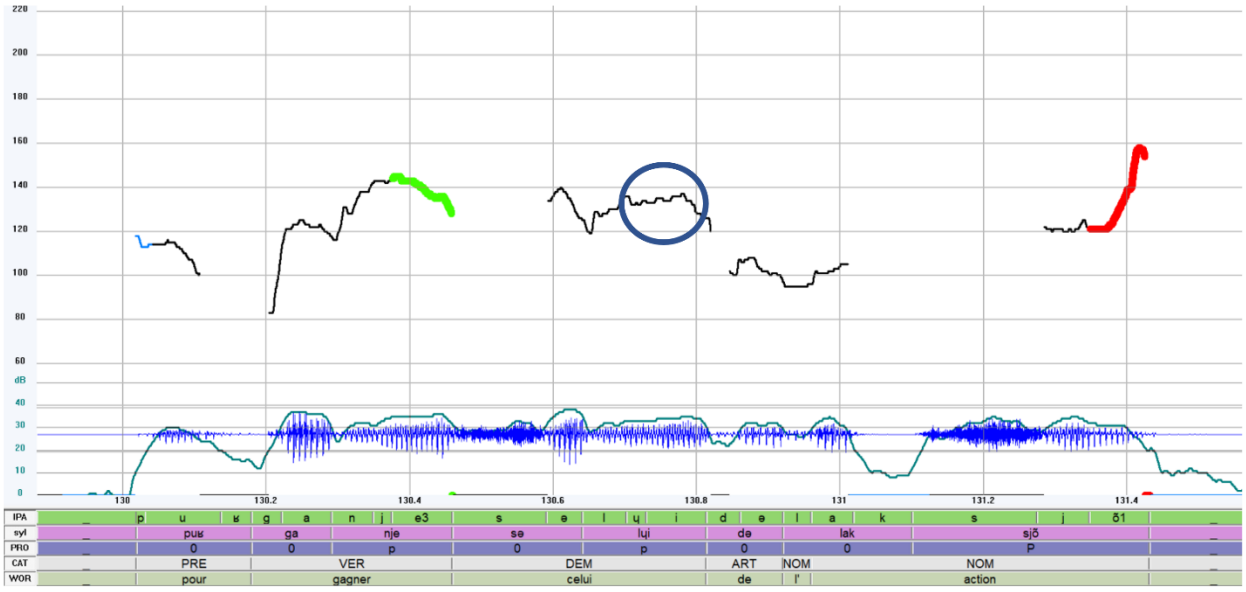

Figure 3 : La syllabe finale de *celui* (cerclée) de l'extrait *pour gagner celu**i** de l'acti**on*** n'est pas marquée comme accentuée car d'une part elle appartient à la catégorie des démonstratifs non accentuables et d'autre part le mouvement mélodique sur sa voyelle finale est inférieur au seuil de glissando.

6.2 Règle des 250 ms

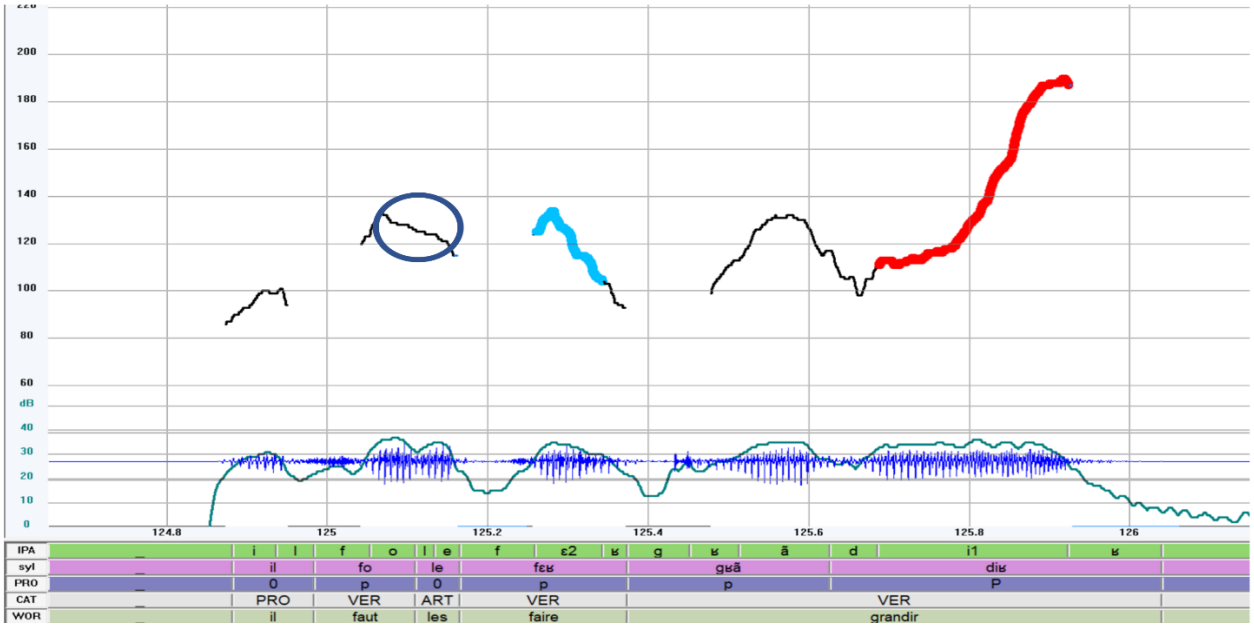

Figure 4 : La syllabe accentuable sur le mot *faut* (cerclée) n'est pas marquée comme accentuée car bien qu'appartenant à la catégorie des verbes, elle est suivie immédiatement à moins de 250 ms par une autre syllabe accentuée sur le verbe *faire* dans l'extrait *il faut les fa**i**re grand**i**r*.

6.3 Règle des 1250 ms

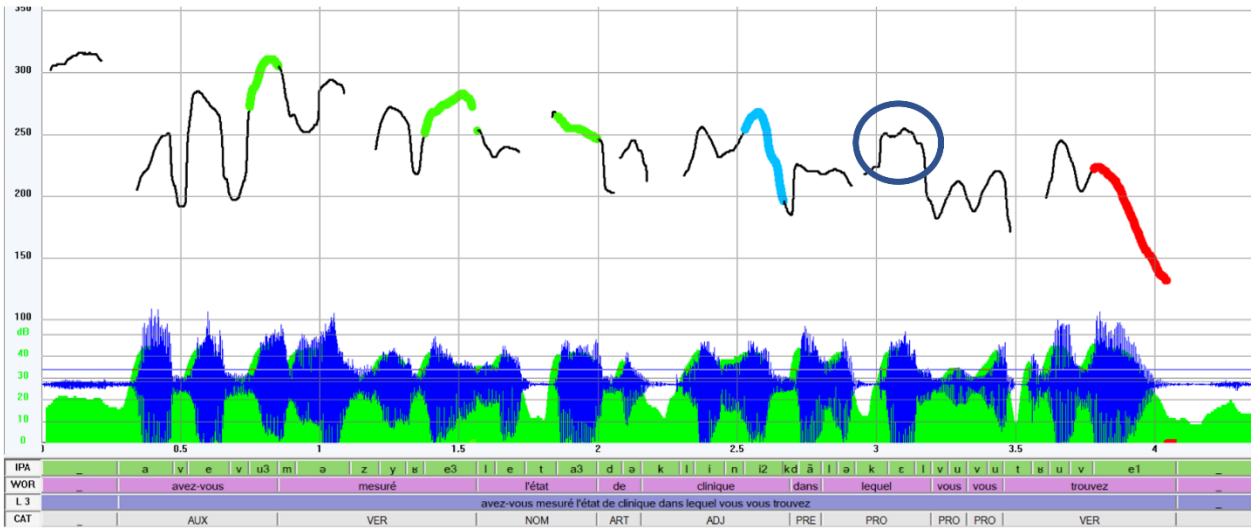

Figure 5 : La dernière syllabe du pronom *lequel*, de catégorie non-accentuable, est clairement perçue comme accentuée mais n'est pas annotée accentuée de par sa catégorie syntaxique dans *avez-v**ou**s mesur**é** l'ét**a**t de clin**i**que dans lequel vous vous trouv**ez***.

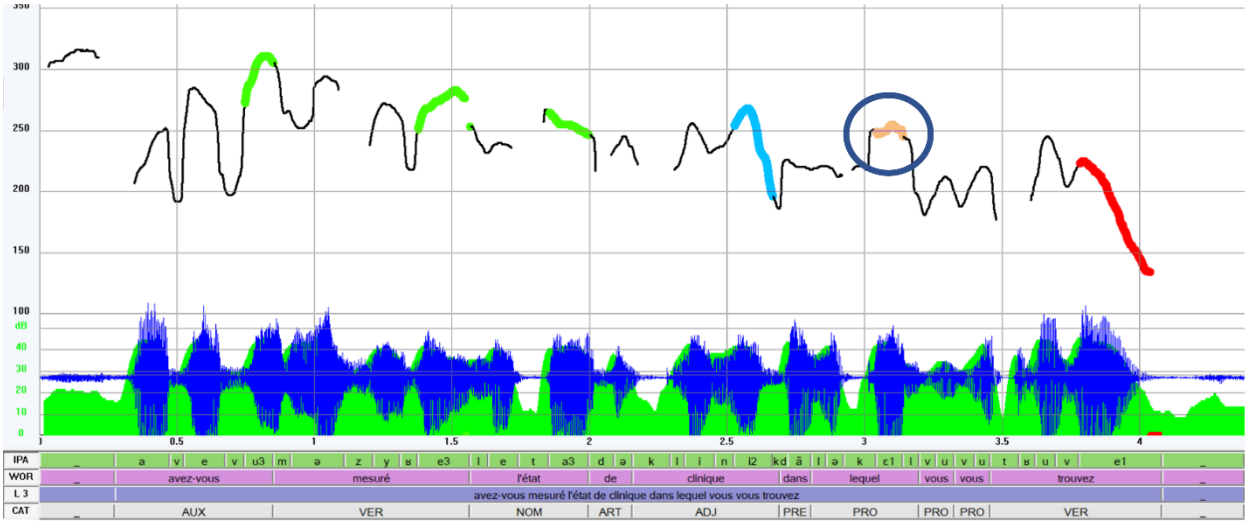

Figure 6 : Application des règles eurythmique et de durée maximale des groupes accentuels. La syllabe finale du pronom *lequel*, annotée comme non-accentuable (Fig. 5) puisque n'ayant ni une valeur de glissando suffisante ni étant suivie d'une pause de plus de 250 ms, est finalement notée comme accentuée en vertu de la règle d'eurythmicité d'une part (la durée moyenne des groupes accentuels précédents est de 599 ms) et l'intervalle entre les syllabes finales de *lequel* et de *trouvez*, égal à 1323 ms, excède la durée maximale d'un groupe accentuel (environ 1250 ms).

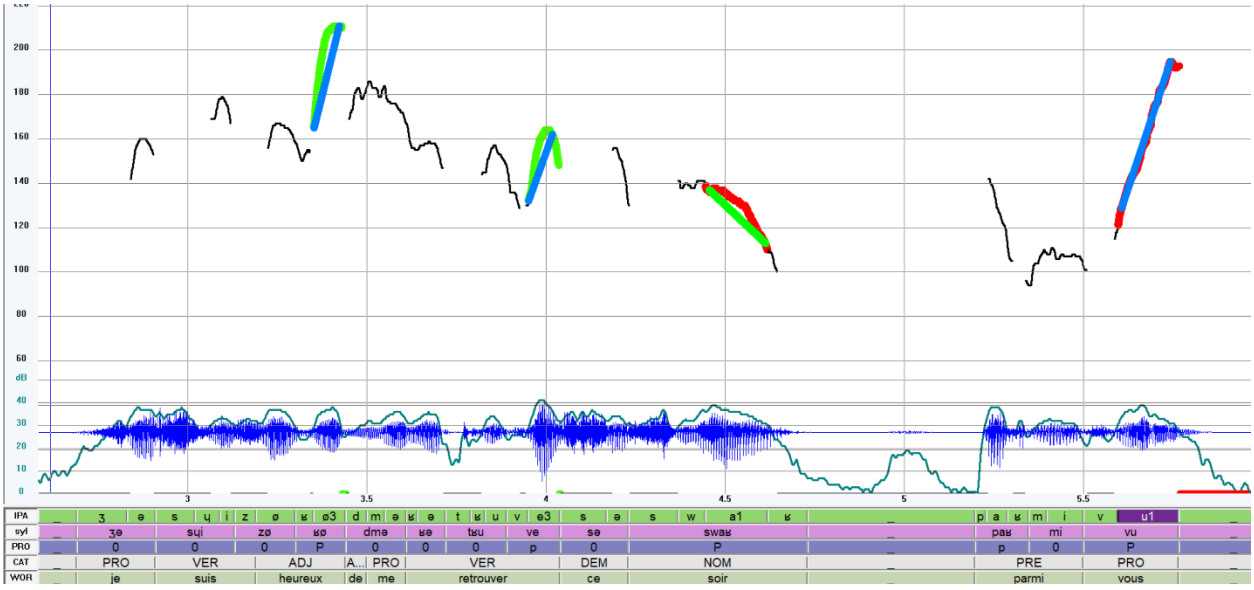

Figure 7 : Positionnement automatique des contours mélodiques à partir du positionnement des syllabes accentuées de l'extrait *je suis heureux de me retrouver ce soir parmi vous*. La structure prosodique résultante est indiquée par les contours

7 Évaluation

On a vu que l'évaluation des performances du programme pose a priori un problème, puisque les résultats sont censés être comparés à des annotations manuelles. La perception de l'accent « tonique » est influencée par le processus de prédiction propre à l'annotateur, tendant à détecter les syllabes accentuées qu'il aurait placées en lisant ou en parlant avec son propre débit. Néanmoins, certaines données de parole lue et spontanée annotées dans le cadre de référence Rhapsodie [22] ont été utilisées comme référence acceptable pour comparer les annotations manuelles avec les annotations obtenues par l'algorithme. Les résultats de cette comparaison sont donnés dans les Table 1 et 2, dont les colonnes correspondent aux classes de syllabes accentuées, et les symbole p et P aux proéminences faibles et fortes perçues par les annotateurs.

Table 1 : Table de correspondance entre les syllabes accentuées détectées par WinPitch (de classe 1, 2 et 3) et des annotation manuelle P (forte proéminence) et p (faible proéminence) (parole lue M201, Corpus Rhapsodie, [21])

	Classe 1	Classe 2	Classe 3
Détectés par l'algorithme	15	42	25
Notés P/p	14	24	22
Non notés P/p	1	18	3

On remarque que ce sont les proéminences de classe 3, c'est-à-dire celles déterminées par la catégorie syntaxique lexicale (nom, adjectif, adverbe, verbe) qui sont sur-détectées par WinPitch et retenues en vertu du critère d'eurythmie, le débit étant de l'ordre de 4 syllabes / seconde. Parmi les exemples de ce type, on a essentiellement des verbes auxiliaires, comme *j'aur**ai** passé, qui **ont** la difficile tâche, qui **sont** parmi nous, elles **sont** souvent confrontées.* On trouve également deux cas de groupes comme ***cher** Hervé* ou ***a**lors*. le nombre de syllabes notées P ou p mais non retenues par l'algorithme est de 2.

Table 2 : Table de correspondance entre les syllabes accentuées détectées par WinPitch (de classe 1, 2 et 3) et des annotation manuelle P (forte proéminence) et p (faible proéminence) (parole spontanée D103, Corpus Rhapsodie, [21])

	Classe 1	Classe 2	Classe 3
Détectés par l'algorithme	15	42	24
Notés P/p	14	25	22
Non notés P/p	1	17	2

Une grande différence apparait pour la classe 2 (critère de glissando), due d'une part aux hésitations (*euh* d'hésitation et allongements) rarement notées comme proéminentes par les annotateurs, mais supérieures au seuil de glissando, et d'autre part à de nombreux exemples de syllabes accentuables séparées de moins de 250 ms (*en **fin** de **compt**e, j'ai chois**i** **L**, quand j'é**tais** pe**tit**e*, etc.), le débit étant dans cet exemple de plus de 7 syllabes / seconde. Ainsi deux syllabes accentuées successives sont perçues comme telles par les annotateurs, alors qu'en réalité ce n'est pas le cas dans l'enregistrement, ce qui est correctement noté par

l'algorithme. D'autre part, le nombre de syllabes notées P ou p mais non retenues par l'algorithme est de 22.

8 Conclusion

Un programme d'annotation des syllabes accentuées en français, suivi d'un générateur de structure prosodique intègre non seulement des critères acoustiques (ici les valeurs de glissando dérivées des durées et des variations mélodiques à l'endroit des voyelles), mais aussi un processus basé sur les catégories syntaxiques et la présence de silence suivant les syllabes accentuées, simulant ainsi les processus de perception de l'auditeur, à la fois bottom-up par les caractères acoustiques de syllabes et top-down par les attentes liées aux catégories syntaxiques ainsi qu'aux durées minimales et maximales des groupes accentuels définis par les syllabes accentuées.

La comparaison avec des données obtenues manuellement par des experts est très satisfaisante, mais fait toutefois ressortir la dépendance des annotations manuelles au débit de parole virtuel implicite nécessairement adopté par les opérateurs.

Références

1. Martin Ph. *Intonation du français*, Paris : Armand Colin, 2009, 256 p.

2. Martin Ph. *The Structure of Spoken Language. Intonation in Romance*, Cambridge: Cambridge University Press, 2015, 206 p.

3. Martin Ph. *Intonation, structure prosodique et ondes cérébrales*, London: ISTE, 2018, 322 p.

4. Garde, P. *L'accent*. Paris : Presses universitaires de France, collection SUP « Le linguiste », n° 5. 1968, 172 p.

5. Martin Ph. Spontaneous speech corpus data validates prosodic constraints, *Proceedings of the 6th conference on speech prosody*, Campbell, Gibbon, and Hirst (eds.), 2014, 525-529.

6. Ghitza, Oded (2011) Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm, *Frontiers in Psychol*. 2, 130.

7. Avanzi, M. Note de recherche sur l'accentuation et le phrasé à la lumière des corpus du français. *Tranel*, 2013, vol. 58, 5-24.

8. Avanzi M., Post B. et Simon A-C. La prosodie du français, accentuation et phrasé, Langue française, 2016/3, No 191,

9. Avanzi M., Lacheret A. et Victorri B. Analor, a Tool for Semi-automatic. Annotation of French Prosodic Structure, *Proc. Speech Prosody 2008*, Campinas, 119-122.

10. Christodoulides, G. & Avanzi, M. An Evaluation of Machine Learning Methods for Prominence Detection in French. *Proc. Interspeech 2014*, 116-119.

11. Blanche-Benveniste C. La naissance des syntagmes dans les hésitations et répétitions du parler, in Araoui J.L. ed. *Le sens et la mesure. Hommages à Benoît de Cornulier*, Honoré Champion, Paris, 2003, 40-55.

12. Deulofeu H-J. Pour une linguistique du rattachement, *Colloque Les linguistiques du détachement*, Nancy, 2006, 7-9 juin 2006.

13. Arnal L. et Giraud A-L. Neurophysiologie de la perception de la parole et multisensorialité, in *Traité de neurolinguistique*, Serge Pinto et Marc Sato éd., Louvain-la-Neuve : De Boeck, 2017, 97-108.

14. Morin Y-C. Histoire des systèmes phoniques et graphique du français, http://ycmorin.net/wp-content/uploads/2012/11/2006-Histoire_phonologie_graphie_du_français.pdf

15. Martin Ph. Quelques changements prosodiques du français parlé de 1900 à 2000 *Verbum*, 2006, Univ. de Nancy II - ATILF.

16. Wioland F. *Les structures rythmiques du français*, Slatkine-Champion, Paris, 1985.

17. Jun, S-A. The Accentual Phrase in the Korean prosodic hierarchy, *Phonology*, (15) 2, 1998, 189-226.

18. Blanche-Benveniste C. *Approches de la langue parlée en français*, Paris : Ophrys, 2000, 164 p.

19. Delattre P., Les dix intonations de base du français, *French Review* 40, 1966, 1-14.

20. Rossi, M. Le seuil de glissando ou seuil de perception des variations tonales pour la parole. *Phonetica*. n° 23, 1971, 1-33.

21. Lexique 3 : http://www.lexique.org/

22. Rhapsodie : corpus prosodique de référence en français parlé, https://www.projet-rhapsodie.fr/

23. WinPitch : Logiciel d'analyse de la parole. www.winpitch.com