Chapitre 8

Mesure de l'impact des pauses et de l'accent lexical sur la compréhensibilité du locuteur

Maintenant que nous disposons d'un prototype permettant d'analyser automatiquement la distribution syntaxique des pauses et l'accentuation lexicale, il est possible d'évaluer quantitativement leur impact sur la compréhensibilité du locuteur. Or, comme nous l'avons vu chapitre 2, cette compréhensibilité est un phénomène perceptif, et ne peut donc être évaluée qu'à travers le jugement d'auditeurs.

Nous proposons un protocole expérimental pour tenter de mesurer l'impact des pauses et de l'accentuation lexicale en temps-réel sur la compréhensibilité. Nous nous inspirons du protocole employé par Nagle et al. (2019) pour évaluer dynamiquement la compréhensibilité du locuteur, qui eux-mêmes adaptent une méthode d'évaluation continue utilisée en psychologie cognitive (MacIntyre, 2012). Nous avons simplifié et adapté leur protocole pour permettre une évaluation de plus grande échelle, en *crowdsourcing*, et en privilégiant l'aspect quantitatif de l'approche.

Nous tenterons de répondre aux deux questions suivantes : Q1) Les auditeurs montrent-ils un comportement cohérent dans l'évaluation dynamique de la compréhensibilité malgré les variations inter- et intra-individuelles ? Q2) Une diminution de la compréhensibilité est-elle observable à la suite d'occurrences de pauses intra-syntagme ou de patterns accentuels inappropriés ?

Après avoir décrit notre protocole expérimental, les stimuli audio sélectionnés, et les participants recrutés pour l'expérience, nous détaillerons les différents traitements effectués sur les données collectées, et les analyses réalisées sur celles-ci. La plateforme expérimentale est présentée avec les résultats, chapitre 13.

8.1 Adaptation du protocole

Nous avons donc choisi de partir du protocole expérimental mis au point par Nagle et al. (2019). Celui-ci tente d'évaluer de manière dynamique le jugement de compréhensibilité, afin de pouvoir observer les fluctuations de ce jugement au fur et à mesure de l'écoute. Si Nagle et al. analysent ces fluctuations de manière globale dans une approche exploratoire, sans cibler de phénomène linguistique précis, nous proposons quant à nous d'exploiter cette méthode pour observer comment varie le jugement des participants à la suite de certaines pauses ou patterns accentuels. Plus concrètement, nous souhaitons observer si le jugement de compréhensibilité a tendance à diminuer à la suite de pauses de bas niveau (intra-syntagme, a priori disfluentes) et de patterns accentuels inappropriés, comparé au jugement mesuré à la suite de pauses de haut niveau (inter-proposition, a priori structurantes) et de patterns accentuels corrects.

Trois modifications majeures du protocole de Nagle et al. (2019) ont été réalisées. Pour permettre à un plus grand nombre d'évaluateurs de participer, nous avons opté pour une passation en ligne, sur une plateforme d'évaluation dédiée. Nous n'avons pas effectué de captation vidéo suivie d'entretiens individuels comme c'est le cas dans le protocole original. L'expérimentation a ainsi été repensée pour permettre une passation en complète autonomie : elle a été simplifiée et raccourcie pour ne pas dépasser un temps théorique de 35 minutes. Une rapide explication de la tâche est donnée à l'écrit en début d'expérience, suivie de trois questions pour vérifier le profil du participant, et d'une phase d'entraînement. La consigne reste écrite jusqu'à la fin de l'expérience. Après chaque stimulus, si l'évaluateur a été jugé trop peu actif, une *pop-up* de rappel s'ouvre avant le stimulus suivant.

La tâche d'évaluation elle-même a été simplifiée de manière à n'avoir plus qu'un seul bouton sur la page au lieu de deux, et donc une seule action possible. Il est simplement demandé à l'auditeur de cliquer sur le bouton dès qu'il sent qu'il fait une effort pour comprendre le locuteur, quelque soit la raison. De plus, il n'y a plus d'incrémentation du jugement comme c'est le cas sur le logiciel utilisé par Nagle et al. Ainsi, au lieu de varier entre 5 et -5, le jugement ne peut plus être que -1. Lorsque l'auditeur clique sur *start* au début de chaque stimulus, celui-ci démarre sans possibilité de mettre pause ni de réécouter. Chaque clic est enregistré sous la forme d'un *timestamp* correspondant à la position du curseur de lecture. À la fin de la lecture, il lui est demandé d'évaluer globalement la performance du locuteur en termes de qualité de prononciation, de fluidité, et de facilité de compréhension à l'aide de curseurs libres. Enfin, une question optionnelle est posée incitant l'évaluateur à expliciter les aspects de la prononciation du locuteur qui l'ont rendu difficile à comprendre, et suggérer des conseils pour s'améliorer. Les stimuli apparaissent dans un ordre aléatoire, et tous les stimuli sont évalués par tous les participants.

8.2 Sélection des stimuli

Afin de mesurer l'impact des différentes catégories de pauses et d'accentuation lexicale, il est nécessaire de présenter des stimuli audio contenant suffisamment d'occurrences de chacune d'elles pour pouvoir observer une tendance significative. Seize segments audio issus des analyses du corpus de locuteurs francophones ont ainsi été sélectionnés pour l'expérimentation. Les critères de sélection sont les suivants : 8 segments de parole présentant une grande proportion de pauses intra-syntagme et de mots au pattern accentuel inapproprié, et 8 segments présentant les conditions inverses. Par ailleurs nous avons veillé à ce que les proportions B1/B2 et homme/femme soient respectées dans les deux groupes.

Pour pouvoir caractériser chaque segment en termes de fluence et de qualité d'accentuation, nous avons calculé deux scores par segment : la proportion de pauses de type intra-syntagme (P_{WP}, présentée dans le chapitre précédent – nombre de pauses intra-syntagme divisé par le nombre de pauses total) et un score accentuel moyen $\overline{S_w}$. Comme le contraste prosodique C_w présenté dans le chapitre précédent, section 7.4, le score accentuel S_w représente le degré de contraste entre la syllabe censée porter l'accent primaire P_s et la moyenne des autres syllabes $\overline{P_u}$, mais normalise la différence des deux valeurs par leur somme pour obtenir une différence relative. La formule est la suivante :

$$S_w = \frac{P_{s,w} - \overline{P_{u,w}}}{P_{s,w} + \overline{P_{u,w}}} \tag{8.1}$$

Ainsi la différence entre 95 et 85 donnera un score S plus petit que la différence entre 15 et 5, tandis que le contraste C correspondant est toujours de 10 points. Il s'est avéré par la suite que ce score est moins approprié pour représenter le degré de contraste entre les syllabes d'un mot, puisqu'il n'y a pas de raison valable de donner moins de poids aux contrastes entre valeurs prosodiques élevées. Par contrainte de temps au moment de la rédaction de ce manuscrit, nous choisissons cependant de présenter les résultats obtenus en l'état, avec ce score accentuel S_w.

En projetant les segments sur un plan défini par ces deux dimensions, on peut alors sélectionner les segments situés aux extrémités : P_{WP} élevé et score accentuel bas, et P_{WP} faible et score élevé (cf. figure 8.1). On appellera le premier groupe « *low* », et le second « *high* ». Le seuil de durée minimale des pauses est fixé à 250 ms et le nombre minimum de tokens à 60 pour éviter les segments trop courts.

Les 16 segments sélectionnés sont produits par 15 locuteurs différents, les segments du groupe *low* sont produits par 4 locuteurs B1 et 4 locuteurs B2, ceux du groupe *high* par 3 locuteurs B1 et 5 locuteurs B2. La répartition homme/femme est res-

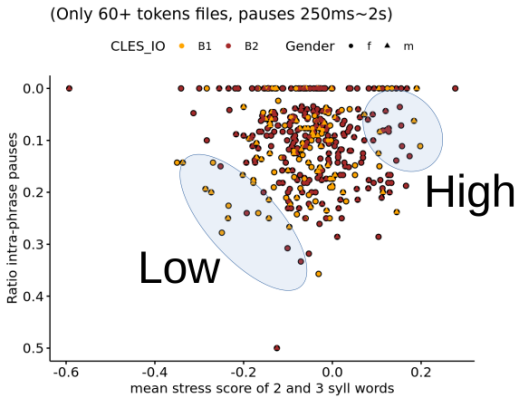

FIG. 8.1 : Choix des stimuli à partir des segments de plus de 60 tokens, projetés en fonction de P_{WP} et du score accentuel moyen

	LOW		HIGH			LOW		HIGH	
	freq	%	freq	%		freq	%	freq	%
BC	59	29,9	73	39,5	**StressO**	1	1,4	22	31,9
BP	99	50,3	98	53	**StressΔ**	35	50	44	63,8
WP	39	19,8	14	7,6	**StressX**	34	48,6	3	4,3
total	**197**		**185**		**total**	**70**		**69**	

TAB. 8.1 : Nombre et proportion de pauses et de mots polysyllabiques de chaque catégorie dans les 16 segments sélectionnés

pectivement de 7 pour 9. La durée des segments s'étend de 26 à 66 secondes (médiane à 38), et le nombre de tokens de 61 à 132 (médiane à 75), sans différence significative entre les groupes *low* et *high*.

Le tableau 8.1 présente le nombre et la proportion de pauses et de mots polysyllabiques de chaque catégorie. Dans le cas de l'accentuation lexicale, les mots sont divisés en trois catégories : *StressO* pour les mots dont le score est élevé ($\overline{S_w} \leq 0{,}2$), *Stress*$\Delta$ pour les mots au contraste peu marqué ($-2 \leq \overline{S_w} < 2$) et *StressX* pour les mots au contraste négatif fort ($\overline{S_w} < -0{,}2$).

Pour s'assurer que l'annotation des pauses et des patterns accentuels est de qualité acceptable, une vérification manuelle a été effectuée sur la moitié des segments, comprenant 193 pauses et 89 mots polysyllabiques. Les pauses dont l'alignement temporel et la catégorie syntaxique sont corrects totalisent 82,4 %, et les mots polysyllabiques correctement reconnus et alignés au niveau du mot et des syllabes totalisent 82,0 %. Au risque de perdre un peu en précision, nous avons décidé de conserver les annotations automatiques en l'état, par soucis de cohérence avec notre démarche tout-automatique (aucune modification manuelle n'a été effectuée depuis le début des traitements).

8.3 Sélection des participants

Les participants ont été recrutés sur la plateforme britannique Prolific[1]. Cette plateforme permet de mettre en relation des chercheurs ou des entreprises avec des personnes de profils variés pour participer à des expérimentations en ligne. Les critères de recrutements que nous avons choisis sont les suivants : être de langue maternelle anglaise, vivre en Angleterre au moment de l'expérience, ne pas avoir déclaré de compétences en langue étrangère (critère "*English speaking monolingual*" sur la plateforme) et respecter une balance de genre. Une rétribution financière a été fixée à hauteur de £10.86 de l'heure, soit £5.25 pour une durée prévue de 35 min (6,14 € (12,7 €/h) au moment de l'expérimentation, en février 2024).

Soixante personnes ont participé à l'expérience, 30 femmes, 30 hommes, de 25 à 72 ans (moyenne à 44, écart type de 12). Seuls les participants qui ont cliqué au moins une fois dans toute l'expérience, et n'ont pas concentré plus de 50 % de leurs clics sur un seul segment ont été retenus pour les analyses.[2]

8.4 Traitement des données

En fin d'expérience, nous avons donc 16 segments audio auxquels sont associés pour chaque évaluateur une liste de *timestamps* correspondant aux clics produits, trois scores globaux numériques et un commentaire textuel optionnel. Nous avons utilisé plusieurs mesures pour évaluer la cohérence et la fiabilité des évaluations globales : le coefficient de corrélation intra-classe (*2-way random model* (ICC2k) du package R psych v2.4.1) pour vérifier la consistance des évaluations entre les évaluateurs, et l'alpha de Cronbach pour évaluer la cohérence interne des évaluations sur l'ensemble des participants. Les deux coefficients ont été calculés à partir des scores bruts de chacune des trois dimensions. Les scores sont ensuite standardisés (z-scores) de manière à les rendre comparables.

Nous proposons d'abord d'analyser les résultats de l'évaluation globale des enregistrements. Le nombre limité d'enregistrements permet difficilement d'envisager une régression linéaire pour calculer un coefficient de corrélation, aussi nous proposons de diviser les segments en deux groupes pour chaque dimension : ceux qui se situent en-dessus et ceux qui se situent en-dessous de la fréquence médiane des pauses intra-syntagme (nombre d'occurrences par token), des pauses inter-proposition et du score accentuel moyen. Les deux distributions sont ensuite comparées à l'aide du test non

[1] https://www.prolific.com/

[2] Trois participants supplémentaires ont été retirés des analyses car leur activité pendant l'expérimentation était trop limitée ou jugée anormale.

paramétrique Wilcoxon-Mann-Whitney et la taille d'effet avec le delta de Cliff. Nous ferons la même chose avec le nombre total de clics normalisés pour voir comment celui-ci évolue globalement en fonction des enregistrements.

Pour l'évaluation dynamique, nous avons calculé la somme des clics par locuteur sur une fenêtre glissante d'une seconde. Afin d'éviter que les « cliqueurs compulsifs », comme les appellent Nagle et al. (2019), ne couvrent les clics des évaluateurs moins actifs, nous proposons d'y soustraire le nombre de clics moyen par minute par locuteur. Nous appellerons ces clics normalisés *m-clics*. Le calcul est effectué de la manière suivante :

$$M_w = \sum_{r=1}^{R} \left(C_{r,w} - \overline{C_r} \right) \tag{8.2}$$

avec M_w le nombre de m-clics dans la fenêtre w, R le nombre d'évaluateurs, $C_{r,w}$ le nombre de clics de l'évaluateur r dans la fenêtre w, et $\overline{C_r}$ la fréquence moyenne de clics de r. Cette normalisation permet de centrer les valeurs autour de 0, et ainsi de considérer les valeurs positives comme anormalement élevées, et les valeurs négatives comme inférieures à la moyenne. Concrètement, cela ne fait que centrer la courbe des patterns de clics sur 0, mais c'est un moyen intéressant de définir un seuil à partir duquel considérer les pics de clics.

La fréquence de m-clics sur les 5 secondes suivant chaque type de pause et de pattern accentuel est ensuite analysée pour déterminer si les clics ont tendance à augmenter, stagner ou diminuer à la suite de l'événement. Le test de rangs non-paramétrique est à nouveau utilisé pour vérifier la significativité de la différence entre la distribution de valeurs à la suite des pauses inter-proposition et intra-syntagme, et des mots à accentuation de type StressO et StressX.