Avanzi Mathieu, Simon Anne-Catherine, Goldman Jean-Philippe, Auchlin Antoine, C-PROM. Un corpus de français parlé annoté pour l'étude des proéminences, In: XXVIIIèmes journées d'étude sur la parole (JEP 2010), Mons, 25-28 mai 2010, 2010, pp. 73-76.

C-PROM :
Un corpus de français parlé annoté pour l'étude des proéminences

M. Avanzi[1, 2], *A.C. Simon*[3], *J.-P. Goldman*[3, 4] & *A. Auchlin*[4]

[1]Chaire de linguistique française, Université de Neuchâtel; [2]MoDyCo, Université Paris Ouest Nanterre ;
[3]Institut Langage & Communication / Valibel Discours & Variation, Université catholique de Louvain;
[4]Département de linguistique, Université de Genève
mathieu.avanzi@unine.ch, anne-catherine.simon@uclouvain.be,
jeanphilippegoldman@gmail.com, antoine.auchlin@unige.ch

ABSTRACT

This paper presents C-PROM, an annotated corpus for studying prominence in French. The corpus includes 3 regional varieties of French (Belgian, Swiss and metropolitan French) and 7 speaking styles (from oral reading to spontaneous conversations); it has a total duration of 70 minutes and was annotated by two phoneticians. The two experts followed a strict annotation protocol, taking into account the previous mistakes encountered by prior research on prominence detection in French and elements of the methodology followed by scholars working on other languages. We conclude by discussing the average consistency between both annotators. The results obtained are quite encouraging, since the F-measure between the two annotators reaches 82.8%, and the kappa-score 0.77.
Keywords: corpus, spontaneous French, prominence, discourse-genre.

1. TRAVAUX ANTÉRIEURS

Les premiers travaux sur la détection de proéminences en français ont vu le jour dans le cadre du projet Phonologie du Français Contemporain (PFC) [1]. Dans le cadre de ce projet, les principes posés pour le protocole d'annotation de la prosodie étaient à l'origine les suivants [2]. La procédure de codage devait (i) être réalisée en dehors de tout cadre théorique spécifique, (ii) reposer sur des bases impressionnistes uniquement, (iii) être reproductible par des annotateurs non experts, (iv) permettre des études transversales à tous les domaines de la prosodie (accentuation, intonation, rythme, etc.).

Dans l'optique de mettre au point une première esquisse de protocole, une expérience pilote a été conduite par [3]. L'auteur a demandé à sept experts phonologues d'annoter les proéminences syllabiques dans un extrait de parole spontanée d'une durée de 3 minutes, sans aucune autre instruction. L'auteur s'attendait à un consensus assez élevé, partant de l'idée que les proéminences correspondaient à des syllabes accentuées, et que les règles de l'accentuation du français étaient bien connues des participants. De façon surprenante, il est apparu que le taux d'accord inter-juges n'était pas aussi bon que ce qui était attendu, puisque sur les 165 syllabes susceptibles de porter un accent primaire dans le corpus, la proportion de syllabes annotées comme proéminentes variait de 19% à 49%. Ces résultats, guère encourageants, ont confirmé les points de vue les plus pessimistes, à savoir que l'annotation des proéminences « s'apparentait plus à un art qu'à une pratique scientifique » [4].

Reprenant ces données, [5] ont cherché à évaluer le caractère robuste de deux paramètres acoustiques (la F0 et la durée) pour une détection automatique des proéminences, dans l'optique de suggérer des mesures objectives pour évaluer les performances des annotateurs. De cette seconde expérience, les auteurs ont conclu que les variations de F0 étaient corrélées avec le taux d'accord inter-juge (plus les variations relatives de F0 sont importantes, plus l'accord entre les codeurs est grand). Par contre ce n'est pas le cas avec les mesures relatives de durée : en effet, ils ont observé qu'au-delà un certain seuil de durée (approximativement entre 175 et 200 ms), la proportion d'accord inter-juge s'inverse et le taux décroît. Cela est dû au fait qu'au-delà d'une certaine durée syllabique, les allongements ne sont plus perçus comme des marqueurs de proéminence mais comme des signaux d'hésitation. Cela signifie en outre que des humains, même experts, ne partagent pas la même définition de la notion de proéminence.

Plusieurs enseignements ont été tirés de ces études pilotes par les initiateurs du codage de la prosodie dans les corpus PFC. Premièrement, il est apparu que le faible taux d'accord provenait vraisemblablement du manque de clarté et de précision dans la consigne et que, par conséquent, si l'on désirait avoir un meilleur taux de consensus entre les annotateurs, il fallait que la notion de proéminence soit correctement définie et qu'elle ne soit pas confondue avec celle d'accent (qui est une notion phonologique impliquant un savoir linguistique spécifique). Ensuite, il est ressorti que le codage perceptif des proéminences devait être borné à un intervalle temporel défini, pour éviter que différentes sous-parties du corpus subissent un codage différent (plus long est l'extrait écouté, moins élevé est la proportion de syllabes perçues comme proéminentes, et inversement). Finalement, l'étude des corrélats acoustiques des proéminences a montré que si la F0 était un indice pertinent pour la détection automatique,

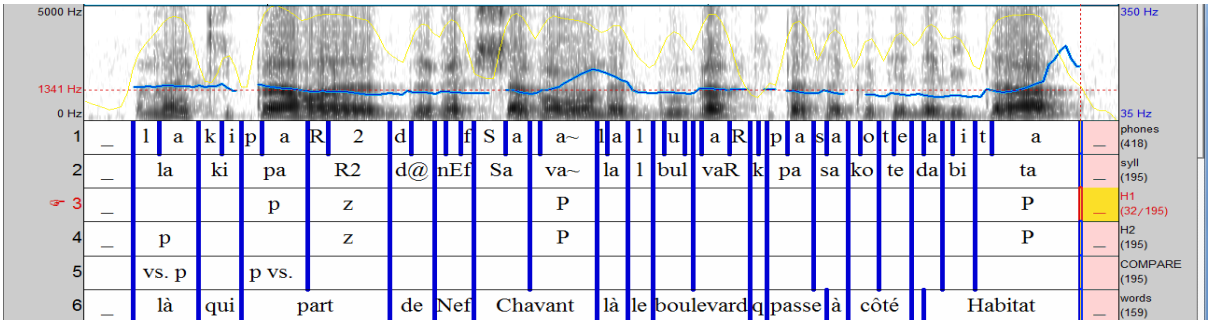

Figure 1 : Copie d'écran Praat, transcription de l'énoncé : « là qui part de Nef Chavant là le boulevard qui passé à côté d'Habitat » [iti-1]. Les couches d'annotations sont, du haut vers le bas, celle des phonèmes, celle des syllabes (toutes deux en SAMPA), celles des annotations manuelles H1 et H2 (tires « delivery »), la tire COMPARE et celle des mots graphiques.

les allongements liés à une hésitation devaient être balisés, d'une part pour ne pas être interprétés différemment selon les codeurs, et d'autre part pour ne pas interférer dans les mesures de durée relative dans une perspective d'automatisation de la détection.

2. DESCRIPTION DU CORPUS

Ces premiers travaux conduits par les participants du groupe Prosodie-PFC ont connu des prolongements dans diverses études menées par un consortium informel de linguistes français, suisses et belges et ont *in fine* abouti à l'élaboration d'un corpus multi-styles et multi-locuteurs annoté pour les proéminences. Ce corpus est présenté dans les sections suivantes. L'annotation perceptive a également permis de développer des algorithmes de détection automatique [6] qui ne sont pas décrits ici. Le corpus est composé d'enregistrements échantillonnés en sept genres avec, du plus formel au moins formel : Journaux radiophoniques (JPA) ; Lectures d'un texte (LEC), Discours politique (POL), Conférences (CNF), Interviews radiophoniques (INT), Descriptions d'itinéraires (ITI) et Récits de vie (NAR). Mis à part les genres ITI et INT, tous les genres de discours sont représentés par 3 enregistrements de 3 minutes environ, de locuteurs francophones natifs de Belgique (BE), de Suisse (CH) et de France métropolitaine (FR) (voir tableau 3 *infra*). Cet ensemble d'enregistrements, d'une durée totale de 70 minutes, a été automatiquement segmenté en phonèmes, syllabes et mots graphiques avec le script EasyAlign [7], implémenté sous Praat [8]. Les transcriptions ont été vérifiées manuellement, et corrigées le cas échéant.

3. ANNOTATION PERCEPTIVE DES PROÉMINENCES

3.1. Protocole

Deux annotateurs (H1 et H2, deux des auteurs de cet article) ont codé la totalité du corpus en suivant un protocole strict. Ce protocole d'annotation du corpus C-PROM tire parti des difficultés rencontrées lors des premières études menées dans le cadre de PFC, et tient compte de certaines des recommandations faites par les superviseurs du corpus de néerlandais parlé [9]. En pratique, chaque annotateur part d'une couche (« tire ») d'annotation vide, dupliquée à partir de la tire syllabe (tires H1 et H2, *cf.* figure 1), et annote les intervalles avec trois types de symboles. La première classe de symboles concerne la perception des proéminences à proprement parler. Deux symboles (p et P) peuvent être utilisés pour annoter les syllabes perçues comme faiblement ou fortement proéminentes. La distinction entre « p » et « P » a une valeur heuristique : elle force le codeur à développer une écoute plus fine et permet d'éviter un marqueur d'indécision du type « ? ». Si aucune proéminence n'est perçue, l'intervalle est laissé vide. Au cours de la phase de comparaison entre annotateurs, les deux symboles « p » et « P » sont confondus en une seule classe. La deuxième catégorie de symboles, caractérisant les marques de travail de formulation (tire « delivery »), est utilisée pour isoler des syllabes qui ont des propriétés spécifiques, susceptibles d'interférer avec la perception et la détection automatique des proéminences. On introduit ainsi dans l'annotation une distinction entre les proéminences supposément liées au système accentuel du français et des phénomènes liés au travail de formulation. Le symbole « z » est utilisé pour noter les syllabes allongées associées à une hésitation (leur longueur gênant les mesures de durée relative, comme l'ont montré [5]). Quant aux schwas post-toniques (@) et aux appendices ($), ils ont été identifiés spécifiquement pour se donner la possibilité d'étudier leurs statuts phonologiques spécifiques : les segments post-focaux sont décrits comme ayant un profil mélodique plat et peu modulé, en théorie; d'autre part, la capacité des schwas en position finale de mot à jouer le rôle de noyau accentuel est controversée. La troisième catégorie de symboles regroupe les « silences » et certains phénomènes para- ou non verbaux. Elle contient les pauses silencieuses repérées automatiquement, les prises de souffle audibles ainsi que les intervalles inexploitables, *i.e.* les segments qui n'ont pu être transcrits (bruits, rire, toux,

chevauchements), susceptibles d'interférer avec le traitement automatique.

3.2. Tâche d'annotation

Dans un premier temps, les deux annotateurs se sont entraînés sur un enregistrement d'indication d'itinéraire d'une minute environ (iti-5), puis ont annoté le reste du corpus chacun de son côté, genre de discours par genre de discours (l'annotation s'est déroulée sur plusieurs mois, de l'automne 2007 à l'automne 2008). En pratique, chaque annotateur écoute des segments de trois à cinq secondes, pas plus de trois fois (l'hypothèse étant qu'une sur-écoute entraînerait un surcodage). Comme l'annotation repose sur la perception de saillances syllabiques et non sur l'analyse visuelle des variations acoustiques (montées de F0 p. ex.), l'affichage du signal de parole est réservé aux cas retors (lors de la phase ultérieure de comparaison). Au fur et à mesure qu'un genre est codé, une tire de comparaison est automatiquement générée, afin de mettre en lumière les cas de désaccord entre les deux codeurs (voir tire COMPARE de la figure 1). Ces désaccords ont ensuite été discutés au cours de sessions communes ; on se reportera à [9] pour une analyse systématique des cas de désaccords, et les règles qui ont émergé pour les arbitrer.

3.3. Evaluation du taux d'accord inter-annotateurs

La tire COMPARE a été utilisée pour estimer le taux d'accord inter-annotateurs. Le Tableau 1 donne le détail de ce taux pour chaque enregistrement du corpus, sur la base du décompte des intervalles impliquant un conflit entre p, P, et un autre symbole.

Tableau 1: Taux d'accord inter-annotateurs : nom du fichier (voir Table 2) ; n00/n11 : syllabes annotées 0 (non proéminent) et P (proéminent) les deux annotateurs ; n10/n01 : syllabes annotées P par un transcripteur, 0 par l'autre; A : accord brut ; R : rappel ; P : précision ; F : F-mesure ; K : Kappa-score.

Fichiers	n00	n11	n10	n01	A	R	P	F	K
jpa-be	924	259	58	29	93	89,9	81,7	85,6	0,81
jpa-ch	587	172	29	50	91	77,5	85,6	81,3	0,75
jpa-fr	668	197	25	46	92	81,1	88,7	84,7	0,80
lec-be	495	87	22	5	96	94,6	79,8	86,6	0,84
lec-ch	390	156	26	33	90	82,5	85,7	84,1	0,77
lec-fr	469	120	32	2	95	98,4	78,9	87,6	0,84
pol-be	225	139	21	26	89	84,2	86,9	85,5	0,76
pol-ch	735	162	20	97	88	62,5	89	73,5	0,66
pol-fr	519	143	13	71	89	66,8	91,7	77,3	0,70
cnf-be	727	182	26	84	89	68,4	87,5	76,8	0,70
cnf-ch	586	186	40	59	89	75,9	82,3	79	0,71
cnf-fr	776	239	39	37	93	86,6	86	86,3	0,82
int-be	722	224	48	44	91	83,6	82,4	83	0,77
int-fr	899	262	56	72	90	78,4	82,4	80,4	0,74
iti-01	124	26	9	2	93	92,9	74,3	82,5	0,78
iti-02	84	41	1	1	98	97,6	97,6	97,6	0,96
iti-03	261	83	22	15	90	84,7	79	81,8	0,75
iti-04	547	154	15	21	95	88	91,1	89,5	0,86
iti-06	281	87	8	7	96	92,6	91,6	92,1	0,89
iti-07	94	24	7	1	94	96	77,4	85,7	0,82
nar-be	582	190	31	36	92	84,1	86	85	0,80
nar-ch	573	172	16	65	90	72,6	91,5	80,9	0,74
nar-fr	468	149	25	40	90	78,8	85,6	82,1	0,76
Total	**11736**	**3454**	**589**	**843**	**91**	**80,4**	**85,4**	**82,8**	**0,77**

Il ressort du Tableau 1 que les taux de syllabes proéminentes de H1 et H2 s'élèvent à 24,3% et 29,3%, respectivement (à comparer à la fourchette 19%-49% de [3]). Le taux d'accord inter-juges a été quantifié à l'aide du coefficient Kappa, et évalué à 0,77 pour l'ensemble du corpus. Quant à la F-mesure, elle indique un taux d'accord de 82,8% (rappel : 80,4 et précision 85,4). A la suite de [9], nous pensons que si le taux d'accord est si bon, c'est parce que les codeurs ont suivi un protocole strict, sur lequel ils se sont entraînés auparavant.

3.4. Annotation de référence

Une annotation consensuelle a émergé de la discussion des désaccords relevés dans la tire COMPARE, et elle a été considérée comme annotation de référence pour l'analyse des syllabes proéminentes. Le Tableau 3 donne le détail de la composition du corpus C-PROM, qui comprend 28 locuteurs (12 femmes, 16 hommes) et 17.778 syllabes. Parmi celles-ci, 805 (4.5%) sont notées spécifiquement dans la tire delivery, 4570 sont annotées proéminentes (25.7%) et 12.403 (69.7%) ne sont ni proéminentes ni notées « delivery ».

4. CONCLUSION

Le but de cet article était de présenter C-PROM, un corpus annoté pour l'étude des proéminences en français. Nous avons essayé de synthétiser en quelques lignes les principaux travaux à l'origine de la mise au point de ce corpus de référence et de présenter la méthodologie suivie pour sa constitution. La méthodologie proposée doit être testée sur de plus larges enregistrements avec davantage d'annotateurs. Cependant cette première expérience s'est révélée encourageante, le consensus entre les deux codeurs atteignant un taux satisfaisant. Des parties du corpus ont déjà été utilisées pour entraîner des systèmes de détection automatique [6] [9] [11], et pour des études de discrimination automatique des genres de discours sur des critères prosodiques [12] **Erreur ! Source du renvoi introuvable.**. Aussi espérons-nous que ce corpus permettra de rendre comparables les prochaines études relatives à la problématique de la proéminence, chose impossible auparavant en raison du manque de données communes.

Le corpus est téléchargeable à l'adresse suivante : http://sites.google.com/site/corpusprom

5. REMERCIEMENTS

Cette recherché a été soutenue par le Fonds National de la Recherche scientifique Suisse, (subsides n° PBNEP1-127788, Université de Neuchâtel) et par le Programme Wist2 Convention n°616422, financé par la Région wallonne (Belgique), Projet *EXPRESSIVE*.

BIBLIOGRAPHIE

[1] Durand, J., Laks, B. & C. Lyche. "La phonologie du français contemporain: usages, variétés et structure", *in* Pusch, C. & W. Raible (eds.), Romance Corpus Linguistics, Tübigen, Gunter Narr Verlag, 93-106, 2002.

[2] Lacheret-Dujour, A., Lyche, Ch. & M. Morel, "Pour une transcription prosodique normalisée au sein du projet PFC (phonologie du français contemporain): champ d'action et limites", Actes des 25^{s} JEP, Fès, Maroc, 2004.

[3] Poiré, P. "La perception des proéminences et le codage prosodique", Bulletin PFC, 6, 69-79, 2006.

[4] Martin, Ph. 2006. « La transcription des proéminences accentuelles : mission impossible ? », Bulletin PFC, 6, 81-87.

[5] Morel M., Lacheret-Dujour, A. Lyche, Ch., Morel, M. & F. Poiré. "Vous avez dit proéminence?", Actes des 26èmes journées d'étude sur la parole, Dinar, Maroc, 2006.

[6] Obin, N., Goldman, J.-P., Avanzi, M. & Lacheret-Dujour, A. "Comparaison de trois outils de détection semi-automatique des proéminences dans les corpus de français parlé", Actes des 22èmes JEP, Avignon, 2008.

[7] Goldman, J.-P. "EasyAlign: a semi-automatic phonetic alignment tool under Praat", http://latlcui.unige.ch/phonetique, 2008.

[8] Boersma, P. & Weenink, D. Praat: doing phonetics by computer (Version 5.2). www.praat.org, 2010

[9] Buhmann, J., Caspers, J., van Heuven, V, Hoekstra, H., Martens, J.-P. & M. Swerts, "Annotation of Prominent Words, Prosodic Boundaries and Segmental Lengthening by Non Expert Transcribers in the Spoken Dutch Corpus", LREC Processing, 779-785, 2002.

[10] Avanzi, M., Goldman, J.-P. Lacheret-Dujour, A. Simon, A.-C. & A. Auchlin, "Méthodologie et algorithmes pour la détection automatique des syllabes proéminentes dans les corpus de français parlé", Cahiers of French Language Studies, vol. 13, no. 2, pp. 2–30, 2007.

[11] Goldman, J.-P.; Avanzi, M.; Lacheret-Dujour, A.; Simon, A.-C.; Auchlin, A. 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, 2007.

[12] Obin, N., Lacheret-Dujour, A., Veaux, C., Rodet, X & A.C. Simon, "A Method for Automatic and Dynamic Estimation of Discourse Genre Typology with Prosodic Features", Interspeech, Brisbane, 2008.

[13] Simon A.C., Auchlin A., Avanzi M., Goldman J.-Ph., "Les phonostyles: une description prosodique des styles de parole en français", in Les voix des Français. En parlant, en écrivant, Abecassi, M. & G. Ledegen (eds), Berne, Peter Lang, 2010, sous presse.

Table 2. Contenu détaillé du corpus C-PROM. Avec, de gauche à droite : genre de discours, cote du sous-corpus, sexe du locuteur, durée, nb. de syllabes, nb. de syllabes non proéminentes, nb. de syllabes non proéminentes, nb. de syllabes associées à un marqueur delivery, nb. de syllabes associées à une hésitation, associées à une séquence postfocale et un schwa post-tonique. Les sous-totaux sont en gras. Les totaux pour l'ensemble du corpus figurent dans la ligne grisée tout en bas du tableau.

	Genres	Fichiers	Loc.	Durée (sec.)	Total syllables	Non-prom syllables	Prom syllables	Delivery syllables			
								total	Z	$	@
+formel	Journaux radiophoniques	JPA-BE	M	253	1315	963	312	40	24	0	16
		JPA-CH	F	180	879	610	242	27	12	0	15
		JPA-FR	M	188	971	683	256	32	19	0	13
		total	**2M/1F**	**621**	**3165**	**2256**	**810**	**99**	**55**	**0**	**44**
	Lectures	LEC-BE	M	114	606	492	111	3	0	0	3
		LEC-CH	M	137	606	403	196	7	0	0	7
		LEC-FR	M	150	618	462	153	3	0	0	3
		total	**3M**	**401**	**1830**	**1357**	**460**	**13**	**0**	**0**	**13**
	Discours politiques	POL-BE	M	188	420	246	160	14	0	0	14
		POL-CH	F	230	1011	753	257	1	0	0	1
		POL-FR	M	217	743	533	209	1	1	0	0
		total	**2M/1F**	**635**	**2174**	**1532**	**626**	**16**	**1**	**0**	**15**
	Conférences	CNF-BE	F	244	1066	776	250	40	35	0	5
		CNF-CH	M	219	950	627	260	63	55	7	1
		CNF-FR	F	224	1117	798	301	18	12	1	5
		total	**1M/2F**	**687**	**3133**	**2201**	**811**	**121**	**102**	**8**	**11**
	Interviews radiophoniques	INT-BE	2F	296	1189	769	317	103	56	32	15
		INT-FR	2M	331	1402	996	346	60	21	30	9
		total	**2M/2F**	**627**	**2591**	**1765**	**663**	**163**	**77**	**62**	**24**
	Itinéraires	ITI-01	M	50	172	117	36	19	15	2	2
		ITI-02	2M	47	142	82	42	18	17	1	0
		ITI-03	F	100	419	270	104	45	30	11	4
		ITI-04	2F	204	790	538	197	55	50	1	4
		ITI-05	M	28	128	92	27	9	8	0	1
		ITI-06	1M/1F	128	431	298	106	27	17	8	2
		ITI-07	M	33	140	98	30	12	10	2	0
		total	**6M/3F**	**590**	**2222**	**1495**	**542**	**185**	**147**	**25**	**13**
- formel	Récits de vie	NAR-BE	F	206	939	634	238	67	55	12	0
		NAR-CH	F	218	949	632	228	89	75	8	6
		NAR-FR	F	198	775	531	192	52	46	2	4
		total	**3F**	**622**	**2663**	**1797**	**658**	**208**	**176**	**22**	**10**
	7 genres	**24**	**16M/12F**	**4183**	**17778**	**12403**	**4570**	**805**	**558**	**117**	**130**