Université Sorbonne Paris III - Sorbonne Nouvelle

Master de Sciences du Langage, spécialité Recherche et Développement

Vers l'annotation automatique de l'hésitation exprimée dans la parole spontanée

MARIN Apolline

Sous la direction de **Nicolas AUDIBERT**, en collaboration avec **Marie TAHON** et **Jane WOTTAWA** (LIUM)

Année universitaire 2018-2019

Remerciements

Mes remerciements reviennent à mon tuteur de stage et directeur de mémoire Nicolas AUDIBERT, Maître de conférences en Sciences Phonétiques, pour tous ses conseils donnés durant le stage ainsi que pour son suivi régulier.

Je remercie Marie TAHON et Jane WOTTAWA ainsi que d'autres collègues du LIUM pour leur aide ainsi que pour le temps accordé à la relecture de mon mémoire.

Mes derniers remerciements reviennent à ma famille qui m'a soutenue pendant la rédaction de ce manuscrit.

Attestation de non-plagiat

Déclaration sur l'honneur

Je, soussignée Apolline MARIN, déclare avoir rédigé ce travail sans aides extérieures ni sources autres que celles qui sont citées.

Toutes les utilisations de textes préexistants, publiés ou non, y compris en version électronique, sont signalées comme telles.

Ce travail n'a été soumis à aucun autre jury d'examen sous une forme identique ou similaire, que ce soit en France ou à l'étranger, à l'université ou dans une autre institution, par moi-même ou par autrui.

Le 16 juin 2019

Résumé

Ce mémoire a pour vocation de présenter pas à pas les différentes étapes de notre stage effectué au laboratoire de Phonétique et Phonologie (LPP) situé à l'ILPGA (Institut de Linguistique et de Phonétique Générales et Appliquées) dont la finalité est la production d'un système automatique capable de reconnaître et de généraliser d'une manière fiable des éléments dans la parole humaine spontanée.

Dans un premier temps, nous avons dû établir un cadre d'annotation pour relever les éléments pertinents et les marques d'hésitation que nous appellerons ici disfluences et ce dans un corpus de parole spontanée.

Puis, après plusieurs corrections apportées à ce cadre d'annotation, nous avons appliqué manuellement cette annotation au corpus NCCFr (Nijmegen Corpus of Casual French) afin d'attribuer des scores de degrés d'hésitation ainsi que de valence, d'activation et de contrôle aux segments identifiés (5833 en tout) comme porteurs de disfluences. Cette étape d'annotation s'est déroulée en plusieurs phases : deux minutes par locuteur puis dix minutes par locuteur sur un total d'une trentaine de fichiers audio.

Le corpus est constitué d'enregistrements audio de 46 locuteurs français en pleine conversation informelle. Une fois les enregistrements sélectionnés et annotés, nous avons effectué des analyses acoustiques, phonétiques et linguistiques afin de faire des regroupements de vecteurs de paramètres qui serviront à l'apprentissage automatique en suivant le principe de l'apprentissage actif.

Les corrélations observées entre le degré d'hésitation et certaines des variables choisies comme la présence de la voyelle [ə] ou [ø] à la fin d'un mot suivi d'une pause pleine transcrite « euh » ou le nombre de conjonctions consécutives indiquent que l'utilisation de ces variables dans la classification automatique pourrait améliorer les performances de l'annotation automatique, réalisée dans un premier temps par le LIUM à partir de paramètres acoustiques de bas niveau. La mise en relation du degré d'hésitation et des dimensions affectives annotées sur un ensemble de 575 segments suggèrent de plus que la classification pourrait être améliorée par l'annotation en degré de contrôle, intégrée dans certains systèmes existants de reconnaissance des émotions.

MOTS CLEFS : annotation manuelle et automatique, corpus, oral spontané, disfluences, français, apprentissage actif, synthèse de la parole

Table des matières

I. Introduction

1. Contexte

Ce mémoire a pour vocation de présenter pas à pas les différentes étapes de mon stage effectué au laboratoire de Phonétique et Phonologie à l'ILPGA. La finalité est de produire un système automatique de type Text-To-Speech (TTS)[1] capable de générer de la parole spontanée, par exemple dans le cadre de conversations naturelles. La mise en place de ce type de système passe nécessairement par la modélisation de la parole naturelle à partir de bases de données audio annotées. A terme, le système sera développé par le LIUM en collaboration avec le LPP.

Notre méthode repose sur des compétences linguistiques, informatiques et statistiques. Dans ce projet, nous portons notre intérêt sur l'hésitation présente dans le cadre de la parole spontanée et conversationnelle. De manière moins approfondie pour des raisons de temps, les dimensions affectives telles que la valence, l'activation et le contrôle de soi, seront abordées.

Afin de modéliser des éléments de l'hésitation caractéristiques de la parole humaine, il est nécessaire d'avoir à disposition un corpus de parole spontanée enrichi par des annotations précises de ces caractéristiques. Pour réduire le coût de l'annotation (en temps et en ressources humaines), un protocole d'apprentissage actif sera étudié par le LIUM en collaboration avec le LPP.

En étant capable d'apprendre les modèles de manière incrémentale, c'est-à-dire au fur-et-à-mesure de l'arrivée de nouvelles données, l'apprentissage actif peut soit augmenter les performances des modèles en augmentant le corpus d'apprentissage, soit adapter les modèles à un domaine en particulier.

Dans le contexte de ce stage, ma tâche principale était d'établir un cadre d'annotation pour relever les éléments pertinents, c'est-à-dire les marques d'hésitation que nous appellerons ici *disfluences*. Il est assez difficile de trouver une définition standard mais la plupart des auteurs s'accordent pour définir les disfluences comme « un phénomène très courant dans la parole spontanée qui se traduit par l'interruption du flux de parole sans l'ajout de contenu propositionnel » (QUADER Raheel, Oct 2018). Une fois le cadre d'annotation fixé, celui-ci a été appliqué au corpus NCCFr[2] constitué d'enregistrements audio de 46 locuteurs français dans une situation de conversation informelle.

[1] Principe de la synthèse vocale qui convertit des données textuelles en input en parole en output. Il en existe deux types : par concaténation de diphones et paramétrique. Voir https://groupeaa.limsi.fr/_media/membres:cda:tech-ing-cda-2013-h7288v2.pld.pdf pour plus de détails.

[2] Nijmegen Corpus of Casual French : http://www.mirjamernestus.nl/Ernestus/NCCFr/NCCFr_draft.pdf

Après différentes phases telles que la sélection des enregistrements, la segmentation automatique, la correction manuelle de la segmentation et l'annotation manuelle en degrés d'hésitation et en dimensions affectives, des analyses acoustiques, phonétiques et linguistiques ont été effectuées sur les données exploitables afin d'identifier les descripteurs discriminants pour cette tâche. Ces descripteurs ont ensuite été regroupés en vecteurs de paramètres puis fournis à plusieurs modèles de régression permettant de prédire une valeur de degré d'hésitation sur des données *non-annotées* en suivant le principe de l'*apprentissage actif*.

2. Motivations

D'une part, il nous semble important de rappeler que le stage devait comporter des problématiques mêlant linguistique et TAL (Traduction Automatique du Langage) et ce projet de recherche nous a paru être une bonne opportunité car celui-ci comporte effectivement des phases requérant des compétences informatiques et linguistiques.

D'autre part, le sujet d'étude nous était inconnu et a tout de suite suscité notre intérêt. Il nous semble pertinent de rappeler à notre lectorat que les systèmes TTS actuels intègrent rarement les disfluences. Cela peut s'expliquer par le peu de ressources de corpus de parole spontanée disponibles en libre accès.

D'ailleurs, la plupart des études sur les disfluences sont réalisées pour la reconnaissance vocale où l'objectif principal est l'amélioration des modèles de langage et la production de transcriptions comportant des disfluences. Les études sur les disfluences sont très rarement menées pour les générer automatiquement dans la synthèse vocale comme le montre l'étude préliminaire réalisée à l'IRISA (QUADER Raheel, Oct 2018).

3. Description des données

Tout d'abord, il est important de signaler que même si les bases de données sont de plus en plus mises à disposition de la communauté, les corpus de parole spontanée et expressive sont encore difficilement accessibles pour des raisons de protection de la vie privée. De plus, ces corpus sont généralement de petites tailles et ne permettent pas de créer des modèles suffisamment robustes pour la commercialisation. Vous pouvez retrouver une copie du contrat d'accès au corpus NCCFr en annexe (Figure 11).

Notre choix s'est porté sur le corpus de parole spontanée, conversationnelle et expressive NCCFr[3] en français. Ce corpus contient 35 heures d'enregistrements de 46 locuteurs tous francophones natifs et discutant entre amis.

[3] Accessible sur le site Language Archive, dossier NCCFr :
https://archive.mpi.nl/islandora/object/lat%3A1839_00_0000_0000_0018_5D51_C

Lors des séances d'essais d'enregistrement, les organisateurs ont constaté qu'il était compliqué d'obtenir de la parole spontanée sur un grand laps de temps quand il n'y a aucune tâche précise à réaliser ou l'absence de changements. C'est pourquoi les sessions d'enregistrement, qui se font à la suite, sont divisées en trois parties de 90 minutes durant lesquelles la paire de locuteurs discute de manière décontractée de leur quotidien. Les 46 locuteurs participants sont des étudiants provenant tous de la même aire géographique (Nord et Centre de la France) qui suivent la même formation professionnelle. Les groupes de locuteurs ne sont jamais mixtes et l'âge moyen des locuteurs avoisine la vingtaine. Tous ces choix ont été établis par les organisateurs de l'expérience pour permettre aux futurs chercheurs d'étudier la variation de la langue dans les interactions selon des paramètres contrôlés (que sont ici les origines sociales et régionales) et garantir la spontanéité des locuteurs.

Le protocole expérimental suivi était de réunir dans une salle en apparence neutre un groupe de trois locuteurs munis de microphones. L'un des locuteurs (le complice des organisateurs) est invité à s'absenter quelques minutes. Les organisateurs font croire aux deux locuteurs restants que l'un des microphones ne fonctionne pas et qu'il doit être remplacé. C'est à ce moment que la paire de locuteurs se met à parler de manière spontanée et plus familière sur des sujets du quotidien. Quelques minutes plus tard, le complice revient et des consignes sont données au groupe. Les participants sont invités à choisir des questions sur une fiche d'activité, débattre ensemble sur les sujets proposés et négocier une réponse commune.

Les enregistrements audio ont été effectués par Francisco Torreira au Laboratoire de Phonétique et Phonologie (UMR7018) à Paris en Novembre 2008 en tant que projet de sa thèse à l'Université Radboud Nijmegen. Quant à la transcription orthographique, elle a été menée en collaboration avec Martine Adda-Decker (CNRS-LIMSI, France). De plus, les sessions ont toutes été filmées pour permettre l'étude des expressions faciales et gestuelles. Dans le cadre de notre projet, nous nous sommes concentrés exclusivement sur les données audio.

Le choix de ce corpus pour notre étude nous a semblé pertinent car il contient une grande quantité de parole spontanée et informelle dans une situation de dialogue donc nous en avons déduit qu'il y avait de grandes chances que les locuteurs produisent des disfluences. De plus, tous les enregistrements ont bénéficié d'une transcription orthographique par des professionnels, ce qui est très utile comme base pour analyser qualitativement le signal et la phase de segmentation automatique.

4. Applications pour le TAL

A une époque où nous souhaitons donner aux intelligences artificielles la capacité de parler, cette étude vise à réaliser un système automatique capable de détecter les hésitations afin d'annoter de manière semi-automatique un plus grand nombre de données pour reproduire une voix plus ou moins naturelle à l'aide de la synthèse vocale. Plus précisément, nous voulons un système capable de générer des patterns de marques de disfluences telles que les ruptures de syntagmes, les allongements et

les pauses remplies à partir de paramètres acoustiques, linguistiques et statistiques prédéfinis. Etudier les disfluences vise effectivement à améliorer les systèmes de synthèse de la parole se basant sur des modèles de langage humain. L'objectif est de générer de la parole comportant des disfluences qui sera estimée « naturelle » par les potentiels utilisateurs de ce système.

Ce projet entretient bien des liens avec le domaine du TAL puisqu'il repose sur le principe de *l'apprentissage actif*. L'objectif de l'apprentissage actif est de réduire les coûts d'annotation humaine. Un humain produit une petite quantité d'annotations, ce qui permet d'apprendre un premier modèle à partir de ce premier sous-corpus. Le système « annote » ensuite automatiquement le reste des données, chaque annotation étant associée à un score de confiance. L'incertitude permet de savoir à quel point le système, par ses calculs, est sûr de lui sur une prédiction donnée. Suivant le score de confiance, nous allons pouvoir soit conserver l'annotation automatique, soit demander à un humain de confirmer ou d'infirmer la prédiction. Nous exploitons alors l'interaction entre l'expertise humaine et les résultats de l'algorithme afin d'annoter le reste du corpus. Dans le cadre global du projet, l'ensemble des données seront annotées par un humain afin d'évaluer différentes stratégies.

Pour être plus précis, les algorithmes n'apprennent pas directement des données brutes (ici le signal audio) mais des vecteurs de paramètres (ou descripteurs) qui décrivent le signal suivant trois aspects : linguistique, phonétique et acoustique. Il s'agit alors de déterminer quels sont les descripteurs les plus pertinents pour la tâche à réaliser.

Evidemment, la performance du système repose sur le facteur humain. Puisqu'il est celui qui produit l'annotation, s'il commet des erreurs, il est fort probable que l'algorithme d'apprentissage ne produise pas de résultats satisfaisants.

5. Contributions

La principale tâche qui m'a été assignée dans le cadre de ce stage est *l'annotation manuelle* des segments (d'abord deux minutes puis dix minutes pour l'ensemble des locuteurs) issus du corpus NCCFr basé sur un cadre d'annotation que j'ai dû établir au préalable et redéfinir à plusieurs reprises.

En parallèle, mes collègues du LIUM et mon tuteur se sont chargés des phases de la segmentation automatique et du protocole d'apprentissage automatique d'après les données que j'ai annotées manuellement et une transcription orthographique fournie par le LIMSI. Par ailleurs, j'ai aussi apporté des corrections manuelles aux données segmentées automatiquement et aux transcriptions orthographiques et phonémiques.

Enfin, en complément des analyses acoustiques réalisées par mes collègues, j'ai effectué des analyses lexicales pour identifier des descripteurs discriminants afin d'expliquer les degrés d'hésitation associés aux segments qui seront fournis aux modèles de régression.

6. Plan du Mémoire

Une première partie vise à faire un état de l'art sur l'étude des disfluences dans le domaine du TAL.

La section suivante correspond à la démarche méthodologique que j'ai employé pour l'annotation manuelle du corpus, les corrections apportées après la phase de segmentation automatique, le formatage et l'extraction des données puis le calcul de certaines caractéristiques de ces données.

La troisième section de ce rapport sera consacrée aux analyses lexicales et aux résultats obtenus en complément des analyses acoustiques produites par mon tuteur de stage.

Enfin, nous ferons part des résultats préliminaires de l'apprentissage actif réalisé par nos collègues du LIUM ainsi que de possibles perspectives pour améliorer les résultats de notre étude.

Nous allons tenter de confirmer ou d'infirmer les hypothèses suivantes :

- Il y a des corrélations positives entre le degré d'hésitation et la durée du segment comportant une disfluence.
- Les scores de valence, d'activation et de contrôle ne sont pas ou très peu corrélés au degré d'hésitation.
- La présence d'une pause remplie « euh » implique un degré d'hésitation fort.

II. Etat de l'art

1. Intérêt de l'étude de l'hésitation

L'étude des hésitations a longtemps été négligée par les sciences de la parole à l'exception des manifestations de la spontanéité comme les erreurs ou les lapsus étudiés par les psychanalystes. En effet, bien que les hésitations soient des manifestations extrêmement fréquentes de la parole spontanée, elles n'étaient pourtant pas considérées à l'époque comme des éléments linguistiques mais plutôt comme des éléments peu porteurs voire dénués d'informations. Il faut attendre la fin des années 50 ainsi qu'une évolution dans le domaine des sciences de la parole pour observer un essor de l'étude des hésitations et ce uniquement en anglais. Les travaux pionniers sont ceux de Goldman Eisler en 1968 (GOLDMAN-EISLER, 1968) et ceux de Maclay et Osgood (MACLAY, 1959) car ce sont les premiers à considérer ces phénomènes comme de véritables objets linguistiques. En effet, ce n'est qu'en 1972 que paraît la première étude de ce phénomène sur le français (GROSJEAN F, 1972-73). Quant à l'intégration des marques d'hésitation dans les systèmes de reconnaissance automatique de la parole, la première démarche revient à Guaïtella (GUAÏTELLA, 1991).

Les linguistes et les phonéticiens décident de dépasser l'étude des phrases lues et artificielles et de porter leur intérêt sur le discours produit dans les situations d'interaction spontanée. « La constitution de nouveaux corpus et de bases de données plus riches, l'élaboration de méthodes de transcription et de représentation des diverses séquences produites, l'approfondissement du lien existant entre les différentes activités langagières et les messages produits, le développement des technologies de la parole pour la reconnaissance automatique ou encore la nécessité de définir des critères fiables pour distinguer les hésitations des séquences sonores constitutives des mots et syntagmes » sont autant de facteurs qui ont incité cette évolution des mentalités. (DUEZ, 2019)

Dans un premier temps, il nous semble pertinent de revoir la terminologie employée pour désigner les hésitations puis de voir comment celles-ci sont appréhendées dans le domaine des sciences de la parole depuis les années 1900.

2. De la notion d'hésitation aux disfluences

En 1956, Mahl emploie le terme de « disturbance » pour insister sur la notion de raté dans la production de la parole. L'hésitation est vue comme une marque d'anxiété, une scorie au même titre que le lapsus ou l'erreur. Mahl répertorie huit catégories dont les pauses remplies (ah), les corrections de phrases, les phrases incomplètes, les répétitions de mots, les bégaiements, l'intrusion de sons incohérents, les erreurs

(néologismes, substitutions, transpositions) et les omissions de mots ou parties de mots.

La fréquence de ces différentes catégories est calculée en relation avec le nombre de mots par phrase par rapport aux temps de pauses silencieuses durant le temps total de parole. Selon Mahl, la fréquence des hésitations est significativement plus élevée dans les phrases marquées par une grande anxiété. Il remet en question la validité de sa typologie et soutient qu'il existe des stratégies individuelles en fonction de la tâche linguistique. Lorsque le niveau d'abstraction est élevé, les locuteurs qui hésitent le plus lors des pauses silencieuses sont ceux qui produisent le moins de pauses remplies, les phrases le plus brèves et les mots les plus rares. En revanche, les locuteurs qui produisent le plus de pauses remplies sont ceux dont les phrases sont les plus longues et les mots les plus fréquents.

Les résultats récents de Christenfeld et Freager (1996) sont en accord avec la typologie de Mahl. « Ils observent une forte augmentation des répétitions et des faux départs chez les locuteurs hésitants mais une diminution des pauses remplies dans la parole produite dans des conditions générant l'anxiété ».

C'est dès 1959 que le terme d'hésitation est employé par Maclay et Osgood. Il est repris dans la majorité des travaux qui visent à établir un lien entre variations temporelles de la parole et activité cognitive. L'occurrence des hésitations paraît liée au degré d'incertitude du mot subséquent. Le nombre d'hésitation semble être plus fort devant les mots lexicaux que devant les mots grammaticaux et plus fort dans les intrasyntagmes qu'à leur frontière. Ils constatent que les faux-départs concernent essentiellement les mots lexicaux et les répétitions des mots grammaticaux situés devant des mots lexicaux. Les pauses remplies et les pauses silencieuses sont quant à elles distribuées devant les mots lexicaux avec une tendance plus marquée pour les pauses silencieuses.

Maclay et Osgood observent également que les hésitations sont des éléments nécessaires à la langue puisqu'ils participent activement au processus de sélection lexicale et grammaticale. Les hésitations aident le locuteur à identifier et à structurer les unités linguistiques. Elles permettent aussi d'attirer l'attention de l'interlocuteur sur les mots à contenu informatif élevé.

Entre 1972-1975, Grosjean et Deschamps différencient les hésitations de la vitesse d'élocution et de la durée des pauses silencieuses. Kowal, O' Connell et Sabin (1975) les désignent comme des hésitations vocales qu'ils opposent aux hésitations gestuelles.

En 1979, Siegman confirme le lien existant entre hésitation et décisions d'ordre cognitif prises par le locuteur et fait l'hypothèse que la nature des différentes hésitations est affectée par les relations sociales et interpersonnelles. Un locuteur a tendance à accélérer son débit de parole et à marquer moins de pauses silencieuses pour garder son tour de parole.

En 1996, Christenfeld et Creager relient les hésitations aux degrés d'attention consciente porté au contenu du message transmis par un locuteur. Cela suggère que les locuteurs contrôlent ce qu'ils disent à plusieurs niveaux et que les pauses remplies sont produites lorsque les locuteurs détectent une erreur ou une erreur à venir et qu'ils sont capables de s'arrêter pour se corriger. Si le processus de parole est un processus conscient alors les pauses remplies en seraient l'un des symptômes.

De nos jours, nous tendons à employer le terme de disfluences (disfluencies en anglais) car elles englobent les hésitations, les erreurs et les lapsus. Nous supposons que Susan C. Meyers, Frances et J. Freeman sont les premiers à avoir employés ce terme dans l'étude de la parole spontanée (MEYERS, 1985).

Contrairement à ce que nous pourrions croire, les disfluences facilitent la synchronisation avec l'interlocuteur dans une situation conversationnelle. Elles permettent notamment d'améliorer la compréhension du message véhiculé en créant des pauses silencieuses dans le flux de parole et signalent à l'interlocuteur la complexité du message en cours de transmission ». Le terme est proche de « disturbance » dans la mesure où il souligne la rupture (produite ou perçue) qui intervient dans la fluidité de la parole. (QUADER Raheel, Oct 2018)

Le terme de disfluence coexiste avec le terme « dysfluencies » mais celui-ci est plutôt réservé au domaine de la parole pathologique, par exemple pour faire référence aux cas de bégaiement (FREEMAN Jackson, 1995).

3. Pauses silencieuses et pauses remplies

De façon traditionnelle, nous distinguons deux types de pauses : les pauses silencieuses dans lesquelles toute production vocale s'interrompt à l'exception de bruits respiratoires et les pauses remplies (ou sonores) constituées d'une unité dite quasi-lexicale comme le euh en français.

Cependant, le terme de pause sonore est souvent jugé comme n'étant pas suffisamment neutre car rien ne permet de classer les marques d'hésitation du côté des pauses, leur rôle n'étant pas uniquement de marquer un temps dans la production d'un énoncé mais aussi d'indiquer à l'auditeur que ce temps est destiné à poursuivre l'encodage sans pour autant céder son tour de parole.

Certaines pauses silencieuses sont considérées comme étant *démarcatives*. Elles apparaissent à la jonction de segments du discours et participent à sa structuration. A l'exception de celles qui sont très brèves, les pauses silencieuses n'ont généralement pas qu'un rôle de marque d'hésitation. Tous les auteurs s'accordent pour dire que ces marques peuvent être combinées : « Pour assurer une telle fonction, elles sont

obligatoirement combinées à d'autres marques de travail de formulation comme un allongement ou une pause remplie euh » (CAMPIONE Estelle & VERONIS, 2004).

Le terme de « marques du travail de formulation *» proposé* par Morel et Danon-Boileau (DANON-BOILEAU, 1998) et défendu par Maria Candea (CANDEA, 2000) tend à montrer que le terme d'hésitation est souvent utilisé de manière abusive pour qualifier les pauses remplies euh et les allongements vocaliques dans les corpus oraux spontanés français. Les auteurs considèrent que les allongements vocaliques en fin de mots n'ont pas la même distribution syntaxique que les euh et avancent l'hypothèse que ces deux marques de travail de formulation pourraient avoir une portée différente (séquence cible plus limitée pour les allongements vocaliques que celles introduites par un euh).

D'autres auteurs comme Grosjean et Deschamps (GROSJEAN F, 1972-73) classent les pauses remplies euh et les allongements vocaliques du côté du temps total d'élocution, considérant que ces deux marques jouent le même rôle et fonctionnent de la même manière (plus il y aurait d'allongements moins il y aurait de euh et inversement). DUEZ conteste partiellement ce point de vue et propose de regrouper les euh du côté du temps total de la pause et de laisser les allongements vocaliques du côté du temps d'élocution.

Dans (DUEZ D. , 1999), l'auteure définit les pauses silencieuses comme « une cessation de l'activité verbale qui se traduit au niveau acoustique par une interruption du signal sonore ». Il est évident que ces pauses silencieuses servent aux phases de respiration du locuteur mais celles-ci peuvent aussi participer à la phase de recherche lexicale, c'est-à-dire lorsque le locuteur structure l'énoncé à venir.

L'auteure tend à distinguer les pauses silencieuses *des pauses silencieuses dites d'hésitation* qui dépendent du degré de spontanéité de l'énoncé et « tendent à être distribuées à l'intérieur des syntagmes ou associées à des hésitations sonores ».

Selon (CAMPIONE Estelle & VERONIS, 2004), « les pauses dites d'hésitation reflètent la difficulté que rencontre ponctuellement le locuteur dans ses opérations de recherche lexicales et d'encodage (travail de formulation), opérations liées à la production du discours ».

Quant aux pauses non silencieuses ou dites *remplies*, DUEZ s'appuie sur la typologie de Maclay et Osgood (MACLAY, 1959) qui regroupe les interjections comme « euh », « um », les faux départs, les syllabes allongées et les répétitions.

Les pauses remplies sont généralement cantonnées au rôle de « signal conventionnel de la part du locuteur », ce qui lui permet de garder son tour de parole et de faire comprendre à son interlocuteur que ce laps de temps lui sert à construire la suite de son énoncé. Or, ce rôle est aussi destiné à certains allongements syllabiques

(généralement affectant une voyelle en finale de mot). Ces « allongements d'hésitation » possèdent des propriétés proches de celles des pauses remplies et pourtant «la terminologie établie regroupe des phénomènes acoustiquement et fonctionnellement différents » sous le terme de pause mais distingue les allongements d'hésitation des « euh ».

Quant à Guaïtella (GUAÏTELLA, 1991), elle choisit de confondre les deux marques et de les classer dans la catégorie des hésitations vocales sans pour autant justifier son choix théorique.

4. Valence, activation et contrôle

Les termes de *valence*, d'*activation* et de *contrôle* sont issus du domaine de la psychologie cognitive des émotions. Au départ, le logicien Russell propose un premier modèle à deux dimensions (RUSSELL, 1980) pour représenter les émotions autour d'un cercle à deux axes : les dimensions de valence (qualité positive ou négative associée à un stimulus ou à une situation) et d'activation (niveau d'excitation faible ou fort associé à un stimulus ou à une situation).

La troisième dimension de contrôle aussi parfois appelée dominance est ajoutée par la suite par Osgood (OSGOOD et al., 1975 cités dans (CLAVEL Chloé, 2006)) qui redéfinit la dimension d'activation comme un état d'excitation, de calme à fortement excité, lié à l'intensité. Il remplace la valence par la dimension d'évaluation (appraisal) pour positionner l'émotion sur un axe négatif et positif. La dimension de contrôle correspond à l'effort du locuteur pour contrôler son émotion.

III. Méthodologie pour l'enrichissement du corpus NCCFr

1. Nos premiers pas dans le projet

Dans un projet de recherche antérieur sur la génération de la parole expressive (TAHON, 2018), notre collègue Marie Tahon était déjà parvenue à reconstruire des spectrogrammes de parole issus de livres audio du genre contes pour enfants. Selon l'étude, « les livres audio constituent un bon corpus pour les systèmes TTS puisqu'ils contiennent du texte, différents personnages avec divers registres de langue et d'émotions et le signal sonore correspondant ». Cependant, il s'agit de parole lue donc ni spontanée ni conversationnelle. Nous allons donc nous inspirer de la méthode utilisée dans cette étude en l'appliquant à la parole spontanée que nous jugeons encore trop peu souvent traitée.

En premier lieu, nous voulons modéliser des caractéristiques de la voix afin de permettre à un système de synthèse vocale d'intégrer certaines de ces caractéristiques. Par exemple de la voix générée à partir de parole lue pourrait être ensuite transformée en parole conversationnelle.

Dans le cadre de mon stage, les unités pertinentes sont les *segments* qui correspondent à des moments de parole d'un locuteur compris entre deux groupes de souffle (pauses respiratoires et/ou silencieuses). Le découpage automatique de ces groupes de souffle a été effectué par mon tuteur à l'aide d'un script Praat. Il s'est appuyé sur l'alignement forcé en phones du système du LIUM en prenant comme seuil de durée la plus courte des pauses silencieuses annotées manuellement dans les productions d'un même locuteur (un seuil de durée par locuteur pour tenir compte des différences interindividuelles de débit de parole). L'alignement du LIUM a permis d'obtenir deux champs de transcription : orthographique et phonétique. Il faudra donc distinguer les transcriptions de l'*annotation* qui désigne ici la tâche d'attribution manuelle de valeurs ou d'étiquettes aux segments considérés.

L'option de rétablir les limites des segments avec une structure prosodique naturelle (la ponctuation) a été mise de côté. A la place, nous avons choisi de construire un cadre d'annotation exploitable pour cibler les segments voulus et faciliter la tâche d'annotation manuelle des extraits du corpus (préalablement segmentés par un système automatique). L'annotation manuelle est une étape complémentaire à l'annotation automatique qui permet de diminuer le taux d'erreurs. Pour donner un exemple concret, le système automatique de segmentation peut ne pas percevoir certains phonèmes car il considère ceux-ci comme des silences. Nous avons pu observer ce cas fréquemment lorsqu'il détectait une consonne sourde.

Le protocole suivi pour l'annotation manuelle peut être décomposé en trois grandes phases : j'ai d'abord élaboré un schéma d'annotation pour segmenter quelques échantillons du corpus NCCFr puis j'ai dû réaliser une annotation manuelle complète

sur les deux premières minutes de chaque locuteur. Enfin, j'ai annoté huit minutes supplémentaires sur l'annotation de deux minutes pour l'ensemble des locuteurs. Ces données serviront à créer un corpus d'apprentissage qui sera fourni à plusieurs modèles de régression dans l'optique d'effectuer un protocole d'annotation interactive. Cependant, pour des raisons de temps, cette phase ne sera pas traitée dans ce mémoire mais reléguée en tant que perspectives du projet.

L'annotation interactive permet d'optimiser l'étape d'annotation des données pour limiter les coûts humains et financiers. Cette approche consiste à utiliser des techniques d'apprentissage automatique pour demander à l'expert humain d'annoter uniquement les échantillons pertinents. Le travail de ce stage s'inscrit dans une étude préliminaire visant à évaluer différentes stratégies d'annotation interactive. C'est pour cela qu'il est nécessaire d'annoter une grande partie des données en amont.

En ce qui concerne la paramétrisation des modèles de régression, mes collègues du LIUM ont fait le choix de se limiter dans un premier temps à un nombre réduit de paramètres MFCC[4]. Les MFCC sont des coefficients spectraux qui permettent de décorréler la source (Fréquence fondamentale F0) du filtre (formants, conduit vocal). Ils reposent sur une échelle de fréquence perceptive logarithmique appelée échelle de Mél. Sur un nombre exponentiel de coefficients, mes collègues ont choisi d'utiliser les vingt premiers coefficients MFCC ainsi que leur dérivée première (leur moyenne et écart-type sur le segment), soit 80 descripteurs. Selon (HACINE-GHARBI, 2012), « il est nécessaire, si l'on veut concevoir un système acceptable en termes de précision, coût de calcul, et d'encombrement mémoire, de limiter le nombre de paramètres en sélectionnant les plus pertinents susceptibles de modéliser le mieux possible les données pour la tâche de reconnaissance ».

2. Etablissement d'un cadre d'annotation

Avant d'avoir accès au corpus complet du NCCFr, nous avions pour mission de segmenter des échantillons audio (fichiers audio et leurs fichiers de transcriptions associés). Deux transcriptions orthographiques étaient à notre disposition : celle diffusée avec les fichiers audio par NCCFr (manuelle) et l'annotation (manuelle) réalisée au LIMSI (Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur). Notre première tâche consistait à créer un cadre d'annotation qui sera ensuite appliqué à une grande partie du corpus NCCFr[5].

Dans un premier temps, nous avons défini les segments (phrases ou mots) que nous voulons cibler à partir d'un petit échantillon du corpus. Nous avons jugé plus raisonnable de délimiter les segments situés entre deux pauses respiratoires pour cibler les marqueurs d'hésitation dans la parole des locuteurs.

Dans la mesure du possible, nous avons choisi d'éviter de segmenter les intervalles comportant du rire bien qu'il soit possible que les pauses silencieuses en contiennent.

[4] Mel Frequency Cepstral Coefficients.

[5] Nous n'avons pas choisi la totalité des fichiers du corpus NCCFr (une trentaine environ).

Ensuite, nous avons dû définir plusieurs champs d'annotation. C'est une étape essentielle qui conditionne toute la suite de l'annotation. Cette étape a été réalisée soigneusement et en collaboration avec mes collègues experts du domaine. Nous avons d'abord défini le maximum de champs afin de caractériser les hésitations pour ensuite les réduire à ceux nous semblant les plus pertinents. Cette étape de réduction était nécessaire pour faciliter l'annotation et éviter de perdre du temps : quand il y a trop d'informations affichées dans Praat, le logiciel a tendance à être moins efficace et il devient difficile de distinguer les intervalles à annoter.

Je vais maintenant décrire l'évolution du cadre d'annotation que j'ai établi en parallèle de mes recherches documentaires. Les trois premiers cadres d'annotation ont été appliqués aux échantillons audio. Le cadre final quant à lui sera conservé pour l'annotation manuelle du corpus NCCFr.

3. Choix et justifications des champs d'annotation

Dans un premier temps, mon cadre d'annotation (Tableau 1) était constitué de six champs. Puis, j'ai proposé un schéma d'annotation constitué de huit champs que j'ai progressivement réduit à ceux me semblant les plus pertinents pour notre étude des hésitations. En effet, au fur-et-à-mesure des modifications apportées, j'ai pu constater que le cadre d'annotation était devenu trop riche et trop fin pour notre sujet d'étude.

De plus, même en réduisant certains champs à une simple légende de trois lettres (par exemple MAD pour les marques discursives), l'annotation de ces champs demeurait une tâche trop complexe et surtout extrêmement chronophage. C'est pourquoi, avec l'approbation de mes collègues, j'ai fait le choix d'alléger le cadre d'annotation voire de fusionner certains champs déterminés comme trop redondants. Par exemple, les champs **« indicateurs d'auto-interruption »** et **« disfluences »** dans le Tableau 2 ont été fusionnés dans le Tableau 3.

Les pauses respiratoires correspondent aux phases de respiration et d'expiration des locuteurs. Généralement, le seuil de durée des pauses respiratoires environne les 200 millisecondes. Quant aux **pauses silencieuses**, ce sont des phases où les locuteurs ne parlent pas et où aucune respiration n'est audible.

Nous avons fait en sorte de mettre de côté le rire en les intégrant dans ces périodes de silence car les locuteurs expriment une forte euphorie sans produire véritablement de parole. Les pauses respiratoires et silencieuses seront par la suite regroupées avec les marqueurs discursifs et les pauses remplies dans le champ d'indicateurs d'auto-interruption puis classées en tant que disfluences car j'ai considéré que les pauses silencieuses pouvaient constituer une marque d'hésitation du locuteur pour la sélection lexicale par exemple.

Le « type d'hésitation » était d'abord une catégorie assez générale me permettant d'ajouter les éléments perçus comme étant pertinents pour l'annotation du degré d'hésitation. Dans un premier temps, cette classe regroupait les faux départs, des éléments au statut particulier comme les interjections, les allongements de mots, les troncations de mots et les pauses remplies tels les « euh ». Par la suite, j'y ai placé toutes les marques de disfluences que j'avais répertoriées après de plus amples recherches documentaires. C'est pourquoi les **« types d'hésitation »** sont devenus les **« Disfluences »**. Ce champ a été enrichi avec les cas de reprises, les tics de langage (tu sais, voilà, enfin…) et les réparations.

Pour annoter les disfluences, j'ai réduit les termes jargonneux à leurs initiales : (PR) pour les pauses remplies, (PS) pour les pauses silencieuses, (MAD) pour les marques discursives, (FD) pour les faux-départs, (RED) pour les réductions, (REP) pour les reprises, (RPA) pour les réparations et (ALL) pour les allongements.

Les « degrés d'hésitation » correspondent aux scores que je dois attribuer aux différents types d'hésitation. Il m'a été conseillé de représenter ces niveaux d'hésitation sous la forme d'une échelle de Likert. J'ai donc opté pour une échelle ordinale à cinq niveaux qui a été étendue au fil de l'annotation à huit niveaux. Après plusieurs écoutes, il m'a semblé pertinent de considérer les hésitations ainsi que la certitude car certains segments étaient difficilement associables à de l'hésitation.

Un niveau d'annotation spécifique aux **émotions et/ou affects** associés aux types d'hésitations a été intégré au schéma d'annotation. Pour des raisons de temps, ce champ n'a été annoté que sur la partie permettant l'apprentissage d'un premier modèle de prédiction du degré d'hésitation (les fichiers de deux minutes par locuteur). Une annotation complètement automatique du reste du corpus est envisagée dans la continuité du projet.

J'ai choisi d'annoter ce champ suivant trois axes affectifs (RUSSELL, 1977, voir la section Etat de l'art) qui sont couramment utilisés dans la littérature et en informatique affective : valence (émotion positive ou négative), activation (passive ou active) et contrôle de soi. Pour des raisons de lisibilité, j'avais choisi un petit système de légende. Par exemple, si la valence est positive, j'ajoutais le symbole « + ». Si elle est négative, le symbole « - » et la lettre « n » pour une valence neutre. Pour le degré d'activation, des étiquettes telles que « calme » ou « excité » et pour le degré de contrôle du locuteur, « c » ou « nc » (contrôlé/non contrôlé). Finalement, ces dimensions ont été annotées sur une échelle de Likert à cinq niveaux afin que celle-ci soit homogène avec celle associée aux degrés d'hésitation.

J'avais aussi créé un champ pour annoter les **« fonctions pragmatiques »** car la situation d'interaction étant un dialogue entre deux personnes, j'avais jugé intéressant d'ajouter des informations sur la fonction des disfluences. En effet, les « hésitations peuvent permettre d'attirer l'attention de l'interlocuteur sur une nouvelle information (O'

CONNELL, KOWAL & HÖRMAN, 1969) ou créer une atmosphère d'intimité et/ou de proximité entre les participants dans un discours » (DUEZ, 1997) cités dans (DUEZ, 2019). Un champ **« Bruits de bouche »** complète cette annotation car nous le considérons comme pertinent pour l'interprétation des affects : l'arrondissement, la protrusion des lèvres ou encore les bruits d'explosion lors de la production de consonnes peuvent influencer le message transmis et la perception de celui-ci par l'interlocuteur.

Le champ « **Structure de surface des hésitations** » a été ajouté ensuite car je pensais que cette terminologie m'aiderait à délimiter les segments contenant les marques de disfluences. Je me suis inspirée de l'article de Pallaud, Rauzy et Blache (2013) qui définit « le phénomène de reprise d'énoncés (par rapport à un abandon) après une auto-interruption disfluente » comme « une caractéristique dominante des énoncés oraux (de parole spontanée) ».

Plus précisément, selon les auteurs, « Les auto-interruptions et les auto-variations dans la fluence verbale sont signalées dans la grande majorité des cas par l'apparition d'une ou plusieurs sortes d'évènements : Les pauses silencieuses ou remplies et les items insérés au milieu d'un syntagme ou même d'un mot qui peuvent s'accompagner de perturbations morphosyntaxiques du flux verbal, c'est-à-dire des reprises d'énoncés, des auto-réparations ou encore l'inachèvement de syntagmes et/ou de mots. »

Les auteurs distinguent deux types d'interruption : « les disfluentes et les suspensives qui ne provoquent qu'une suspension temporaire et non une réorganisation de l'énoncé ». Parmi les interruptions disfluentes, ils relèvent deux catégories. D'une part, celles qui provoquent « la reprise d'une partie de l'énoncé interrompu avec une éventuelle modification de l'énoncé repris » (exemple donné par les auteurs : mais les euh les nanas du foyer). D'autre part, « Celles qui, l'énoncé ayant été laissé inachevé, sont suivies d'une nouvelle construction ou d'un nouveau syntagme » (nouvel exemple fourni par les auteurs : ah si mais j'ai// c'est un truc qui m'avait fait bien rire).

Ils regroupent dans la catégorie auto-interruptions les espaces interregnum : (« interruptions dans le flux verbal d'une durée variable »), les pauses silencieuses (« d'une durée supérieure ou égale à 200ms »), les pauses remplies et les marqueurs discursifs.

L'interruption délimite trois espaces formels : « Le Reparandum qui est ce qui, avant le point de rupture, contient une perturbation et sera simplement poursuivi, repris, répété, modifié ou abandonné lors du Reparans ».

Ensuite, « L'Interregnum qui correspond au moment potentiel avant que n'intervienne le Reparans. Il peut être vide ou contenir des indices de disfluence le plus souvent non lexicalisés comme les pauses remplies et/ou silencieuses, les

répétitions de troncation, les éléments discursifs et/ou parenthétiques plus ou moins longs, les onomatopées… »

Enfin, « Le Reparans : la partie potentielle de l'énoncé prononcé qui peut poursuivre, répéter ou modifier ce qui a été dit lors du Reparandum. Cet élément comporte deux situations selon qu'il est vide ou rempli. Si le Reparans est vide, l'énoncé est alors laissé inachevé. Dans le cas contraire, trois cas de figure peuvent se présenter : soit le syntagme interrompu est complété, soit l'énoncé déjà prononcé est repris partiellement (entassement paradigmatique), soit l'énoncé est repris avec des modifications ».

Voici un de leurs exemples : « Tu perds un peu comment dire euh + un peu des repères »

Tu perds un peu	comment dire euh +	un peu	des repères
REPARANDUM	INTERREGNUM	REPARANS	

A propos de la localisation de ces éléments, « L'analyse morphosyntaxique de ces disfluences lexicalisées (PALLAUD 2002 ; PALLAUD & HENRY, 2004) a montré que le Reparandum ne peut être identifié qu'à l'aide des éléments qui vont lui succéder et tout particulièrement ce qui va être repris de l'énoncé avant le point de rupture (le Reparans). Lorsqu'il y a inachèvement de l'énoncé, le Reparandum correspond à l'item tronqué ou le dernier item du syntagme laissé inachevé. »

Un champ **« Tours de parole »** me semblait pertinent pour voir si les locuteurs ont tendance à se couper la parole. En combinant les fichiers audio et de transcriptions, j'ai pu remarquer des cas de chevauchement de tours de parole et donc déterminer ce qui était perceptible et possible d'annoter. Cependant, comme nous avons finalement traité les fichiers de manière isolée (locuteur par locuteur), ce champ était devenu superflu et a été retiré du schéma d'annotation.

Le cadre d'annotation final (Tableau 4) a été validé par l'ensemble de l'équipe et appliqué pour annoter manuellement deux minutes de parole dans un fichier audio combinant deux locuteurs (32 fichiers en tout) sur l'ensemble des données audio disponibles soit environ 64 minutes en tout et 23 minutes de parole après alignement. Quand nous avons décidé d'étendre les deux minutes de parole par locuteur à dix minutes par fichiers audio soit 320 minutes en tout et 116 minutes de parole, nous avons finalement conservé les deux derniers champs **« valence, activation, contrôle »** et **« Degré d'hésitation/certitude »** du cadre d'annotation après l'alignement en mots et en phones réalisé par nos collègues du LIUM.

4. Evolution des cadres d'annotation illustrés d'exemples

Les quatre tableaux ci-dessous retracent l'évolution du cadre d'annotation que j'ai dû établir et servent à illustrer mes choix de champs d'annotation avec des exemples.

Champs d'annotation	Exemples et/ou étiquettes associées
Pauses respiratoires et silencieuses	PR et PS
Type d'hésitation Faux-départs Interjections Allongements de mots Troncations Pauses remplies	 *et le la maison* *&hum, &ben…* *Mais tu as (à l'écoute la voyelle [a] est allongée)* *c'est le pro(fesseur)* *euh*
Degré d'hésitation	Compris entre 1 et 5
Émotion/Affect Associé + Bruits de bouche	Joie, Colère… B pour bruits de bouche
Fonction pragmatique	Processus de sélection lexicale, exprimer son incertitude, emphase sur une information…

Tableau 1 : Premier cadre d'annotation

Champs d'annotation	Exemples et/ou étiquettes associées
Structure de surface des hésitations (Reparandum, Interregnum, Reparans)	*Tu perds un peu comment dire euh + un peu des repères* (voir plus haut)
Indicateurs d'auto-interruption Marqueurs discursifs Pauses respiratoires Pauses remplies	 *Tu vois, bon…* RSP PR
Disfluences les faux-départs les réductions (=troncation) les reprises les tics de langage les réparations	 *Le la souris* *C'est pas moi qui pa()* *Tu sais je// je l'ai vu hier* *T'sais, enfin, voilà quoi* *Je voulais dire euh // je voulais dire ça*.
Tours de parole	Tdp loc1/2
Bruits de bouche	B pour bruits de bouche
Degrés d'hésitation	Echelle de Likert à cinq points de 1 à 5 (hésitation faible à hésitation très forte)
Fonction pragmatique	Processus de sélection lexicale, exprimer son incertitude, emphase sur une information…
Emotions/Affects associés	Joie, Colère…

Tableau 2: Deuxième cadre d'annotation

Champs d'annotation	Exemples et/ou étiquettes associées
Bruits de bouche	B
Disfluences	(PR) pour les pauses remplies, (PS) pour les pauses silencieuses, (MAD) pour les marques discursives, (FD) pour les faux-départs, (RED) pour les réductions, (REP) pour les reprises, (RPA) pour les réparations (ALL) pour les allongements
Valence, Activation, Contrôle	Echelle de Likert à cinq points : -2 à 2
Degré d'hésitation	Echelle de Likert à cinq points de 1 à 5

Tableau 3 : Troisième cadre d'annotation

Champs d'annotation	Exemples et/ou étiquettes associées
Mots	Se base sur la transcription orthographique manuelle fournie avec le corpus
Phones	Se base sur la transcription en phonèmes fournie avec le corpus et l'alignement en phones du LIUM
Valence, Activation, Contrôle	Echelle de Likert à cinq points : -2 à 2
Degré d'hésitation et/ou de certitude	Echelle de Likert à huit points de -3 à 5

Tableau 4 : Cadre d'annotation final

5. Annotation manuelle

Une fois le corpus NCCFr disponible (Voir la section I-3) et avant de me lancer dans l'annotation, j'ai effectué quelques analyses qualitatives des fichiers audio pour me familiariser au contenu. J'ai écouté des extraits à plusieurs reprises afin de déterminer lesquels étaient les plus pertinents à sélectionner pour l'annotation. Par exemple, je n'ai pas sélectionné les fichiers comportant beaucoup de silences ou de rires.

Chaque sous-partie comporte deux canaux d'enregistrement correspondant au microphone porté par chaque participant. Une étape de vérification des correspondances entre les fichiers de transcriptions et chacun des canaux a été nécessaire pour s'assurer de la bonne qualité de l'alignement automatique. En effet, l'alignement automatique nécessite d'associer la transcription orthographique avec le signal audio correspondant.

Dans le cas où l'audio et la transcription ne correspondent pas, l'alignement va devoir chercher le signal de la transcription du locuteur qui ne porte pas le microphone mais qui est enregistré en arrière-plan dû à la proximité des locuteurs dans la pièce, ce qui risque d'engendrer des erreurs concernant l'alignement phonétique.

Bien que nous disposions de deux transcriptions orthographiques (celle diffusée avec les fichiers audio par NCCFr et l'annotation réalisée au LIMSI) nous avons opté pour les fichiers de transcriptions préalablement annotés par le LIMSI car après vérification, ils correspondaient davantage à nos besoins. De plus, la fréquence d'échantillonnage choisie par le LIMSI et celle du corpus d'origine NCCFr sur le site de Language Archive diffèrent : 16000 Hz pour le LIMSI contre 48000 Hz pour Language Archive. Cependant même constat pour tous les fichiers audio, il semblerait y avoir des problèmes d'en-têtes mais nous en ignorons la source.

Afin d'identifier la durée cumulée des pauses remplies et des intervalles autres que les silences, les traitements ont été réalisés à l'aide de scripts Praat pour extraire des valeurs temporelles à partir des fichiers de transcription et des valeurs acoustiques à partir des fichiers audio.

Pour l'annotation des échantillons de test, j'avais combiné les parties afin d'évaluer les performances sur une paire de locuteurs. Après de nombreuses écoutes, j'ai remarqué que certains enregistrements n'étaient pas coupés au même moment. Il y avait donc des petits décalages et de l'écho, ce qui me compliquait un peu la tâche d'annotation. C'est pourquoi nous avons finalement opté pour l'annotation des fichiers de manière isolée (locuteur par locuteur).

La tâche d'annotation manuelle requiert de la concentration et de la constance. A titre indicatif, l'annotation et la segmentation de deux minutes de parole (silences inclus) sur l'ensemble des champs du schéma défini plus haut prend entre deux et trois heures. Ainsi, l'annotation de deux minutes de parole pour l'ensemble des locuteurs pouvait atteindre les 45 heures.

Effectivement, la tâche d'annotation ne consiste pas uniquement à compléter des cases selon un cadre d'annotation préétabli. C'est bien un travail qui demande de la concentration (beaucoup de réécoute du signal sonore), de la réflexion pour choisir les valeurs et les scores à attribuer aux segments et qui demande beaucoup de patience (pour corriger les erreurs de segmentation par exemple).

Par ailleurs, j'ai aussi hésité sur les valeurs à attribuer aux segments, notamment pour le champ de **« Valence, activation et contrôle »**. En effet, la perception des affects est complètement dépendante des individus car elle repose sur leurs expériences et leur perception. La question de l'objectivité des annotations émotionnelles est généralement résolue en prenant le vote majoritaire obtenu sur plusieurs annotateurs mais elle reste une problématique centrale dans le domaine (DEVILLERS Laurence, 2005).

De plus, l'élément neutre prédomine largement le spectre des annotations dans le cas de la parole spontanée dans toutes les études menées dans ce domaine. Par conséquent, je sais que ma manière d'annoter n'était pas toujours la même bien que je sois parvenue à conserver une stabilité en ce qui concerne les longueurs des segments. Mon annotation est nécessairement subjective et il ne servirait à rien d'essayer de la rendre objective. Qu'elle ne soit pas constante au cours du temps est un autre élément à prendre en compte mais qui est fatalement humain : Il n'existe pas de VRAI mais une marge de probabilité de tomber dans le PROBABLEMENT VRAI. C'est donc ma perception que nous allons modéliser et celle-ci prévaudra ensuite pour le système de synthèse.

Ainsi, le champ des affects m'a paru être le plus difficile à annoter car pour des émotions telles que la joie ou la colère, il est facile de dire si un segment est porteur d'une valeur positive ou négative mais pour les hésitations, cela me semblait plus compliqué. A tel point que je mettais presque partout des valeurs nulles pour indiquer la neutralité des segments.

Grâce à tout ce travail, nous disposons d'un premier sous-corpus consistant enrichi d'annotations précises sur l'hésitation et les affects. Afin d'avoir suffisamment de données pour étudier différentes stratégies d'apprentissage actif, nous avons jugé pertinent d'étendre l'annotation à dix minutes de parole par locuteurs, ce qui a étendu considérablement le temps dédié à l'annotation : une trentaine de fichiers à annoter soit plus d'un mois de travail d'annotation.

Pour annoter ces huit minutes supplémentaires, nous avons fait le choix de ne conserver que le dernier champ « Degré d'hésitation ». Les bruits de bouche ont été mis de côté car ils appartiennent à une catégorie bien trop fine et trop complexe pour être détectée par notre système car celui-ci est avant tout entraîné sur des catégories de parole comme les phones. Comme dit précédemment, nous n'avons également pas poursuivi l'annotation de la valence, de l'activation et du degré de contrôle des segments mais les données des segments annotés sur deux minutes par locuteur seront tout de même utiles par la suite, notamment pour compléter nos analyses.

En parallèle, un collègue du LIUM s'est chargé de relancer un alignement en phones et en mots avec le système du LIUM en s'appuyant sur la transcription orthographique fournie par le LIMSI. Cet alignement avait pour but de délimiter les unités temporelles auxquelles attribuer des scores de degré d'hésitation. En attendant que l'alignement soit terminé, notre mission suivante consistait à définir des descripteurs linguistiques, notamment en faisant le repérage d'indices linguistiques comme des séquences de mots qui se répètent ou des motifs particuliers associés de façon récurrente à des degrés d'hésitation plus ou moins élevés.

Par ailleurs, le problème de correspondance entre fichiers audio et de transcriptions n'était pas totalement résolu. Nous avons donc décidé de faire un fichier

tabulaire afin de répertorier tous les noms des fichiers afin de déterminer lesquels présentent des erreurs de correspondance et les harmoniser entre eux. L'initiative n'a pas été vaine puisque sur la trentaine de fichiers, un peu plus d'une dizaine présentait cette erreur.

Ce n'était pas sans peines mais nous disposons maintenant d'un corpus exploitable ! Nous pouvons maintenant passer à l'étape de l'analyse des données de notre corpus.

6. Analyses linguistiques en complément des analyses acoustiques

En parallèle des analyses acoustiques réalisées par mes collègues et mon tuteur, je me suis chargée de relever des descripteurs linguistiques pour prédire le degré d'hésitation d'un segment.

Une première session d'analyse du corpus a été réalisée à l'aide de l'outil lexicométrique Lexico5[6]. Pour récupérer le contenu textuel qui nous intéressait (les segments), nous avons procédé au nettoyage de certains symboles spéciaux tels les marqueurs d'annotation comme « {breath} » et « & » qui pourraient gêner la bonne segmentation des données dans Lexico5.

Formes (ordre lexicométrique)	Formes (o...
est	12063
c'	10826
euh	9003
je	7495
tu	7123
de	6725
pas	6651
que	6533
et	6320
ça	6281
ouais	6260
mais	5862
le	5115
il	4721
en	4649
la	4416
un	4082
a	4013

Figure 1 : Dictionnaire des formes et des occurrences, obtenu dans Lexico5 après segmentation du fichier texte regroupant tous les segments

[6] Notice à cette adresse : http://lexi-co.com/ressources/manuel-3.41.pdf

L'outil Lexico5 m'a permis d'avoir une première vue d'ensemble sur le corpus et de définir la liste des mots les plus récurrents comme on peut le voir sur la Figure 1. Il permet aussi de visualiser un concordancier mais non d'appréhender leur contexte linguistique en tenant compte des degrés d'hésitation. C'est pourquoi je me suis servie des langages python et R pour une étude plus en profondeur. Je me suis appuyée sur des scripts python dont certains m'ont été fournis par ma collègue Marie TAHON et je les ai adaptés pour extraire les variables lexicales à partir de l'annotation des segments que j'ai effectuée.

Pour simplifier les futurs traitements, il m'a fallu formater les données autrement : à l'aide d'un script python, les fichiers de transcriptions ont été convertis en fichier tsv[7].

Cela a facilité tous les traitements aussi bien en python qu'avec le logiciel R[8] car j'ai désormais à disposition des tableaux de données que je peux manipuler avec la structure dataframe[9].

Avec un autre script python, j'ai déterminé des caractéristiques des segments annotés me permettant d'effectuer quelques observations afin de déceler des motifs significatifs ou bien de préparer des variables statistiques. Il m'a été conseillé d'utiliser une base de données (Lexique 3.82[10]) disponible en ligne qui répertorie entre autres les classes grammaticales (POS[11]), les phonèmes, le genre et le nombre d'un très grand nombre de mots français. La base Lexique 3.82 m'a servi pour l'appariement au corpus afin d'obtenir un étiquetage en parties du discours. Nous nous sommes basés sur la transcription isolée, les entrées orthographiques, ainsi que sur divers champs de la base comme le genre, le nombre, la fréquence du mot considéré dans les livres et les films afin d'extraire la forme associée à la POS la plus probable.

En première approximation et pour la plupart des requêtes adressées au lexique, lorsqu'une entrée orthographique correspond à plusieurs lignes dans le lexique, étant donné que chaque ligne est associée à une probabilité d'apparition, le mot affecté de la plus grande probabilité est sélectionné. Les variables statistiques présentées dans la section suivante de ce mémoire ont été calculées à ce moment-là mais d'autres caractéristiques ont été extraites comme le nombre d'onomatopées par segment ou le nombre de mots par segment.

Les statistiques tels que les calculs des coefficients de corrélation et la création de graphes (Figure 13, Figure 15 et Figure 17) ont pu être effectuées avec R. Nous pouvons maintenant passer à l'étape de l'analyse des données de notre corpus.

[7] Tab-separated values : fichier dont les valeurs sur une même ligne sont séparées par une tabulation
[8] Langage R permettant de faire des statistiques
[9] Cette structure peut être utilisée avec le module pandas en python et les objets dataframe et tibble en R.
[10] http://www.lexique.org/
[11] Parts Of Speech : parties du discours

IV. Analyses et résultats

Dans cette section, nous allons présenter quelques analyses statistiques obtenues à partir des données issues de notre tâche d'annotation manuelle avec Praat et de prétraitements avec des scripts python. L'objectif est de déterminer des descripteurs acoustiques et linguistiques afin de prédire le degré de certitude ou au contraire d'hésitation dans les segments. Pour rappel, les segments correspondent aux productions des locuteurs délimitées par des pauses silencieuses, segmentées automatiquement et annotées en certitude/hésitation lors de la dernière étape d'annotation manuelle sur dix minutes pour tous les locuteurs (17 sessions d'enregistrement, deux locuteurs considérés par session).

1. Choix des variables lexicales

En m'appuyant sur un fichier regroupant un grand nombre de descripteurs acoustiques et sur une analyse lexicale et morpho-syntaxique du corpus, j'ai pu révéler certains patterns associés à de forts degrés d'hésitation et/ou de certitude. J'ai défini les variables statistiques utilisées comme suit :

Nom de la variable	Définition	Exemples
Ajust	Variable booléenne valant VRAI si le segment contient un mot de classe adjectif, article, préposition ou pronom, suivi d'un mot de la même classe dont le genre ou bien le nombre varie	« Un une villa euh » « des un des des enquêtes tu sais euh »
Eeuh	Variable booléenne valant VRAI si le segment contient un mot se terminant par le son [œ] ou [ø] suivi du mot « euh »	« enfin moi je le euh je le sens enfin »
Nbconj	Nombre maximal de conjonctions qui se succèdent dans le segment	La variable prend la valeur 3 pour « et puis sinon c'est sport »
Repet	Nombre d'apparition d'un mot suivi de lui-même dans le segment	La variable prend la valeur 2 pour « le le le fait d'être enregistré »

Tableau 5 : Récapitulatif des variables statistiques lexicales

Dans un premier temps, j'ai extrait les 20 mots les plus fréquents dans les segments pour les degrés d'hésitation de 1 à 5 (Figure 12). Ensuite, j'ai considéré un ensemble H (Figure 13) constitué des 19 mots[12] repérés dans la liste des 20 mots pour tous les degrés d'hésitation positifs. Il est défini par

$$H := \{\text{euh}, \text{c'est}, \text{hum}, \text{enfin}, \text{que}, \text{ça}, \text{à}, \text{et}, \text{sais}, \text{de}, \text{mais}, \text{je}, \text{j'}, \text{le}, \text{les}, \text{la}, \text{par}, \text{tu}, \text{en}\}.$$

J'ai commencé par m'intéresser à deux variables définies sur l'ensemble des segments qui sont la somme des occurrences des mots de l'ensemble H et la durée du segment.

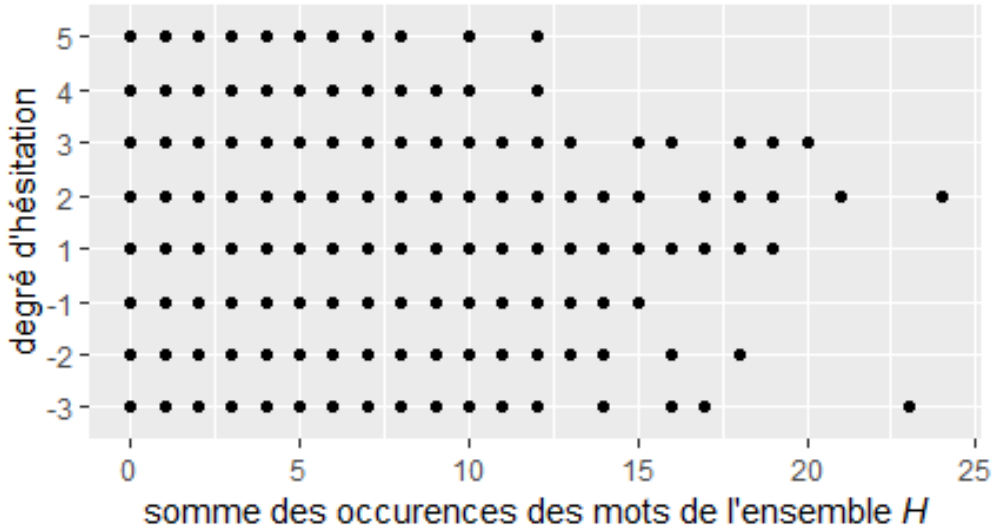


Figure 2 : Nuage de points du degré d'hésitation contre le nombre d'occurrences des mots de l'ensemble H

La corrélation de Spearman entre le degré d'hésitation et cette première variable, le nombre d'occurrences des mots de l'ensemble *H*, vaut environ 0,245 : le degré d'hésitation a donc tendance à augmenter lorsque le nombre d'occurrences considéré croît. Sur la Figure 2, la corrélation entre les deux variables n'apparaît pas clairement mais elle est tout de même significative ($p < 1^{e}\text{-}79$) quoique modérée.

[12] Les mots les plus fréquents ont été choisis tous degrés d'hésitation confondus. Ce sont les 19 mots récurrents retrouvés dans chaque liste.

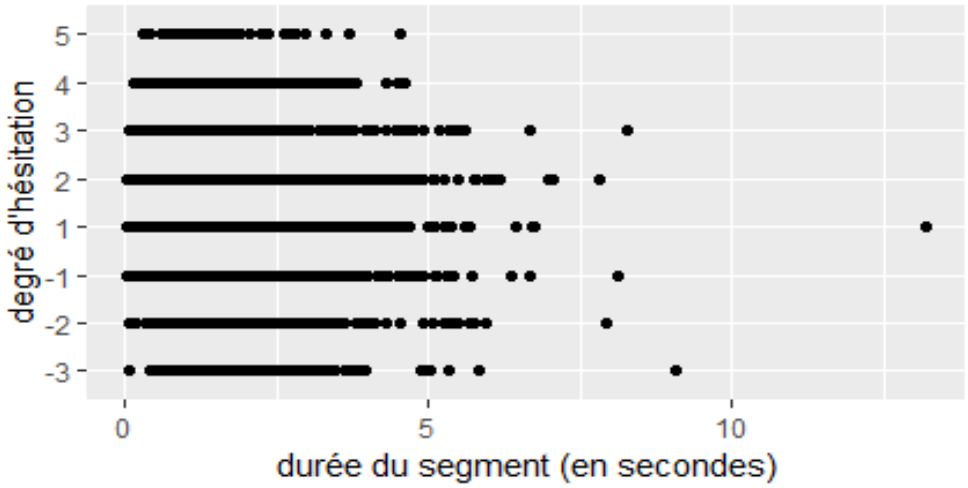


Figure 3 : Nuage de points du degré d'hésitation contre la durée du segment

Sur la Figure 3, la corrélation entre les deux variables, le degré d'hésitation et la durée du segment, n'apparaît pas non plus de façon claire. Il n'y a pas de très forte dépendance entre les degrés d'hésitation et la durée des segments même si la corrélation reste significative ($p<1^{e}$-46). Ce résultat correspond à la corrélation de Spearman entre ces deux variables (approximativement 0,188).

A ce stade, je n'ai pas encore trouvé *a priori* de variables satisfaisantes qui puissent prédire le degré d'hésitation de façon *isolée*. J'ai donc procédé à une analyse du corpus regroupant tous les segments annotés pour y déceler certains motifs qui me semblaient pertinents. C'est à partir de cette analyse que j'ai pu définir les variables lexicales présentées plus haut.

2. Etude des variables lexicales

Pour les variables du Tableau 5, le nombre de valeurs prises par les variables étant trop faible devant le nombre de segments, les nuages de points ne nous ont pas semblé adaptés et ont été remplacés par des boîtes à moustache (par catégorie du prédicteur considéré) superposées aux droites de régression linéaire. Leur couleur renseigne sur le nombre de segments correspondants et la corrélation de Spearman est mentionnée en haut et désignée par rho. Quant à la partie grisée, elle correspond à l'intervalle de confiance gaussien. 5833 segments ont été considérés.

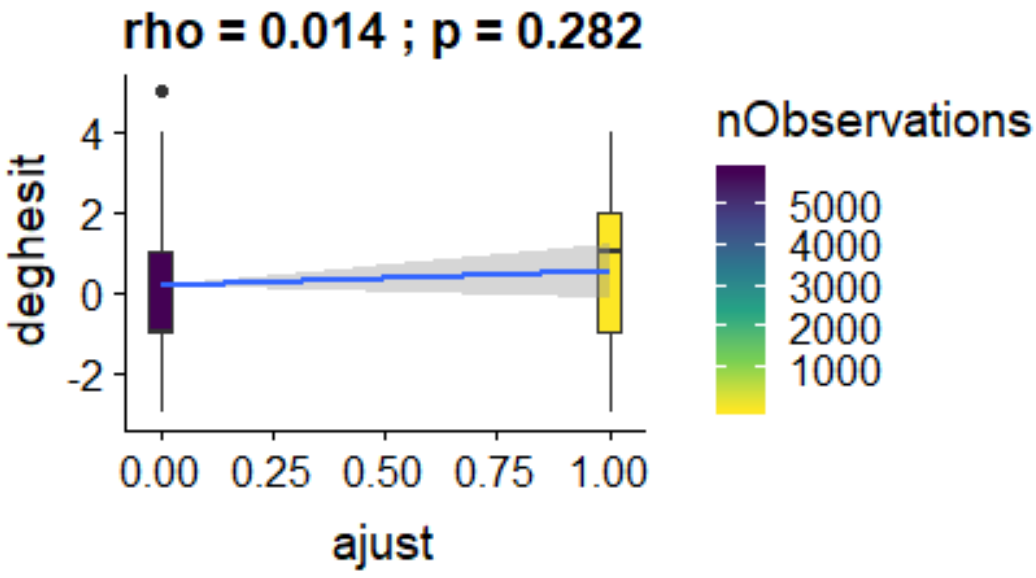


Figure 4 : Degré d'hésitation contre la variable ajust

Sur la Figure 4, nous pouvons observer une très faible dépendance entre le degré d'hésitation et la variable ajust qui se traduit par la rectification en genre et/ou en nombre d'un article par exemple. Comme l'indique la valeur p au-dessus du graphique, la corrélation entre le degré d'hésitation et le phénomène d'ajustement en genre et/ou en nombre reste extrêmement faible et non significative.

Il y a relativement peu de segments pour lesquels il y a un « ajustement » (moins de 500). Qu'il y ait un ajustement ou non, il n'y a pas de fortes influences sur le degré d'hésitation.

Contrairement à nos attentes, Ajust semble donc être un critère insuffisant pour expliquer les degrés d'hésitation associés aux segments. J'en déduis que les locuteurs se corrigent assez rapidement, même lorsqu'il y a production de faux-départs.

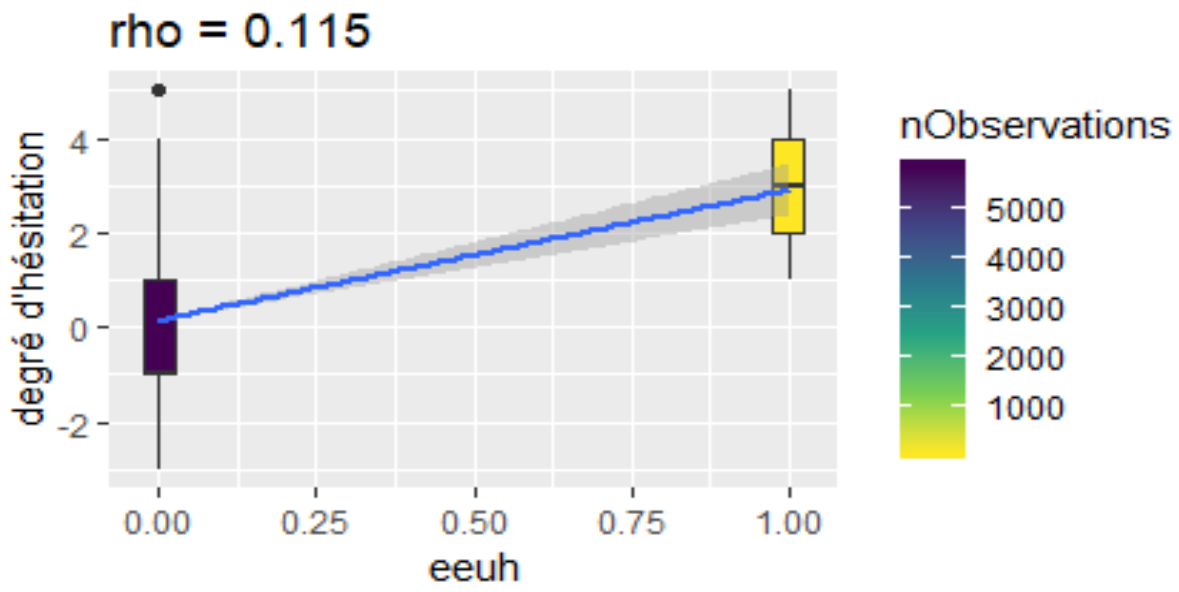


Figure 5 : Degré d'hésitation contre la variable eeuh

La Figure 5 représente le degré d'hésitation contre la variable eeuh. Je me suis inspirée des disfluences telles que les pauses remplies suivies d'un allongement et/ou parfois d'une réparation pour constituer cette variable. J'ai choisi d'étudier ce contexte de voyelle centrale car j'ai pu constater que les locuteurs avaient tendance à prononcer un « euh » après des mots se terminant par les sonorités [ə], [œ] ou [ø]. J'ai notamment observé un tel phénomène pour les sonorités [ɛ] ou [e].

L'interjection « euh » se retrouve fréquemment dans le segment avec un fort degré d'hésitation. En conséquence, il est naturel de penser qu'un locuteur qui hésite va systématiquement prononcer « euh » lorsqu'il cherche ses mots après avoir prononcé un mot.

Quand la variable eeuh vaut « VRAI », nous pouvons observer une forte tendance à être incertain. Cependant si le segment ne contient pas le mot « euh » précédé des sons [ə], [œ] ou [ø], nous observons autant de la certitude que de l'incertitude. Dans les cas où il y a de l'incertitude, l'hésitation est en fait moyenne voire faible (la médiane de la boîte à moustache à gauche est au niveau -1, signe de certitude).

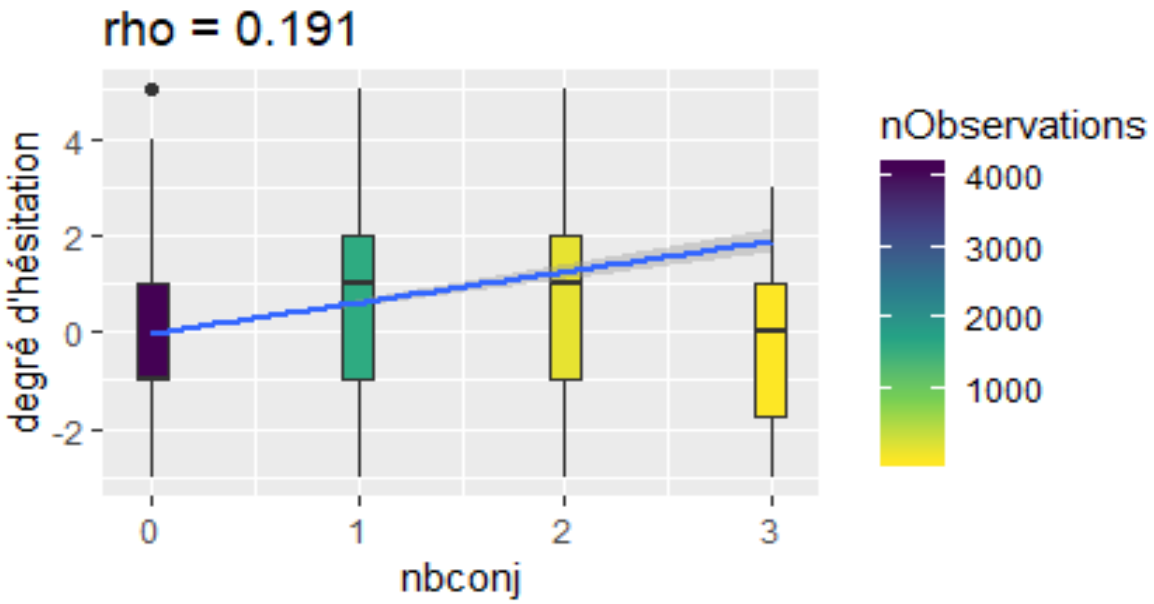


Figure 6 : Degré d'hésitation contre la variable nbconj

Sur la Figure 6, nous observons qu'il y a peu de cas de répétitions de conjonctions. Lorsqu'il n'y a aucune conjonction dans le segment, la certitude et l'hésitation sont quasiment autant probables.

Cependant, cette variable pourrait aider à la prévision du degré d'hésitation : lorsque le nombre maximal de conjonctions successives vaut un ou deux, on retrouve davantage d'hésitation que de certitude.

Mais la capacité de la variable nbconj (nombre de conjonctions successives) à expliquer le degré d'hésitation reste modeste : en effet, les boîtes à moustache présentent une grande étendue ainsi qu'un grand écart interquartile. Différents degrés d'hésitation sont en proportion non négligeables pour chaque valeur de nbconj.

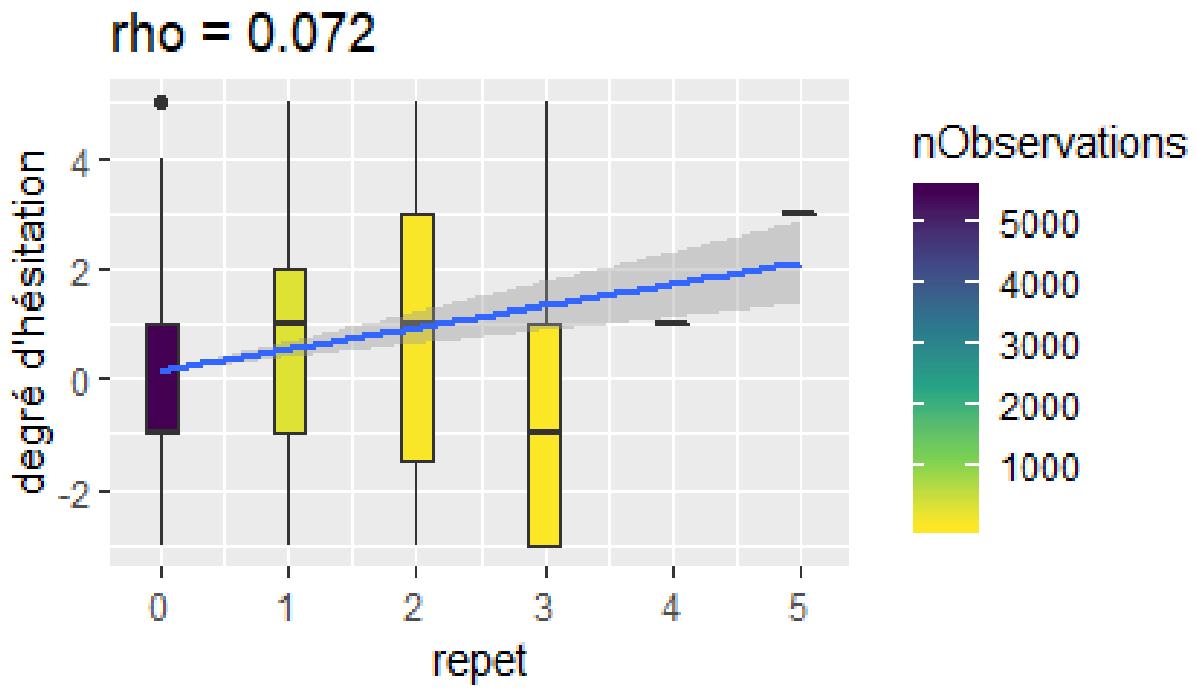


Figure 7 : Degré d'hésitation contre la variable repet

Pour un nombre de répétitions égal ou supérieur à 4 ou 5, nous pouvons observer sur la Figure 7 une incertitude moyenne bien que ces cas soient extrêmement rares (un seul cas pour ces deux valeurs) : l'impact de ces cas sur la corrélation est donc négligeable.

Les cas où le nombre de répétition vaut 2 ou 3 sont peu fréquents et sont associés autant à la certitude qu'à l'incertitude.

Cependant l'essentiel des segments ne contiennent pas de répétitions : la variable repet ne permet pas d'expliquer de manière satisfaisante le degré d'hésitation. La faible hésitation majoritairement associée aux cas légèrement plus nombreux dans lesquels on observe une seule répétition explique toutefois que la corrélation, bien que très faible, soit positive et significative (la valeur p est inférieure à 10^{-7}).

3. Valence, activation et contrôle

Pour les variables suivantes que nous définissons dans le Tableau 6, j'ai récupéré les scores de valence, d'activation et de contrôle que j'avais attribués aux segments de deux minutes annotés manuellement (Figure 14).

Nom de la variable	Définition	Exemple
Activation	Niveau d'excitation du locuteur dans l'intervalle d'entiers [-2, 2]	-2 : locuteur très calme 0 : niveau d'excitation neutre 2 : locuteur très agité
Contrôle	Degré de contrôle du locuteur, prenant les valeurs dans {-2, -1, 1, 2}	-2 : Perte de contrôle de soi 2 : Contrôle parfait de soi, capacité à terminer son propos malgré l'hésitation
Valence	Caractère positif ou négatif du segment, donné par une valeur dans {-2, -1, 0, 1, 2}	-2 : segment à connotation très négative 0 : segment neutre 2 : segment très positif (état émotionnel plaisant, situation agréable…)

Tableau 6 : Variables valence, activation et contrôle

Seules deux minutes sur dix minutes d'annotations par locuteur sont concernées par un renseignement de degré d'hésitation, de la valence, de l'activation et du contrôle, soit un total de 575 segments sur les 5833 auxquels un degré de certitude ou d'hésitation a été attribué.

A noter que certains segments ont été fusionnés, le degré d'hésitation étant une moyenne des degrés d'hésitation des segments initiaux pondérée par leur durée. Cependant, lors de ces fusions, les valeurs de la Valence, de l'Activation et du Contrôle n'ont pas été calculées, ces cas étant trop minoritaires pour présenter un intérêt pour l'analyse finale. Les segments issus de telles fusions ne sont donc pas intégrés aux études statistiques qui suivent.

Nous considérons à partir de maintenant 575 segments.

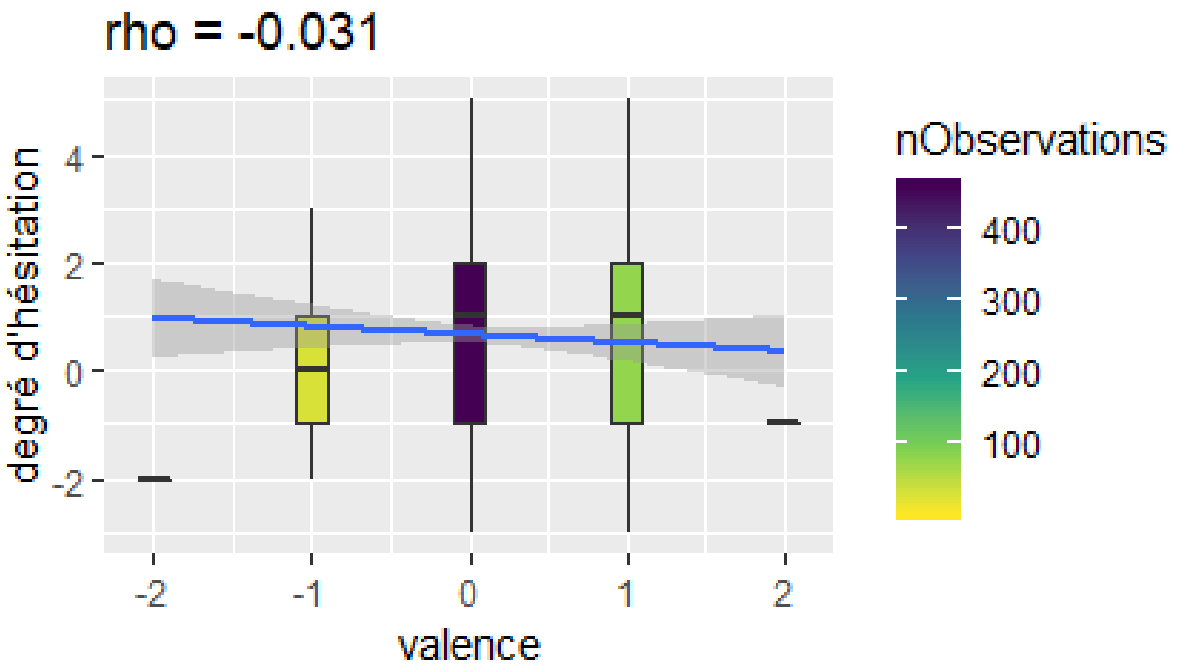


Figure 8 : Degré d'hésitation contre la variable valence

La première chose à signaler sur la Figure 8 est la surreprésentation du niveau neutre qui concerne tous les degrés d'hésitation. Effectivement, la valence a été une des valeurs les plus difficiles à évaluer. Il est bien plus aisé d'évaluer des affects que des disfluences avec ce type de dimension.

Les cas où la valence vaut -2 ou 2 sont très rares et sont associés presque à chaque fois à de la certitude. Le graphe est plutôt symétrique.

Nous pouvons dire qu'il y a très peu de dépendance entre le degré d'hésitation et la valence : les cas où la valence vaut 1 ou -1 présentent une boîte à moustache avec une grande étendue et un grand écart interquartile (donc une dispersion des valeurs très importante).

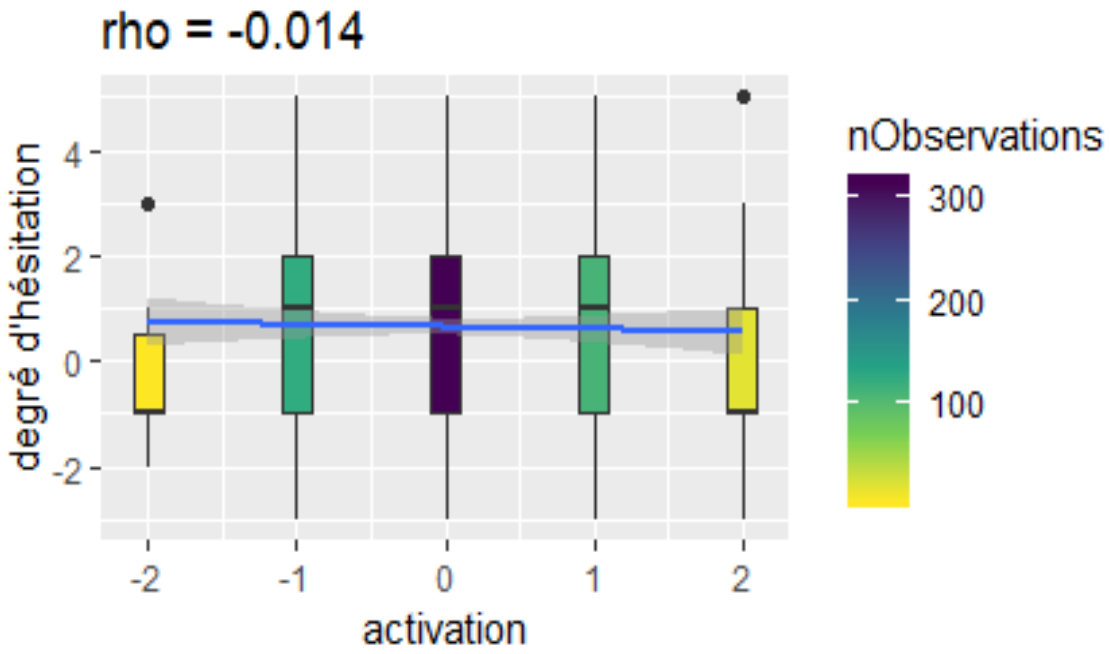


Figure 9 : Degré d'hésitation contre la variable activation

La Figure 9 montre le degré d'hésitation en fonction de l'activation. Etant donné le coefficient de corrélation de Spearman, la dépendance entre ces variables semble faible. Pour une activation qui vaut -1, 0 ou 1, c'est-à-dire les cas les plus répandus, nous retrouvons autant de certitude que d'incertitude.

En revanche, pour une activation qui vaut -2 ou 2, les cas extrêmes, il est très probable de retrouver de la certitude (voir les médianes en bas des boîtes).

L'activation ne permet pas de bien expliquer l'hésitation.

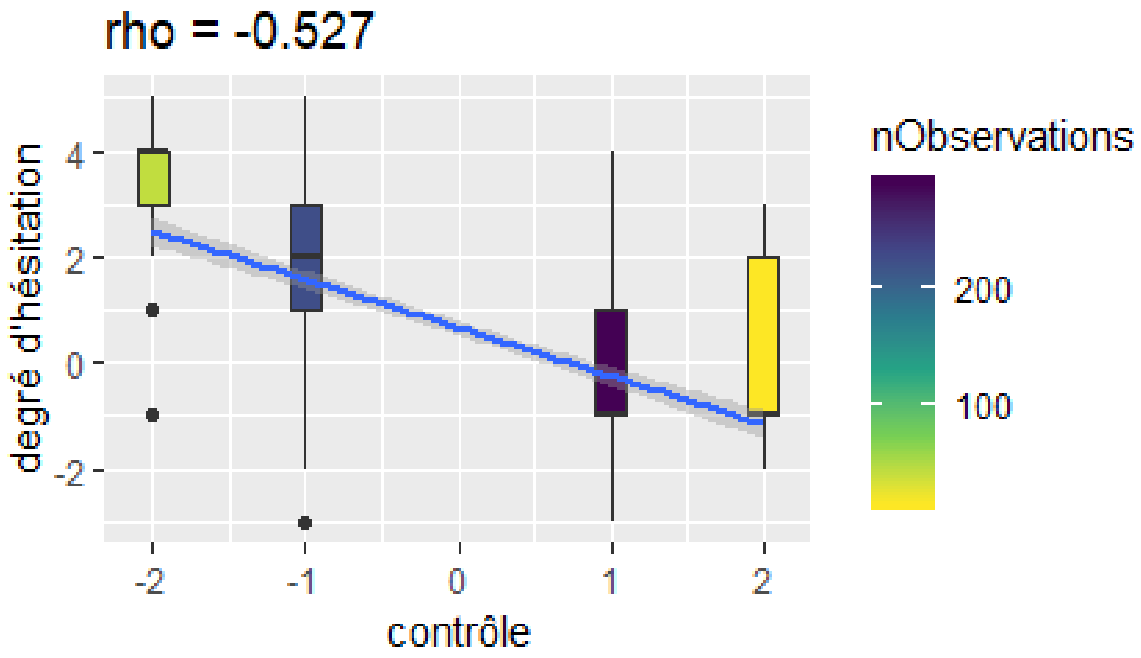


Figure 10 : Degré d'hésitation contre la variable contrôle

Lorsque le contrôle vaut -2, nous avons très probablement une incertitude moyenne. Pour un niveau de contrôle qui vaut 2, cas assez rare, nous retrouvons autant de certitude que d'incertitude (médiane au niveau -1). Lorsque le niveau de contrôle vaut -1 ou 1, nous retrouvons de la certitude et de l'incertitude (voir l'étendue) mais ces deux catégories correspondent respectivement à de l'incertitude moyenne et une faible certitude (écart interquartile plutôt faible devant le nombre d'observations considérées).

La Figure 10 est le graphe où nous observons la plus forte relation de dépendance entre deux variables. Lorsque la valeur de contrôle augmente, le degré d'hésitation a tendance à diminuer.

4. Bilan des analyses

L'une des variables statistiques présentées dans cette section permettant de prédire le mieux l'hésitation est la variable eeuh, qui indique la présence dans le segment d'un mot finissant par une sonorité [ə], [œ] ou [ø] suivi du mot « euh ». La variable contrôle est la variable la plus corrélée au degré d'hésitation, néanmoins le contrôle ne peut pas servir de prédicteur puisqu'il n'est pas possible de l'estimer uniquement grâce à une liste de mots. Le signal acoustique est nécessaire afin d'estimer le contrôle d'une personne sur ses émotions.

Les autres variables statistiques sont très peu corrélées au degré d'hésitation et constituent de moins bons prédicteurs. En ce qui concerne la variable eeuh, les résultats sont à mettre en parallèle avec ceux issus des analyses acoustiques effectuées par nos collègues : les voyelles [œ] ou [ø] sont effectivement les plus importantes dans les segments associés à de l'hésitation.

En ce qui concerne les 5833 segments étudiés ici, nous pouvons voir d'après le Tableau 7 que plus le degré d'hésitation considéré est élevé, plus le nombre de segments se terminant par la sonorité [œ] ou [ø] est grand (Figure 16).

Degré d'hésitation	Nombre de segments	Nombre de segments se terminant par la sonorité [œ] ou [ø]	Ratio (%)
1	1161	88	7, 58
2	700	157	22, 43
3	468	179	38, 25
4	221	108	48, 87
5	51	23	45, 1
Total	2601	555	21, 34

Tableau 7 : Répartition des segments se terminant par la sonorité [œ] ou [ø]

V. Discussions

I. Présentation de quelques résultats de l'apprentissage actif

Nous rappelons à notre lectorat que le protocole de l'apprentissage automatique a été réalisé par nos collègues du LIUM. Par conséquent, pour éviter le plagiat, certains résultats ne seront pas présentés dans cette section mais conservés pour une de leur prochaine publication. Le corpus annoté a été divisé en trois sous-corpus référencés dans le Tableau 8.

Le premier corpus (train) contient les segments annotés entre zéro et deux minutes pour tous les locuteurs à l'exception de deux extraits (28_11_07_1_1 et 04_12_07_1_1). Le second (online) contient les segments annotés restant entre deux et dix minutes pour tous les locuteurs à l'exception des deux extraits précédents (28_11_07_1_1 et 04_12_07_1_1). Enfin, le troisième (test) contient les segments annotés entre zéro et dix minutes des deux locuteurs de test. Cette répartition nous assure de faire des tests indépendamment du locuteur. Quelle que soit la stratégie choisie, le test sera toujours réalisé sur le corpus de test.

	Version 2min	Version 10min
Train (80%)	832	4040
Dev (20%)	208	1010
Train + dev	1040	5050
Online	4010	0
Test	360	360

Tableau 8 : Récapitulatif des sous-corpus Train, Online et Test

Dans une première approche, seuls les coefficients cepstraux (MFCC) et leur dérivées premières (moyenne et écart-type sur l'ensemble du segment) soit 80 descripteurs, sont pris en compte. Les descripteurs sont systématiquement normalisés par rapport au corpus d'apprentissage.

Six modèles de régression ont été testés pour la prédiction d'une valeur de degré d'hésitation : SVR (SVM pour la régression) avec un noyau gaussien (rbf), polynomial (poly) ou linéaire (lin), une régression linéaire, une régression de type Lasso et une de type Ridge. L'ensemble de ces modèles seront évalués avec une erreur quadratique moyenne (RMSE) entre la valeur prédite, et la valeur annotée. Une valeur moyenne (AVG) obtenue pour l'ensemble des modèles sur chaque segment est également fournie.

II. Résultats RMSE

La RMSE (root mean square error) a été appliquée pour mesurer l'erreur entre la prédiction (valeur entre -3 et 6) et la valeur annotée manuellement. Contrairement à nos attentes, l'ajout des huit minutes supplémentaires de données d'apprentissage, n'a pas amélioré la performance des modèles de régression mais la stratégie de faire annoter manuellement les segments qui reçoivent des valeurs très différentes selon les modèles semble apporter des résultats satisfaisants.

Pour le corpus d'apprentissage contenant les segments annotés sur deux minutes par locuteur, le meilleur résultat obtenu est de 3,46 (sur une échelle de -3 à 6) avec le modèle Lasso tandis que pour le corpus d'apprentissage contenant les segments annotés sur dix minutes par locuteur le meilleur résultat obtenu est de 3,47 (sur une échelle de -3 à 6) avec le modèle SVR, soit une légère baisse des performances.

III. Pistes d'amélioration des résultats de notre étude

Les MFCC étant des mesures acoustiques à très court terme, il nous semble raisonnable de considérer que l'intégration de nos descripteurs lexicaux dans les vecteurs de paramètres sont susceptibles d'apporter des informations indépendantes des caractéristiques de l'hésitation déjà capturées par les MFCC et par conséquent d'améliorer les performances des modèles de régression.

Pour la suite du projet, il serait judicieux d'enrichir l'annotation avec les valeurs des dimensions affectives valence, activation et contrôle car nous avons pu constater que le contrôle pouvait constituer un bon prédicteur de l'hésitation dans la parole spontanée. D'ailleurs, il serait intéressant d'exploiter des scores VAC qui seraient cette fois-ci attribués par un système de reconnaissance automatique des émotions pour comparer les résultats.

Etant données quelques difficultés rencontrées lors du stage et la limite de temps, nous n'avons pas pu procéder à une phase d'évaluation du système. C'est pourquoi nous envisageons dans la continuité de ce projet une phase d'annotation interactive qui consisterait à demander à un groupe d'individus d'attribuer des scores d'hésitation et de VAC à des segments associés à des extraits audio afin de vérifier si la variabilité interjuge rend de meilleurs résultats, afin d'évaluer dans quelles mesures d'autres locuteurs tendent à attribuer des scores proches de ceux que j'ai choisis.

Par ailleurs, nous avons l'ambition de développer une interface homme/machine plus conviviale pour orienter le public non expert et lui permettre de naviguer dans un morceau de signal du corpus ou d'afficher des étiquettes selon un code couleur et avec le degré de certitude qui lui est associé.

VI. Conclusion

Nous pouvons maintenant confirmer ou infirmer les hypothèses émises.

Il semblerait qu'il y ait peu de dépendance entre le degré d'hésitation perçu et la durée du segment comportant une disfluence.

Le degré d'hésitation semble augmenter lorsque les locuteurs sont perçus comme ayant un moindre contrôle de la situation mais il faudrait compléter l'annotation de ces scores pour s'en assurer.

Quant au lien entre la présence d'une pause remplie « euh » et le degré d'hésitation, celui-ci est confirmé bien qu'il reste modéré.

Bibliographie

CAMPIONE Estelle & VERONIS, J. &. (2004). *Pauses et hésitations en français spontané.*

CANDEA, M. (2000). Les euh et les allongements dits "d'hésitation": deux phénomènes soumis à certaines contraintes en français oral non lu. *XXIIIèmes Journées d'Etude sur la Parole, Aussois, 19-23 Juin.*

CLAVEL Chloé, V. I. (2006, Octobre). De la construction du corpus émotionnel au système de détection. Le point de vue applicatif de la surveillance dans les lieux publics. *Revue d'Intelligence Artificielle, 20*, 529-551. doi:10.3166/ria.20.529-551

DANON-BOILEAU, M. &. (1998). *Grammaire de l'intonation: l'exemple du français.* Ophrys.

DEVILLERS Laurence, V. L. (2005). Challenges in real life emotion annotation and machine learning based detection. *Neural Networks, 18*(Issue 4), 407-422.

DUEZ. (2019). *Signification des hésitations dans la parole spontanée.* Université de Provence, laboratoire Parole et Langage. Récupéré sur http://www2.lpl-aix.fr/~fulltext/1198.pdf

DUEZ, D. (1999, Mars). La fonction symbolique des pauses dans la parole de l'homme politique. *Faits de Langues*(13), 91-97. doi:https://doi.org/10.3406/flang.1999.1242

GOLDMAN-EISLER. (1968). *Psycholinguistics: experiments in spontaneous speech, Academic.*

GROSJEAN F, D. A. (1972-73). Analyse des variables temporelles du français spontané. *Phonetica, 26*(130-156 & 28), 191-226.

GUAÏTELLA. (1991). Hésitations vocales en parole spontanée: réalisations acoustiques et fonctions rythmiques. *Travaux de l'Institut de Phonétique d'Aix, 14*, 113-130.

HACINE-GHARBI, A. (2012). *Sélection de paramètres acoustiques pertinents pour la reconnaissance de la parole.* Université d'Orléans. Récupéré sur https://tel.archives-ouvertes.fr/tel-00843652/document

MACLAY, H. &. (1959, Janvier). Hesitation Phenomena in Spontaneous English Speech. *Word, 15*, 19-44. doi:10.1080/00437956.1959.11659682

MEYERS, C. F. (1985). *Interruptions as a Variable in Stuttering and Disfluency.*

PALLAUD, B., RAUZY, S., & BLACHE, P. (2013). Auto-interruptions et disfluences en français parlé dans quatre corpus du CID. *Travaux Interdisciplinaires sur la Parole et le Langage (TIPA)*. doi:10.4000/tipa.995

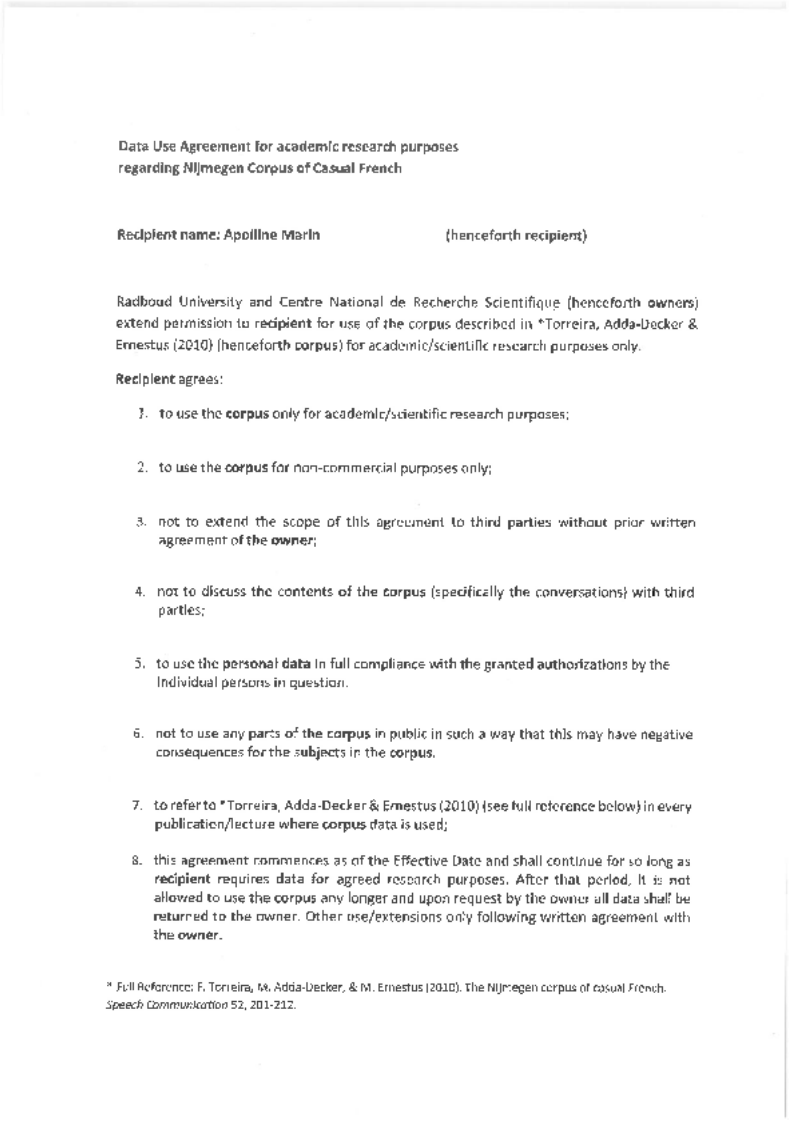
QUADER Raheel, L. G. (Oct 2018). Disfluency Insertion for Spontaneous TTS: Formalization and Proof of Concept. *6th International Conference on Statistical Language and Speech Processing*, (pp. pp.1-12). Mons, Belgium. Récupéré sur https://hal.inria.fr/hal-01840798/document

TAHON, M. L. (2018). *Can we Generate Emotional Pronunciations for Expressive Speech Synthesis?* Récupéré sur https://hal.archives-ouvertes.fr/hal-01802463/document

Table des figures

Liste des tableaux

Annexes



Data Use Agreement for academic research purposes
regarding Nijmegen Corpus of Casual French

Recipient name: Apolline Marin (henceforth **recipient**)

Radboud University and Centre National de Recherche Scientifique (henceforth **owners**) extend permission to **recipient** for use of the corpus described in *Torreira, Adda-Decker & Ernestus (2010) (henceforth **corpus**) for academic/scientific research purposes only.

Recipient agrees:

1. to use the **corpus** only for academic/scientific research purposes;
2. to use the **corpus** for non-commercial purposes only;
3. not to extend the scope of this agreement to third **parties** without prior written agreement of the **owner**;
4. not to discuss the contents of the **corpus** (specifically the conversations) with third parties;
5. to use the **personal data** in full compliance with the granted **authorizations** by the individual persons in question.
6. not to use any parts of the **corpus** in public in such a way that this may have negative consequences for the **subjects** in the **corpus**.
7. to refer to *Torreira, Adda-Decker & Ernestus (2010) (see full reference below) in every publication/lecture where **corpus** data is used;
8. this agreement commences as of the Effective Date and shall continue for so long as **recipient** requires data for agreed research purposes. After that period, it is not allowed to use the **corpus** any longer and upon request by the owner all data shall be returned to the owner. Other use/extensions only following written agreement with the **owner**.

* Full Reference: F. Torreira, M. Adda-Decker, & M. Ernestus (2010). The Nijmegen corpus of casual French. *Speech Communication* 52, 201-212.

Figure 11 : Contrat permettant l'accès légal au corpus NCCFr

Figure 12 : Script extract_frequenciesV2.py (adaptation du script de Marie TAHON)

```
import argparse
import sys
import os
import numpy as np
import textgrid
from decimal import *

def get_content_from_interval(tdeb, tfin, intervalTier, eps):
    kdeb = 0
    while (kdeb < len(intervalTier)):
        if (intervalTier[kdeb].minTime > tdeb - eps):
            break
        else:
            kdeb += 1
    kfin = kdeb
    while (kfin < len(intervalTier)):
        if (intervalTier[kfin].maxTime > tfin + eps):
            break
        else:
            kfin += 1
    #print(tdeb, tfin, kdeb, kfin)
    text = ''
    for k in range(kdeb, kfin):
        text += ' ' + str(intervalTier[k].mark.encode('utf-8'))
    return text


def extract_write_words_phones(work_dir, filename, eps):
    f = open(filename, 'w')

    for gridname in os.listdir(work_dir): #on parcourt le repertoire (on considere
qu'il ne contient que des textgrid)
        name, ext = os.path.splitext(gridname)
        print(name)
        grid_emo = textgrid.TextGrid.fromFile(work_dir + gridname)#gridemo est un
object qui contient tout le textgrid
        intervalTier_hesit = grid_emo[5] #donne le 6eme Tier qui normalement
correspond au degre d'incertitude cale sur les segments obtenus automatiquement
eventuellement corrige par Appoline.
        intervalTier_word = grid_emo[0] #donne le 1er Tier qui normalement contient
les mots
        intervalTier_phone = grid_emo[1] #donne le 2em Tier qui normalement
contient les phonemes
        ideb_w = np.array([i.minTime for i in intervalTier_word],dtype=float)
        ifin_w = np.array([i.maxTime for i in intervalTier_word], dtype=float)
        for k, i in enumerate(intervalTier_hesit): #boucle sur tous les intervalles
du Tier hesitation
            ideb = i.minTime #temps de debut de l'intervalle
            ifin = i.maxTime #temps de fin de l'intervalle
            list_words = get_content_from_interval(ideb, ifin, intervalTier_word,
eps)
            list_phones = get_content_from_interval(ideb, ifin, intervalTier_phone,
eps)
            f.write('{0}\t{1}\t{2}\t{3}\t{4}\t{5}\n'.format(name, ideb, ifin,
i.mark, list_words, list_phones))

    f.close()
    return True

def extract_item_frequencies(filename, degre):
    freq_w = {}
```

```
    with open(filename, 'r') as f:
        for line in f:
            tab = line.strip().split('\t')#on recupere la ligne sous forme de
tableau
            if len(tab) < 5:
                tab.append('None')
                tab.append('<eps>')
                tab.append('sil')
            if (tab[3] == 'silence') | (tab[3] == '') | (tab[3] == '<eps>'): #ce
sont les segments pour lesquels il n'y a pas d'annotation hesitation
                continue
            if '.' in tab[3]:
                hesit = float(tab[3])
            else:
                hesit = int(tab[3])
            if hesit != degre:#il faut modifier ce degre pour pouvoir calculer les
frequences sur les mots d'un degre donne
                continue
            text = tab[4].split(' ')#on recupere les mots separes par des espaces
            for w in text:
                if w in freq_w:
                    freq_w[w] += 1#si le mot est deja dans le dictionnaire on lui
ajoute une apparition
                else:
                    freq_w[w] = 1#sinon on cree une entree dans le dictionnaire
initialisee a 1

    freq_w_sorted = sorted(freq_w.items(), key=lambda kv: kv[1], reverse=True)
    #print 20 most frequent words
    f_tot = 0
    for i, w in enumerate(freq_w_sorted):
        #print(w)
        if i < 20:
            print('mot {0}: {1}'.format(w[0], w[1]))
        f_tot += w[1]
    print('nb total de mot pour le degre {0}:{1}'.format(degre, f_tot))

    return True

def main():
    parser = argparse.ArgumentParser(description='Write file with words and phones
per segments')
    parser.add_argument('work_dir', type=str, help='working directory where are
located TextGrid files')
    parser.add_argument('cues_name', type=str, help='names of the file where are /
will be linguistic and phonetic cues')
    parser.add_argument('--time_ths', type=float, default=0.01, help='time
threshold for time ntervals overlap')
    parser.add_argument('phase', type=int, help='phase: 1 extract features, 2
compute frequencies')
    args = parser.parse_args()

    degre = 5
    print(args.time_ths, type(args.time_ths))
    if args.phase == 1:
        extract_write_words_phones(args.work_dir, args.cues_name,
Decimal(args.time_ths))
        print('extraction done')
    elif args.phase == 2:
        extract_item_frequencies(args.cues_name, degre)

if __name__ == "__main__":
    main()
```

Deux fonctions : extraction de tous les segments annotés avec des scores de degrés d'hésitation et de VAC et création de la liste des 20 premières occurrences des mots porteurs d'hésitation

Figure 13 : Script stats.py (calcul de la corrélation de Spearman et nuage de points)

```
###################
#variables
###################

f = open("result")
#variable Y : degré d'hésitation
Y = [0 for i in range(nb)]
#variable X1
X1 = [0 for i in range(nb)]
#variable X2
X2 = [0 for i in range(nb)]

hesit = {'euh', "b\"c'\" b'est'", 'hum', 'enfin', 'que', '\xc3\xa0', '\xc3\xa7a',
'et', 'sais', 'de', 'mais', 'je', 'j\'', 'le', 'les', 'la', 'pas', 'tu', 'en'}
def calc_X1(words):
    """
    nombre d'occurences des mots de l'ensemble hesit dans words (liste de mots du
segment)
    """
    nb = 0
    for w in hesit:
        nb += words.count(w)
    return nb

line = f.readline()
i = 0
while line!='':
    res = re.findall("[^\t]+\t[^\t]+\t[^\t]+\t([^\t]*)[ |\t]", line)
    #calcul de Y
    deg = float(res[0])
    if deg>=0:
        Y[i] = np.ceil(float(res[0]))
    else:
        Y[i] = np.floor(float(res[0]))
    #détermination de X1
    res = re.findall("[^\t]+\t[^\t]+\t[^\t]+\t[^\t]+\t([^\t]*)\t[^\n]+\n",
line)#liste des mots du segment
    X1[i] = calc_X1(res[0])

    #X2
    res = re.findall("[^\t]+\t([^\t]+)\t([^\t]+)\t", line)
    X2[i] = float(res[0][1]) - float(res[0][0])
    line = f.readline()
    i += 1

f.close()

mess = input("Do you want tot create a file for Y, X1...? (y/n)\n")
if mess=='y':
    f = open("vars", mode="w")
    f.write("Y X1 X2\n")
    for i in range(nb):
        f.write(str(Y[i])+" "+str(X1[i])+" "+str(X2[i])+"\n")
    f.close()
    print("file created")
```

Figure 14 : Script extract_frequenciesVAC.py (adaptation du script de Marie TAHON pour extraire les valeurs VAC des segments annotés et calculer le coefficient de Spearman)

```
import argparse
import sys
import os
import numpy as np
import textgrid
from decimal import *

def get_content_from_interval(tdeb, tfin, intervalTier, eps):
    kdeb = 0
    while (kdeb < len(intervalTier)):
        if (intervalTier[kdeb].minTime > tdeb - eps):
            break
        else:
            kdeb += 1
    kfin = kdeb
    while (kfin < len(intervalTier)):
        if (intervalTier[kfin].maxTime > tfin + eps):
            break
        else:
            kfin += 1
    #print(tdeb, tfin, kdeb, kfin)
    text = ''
    for k in range(kdeb, kfin):
        text += ' ' + intervalTier[k].mark
    return text


def extract_write_words_phones(work_dir, filename, eps):
    f = open(filename, 'w')
    f.write("file\tbeg\tend\tdeghesit\tV\tA\tC\twords\tphones\n")
    for gridname in os.listdir(work_dir): #on parcourt le repertoire (on considere
qu'il ne contient que des textgrid)
        name, ext = os.path.splitext(gridname)
        print(name)
        grid_emo = textgrid.TextGrid.fromFile(work_dir + gridname)#gridemo est un
object qui contient tout le textgrid
        intervalTier_hesit = grid_emo[5] #donne le 6eme Tier qui normalement
correspond au degre d'incertitude cale sur les segments obtenus automatiquement
eventuellement corrige par Appoline.
        intervalTier_VAC = grid_emo[4]
        intervalTier_word = grid_emo[0] #donne le 1er Tier qui normalement contient
les mots
        intervalTier_phone = grid_emo[1] #donne le 2em Tier qui normalement
contient les phonemes
        ideb_w = np.array([i.minTime for i in intervalTier_word],dtype=float)
        ifin_w = np.array([i.maxTime for i in intervalTier_word], dtype=float)

        for k, i in enumerate(intervalTier_VAC): #boucle sur tous les intervalles
du Tier hesitation
            if i.mark=='' or i.mark=="silence" or i.mark=="INTERV_MULTIPLES":
                continue
            ideb = i.minTime #temps de debut de l'intervalle
            ifin = i.maxTime #temps de fin de l'intervalle
            list_words = get_content_from_interval(ideb, ifin, intervalTier_word,
eps)
            list_phones = get_content_from_interval(ideb, ifin, intervalTier_phone,
eps)
            deghesit = get_content_from_interval(ideb, ifin, intervalTier_hesit,
eps)
            f.write('{0}\t{1}\t{2}\t{3}\t{4}\t{5}\t{6}\n'.format(name, ideb, ifin,
deghesit.replace(" ", ""), i.mark.replace(", ", "\t"), list_words, list_phones))

    f.close()
    return True
```

```
def extract_item_frequencies(filename, degre):
    freq_w = {}
    with open(filename, 'r') as f:
        for line in f:
            tab = line.strip().split('\t')#on recupere la ligne sous forme de
tableau
            if len(tab) < 5:
                tab.append('None')
                tab.append('<eps>')
                tab.append('sil')
            if (tab[3] == 'silence') | (tab[3] == '') | (tab[3] == '<eps>'): #ce
sont les segments pour lesquels il n'y a pas d'annotation hesitation
                continue
            if '.' in tab[3]:
                hesit = float(tab[3])
            else:
                hesit = int(tab[3])
            if hesit != degre:#il faut modifier ce degre pour pouvoir calculer les
frequences sur les mots d'un degre donne
                continue
            text = tab[4].split(' ')#on recupere les mots separes par des espaces
            for w in text:
                if w in freq_w:
                    freq_w[w] += 1#si le mot est deja dans le dictionnaire on lui
ajoute une apparition
                else:
                    freq_w[w] = 1#sinon on cree une entree dans le dictionnaire
initialisee a 1

    freq_w_sorted = sorted(freq_w.items(), key=lambda kv: kv[1], reverse=True)
    #print 20 most frequent words
    f_tot = 0
    for i, w in enumerate(freq_w_sorted):
        #print(w)
        if i < 20:
            print('mot {0}: {1}'.format(w[0], w[1]))
        f_tot += w[1]
    print('nb total de mot pour le degre {0}:{1}'.format(degre, f_tot))

    return True

def main():
    parser = argparse.ArgumentParser(description='Write file with words and phones
per segments')
    parser.add_argument('work_dir', type=str, help='working directory where are
located TextGrid files')
    parser.add_argument('cues_name', type=str, help='names of the file where are /
will be linguistic and phonetic cues')
    parser.add_argument('--time_ths', type=float, default=0.01, help='time
threshold for time ntervals overlap')
    parser.add_argument('phase', type=int, help='phase: 1 extract features, 2
compute frequencies')
    args = parser.parse_args()

    degre = -1
    print(args.time_ths, type(args.time_ths))
    if args.phase == 1:
        extract_write_words_phones(args.work_dir, args.cues_name,
Decimal(args.time_ths))
        print('extraction done')
    elif args.phase == 2:
        extract_item_frequencies(args.cues_name, degre)

if __name__ == "__main__":
    main()
```

Figure 15 : Script lexi.py

```
def get_nbconj(cgram_list):
    nbmax = 0
    i = 0
    while i<len(cgram_list):
        nb = 0
        while i<len(cgram_list) and cgram_list[i]!='CON':
            i = i+1
        while i<len(cgram_list) and cgram_list[i]=='CON':
            nb += 1
            i += 1
        if nb>nbmax:
            nbmax = nb
    return nbmax

def get_ajust(seg):
    for i in range(len(seg)-1):
        dat = np.asarray(lexi[lexi.ortho==seg[i]]['freqlivres'])
        if dat.shape[0]==0:
            continue
        ind = np.argmax(dat)#mot le plus probable
        cgram1 = lexi[lexi.ortho==seg[i]]['cgram'].iloc[ind]
        if cgram1 in {'VER', 'ONO', 'LIA', 'CON', 'AUX', 'ADV', 'NOM'}:
            continue
        #répétition de la classe grammaticale?
        aux = np.asarray(lexi[lexi.ortho==seg[i+1]]['cgram'])
        if not cgram1 in aux:
            continue
        ind2 = np.nonzero(aux==cgram1)[0][0]
        #changement de genre/nombre?
        genre1 = lexi[lexi.ortho==seg[i]]['genre'].iloc[ind]
        nombre1 = lexi[lexi.ortho==seg[i]]['nombre'].iloc[ind]
        genre2 = lexi[lexi.ortho==seg[i+1]]['genre'].iloc[ind2]
        nombre2 = lexi[lexi.ortho==seg[i+1]]['nombre'].iloc[ind2]
        if ({"f", "m"}=={genre1, genre2} and nombre1==nombre2) or ({"s",
"p"}=={nombre1, nombre2} and genre1==genre2):
            return True
    return False

def get_eeuh(seg):
    for i in range(len(seg)-1):
        aux = lexi[lexi.ortho==seg[i]]['phon']
        if aux.shape[0]==0:
            continue
        if seg[i+1]=='euh' and aux.iloc[0][-1] in {'2', '9'}:
            return True
    return False

def get_cgram(seg):
    cgram = ""
    for w in seg:
        if w[-1]=="'":#c'...
            w = w[:-1]
        if w[0] in {'&', '-'}:#&hm, &ouais, -là...
            w = w[1:]

        dat = np.asarray(lexi[lexi.ortho==w]['freqlivres'])

        if dat.size==0:
            cgram += "- "#pas d'entrée dans le lexique
        else:
            ind = np.argmax(dat)#mot le plus probable
            cgram += lexi[lexi.ortho==w]['cgram'].iloc[ind] + " "
    return cgram
```

```
def get_nbmots(seg):
    return len(seg)

def get_ono(seg):
    nb = 0
    for w in seg:
        if 'ONO' in np.asarray(lexi[lexi.ortho==w]['cgram']):
            nb += 1
    return nb


nbseg = len(segments[0])#nombre de segments

#pour extraire les mots des segments
regexp1 = re.compile(" b[\'](.*?)\'")
regexp2 = re.compile(" b[\"](.*?)\"")
for i in range(nbseg):
    #recodage
    seg = [w for w in " ".join(regexp2.split("
".join(regexp1.split(str(segments.iloc[i][4]).encode("latin1").decode()))))).split("
") if w != ""]

    info['mots'].append(" ".join(seg))
    info['cgram'].append(get_cgram(seg))
    info['ono'].append(get_ono(seg))
    info['nbmots'].append(get_nbmots(seg))
    info['eeuh'].append(get_eeuh(seg))
    info['ajust'].append(get_ajust(seg))
    info['repet'].append(get_repet(seg))
    info['nbconj'].append(get_nbconj(info['cgram'][-1].split(" ")))
    info['fichier'].append(segments.iloc[i][0])
    info['tdeb'].append(segments.iloc[i][1])
    info['tfin'].append(segments.iloc[i][2])
    deg = segments.iloc[i][3]
    if deg > 0:
        info['deghesit'].append(np.ceil(deg))
    else:
        info['deghesit'].append(np.floor(deg))
#print(segments.iloc[3][4].split("'")[5].encode("latin1").decode())#.encode("raw_un
icode_escape"))#.decode('unicode-escape'))#.encode("latin1").decode())


#écriture du résultat + graphes
df = pd.DataFrame(info)
df['ajust&eeuh']=df['ajust']&df['eeuh']

df['ajust&eeuh']=df['ajust&eeuh'].astype('int8')
df['ajust']=df['ajust'].astype('int8')
df['eeuh']=df['eeuh'].astype('int8')
df.plot.scatter('ajust&eeuh', 'deghesit')
plt.show()
df.plot.scatter('ajust', 'deghesit')
plt.show()
df.plot.scatter('eeuh', 'deghesit')
plt.show()

df.plot.scatter('nbconj', 'deghesit')
plt.show()
df.plot.scatter('repet', 'deghesit')
plt.show()
df.to_csv("result.tsv", sep="\t")
```

Figure 16 : Script lexi2.py (calcul de la répartition des segments se terminant par [œ] ou [ø])

```
# -*- coding: utf-8 -*-
import pandas as pd
import re
import numpy as np
import matplotlib.pyplot as plt



segments = pd.read_csv("../../../Stats/result", sep='\t', header=None,
encoding="unicode_escape")
lexi = pd.read_csv("Lexique382.tsv", sep='\t', encoding="utf8")


tab=np.zeros((5, 2))#occ du degré, occ de la sonorité e
nbseg = len(segments[0])#nombre de segments

#pour extraire les mots des segments
regexp1 = re.compile(" b[\'][^\']*)\'")
regexp2 = re.compile(" b[\"][^\"]*)\"")
for i in range(nbseg):
    #recodage
    seg = [w for w in " ".join(regexp2.split("
".join(regexp1.split(str(segments.iloc[i][4]).encode("latin1").decode())))).split("
") if w != ""]
    if seg==[]:
        continue

    deg = segments.iloc[i][3]
    deg = np.int32(np.ceil(deg))

    if deg > 5:
        continue
    if deg>0:
        tab[deg-1][0] += 1
    else:
        continue

    dat = np.asarray(lexi[lexi.ortho==seg[-1]]['phon'])
    if dat.shape[0]==0:
        continue
    if dat[0] in {"2", "9"}:
        tab[deg-1][1] += 1

print(tab)
```

Figure 17 : Exemple d'un script python utilisé dans R pour réaliser les boîtes à moustaches des variables lexicales, adapté d'après un script de Nicolas AUDIBERT

```
library(tidyverse)
library(cowplot)

donnees <- read_tsv("result.tsv")

variableRef = "deghesit"
variablesLex = c("eeuh","ajust", "repet", "nbconj")
largeur_relative_boxplots = .05
donnees<-filter(donnees, deghesit<=5)
afficher_p_value = TRUE

# boite à moustaches (boxplot) pour représenter la distribution du degré
d'hésitation en fonction des valeurs de chacun des prédicteurs
for(variableLex in variablesLex) {
  # calcul de la corrélation de Spearman avec le degré d'hésitation, ainsi que son
degré de significativité
  # on utilise suppressWarnings pour ne pas afficher à chaque fois le message
d'avertissement sur l'approximation effectuée pour le calcul de la p-value
  corrTestSpearman = suppressWarnings(cor.test(donnees %>% pull(variableLex),
donnees %>% pull(variableRef), method = "spearman"))
  corrSpearman = corrTestSpearman$estimate
  corrSpearmanSignificativite = corrTestSpearman$p.value

  # calcul du nombre d'observations par valeur et ajout au jeu de données
  tailleGroupes = donnees %>%
    group_by_at(variableLex) %>%
    summarise(nObservations = n())
  donnees = donnees %>%
    left_join(tailleGroupes, by = variableLex)

  titre_figure = paste0("rho = ", round(corrSpearman, digits = 3))
  if(afficher_p_value) {
    if(corrSpearmanSignificativite>=.001)
      texte_p_value = paste0("p = ", round(corrSpearmanSignificativite, digits =
3))
    else
      texte_p_value = paste0("p < 1e", ceiling(log10(corrSpearmanSignificativite)))
    titre_figure = paste0(titre_figure, " ; ", texte_p_value)
  }

  # tracé de la figure
  fig <- ggplot(donnees) +
    # 1e partie de la figure : boxplots avec couleur de remplissage qui dépend du
nombre d'observations
    geom_boxplot(aes_string(x = variableLex, y = variableRef, group = variableLex,
fill = "nObservations"), width = diff(range(donnees %>%
pull(variableLex)))*largeur_relative_boxplots) +
    scale_fill_viridis_c(direction = -1) +
    # 2e partie de la figure : droite de régression
    geom_smooth(aes_string(x = variableLex, y = variableRef), method = "lm") +
    ggtitle(titre_figure) # affichage de la corrélation dans le titre
  print(fig)

  # export en fichier image
  fichierFig <- paste0("boxplot_reg_", variableLex, ".png")
  ggsave(fig, filename = fichierFig, width = 20, height = 14, units = "cm")

  # on retire la colonne nObservations
  donnees = donnees %>%
    select(-nObservations)
```