DOCTORAT AIX-MARSEILLE UNIVERSITÉ
délivré par
Université Aix-Marseille I
U.F.R. LACS
ÉCOLE DOCTORALE COGNITION, LANGAGE, EDUCATION E.D. 356

N° attribué par la bibliothèque
| | | | | | | | | | |

THÈSE

pour obtenir le grade de
DOCTEUR D'AIX-MARSEILLE UNIVERSITÉ
Formation doctorale : Langage et Parole
Spécialité : Phonétique

Présentée et soutenue publiquement
par

Vincent AUBANEL

le 21 janvier 2011

Variation phonologique régionale en interaction conversationnelle

Directeur de Thèse :
M. Noël NGUYEN

Jury :

Mme Martine ADDA-DECKER	CNRS/Université Paris 3	Rapporteur
M. Gérard BAILLY	CNRS/Université Stendhal	Rapporteur
Mme Véronique DELVAUX	FNRS/Université de Mons-Hainaut	Examinateur
Mme Christine MEUNIER	CNRS/Université Aix-Marseille I	Examinateur
M. Noël NGUYEN	CNRS/Université Aix-Marseille I	Directeur de Thèse

Remerciements

Je voudrais ici exprimer toute ma gratitude aux personnes qui, à un moment ou à un autre, m'ont soutenu et accompagné dans ce long cheminement que représente une thèse.

Ce travail n'aurait ainsi jamais vu le jour sans le soutien toujours renouvelé et la confiance sans faille de mon directeur de thèse, Noël Nguyen. Au delà de l'encadrement scientifique tout simplement idéal qu'a constitué son conseil, son ouverture d'esprit, sa disponibilité, son honnêteté sauront encore m'inspirer bien après la fin de cette aventure commune.

Je remercie vivement Martine Adda-Decker et Gérard Bailly d'avoir accepté de rapporter ce travail, ainsi que Véronique Delvaux et Christine Meunier d'avoir bien voulu être membres de mon jury. Par ailleurs, mes discussions avec chacun d'entre eux m'auront bien aidé.

Je remercie la région Provence-Alpes-Cöte d'Azur pour m'avoir accordé une bourse doctorale, et toute l'équipe de France Bleu Provence à Aix-en-Provence qui m'a accueilli dans ses locaux et donné à voir une autre perspective pour s'intéresser aux accents régionaux.

J'ai eu la chance d'évoluer dans un contexte stimulant et chaleureux. Ainsi je suis reconnaissant au LPL pour son accueil, en particulier à Philippe, et à Simone, Armelle, Rachida et Nadia pour leur efficacité et leur disponibilité, ainsi qu'au support amélioré d'Alain, Thierry, les deux compères Cyril et Seb, Loundou, Carine et Bernard Bel. Merci également aux gardiennes de la connaissance, massicots et agrafeuses Isabelle, Joëlle et Claudia pour leur efficacité et les agréables moments passés ensemble.

Il a fait bon partager repas, discussions et plages avec Sophie, Marion, Roxane et Yohann, trajets avec Christine et Cheryl, l'ombre des néfliers avec Daniel, Christian, Bernard Teston et Peter, explorations primatologiques avec Berthille et planchas avec tous. Merci à Cristel pour sa présence, à Robert pour sa disponibilité et son aide significative, à Muriel pour son aide organisationnelle, Médéric pour sa connaissance universelle de la région marseillaise.

Je remercie Guillaume Gravier et Jean-Philippe Goldman pour m'avoir gracieusement laissé utiliser leurs modèles acoustiques et pour leurs conseils avisés.

Je remercie également tous les étudiants du Lycée Thiers à Marseille qui ont participé à mon étude, ainsi qu'à l'équipe pédagogique, M. Chalumeau en tête, qui m'a accueillie dans l'établissement et facilité la conduite des enregistrements.

Je remercie tout particulièrement Olivier, Pauline et Stéphane pour leur soutien continu et inestimable tout au long des derniers moments de la pré-

paration du manuscrit.

Un grand merci à mes amis Katia, Céline, Manu, Sil, Kizzi, Fida et aux doctorants et post-doctorants du LPL, en particulier Pauline, Haydée, Amandine, Anne, Cécile, Francesco, Mathilde, Yohana, Angèle, James et Elisa pour leur soutien, leur compréhension et leur amitié précieuse.

Merci à mes parents pour leur dévouement, à Lucia, à mes frères, à ma famille de Marseille et d'ailleurs.

Merci Daria pour ton amour.

Je remercie enfin très chaleureusement Martin Cooke et Maria Luisa Lecumberri ainsi que l'équipe du laboratoire Laslab pour m'avoir accueilli et donné cette magnifique nouvelle impulsion pour la suite de mes recherches.

Résumé

C'est dans l'interaction sociale, lieu d'occurrence premier du langage parlé (Local, 2003) que la parole est apprise, qu'elle est produite quotidiennement et qu'elle évolue. De nouvelles approches interdisciplinaires de l'étude de la parole, notamment la sociophonétique ou les récents développements de l'interaction conversationnelle, ouvrent de nouvelles perspectives dans la modélisation du traitement de la parole. Une question centrale à cette entreprise est la caractérisation des représentations mentales associées aux sons de la parole. Pour traiter cette question, nous utilisons l'approche exemplariste du traitement de la parole, qui propose que les sons de la parole sont mémorisés en incorporant des informations contextuelles détaillées. Nous présentons une nouvelle tâche interactionnelle, GMUP (pour "Group 'em up"), destinée à recueillir les réalisations de matériel phonologique finement contrôlé produit par deux interactants dans un cadre expérimental écologiquement valide. Les variables phonologiques décrivent les différences existant entre deux variétés de français parlé, le français standard et le français méridional. Des outils de reconnaissance automatique de la parole ont été développés pour évaluer la convergence phonétique, observable de l'évolution des représentations mentales, à deux niveaux de granularité : au niveau catégoriel de la variable phonologique et au niveau plus fin, subphonémique. L'emploi de mesures acoustiques détaillées à grande échelle permet de caractériser finement les différences inter-individuelles dans l'évolution de la forme des réalisations acoustiques associées aux représentations mentales en interaction conversationnelle.

Mots-clés Phonétique et phonologie, variation phonologique régionale, interaction conversationnelle, modèles à exemplaires, sociophonétique, convergence phonétique, reconnaissance automatique de la parole.

Regional phonological variation in conversational interaction

Abstract

It is in social interaction, the primary site of the occurrence of spoken language (Local, 2003) that speech is learned, that it is produced everyday and that it evolves. New interdisciplinary approaches to the study of speech, particularly in sociophonetics and in recent developments in conversational interaction, open new avenues for modeling speech processing. A central question in this enterprise relates to the caracterization of the mental representations of speech sounds. We address this question using the exemplarist approach of speech processing, which proposes that speech sounds are stored in memory along with detailed contextual information. We present a new interactional task, GMUP (which stands for "Group 'em up"), designed to collect realizations of highly-controlled phonological material produced by two interactants in an ecologically valid experimental setting. The phonological variables describe differences between two varieties of spoken French, Northern French and Southern French. Automatic speech recognition tools were developed to evaluate phonetic convergence, an observable of the evolution of the mental representations of speech, at two levels of granularity: at the categorical level of the phonological variable and at a more fine-grained, subphonemic level. The use of large-scale detailed acoustic measures allows us to finely caracterize interindividual differences in the evolution of the acoustic realizations associated with the mental representations of speech in conversational interaction.

Keywords Phonetics and phonology, regional phonological variation, conversational interaction, exemplar models, sociophonetics, phonetic convergence, automatic speech recognition.

Thèse préparée au :

Laboratoire Parole et Langage (LPL)
UMR 6057, CNRS & Université Aix-Marseille I
5, avenue Pasteur
13100 Aix-en-Provence

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

La parole est un objet d'étude fascinant car elle est un élément essentiel de l'existence de tout être humain. De la même façon que chaque individu présente des attributs qui lui sont propres et en font des êtres uniques, leur parole présente aussi cette richesse et cette unicité. Nous n'avons par exemple aucune difficulté à reconnaître un proche au téléphone, du premier "coup d'oreille". L'unicité de la parole n'est pas une donnée préétablie, arbitraire, qui pourrait être déterminée par le codage ADN de l'individu, comme certaines de ses caractéristiques physiques (couleur des yeux, taille par exemple). Elle comporte une composante construite, qui est le résultat de processus conscients ou inconscients chez les locuteurs et qui peut dans une large mesure conduire la parole des individus à évoluer indépendamment de leurs caractéristiques physiques. Même si la parole présente bien des liens stables avec la physiologie du locuteur, comme par la relation qui existe entre la longueur du conduit vocal et la hauteur de voix moyenne du locuteur, elle évolue dans un système largement indépendant de ces seules caractéristiques. Par exemple, il est toujours surprenant de découvrir le visage d'un présentateur de radio dont la voix nous est familière, et en général il est difficile de prédire l'apparence visuelle d'un locuteur à partir de sa seule voix, alors même que celle-ci peut véhiculer une quantité d'informations permettant d'identifier, c'est-à-dire attribuer des caractéristiques identitaires au locuteur. Si la parole de chaque individu est unique, la distance qui sépare la parole de deux individus est variable, et peu aller d'une ressemblance presque complète, par exemple dans le cas de deux amis du même sexe qui fréquentent les mêmes milieux sociaux, à des écarts très importants, si l'on compare par exemple la parole d'une fillette et d'un adulte mâle qui ne parlent pas la même langue.

La parole est un support privilégié de l'expression de l'identité. C'est à travers celle-ci que les locuteurs expriment, de façon quotidienne, des prises de position, des choix, des pensées, en somme tout ce qui constitue l'identité

de l'individu. Elle contient ainsi toute la diversité nécessaires à l'expression de ces aspect de l'identité dans toute leurs nuances. De façon importante, ces moyens incluent d'autres dimensions que la seule dimension du langage, que l'on pourrait voir pour contraster la comparaison comme un système de symboles pouvant être combinés pour former des séquences ayant un sens et pouvant satisfaire des besoins de communication. Une de ces dimensions qui caractérise la parole est la variation phonologique, qui peut entre autre porter des informations sur l'appartenance régionale et sociale du locuteur. Cette dimension est primordiale dans la détermination de l'identité du locuteur. Dans les rares cas où cette seule dimension vient à disparaître – cette perte sélective de faculté étant d'ailleurs remarquable –, comme dans le syndrome de l'accent étranger, un trouble du langage d'origine neurologique, c'est toute l'identité de la personne qui est remise en question. Un tel cas a récemment été relaté dans les médias [1] : une personne habitante du sud-ouest de l'Angleterre s'est réveillée un beau matin avec ce que son entourage définissait comme un accent français. Bien que n'ayant pas perdu la faculté de produire des phrases bien formées, ce changement de « surface » a radicalement changé la vie de cette personne, qui a perdu son emploi par exemple. Il est intéressant de noter que la parole pathologique ne présentait pas strictement les traits phonétiques et phonologiques du français, mais la déviance par rapport aux patrons habituels définissant la variété parlée par la locutrice induisait chez ses auditeurs une interprétation d'un accent étranger.

Cette extraordinaire richesse de la parole qui permet à chaque individu de se positionner individuellement dans son environnement social a reçu récemment une attention grandissante dans plusieurs domaines. C'est le cas notamment de la sociophonétique, qui se donne pour objectif « [d']identifier, et à terme d'expliquer les sources, loci, paramètres et fonctions communicatives de la variation socialement structurée dans la parole » (Foulkes *et al.*, 2010, notre traduction). Voir aussi la présentation de cette discipline par Hay et Drager (2007) et les travaux rassemblés dans le numéro spécial du *Journal of Phonetics* (n° 34, 2006) dédié à la variation sociophonétique. Cette discipline est issue de la tradition variationniste qui vise à mettre en relation des formes phonétiques et des facteurs sociaux, et incorpore les méthodes expérimentales de la phonétique. Ce rapprochement est intéressant car il permet de dépasser la vision (par ailleurs passionnante) d'un langage, (et en particulier une parole) stratifié en grands ensembles de traits associés à des catégories sociales. Un des objectifs de la sociolinguistique variationniste est de découvrir les facteurs qui président au changement linguistique. Elle se foca-

[1] URL : http://www.guardian.co.uk/uk/2010/sep/14/woman-awoke-migraine-\french-accent, site consulté le 7 novembre 2010

lise ainsi sur des tendances centrales, et des empans temporels relativement importants pour tester ses prédictions. Il en résulte une modélisation qui, bien qu'elle ait l'avantage d'être généralisable, atténue les différences inter-individuelles pour dégager les tendances centrales associées à des catégories sociales. L'apport des méthodes expérimentales de la phonétique dans la vision d'un langage d'abord vu comme situé socialement, contribue à pouvoir s'intéresser de plus près à des phénomènes qui, s'ils échappaient à des modélisations plus globales, ne portent pas moins des informations pertinentes sur la façon dont les locuteurs/auditeurs utilisent la matière sonore que constitue la parole pour structurer, enrichir et faciliter leurs échanges quotidiens (voir par ex. Hay et Drager, 2007; Docherty, 2007). La forme donnée à la variation phonologique a ainsi bénéficié d'un changement d'une approche à l'autre. La variable phonologique, définie comme le lieu de la réalisation de la variation de réalisation phonétique, était ainsi majoritairement exprimée dans la tradition variationniste comme une alternance catégorielle entre différentes réalisations. Par exemple, les fameux résultats de Labov (1966) sur la stratification sociale du /r/ dans les grands magasins new-yorkais mettait en évidence une association entre une réalisation rhotique et une classe sociale élevée d'une part, et une réalisation non-rhotique avec la classe moyenne d'autre part. Même si les patrons observés ne présentent pas toujours des relations catégorielles, la formulation des associations, elle, l'est. L'approche sociophonétique de son côté, en ne se limitant pas à essayer de mettre en évidence des variations phonologiques avec des groupes de locuteurs, et en adoptant une méthodologie expérimentale plus proche du signal de parole, dans la tradition phonétique, a pu ainsi mettre au jour des patrons de variation entre facteurs sociaux et traits phonétiques plus détaillées. Le travail de Docherty et Foulkes (1999) sur la réalisation de /t/ en anglais de Tyneside en fournit un exemple : si la variation pouvait dans une première analyse être définie comme une alternance entre forme relâchée et glottalisée, des analyses phonétiques plus fines ont mis en évidence une série de propriétés acoustiques, incluant la pré-aspiration, le voisement continuatif, l'affaiblissement de la voyelle, la pré-affrication et la spirantisation. L'analyse la plus informative s'est révélée être en terme de coordination de gestes articulatoires dans la transition voyelle – plosive alvéolaire.

La possibilité de considérer les patrons de variation individuels et la reconnaissance que ceux-ci sont informatifs laisse entrevoir des avancées majeures dans la modélisation de la parole et du langage. Les chercheurs de différents domaines appellent ainsi de plus en plus à prendre en compte cette information individuelle, en argumentant que les modèles actuels de perception / production de la parole sont incomplets s'ils se focalisent exclusivement sur la dimension lexicale (Local, 2003; Docherty, 2007; Hawkins, 2010). On peut

voir ainsi une évolution dans la modélisation du langage qui se complexifie et s'enrichit : d'une linguistique autonome voyant la langue comme un système abstrait de signes, et reléguant la variation phonologique au domaine de la compétence, la sociolinguistique a introduit la dimension sociale comme partie intégrante du langage, et la sociophonétique, qui revendique des liens de plus en plus fort avec une linguistique interactionnelle, socialement située, ouvre la voie vers une modélisation de la parole plus en accord avec la réalité des usages du point de vue des locuteurs eux-mêmes.

L'approche théorique qui semble le mieux à même d'incorporer toute la finesse de la langue dans sa réalisation acoustique, tout en maintenant un pouvoir de modélisation systémique qui est bien sûr un aspect essentiel de la langue, est l'approche par modèles à exemplaires. Développée à l'origine comme une modélisation de la mémoire, elle propose que les sons du langage sont perçus, stockés et produits sous la forme de représentations détaillées (voir entre autres Goldinger, 1997; Johnson, 1997; Pierrehumbert, 2001; Hawkins, 2003; Nguyen, 2010). Ces représentations incorporent à la fois les informations linguistiques distinctives comme les catégories phonologiques qui permettent de différencier les différentes entités qui forment le lexique, mais aussi nombre d'informations traditionnellement considérées comme extra-linguistiques, appelées informations *indexicales* (Abercrombie, 1967). Ces informations décrivent des différences interlocuteurs comme le genre, l'âge, le statut économique, l'appartenance régionale mais aussi des informations personnelles comme le statut émotionnel, l'attitude, et tout un panel de facteurs spécifiques au contexte pouvant être rassemblés sous le terme d'effets de style. Si les composantes « linguistiques » et « extra-linguistiques » sont présentées comme qualitativement différentes, l'approche exemplariste fait précisément l'hypothèse que ces deux types d'information ne sont pas séparés dans les processus de perception, de représentation et de production de la parole.

Le reste de la thèse est organisé comme suit. Dans le **chapitre 2** nous présentons l'approche exemplariste du traitement de la parole et les fondements théoriques apportés par l'étude de la parole en interaction. Nous introduisons ensuite les variétés de français parlé étudiées qui serviront de support à l'étude des représentations mentales des sons de la parole.

Dans le **chapitre 3** nous présentons la démarche méthodologique adoptée pour l'étude des représentations mentales en interaction conversationnelle. Nous présentons une nouvelle tâche interactive qui permet de recueillir les réalisations répétées de mot-cibles en interaction conversationnelle. Ces mot-cibles incorporent des segments critiques qui illustrent les principales différences phonétiques et phonologiques entre les variétés du français standard et méridional. Nous présentons ensuite les techniques utilisées pour localiser et catégoriser ces réalisations, et pour relever l'information acoustique

pertinente.

Les résultats sont présentés dans les deux chapitres suivants. Le **chapitre 4** offre une caractérisation de la variété parlée des locuteurs à partir de l'alignement des segments critiques en variantes de prononciation, tandis que le le **chapitre 5** s'intéresse à l'évolution de la réalisation de ces segments critiques à un niveau plus détaillé, sub-phonémique.

Le **chapitre 6** présente enfin une conclusion de ce travail.

Chapitre 2

Contexte

2.1 Modèles à exemplaires

2.1.1 Introduction

Les modèles à exemplaires constituent un cadre théorique récent dans les théories linguistiques dont le but est d'expliciter la forme associée aux unités linguistiques mais aussi les mécanismes de leur formation. Ils tirent leur origine de travaux visant à modéliser la mémoire, comme ceux de Semon à la fin du XIXe siècle. Ils sont tombés dans l'oubli en raison de l'inadéquation avec la mode de l'époque pour des représentations économiques, minimales et symboliques de la mémoire. La linguistique voyait de son côté se développer le structuralisme, attaché à décrire la langue comme un système de contrastes, en écartant volontairement toute composante cognitive (Bloomfield, 1933), puis le génerativisme, qui bien que mettant l'accent sur le système linguistique comme un système de connaissance, mettait surtout en avant l'aspect économique et efficace de celui-ci, et caractéristique d'un « locuteur-auditeur idéal, dans une communauté linguistique complètement homogène » (Chomsky, 1965, p. 3, notre traduction). Récemment, les modèles à exemplaires ont reçu un regain d'intérêt, notamment avec la découverte et l'intérêt apporté à la capacité du cerveau à mémoriser un grand nombre d'informations détaillées, y compris des événements uniques ou peu courants.

Ils ont été adaptés pour la perception de la parole par Goldinger (1996) et Johnson (1997), en utilisant les modèles de Hintzman (1986) et Nosofsky (1986) respectivement, puis étendus à la production de la parole par Pierrehumbert (2001, 2002, 2003).

Ils sont intéressants pour la perception de la parole car ils permettent de modéliser certains phénomènes autrement inexpliqués, comme le caractère très détaillé de la mémoire des sons de la parole ou l'influence des informa-

tions indexicales dans la reconnaissance des mots, et apportent des explications novatrices pour d'autres phénomènes comme les effets de fréquence dans la reconnaissance des mots. Ils ont suscité un débat nourri entre leurs défendants et les tenants de l'approche abstractionniste, qui tend à se résoudre à travers le développement de modèles hybrides. Ces derniers incorporent la capacité des modèles à exemplaires à prendre en compte explicitement le traitement et la mémorisation des détails, et les capacités formalisatrices des modèles abstractionnistes, qui reflètent l'aptitude indéniable des sujets parlants à manipuler des formes abstraites de connaissance (voir par ex. Hawkins, 2010).

Nous avons choisi le cadre des modèles à exemplaires car contrairement aux modèles abstractionnistes, traditionnellement utilisés en phonétique et phonologie, qui considèrent la langue comme étant homogène pour un groupe donné, l'approche exemplariste voit la langue comme étant d'abord le propre de l'individu, et la modélise comme des connaissances partagées par plusieurs membres d'une communauté. Elle modélise donc dans ses principes mêmes les différences inter- et intra-individuelles.

2.1.2 Principes

La caractéristique principale des modèles à exemplaires est qu'ils considèrent que les représentations associées aux sons de la parole sont stockées sous forme détaillée et continue. Ceci est en contraste fort avec les modèles traditionnels qui considèrent que les sons de la parole sont reconnus et stockés sous la forme d'entités discrètes et invariantes, constituées de symboles sur des alphabets réduits (inventaire phonémique, inventaire lexical, etc.). Par exemple, si le mot '*vert*' peut admettre, dans une approche abstractionniste, une représentation symbolique constituée de la séquence de 3 phonèmes /vɛʁ/, les modèles à exemplaires considèrent que cette entité est représentée par la somme de toutes les expériences que l'auditeur a eues de l'écoute de la forme prononcée de ce mot.

Dans leur formulation initiale, les modèles à exemplaires considèrent que le signal de parole est présenté au système perceptif dans son intégralité, puis est comparé avec la totalité des expériences, appelées *exemplaires*, en vue de sa reconnaissance et de son stockage. Ces expériences sont stockées sur une carte cognitive multi-dimensionnelle qui admet une métrique de similarité : des zones proches sur cette carte ont une valeur similaire. Pour simplifier la formulation, les auteurs considèrent le stimulus comme une matrice à deux dimensions (par ex. Johnson, 1997), temporelle et spectrale, chaque valeur de cette matrice pouvant être comparée avec une zone de la carte cognitive. L'échantillonnage des dimensions est défini par la sensibilité spectrale

et temporelle du système auditif périphérique. Ce patron d'activation spatio-temporel des zones de la carte cognitive correspond à ce que Goldinger (1996, p. 46) a appelé un *écho*, « agrégat de toutes les traces activées, envoyé à la conscience depuis la mémoire à long terme » (notre traduction). Goldinger fait donc apparaître à la conscience seulement à ce stade de traitement la perception du stimulus : ce ne sont pas les dimensions individuelles et physiques du stimulus qui sont perçues, mais le résultat de sa comparaison avec les expériences précédentes de stimuli similaires. Un mot est donc reconnu par sa comparaison avec les traces similaires stockées en mémoire, et est stocké par ce même processus, par le renforcement de l'activation des zones auxquelles il est le plus similaire.

Les zones activées de la carte cognitive activent à leur tour des *catégories* ou étiquettes qui leur sont associées. Ces catégories sont des unités de *sens*, qui sont façonnées dynamiquement par ce processus de comparaison entre un stimulus et la carte cognitive. Ces catégories sont potentiellement infinies en nombre, sont hiérarchisées, et décrivent des domaines temporels différents. Dans les implémentations des modèles, ils sont restreints à des valeurs inférieures ou égales au mot.

Cette description exclusivement montante de la perception des sons de parole doit en réalité être ajustée pour la modéliser de façon complète. On a vu que le cœur du mécanisme de perception selon les modèles à exemplaires se situe du côté de l'activation des zones de la carte cognitive. Mais cette activation n'est pas exclusivement le résultat de la stimulation par le signal de parole ; elle est également modulée de façon descendante par d'autres facteurs que Pierrehumbert (2001) regroupe sous le terme de facteurs *attentionnels*. Ce sont ces facteurs qui vont pré-activer des zones, toujours de façon spatio-temporelle, et biaiser l'association de certaines zones de la carte avec certaines catégories. Par exemple, la portion de signal de parole immédiatement située après la première voyelle de la forme prononcée du mot '*médecin*' (transcrite par le son [t] dans une transcription phonétique étroite) aura toutes les chances d'être mise en relation avec la catégorie correspondant au phonème /d/, et qui par là conduit à la reconnaissance du mot, même si cette portion présente les mêmes caractéristiques phonétiques que la portion de signal suivant la première voyelle de la forme prononcée de 'jet-set'.

Ces influences descendantes que l'on peut attribuer à la connaissance de la langue et donc leur conférer un caractère acquis et stable sont également dynamiques et peuvent révéler un ajustement ponctuel de l'auditeur à la parole. C'est dans ce cadre que l'on peut interpréter les phénomènes de changement de décision de catégorisation de voyelle mis en évidence par Evans et Iverson (2004). Dans leur étude, les auditeurs devaient choisir de façon itérative la meilleure correspondance entre un stimulus de voyelle syn-

thétique et une voyelle contenue dans une phrase porteuse, produite par des locuteurs de deux variétés régionales différentes. L'ajustement du meilleur exemplaire était dépendant de l'accent régional associé à la phrase porteuse, et de l'origine dialectale de l'auditeur. Ces processus d'ajustement dynamiques de l'activation des zones de la carte cognitive peuvent également être mis en parallèle avec les premières observations de l'ajustement rapide de la perception de la parole de l'interlocuteur. Par exemple, Joos (1948, p. 61) propose que l'auditeur construit à partir des premiers échantillons de parole perçus un cadre de référence dans lequel les productions suivantes sont projetées pour dévoiler leur identité. On peut aussi citer les résultats bien connus de Ladefoged et Broadbent (1957), qui montrent que l'identité d'une voyelle ne dépend pas exclusivement de la valeur absolue des formants qui la composent, mais est mise en relation avec les valeurs de formants des autres voyelles du locuteur qui produit la phrase.

Une caractéristique importante des modèles à exemplaires est l'adoption d'une approche probabiliste de catégorisation et de représentation des sons de la parole. Les stimuli qui présentent des similarités avec des exemplaires déjà stockés renforcent les zones associées, ce qui entraîne la formation graduelle des distributions d'activation. Ces distributions correspondent aux régularités présentées par les multiples expériences d'audition des sons de la parole. Les pics des distributions identifient des catégories, auxquelles sont associées les étiquettes (phonème, mot, etc.). Un stimulus qui activera une zone à proximité d'un pic de distribution va activer la catégorie qui correspond à ce pic de distribution et renforcer la distribution à l'endroit de la zone. C'est de cette façon qu'est modélisée l'abstraction dans les modèles à exemplaires ; avec une mesure de similarité entre le stimulus et la distribution, dont le pic correspond à la catégorie abstraite, la catégorie phonétique par exemple. Le caractère abstrait des catégories est un phénomène émergent et dynamique, qui est modélisé lors de l'accès aux représentations, et non lors du stockage comme pour les modèles abstractionnistes. Johnson (1997, p. 111) suggère que « la structure phonologique abstraite est un phénomène flottant, émergeant et disparaissant au fil de la reconnaissance des mots ».

Multi-modalité, poids

Comme nous l'avons vu, les catégories phonologiques, qui constituent un sous-ensemble des distributions contenues sur la carte cognitive, se dégagent des propriétés statistiques de la substance acoustique de la parole. Cette propriété permet de modéliser les différences très ténues qui existent entre les inventaires des sons des langues du monde, comme le montrent les exemples proposés par Pierrehumbert (2001). Ainsi, le français québécois diffère à la

fois de l'anglais canadien et du français métropolitain dans la distribution des temps de VOT des plosives voisées et non voisées (Caramazza et Yeni-Komshian, 1974). En fait, aucune langue ne possède deux catégories sonores identiques, et cela peut s'expliquer par le fait que les matériaux sonores qui constituent les langues s'actualisent et évoluent à travers la parole de façon indépendante. Ainsi les distributions qui résultent de ces régularités n'ont *a priori* pas de raison de présenter les mêmes caractéristiques. Cette vision détaillée du stockage des catégories phonologiques remet en question l'existence d'universaux segmentaux caractérisant les langues, comme les traits distinctifs, et servant de base à la création d'un inventaire phonémique par combinaison de traits. Cette vision des catégories phonologiques considère plutôt un espace continu, de haute dimensionalité et résolution spatiale (déterminée par la sensibilité physiologique du système auditif par exemple), dans lequel n'importe quelle discrétisation, au sens de renforcement de zones, est possible. Le fait d'observer des similarités dans l'inventaire des langues apparaît ainsi plutôt comme le résultat d'une contrainte articulatoire du système phonatoire commun à tous les sujets parlants, et qui limite les possibilités articulatoires et acoustiques de l'espace vocalique à l'intérieur du quadrilatère défini par les voyelles cardinales, par exemple.

Comme n'importe quelle régularité dans la forme sonore entraîne de fait la formation d'une distribution, on peut considérer l'existence d'unités temporelles (à savoir la présence de pics de distribution) à n'importe quelle granularité. En particulier, on verra émerger les unités correspondant aux descriptions traditionnelles de la langue, comme le mot ou la catégorie phonétique, qui sont les plus couramment utilisées dans la présentation des modèles à exemplaires. L'exposition répétée à des mots forme le lexique, tandis que l'exposition répétée aux sons composant les mots forme les catégories phonétiques : les réalisations de la voyelle dans les mots '*vert*' et '*terre*', de par leur similarité, vont renforcer la même représentation et s'unifier dans la catégorie phonétique [ɛ]. Cette présentation des modèles à exemplaires s'apparente en fait à une modélisation abstractionniste, où les mots seraient composés d'une séquence d'unités invariantes et abstraites : les pics de distribution. Mais s'il est vrai que cette régularité des catégories phonétiques, à la fois temporelle (d'une répétition d'un même mot à l'autre) et entre les mots est frappante, et constitue la base de la formation des catégories dans les modèles à exemplaires, la parole est en même temps caractérisée par une grande variabilité. Celle-ci n'est pas uniquement le fait de variations aléatoires autour des réalisations canoniques, mais elle est au contraire structurée. Les sources de cette variabilité sont multiples, et incluent les processus phonétiques naturels (universels, d'une certaine façon), comme la coarticulation, et les caractéristiques du locuteur, comme sa hauteur de voix moyenne ou son

accent régional. Ces variations, comme elles sont régulières, doivent entraîner selon les modèles à exemplaires la formation de distributions associées à ces régularités dans la carte cognitive. Ainsi, même en se limitant au niveau de granularité temporelle des segments, on ne peut pas considérer un mot comme une succession de distributions identifiant chacune une catégorie phonétique : il faut considérer des distributions *multi-modales*, chaque mode correspondant à une régularité de variation rencontrée au fil des expositions au mot parlé. Ainsi, la représentation mentale pour le mot '*lait*' peut être constituée d'autant de distributions (ou de catégories) que les régularités que l'on peut énoncer pour décrire les différentes expériences du mot : mot produit par une voix d'homme, de femme ou d'un locuteur spécifique, prononcé avec des intonations différentes selon le contexte dans lequel il est utilisé, avec une qualité de voyelle indiquant l'appartenance régionale du locuteur... Dans cette présentation exclusivement probabiliste, n'importe quelle régularité donnera lieu à la création d'une catégorie, et induira donc des unités de *sens* associées à ces régularités. Cette vision est un peu trop extrême, et illustre plutôt un *mécanisme* par lequel les unités de sens peuvent se former plutôt qu'une description déterministe de la formation effective des catégories. Ce point de vue est équivalent à dire que le système perceptif peut potentiellement attribuer du sens à n'importe quelle régularité observée, mais ce n'est pas pour autant qu'il le fait. Par exemple, on ne peut pas prédire que des locuteurs ayant passé une longue période dans un environnement linguistique non natif acquièrent toutes les caractéristiques de celui-ci, et avec le même degré de précision. On observe ce phénomène chez certains sujets, validant les possibilités offertes par ce mécanisme, mais la non systématicité de l'effet de l'exposition sur les réalisations amène à considérer d'autres facteurs. Tous les modèles à exemplaires proposent ainsi des facteurs de plus haut niveau, d'influence descendante, qui modulent ces mécanismes probabilistes et déterministes de création de sens à partir des régularités statistiques présentées par la substance de la parole. Ces facteurs sont modélisés par des *poids* qui influencent la sélection d'une catégorie parmi plusieurs possibles, et même leur formation. Ce sont ces ajustements de poids qui peuvent pré-activer une catégorie phonétique permettant la reconnaissance d'un mot, lorsque celui-ci est présenté dans du bruit par exemple. Ce sont ces mêmes poids qui peuvent pré-activer les distributions associées à la voix du locuteur, et faciliter ainsi la reconnaissance de ses productions. Enfin, c'est par une pondération de l'activation des catégories que l'on peut modéliser le fait qu'à certaines régularités du signal ne soient pas attribuées de distributions, bien qu'elles soient majoritaires dans la parole ambiante et exploitées par les locuteurs, et qu'en résultat l'auditeur manifeste une insensibilité à des oppositions basées sur ces paramètres. Si tous les modèles exemplaristes ont

recours à ces poids pour expliquer ces faits de perception (et de production, voir ci-après), aucun d'entre eux n'explicite comment ils sont intégrés dans les modèles (Pierrehumbert 2001, Foulkes et Docherty 2006, p. 431). Une raison possible à cela est que l'ajustement des poids peut être rapproché des procédures de normalisation chères aux modèles abstractionnistes, qui considèrent qu'une partie de l'information est écartée dans les processus de perception de la parole, comme c'est le cas des exemples présentés ci-dessus.

Les modèles à exemplaires sont dotés d'un mécanisme d'*oubli*, qui s'exprime simplement dans l'approche probabiliste par une atténuation des activations les plus anciennes (Pierrehumbert, 2001). Ce mécanisme permet de modéliser la façon dont les locuteurs adaptent leur système linguistique au fil du temps et de leurs interactions sociales, donnant une explication simple au changement linguistique à long terme. Combinés avec la capacité de représentation détaillée de la substance sonore, ces mécanismes d'oubli permettent de la même façon, à travers des petits effets d'imitation dans la boucle perception-production au niveau de la communauté (Pierrehumbert, 2006, p. 521), la formation des dialectes.

Production

Les modèles à exemplaires proposent un mécanisme similaire pour la production, à savoir une sélection probabiliste des cibles et l'ajustement de poids pour une modélisation complète de la production. Ils supposent que la perception et la production partagent les mêmes représentations mentales. Les cibles, qu'elles soient considérées comme étant des mots ou des segments, sont sélectionnées par l'activation de zones dans le voisinage du pic de distribution. Cette sélection aléatoire modélise d'une part la variabilité de production rencontrée dans la parole, et d'autre part le renforcement de la catégorie activée. En effet, l'accès à une catégorie pour la production modifie les représentations stockées par l'activation des zones, et mathématiquement, la sélection aléatoire d'une zone autour du pic renforce le pic de cette distribution.

En fait, une formulation plus rigoureuse pourrait se passer de la sélection aléatoire d'une zone et postuler simplement l'activation du pic de la distribution comme mécanisme de base. La variabilité aléatoire dans la production peut se modéliser alors à travers des processus subséquents de production (imprécision articulatoire par exemple), et les autres types de variabilité par l'ajustement des poids. Le fait que les distributions soient modifiées lors de l'accès pour la production rend compte d'un aspect essentiel de la production : il faut donner une importance disproportionnée aux productions du locuteur par rapport aux expériences entendues dans les activations des distributions. Si on ne prend pas en compte cet aspect, alors les productions

du locuteur seraient totalement déterminées par la parole ambiante, et un homme pourrait adopter une hauteur de voix de femme s'il n'est exposé qu'à de la parole produite par des femmes. Au contraire, le fait que les pics de distribution soient renforcés également par la production permet au locuteur de construire et de maintenir des régularités dans sa propre production. Les poids sont aussi primordiaux pour décrire les patrons de production observés. La forme produite des mots n'est évidemment pas constante, et manifeste des changements catégoriels qui peuvent être observés dans les changements de style par exemple. Ainsi, des changements dans l'intonation peuvent être observés entre la parole conversationnelle et la lecture de texte. Au niveau segmental, on peut mettre en évidence des différences entre le style formel et le style familier, par des réalisations plus articulées pour le premier. Ces changements sont modélisés par les modèles à exemplaires par l'activation de distributions différentes au cours de la production.

L'approche exemplariste a également suggéré des pistes intéressantes dans le domaine de l'acquisition de la parole (Foulkes et Docherty, 2006) : la parole adressée aux enfants est caractérisée par un vocabulaire et une syntaxe simplifiée, un débit plus lent (Snow, 1995) et des contrastes phonologiques exagérés (Kuhl *et al.*, 1997; Malsheen, 1980). On peut voir ça comme un comportement qui contraste au maximum les différences existant dans la parole, et qui facilite l'émergence de catégories distinctes. La parole de l'enfant est plus similaire à la parole de la mère au début de l'acquisition : l'enfant parle peu, et ses distributions proviennent en plus grande partie de son exposition. Par la suite, l'enfant attribue des poids plus importants aux caractéristiques de la parole correspondant à son genre, il biaise donc sa production vers les distributions caractéristiques de son genre, à travers l'ajustement des poids.

2.1.3 Succès de l'approche exemplariste

Une des contributions les plus claires de l'approche exemplariste à la modélisation du traitement de la parole concerne les effets de fréquence dans la reconnaissance des mots parlés, les mots les plus fréquents étant reconnus le plus rapidement (voir Jurafsky, 2003, pour une revue). En comparaison, les modèles générativistes attribuent les effets de fréquence au domaine de la performance linguistique. Dans des tâches de shadowing, Goldinger (1998) a mis en évidence que les mots les plus fréquents étaient répétés plus rapidement. Il avance que les mots fréquents sont mieux représentés en mémoire et ont donc un niveau d'activation plus important lors de la présentation d'un nouvel exemplaire. Cela facilite sa reconnaissance, et donc la rapidité de répétition. Il a aussi trouvé que les mots les moins fréquents étaient répétés avec une meilleure fidélité. En manipulant la latence imposée de répétition,

Goldinger a également trouvé une meilleure fidélité de répétition pour les latences les plus courtes. Ceci est dû au fait que l'écho retourné par la comparaison avec les exemplaires stockés comporte une plus grande influence de la dernière présentation et sa production en est une meilleure approximation.

Les modèles à exemplaires ne privilégient pas une unité temporelle particulière, celles-ci émergeant de la structure du langage, suivant les principes probabilistes évoqués plus haut. Il semblerait donc possible d'observer les effets de fréquence observés sur les mots par Goldinger à d'autres niveaux de granularité temporelle, et c'est précisément ce qu'a observé Connine (2004) pour le niveau du segment. Dans une tâche d'identification de phonème en position initiale de mot, elle a trouvé que les réponses correspondant à des mots du lexique étaient plus fréquentes lorsque le stimulus contenait en position médiane une variante phonologique de fréquence élevée, par opposition à une variante moins courante. Ainsi, la séquence sonore /PreTi/ , où /P/ est le phonème ambigu à identifier sur un continuum /b/–/p/ et /T/ la variante alternant entre /t/ et /ɾ/, était plus souvent interprétée comme un mot du lexique (*pretty*, 'joli' (mot) vs. *bretty* (non mot)) lorsque la variante était /ɾ/, plus fréquente en position intervocalique en anglais américain. Cela montre que la fréquence d'une variante phonologique a une influence sur l'accès aux représentations lexicales, et est compatible avec un stockage détaillé et explicite des variantes. Selon Connine, une telle observation est incompatible avec un stockage abstrait et économique des unités lexicales sous la forme de suite de symboles, suivant des processus de normalisation qui feraient correspondre plusieurs variantes phonologiques à une seule entité symbolique.

La postulation de procédures de normalisation fait apparaître un autre problème. Si on postule une seule entité correspondant à la variable phonologique, comment expliquer que l'auditeur puisse être exposé à plusieurs variantes ? Pour être produite, chez un locuteur de la langue, et étant écarté le fait que cette variante soit le résultat de processus phonétique naturel comme la coarticulation, la variante doit être mémorisée, et donc identifiée comme telle. On pourrait postuler des mécanismes qui effectuent une séparation entre la forme particulière qui identifie la variante et la forme stable qui identifie le phonème, c'est-à-dire la fonction du son dans le système phonologique, et qui les présentent conjointement à un module d'accès au lexique. Lors de la production, ces mécanismes combineraient à nouveau ces deux composantes pour restituer la variante en question. Mais cela semble assez consommateur de ressources et d'une implémentation compliquée, artificielle, et ainsi contradictoire avec les capacités de simplification et d'abstraction prêtées au système perceptif.

Les phénomènes d'assimilation ont été le sujet de nombreuses recherches alimentant le débat sur les différentes approches du traitement de la parole

et la forme des représentations mentales (voir Nguyen, 2010, pour une revue récente). L'approche abstractionniste cherche à mettre en évidence des mécanismes mis en œuvre par les locuteurs pour retrouver une forme invariante sous-jacente, en postulant par exemple une sous-spécification des traits (Lahiri et Reetz, 2002) ou l'application de transformation inverse (hypothèse d'inférence phonologique de Gaskell, 2003). Elle modélise ainsi des processus qui caractérisent la « parole connectée », vue comme une parole canonique contaminée par des phénomènes d'approximation. A l'inverse, l'approche exemplariste (voir par ex. Local, 2003, dans une cadre un peu différent mais compatible) considère que les phénomènes d'assimilation sont porteurs d'indices acoustiques distribués portant sur le mot dans son ensemble, et qui invitent à se dégager d'une vision segmentale réductrice du traitement de la parole.

D'autres phénomènes ne peuvent être expliqués qu'avec les modèles à exemplaires. Par exemple, Yaeger-Dror et Kemp (1992) ont montré que certaines voyelles du français québécois, considérées comme un phonème unique dans une phonologie purement distinctive, adoptent en fait des réalisations acoustiques différentes selon l'item lexical dans lequel elles apparaissent (et selon des caractéristiques individuelles des locuteurs, comme l'âge, le sexe, le rang social). Une régularité a de plus été mise en évidence dans les patrons de variation : les mots qui partageaient des réalisations similaires de la voyelle appartenaient tous à un champ lexical précis, et se référaient plus souvent au passé. Ceci est en contradiction avec une séparation supposée de la grammaire phonologique et le niveau lexical.

Dans une autre étude, Hay *et al.* (2009) ont examiné la réalisation de la consonne /t/ en position finale de mot en anglais de Nouvelle Zélande, qui présente actuellement un changement phonétique, vers une forme non relâchée. Ils ont observé que les adultes produisaient plus de formes relâchées lorsqu'ils relataient des événements appartenant au passé, la forme alors majoritaire. Ces patrons de variation sont incompatibles avec un phonème unique /t/ éventuellement soumis à un changement linéaire d'une forme relâchée vers une forme non relâchée. En plus de fournir une explication intégrée du changement phonétique en le modélisant comme un déplacement du pic de distribution d'une forme à une autre reflétant les changements dans la parole environnante, l'approche exemplariste propose que l'effort de souvenir par les locuteurs d'événements passés à mobilisé les exemplaires appris à cette époque, donnant une forme relâchée à la consonne /t/. Un mécanisme possible pour cette réalisation est l'ajustement des poids qui activent les pics de distribution associés. Ces patrons pointent donc vers des représentations mentales détaillées, comme en témoigne la conservation de plusieurs réalisations associées à la même unité fonctionnelle, et contextualisées, par

la mise en évidence d'un lien entre réalisation acoustique et son contexte d'apprentissage.

Propriétés indexicales

Un sujet de débat entre les approches abstractionnistes et exemplaristes concerne l'information associée au locuteur dans le signal de parole, ou informations *indexicales* (Abercrombie, 1967). Les approches abstractionnistes considèrent cette variation comme non désirable, devant être éliminée par des processus de normalisation pour atteindre le contenu lexical et sémantique du mot. Ces processus de normalisation incluent par exemple la suppression de l'influence de la variation de longueur du conduit vocal sur le spectre de parole. En général, cette influence est assez bien décrite par une corrélation entre la hauteur des formants et la longueur du conduit vocal. Un grand nombre de techniques se sont ainsi développées (voir Adank *et al.*, 2004, pour une revue récente) visant à appliquer aux mesures acoustiques la transformation inverse de celle introduite par les caractéristiques du locuteur, principalement la longueur du conduit vocal. L'objectif de cette transformation est d'atteindre des patrons formantiques *indépendants du locuteur*, présentant une forme d'invariance, et qui peuvent par là être plus facilement mis en correspondance avec des catégories phonétiques ou phonologiques supposées fixes. Le problème, d'une certaine façon, c'est que ces différences hommes/-femmes introduites par des différences de longueur de conduit vocal existent bien, et donc ces procédures de normalisation atteignent un certain degré de performance quand il s'agit par exemple de discriminer les catégories de voyelles. Une classification donne en effet de meilleurs résultats à partir des données transformées (ex : normalisation de Lobanov : 92% de classification correcte) qu'avec des données non transformées (données en Hertz : 79%). En réalité, les différences associées au genre du locuteur, pour ne prendre que cette dimension des différences interlocuteurs, sont plus subtiles, comme le montre la grande diversité des différences hommes / femmes dans la comparaison interlangues présentée dans Johnson (2006). Ces résultats montrent d'une part une certaine indépendance vis-à-vis des différences de longueur de conduit vocaux – même si des valeurs de formants globalement plus élevées se vérifient pour les femmes –, et d'autre part que les locuteurs *exploitent* ces différences biologiques pour construire la dimension sociale du genre Caplan (1987) : dans la même étude, Johnson montre que les hommes adoptent un formant F_2 qui est légèrement corrélé *positivement* avec la taille du conduit vocal, ce qui indique en particulier que les hommes ayant un conduit vocal plus petit que la moyenne présente un formant F_2 plus *bas* que la moyenne. Ainsi, la recherche de paramètres pouvant être normalisés, même sur des

dimensions qui semblent aussi robustes que le genre, et même si elle donne de bonnes performances dans certaines tâches spécifiques, apparaît comme invalide, surtout si l'on suppose que ce sont de tels processus qui participent à l'extraction supposée de l'information linguistique pertinente dans la parole de l'interlocuteur. Pour aller plus loin, il est maintenant admis que non seulement les informations indexicales n'empêchent pas les locuteurs de décoder l'information lexicale (elles empêchent bien les outils développés pour accomplir ces tâches...), elles permettent au contraire de *faciliter* l'accès à cette information, et à sa mémorisation. Strand (2000) a par exemple montré que le délai de répétition de mots était plus court lorsque ceux-ci étaient prononcés avec une voix stéréotypique d'homme ou de femme, par rapport à des voix non stéréotypiques, ce qui indique une possible facilitation de l'identification du genre du locuteur sur la reconnaissance du mot. Goldinger (1997) est de son côté arrivé à la conclusion que l'information indexicale du locuteur devrait être conservée en mémoire et permettrait ainsi une meilleure mémorisation de mots prononcés : une liste de 10 mots prononcés par 10 voix différentes pouvait être mieux mémorisée que lorsque les mots étaient prononcés par une seule voix. Enfin, Johnson (1997) montre qu'il n'est pas nécessaire de *supprimer* (ce à quoi revient d'appliquer une procédure de normalisation) l'information associée au locuteur pour implémenter des procédures de classification permettant de différencier avec succès un ensemble de voyelles. Il présente pour cela une simulation d'un modèle à exemplaires simplifié dans lequel les stimuli sont représentés par 5 dimensions acoustiques (durée, f_0, F_1, F_2, F_3) et obtient des performances d'identification de voyelle qui sont comparables à celles d'auditeurs se basant sur l'information formantique au milieu de la voyelle. Le modèle permet également de prédire correctement le sexe du locuteur, sans que soient appliquées des transformations explicites dans les paramètres acoustiques.

Reconnaissance automatique de la parole

La reconnaissance automatique de la parole peut être considérée comme implémentant certains principes des modèles à exemplaires, dans leur plus simple formulation. La proximité vient des principes probabilistes qui dans les deux cas déterminent les catégories. Dans les modèles à exemplaires, les catégories abstraites émergent comme moyennage de toutes les catégories présentées au système perceptif. Dans les systèmes de reconnaissance basés sur les modèles de Markov cachés, les modèles acoustiques, entités abstraites associées aux phones (catégories phonétiques) sont explicitement calculés par moyennage multidimensionnel des paramètres acoustiques décrivant les portions de signal de parole associés aux phonèmes. Cependant,

dans leur implémentation avec des modèles monophones, les systèmes de reconnaissance automatique font des hypothèses simplificatrices significatives, en supposant une équivalence de réalisation acoustique des phonèmes dans tous les contextes (définissant par là le phone). Or on sait qu'en plus de l'influence contextuelle segmentale de réalisation, qui est modélisée par les modèles triphones, les réalisations des phonèmes présentent une part importante de variation non attribuable au contexte phonétique immédiat (Pierrehumbert, 2003, p. 129). Par exemple le même phonème défini dans une langue par les oppositions de sens qu'il permet peut admettre des réalisations différentes et *régulières* en fonction du style de parole, du locuteur qui le produit, de l'accent, etc. Alors que les systèmes de reconnaissance automatique rassemblent ces réalisations dans une classe d'équivalence, les modèles à exemplaire permettent précisément, à partir des régularités physiques (vecteurs de paramètres acoustiques/auditifs) de construire autant de catégories qu'il en existe dans la parole rencontrée.

2.1.4 Limitations

Pierrehumbert maintient l'existence d'une « grammaire phonologique », à l'interface entre la perception et le lexique. Il est en effet indéniable que les locuteurs ont conscience que le langage est composé d'unités symboliques (phonèmes, syllabes...) qu'ils peuvent manipuler, en créant des nouveaux mots par combinaison d'unités plus petites, et en apprenant des nouveaux mots. Il est en effet peu probable que l'apprentissage d'un nouveau mot d'une langue ne soit possible que grâce à l'écho renvoyé par la comparaison entre le stimulus et les exemplaires déjà présents dans le lexique, même si celui-ci évoque une trace plus importante du fait de sa rareté, puisque ces exemplaires n'existent pas. Par contre, les unités le composant sont bien présentes dans le lexique, à travers les distributions associées à d'autres mots, à savoir les exemplaires non pas du mot mais d'unités plus petites. Mais permettre la mise en relation d'une succession d'exemplaires (associés à des catégories phonétiques, pour fixer les idées) qui n'a jamais été rencontrée dans la langue avec la mémorisation de celle-ci implique forcément une capacité à combiner des entités abstraites. Cependant, on peut pousser un peu plus loin la vision probabiliste de l'approche exemplariste, et ne pas s'arrêter à considérer des distributions associées aux seules catégories phonétiques, et aux mots, car cela revient assez rapidement à tomber dans une vision symbolique des représentations de la parole. En effet, si les niveaux du mot et de la catégorie phonétique sont souvent invoqués comme granularité, aucun niveau n'est fixé *a priori*. En particulier, il faut également considérer les probabilités de transition entre des catégories phonétiques qui s'enchaînent dans la parole.

Même si la conceptualisation n'est pas aisée, on peut très bien considérer une séquence de deux catégories phonétiques non pas comme une succession d'entités discrètes, ni comme une réalisation qui donnerait lieu à une distribution indépendante de celle qui la compose, et qui serait hautement inefficace pour le stockage des représentations, mais comme un exemplaire composé de ces deux éléments eux-mêmes constituant deux exemplaires, et faisant partie d'exemplaires de plus haut niveau. L'activation résiduelle de cet exemplaire est composée, selon les principes énoncés plus haut, de l'activation des éléments qui le composent, c'est-à-dire de la fréquence d'occurrence des éléments, mais aussi de la fréquence d'occurrence de la *séquence*. En appliquant ce principe à la totalité du nouveau mot, on réalise alors que l'on peut considérer ce nouveau mot comme un nouvel exemplaire, dont la nouveauté, et donc les ressources à mettre en œuvre, est bien moins importante que l'on pourrait penser : ce n'est pas une séquence totalement nouvelle puisqu'elle est composée d'éléments déjà présents sous la forme d'autres exemplaires, et même l'enchaînement de ces symboles n'est pas à retenir d'une façon abstraite et indépendante de la connaissance du lexique, mais au contraire dépend largement des représentations déjà contenues dans le lexique selon le même principe probabiliste, mais que l'on exprime ici comme la connaissance des probabilités de transition. Le degré de nouveauté associé à ce nouveau mot est donc minime, et on n'a pas besoin d'invoquer une grammaire probabiliste de fonctionnement qualitativement différent, qui fournirait cette facilitation soudaine pour la mémorisation et production de ce nouveau mot.

Pour le formuler d'une autre façon, il faut appliquer le principe probabiliste également à la dimension temporelle, de la même façon que pour la dimension spectrale, ce qui n'a pas été le cas dans les modèles développés jusqu'ici : si Pierrehumbert a recours à un niveau abstrait intermédiaire, Johnson (1997) invoque quant à lui un *détecteur de surprise*, qui fournirait les informations nécessaires à la segmentation temporelle du signal et ainsi les limites nécessaires pour l'apprentissage des exemplaires. Ainsi, si les modèles à exemplaires ont proposé un changement radical et opposé à une vision traditionnelle abstractionniste du traitement de la parole, une partie de leur formulation est toujours empreinte d'une vision symbolique et économique. Il reste en effet à développer un modèle qui intégrerait dès le départ la dimension temporelle au même titre que la dimension spectrale dans la formulation probabiliste des représentations des sons de la parole.

En réalité, la faiblesse attribuée à l'approche exemplariste pour la modélisation des aspects combinatoires de la langue est même défendue comme une force par Bybee et McClelland (2005), précisément car les langues telles qu'elles sont utilisées ne présentent pas d'aspect combinatoire systématique.

Inclusion d'autres modalités sensorielles

Pour des raisons évidentes de simplicité, les premières expositions des modèles à exemplaires se sont concentrées sur les informations ayant principalement trait à la parole, et même exclusivement auditives. Pourtant, la démarche met au centre de la modélisation l'intégration du contexte dans la formation et l'accès aux exemplaires. On a vu par exemple que les informations indexicales (sexe, appartenance régionale, âge, état émotionnel,...) trouvaient leur place dans les représentations. De même l'association entre la réalisation phonétique des /t/ de l'étude de Hay *et al.* (2009) montre une association entre des catégories phonétiques et une période temporelle remontant à plusieurs années chez les locuteurs adultes. En réalité, la frontière qui sépare les informations pouvant être dérivées directement du signal et celles qui sont ajoutées par l'auditeur comme résultat de sa perception globale, multisensorielle de la séquence de parole, est pour le moins floue. On peut par exemple considérer que l'évaluation de l'appartenance dialectale du locuteur peut ne pas trouver des corrélats acoustiques dans la portion de signal considérée, mais être dérivée de l'exposition à de telles caractéristiques dans une portion précédente de parole (par exemple, une façon caractéristique de dire « Bonjour » qui peut être perçue comme un shibboleth), ou même être dérivée d'indices non verbaux, comme la tenue vestimentaire du locuteur. En fait, il apparaît que toutes les informations perceptibles par l'auditeur peuvent potentiellement contribuer à la formation des exemplaires comme unité de sens associé au langage. Dans les faits, leurs associations ne sont que peu retenues parmi les multitudes possibles, simplement car elles sont trop peu fréquentes. Par exemple, l'association créée par l'audition simultanée de la voyelle /a/ et le toucher d'une feuille de figuier, si même elle se produit, a peu de chance d'être reproductible et s'atténuera rapidement. Si le langage parlé présente une certaine homogénéité sensorielle qui est l'audition, c'est simplement parce qu'il s'est développé en utilisant principalement cette modalité. Mais les langues varient dans leur proportion d'utilisation de la modalité auditive, certaines langues font par exemple une utilisation plus importante de gestes, d'autres écartent complètement la modalité auditive, comme les langues des signes – qui présentent par ailleurs les mêmes caractéristiques structurelles (lexique, syntaxe, dialectes, etc.) que les langues orales. C'est dans cette optique que l'on peut voir l'intégration relativement récente à la langue de facteurs considérés traditionnellement dans le domaine de la linguistique générale comme extra-linguistiques, comme les signaux non-verbaux (backchannels, mimogestualité) ou la qualité de la voix (signalement de la fin de tour de parole en finlandais, Ogden (2004)). Même l'amplitude globale de la voix, un paramètre acoustique que l'on ne considère pas traditionnellement comme gardé

en mémoire en tant qu'aspect pertinent de la forme de surface des sons de la parole (Bradlow *et al.*, 1999; Church et Schacter, 1994), pourrait potentiellement l'être, si cet aspect est utilisé de façon distinctive. Ce paramètre peut par exemple faire partie des caractéristiques indexicales d'un locuteur, et on pourrait facilement imaginer une expérience qui fait varier systématiquement ce paramètre dans l'apprentissage de catégories (artificielles) de sens. On trouve de nombreux arguments supplémentaires dans l'expérience quotidienne du langage qui renforcent l'association multi-sensorielle avec les représentations des sons de la parole du fait de leur co-occurrence répétée dans l'exposition au langage, comme l'activation mentale de la forme écrite des mots lors de la perception auditive, ou encore l'évocation auditive de sons de la parole lors de la présentation de mouvements articulatoires exclusivement visuels, lorsque le son d'un film est coupé par exemple.

Il semble que l'on puisse pousser le principe probabiliste encore plus loin, en l'étendant explicitement à d'autres modalités dans l'approche exemplariste. Cela permettrait peut-être d'intégrer la présence des *poids* régulièrement invoqués comme influence descendante pour ajuster des faits de perception non comme des règles externes, composantes abstraites, pouvant moduler arbitrairement la perception et la production de la parole, mais comme des composantes intégrées à la construction et à l'accès aux exemplaires, éventuellement avec des empans temporels bien plus grands que le segment ou le mot.

Les modèles à exemplaires ne sont pas en premier lieu une modélisation de la langue, mais une modélisation de la représentation cognitive de la langue. Comme ce mécanisme sert à représenter la langue, et que cette dernière admet une description systémique, cette structure systémique est transférée dans les représentations mentales du langage dérivées par le mécanisme probabiliste de perception par exemplaires.

2.1.5 Conclusion

Les modèles à exemplaires se présentent ainsi comme un mécanisme approprié pour modéliser les processus de perception et de représentation de la parole. Dans leur formulation actuelle, ils nécessitent l'inclusion de mécanismes de plus haut niveau pour rendre compte complètement du traitement de la parole. Une extension du principe probabiliste à la dimension temporelle et à d'autres modalités sensorielles semblerait néanmoins pouvoir contribuer à combler les limitations de cette approche.

Les exemplaires se construisent par exposition répétée à des échantillons de parole. Ces échantillons renforcent les représentations similaires déjà présentes dans la mémoire de l'auditeur. Cependant, on a vu que la seule expo-

sition répétée n'était pas suffisante à rendre compte de la complexité de ce mécanisme. Par exemple, d'autres facteurs influençant la forme sonore que peuvent présenter les sons de parole, comme les caractéristiques indexicales de l'interlocuteur ou le contexte de la communication, jouent un rôle déterminant dans la formation des exemplaires. Les informations fournies par d'autres modalités sensorielles, par exemple visuelles, ou projetées par l'auditeur sur la situation de communication, ont également une influence sur la perception et la représentation des sons de la parole.

Cet ensemble de forces qui diversifient la forme sonore des échantillons de parole sont donc des éléments intégrants et nécessaires au processus de traitement de la parole chez l'auditeur, et c'est à travers ce processus que l'auditeur opère le passage des stimuli auditifs qui lui sont présentés à des catégories de sens associées à ceux-ci, lui permettant de comprendre son interlocuteur et de communiquer avec lui.

La diversité de ces forces constitue en réalité les conditions naturelles de l'exposition à la parole dont les locuteurs font l'expérience quotidiennement, c'est-à-dire l'interaction sociale. Il apparaît ainsi nécessaire de considérer ce cadre là pour l'étude de la formation et de l'évolution des représentations mentales. Nous présentons dans la section suivante les caractéristiques, les avantages et les enjeux de la parole en interaction.

2.2 Parole en interaction

La parole en interaction présente un certain nombre de caractéristiques qui rendent son étude nécessaire dans l'exploration des représentations mentales de la parole.

L'intérêt pour la parole en interaction est un élément central de l'analyse conversationnelle. Développée d'abord dans une perspective sociologique, visant à explorer les compétences et pratiques développées par les locuteurs dans l'interaction sociale, elle se propose d'étudier les régularités qui structurent la parole en interaction et qui lui sont propres, c'est-à-dire indépendantes de l'individu ou de l'institution sociétale. Elle contraste en cela avec les approches traditionnelles individualistes et sociologiques, qui tentent de l'expliquer tour à tour par les forces internes vs. externes, mais la considère dans les deux cas comme un épiphénomène.

L'analyse conversationnelle se centre sur les pratiques des interlocuteurs, et s'est développée en défendant une méfiance et une remise en question systématique des catégories présupposées associées aux comportements humains. L'objectif poursuivi est de proposer des explications qui soient cohérentes avec le point de vue des praticiens. Les locuteurs explicitent rarement les

interprétations qu'ils font de la parole de leur interlocuteur mais l'encodent implicitement dans leur réponse. C'est ce qui permet de construire l'intersubjectivité, de co-construire le sens et de développer un socle commun pour des actions émergentes (Heritage, 1984).

Cette approche, mise en regard de celle défendue par une linguistique autonome qui considère que la langue est caractéristique d'un individu idéal, a des implications importantes pour l'exploration de la modélisation des représentations mentales.

2.2.1 Variabilité

Un des changements de perspective introduits par l'analyse conversationnelle concerne la variabilité, notoirement présente dans la parole, à tous les niveaux. Depuis longtemps considérée comme un obstacle à la modélisation de la parole, elle se révèle être porteuse d'information pour la compréhension du traitement de la parole par les locuteurs. Les processus observés dans les interactions, comme la gestion des tours de parole, ne doivent pas être considérés comme des épiphénomènes se rajoutant aux structures langagières traditionnelles permettant de décrire la parole monologique, et l'écrit. Au contraire, les régularités observées sont définitoires de l'interaction, et les systèmes linguistiques sont mobilisés pour satisfaire les buts interactionnels considérés comme des buts pratiques. Les processus interactionnels contribuent à former, à reformater les systèmes linguistiques mobilisés (grammaire, syntaxe, phonétique/phonologie) en fonction des nécessités locales de l'interaction, qui sont déterminées par les locuteurs, le sujet de conversation, etc. Ce faisant, ils affectent la forme de surface de ces systèmes, et c'est pour cela que dans une perspective où ces systèmes sont considérés comme figés, intériorisés par chacun des locuteurs et indépendants dans une certaine mesure, ces transformations locales, dynamiques, contingentes ne peuvent être expliquées autrement que comme des déviances qui sont perturbatrices.

On peut contraster deux visions de la variabilité présente dans la parole en interaction ou conversationnelle. La première considère que les auditeurs font face à des perturbations introduites par un relâchement lié à la situation d'élocution, et qui se traduit par des phénomènes de coarticulation et/ou d'hypo-articulation. Ces phénomènes détériorent la production d'un message hypothétique, qui aurait une forme canonique de citation, à rapprocher avec une forme écrite. Le comportement attribué aux auditeurs est de restaurer cette forme canonique, en normalisant ces perturbations. Au contraire, l'approche interactionniste considère que la variabilité est structurante. Elle voit la parole conversationnelle comme une parole riche, première, dont la parole lue est une réduction, un appauvrissement. Les régularités interactionnelles

mises en évidence ne sont pas des béquilles que les locuteurs exploiteraient pour résoudre des problèmes de décodage d'une séquence d'éléments lexicaux composant le message linguistique, mais constituent au contraire la substance même de l'échange communicationnel.

Comme souligné par Ford et Couper-Kuhlen (2004), la dimension phonétique a été centrale dans l'analyse conversationnelle, et ce depuis le début, comme le témoigne le système de transcription de Gail Jefferson développé dans les années 70 (mais voir des présentations actualisées dans Jefferson (2004); Atkinson et Heritage (2006)) qui porte une attention particulière à tous les détails audibles de l'interaction. Mais l'importance du « design phonétique » de la parole en interaction a trouvé ses lettres de noblesse dans les travaux de phonéticiens à l'université de York à travers l'ouvrage *Doing Phonology* de Local et Kelly (1989) qui pose les bases d'une phonologie de la conversation. En adoptant une perspective inspirée de la linguistique Firthienne, ils proposent une analyse paramétrique, dynamique et relative de la substance phonétique. Eux aussi défendent une approche qui ne délaisse aucun niveau d'analyse : ils préconisent dans une première analyse une écoute impressionniste attentive aux moindres détails, et à différents ordres : détails phonétiques, setting articulatoire, résonances, tempo, intensité perçue ; variabilité, co-occurrence, phase des paramètres. Seulement en deuxième lieu vient l'interprétation, qui consiste à dégager empiriquement des tâches interactives que les participants pourraient être en train d'exécuter. Cette approche contraste avec les analyses phonémiques, qui selon eux sont des systèmes développés pour modéliser la prose écrite et supporter une notation orthographique. Ces systèmes sont ainsi mal équipés et biaisent l'analyse de la parole conversationnelle, en s'accompagnant d'hypothèses non garanties, comme la segmentabilité de la parole, le fait qu'elle présente peu de chevauchements ou encore que la coarticulation est due au contexte phonétique immédiat. Cette approche a permis à Local et ses collègues de mettre en évidence qu'un nombre jusqu'alors insoupçonné de détails phonétiques, traditionnellement considérés comme de la variation indésirable, ou simplement non remarqués, existent et sont structurés dans les conversations. Ces détails phonétiques incluent des changements de registre mélodique et d'intensité perçue, des convergences de hauteur de voix et d'intensité perçue, des pauses de tenue glottale, ou encore l'aspiration des plosives. Il est important de noter que dans tous les cas, ces régularités phonétiques sont associées à des activités interactives, comme la délimitation de tour de parole, l'occupation de la scène, la retenue du tour de parole, etc. De plus, ces études montrent à chaque fois que ces « exposants phonétiques » associés à des tâches interactives sont utilisés par les interactants : leur comportement montre une orientation vers ces détails fins.

2.2.2 Importance de la situation

Un aspect important des études menées en analyse conversationnelle est qu'elles ont toutes en commun de montrer que les associations mises en évidence entre détails phonétiques et activités interactives sont à chaque fois spécifiques à la langue ou au dialecte, et à la situation. Par exemple, si le contour mélodique final semble être l'indice principal utilisé par les locuteurs de l'anglais dans le signalement de la fin de tour de parole (au moins pour des variétés non standard, voir Reed (2004)), ce sont des indices appartenant à un registre tout différent, de qualité de la voix, qui semblent jouer ce rôle-là de façon prépondérante en finnois et en japonais (Ogden, 2004; Tanaka, 2004). Ceci illustre bien la démarche de l'analyse conversationnelle : plutôt que de chercher des universaux phonétiques associés à des tâches interactives, les chercheurs reconnaissent dès le départ qu'il est illusoire et même contre productif de chercher à calquer un modèle avec des catégories pre-supposées sur des comportements, qu'ils soient interactionnels ou autre. Au contraire, l'accent est porté sur la reconnaissance que l'innovation et l'improvisation sont le propre de l'échange communicationnel. C'est une approche qui peut paraître difficile à adopter, car l'idée est d'engager une recherche sans certitude de la forme des résultats qui peuvent être trouvés. Mais l'objectif du chercheur est en réalité bien précis, et consiste à identifier les régularités qui se manifestent dans la forme sonore de la parole, parfois à des niveaux temporel et/ou acoustiques inattendus. Par exemple, une différence systématique dans le degré de frication (entre autres paramètres phonétiques) de la dentale dans la séquence *"I think"* en anglais a été mise en évidence par Local (2003) selon que cette séquence portait une fonction lexicale (frication non voisée) ou interactionnelle (pas de frication). Dans ce rôle de marqueur de discours, cette caractéristique phonétique est indissociable d'une fenêtre temporelle concernant la parole produite précédemment relativement longue (une dizaine de syllabes), et est donc à analyser dans un domaine temporel qui dépasse le segment.

L'aspect important de cette démarche est que toute régularité découverte dans la substance phonétique de la parole est potentiellement informative pour la compréhension de la production, perception et représentation de la parole. On voit ainsi se profiler des systèmes phonologiques émergents, propres à des contextes spécifiques, qui de par leur nature éphémère et/ou trop locale, n'ont pas pu trouver des descriptions et des prises en compte au même niveau que le système phonologique décrivant les contrastes phonémiques d'une langue par exemple, car bien plus facilement généralisables. Pourtant, ces systèmes d'association entre régularités phonétiques et catégories de sens (qui ne sont donc pas typologiquement différents des systèmes

phonologiques phonémiques) ont été établis comme porteurs de sens pour les interactants, au moins au même titre d'importance que les systèmes de contrastes lexicaux : les interactions homme-machines sont encore bien loin de pouvoir approcher la fluidité et, de façon concomitante, la richesse de sens présentée par les interactions humaines.

2.2.3 Insuffisance d'une approche lexicale

John Local soutient ainsi que « le sens est bien plus que le sens lexical » (Local, 2003, p. 322). En effet, un dialogue est composé de bien plus qu'un enchaînement de phrases et de mots porteurs de sens lexical. En pratique d'ailleurs, le décodage lexical précis des enchaînements de parole est d'un intérêt bien secondaire pour les locuteurs dans leurs interactions quotidiennes, et la forme prononcée qui permet sa distinctivité phonologique par rapport à d'autres éléments du lexique diverge en général notoirement de la forme de citation, telle qu'elle serait obtenue par une lecture du mot isolé. Les buts véritables des interlocuteurs sont naturellement bien différents du simple décodage lexical de leurs échanges, qui n'en sont qu'un support. Les buts poursuivis par les interactants dans les interactions quotidiennes qui forment la majorité des échanges sont multiples et simultanés, et incluent généralement un positionnement social : l'adolescent qui répond en marmonnant à une question cherche plus à transmettre une attitude qu'à transmettre une séquence de phonèmes bien formés mais est néanmoins bien compris (précisément pas de façon exclusivement lexicale...) par son interlocuteur. Même dans une situation interactionnelle simple qui est de demander un rayon dans un supermarché, et qui pourrait être considérée comme un échange dont le succès dépend essentiellement d'une communication lexicale réussie, de nombreuses informations « extra lexicales » sont fournies et utilisées par les interactants.

Pour Pickering et Garrod (2004), ce sont précisément ces variations, considérées traditionnellement comme indésirables, déviances par rapport à une norme supposée, qui permettent qu'un dialogue soit si facile, et en particulier plus facile (à préparer et à comprendre) qu'un monologue. Ils postulent l'existence de mécanismes d'alignement à plusieurs niveaux linguistiques, dont l'établissement est interdépendant, et qui permettent aux interlocuteurs de simplifier énormément les efforts de compréhension de l'autre en partageant les mêmes représentations linguistiques. Ces représentations partagées ont en grande partie une origine dans les représentations *préexistantes* à l'interaction, comme par exemple un lexique commun, un système phonologique commun, une connaissance de la grammaire et des ressources pour gérer l'interaction commune, qui fournissent en quelques sorte les conditions nécessaires et initiales de l'interaction. Mais ce qui est remarquable de constater

est que nombre de représentations émergent localement ne préexistent pas à l'interaction, et sont inconnues des interlocuteurs avant celle-ci. Il semblerait que les locuteurs aient potentiellement une grande facilité à co-construire ces représentations émergentes, qui permettent d'évaluer le succès d'une interaction, et qui, de par leur aspect imprévisible, posent des problèmes aux modèles traditionnels. Ces représentations émergentes concernent tous les niveaux linguistiques de l'échange, comme la réutilisation de la structure syntaxique lors d'une réponse (Branigan *et al.*, 2000), le choix et la répétition de mots lexicaux (Tannen, 1989), et, pour le niveau sémantique, l'établissement spontané de l'utilisation d'un mot ou groupe de mots comme syntagme pendant la seule durée de l'interaction. Pickering et Garrod (2004) présentent un exemple de l'étude de Garrod et Anderson (1987) qui montre qu'un sens différent était attribué au mot *row* ('ligne') pendant des sessions de jeu interactif de résolution de labyrinthe. Ce mot pouvait, selon les paires d'interactants, faire soit référence à un ensemble ordonné (ex : 1ere ligne) ou non ordonné (ex : ligne du dessus), mais une fois qu'il était associé à un de ces deux sens, celui-ci restait fixé pendant toute la durée de l'interaction. Pour le niveau phonétique, l'étude de Pardo (2006) a recueilli le jugement de similarité perçue entre les formes prononcées de mot-cibles par deux locuteurs au cours d'une tâche interactive. Le degré de similarité perçue augmentait au fil de l'interaction, et persistait après sa conclusion. Si les indices phonétiques utilisés par les auditeurs ne sont pas explicités, ce résultat montre également que le niveau phonétique pourrait participer à ces mécanismes d'alignement entre les participants d'une interaction.

2.2.4 Validité écologique

Les chercheurs en analyse conversationnelle attachent une grande importance à utiliser des données naturelles, collectées dans des situations interactionnelles réelles et dont les objectifs d'analyse n'influencent pas les tâches interactionnelles dans lesquelles les locuteurs sont engagés. Une des raisons à cela est probablement que l'objectif de ces derniers est précisément de découvrir des tâches interactionnelles dans toute la contingence et la richesse de l'échange, et de mettre en évidence des corrélats réguliers. Concernant la dimension phonétique, ces corrélats incluent la qualité de voyelle, le setting articulatoire, la qualité de voix, la hauteur de voix, le tempo, etc.

Cette approche « non motivée » conduit les chercheurs à imaginer des situations expérimentales très variées, comme l'enregistrement des interactions entre un pilote d'avion et la tour de contrôle (Auvinen, 2009), les échanges dans un bloc opératoire (Mondada, 2003) ou des conversations libres. Cette diversité, qui contraste avec les tâches traditionnellement utilisées encore au-

jourd'hui comme la lecture de liste de mots ou de phrases isolées pour examiner des phénomènes de coarticulation ou de contours intonatifs par exemple, n'implique pas pour autant que ces situations expérimentales ne présentent pas un caractère systématique. Ces types de tâche ont toutes en commun d'avoir une pertinence interactionnelle pour les locuteurs. Cette question a été centrale à la réflexion en ethnométhodologie sous le terme de *accountability*, notamment pour les situations interactionnelles sur le lieu de travail. Elle s'intéresse à la « possibilité pour les co-participants de reconnaître mutuellement le sens de leurs actions » (Mondada, 2006, p. 12). La validité écologique semble être ainsi vérifiée lorsque les locuteurs comprennent et s'impliquent dans une tâche qui a une pertinence interactionnelle pour eux : renforcer le lien social au cours d'une conversation téléphonique, conduire avec succès un entretien d'embauche, etc. L'objectif du chercheur est alors de mettre en évidence des corrélats (phonétiques par exemple) associés à des tâches interactionnelles secondaires, pratiques, et éventuellement inconscientes du côté des participants qui servent la réussite de la tâche principale. Par exemple, dans les entretiens dialectaux sur l'anglais de Tyneside, Kelly et Local (1989) montrent que lorsqu'un interviewé répète un mot, il exécute l'une des trois tâches interactives suivantes : reconnaissance, demande de confirmation, appropriation. Ces trois tâches sont signalées par un ensemble de caractéristiques phonétiques, qui entraînent une réaction de la part de l'interviewer (prise d'acte de la reconnaissance, répétition, retenue des questions)

On peut donc opérer une distinction entre deux niveaux de tâches interactionnelles : les unes constituent les objectifs principaux de la communication, elles sont conscientes et exprimables par les interlocuteurs (comme réussir un entretien d'embauche), les autres sont des tâches accessoires, pratiques, qui servent la tâche principale, et qui sont l'objet des chercheurs en analyse conversationnelle. Ces tâches présentent un intérêt car elles ne sont pas forcément conscientes chez les interactants, bien qu'ils présentent – et c'est un aspect essentiel des résultats des études dans ce domaine – une orientation comportementale vers ces tâches.

C'est comme cela que nous comprenons la validité écologique défendue par les interactionnistes : si la tâche proposée par l'expérimentateur ne revêt pas une pertinence interactionnelle aux yeux des participants, il y a peu de chances de voir ceux-ci s'engager dans des tâches accessoires pour résoudre la tâche principale. Or la parole est dans la plupart des cas produite dans des contextes qui comportent une pertinence interactionnelle globale, elle est en réalité l'outil qui permet d'accomplir l'interaction. Il faut donc écarter de ces cas les situations d'enregistrement au cours d'expérience de laboratoire, ou les situations d'enseignement de la langue qui utilisent des méthodes traditionnelles. Il est d'ailleurs à noter que des méthodologies actuelles en classe

de langue essaient de proposer aux apprenants des activités interactives avec de réels enjeux interactionnels, des jeux collaboratifs, par exemple, acceptant ainsi l'idée d'un apprentissage de la langue comme un processus foncièrement sociocognitif (Doehler, 2010). Si on se donne la parole comme objet d'étude, il apparaît donc intéressant de ne pas écarter le contexte interactionnel (« site d'occurrence premier » selon J. Local) indissociable de sa réalisation.

Il est donc possible d'imaginer des tâches qui présentent un intérêt interactionnel pour les locuteurs, et qui permettent de contrôler finement le matériel sonore qui sera prononcé par les participants. Un paradigme souvent utilisé qui combine ces deux contraintes est le jeu, puisqu'il permet aux participants de se divertir à partir d'éléments qui sont propres au jeu : le matériel (noms des rues du Monopoly par exemple), les règles du jeu, et aussi un objectif à atteindre. En manipulant ces éléments, préexistants à l'interaction, on peut donc apporter le degré de contrôle nécessaire à l'obtention de productions sonores voulues tout en maintenant un intérêt interactionnel chez les participants, et permettant ainsi l'émergence de tâches interactives accessoires comme l'occupation de la scène de parole, la réalisation phonétique du matériel du jeu lorsqu'il est prononcé pour la première fois, etc.

2.2.5 La parole en interaction comme cadre privilégié pour l'étude des représentations mentales

Nous avons ici un objectif différent de celui de mettre au jour des tâches interactionnelles à travers leur relation avec des régularités phonétiques : nous cherchons à expliciter la dynamique (formation, évolution, actualisation) des représentations mentales des sons de la parole. Nous prenons acte en revanche que la parole en interaction, sous la condition de validité écologique énoncée ci-dessus, permet de faire apparaître des tâches interactives accessoires, qui, du fait de leur caractère improvisé notamment, forment l'essence du langage. Si nous voulons observer la forme sonore du langage de façon non biaisée, nous avons donc plus de chance de le faire dans cette situation.

Le choix de la parole en interaction pour observer les représentations mentales va bien sûr au-delà d'une simple observation des spécificités de la parole spontanée par rapport à sa forme de citation, qui est l'objet des études s'intéressant à la parole connectée, comme la coarticulation.

L'approche interactionnelle a montré l'existence de systèmes de sens autres que ceux traditionnellement posés et indépendants que seraient la phonologie, la syntaxe ou le lexique, et que les locuteurs combineraient à loisir pour communiquer. Ces systèmes de sens, que l'on peut regrouper sous le terme d'interactionnels, émergent localement et utilisent les différentes modalités des

systèmes traditionnellement posés (tournure syntaxique particulière combinée avec un faisceau de traits phonétiques pour signaler la fin de tour par ex.). En montrant leur interdépendance et en décloisonnant la formalisation de ces niveaux (des caractéristiques acoustiques traditionnellement associées à un « segment » se retrouvent aussi loin que 5 syllabes du segment en question), cette approche permet aussi de se libérer du problème posé par l'omniprésence de la variabilité de la parole : elle est considérée comme structurante, plutôt qu'indésirable. La parole ainsi libérée peut dorénavant se laisser modéliser dans toute sa richesse.

Le langage se crée et évolue dans l'interaction. Les représentations mentales de la parole ne sont pas totalement préexistantes à l'interaction, et sont mises au jour, modifiées, actualisées lors de l'interaction. Observer ces changements n'est pas quelque chose de temporaire, localement déviant, et à écarter dans la tentative d'accéder aux « vraies » représentations mentales des sujets ; au contraire ce sont ces changements, cette non rigidité, qui constituent à notre sens la nature réelle des représentations mentales. De plus, elles sont observables publiquement dans les échanges, à travers l'enregistrement sonore. La quête même de vouloir accéder aux représentations mentales enfermées de façon présumée dans le cerveau de l'individu apparaît comme mal posée, puisque cette nature dépend des contingences de l'interaction, comme l'influence de l'interlocuteur, partenaire ponctuel de l'interaction. Paradoxalement, pour bien comprendre les mécanismes de traitement de la parole d'un locuteur (perception, production, représentation), il est nécessaire de considérer l'environnement et en particulier son partenaire.

Une approche qui consisterait à enregistrer individuellement des locuteurs (productions orales ou réponses de perception) pour accéder « plus directement » aux représentations mentales apparaît ainsi comme biaisée, et limitée dans ses résultats possibles. En effet, si les résultats sont *a fortiori* plus réguliers, moins variables, car les contextes extérieurs ont été volontairement écartés, cette approche ne garantit pas de mettre au jour les « vraies » natures du traitement de la parole : au mieux, le locuteur exécute correctement l'exercice et produit des réponses régulières. Mais ces réponses, par leur conservatisme, ne pourront pas refléter les mécanismes de formation, et constituent en définitive un échantillon artificiel, peu représentatif de l'activité langagière majoritaire des sujets. L'enregistrement individuel est plutôt à considérer comme un cadre inhibant pour l'observation des représentations mentales de la parole.

La parole en interaction apparaît donc comme un cadre privilégié pour l'observation des représentations mentales, en ce qu'elle permet de favoriser l'innovation, propre des conditions naturelles et majoritaires de production de la parole. Dans ce cadre, la variabilité n'est pas un problème mais une

solution pour la compréhension des mécanismes de traitement de la parole.

Chapitre 3

Méthode

3.1 Introduction

Nous présentons dans ce chapitre la méthode utilisée pour explorer l'évolution des représentations mentales associées aux sons de la parole en interaction conversationnelle. Nous présentons une tâche interactive qui permet de recueillir les réalisations répétées de nouveaux mots par deux interactants, locuteurs de deux variétés de français parlé différentes, le français standard et méridional respectivement. Les mots présentent des caractéristiques phonologiques finement contrôlées qui illustrent les différences existant entre ces deux variétés, et se présentent comme des nouveaux exemplaires que les locuteurs vont être amenés à intégrer dans leur lexique, du moins pendant la durée de l'interaction.

Nous commençons par présenter le niveau de la variation phonologique régionale, un niveau d'analyse privilégié pour l'exploration de la formation et de l'évolution des exemplaires. Nous détaillons ensuite les principales caractéristiques phonétiques et phonologiques qui différencient les deux variétés de français parlé, que nous regroupons suivant 5 dimensions phonologiques. Nous présentons ensuite la construction des mots, stimuli sur lesquels sont distribuées les différents segments associés aux variables phonologiques. Nous passons ensuite à la description de la tâche interactive développée. La collection du corpus ainsi que son alignement est ensuite détaillée, et nous finissons par quelques données chiffrées sur le corpus.

3.2 Variation phonologique régionale

Nous présentons dans cette section la pertinence du niveau de la variation phonologique régionale pour l'étude de l'évolution des représentations men-

tales, ainsi que les caractéristiques phonétiques et phonologiques des deux variétés de français parlé sous étude, à savoir le français standard et le français méridional (FS et FM ci-après).

Traditionnellement considérée comme une dimension séparée des informations purement linguistiques dans les modèles de traitement de la parole, la variation phonologique régionale commence à trouver sa place, en parallèle avec les informations indexicales caractérisant le locuteur, dans des modèles intégrés de traitement de la parole, grâce aux apports de la sociophonétique (voir par ex. Docherty, 2007). Si de nombreuses études ont été conduites sur le traitement de la parole comportant un accent étranger, une attention grandissante a aussi été apportée dans les dernières années à l'impact que la variation phonologique interne à la langue du locuteur pourrait avoir sur la communication parlée (Brunellière *et al.*, 2009; Clopper et Bradlow, 2008; Conrey *et al.*, 2005; Cutler *et al.*, 2005; Delvaux et Soquet, 2007; Dufour *et al.*, 2007; Evans et Iverson, 2004; Floccia *et al.*, 2006; Hay *et al.*, 2006; Kraljic *et al.*, 2008; Sumner et Samuel, 2009). Par exemple, le répertoire phonémique peut être différent entre deux variétés régionales d'une langue, ce qui peut rendre plus difficile pour un locuteur d'une variété de reconnaître les mots d'une variété dont il n'est pas un locuteur natif. Dans un récent travail sur le français, Dufour *et al.* (2007) ont considéré les différences qui existent entre le système vocalique du français standard et celui du français méridional, et ont montré que ces différences affectaient de manière significative la façon dont les mots parlés étaient identifiés par les locuteurs des deux accents. Floccia *et al.* (2006) ont mis en évidence un coût initial temporaire de traitement associé avec la reconnaissance de mots parlés dans un accent régional non familier. Delvaux et Soquet (2007) ont examiné l'effet que l'exposition à un accent non-natif peut avoir sur la production de la parole en français parlé en Belgique, et ont trouvé des patrons d'imitation phonétique entre les accents. Des recherches plus approfondies doivent cependant être conduites pour mettre au jour pleinement les mécanismes qui permettent de caractériser la facilité avec laquelle la communication parlée peut se produire entre des accents régionaux différents.

3.2.1 Accents FS et FM

Le français parlé présente une variation phonologique qui est traditionnellement décrite en terme de variation géographique (variation diatopique) plutôt que sociale (variation diastratique) (Walter, 1988; Woehrling, 2009). Ceci est probablement à mettre en relation avec la disparition abrupte et récente des dialectes ou « patois » du français, due à la politique linguistique de la fin du XIXe siècle qui visait à affaiblir les langues régionales, et qui a conduit

naturellement les générations futures, tant au niveau politique par la mise en place de mesures visant à sauvegarder les langues régionales qu'au niveau des locuteurs de la langue et des chercheurs du domaine à se focaliser sur cette dimension (Martinet (1945); Carton *et al.* (1983); Walter (1982), voir aussi Coveney (2001), p. 4).

Pourtant, la variation sociale existe bien en français, comme le montre Gadet (2003) avec un ouvrage dédié à cette question, et les nombreuses études menées sur le français des banlieues (Binisti et Gasquet-Cyrus, 2003; Jamin *et al.*, 2006; Trimaille, 2008; Fagyal, 2010).

Des études actuelles font apparaître un nivellement des accents en France, de façon similaire à ce qui est observé en Grande-Bretagne (Armstrong, 2002; Foulkes et Docherty, 1999; Kerswill, 2003). Dans une étude sur les variétés parlées de Rennes et de Nancy, Boughton (2005) trouve ainsi que les différences régionales marquées tendent à disparaître, et que les locuteurs sont mieux caractérisés par des différences attribuées à une variation sociale. Ces différences ont été confirmées par une étude de perception (Boughton, 2007), dans laquelle des confusions entre les régions étaient facilement faites. Cette confusion perceptive a été également observée dans l'étude de Woehrling (2009) considérant des échantillons de parole relevés en cinq points du territoire français, tirés du corpus « *Phonologie du Français Contemporain* » (PFC) (Durand *et al.*, 2003a). La matrice de confusion de l'identification perceptive de l'origine géographique met une évidence une bi-partition Nord–Sud (correspondant aux variétés FS et FM décrites dans ce travail), avec une plus grande proportion de confusion entre variétés à l'intérieur de ces catégories qu'entre ces catégories.

Dimensions phonologiques

Les variétés standard et méridionales du français parlé présentent des différences bien établies aux niveau phonologique, phonétique et prosodique (voir entre autres Martinet, 1945; Carton *et al.*, 1983; Durand, 1990; Coveney, 2001; Durand et Lyche, 2004; Coquillon, 2005; Eychenne, 2006). Nous présentons ci-dessous 5 dimensions phonologiques illustrant les principales différences au niveau phonétique/phonologique décrites dans la littérature et qui font l'objet de notre étude. Les notations en italique sont utilisées comme identifiants de ces dimensions dans la suite de ce travail.

Schwa (*Schw.*) Les schwas en position finale de mot sont plus susceptibles d'être réalisés en FM qu'en FS. En position interne de mot, entre deux consonnes simples, les schwas devraient être plus fréquemment réalisés en FM qu'en FS (Eychenne, 2006). On s'attend également à une fréquence de

réalisation plus élevée en FM lorsqu'ils correspondent à un 'e' dans la forme écrite des mots (Durand *et al.*, 2003b).

Voyelles moyennes postérieures (*Post.*) Les voyelles moyennes postérieures tendent à être antériorisées en FS (Boula de Mareüil *et al.*, 2008; Coveney, 2001; Fonagy, 1989; Martinet, 1958). C'est d'ailleurs le trait phonétique qui semble le mieux différencier les variétés FS et FM dans des caractérisations automatiques à grande échelle (Boula de Mareüil *et al.*, 2008).

Voyelles moyennes en syllabe finale de mot (*Moy.*) Alors que des distinctions contrastives existent en FS entre /e/–/ɛ/, /ø/–/œ/, /o/–/ɔ/, la distribution des variantes mi-hautes et mi-basses en FM est déterminée entièrement par une variante de la *loi de position* : l'allophone mi-haut se produit en syllabe ouverte et l'allophone mi-bas en syllabe fermée et lorsque la syllabe suivant contient un schwa (Durand, 1990).

Plosives coronales (*Cor.*) Les plosives coronales tendent à être produites comme des affriquées postalvéolaires devant les voyelles hautes en FM, particulièrement dans la variété parlée à Marseille (un phénomène caractérisé comme une palatalisation dans Binisti et Gasquet-Cyrus (2003)). Cette réalisation particulière est traditionnellement associée avec des locuteurs de classe sociale inférieure, ou des locuteurs issus de l'immigration, mais s'étendrait actuellement aux classes sociales plus élevées (Trimaille, 2008).

Voyelles nasales (*Nas.*) Les voyelles nasales en FS correspondraient à des séquences $V + N$ en FM (Durand, 1988). Le FS ne ferait plus la distinction entre /ɛ̃/ et /œ̃/ (Martinet, 1945; Malécot et Lindsay, 1976; Walter, 1976; Fagyal, 2006). Dans cette variété, les voyelles nasales seraient également engagées dans un changement en chaîne : /ɛ̃/→ /ɑ̃/ ; /ɑ̃/→ /ɔ̃/ et /ɔ̃/→ /õ/ (Hansen, 2001).

3.3 Stimuli

Cette section décrit les étapes de construction des stimuli qui serviront de support à la caractérisation acoustique de l'évolution des représentations mentales. Ils se présentent sous la forme de pseudo-mots, et sont présentés aux participants comme des noms propres. On peut considérer ces mot-cibles comme une collection de nouveaux exemplaires, à la fois au niveau du mot et à des niveaux inférieurs comme le phonème, qui seront répétés et dans

une certaine mesure mémorisés par les participants au cours de la tâche. Un autre avantage à l'utilisation de pseudo-mots est que l'on peut s'attendre à une réalisation moins réduite par rapport à des mots appartenant au lexique (Clopper et Pierrehumbert, 2008). Nous évitons par ailleurs la nécessité de contrôler la fréquence lexicale des mots, qui a une influence sur la variation de prononciation (Bybee, 2002). Ces mot-cibles sont construits de manière à incorporer les variables phonologiques décrites à la section 3.2.1 qui différencient les variétés FS et FM, tout en respectant les contraintes phonotactiques du français. Ce dernier point permet de donner une apparence familière aux mot-cibles pour des locuteurs du français, et favorise ainsi une prononciation homogène et aisée de ceux-ci.

3.3.1 Variables phonologiques

Les variables phonologiques sont tirées des cinq dimensions phonologiques décrites dans la section 3.2.1. Elles sont insérées dans les stimuli et alternent avec des segments qui ne sont pas analysés, mais qui sont choisis pour assurer un bon degré d'approximation du stimulus avec la phonotactique du français. On utilisera le terme de *segment critique* pour se référer aux variables phonologiques incluses dans la séquence de segments constituant le mot, par opposition aux segments non analysés.

Les variables phonologiques ont été choisies pour décrire complètement la dimension phonologique dont ils sont tirés. Elles doivent donc toutes partager les mêmes caractéristiques qui différencient ces dimensions phonologiques, afin de les regrouper sous le même phénomène et d'obtenir une même mesure pour les analyses subséquentes. Effectuer ces regroupements revient à faire l'hypothèse que les caractéristiques qui différencient les accents FS et FM sur les dimensions phonologiques identifiées sont bien des caractéristiques phonologiques : elle s'appliquent uniformément aux différents segments critiques qui composent les catégories phonologiques. Par exemple, si cette cohérence phonologique se vérifie, nous serons en mesure d'évaluer le degré de palatalisation/affrication pour un locuteur des plosives coronales devant les voyelles hautes, qu'il s'agisse de la plosive voisée ou non voisée, précédant la voyelle arrondie ou non-arrondie. Bien évidemment la définition de cette mesure unique de la dimension phonologique est sujette d'une part à la cohérence phonologique, et d'autre part à la pertinence de la mesure obtenue pour les différentes sous-dimensions. Les différents segments critiques qui composent les 5 dimensions phonologiques sont détaillés dans la table 3.1.

La dimension phonologique *Moy.* diffère des autres dimensions phonologiques dans la mesure où elle décrit une différence distributionnelle des variantes entre les variétés FS et FM. Les deux variétés possèdent les deux

DP	SC	Description	Exemple	Prononciation attendue	
				FM	FS
		Schwa			
Schw.	S1	non final, graphié	*Corre<u>f</u>ère*	[ə]	[ ]
	S2	final, graphié, après consonne voisée	*Botonn<u>e</u>*	[ə]	[ ]
	S3	final, graphié, après consonne non voisée	*Adurauqu<u>e</u>*	[ə]	[ ]
	S4	final, non graphié, après consonne voisée	*Jeanbri<u>l</u>*	[ə]	[ ]
	S5	final, non graphié, après consonne non voisée	*Dévo<u>c</u>*	[ə]	[ ]
		Voyelles moyennes postérieures			
Post.	B1	en syllabe ouverte	*L<u>o</u>dini*	[o]	[o̞]
	B2	en syllabe fermée	*Vic<u>o</u>lfi*	[ɔ]	[ɔ̟]
		Voyelles moyennes			
Moy.	M1	antérieure non arrondie en syllabe ouverte, graphiée 'é'	*Noturé*	[e]	[e]
	M2	antérieure non arrondie en syllabe ouverte, graphiée 'ais'	*Pandur<u>ais</u>*	[e]	[ɛ]
	M3	postérieure en syllabe fermée, graphiée 'o'	*Cont<u>o</u>r*	[ɔ]	[ɔ]
	M4	postérieure en syllabe fermée, graphiée 'au'	*Samb<u>au</u>le*	[ɔ]	[o]
		Plosives coronales suivies d'une voyelle haute			
Cor.	C1	séquence /ti/	*Ou<u>t</u>imil*	[tʃi]	[ti]
	C2	séquence /ty/	*Ma<u>t</u>uca*	[tʃy]	[ty]
	C3	séquence /di/	*A<u>d</u>inac*	[dʒi]	[di]
	C4	séquence /dy/	*In<u>d</u>ucar*	[dʒy]	[dy]
		Voyelles nasales			
Nas.	N1	antérieure non arrondie	*Ollevi<u>n</u>té*	[ɛN]	[ɛ̃]
	N2	antérieure arrondie	*D<u>un</u>duco*	[œN]	[ɛ̃]
	N3	ouverte	*S<u>an</u>tère*	[aN]	[ɑ̃]
	N4	postérieure	*F<u>on</u>dula*	[ɔN]	[ɔ̃]

TABLE 3.1 – Segments critiques (SC) associés avec les cinq dimensions phonologiques (DP). Pour chaque segment critique, la transcription orthographique d'un mot-cible est donnée en exemple (la séquence de lettres soulignée correspond au segment critique). La prononciation attendue du segment critique est données pour les accents FS et FM. N représente une consonne nasale de même lieu d'articulation que la consonne suivante dans le mot. Elle est remplacée par [n] devant [t] and [d], par [m] devant [p] and [b] et par [ŋ] devant [k] et [g].

allophones mi-haut et mi-bas, mais le choix de l'un ou l'autre dépend de la structure syllabique. Les segments critiques M1 et M3 concernent des structures syllabiques pour lesquelles la même variante de prononciation est attendue dans les deux variétés, et fournit donc une condition de contrôle pour la réalisation de ces catégories phonétiques.

Par ailleurs, les segments critiques M3 et M4 inclus dans la dimension phonologique *Moy.* caractérisant les différences de distribution entre les deux variétés peuvent également être associés à la dimension phonologique *Post.*, et décrire les différences d'antériorisation associée à cette dimension entre les deux accents.

3.3.2 Génération des stimuli

La méthode utilisée pour la construction des pseudo-mots est celle des grammaires probabilistes. Cette approche permet de construire des pseudo-mots qui respectent les règles phonotactiques du français. Ces pseudo-mots sont composés par des fragments de mots appartenant à une base de données d'apprentissage qui modélise la langue, agencés de manière similaire à celle observée dans la base d'apprentissage.

Une grammaire probabiliste est une sous-classe des modèles de Markov cachées (Hidden Markov Models, HMM, voir Rabiner (1989)). Elle associe à une séquence de symboles sur un alphabet une probabilité qui est le produit des probabilités conditionnelles de chaque symbole étant donnée la probabilité de la sous-séquence de symboles précédent. La longueur de la sous-séquence est un paramètre qui définit l'ordre du modèle ; les grammaires n-gramme considèrent les sous-séquences de longueur $n-1$. Ainsi, dans un modèle trigramme, la probabilité de la séquence

$$S = \sigma_1\sigma_2\sigma_3\sigma_4$$

où les σ_i sont des symboles de l'alphabet Σ, est calculée par

$$P(S) = \pi_1 * \pi_2 * \pi_3 * \pi_4 * \pi_5$$

où $\pi_1 = p(\sigma_1|\varnothing\varnothing)$, $\pi_2 = p(\sigma_2|\varnothing\sigma_1)$, $\pi_3 = p(\sigma_3|\sigma_1\sigma_2)$, $\ldots$, $\pi_5 = p(\varnothing|\sigma_3\sigma_4)$. $\varnothing$ est l'état initial et final de la séquence, et correspond au début et à la fin du mot.

Les probabilités de transition sont calculées en combinant les occurrences des sous-séquences des mots de la base d'apprentissage, éventuellement pondérée par la fréquence des mots dans la base, si cette information est disponible. Par exemple, la probabilité de transition associée au début du mot 'tsarine' /tsarin/ $p(/s/|\varnothing/t/)$ sera vraisemblablement faible étant donné le

peu de mots du français qui commencent par /ts/ ('tsunami', 'tzigane', ...), tandis que la probabilité de transition associée à la sous-séquence /sar/, p(/r/ | /sa/) sera plus élevée car elle est plus représentée dans la langue française ('hussard', 'sarabande', ...), tout comme la fin du mot /in/ qui correspond à un grand nombre de mots du français notamment parce qu'elle est utilisée pour indiquer l'indication du genre féminin ('cousine', 'fine', ...).

Une fois la grammaire construite, elle peut être utilisée pour générer de nouvelles séquences, à partir de la matrice des probabilités de transition et de l'algorithme de Viterbi (voir Rabiner, 1989, par ex.). On peut par exemple générer les N séquences de n phonèmes de plus forte probabilité, qui seront composées par les sous-séquences les plus observées (et éventuellement des mots les plus fréquents) de la base d'apprentissage. Les nouvelles séquences présenteront les caractéristiques phonotactiques du français et notamment la structure syllabique, car cette information est implicitement capturée dans le calcul de la matrice des probabilités de transition dans la phase d'apprentissage.

Une méthode similaire a été employée pour l'anglais par Nye et Fowler (2003). En utilisant les tables de fréquences transitionnelles des phonèmes de l'anglais de différentes longueurs compilées par Hultzén *et al.* (1964), ces auteurs ont pu générer des séquences phonétiques qui varient dans leur degré d'approximation avec des séquences phonétiques appartenant à l'anglais.

Dans notre travail, l'inclusion des variables phonologiques a été effectuée de la façon suivante : deux occurrences de chaque variable phonologique (38 au total) ont été distribués semi-aléatoirement sur un ensemble de 16 séquences, pour donner des patrons constitués d'emplacement vides alternant avec des segments critiques. La longueur des séquences a été initialement fixée à 6 segments, qui est la longueur minimale pour inclure la totalité des segments critiques tout en garantissant une génération diversifiée et naturelle de stimuli.

Les emplacements vides seront remplis par la grammaire probabiliste, qui sélectionne les meilleurs candidats parmi l'ensemble des sons du français pour satisfaire le patron. Certaines contraintes ont été prises en compte dans cette répartition. Premièrement, l'occurrence d'une variable phonologique a été restreinte à l'intérieur du mot, à l'exception des schwas finaux et des voyelles moyennes en syllabe ouverte en position finale de mot. Ceci a été fait pour contrôler l'environnement segmental des variables phonologiques, le contexte segmental aux frontières des mots ne pouvant pas être contrôlé. Par ailleurs, une précaution a été prise pour éviter que la répartition des variables phonologiques ne donne lieu à des séquences voyelle-voyelle, car les mots ainsi générés par la grammaire probabiliste auraient de fortes chances d'avoir une probabilité très faible. Finalement, certains emplacements dans la

séquence ont été restreints à une classe de segments, dans le but d'équilibrer l'occurrence des segments critiques et des autres types de segments, et ainsi donner une forme plus diversifiée aux mots. Par exemple, parmi la moitié des 16 mots à générer se terminant par une voyelle – les mots de l'autre moitié se terminent par une consonne éventuellement suivie d'un schwa – la moitié d'entre eux se terminent par une voyelle moyenne antérieure. Nous avons ainsi restreint les segments possibles pour les 4 mots restant aux voyelles à l'exception des voyelles moyennes antérieures (voir table 3.2).

Le codage phonétique adopté pour l'écriture des patrons a été adapté d'après celui utilisé dans la base *lexique* [1], lui-même tiré du code SAMPA [2], qui permet de représenter les caractères phonétiques d'une forme manipulable simplement par l'ordinateur. Ce codage a été spécifié de manière à pouvoir représenter les sons du français ainsi que des classes de son évoquées ci-dessus sur un seul caractère, pour des raisons de simplicité dans l'écriture des scripts de construction des patrons et de génération des stimuli. Ce codage avec ses correspondances avec l'alphabet phonétique international (IPA) est donné en appendice, dans la table 6.2.

La table 3.2 montre le processus de génération pour une liste de 16 mots. Cette liste sera utilisée dans la première phase de la tâche interactive présentée à la section 3.4.

Nous avons utilisé trois grammaires probabilistes, construites sur 3 bases différentes. Ces 3 bases ont été choisies en raison du type de mot qu'elles contiennent, et par suite de la forme attendue des séquence générées : la forme attendue des mots générés à partir la base `DicoLPL` (VanRullen *et al.*, 2005) est celle des mots les plus courant du français. Nous nous sommes également intéressés à obtenir des mots qui puissent partager des caractéristiques avec des noms propres et des prénoms rencontrés en français, à travers l'utilisation de la base `prénoms` (New *et al.*, 2004) et une base construite par nos soins à partir d'une liste de noms propres. Les caractéristiques principales qui différencient ces bases sont présentées dans la table 3.3.

Les 16 patrons ainsi construits, composés de 2 à 3 variables phonologiques et d'emplacements vides ont été soumis aux 3 grammaires probabilistes. La sélection des stimuli a été faite en examinant les 10 séquences de probabilité maximale générées par chaque grammaire. Le degré d'approximation avec la phonotactique du français était donné par la probabilité de la séquence. Celle-ci a été contrôlée pour qu'elle appartienne à la plage de valeur supérieure des probabilités des mots de la base à partir de laquelle ils avaient été générés. Lorsque ce n'était pas le cas, ce qui était attribuable à la distribu-

[1] URL : www.lexique.org/outils/Manuel_Lexique.pdf, page consultée le 3 août 2010
[2] URL : www.phon.ucl.ac.uk/home/sampa, page consultée le 3 août 2010

Patron	Forme phonologique	Forme orthographique
.§di.e	sɔ̃dite	Sondité
.oty.e	notyʁe	Noturé
.1dy.E	lœ̃dyʁɛ	Lundurais
...5.E	bazɛ̃tɛ	Bazintais
..ty.X	matyka	Matuca
...O..X	vikɔlfi	Vicolfi
.1dy.X	dœ̃dyko	Dunduco
..§..X	leɔ̃stɑ̃	Léonstan
.@.°.oT°	dɑ̃təʁokə	Danterauque
.di.WT	adinak	Adinac
.@.oD°	sɑ̃bolə	Sambaule
.ti.WD	utimil	Outimil
.o.OD°	botɔnə	Botonne
.ti.WD	stimɛn	Stimen
.O.°.WT°	kɔʁəfɛʁə	Correfère
.5..OT	vɛ̃kjɔʁ	Vinquior

TABLE 3.2 – Génération d'une liste de 16 stimuli. A partir d'un patron contenant des segments critiques, une forme phonologique est générée par la grammaire probabiliste puis convertie sous forme orthographique. Le codage phonétique des patrons suit la convention adoptée par *lexique*. Des symboles supplémentaires ont été ajoutés pour spécifier des ensembles de segments (voir texte). La liste complète des stimuli est donnée en appendice, dans la table 6.1.

tion des variables phonologiques dans les séquences, le patron a été modifié manuellement par l'ajout d'un emplacement vide, et un nouvel ensemble de mots a été calculé pour ce patron. La densité des variables phonologiques à inclure dans les stimuli n'a pas permis de contrôler plus finement d'autres critères qui pourraient se révéler utiles dans la construction de pseudo-mots, comme leur voisinage phonologique. Ce critère n'a été pris en compte qu'intuitivement, en rejetant les mots qui partageaient des séquences trop longues avec des mots fréquents du français.

Finalement, nous avons converti ces formes phonologiques en formes orthographiques, en suivant les règles phonographiques du français (voir par ex. Blanche-Benveniste et Chervel, 1978, chap. 3). Il n'existe pas à notre connaissance de base de données des relations graphèmes–phonèmes pour le français. On peut trouver certaines listes, comme celle maintenue par S.

	DicoLPL	prénoms	noms
Nombre total d'observations	136000000	494000	747
Séquences de phonèmes uniques	75000	541	708

TABLE 3.3 – Caractéristiques des 3 bases utilisées pour la construction des pseudo-mots.

Lavoie [3], qui cherche plutôt à lister les différentes graphies associées aux phonèmes du français dans tout leur exotisme, mais ne fournit pas de fréquence de ces relations.

Il est à noter que le fait que les stimuli soient présentés sous forme orthographique aux participants entraîne que leur réalisation acoustique est, dans un premier temps en tout cas, le résultat de la conversion graphème-phonème par les locuteurs. Celle-ci peut être différente de celle attendue, et la réalisation acoustique ne pourra donc pas être considérée comme une réalisation phonétique de la variable phonologique attendue. Par exemple, il se peut que la voyelle nasale /ɛ̃/, présentée sous forme orthographique par le digramme 'in', puisse être interprétée par le locuteur comme la séquence des graphèmes 'i' et 'n' et réalisée phonétiquement [in]. De telles réalisations non-attendues sont cohérentes avec les règles phonographiques du français, et peuvent par ailleurs révéler des phénomènes intéressants de recherche de consensus entre les participants sur la prononciation adoptée pour les stimuli. Nous les avons cependant considérées comme erronées dans le cadre de notre étude et laissées de côté dans les analyses.

Cinq listes de mots ont été générées et sont jointes en appendice (table 6.1). En moyenne, les mots ont une longueur de 6.3 phonèmes (SD = 0.7) ou 2.7 syllabes (SD = 0.5). La table 3.4 détaille le nombre moyen d'occurrence des segments critiques par liste de 16 noms, sur chaque dimension phonologique. La prise en compte des nombreux critères et contraintes pour l'inclusion des variables phonologiques a donné lieu à quelques déséquilibres dans le matériel (segments critiques en moins ou en plus), dont on prévoit qu'ils seront atténués par la variabilité du nombre de répétition des mots par les locuteurs au cours de la tâche interactive.

[3] http://www.benoit-lavoie.ca/graphies/index_fr.html, page consultée le 3 août 2010.

	SC 1	SC 2	SC 3	SC 4	SC 5	total
Schw.	1.6 (0.9)	2.8 (0.4)	1 (0.7)	2 (0)	2.2 (0.4)	9.6 (0.9)
Post.	2.4 (0.5)	2 (0)	–	–	–	4.4 (0.5)
Moy.	2 (0)	2 (0)	1.8 (0.4)	2.2 (0.4)	–	8 (0)
Cor.	2.2 (0.4)	2 (0)	2 (0)	2 (0)	–	8.2 (0.4)
Nas.	2 (0)	2 (0)	2.2 (0.4)	2 (0)	–	8.2 (0.4)

TABLE 3.4 – Nombre d'occurrence moyen et déviation standard entre parenthèses de chaque segment critique (SC) pour les cinq dimensions phonologiques. Voir la table 3.1 pour l'intitulé des segments critiques. Le nombre total moyen pour chaque dimension phonologique est donné dans la colonne de droite.

3.4 GMUP : tâche interactive

L'objectif poursuivi par ce travail est d'examiner la forme sonore et l'évolution d'un ensemble de variables phonologiques produites par deux locuteurs engagés dans une interaction conversationnelle. Les variables phonologiques sont distribuées sur un ensemble de mots, que les locuteurs sont amenés à produire et à répéter au cours de l'interaction.

Nous attachons une importance particulière à ce que les mots soient prononcés en interaction conversationnelle, par opposition à un enregistrement de parole préparée, dans une tâche de lecture par exemple (voir section 2.2.5). La langue est développée par les êtres humains pour communiquer et « naviguer dans leurs actions sociales quotidiennes » Ford et Couper-Kuhlen (2004), et déploie donc toutes ses caractéristiques (phonétiques, prosodiques, syntaxiques, lexicales...) pour satisfaire ces buts d'interaction. En examinant de la parole préparée, on peut donc examiner au mieux une *réduction* des ressources et mécanismes disponibles aux locuteurs, c'est-à-dire principalement celles nécessaires à l'exécution de la tâche demandée. On a donc toutes les chances d'accéder à une version allégée et standardisée de la substance de la langue. La langue est par contre un système dynamique en évolution permanente, et l'observation de celle-ci dans son site premier d'occurrence, l'interaction sociale (Local, 2003), permet non seulement de libérer ces contraintes d'accès simple aux ressources, mais également de favoriser l'apparition des mécanismes mêmes qui président à la construction et l'évolution de la langue. Ces mécanismes incluent la formation de nouvelles catégories phonologiques. Même si la durée d'une interaction que l'on peut raisonnablement enregistrer dans le cadre d'une tâche interactive apparaît comme infinitésimale en comparaison avec la somme de toutes les interactions qui constituent le quotidien

de locuteurs et qui ont façonné le système phonologique de ceux-ci, nous nous sommes intéressés à découvrir les signes qui pourraient relever de ces mécanismes. Ainsi, nous considérons la parole spontanée produite en interaction conversationnelle comme le lieu privilégié pour observer la forme sonore des catégories phonologiques et de leur évolution qui contribue à définir la langue des locuteurs.

Parmi les protocoles existant pour analyser la parole conversationnelle on trouve les conversations conduites par un expérimentateur (Natale, 1975), les conversations entre deux interactants dont on impose le sujet de discussion (Bertrand *et al.*, 2008), les tâches de construction collaborative d'histoire en complétant des fragments de phrases (Kraljic *et al.*, 2008), ou encore la tâche développée par Bradlow *et al.* (2007), dans laquelle les participants doivent identifier les différences entre deux images. Le protocole le plus utilisé est probablement la Map Task (Anderson *et al.*, 1991), dans lequel les locuteurs doivent collaborer verbalement pour reproduire sur la carte d'un des participants le chemin dessiné sur celle de son partenaire. Sur la base de ces protocoles existants, nous avons développé une nouvelle tâche interactive permettant aux participants d'échanger librement autour de mots dont les caractéristiques phonologiques sont finement contrôlées.

La tâche interactive que nous avons développée s'intitule GMUP (pour "Group 'em up !", *Regroupez-les !*) qui permet d'obtenir une production répétée de pseudo-mots dans une interaction conversationnelle. GMUP prend la forme d'un jeu qui consiste à découvrir de manière interactive les relations qui existent entre des personnages présentés comme formant un réseau social. A chaque personnage est associé un nom qui est attribué par les participants, qui sont amenés à le prononcer un certain nombre de fois au cours de la tâche.

Les participants sont assis de part et d'autre d'une table, au milieu de laquelle un écran est placé qui empêche les participants de se voir. Chaque participant utilise un ordinateur pour jouer au jeu, et porte un microphone serre-tête de haute qualité.

GMUP est constitué de deux parties. Dans la première partie, les participants doivent attribuer ensemble un nom parmi une liste de 16 noms pré-établis à chacune des 16 photos qui représentent des portraits d'homme ou de femme. Les photos sont présentées tour à tour par groupe de deux, ainsi que trois noms tirés aléatoirement de la liste. A chaque tour, les participants doivent décider ensemble quel nom est le plus approprié à chacune des photos. Les photos et les noms sont les mêmes pour les deux participants, mais leur ordre est aléatoirement déterminé pour chaque participant. Ceux-ci doivent ainsi décrire verbalement les photos et prononcer les noms. Le nom qui n'est pas choisi est remis dans la liste et attribué plus tard à une autre photo. Cette première phase du jeu permet aux participants de s'engager

dans une collaboration ainsi que de se familiariser avec le matériel. Les participants s'approprient ainsi dans cette première phase le réseau social partagé des personnages.

Dans la deuxième partie, les participants sont informés que les 16 personnages sont divisés en 4 groupes de taille variable. La tâche assignée aux participants est d'identifier ces groupes et de déterminer l'appartenance de chacun des personnages à un des 4 groupes. L'information d'appartenance aux groupes est donnée par des déclarations écrites, disponibles pour certains des personnages. La déclaration concerne soit le personnage qui la fait, par ex. : « J'ai 100 paires de chaussures », ou à un autre personnage, par ex. « X n'achète que des yaourts 0% » (où X est ce dernier personnage). Les participants doivent inférer à partir de ces déclarations données en exemple l'existence d'un groupe d'« accros de la mode » et un groupe de personnages qui suivent un régime (les noms donnés aux groupes sont créés par les participants). Pour un personnage donné, les participants n'ont pas accès à la même déclaration : les participants doivent ainsi, pour classer correctement les personnages, échanger l'information disponible à chacun d'eux avec leur partenaire, et prononcer à la fois le nom du personnage et la déclaration associée. Il n'y a qu'une seule façon d'associer les personnages avec les groupes.

La figure 3.1 montre la structure du réseau social des personnages que les participants doivent découvrir, à savoir l'appartenance de chaque personnage à l'un des quatre groupes. Elle montre la distribution de l'information contenue dans les déclarations entre les personnages, et entre les participants.

GMUP partage un certain nombre de caractéristiques avec les autres protocoles expérimentaux mentionnés plus haut, mais c'est à la Map Task (Anderson *et al.*, 1991) que GMUP ressemble le plus. Il y a cependant des aspects importants qui sont spécifiques à GMUP. Les deux participants ont des rôles symétriques dans GMUP, contrairement à l'asymétrie Donneur-Receveur dans la Map Task. Pardo (2006) a en effet trouvé que le rôle du locuteur avait un impact sur la façon dont les participants interagissent dans une tâche de Map Task, les Donneurs adoptant plus les caractéristiques phonétiques que les Receveurs que le contraire. GMUP est aussi construite de façon à ce que le résultat de l'interaction, à savoir la structure du réseau social virtuel, émerge progressivement au cours du jeu comme le résultat du processus collaboratif entre les deux participants, alors que ce résultat est établi dès le départ dans la Map Task, sous la forme de la carte géographique étiquetée et comportant le chemin à dupliquer donnée au Donneur. De plus, les participants ont des informations complémentaires et consistantes concernant les personnages dans GMUP, alors que des inconsistances informationnelles sont introduites à dessein entre les participants de la Map Task. De plus, GMUP favorise autant le discours direct qu'indirect (à travers la lecture des

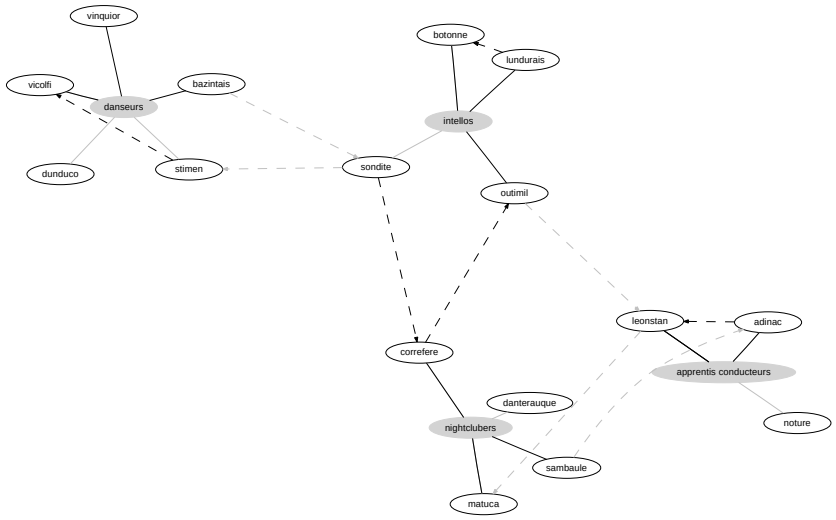

FIGURE 3.1 – Structure du réseau social dévoilé par les déclarations. Les traits pleins indiquent l'appartenance d'un personnage (étiquettes cerclées de noir) à un groupe (étiquettes pleines). Les flèches en pointillés indiquent que l'information sur l'appartenance d'un personnage (pointe de la flèche) est donné par un autre personnage (origine de la flèche). Lien de couleur noire [resp. grise] : l'information est disponible au participant A [resp. B].

déclarations). Finalement, une autre caractéristique de GMUP, et peut-être la plus importante, est qu'elle permet la production de pseudo-mots présentant des caractéristiques phonologiques finement contrôlées par des locuteurs en interaction conversationnelle.

Si elle a été développée spécifiquement pour notre étude, on peut voir à partir de la figure 3.1 les possibilités d'extension de GMUP et la flexibilité qu'elle offre pour traiter différentes situations interactionnelles. On peut par exemple facilement varier la difficulté de la tâche en variant le nombre de personnages et/ou de groupes, ainsi que le degré de redondance de l'information partagée, ou encore en introduisant un certain niveau d'incohérence dans les déclarations. GMUP peut aussi être étendue aux interactions à plus de 2 participants, en répartissant l'information partagée sur un nombre arbitraire de participants. Enfin, en variant le degré d'information détenue par les participants, on peut contrôler le rôle qui leur est attribué. Un protocole Donneur – Receveur sera ainsi obtenu en attribuant la totalité de l'information au Donneur, le Receveur ayant alors pour tâche de découvrir la structure du réseau social en se basant exclusivement sur ce que lui communique le Donneur.

3.5 Participants

Les participants ont été sélectionnés dans un lycée public du centre ville de Marseille, que nous avons choisi pour avoir accès à une population de niveau socioéconomique et d'origine géographique variés. Par ailleurs, étudier la dynamique des systèmes phonologiques chez une population lycéenne représente un intérêt particulier car la période de l'adolescence se caractérise linguistiquement par un éloignement du modèle familial en faveur de l'utilisation importante de formes non-standard sous l'effet du rôle social grandissant du groupe de pairs (Foulkes *et al.*, 2010, paragraphe 3.1.2).

Dans une première étape, 41 élèves de 7 classes différentes se sont portés volontaires pour compléter deux questionnaires. Le premier questionnaire a été établi d'après celui utilisé dans le projet « *Phonologie du Français Contemporain* » (PFC) (Durand *et al.*, 2003a), dont le but est d'évaluer la variété de français parlé par l'élève. Ce questionnaire est donné en Appendice, dans la figure 6.1. Nous avons donné autant d'importance pour décider de l'affectation d'un locuteur au groupe FM ou FS à son origine géographique qu'à son temps passé dans la région PACA, son niveau socioéconomique tel que déterminé par la profession des parents et sa propre perception de son accent. Par la suite, une vérification auditive sur la base des enregistrements de pre-test (voir section 3.6) a permis d'ajuster ces classements.

Le deuxième questionnaire était destiné à donner une mesure de com-

pétence sociale des élèves, à travers la traduction française de l'échelle de Désirabilité Sociale (Crowne et Marlowe, 1960; Marlowe et Crowne, 1961). La désirabilité sociale est définie comme un besoin d'approbation sociale et la conviction que celui-ci peut être satisfait à travers des comportements culturellement acceptables et appropriés (Marlowe et Crowne, 1961). L'échelle prédit que les sujets qui obtiennent des scores élevés ont un besoin élevé d'approbation sociale, et expriment des jugements plus favorables, dans le sens de comportements sociaux attendus. Les individus en bas de l'échelle, au contraire, montrent une plus grande liberté par rapport aux attitudes socialement correctes. L'échelle traduite a été validée pour le français en vérifiant que les moyennes et déviations standard obtenues pour un ensemble de sujets français ne participant pas à l'expérience étaient comparables à celles données par l'échelle originale (échelle originale/traduite : N = 120/82 ; Moyenne = 4.16/4.77 ; Déviation standard = 1.75/1.57)[4]. Ce questionnaire est joint en Appendice, dans les figures 6.2 et 6.3.

Nous nous sommes intéressés en utilisant cette mesure de compétence sociale à savoir si l'ajustement des comportements prédit par l'échelle concerne aussi les comportements langagiers, et en particulier le niveau phonétique. Nous nous intéressons donc à savoir si les participants qui obtiennent un score élevé de désirabilité sociale vont manifester un ajustement des caractéristiques phonétiques et phonologiques de leur parole vers celle de leur partenaire et/ou vers les prototypes accentuels véhiculés par celui-ci. Au contraire, dans le cas d'un score faible, la plus grande indépendance de comportement prédite par l'échelle conduirait les participants à manifester un comportement langagier plus stable et, en particulier, qui ne serait pas influencé par celui de leur partenaire ou par l'accent qu'il incarne.

De tels ajustements dans la parole en relation avec la désirabilité sociale ont déjà été mis en évidence pour l'intensité vocale. Natale (1975) a ainsi montré qu'au cours d'une conversation durant laquelle l'intensité vocale d'un participant était modifiée, son partenaire manifestait un changement de sa propre intensité vocale dans le sens de la modification, et la magnitude de l'effet était positivement corrélée avec le score de désirabilité sociale du sujet. Gregory et Webster (1996), en utilisant une mesure acoustique LTAS (spectre moyen à long terme), une mesure globale des caractéristiques vocales du locuteur, ont montré que dans des communications dyadiques, le locuteur de niveau social plus bas convergeait vers son partenaire sur la dimension du signal non verbal.

[4]Les participants ont aussi complété l'échelle de Self-Monitoring (Snyder, 1974), révisée par (Lennox et Wolfe, 1984), mais seule l'échelle de Désirabilité Sociale a été utilisée pour apparier les participants.

Dans une deuxième étape, 24 élèves ont été sélectionnés sur la base de leur réponse aux deux questionnaires (20 filles et 4 garçons, âge moyen 15.8 ans, SD=0.9). Douze d'entre eux étaient locuteurs du français standard, et les douze autres locuteurs du français méridional. Les locuteurs avaient un score allant de 1.8 à 8.2 sur l'échelle de Désirabilité Sociale qui s'étend de 0 (faible besoin d'acceptation sociale) à 10 (besoin élevé d'acceptation sociale).

Finalement, chacun des locuteurs du français standard a été apparié avec un des locuteurs du français méridional pour former 12 paires (dyades, ci-après). Dans une dyade donnée, les deux sujets étaient du même sexe et ne se connaissaient pas. Ils avaient une position similaire sur l'échelle de Désirabilité Sociale (différence absolue maximale de 1.8 entre les deux sujets).

3.6 Procédure

Les enregistrements ont été menés dans une pièce calme de l'établissement et en présence de l'expérimentateur. Ils ont été divisés en deux sessions séparées d'une semaine d'intervalle. Durant la première session, les sujets ont été enregistrés individuellement durant la lecture de 3 listes de 16 noms, à l'intérieur d'une phrase porteuse (« Le numéro ___ c'est _____ ») dans un ordre aléatoire qui n'était pas le même pour tous les sujets (pre-test).

Dans une deuxième session, chaque dyade a effectué un court entraînement qui permettait aux participants de se familiariser avec les règles du jeu. Ils ont ensuite participé à 3 jeux successifs, chacun consistant en un ensemble différent de 16 mots, 16 photos et 16 déclarations (phase de test). En moyenne, la durée pour compléter les 3 jeux était de 20 minutes. Après la fin du troisième jeu, les deux participants étaient enregistrés séparément durant la lecture des 3 listes de noms mélangés avec 32 nouveaux noms qui n'avaient pas été utilisés dans les jeux, dans un ordre aléatoire qui n'était pas le même pour tous les participants, comme dans la première session (post-test).

Les sujets étaient enregistrés en utilisant des microphones serre-tête AKG C 420 reliés à un ordinateur par une interface audionumérique de haute qualité (EDIROL UA-25). Durant les jeux, la parole de chaque sujet était enregistrée sur un des deux canaux d'un fichier stéréo pour permettre la synchronisation des enregistrements entre les sujets.

Les sujets ne se sont pas vus pendant la totalité des deux sessions.

3.7 Alignement

Une procédure semi-automatique a été développée en utilisant le toolkit HTK de reconnaissance automatique de la parole[5]. L'objectif est double : localiser temporellement les réalisations acoustiques des variables phonologiques incluses dans les mot-cibles, et décrire leur réalisation en termes de variantes de prononciation. La modélisation de la variation de prononciation avec des techniques de reconnaissance automatique de la parole a reçu un intérêt grandissant (voir Strik et Cucchiarini, 1999, pour une revue), notamment en raison des travaux de plus en plus nombreux effectués sur de la parole conversationnelle, qui présente une part plus importante de variation. La façon la plus simple de modéliser la variation de prononciation interne au mot, que nous adoptons ici, consiste à ajouter des variantes de prononciation dans le dictionnaire de prononciation. Il ne suffit cependant pas de multiplier les ajouts de variantes pour améliorer la reconnaissance, et il existe un compromis entre le nombre de variantes nécessaires et la qualité des modèles acoustiques (Adda-Decker et Lamel, 1999). D'autres approches existent, comme une modélisation utilisant des modèles basés sur des traits plutôt que sur des phones (Bates *et al.*, 2007).

Les modèles acoustiques utilisés sont des monophones indépendants du contexte entraînés pour le français par J.-P. Goldman à l'Université de Genève (ensemble MA1)[6]. Dans la suite de ce travail nous avons été amenés à utiliser un autre ensemble de modèles acoustiques (ensemble MA2, cf. section 3.7.3). Les caractéristiques de ces deux ensembles sont présentés dans la table 3.5.

	MA1	MA2
Taille du corpus d'apprentissage	20 minutes	72 heures
Nombre de locuteurs	5	1564
Type de parole	lecture	radiophonique
Nombre de Gaussiennes	1	128

TABLE 3.5 – Caractéristiques des deux ensembles de modèles acoustiques utilisés pour l'alignement forcé.

[5]URL : htk.eng.cam.ac.uk, page consultée le 03 août 2010.
[6]URL : latlcui.unige.ch/phonetique, page consultée le 03 août 2010.

3.7.1 Word-spotting

Dans un premier temps, une routine de reconnaissance automatique de la parole a été écrite pour localiser les mot-cibles dans les enregistrements. Le modèle de langue utilisé, qui décrit les séquences de mots possibles dans le signal de parole, est composé exclusivement des mot-cibles utilisés pour l'enregistrement analysé et de la totalité des modèles acoustiques décrivant les phonèmes du français. La grammaire, qui utilise la notation EBNF (Extended Backus-Naur Form) (cf. Young *et al.*, 2006, p. 169), s'écrit :

```
( < {$phon} $motCible {$phon} > )
```

où `$phon` est l'ensemble des modèles acoustiques y compris les modèles de respiration et de silence, et `$motCible` est la liste de mot-cibles spécifiques à l'enregistrement considéré : pour les jeux elle contient 16 mots, pour le pre-test 48 mots correspondant aux 3 jeux, et pour le post-test 80 mots, correspondants aux mots du pre-test plus deux listes de 16 mots. Cette grammaire décrit le signal de parole comme composé de séquences répétées de mot-cibles éventuellement précédées et suivies de n'importe quel son du français. Les mot-cibles sont composés d'une séquence unique de modèles acoustiques, qui correspondent à une prononciation standard. Cela n'est pas gênant dans la mesure ou la majorité des réalisations phonétiques attendues pour les locuteurs du français méridional sont assez proches sur la dimension phonétique de leur correspondantes du français standard, et peuvent donc être approximées par ces modèles acoustiques.

La pénalité d'insertion de mot est un paramètre utilisé pendant la reconnaissance qui permet d'ajuster le nombre de mots à reconnaître dans le signal de parole : une pénalité élevée entraîne le découpage du signal en mots de longueur plus importante, et permet ainsi d'ajuster l'équilibre entre fausses alertes (mot-cibles détectés pour des portions de signal ne correspondant pas à un mot) et non-détection des mots. La valeur qui donnait le meilleur équilibre était $p = -80$.

Pour obtenir une reconnaissance complète et correctement alignée de l'ensemble des mot-cibles, une seconde passe est effectuée, à l'aide d'un éditeur de signal interactif développé avec Matlab. Les portions de signal correspondant à de la parole, obtenues en éliminant les séquences alignées avec le modèle acoustique de silence à la première passe, sont jouées en accéléré à l'utilisateur qui interrompt l'écoute lorsqu'il identifie un mot-cible prononcé. Le spectrogramme correspondant à la portion de signal de parole est alors présenté à l'utilisateur sur lequel sont posées les frontières du candidat identifié par la procédure de word-spotting. L'utilisateur peut alors valider le candidat et poursuivre l'écoute, ajuster les frontières du candidat proposé, ou exécuter

à nouveau la reconnaissance en mot-cibles avec la grammaire ci-dessus sur la portion de signal, en ajustant éventuellement la fenêtre de signal considéré. Dans quelques cas, la procédure n'est pas parvenue pas à détecter correctement le mot-cible dans la fenêtre sélectionnée, par exemple lorsque le mot était tronqué, ou prononcé d'une façon qui divergeait significativement de la prononciation attendue (erreurs de lecture). Dans ces cas, l'utilisateur peut manuellement sélectionner le mot-cible et spécifier les frontières correspondant à la séquence, et marquer ce mot comme « erroné » pour un codage futur des erreurs de prononciation. Cette deuxième passe a permis de s'assurer que la totalité des mots produits dans l'interaction étaient identifiés, y compris les mots prononcés de façon erronée, et que les frontières temporelles étaient bien ajustées. Une capture d'écran de l'interface utilisée est montrée dans la figure 3.2.

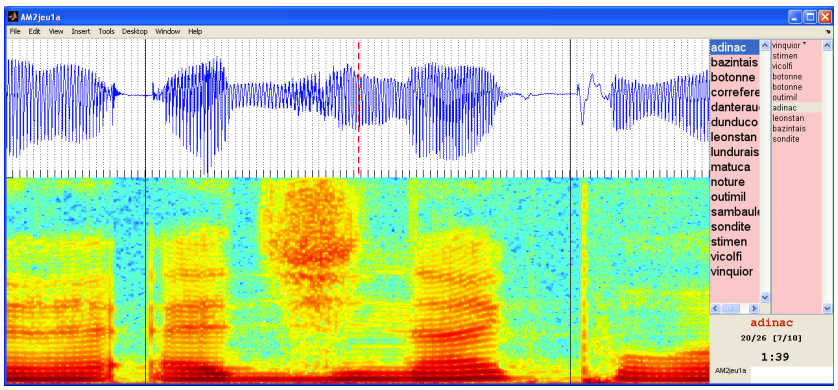

FIGURE 3.2 – Capture d'écran de l'interface développée pour aligner les mot-cibles. A gauche : forme d'onde et spectrogramme, avec des repères temporels d'une granularité de 10 ms. A droite, liste des mot-cibles candidats pour la reconnaissance (colonne de gauche) et liste des mot-cibles reconnus (colonne de droite). Divers raccourcis clavier sont également disponibles pour ajuster la reconnaissance des mots sur la portion de signal courante.

Une troisième passe a été nécessaire pour coder séparément les segments qui dévient de la prononciation attendue et qui nécessitent une transcription *ad-hoc* en modèles acoustiques. Nous avons jugé opportun de conserver la trace complète de la production des mots, même quand la totalité ou une partie du mot n'était pas exploitable, pour laisser la possibilité d'analyser les parties non-erronées des mots, et comptabiliser les répétitions des mots au cours de l'interaction. Nous distinguons deux niveaux d'annotation des erreurs. Le premier, au niveau du mot, est un indicateur booléen qui indique

si la totalité ou une partie du mot a été correctement prononcé. Il est utilisé aussi pour marquer les mots dont la forme acoustique ne permet pas d'obtenir des mesures acoustiques fiables, comme les mots chuchotés ou produits en même temps qu'une autre source de bruit (mouvement du micro par exemple). L'autre niveau d'annotation des erreurs concerne le niveau des segments. Il permet de coder séparément pour chaque segment qui compose le mot ceux qui sont à écarter des analyses subséquentes. Nous précisons que la notion de correction de prononciation fait référence à la conversion graphème-phonème attendue, qui permet de recevoir ce jugement binaire. Les variations de prononciation qui relèvent de la façon de prononcer du locuteur, et qui incluent les variations régionales, sont évidemment non marquées comme erronées. L'ensemble des mots marqués comme erronés à la précédente étape sont présentés tour à tour à l'annotateur dans une seconde interface à l'aide de laquelle l'utilisateur peut écouter le mot et spécifier la prononciation déviante ou absente d'un ou plusieurs segments le composant. A cet effet, chaque mot est présenté sous la forme de la séquence de segments critiques et de segments « fillers » utilisée pour construire la séquence de modèles acoustiques qui définit le mot-cible dans la grammaire. L'utilisateur peut alors marquer les segments erronés en choisissant une transcription erronée déjà rencontrée pour le type de segment considéré (y compris l'élision), ou ajouter une nouvelle transcription déviante pour ce type de segment. La figure 3.3 montre une capture d'écran de cette interface de contrôle.

Ces deux étapes nécessitent l'intervention d'un annotateur humain et sont donc coûteuses en temps. La présentation des séquences de parole en accéléré, la possibilité d'effectuer le codage des erreurs en deux étapes, et la présentation de mots identiques en séquence lors du codage des erreurs ont permis d'obtenir l'identification de la totalité des productions des mot-cibles avec un codage des erreurs de prononciation et leur localisation temporelle vérifiée individuellement en environ deux fois le temps réel correspondant à l'interaction.

Dans cette étape, l'intervention de l'annotateur a permis d'obtenir une identification complète des mot-cibles, les localisations précises des frontières et le codage des erreurs de prononciation. Ces trois points permettent d'assurer un maximum de succès à la procédure d'alignement automatique en variantes de prononciation.

3.7.2 Alignement en variantes de prononciation

Une fois les mot-cibles correctement identifiés et localisés, nous avons procédé à leur alignement au niveau segmental. Notre objectif est, comme pour les mot-cibles, d'obtenir l'identification et la localisation temporelle des seg-

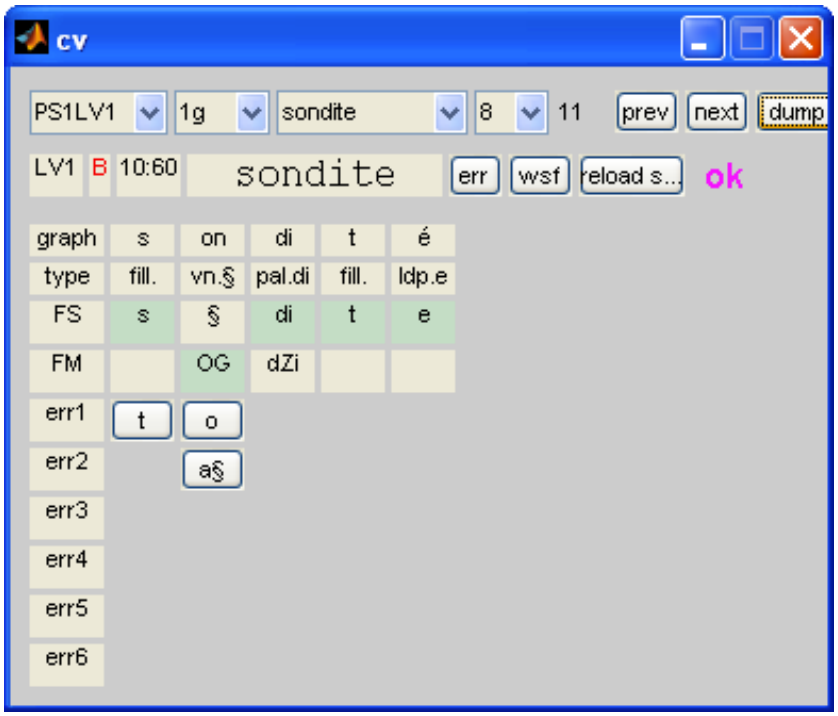

FIGURE 3.3 – Capture d'écran de l'interface développée pour coder les erreurs de prononciation individuellement pour chaque mot-cible. En vert : variantes choisies par la procédure. Lignes err1 . . . err6 : boutons permettant de coder les erreurs.

ments qui composent les mots. L'identification, c'est-à-dire le choix d'une variante de prononciation plutôt qu'une autre par l'aligneur, peut déjà fournir des informations sur la façon dont sont réalisés les segments critiques, et la détermination de leurs limites temporelles peut nous permettre de conduire des analyses acoustiques plus détaillées pour caractériser leur forme et leur évolution.

La tâche d'alignement forcé consiste à fournir à l'aligneur une transcription pre-établie et un signal de parole associé. Le résultat de l'aligneur sera la liste des étiquettes qui correspondent aux frontières entre les modèles acoustiques telles que le produit des scores des modèles acoustiques associés aux portions de signal ainsi délimitées soit minimal. L'algorithme de minimisation est celui de Viterbi. HTK permet d'étendre la fonctionnalité de l'alignement forcé en spécifiant plusieurs alternatives de transcription. Le score de chaque transcription est calculé de la même façon, et la meilleure transcription est celle qui donne le meilleur score.

L'alignement forcé est exécuté sur les mot-cibles en associant à chaque mot et à chaque réalisation des mots erronés un dictionnaire de prononciation spécifique. Le dictionnaire de prononciation pour un mot-cible donné contient la liste de toutes les transcriptions possibles obtenues par combinaison des variantes de prononciation décrivant les réalisations des segments critiques dans les deux variétés de français considérées. Par exemple, pour le mot '*Santinais*', la liste des transcriptions possibles contient l'énumération de tous les chemins possibles qui traversent le graphe des variantes de prononciation, présenté dans la figure 3.4. Si on voit à partir de cette exemple

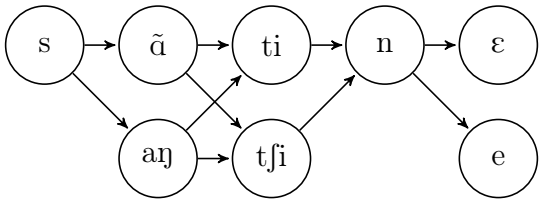

FIGURE 3.4 – Graphe listant les transcriptions possibles du mot '*Santinais*' incluses dans le dictionnaire de prononciation (8 combinaisons possibles). La prononciation standard [resp. méridionale] est obtenue en choisissant les alternatives sur la ligne du haut [resp. du bas].

qu'il existe deux transcriptions remarquables, caractéristiques de chacune des variétés de français étudiées, le processus d'alignement forcé ne les traitera pas différemment, car le dictionnaire contient la liste de toutes les combinaisons. Le résultat de l'alignement permet donc d'évaluer indépendamment

la réalisation de chacun des segments critiques contenus dans un mot. Par ailleurs, cette indépendance de réalisation a aussi une validité linguistique, les locuteurs d'une variété donnée pouvant tout à fait, ponctuellement ou de façon générale, présenter pour certains traits les caractéristiques phonétiques d'une autre variété.

Si la description phonologique que nous avons faite des deux variétés est binaire, en décrivant pour chacune d'elle la réalisation la plus caractéristique (voir section 3.2.1 et table 3.1), la réalisation acoustique est évidemment moins tranchée, et peut se modéliser avec plusieurs modèles acoustiques. C'est pourquoi nous avons inclus dans les variantes de prononciation un nombre de modèles acoustiques supérieur à 2 pour certains segments critiques. Au delà de donner une caractérisation plus précise de la réalisation des segments critiques, l'inclusion de plusieurs modèles acoustiques pourrait permettre de révéler d'autres phénomènes intéressants concernant leur réalisation (voir section 4.4). De plus, les modèles acoustiques en tant qu'objets statistiques ne sont représentatifs d'une catégorie phonétique que dans la mesure où ils ont été appris sur des segments de parole représentant effectivement les catégories phonétiques en question. La pertinence phonétique des modèles acoustiques dépend ainsi du corpus d'apprentissage (style de parole, origine géographique des locuteurs, nombre des locuteurs...). La liste des modèles acoustiques utilisés pour décrire les segments critiques est donné dans la table 3.6.

La procédure d'alignement forcé est effectuée totalement automatiquement et associe à chaque réalisation des segments critiques une variante de prononciation. Le contexte environnant les segments critiques a été décrit le plus précisément possible au cours de la tâche de word-spotting (identification des mot-cibles, vérification des frontières de mot, codage des déviations de prononciation) pour limiter au maximum les erreurs d'alignement, mais le principe même de la procédure d'alignement forcé ne garantit pas dans son principe une précision temporelle de l'alignement.

3.7.3 Evaluation de la qualité de l'alignement

La qualité d'un alignement phonétique dépend de l'utilisation qu'on veut en faire. Si l'on se propose d'effectuer des comparaisons de qualité de voyelle, on voudra peut-être localiser le noyau de la voyelle avec certitude. Dans ce cas, la précision temporelle de la position des frontières tolère une certaine variabilité d'une part parce que l'information à recueillir n'est pas aux frontières, et d'autre part parce que l'opération qui permet de calculer le milieu de la voyelle limite l'influence de l'imprécision des frontières, si la distribution des erreurs est identique pour les deux frontières. Si en revanche on veut calculer

DP	SC	Prononciation attendue		Modèles acoustiques
		FM	FS	
Schw.	S1	[ə]	[ ]	{ə} ; {ø} ; {œ} ; {}
	S2	[ə]	[ ]	{ə} ; {ø} ; {œ} ; {}
	S3	[ə]	[ ]	{ə} ; {ø} ; {œ} ; {}
	S4	[ə]	[ ]	{ə} ; {ø} ; {œ} ; {}
	S5	[ə]	[ ]	{ə} ; {ø} ; {œ} ; {}
Post.	B1	[ɔ]	[o̞]	{ø} ; {œ} ; {o} ; {ɔ}
	B2	[o]	[o̞]	{ø} ; {œ} ; {o} ; {ɔ}
Moy.	M1	[e]	[e]	{e}
	M2	[e]	[ɛ]	{e} ; {ɛ}
	M3	[ɔ]	[ɔ]	{ø} ; {œ} ; {o} ; {ɔ}
	M4	[ɔ]	[o]	{ø} ; {œ} ; {o} ; {ɔ}
Cor.	C1	[tʃi]	[ti]	{ti} ; {tʃi} ; {tsi}
	C2	[tʃy]	[ty]	{ty} ; {tʃy} ; {tsy}
	C3	[dʒi]	[di]	{di} ; {dʒi} ; {dzi}
	C4	[dʒy]	[dy]	{dy} ; {dʒy} ; {dzy}
Nas.	N1	[ɛN]	[ɛ̃]	{ɛ̃} ; {ɑ̃} ; {ɔ̃} ; {ɛN} ; {eN}
	N2	[œN]	[ɛ̃]	{ɛ̃} ; {ɑ̃} ; {ɔ̃} ; {œN}
	N3	[aN]	[ɑ̃]	{ɛ̃} ; {ɑ̃} ; {ɔ̃} ; {aN}
	N4	[ɔN]	[ɔ̃]	{ɛ̃} ; {ɑ̃} ; {ɔ̃} ; {ɔN} ; {oN}

TABLE 3.6 – Modèles acoustiques utilisés pour chaque segment critique dans la construction du dictionnaire de prononciation. Le modèle vide ({}) correspond à l'élision. Les séquences de plus de deux symboles correspondent à la concaténation de plusieurs modèles acoustiques. N représente une consonne nasale de même lieu d'articulation que la consonne suivante dans le mot. Elle est remplacée par [n] devant [t] et [d], par [m] devant [p] et [b] et par [ŋ] devant [k] et [g].

la durée du Voice Onset Time (VOT) d'une plosive, la précision de la localisation de l'explosion de la consonne et du début de vibration des cordes vocales est cruciale.

L'utilisation la plus courante des modèles acoustiques comme objets statistiques décrivant des sons de la parole est la *reconnaissance* de la parole. L'objectif recherché est d'identifier avec le moins d'erreur possible une cible, comme un segment phonétique ou une syllabe, et par suite les unités qui les regroupent comme des mots et des phrases. Cet objectif est atteint lorsque la séquence des unités est détectée, c'est-à-dire qu'une correspondance ordonnée des unités est trouvée avec le signal de parole. Autrement dit, le résultat du processus de reconnaissance automatique, et les étiquettes temporelles qui en découlent, garantissent une description la plus fidèle possible de l'enchaînement des cibles, mais pas de leur frontières. Les frontières obtenues pour une cible garantissent que la cible est localisée quelque part après le début de la frontière gauche et quelque part avant la frontière droite. Elle n'est que secondairement un indicateur d'un évènement d'intérêt, qui est la frontière entre deux cibles, calculée comme le milieu de celles-ci. Cela n'empêche néanmoins pas d'obtenir des résultats cohérents de localisation des frontières entre deux alignements dont les cibles qui encadrent la frontière présentent des caractéristiques similaires. C'est d'ailleurs cet aspect du processus de reconnaissance qui implique que les modèles acoustiques monophones indépendants du contexte, s'ils affichent des performances de reconnaissance moins bonnes que les modèles triphones qui modélisent des cibles avec leur contexte (voir par ex. Deshmukh *et al.*, 1999; Ney et Ortmanns, 1999), présentent une précision de localisation de frontière meilleure que ces derniers (voir aussi Bürki *et al.*, 2008, pour une étude comparative de différents systèmes pour l'alignement du schwa en français). D'autres approches en technologie de la parole (voir par ex. Hosom, 2008) tentent de développer des approches qui visent à améliorer la précision de l'alignement en incorporant l'information de transition entre les segments dans le processus de décodage.

Nous avons évalué la qualité de l'alignement obtenu avec la méthode décrite plus haut en vérifiant manuellement un sous-ensemble des alignements décrivant d'une façon la plus large possible le corpus. Nous avons pour cela sélectionné aléatoirement 4 segments par sujet, soit 96 alignements en tout, qui décrivent les quatre classes de son représentées par les segments critiques, et qui possèdent des caractéristiques acoustiques distinctes. Ces classes de son sont identiques aux dimensions phonologiques à l'exception de la classe *Voyelles orales*, qui rassemble les dimensions phonologiques *Post.* et *Moy.*. Les alignements ont été tirés de la phase de test, les tâches de lecture (phases pre- et post-test) donnant de meilleurs résultats étant donné le style de parole plus soigné. Les résultats de cette procédure de vérification sont présentés

dans la table 3.7.

imprécision	Schwa	Voyelles orales	Séq. Cor. Voy. haute	Voyelles nasales
$<$ 20 ms.	2	2	4	3
$>$ 20 ms.	5	1	2	1

TABLE 3.7 – Qualité de l'alignement, évaluée par le nombre de segments incorrectement alignés. La résolution temporelle de l'aligneur est de 10ms. Les erreurs excédant 20 ms. comprennent également les élisions.

Les résultats de cette vérification permettent de s'assurer que l'alignement effectué décrit bien les réalisations acoustiques des segments critiques produits par les participants au cours des interactions, et permettent de mener des analyses sur leur forme et leur évolution au cours de la tâche. Ces analyses et les résultats sont présentés à la section 4.

Si les résultats d'alignement assurent une qualité suffisante pour une caractérisation au niveau catégoriel des réalisations des segments critiques, les imprécisions temporelles qui en résultent rendent difficile une mesure automatique de l'information acoustique détaillée. Nous avons pu utiliser plus tard dans notre travail un autre ensemble de modèles acoustiques, entraînés par G. Gravier à l'IRISA. Ce sont, comme les premiers, des monophones indépendants du contexte, mais ils sont entraînés sur une plus grande base d'apprentissage et sont modélisés par un plus grand nombre de gaussiennes (voir table 3.5, p. 63).

Cet ensemble de modèles présente une amélioration sensible de la qualité d'alignement, en terme de nombre d'erreurs d'alignement et de précision temporelle de celui-ci. C'est à partir des résultats de ce dernier alignement que sont développées les méthodes de localisation des informations acoustiques pertinentes caractérisant la réalisation des segments critiques présentés à la section suivante.

3.8 Mesures acoustiques

Après nous être assurés de la « précision relative » de l'alignement, nous nous intéressons maintenant à un niveau plus fin d'analyse que celui donné par l'alignement en variantes de prononciation, et qui est obtenu par la caractérisation phonétique détaillée de la réalisation des segments critiques. Cette section présente la méthode utilisée pour localiser dans un premier temps les positions les plus pertinentes pour la mesure de l'information acoustique, et

dans un deuxième temps les mesures choisies pour capturer au plus près l'information acoustique associée aux segments critiques pour caractériser leur forme et évolution au cours de l'interaction.

Nous détaillons les critères utilisés pour localiser l'information acoustique pertinente pour les 4 classes de son que constituent les dimensions phonologiques étudiées. Le point commun dans tous les cas est de partir de la position temporelle donnée par l'alignement et de caractériser le signal acoustique au voisinage de cette position. Cette méthode s'apparente à celle développée par Stevens et ses collègues (par ex. Stevens, 2002; Liu, 1996; Chen, 2000) dans le développement de leur système de reconnaissance de la parole basée sur la détection de traits. Nous avons cependant utilisé des approches spectrales globales (« whole-spectrum »), plus robustes et mieux adaptées pour traiter la parole spontanée. Une comparaison détaillée entre l'approche spectrale globale et des méthodes traditionnelles de mesures de formants dans la caractérisation acoustique de sons de la parole est présentée dans Harrington (2010, p. 89). Ces approches relativement récentes permettent de caractériser avec une dimensionalité faible la forme générale du spectre, et de conserver ainsi une grande partie de l'information phonétique pertinente. Par exemple, si la mesure des formants reste la méthode de choix pour évaluer la qualité de voyelles statiques, c'est-à-dire mesurée en un point du signal, dans la mesure où elle donne la meilleure séparation des catégories de voyelles dans un plan (F_1, F_2) par exemple, la paramétrisation du signal de parole au même point du signal avec des coefficients MFCC par exemple, si elle donne une séparation moins bonne, présente le grand avantage d'être une méthode robuste car elle n'implique pas une procédure de détection à laquelle il faut apporter beaucoup de soin, comme c'est la cas pour la mesure des formants. Ce sont pour ces raisons que ces mesures sont utilisées dans les processus de reconnaissance automatique de la parole.

Les formants ont par contre l'avantage, dans une perspective d'interprétation phonétique, d'être compatibles avec la théorie source-filtre de production de la parole (Fant, 1960). Les approches spectrales globales trouvent en revanche une pertinence théorique dans la théorie de l'information (Shannon, 1948) qui a servi de cadre théorique au développement du traitement automatique de la parole.

3.8.1 Localisation et mesure de l'information acoustique pertinente

Schwas

Les résultats de l'alignement en variantes de prononciation ont révélé des inconsistances dans la façon dont été alignés les schwas finaux. Cela est dû en partie au fait que l'aligneur ne faisait pas la différence entre un schwa en position finale de mot et une marque d'hésitation ou une pause remplie se trouvant après le mot-cible. Par ailleurs, les différences prédites entre les deux variétés concernent la fréquence de réalisation de la variable, et non le timbre. Des différences sur la dimension de la qualité de voyelle ont bien été relevées pour la variété méridionale, avec une tendance à la postériorisation. Cette tendance est attribuée à l'influence du substrat franco-provençal qui a conservé la voyelle /a/ du latin en position finale de mot plus longtemps que les variétés du nord (Carton *et al.*, 1983; Fagyal, 2006). Cependant cette réalisation postérieure a été également observée pour le français parlé de Lille, une variété du Nord (Carton *et al.*, 1983). Si la dimension du timbre du schwa aurait ainsi pu se révéler intéressante pour caractériser les différences de réalisation entre les deux variétés, elle a été écartée dans la suite des analyses.

Voyelles orales

Les différences de réalisation des voyelles entre les variétés du français standard et du français méridional concernent des différences de qualité de voyelle (d'antériorité pour les voyelles moyennes postérieures et de hauteur pour les voyelles moyennes). Nous nous intéressons donc à mesurer la qualité de voyelle, que nous prenons au point de maximum de stabilité. Ce point a été déterminé aux alentours de l'étiquette proposée par l'alignement. Les durées des étiquettes sont ajustées autour du milieu du segment pour avoir une durée minimale de 60 ms. Nous avons extrait sur cette durée les 3 premiers coefficients DCT, calculés sur les spectres échantillonnés tous les 5 ms. Les 3 premiers coefficients correspondent respectivement à la moyenne du spectre, c'est-à-dire l'intensité globale de la portion de signal de parole associée, à la pente spectrale et au degré de courbure du spectre. Ces informations sont suffisantes pour caractériser les changements de forme spectrale au cours de la réalisation de la voyelle : dans le cas d'une séquence plosive – voyelle par exemple, le premier coefficient présentera une variation temporelle importante à l'endroit de la transition entre les segments. Pour le cas d'une séquence fricative – voyelle, c'est plutôt la pente spectrale et la courbure du spectre qui vont présenter des variations importantes. Dans le cas d'enchaî-

nement de voyelles, si la variation des 3 coefficients est moins importante, elle sera néanmoins maximale à l'endroit de la transition. La partie la plus stable a été déterminée en calculant la dérivée de ces 3 signaux temporels, puis en localisant la position du minimum du signal résultant de la somme de la valeur absolue des 3 signaux. Autant d'importance est donc attribuée aux 3 coefficients dans la détermination de la partie la plus stable. La dérivée a été calculée par différenciation d'ordre 3, c'est-à-dire que la valeur de la dérivée à un point t du signal est prise comme la différence entre la valeur du coefficient à l'instant $t + 3$ et $t - 3$ (l'échantillonnage est celui des spectres). Les signaux étant bruités, et la mesure d'un extremum représentatif sujet aux variations locales du signal, nous avons sélectionné comme cible représentant le maximum de stabilité le milieu de la plus longue portion de signal au dessous d'un seuil ajusté empiriquement. La figure 3.5 montre les 3 étapes du calcul de la dérivée cumulée des 3 premiers coefficients DCT pour la voyelle moyenne postérieure en position finale du mot '*Sambaule*' produit par le locuteur m03 pendant la première partie de l'interaction. Si les signaux associés aux 3 coefficients DCT ne manifestent pas individuellement de variation marquée, la somme des 3 signaux dérivés dont on a pris la valeur absolue permet de dégager le degré de variation spectrale globale. On notera par ailleurs que les 3 signaux ont un sens de variation différent – illustrant le fait qu'ils encodent bien des dimensions différentes du spectre (voir plus haut) – mais que la variation simultanée sur ces dimensions est bien capturée par la procédure. La figure 3.6 montre la détermination de la position de la cible sur ce signal, et la figure 3.7 montre la position de cette cible en référence à la forme d'onde et au spectre. Le changement spectral correspondant à la transition de la voyelle vers la liquide est écarté par la procédure.

Une fois la cible localisée, l'information acoustique pertinente qui caractérise le segment a été mesuré par les 5 premiers coefficients DCT en ce point du signal, dans la plage spectrale 200–4000Hz.

Séquences Coronales – Voyelles hautes

L'élément qui nous intéresse dans la caractérisation des séquences plosives coronales suivies de voyelles hautes est le degré d'affrication / palatalisation des plosives (Binisti et Gasquet-Cyrus, 2003; Woehrling et Boula de Mareüil, 2006), caractéristique des réalisations de la variété méridionale. Ce phénomène est localisé acoustiquement entre l'explosion de la plosive et s'étend sur la réalisation de la voyelle suivante. Nous nous proposons d'examiner la forme spectrale de ces séquences au point d'explosion de la plosive, à l'endroit où l'énergie est maximalement localisée dans la partie haute du spectre. Une

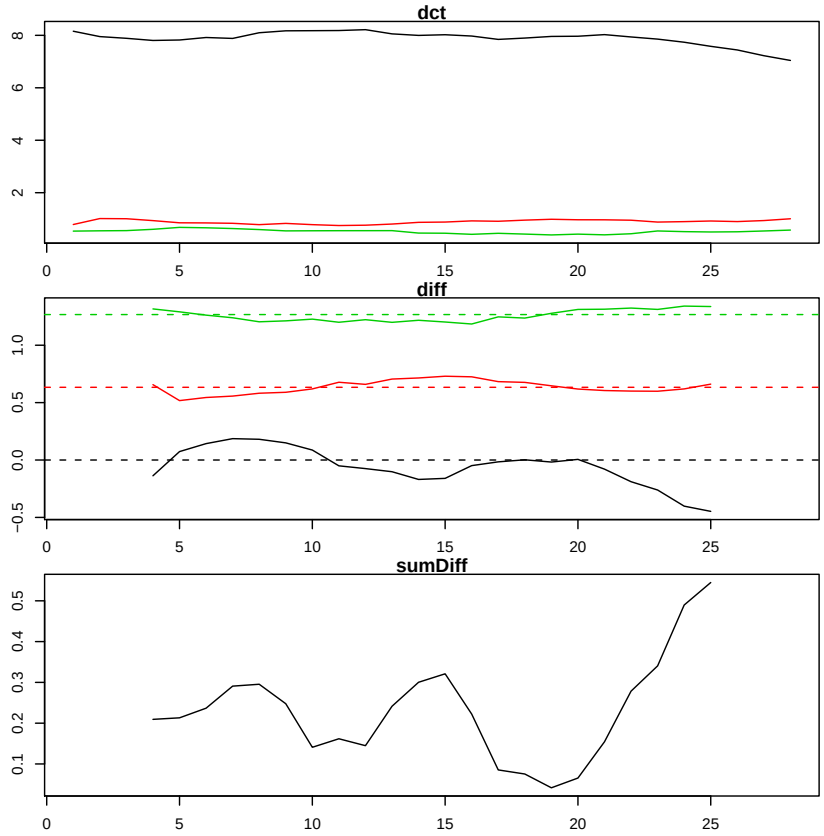

FIGURE 3.5 – Etapes de calcul de la dérivée cumulée des 3 premiers coefficients DCT pour la voyelle moyenne postérieure en position finale du mot '*Sambaule*' produit par le locuteur m03 pendant la première partie de l'interaction. Panneau du haut : 3 premiers coefficients DCT. Panneau du milieu : dérivées des signaux. Un décalage est introduit pour faciliter la visualisation. Panneau du bas : somme de la valeur absolue des signaux. Abscisses : temps (ms.).

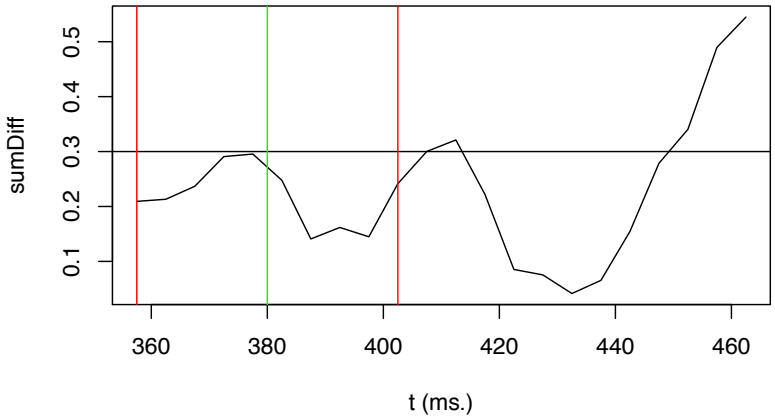

FIGURE 3.6 – Détermination de la position identifiant la partie la plus stable de l'extrait de parole, à partir du signal des dérivées cumulées des 3 premiers coefficients DCT (voir figure 3.5). La cible est calculée comme le milieu du plus long intervalle de valeurs inférieures à un seuil déterminé empiriquement (ici 0.3). Trait horizontal noir : seuil. Traits rouges : frontières de l'intervalle. Trait vert : cible.

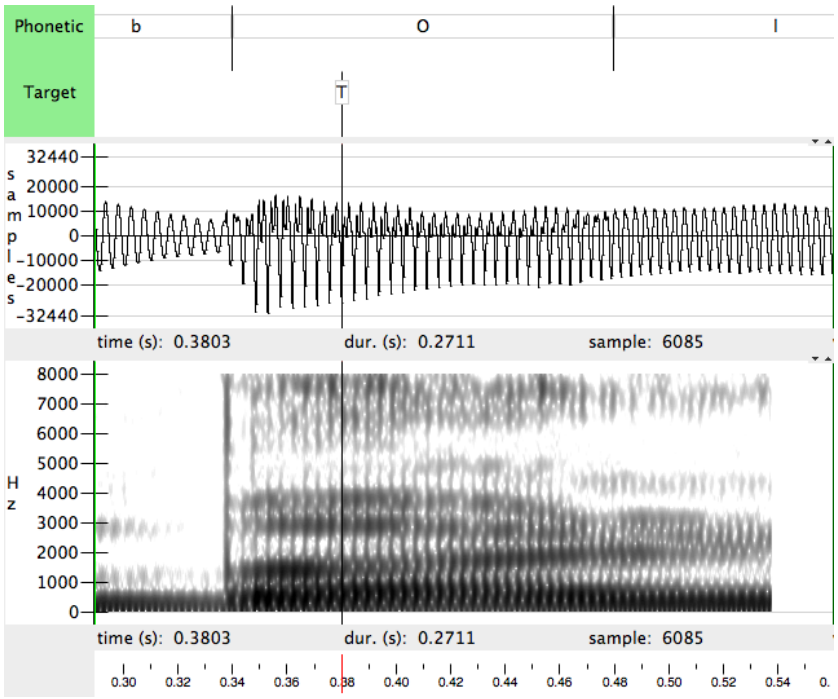

FIGURE 3.7 – Position de la cible de la partie la plus stable de la voyelle moyenne postérieure superposée sur la forme d'onde et le spectre. Ici, la variante de prononciation choisie par l'aligneur est {ɔ}.

caractérisation détaillée du lieu d'articulation des obstruentes par la forme spectrale est présentée dans Harrington (2010, p. 103).

Le premier moment spectral m_1 permet précisément d'obtenir cette mesure. Cette caractérisation statistique d'une distribution donne la position dans une distribution du maximum de densité. Nous avons donc calculé la position du maximum de m_1 sur la portion de signal correspondant à la totalité des modèles acoustiques décrivant le segment critique (composée de séquences de 2 ou 3 modèles, selon le choix effectué par l'aligneur). Comme le signal temporel du m_1 présente pour ces séquences une trajectoire montante-descendante relativement claire, nous n'avons donc pas eu recours à des critères supplémentaires pour extraire le maximum. Des techniques similaires utilisant les moments spectraux pour la caractérisation des fricatives ont été utilisées par Forrest *et al.* (1988); Jongman *et al.* (2000); Tabain (2001).

La figure 3.8 présente le maximum de m_1 associé à la séquence Coronale – Voyelle haute dans le mot '*Lundurais*' produit par le locuteur m01 pendant la première partie de l'interaction. La figure 3.9 montre la position de cette cible en référence à la forme d'onde et au spectre.

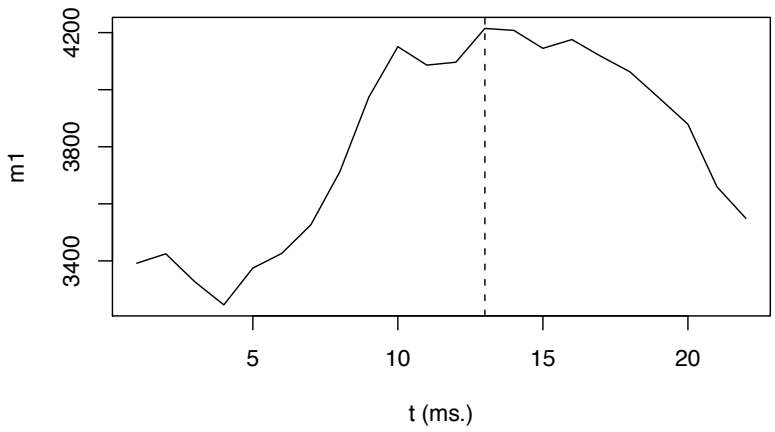

FIGURE 3.8 – Courbe du premier moment spectral m_1 pour la séquence Coronale – Voyelle haute du mot '*Lundurais*' produite par le locuteur m01 pendant la première partie de l'interaction. Le maximum est identifié par le trait en pointillés.

Une fois la cible localisée, l'information acoustique pertinente qui caractérise le segment a été mesuré par les 5 premiers coefficients DCT en ce point

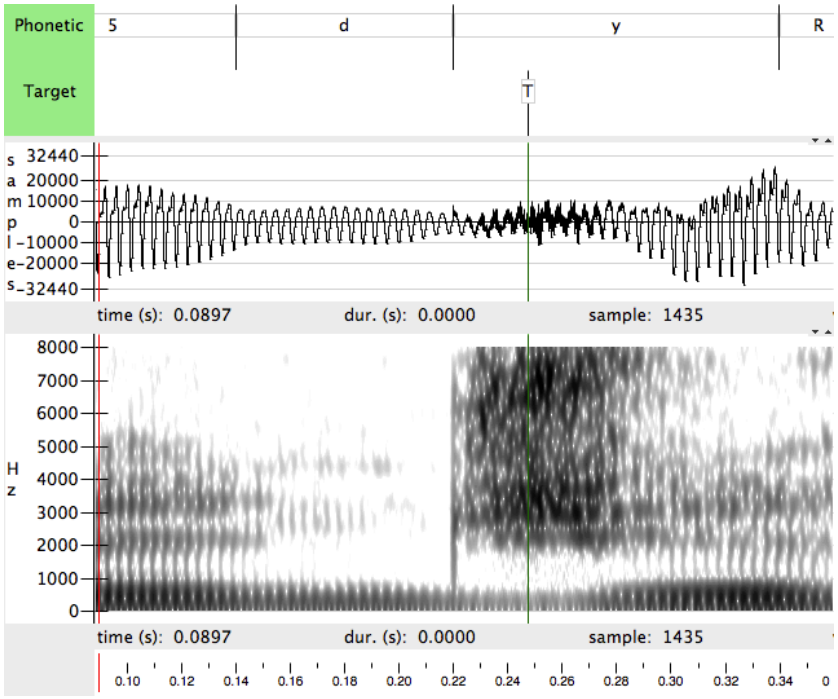

FIGURE 3.9 – Localisation du maximum du moment spectral m_1 pour la séquence Coronale – Voyelle haute du mot '*Lundurais*' produite par le locuteur m01 pendant la première partie de l'interaction. La séquence de variantes de prononciations choisie par l'aligneur est {dy}.

du signal, dans la plage spectrale 500–8000Hz.

Voyelles nasales

La localisation des cibles acoustiques pour les voyelles nasales a été effectuée de la même façon que pour les voyelles orales, en déterminant la partie stable de l'extrait de parole. Lorsque la voyelle était alignée avec une variante de prononciation composée de deux modèles, qui décrit alors les réalisations méridionales en séquence Voyelle orale – Consonne nasale, seule la cible associée à l'élément vocalique a été localisée.

3.9 Corpus

Nous présentons, pour finir ce chapitre, quelques données chiffrées sur notre corpus.

Le corpus comprend la réalisation de 8228 mots par 24 locuteurs. Sur ces mots, 61 ont été annotés comme erronés. Les mots erronés sont les mots qui ne sont pas complets, qui comportent du bruit, qui sont d'un niveau très faible ou dont la prononciation s'éloigne trop de la forme de citation. Ces mots totalisent la réalisation de 19479 segments, dont 380 (1.9%) ont été codés comme erronés. Les segments erronés sont les segments qui ont été exclus manuellement de l'alignement en variantes de prononciation. Ces prononciations erronées viennent pour la plupart d'une mauvaise conversion graphème–phonème par les locuteurs.

La table 3.8 présente le nombre moyen de répétition par locuteur de segments critiques associé à chaque dimension phonologique. Ces chiffres ne prennent pas en compte les réalisations des schwas finaux qui ont été écartés des analyses.

Dimension Phonologique	*N*
Schw.	26
Post.	78
Moy.	57
Cor.	108
Nas.	108

TABLE 3.8 – Nombre moyen de répétition par locuteur de segments critiques associé à chaque dimension phonologique.

La table 3.9 récapitule les critères utilisés pour la caractérisation acoustique des segments critiques.

Dimension Phonologique	Localisation	Mesure
Schw.	–	–
Post.	min. du signal de stabilité	5 coef. DCT sur la plage 200-4000Hz.
Moy.	min. du signal de stabilité	5 coef. DCT sur la plage 200-4000Hz.
Cor.	pic de m_1	5 coef. DCT sur la plage 500-8000Hz.
Nas.	min. du signal de stabilité	5 coef. DCT sur la plage 200-4000Hz.

TABLE 3.9 – Critères utilisés pour la caractérisation acoustique des segments critiques : localisation temporelle et type de mesure effectuée.

Chapitre 4

Classification automatique de la variété régionale

Ce chapitre présente les résultats de l'alignement automatique en variantes de prononciation des segments critiques qui composent les mot-cibles. Nous commençons par rappeler la problématique de ces analyses, puis nous présentons la méthode utilisée. Les résultats sont ensuite présentés puis discutés.

Une partie des résultats présentés dans ce chapitre a fait l'objet d'une publication dans la revue *Speech Communication* (Aubanel et Nguyen, 2010).

4.1 Problématique

L'objectif poursuivi est de savoir si les participants qui sont engagés dans une interaction conversationnelle adoptent les patrons de parole de leur partenaire, et en particulier sur la dimension de l'accent, décrit en termes de différences segmentales. Si c'est le cas, la procédure d'alignement forcé en variantes de prononciation, en effectuant un alignement indépendant de chacune des variantes de prononciation, doit capturer les évolutions qui se manifestent dans la parole. Des phénomènes similaires ont été observés dans différents cadres expérimentaux : Pardo (2006) a trouvé que la forme sonore de mot-cibles produit dans une interaction par deux participants étaient plus similaires à la fin de l'interaction qu'au début de celle-ci. La mesure de similarité était donnée par un groupe d'auditeurs, dans une tâche de discrimination AXB. On ne sait donc pas quels critères étaient utilisés par les auditeurs pour faire ces jugements.

Nous nous intéressons spécifiquement à ce qui se passe au niveau des catégories phonétiques, et en particulier si les changements, s'ils se produisent, provoquent un changement de choix de variantes de prononciation dans l'ali-

gnement. Une autre question d'intérêt que l'on peut examiner avec un alignement en variantes de prononciation est la caractérisation individuelle de chaque locuteur sur cette dimension. On peut ainsi dresser le profil accentuel des sujets, et voir par exemple si être locuteur d'un accent implique une adhésion aux prototypes de l'accent considéré, de façon égale sur les dimensions phonologiques individuelles.

Une autre question d'intérêt est de savoir dans quelle mesure des systèmes automatiques permettent de caractériser la parole en interaction conversationnelle.

4.2 Méthode

Nous avons évalué la variété régionale parlée par les participants à travers un questionnaire (section 3.5 et figure 6.1), puis nous avons recueilli leurs productions en interaction. Une question qui se pose est de savoir si les réalisations des variables phonologiques, telles que modélisées par l'alignement en variantes de prononciation, correspondent bien à cette évaluation de l'accent par le questionnaire. Nous avons donc entrepris de prédire la variété régionale parlée en fonction des alignements en variantes de prononciation. Nous avons pour cela entraîné un classifieur naïf de Bayes qui apprend les correspondances entre choix des variantes de prononciation et variété régionale attribuée, et permet de prédire la variété régionale sur présentation d'un nouvel ensemble de variantes de prononciation.

4.2.1 Classifieur naïf de Bayes

Le classifieur naïf de Bayes est une technique d'apprentissage supervisé qui attribue à une observation w définie par un vecteur d'attributs X une classe parmi un ensemble de classes C. L'attribution de la classe utilise la règle bayesienne d'affectation optimale, qui garantit le taux d'erreur de classification minimal en maximisant la probabilité *a posteriori* d'appartenance aux classes. Elle s'écrit

$$y(w) = \underset{C=c_1 \ldots c_q}{Argmax}\, P(C|X)$$

où $c_1, \cdots, c_q$ désigent les q classes, et $X = x_1, \cdots, x_p$ le vecteur des p attributs.

La probabilité conditionnelle $P(C|X)$ s'écrit, avec la règle de Bayes,

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

L'objectif étant de détecter le maximum de cette quantité en fonction de C, on peut supprimer le dénominateur, qui n'en dépend pas. La règle d'affectation se récrit

$$y(w) = \underset{C=c_1,...,c_q}{Argmax} \left[P(X|C)P(C)\right]$$

Avec l'hypothèse d'indépendance conditionnelle des attributs $x_1, ..., x_p$, qui implique que

$$p(x_i|C, x_j) = p(x_i|C), \qquad \forall i \neq j,$$

le terme $P(X|C)$ se trouve simplifié en

$$\begin{aligned} P(X|C) &= p(x_1|C)p(x_2, \cdots, x_p|C, x_1) \\ &= p(x_1|C)p(x_2|C, x_1) \cdots p(x_p|C, x1, x_2, \cdots, x_{p-1}) \\ &= p(x_1|C)p(x_2|C) \cdots p(x_p|C) \\ &= \prod_{i=1}^{p} p(x_i|C) \end{aligned}$$

Ainsi la fonction de décision se récrit en

$$y(w) = \underset{C=c_1,...,c_q}{Argmax} \left[P(C) \prod_{i=1}^{p} p(x_i|C)\right]$$

Les probabilités sont estimées en calculant sur le corpus d'apprentissage les proportions des observations x_i sur les modalités de C. La probabilité $P(C)$ est la proportion des éléments de chaque classe dans le corpus d'apprentissage : $p(c_k) = n_k/n$ où n_k est le nombre des observations de la classe c_k et n le nombre total des observations. En pratique, on utilise l'estimateur Laplacien des probabilités $p(c_k) = \frac{n_k+m}{n+mq}$ où m est le *m probability estimate*, en général égal à 1, et q le nombre de classes. Ce « lissage » permet d'éviter d'avoir des probabilités nulles $P(X|C)$ qui invalideraient le calcul de la fonction de décision.

Un des avantages du classifieur de Bayes, par rapport à un simple compte des différentes réalisations en variantes de prononciation est la possibilité de combiner les résultats de classification obtenus pour un vecteur d'attributs avec les résultats de classification obtenus sur d'autres vecteurs d'attributs (associés bien sûr à la même observation). On considère alors les différents résultats qui sont des probabilités, comme un nouveau vecteur d'attributs et on applique à nouveau la fonction de classification. Cette application à

plusieurs niveaux, dont le résultat final est par ailleurs identique à celui donné par une classification effectuée simultanément sur tous les vecteurs d'attributs associés à une observation, permet de contraster les performances de différentes combinaisons d'attributs, et d'identifier les plus performantes, par exemple.

Apprentissage

L'apprentissage du classifieur a donc consisté à calculer dans un premier temps la proportion d'alignement de chaque variante de prononciation associée à chaque segment critique (voir table 3.6), séparément pour les variétés FM et FS, qui constituent nos deux classes.

L'estimation des paramètres du classifieur consiste, à partir des proportions d'alignement de chaque variante de prononciation pour les deux classes de locuteurs, à établir le seuil utilisé pour la règle d'affectation optimale. Ce seuil est la valeur de probabilité sur la variante de prononciation qui maximise l'affectation *a posteriori* des locuteurs dans la classe à laquelle ils appartiennent, sur la base de la proportion de choix de cette variante de prononciation dans l'alignement. Ce seuil sépare sur la fonction de probabilité de densité cumulée les deux groupes de locuteurs en 2 parties égales. Un exemple de la détermination de ce seuil est donné dans la figure 4.1, pour le segment critique constitué de la voyelle moyenne postérieure en syllabe ouverte (B2).

Ainsi, à chaque variante de prononciation est associée un seuil de décision et un indice de confiance, donné par la formule de Bayes.

$$\begin{aligned} p(c_k|x) &= \frac{p(x|c_k)p(c_k)}{p(x)} \\ &= \frac{p(x|c_k)p(c_k)}{\sum_{i=1}^{q} p(x|c_i)p(c_i)} \end{aligned}$$

Les indices de confiance des variantes de prononciation associées à un segment critiques sont ensuite combinés pour fournir un indice de confiance représentatif du segment critique. Il est calculé par

$$p(C|X) = \frac{\prod_{i=1}^{N-1} p(x_i|C)}{\sum_{i=1}^{q} \prod_{j=1}^{N-1} p(x_j|c_i)}$$

où $x_1, \cdots, x_N$ est le vecteur décrivant les variantes de prononciation du segment critique. Dans notre cas $2 \leq N \leq 4$. On ne considère que les $N-1$

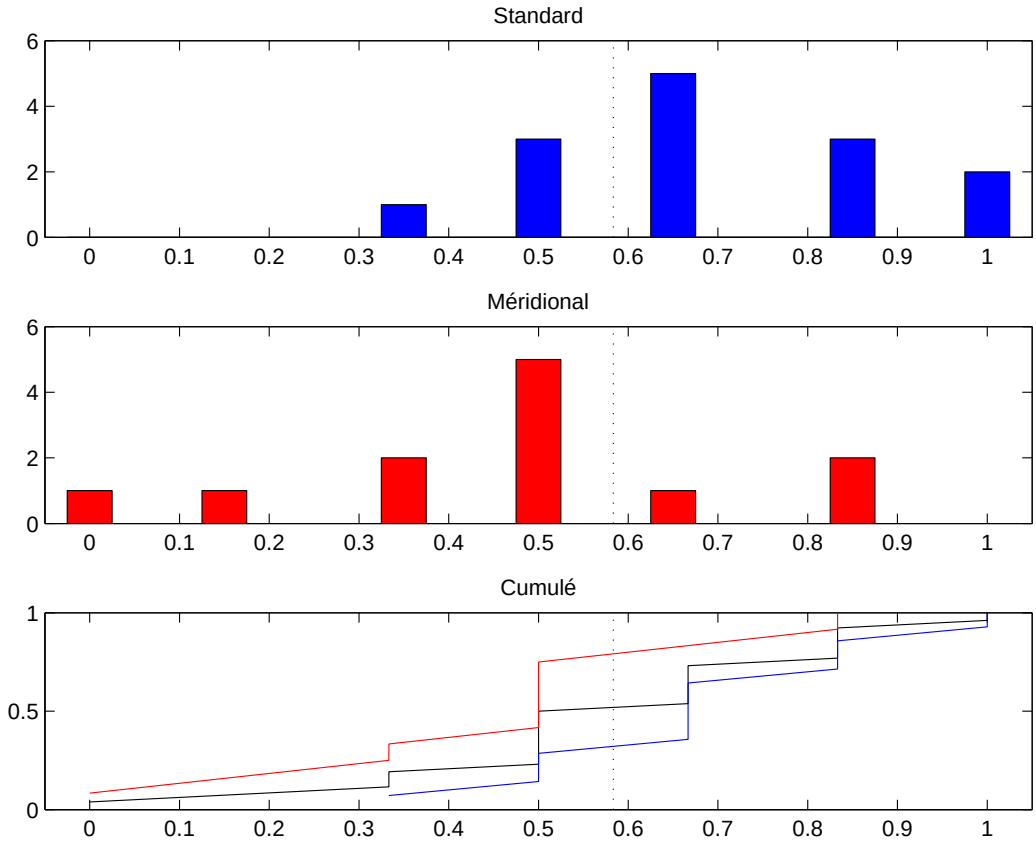

FIGURE 4.1 – Densité de probabilité pour la variante de prononciation {ɔ} du segment critique B2, pour les locuteurs du français standard (panneau du haut, couleur bleue) et méridionale (panneau du milieu, couleur rouge). La fonction de densité de probabilité cumulée pour les deux groupes de locuteurs est donnée dans le panneau du bas. Le seuil estimé pour la règle de décision est matérialisé par la ligne en pointillés.

premières distributions car la Nième est prédictible par les précédentes : si lors de l'alignement, le segment critique C1 (séquence /ti/) admet un alignement constitué de 20% de {ti} et 50% de {tʃi}, la proportion d'alignement sur la variante {tsi} peut se déduire à 30%.

Enfin, les probabilités *a posteriori* sont combinées à nouveau pour donner le score global du classifieur. Une fois les paramètres du classifieur estimés, on peut utiliser les probabilités *a posteriori* pour prédire l'appartenance d'un nouveau vecteur d'observation.

Test

Nous avons procédé à l'évaluation du classifieur sur l'ensemble des sujets par validation croisée (ou *leave-one-out*) : on choisit un locuteur dont on veut prédire la variété parlée à partir des alignements en variantes de prononciation, on effectue un apprentissage sur les variantes de prononciation de la façon exposée ci-dessus sur l'ensemble des locuteurs à l'exception du locuteur choisi, et on affecte une classe à ce locuteur d'après sa probabilité maximale d'appartenance *a posteriori* à cette classe.

Lors de l'élaboration du score global qui détermine l'attribution d'une classe à un locuteur, on effectue donc un premier test sur la première variante de prononciation, par exemple la variante {ə} du segment critique S1, pour fixer les idées. L'attribution d'une classe suivant ce premier test fournit déjà un indice de l'appartenance à l'une des deux variétés de français parlé, selon la probabilité associée à cette dimension. Par la suite, les probabilités de toutes les variantes seront combinées et fourniront le score global. Si un locuteur présente toutes les caractéristiques d'une variété donnée dans l'alignement en variantes de prononciation, alors les scores obtenus à chaque étape de classification ne feront que renforcer le score global, qui augmentera pour la classe en question. Si par contre les alignements suivent des distributions qui sont tantôt caractéristiques d'une variété, tantôt de l'autre, le score global restera proche du niveau de chance (50%).

Evaluation de la performance du classifieur

On obtient donc au terme du test par validation croisée des locuteurs deux attributions de classe pour chaque sujet : l'une est la classe *observée*, qui est la variété parlée par le locuteur évaluée par le questionnaire, l'autre est la classe *prédite* par le classifieur, qui est l'estimation de la variété parlée par le locuteur en regard de l'alignement en variantes de prononciation en comparaison de l'alignement des autres locuteurs des deux classes. On peut évaluer les performances du classifieur en comparant les écarts entre valeurs

observées et prédites. On construit pour cela la *matrice de confusion*, qui est la table de contingence pour les classes observées (dimension ligne) et les classes prédites (dimension colonne) :

		prédiction	
		S	M
observation	S	$n_{S,S}$	$n_{S,M}$
	M	$n_{M,S}$	$n_{M,M}$

A partir de cette matrice, on calcule la *précision* de la classification, qui est, pour chaque classe, le nombre d'éléments correctement prédits divisé par le nombre total d'éléments prédits de la classe. Pour la classe S,

$$precision = \frac{n_{S,S}}{n_{S,S} + n_{M,S}}$$

Le *rappel* est une mesure de qualité de la classification, donnée, pour chaque classe, par le nombre d'éléments correctement prédits divisé par le nombre total d'éléments observés de la classe. Pour la classe S,

$$rappel = \frac{n_{S,S}}{n_{S,S} + n_{S,M}}$$

Ces deux grandeurs sont des mesures indépendantes de la performance du classifieur : si celui-ci attribue systématiquement la classe S à tous les locuteurs testés, le rappel atteindra sa valeur maximale (on classe correctement tous les locuteurs S) pour la classe S. La précision est maximale lorsque le classifieur ne commet pas d'erreur d'affectation de classe sur la population prédite.

Finalement, on obtient une mesure globale de performance de la classification par la moyenne harmonique de ces deux grandeurs, appelée F-mesure :

$$F = 2\frac{precision.rappel}{precision + rappel}$$

La F-mesure vaut 1 lorsque la classification est parfaite (il y a identification entre la classe observée et prédite pour tous les locuteurs) et 0 lorsque tous les locuteurs sont classés dans la classe opposée. La valeur de 0.5 ($1/N$ dans le cas de N classes) correspond au niveau de hasard : le classifieur a les mêmes performances qu'une classification qui consisterait à attribuer aléatoirement une classe à un locuteur.

4.3 Résultats

Cette section présente les étapes de la construction du classifieur puis les résultats de classification sur l'ensemble des locuteurs sur les données d'alignement correspondant à l'interaction. Vient ensuite une comparaison des performances du classifieur sur les différentes phases de l'interaction, et avec les phases de pre- et post-test. Finalement, une comparaison de différents classifieurs construits à partir des alignements donnés par combinaison des deux ensembles de modèles acoustiques et diverses spécification de variantes de prononciation est donnée en fin de section.

4.3.1 Apprentissage des seuils de décision

Nous présentons ci-après quelques uns des 68 seuils de décision (un par variante de prononciation, voir table 3.6) pour la classe FS ou FM, calculés dans l'apprentissage d'un classifieur sur la totalité des sujets. Ces seuils sont déterminés par la probabilité maximale d'appartenance *a posteriori* aux classes FS et FM. Les 24 classifieurs entraînés pour la validation croisée ne diffèrent de celui-ci que par la mise de côté d'une proportion pour chaque variante de prononciation, correspondant au locuteur à tester. Le segment critique M1 a également été mis de côté pour l'apprentissage, car une seule variante de prononciation y était associée (voir paragraphe 3.3.1).

Les distributions des alignements en variantes de prononciation et les seuils de décision associés sont regroupés par dimension phonologique et par segment critique ci-dessous. La table 4.1 synthétise ces résultats en moyennant pour chaque groupe de locuteurs les fréquences d'alignement en variantes de prononciation, pour chaque segment critique et chaque dimension phonologique (les pourcentages présentés ci-dessous sont tirés de cette table).

Schwas

La figure 4.2 présente les distributions des alignements en variantes de prononciation pour le segment critique S1 (Schwa non final, graphié).

La variante de prononciation qui sépare le mieux les deux groupes de locuteurs est {ø} : si les locuteurs méridionaux sont relativement uniformément répartis sur cette dimension, ce segment critique n'a jamais été aligné avec cette variante pour une bonne bonne partie des locuteurs du français standard. Une répartition intéressante apparaît pour la variante {}, qui est la détection d'une absence de réalisation du schwa par l'aligneur. L'alignement suggère d'une part que les deux groupes de locuteurs admettent un alignement moyennement fréquent avec cette variante, celle-ci étant choisie un peu

DP	SC	var.	freq. moy. FS	freq. moy. FM	F-m.
Schw.	S1	{ə}	0.11	0.21	0.58
		{ø}	0.18	0.17	
		{œ}	0.05	0.02	
		{}	0.66	0.61	
Post.	B1	{ø}	0.30	0.15	0.71
		{œ}	0.01	0.01	
		{o}	0.53	0.77	
		{ɔ}	0.16	0.07	
	B2	{ø}	0.08	0.00	0.75
		{œ}	0.08	0.04	
		{o}	0.29	0.69	
		{ɔ}	0.55	0.27	
	B3	{ø}	0.14	0.04	0.66
		{œ}	0.38	0.21	
		{o}	0.10	0.23	
		{ɔ}	0.38	0.51	
Moy.	M2	{e}	0.69	0.86	0.75
		{ɛ}	0.31	0.14	
	M4	{ø}	0.11	0.05	0.20
		{œ}	0.11	0.08	
		{o}	0.28	0.40	
		{ɔ}	0.50	0.47	
Cor.	C1	{ti}	0.05	0.13	0.71
		{tʃi}	0.00	0.02	
		{tsi}	0.95	0.86	
	C2	{ty}	0.39	0.38	0.08
		{tʃy}	0.02	0.03	
		{tsy}	0.60	0.58	
	C3	{di}	0.52	0.72	0.83
		{dʒi}	0.03	0.10	
		{dzi}	0.45	0.18	
	C4	{dy}	0.60	0.74	0.50
		{dʒy}	0.22	0.19	
		{dzy}	0.18	0.08	
Nas.	N1	{ɛ̃}	0.23	0.09	0.75
		{ɑ̃}	0.46	0.15	
		{ɔ̃}	0.14	0.26	
		{ɛN}	0.16	0.36	
		{eN}	0.01	0.15	
	N2	{ɛ̃}	0.11	0.09	0.79
		{ɑ̃}	0.25	0.04	
		{ɔ̃}	0.39	0.44	
		{œN}	0.26	0.43	
	N3	{ɛ̃}	0.02	0.08	0.70
		{ɑ̃}	0.62	0.44	
		{ɔ̃}	0.36	0.38	
		{aN}	0.01	0.10	
	N4	{ɛ̃}	0.00	0.01	0.62
		{ɑ̃}	0.21	0.21	
		{ɔ̃}	0.68	0.51	
		{ɔN}	0.03	0.09	
		{oN}	0.08	0.17	

TABLE 4.1 – Fréquences moyennes d'alignement en variante de prononciation pour chaque segment critique (SC) associé à chaque dimension phonologique (DP). La F-mesure moyenne pour chaque segment critique est donnée dans la colonne la plus à droite. Noter que le segment critique B3 correspond au segment critique M3 dans le tableau 3.1 (voir section 4.3.1).

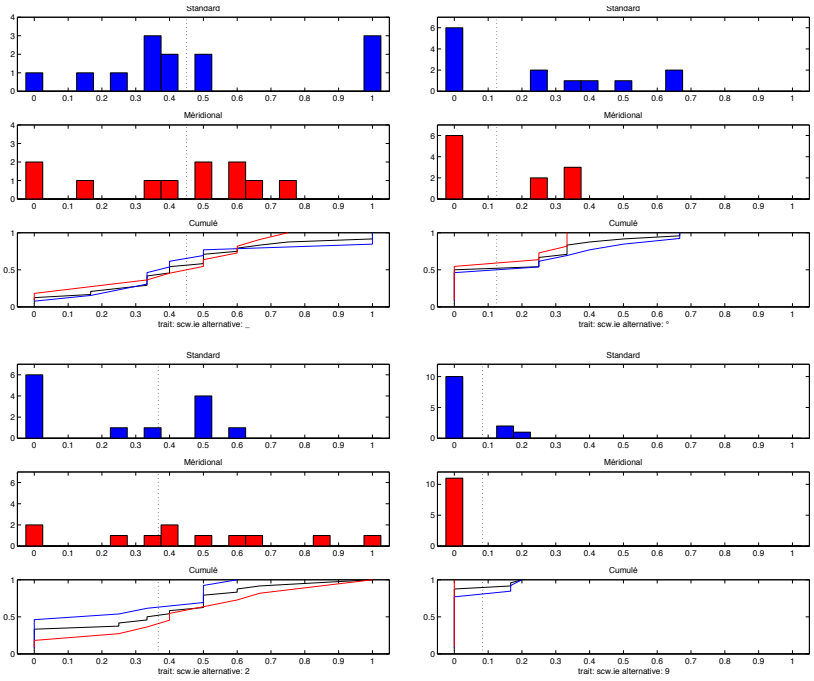

FIGURE 4.2 – Distributions des alignements en variantes de prononciation pour le segment critique S1. De gauche à droite et de haut en bas : {}, {ə}, {ø}, {œ}. Chaque panneau montre les distributions pour les locuteurs du français standard (en haut), du français méridional (au milieu) et les distributions cumulées (en bas). Le seuil de décision est représenté par un trait en pointillés.

plus fréquemment pour les locuteurs méridionaux, ce qui est contraire à la réalisation attendue. En revanche, pour 3 locuteurs du français standard, cette variante a *toujours* été choisie par l'aligneur, et pour 2 locuteurs du français méridional, n'a *jamais* été choisie suggère un statut phonologique spécial du schwa selon les locuteurs : il serait systématiquement non réalisé (vs. réalisé) pour une partie des locuteurs du français standard (resp. du français méridional) et adopterait un taux de réalisation plus variable pour une autre partie des locuteurs.

Globalement, le taux de non réalisation du schwa est plus important pour les locuteurs du FS (66%) que pour les locuteurs du FM (61%), ce qui correspond à la description faite dans la littérature, à savoir une élision plus fréquente en FS, et une grande variabilité de réalisation (par ex. Racine, 2008). Il semble de plus que cette variabilité s'étende aux locuteurs du français méridional, comme le montre le nombre important de non réalisation de cette variante pour 4 locuteurs. Les variantes de prononciation {ə} et {œ} n'ont que rarement été choisies par l'aligneur pour les deux groupes de locuteurs.

En ce qui concerne les schwas finaux (segments critiques S2, ⋯, S5), l'alignement se répartit globalement entre les variantes {} et {ø} pour les deux groupes de locuteurs, et dans une moindre mesure {ə}. Dans les 4 cas, la distribution des proportions de réalisation est similaire pour les deux groupes de locuteurs, mais un plus grand nombre d'alignement majoritaire en {} est obtenu pour le groupe de locuteurs du français standard. Par ailleurs, un nombre plus important de locuteurs du français standard que de locuteurs de français méridional admettaient un faible pourcentage d'alignement avec la variante {ø}. Cependant, comme noté à la section 3.8.1, la position finale de ce segment dans le mot-cible et l'impossibilité qui en découle de contrôler le contexte phonétique droit entraînait la confusion par l'aligneur avec d'autres segments de parole (pauses remplies et marques d'hésitation). Les schwas finaux ne sont donc pas inclus dans les analyses suivantes.

Voyelles moyennes postérieures

La figure 4.3 montre les distributions des alignements pour le segment critique B2 (voyelle moyenne postérieure en syllabe fermée). Ici, la différentiation des groupes de locuteurs se fait sur 3 variantes de prononciation : {o}, {o} et {ø}. La variante {œ} n'est utilisée que pour un locuteur, dans 15% de ses réalisations. La différence entre les groupes de locuteurs est caractérisée par un alignement plus fréquent avec les variantes {o} et {ø} pour les locuteurs du français standard, et un alignement plus fréquent en {ɔ} pour les locuteurs du français méridional.

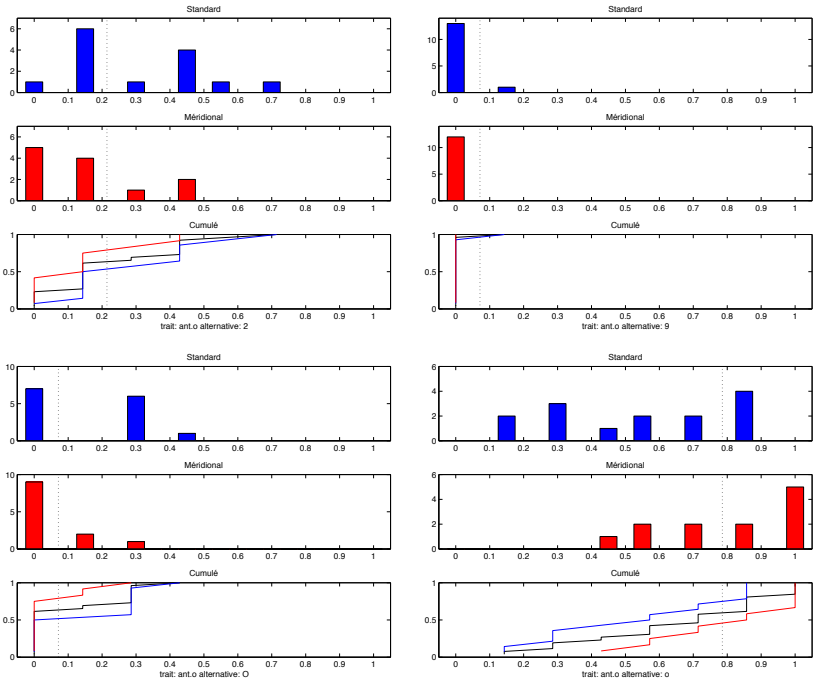

FIGURE 4.3 – Distributions des alignements en variantes de prononciation pour le segment critique B2. De gauche à droite et de haut en bas : {ø}, {œ}, {ɔ}, {o}. Chaque panneau montre les distributions pour les locuteurs du français standard (en haut), du français méridional (au milieu) et les distributions cumulées (en bas). Le seuil de décision est représenté par un trait en pointillés.

Pour le segment critique B2, la différence entre les deux groupes se fait principalement sur les variantes {o} et {ɔ}, et dans une moindre mesure {œ}, les locuteurs du français standard admettant un plus grand nombre d'alignement sur les variantes {ɔ} et {œ}, et les locuteurs du français méridional un plus grand nombre d'alignement sur la variante {o}.

Sur les deux segments critiques, la voyelle moyenne postérieure a donc tendance a être catégorisée comme antérieure par l'aligneur plus fréquemment pour les locuteurs du FS (33%) que pour les locuteurs FM (15%). Si ces distributions sont cohérentes avec l'antériorisation des voyelles moyennes postérieures relatée dans la littérature (Boula de Mareüil *et al.*, 2008; Martinet, 1958), elles révèlent en outre une différentiation des deux groupes de locuteurs sur une opposition de hauteur.

Voyelles moyennes

La figure 4.4 montre les fréquences d'alignement pour le segment critique M2, la voyelle moyenne antérieure en position finale de mot. On y voit

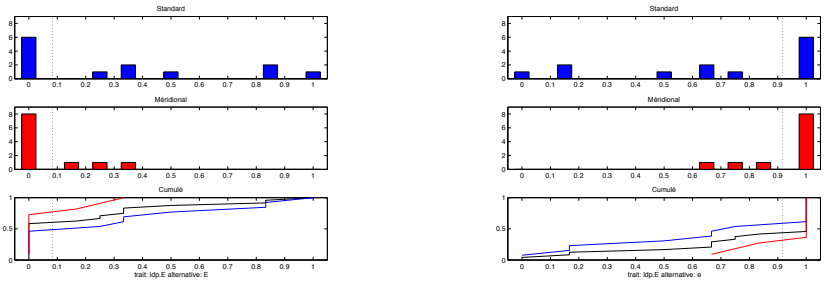

FIGURE 4.4 – Distributions des alignements en variantes de prononciation pour le segment critique M2. De gauche à droite : {ɛ}, {e}. Chaque panneau montre les distributions pour les locuteurs du français standard (en haut), du français méridional (au milieu) et les distributions cumulées (en bas). Le seuil de décision est représenté par un trait en pointillés.

que l'alignement majoritaire a été fait sur la variante {e} pour les deux groupes de locuteurs. Cependant, 7 locuteurs du français standard, contre 3 pour le groupe méridional admettaient un alignement plus fréquent pour la variante {ɛ}, ce qui permet de différencier les deux groupes. Globalement, un l'alignement était de 31% pour la variété standard et 14% pour la variété méridionale.

Pour le segment critique M3 (voyelle moyenne postérieure en syllabe fermée, graphiée 'o'), la différence entre les deux groupes se manifeste pour la

variante et {œ}, celle-ci étant plus fréquemment choisie pour les locuteurs du français standard, ce qui est cohérent avec une antériorisation de ce type de voyelle en français standard (nous n'avions pas de prédiction de différenciation sur la dimension de hauteur pour ce segment). Nous la comptabilisons en revanche dans la dimension phonologique *Post.*, comme segment critique B3.

Les variantes {o} et {ɔ} sont celles qui permettent le mieux de différencier les réalisations du segment critique M4 (voyelle postérieure en syllabe fermée, graphiée 'au'). Conformément à ce qui est attendu pour les locuteurs du français standard, les réalisations de ces derniers sont plus souvent alignés avec la variante {o} que les locuteurs du français méridional, et le contraire est vrai pour la variante {ɔ}. Les fréquences moyennes de ces alignements pour les deux groupes de locuteurs ne relatent cependant pas cette distinction, les réalisations des locuteurs de français méridional étant en moyenne alignés à 40% avec la variante mi-fermée contre 47% avec la variante mi-ouverte, alors que les réalisations des locuteurs FS sont plus souvent alignées avec la voyelle ouverte (50%) que mi-fermée (28%).

Séquences Coronales – Voyelles hautes

Pour les 4 segments critiques, la variante comprenant le modèle acoustique représentant une fricative postalvéolaire a été rarement choisie par l'aligneur pour modéliser une séquence constituée d'une plosive coronale suivie d'une voyelle haute (voir par ex. figure 4.5). En revanche, une différenciation claire des deux groupes des locuteurs apparaît en comparant les fréquences d'alignement entre les variantes composées d'une séquence de modèles acoustique plosive – voyelle, celle-ci étant plus fréquemment choisie pour les locuteurs méridionaux, et celles utilisant les séquence de modèles acoustique plosive – fricative alvéolaire – voyelle, plus fréquemment choisie pour les locuteurs du français standard. Dans le cas du segment critique C3, dont les distributions d'alignement sont présentées dans la figure 4.5, la variante avec friction alvéolaire a été choisie en moyenne 45% des fois pour les locuteurs de français standard et seulement 18% des fois pour les locuteurs du français méridional. Nous proposons une interprétation de ce résultat dans la discussion.

Voyelles nasales

Globalement, les variantes composées de deux segments étaient choisies plus fréquemment pour les locuteurs du français méridional (33%) que pour les locuteurs du français standard (14%). De plus, certaines particularités de manifestent dans la différenciation des deux groupes. Pour le segment cri-

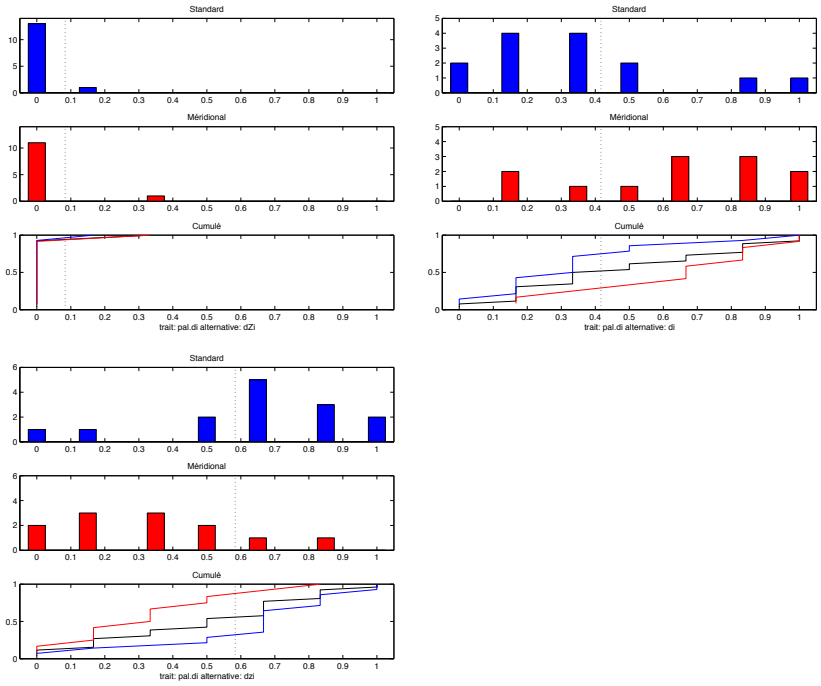

FIGURE 4.5 – Distributions des alignements en variantes de prononciation pour le segment critique C3. De gauche à droite, et de haut en bas : {dʒi}, {di} et {dzi}. Chaque panneau montre les distributions pour les locuteurs du français standard (en haut), du français méridional (au milieu) et les distributions cumulées (en bas). Le seuil de décision est représenté par un trait en pointillés.

tique N1 (voyelle nasale antérieure non arrondie, ou /ɛ̃/), les variantes de prononciation plus fréquemment choisies pour les locuteurs du français méridional étaient {ɛN} et {ɔ̃}. Les locuteurs du français standard étaient en revanche caractérisés par un alignement plus fréquent avec les variantes {ɑ̃} et {ɛ̃}. Pour le segment critique N2 (voyelle nasale antérieure arrondie), dont le fusionnement en français standard avec la voyelle nasale antérieure non arrondie est décrit comme achevé (Martinet, 1945; Malécot et Lindsay, 1976; Walter, 1976; Fagyal, 2006), c'est sensiblement la même différenciation qui est à l'œuvre : les alignements majoritaires en {ɛ̃} et {ɑ̃} sont caractéristiques des locuteurs du français standard. Pour la voyelle nasale ouverte (segment critique N3), la différenciation se fait sur les variantes {ɑ̃} et {aN}, les alignements avec la variante {ɑ̃} étant plus fréquents pour les locuteurs du français standard, et ceux avec la variante {aN} plus fréquents pour les locuteurs du français méridional. Enfin, la même différenciation apparaît pour le segment critique N4 (voyelle nasale postérieure), la variante {ɔ̃} étant plus souvent choisie pour les locuteurs du français standard, et la variante {oN} plus souvent choisie pour les locuteurs du français méridional.

4.3.2 Résultats du classifieur par validation croisée

La table 4.2 montre les probabilités normalisées pour chaque locuteur d'être catégorisé comme locuteur du français standard, obtenus par la classification par validation croisée pour les 24 locuteurs.

La probabilité normalisée P_n combine l'information de l'indice de confiance CL associée à l'attribution d'une classe et la valeur de cette classe. Elle est définie par

$$P_n = \begin{cases} CL & \text{si } c_k = S \\ 1 - CL & \text{si } c_k = M \end{cases}$$

Pour chaque dimension phonologique et chaque locuteur, l'indice de confiance est calculé en combinant les indices de confiance des segments critiques la composant :

$$CL_j = \frac{\prod_{i=1}^{n-1}(CL_i)}{\prod_{i=1}^{n-1}(CL_i) + \prod_{i=1}^{n-1}(1 - CL_i)}$$

où CL_j est l'indice de confiance associée à la j-ème dimension phonologique, et CL_i l'indice de confiance associé au i-ème segment critique. Seuls les $n-1$ premiers segments critiques sont pris en compte.

Le score global SG est donné en combinant les niveaux de confiance obtenus pour les dimensions phonologiques selon la formule :

Loc.	*Schw.*	*Post.*	*Moy.*	*Cor.*	*Nas.*	SG
s01	0.12	0.87	0.76	0.82	0.97	1.00
s02	0.20	0.27	0.73	0.99	0.00	0.09
s03	0.37	0.94	0.80	0.99	0.99	1.00
s04	0.69	0.99	0.76	0.87	0.99	1.00
s05	0.34	0.22	0.66	0.99	0.99	1.00
s06	0.12	0.29	0.22	0.96	1.00	0.99
s07	0.69	0.97	0.25	0.99	0.81	1.00
s08	0.69	0.99	0.66	0.38	1.00	1.00
s09	0.69	0.92	0.69	0.04	0.97	0.99
s10	0.55	1.00	0.82	0.99	0.91	1.00
s11	0.69	1.00	0.36	0.58	0.01	0.98
s12	0.34	0.92	0.82	0.00	0.88	0.27
F-m.	0.55	0.75	0.69	0.75	0.77	0.80

Loc.	*Schw.*	*Post.*	*Moy.*	*Cor.*	*Nas.*	SG
m01	0.31	0.99	0.41	0.97	0.01	0.93
m02	0.31	0.67	0.55	0.04	0.01	0.00
m03	0.69	1.00	0.92	0.46	0.75	1.00
m04	0.21	0.02	0.46	0.01	0.98	0.00
m05	0.69	0.02	0.50	0.25	0.02	0.00
m06	0.48	0.01	0.82	0.01	1.00	0.51
m07	0.21	0.01	0.41	0.19	0.12	0.00
m08	0.21	0.01	0.41	0.91	0.04	0.00
m09	0.31	0.00	0.89	0.04	0.90	0.00
m10	0.79	0.00	0.33	0.75	0.01	0.00
m11	0.79	0.00	0.33	0.04	0.03	0.00
m12	0.21	0.07	0.37	0.01	0.02	0.00
F-m.	0.62	0.75	0.64	0.75	0.73	0.78

TABLE 4.2 – Probabilité normalisée pour chaque locuteur d'être catégorisé comme locuteur du français standard, obtenues par la classification en validation croisée. Les cellules grisées contiennent des valeurs supérieures à 0.5.

$$SG = \frac{\prod_{j=1}^{m}(CL_j)}{\prod_{j=1}^{m}(CL_j) + \prod_{j=1}^{m}(1 - CL_j)}, m = 5$$

où CL_j est le niveau de confiance associé la j-ème dimension phonologique.

La matrice de confusion associée au classifieur est

		prédiction	
		S	M
observation	S	10	2
	M	3	9

Environ 79% des locuteurs, soit 19 sur 24, se voient attribuer la classe correcte par le classifieur. La précision associée au groupe des locuteurs du français standard (resp. du français méridional) est $p_S = 0.77$ (resp. $p_M = 0.82$). Le rappel associé au groupe FS (resp. FM) est $r_S = 0.83$ (resp. $r_M = 0.75$). Ces valeurs nous donnent par suite les F-mesures associées aux deux groupes : $F\text{-m.}_S = 0.80$ et $F\text{-m.}_M = 0.78$. La F-mesure moyenne évaluant la qualité du classifieur est 0.79.

4.3.3 Pouvoir de discrimination des segments critiques et dimensions phonologiques

La figure 4.6 montre le pouvoir de discrimination des segments critiques, donné par la F-mesure obtenue par le classifieur pour chacun d'entre eux. La figure montre aussi le pouvoir de discrimination de chaque dimension phonologique, obtenue de la même façon en combinant les scores obtenus pour les segments critiques la composant. Les segments critiques qui présentent le meilleur score sont C3, (F=.83), N2 (F=.79) et B2 (F=.75). Celles qui ont le score le plus faible sont C2 (F=.08), M3(F=.20) et C4 (F=.50). [1]

Ces scores sont les résultats de la classification effectuée sur chaque segment critique individuellement. Si l'on présente au classifieur les réalisations d'un locuteur uniquement sur les alignements de la variable C3, son appartenance à l'un des deux groupes de locuteurs sera correctement prédite avec un niveau de confiance de 83%. L'intérêt du classifieur est de pouvoir combiner ces contributions individuelles des segments critiques et construire ainsi un score global pour chaque dimension phonologique, et par la suite pour la totalité des réalisations.

4.3.4 Evolution des performances

Une question importante est de savoir dans quelle mesure la réalisation phonétique des mots d'un locuteur a tendance à ressembler à celles de l'autre locuteur pendant l'interaction. Pardo (2006) a trouvé que la similarité perçue de la prononciation augmentait au cours d'une conversation, pour un ensemble présélectionné de mots. Pour traiter cette question, nous avons entraîné le classifieur et évalué ses performances indépendamment pour chacun des 3 jeux successifs. Nous faisons l'hypothèse que si la similarité de prononciation entre les locuteurs augmente au cours de l'interaction, les performances du classifieur devraient diminuer, du fait de l'attribution par le

[1] Trois de ces scores semblent incohérents : le seul segment critique C3 donne un meilleur score que le classifieur global (F=.83 contre F=.79), et les scores de C2 et M2 indiquent une classification inversée (scores inférieurs au hasard). Ces phénomènes sont par construction impossibles. L'incohérence provient d'un biais de la procédurte de validation croisée : les segments C2 et M2 ayant un pouvoir classifiant proche du hasard, la suppression dans cette procédure d'une observation déséquilibre les données en faveur de la classe opposée, et induit l'attribution aberrante systématique à cette classe. Les scores de C2 et M2 en découlent, et le score global s'en trouve ainsi abaissé. Il est à noter que seule la valeur du score en est affectée, et non la performance effective du classifieur, ni les performances comparées sur différents ensembles de test. Une occurrence de ce biais est symptomatique de la concomitance d'une petite population et d'une variante faiblement discriminante : le classifieur naïf de Bayes est une procédure de classification asymptotiquement non biaisé.

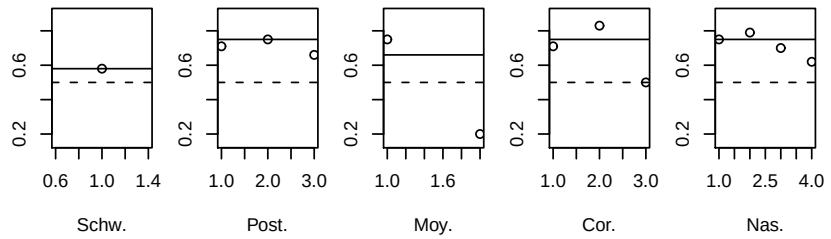

FIGURE 4.6 – F-mesures associées aux segments critiques et aux dimensions phonologiques. La ligne horizontale en trait plein indique le niveau de confiance pour chaque dimension phonologique. Le niveau de hasard est donné par la ligne en pointillés

classifieur à un locuteur donné la variété parlée de son partenaire. La table 4.3 montre les résultats du classifieur entraîné et testé sur chacune des jeux successifs composant l'interaction.

phase	pre	jeu1	jeu2	jeu3	post
F-m.	0.63	0.79	0.79	0.83	0.71

TABLE 4.3 – Performance du classifieur entraîné et testé sur chacun des 3 jeux successifs composant l'interaction, donné par la F-mesure moyenne entre les 2 groupes de locuteurs

Les résultats montrent que la performance du classifieur reste stable au cours des 3 jeux. Bien qu'ils n'écartent pas la possibilité que des effets de convergence aient pris place d'un jeu à l'autre, ces résultats suggèrent que de tels effets n'étaient pas suffisants pour que le classifieur bascule un locuteur dans le groupe de son partenaire au cours de l'interaction.

Une autre possibilité est que des effets de convergence se soient produits dès le début de l'interaction, comme dans l'expérience de Delvaux et Soquet (2007). Pour vérifier cette possibilité, nous avons évalué les performances de classification sur les données enregistrées dans le pre-test. Des effets de convergence au début de l'interaction devraient être caractérisés par une baisse de performance du classifieur entre le pre-test et le premier jeu. La table 4.3 montre que c'est précisément le contraire qui se produit : les performances sont meilleures dans le premier jeu (et généralement dans l'interaction) que dans le pre-test. L'examination des alignements en variantes de prononciation pour chaque groupe a révélé que la tâche de lecture utilisée dans le pre-test a peut-être eu une influence confondue sur la façon dont les

noms ont été produits, au moins pour un sous-ensemble des dimensions phonologiques. Par exemple, les deux locuteurs de chacune des dyades avaient tendance à produire plus de schwas interne durant le pre-test (53%) en comparaison avec l'interaction (37%).

4.3.5 Locuteurs mal classés

La table 4.2 montre que deux locuteurs du français standard (s02, s12) et trois locuteurs du français méridional (m01, m03, m06) ont été mal classés. L'examination individuelle de l'étiquetage effectué par l'aligneur suggère que ces locuteurs manifestent des caractéristiques associées aux deux variétés de français. Pour s02, le type de variante de prononciation choisie par l'aligneur pour les plosives coronales étaient celle décrivant une plosive alvéolaire, consistant avec notre caractérisation de la variété FS, mais des variantes de prononciation composées de deux segments pour les voyelles nasales, ce qui est consistant avec notre caractérisation de la variété FM. Pour le locuteur s12, l'alignement automatique est en accord général avec la prononciation FS pré-établie, mais s'en démarque fortement pour la réalisation des plosives coronales, qui a été catégorisée comme post-alvéolaire. Les voyelles nasales du locuteur m01 ont été traitées comme représentatives de la variété FM, mais ses voyelles moyennes postérieures étaient principalement associées avec des réalisations antériorisées, et ses plosives coronales associées avec une réalisation alvéolaire. m03 s'est vu attribuer des variantes typiques de la variété FS sur toutes les dimensions phonologiques sauf sur les plosives coronales. Les formes prononcées des mots du locuteur m06 sont catégorisées comme étant à mi-chemin entre la variété FS (voyelles nasales, distribution des voyelles moyennes en position finale de mot) et la variété FM (voyelles moyennes postérieures, plosives coronales).

4.3.6 Utilisation d'autres modèles acoustiques / variantes de prononciation

Les variantes de prononciation qui sont rarement choisies durant l'alignement appellent à évaluer des alignements utilisant différentes combinaisons de variantes de prononciation. Nous avons également voulu connaître les performances du classifieur en utilisant l'autre ensemble de modèles acoustiques à notre disposition. La présentation des 2 combinaisons de variantes de prononciation est présentée dans la figure 4.7. La figure présente dans les cases colorées les variantes de prononciation sur lesquelles les deux groupes de locuteurs présentent une différence de distribution significative, suivant le test non paramétrique de Kolmogorov-Smirnov. Lorsqu'une telle différence existe, elle

var. phon.	JPG_all							JPG_2		GG_all							GG_2	
ant.o	2	9	O	o				2	o	2	9	O	o				2	o
ant.O	9	2	O	o				9	O	9	2	O	o				9	O
ldp.o	o	2	9	O				2	O	o	2	9	O				2	O
ldp.O	9	2	o	O				9	O	9	2	o	O				9	O
ldp.e								e	E								e	E
ldp.E	E	e						E	e	E	e						E	e
pal.di	di	dzi	dZi					di	dzi	di	dzi	dZi					di	dzi
pal.dy	dy	dzy	dZy					dy	dzy	dy	dzy	dZy					dy	dzy
pal.ti	ti	tsi	tSi					ti	tsi	ti	tsi	tSi					ti	tsi
pal.ty	ty	tsy	tSy					ty	tsy	ty	tsy	tSy					ty	tsy
scw.ie	_	°	2	9				_	°	_	°	2	9				_	°
vn.1	1	§	@	9G				1	9G	1	§	@	9G				1	9G
vn.5	5	@	§	EG	En	EG	en	5	EG	5	@	§	EG	En	EG	en	5	EG
vn.@	@	§	5	aG	an			@	aG	@	§	5	aG	an			@	aG
vn.§	§	@	5	OG	On	oG	on	§	OG	§	@	5	OG	On	oG	on	§	OG
nb disc.							5		5							8		4

nb disc. : Nombre de variables phonologiques contenant au moins une alternative qui donne des distributions différentes.

Case colorée : la différence entre les distributions et significative (test de Kolmogorov-Smirnov positif, $p < .05$) et la proportion la plus élevée caractérise l'accent :
Standard
Méridional

FIGURE 4.7 – Comparaison des performances de différentes combinaisons d'alignements.

sépare les groupes de locuteurs dans une direction qui n'est jamais contradictoire avec la description phonologique des différences entre les accents. Cette comparaison met en évidence l'influence du choix du nombre de variantes de prononciation incluses dans l'alignement : la réduction des variantes de prononciation ne permet pas systématiquement une meilleure différenciation des deux groupes de locuteurs. D'autre part, les deux ensembles de modèles acoustiques apparaissent modéliser de façon similaire les différences entre les deux groupes de locuteurs, mais présente néanmoins des différences notables. Par exemple les modèles acoustiques de l'ensemble MA1 semble mieux capturer la différence entre les accents pour les voyelles moyennes, que l'ensemble MA2. En revanche, dans la modélisation des séquences plosive coronale – voyelle haute et pour les voyelles nasales non antérieures, ce sont ces derniers qui semblent plus à même de capturer les différences entre les deux accents.

4.4 Discussion

Les résultats présentés à la section précédente permettent tout d'abord de valider l'approche de la caractérisation de la variation phonologique régionale par l'alignement automatique en variantes de prononciation décrivant les variétés considérées. En général, les choix de variantes de prononciation

effectués par l'aligneur étaient en accord avec les descriptions phonétiques et phonologiques du français standard et du français méridional trouvées dans la littérature. Ces résultats apportent en outre des nouveautés dans la caractérisation des deux variétés. Dans une exploration à grande échelle sur la caractérisation automatique des accents régionaux du français, Boula de Mareüil *et al.* (2008) ont montré que la caractéristique phonétique qui permettait le mieux de distinguer les variétés du français standard et méridional était l'antériorisation des voyelles moyennes postérieures en français standard. Si nos résultats sont globalement consistants avec cela, nous avons observé que l'antériorisation des voyelles postérieures en français standard était dépendante de la position dans le mot de la syllabe qui contient la voyelle. Dans les syllabes fermées, la voyelle associée avec le graphème 'o' était effectivement plus fréquemment antériorisée en FS qu'en FM, mais dans une plus grande mesure lorsque la syllabe était en position finale de mot (FS : 52%, FM : 25%) qu'en position interne (FS : 16%, FM : 4%).

Pour les voyelles nasales, les choix de l'aligneur ont permis de distinguer les deux variétés, en associant plus fréquemment des séquences V+N pour les locuteurs du français méridional que pour les locuteurs du français standard. Les résultats sont en ligne avec ceux obtenus par Boula de Mareüil *et al.* (2007) qui emploient également une caractérisation de la variation des variétés standard et méridionale avec des variantes de prononciation incluant notamment des séquences voyelle nasale–consonne nasale. Les distributions des fréquences d'alignement obtenues ici peuvent en outre être mis en relation avec deux aspects de changement en cours décrits pour le français standard. Le premier est le fusionnement de la voyelle nasale antérieure non arrondie /ɛ̃/ et arrondie /œ̃/ au profit de /ɛ̃/ , prédit de longue date (Martinet, 1945; Malécot et Lindsay, 1976; Walter, 1976) et qui semble finalement achevé Fagyal (2006). Les fréquences d'alignement obtenues semblent ici suggérer qu'une opposition entre ces deux voyelles reste productive. En effet, pour les deux groupes de locuteurs, la voyelle arrondie est alignée le plus fréquemment avec la variante représentant la voyelle nasale arrondie postérieure {ɔ̃}. (/œ̃/ : 42%, /ɛ̃/ : 20%). Il faut noter ici que la voyelle /œ̃/ était présentée sous forme écrite dans les noms, et que les locuteurs ont pu, dans la conversion graphème–phonème du digramme *'un'*, être influencés par la valeur phonographique de voyelle arrondie associée au graphème *'u'* et aux digrammes qu'il forme en français (*'ou', 'eu',...*). Les modèles acoustiques utilisés ne comportant pas la variante de prononciation {œ̃}, il n'est peut-être pas surprenant de voir associée la prononciation de ce segment avec la variante {ɔ̃}, qui est la meilleure approximation d'une voyelle arrondie nasalisée. Les alignements confirment en outre un deuxième aspect d'évolution des voyelles nasales, à savoir le changement en chaîne /ɛ̃/ → /ɑ̃/ → /ɔ̃/ relaté pour le

français standard (Hansen, 2001; Fagyal, 2006). Si la variante {õ} est la plus fréquemment choisie pour les locuteurs du français standard pour la voyelle /õ/ (68%), cette variante arrive en deuxième position pour la voyelle /ã/ (36%, après {ã} : 62%), et, de façon plus significative, la variante la plus souvent choisie pour la voyelle /ɛ̃/ est {ã} (46%). Les locuteurs de la variété méridionale manifestaient des tendances similaires, bien que modulées par les fréquences d'alignement avec les variantes en deux segments.

Un autre résultat important donné par la caractérisation des variétés régionales à travers la modélisation par variantes de prononciation concerne la réalisation des plosives coronales devant une voyelle haute. La caractérisation qui est donnée dans la littérature décrit une palatalisation/frication de ces séquences pour le français méridional Binisti et Gasquet-Cyrus (2003); Woehrling et Boula de Mareüil (2006), dans les séquences comme *tu as* [tja]. Le phénomène que nous avons observé ici et qui sépare le mieux les deux variétés semble être mieux décrit en terme d'une différence de lieu d'articulation des plosives, postalvéolaire pour la variété FM et alvéolaire pour la variété FS. Dans les deux variétés, cette réalisation peut s'accompagner d'un bruit de friction qui adopte alors le lieu d'articulation de la plosive. De plus, les réalisations des locuteurs du français méridional ont plus souvent été associées avec des variantes sans bruit de friction que celles du français standard (fréquences supérieures de 10% en moyenne d'alignement en variante sans bruit de friction pour la variété FM). Les fréquences d'alignement obtenues pour la variété FS pointent vers un processus d'assibiliation, caractéristique du français du Québec (Walker, 1984) et relevé pour le français métropolitain chez des jeunes parisiens de banlieue en contact avec des langues d'immigration (Armstrong et Jamin, 2002). La séquence /di/ illustre le mieux ce phénomène, et de façon remarquable, c'est ce segment critique qui présente la meilleure contribution individuelle au score global de classification. A notre connaissance, le fait que la réalisation de cette séquence soit un marqueur régional si fort apparaît avoir été peu remarqué dans la littérature sociophonétique jusqu'ici.

La bonne performance générale du classifieur apporte des éclaircissements intéressants sur les facteurs sociolinguistiques qui pourraient influencer la parole des locuteurs. Il est en effet intéressant de noter qu'une majorité de nos locuteurs du français standard sont nés à Marseille, et que la plupart d'entre eux y ont vécu au moins 10 ans. Pourtant, ces derniers adoptent dans la prononciation des noms la plupart des caractéristiques phonétiques et phonologiques associées au français standard. Ces caractéristiques incluent le contraste /e/–/ɛ/ en position finale de mot, alors que cette opposition tend à disparaître dans le parler des jeunes parisiens Fagyal *et al.* (2002). Une explication possible à cela est qu'en plus d'une exposition à long terme

à une variété non native, les réalisations des locuteurs sont conditionnées par d'autres facteurs, comme l'identification à un groupe social que la variété native des locuteurs pourrait incarner. Ceci pourrait être d'autant plus vrai pour les locuteurs du français standard, étant donné que la variété FS est souvent considérée comme ayant un statut social plus élevé que la variété FM.

L'utilisation d'un classifieur naïf de Bayes, avec son hypothèse d'indépendance des attributs se prête particulièrement bien à la modélisation phonétique des variétés régionales. Premièrement, l'hypothèse d'indépendance est vérifiée techniquement par l'indépendance des observations recueillies dans les formes prononcées : le choix de catégorisation d'un segment critique en une variante de prononciation donnée n'a aucune influence sur le choix de catégorisation d'un autre segment critique, même si ces deux segments font partie du même mot. D'autre part, la caractérisation phonétique et phonologique d'une variété comme la sommation de descriptions de réalisations particulières d'un ensemble de variables phonologiques suit également cette hypothèse d'indépendance : il n'y a pas *a priori* de raison (articulatoire par exemple) de supposer que l'antériorisation des voyelles moyennes postérieures s'accompagne d'une réalisation moins fréquente de schwas. Précisément, un accent ne peut acquérir une identité que si cette hypothèse d'indépendance est vérifiée : c'est la cohérence sur ces dimensions individuelles qui contribueront à le démarquer d'autres variétés. Bien sûr, les frontières entre les variétés ne sont ni pre-établies ni fixées dans le temps, et toutes les caractéristiques ne présentent pas le même degré de salience perceptive et/ou acoustique (Labov, 2001, p. 78). L'approche utilisant le classifieur naïf de Bayes permet d'examiner d'une part le degré de cohérence globale d'un locuteur pour une variété et de fournir les scores intermédiaires associés à chacune des dimensions phonologiques décrivant la variété, et d'autre part d'estimer lesquelles de ces dimensions sont les plus salientes acoustiquement pour différencier les deux variétés. On a ainsi vu que les formes prononcées des mots de certains locuteurs présentaient des caractéristiques phonétiques et phonologiques associées simultanément aux deux variétés décrites, et que ces locuteurs se voyaient alors attribuer par le classifieur une variété parfois différente de celle établie par le questionnaire. La procédure de classification a aussi fourni un classement des 3 dimensions les plus discriminantes pour différencier les deux accents (réalisation des séquences coronales, voyelles moyennes postérieures, voyelles nasales), mais aussi un classement des segments critiques les plus discriminants. Cette interprétabilité aisée des scores de discriminabilité pourrait ouvrir des voies vers une évaluation comparée de salience acoustique/perceptive, en les mettant en relation avec les jugements d'auditeurs d'identification d'accent, comme ceux menés par Woehrling (2009).

Notre corpus était composé d'un échantillon relativement restreint des variétés de français standard et méridional. Les résultats montrent une différenciation robuste de ces deux classes globales, et suggèrent une caractérisation plus fine et linguistiquement pertinentes de chacune des variétés. On peut facilement étendre cette méthode à plus grande échelle pour caractériser de nouvelles variétés ou sous-variétés, dont les différences doivent pouvoir être décrites en terme de variantes de prononciation. Le clustering Bayesien (voir par ex. Heller, 2008), une technique d'apprentissage non supervisée qui utilise la même métrique de similarité entre les classes pourrait permettre par ailleurs de voir émerger des regroupements entre les locuteurs qui, grâce à l'interprétation facile des résultats sur les différentes dimensions, pourraient contribuer à dresser une image détaillée, contrastée et hiérarchique des multiples variétés parlées d'une langue.

Convergence

Une question d'intérêt de ce chapitre était d'évaluer les changements qui pourraient se produire dans les réalisations des locuteurs au cours de l'interaction avec leur partenaire. Bien qu'il y existe une distinction claire entre les deux groupes de locuteurs dans la façon dont les noms ont été produits, cela n'a pas empêché une influence mutuelle entre les locuteurs, particulièrement dans le cadre d'un jeu comme GMUP, qui nécessite une collaboration étroite ainsi qu'un transfert d'information efficace à travers la parole. L'imitation phonétique d'une variété non-native de la langue du locuteur a récemment été observée en français par Delvaux et Soquet (2007). Dans cette partie, nous nous sommes proposés d'évaluer des effets de convergence potentiels en examinant les performances du classifieur entre les trois jeux successifs. Nous avons fait l'hypothèse qu'un plus grand nombre de cas d'un locuteur étant incorrectement associé avec le groupe de son partenaire, au fil du développement de l'interaction, pouvait être attribué à en effet de convergence entre les partenaires. Nos résultats n'ont pas révélé une telle tendance, car il n'y a pas eu de changement substantiel dans les performances du classifieur d'un jeu sur l'autre. Il y a cependant une possibilité que des effets de convergence se soient produits à un niveau plus fin, sub-catégoriel, que l'alignement forcé du signal de parole en catégories phonétiques n'a pas permis de capturer. Il se pourrait également qu'une convergence se soit produite dès le début du premier jeu, ce qui serait consistant avec les résultats de Delvaux et Soquet (2007), bien que leur étude ait été conduite dans un cadre expérimental non interactif. Ceci est d'autant plus vraisemblable du fait que les locuteurs des deux groupes ont une connaissance passive de la variété de leur interlocuteur : les locuteurs du français standard enregistrés sont exposés quotidiennement

à de la parole de la variété méridionale, habitant dans la région, et les locuteurs du français méridional se voient eux exposés à la variété standard à travers les médias, dans lesquels le français standard est la variété la plus utilisée. Dans un tel scénario, les scores de classification devraient être plus élevés pour le pre-test en comparaison avec le premier jeu. Cependant, nos analyses ont révélé précisément le contraire. Les scores de classification plus faibles dans le pre-test peuvent être attribués, en partie, à la tâche de lecture utilisée dans ce pre-test. Une meilleure façon d'identifier des effets de convergence immédiats serait d'enregistrer les locuteurs en interaction avec un partenaire de la même variété dans une phase préliminaire. Des analyses plus approfondies sont nécessaires pour savoir si une convergence phonétique s'est produit dès le début de l'interaction, ou à un niveau plus fin, ou aux deux.

4.5 Conclusion

Nous avons présenté dans ce chapitre une procédure de classification bayesienne qui permet de prédire l'appartenance des locuteurs à la variété de français parlé standard ou méridional, à partir des résultats de l'alignement des réalisations des variables phonologiques en variantes de prononciation.

Dans un premier temps, le classifieur apprend la correspondance, pour chaque segment critique composant chaque dimension phonologique, entre les proportions de choix d'alignement en variantes et chaque variété. Cet apprentissage est effectué sur les réalisations des variables phonologiques de l'ensemble des locuteurs. Dans un deuxième temps, cette information est utilisée pour prédire individuellement l'appartenance des locuteurs à l'une ou l'autre des deux variétés. Les informations sont combinées sur l'ensemble des segments critiques appartenant à la même dimension phonologique, puis combinées à nouveau sur l'ensemble des dimensions phonologiques pour décider de l'attribution de la variété parlée au locuteur. Dans son ensemble, le classifieur permet de prédire correctement la variété parlée pour 79% des locuteurs, soit 19 sur 24.

La variation phonologique régionale en interaction conversationnelle peut ainsi se modéliser par une caractérisation en variantes de prononciation. Cela a été possible grâce à un contrôle fin en amont du matériel devant être prononcé par les locuteurs, et au développement d'outils d'alignement dédiés.

Cette caractérisation permet, au delà d'identifier correctement les variétés FS et FM, de dégager des propriétés plus fines sur la dimension de l'accent des locuteurs. Il a ainsi été possible de dégager l'importance relative des variables phonologiques dans la distinction des variétés, en mettant par

exemple en évidence le statut particulier de la variable phonologique /di/ dans la différenciation des deux variétés, un trait peu remarqué dans la littérature sociophonétique. Un maintien de l'opposition /ɛ̃/–/œ̃/ a également été observé dans les deux variétés, ainsi qu'un changement en chaîne des voyelles nasales.

L'alignement en variantes de prononciation des variables phonologiques des deux participants engagés dans une tâche interactionnelle n'a pas révélé un changement catégoriel de celles-ci, au cours de la vingtaine de minutes que durait l'interaction. Cela semble indiquer une relative stabilité des représentations associées à ces formes sonores, porteuses entre autres d'informations socio-indexicales. Nous avons cependant évoqué la possibilité que des phénomènes d'adaptation se soient produits dès le début de l'interaction, et n'aient pas pu être évalués en raison du style de parole de lecture impliqué par la tâche de contrôle de pre-test.

Une autre possibilité est que des modifications plus fines de ces formes sonores se soient produites mais n'aient pas été capturées par l'alignement catégoriel en variantes de prononciation. Cette question est l'objet du chapitre suivant.

Chapitre 5

Convergence sub-phonémique

5.1 Problématique

Les résultats du chapitre précédent ont appelé à dépasser le niveau catégoriel des alignements en variantes de prononciation, pour évaluer si les formes prononcées des noms des locuteurs se modifiaient lorsque ceux-ci sont engagés dans une interaction avec un partenaire. Nous rappelons que l'attribution d'une variante de prononciation à une portion de signal de parole par la procédure d'alignement forcé se fait en comparant les vecteurs de traits calculés sur cette portion de signal avec les vecteurs de traits associés aux variantes de prononciation candidates spécifiées dans le dictionnaire de prononciation. La variante de prononciation choisie est celle qui présente la distance minimale avec la portion de signal. On voit ainsi que plusieurs portions de signal peuvent correspondre à la même variante de prononciation, du moment que leurs distances avec cette variante de prononciation restent inférieures à celles calculées avec les autres variantes de prononciation incluses dans le dictionnaire. Ainsi, une modification progressive d'une portion de signal de parole associée par l'aligneur à une variante de prononciation en direction d'une autre variante de prononciation ne se verra attribuer cette variante de prononciation que lorsque la différence des distances aux deux variantes dépassera un seuil, alors que les réalisations qui resteront en deçà de ce seuil se verront toujours attribuer la première variante.

Pardo (2006) a mis en évidence une augmentation de la similarité des formes de mot prononcés au cours de l'interaction. On ne sait cependant pas quels indices étaient utilisés par les auditeurs pour faire ces jugements (qualité segmentale, contour prosodique, intensité, hauteur de voix,...). Nous nous proposons d'évaluer les paramètres acoustiques qui permettent de caractériser les deux accents pour tenter de voir si une convergence s'est produite

sur la dimension phonétique segmentale des accents. D'autres dimensions différencient ces variétés, comme le contour intonatif par exemple (Coquillon, 2005) et d'autres paramètres acoustiques comme la fréquence fondamentale ou l'intensité pourraient exister, mais nous nous sommes limités à ceux détaillés dans la partie 3.8.

Delvaux et Soquet (2007) ont de leur côté examiné l'effet que l'exposition à un accent non-natif peut avoir sur la production de la parole en français parlé en Belgique, et ont trouvé des patrons d'imitation phonétique à un niveau sub-phonémique entre les accents. Les analyses portaient sur des mot-cibles insérés dans des phrases porteuses, dans un cadre expérimental non-interactif. Babel (2009) a aussi montré, également dans un cadre non-interactif, des patrons de convergence phonétique dans une tâche de shadowing sur des mots isolés, en employant les formants comme mesure acoustique.

Enfin, Bailly et Lelong (2010) ont récemment mis en évidence des phénomènes d'ajustement phonétique entre des participants à un jeu interactif. Ils ont développé une méthode originale de mesure de la convergence à partir de techniques de reconnaissance automatique de la parole, qui consistait à entraîner des modèles acoustiques sur la parole des interactants dans une phase de pre-test puis à les utiliser pour aligner les productions de leur partenaire dans la phase de l'interaction. La convergence est alors évaluée d'après les performances de reconnaissance des modèles du partenaire. Si les résultats présentaient des différences interindividuelles marquées, le sexe ainsi que la familiarité des locuteurs semblaient jouer un rôle déterminant dans les patrons de convergence.

Nous présentons dans la section suivante les mesures acoustiques utilisées pour obtenir une image plus détaillée des réalisations des locuteurs, la métrique utilisée pour mesurer la convergence et les outils statistiques employés à cette fin.

5.2 Méthode

5.2.1 Mesures acoustiques

La localisation des mesures acoustiques pertinentes est présentée en détails dans la section 3.8. Les mesures acoustiques utilisées pour caractériser les segments critiques ont été effectuées en relevant les 5 premiers coefficients DCT dans une plage de spectre spécifique : entre 200 Hz et 4000 Hz pour les sons vocaliques et entre 500 Hz et 8000 Hz pour les séquences plosives coronales – voyelles hautes. Le premier coefficient de la DCT a été écarté

car il encode l'énergie moyenne du signal, et ne porte donc pas d'information pertinente concernant la forme du spectre.

5.2.2 Métrique pour la convergence

Nous définissons ici la convergence par la variation simultanée de deux quantités : le rapprochement vers une cible et l'éloignement d'un point de référence. Le point de référence décrit les productions du locuteur tandis que la cible est caractéristique des productions de son interlocuteur. Il est donc souhaitable d'avoir un moyen de représenter des informations décrivant avec pertinence les réalisations des locuteurs d'une façon qui permette de visualiser simultanément ces variations. La technique d'analyse discriminante linéaire peut nous aider en cela. Elle peut être vue comme une technique de réduction de la dimension [1]. Elle cherche, dans l'espace des données, une direction de projection qui maximise la variance inter-classes. Précisément, elle construit dans cet espace les centres de gravité des sous-ensembles des données associées à chacune des classes, et définit le plan discriminant comme le plan de dimension minimale passant par ces centres de gravité. Celui-ci possède une dimension de moins que le nombre de classes. La projection des données se fait alors sur ce plan, orthogonalement à celui-ci. Le plan discriminant possédant une dimension de moins que le nombre de classes, la réduction de dimension obtenue est de $d - c + 1$ si d est la dimension de l'espace des données et c le nombre de classes. Dans le cas d'un nombre de classes égal à deux, la projection se fait selon un hyperplan et le plan discriminant est une droite. Appliquée à notre cas de convergence entre deux locuteurs, l'analyse discriminante associera un point sur une droite à chaque observation composée des 4 coefficients DCT caractérisant la production d'un des locuteurs. Par construction, les locuteurs seront maximalement différenciés sur cette droite. Une variation sur cette droite du centre de gravité de la première classe à la deuxième peut donc, conformément à notre définition de la convergence, s'interpréter comme un éloignement d'une observation de la première classe et un rapprochement de la deuxième classe.

On peut ensuite examiner les différences entre les classes soit en comparant les centres de gravité des classes, mais selon la dimensionalité et le type des données l'interprétation n'est pas aisée, ou en examinant directement les coefficients des variables canoniques du plan discriminant, qui informent sur la contribution individuelle de chaque dimension.

[1] Vue comme une technique de classification, cette technique est identique à un classifieur Bayesien : elle attribue à une observation la classe qui maximise sa probabilité *a posteriori* d'appartenance.

Pour toutes les analyses présentées, la construction de la variable canonique a été calculée à partir des données du pre-test. Dans cet espace, par construction, la distance inter-groupes est maximale. Les données des autres phases des enregistrements sont ensuite transposées dans cet espace de référence en appliquant la variable canonique, c'est-à-dire en effectuant une combinaison linéaire de chaque vecteur de données avec les coefficients de la variable canonique. Une réduction de la distance entre les deux nuages de points signifie donc un rapprochement des réalisations entre les groupes, étant donné l'importance attribuée à chacune des dimensions des données (les coefficients de la variable canonique) qui ont servi pour la construction de l'espace de référence. Les coefficients ont été ajustés pour les données du pre-test, on s'attend donc naturellement à un rapport entre variance intra-groupe et variance inter-groupes moins bon pour les autres phases de l'expérience, les coefficients n'étant pas « optimisés ». Une comparaison de la distance inter-groupes entre le pre-test et tout autre ensemble de données sera donc systématiquement biaisée, en donnant en général une distance plus grande pour le pre-test. En revanche, les comparaisons entre les phases différentes du pre-test ne présenteront pas cet effet d'optimisation et pourront donc être comparées sans problème.

Nous avons fait le choix de fixer un espace de projection des données multidimensionnelles par l'évaluation de la variable canonique pour le pre-test. Cette contrainte d'un espace fixe peut sembler mener à des artefacts de mesure. Imaginons par exemple que deux nuages de points dans l'espace à deux dimensions des formants F_1, F_2 décrivent des réalisations centrées autour des voyelles [u] et [o] respectivement. L'axe discriminant passe donc par ces deux voyelles, et la distance séparant la projection des deux nuages de points sur cet axe est une indication de la séparation entre ces deux voyelles. Si maintenant on cherche à évaluer la distance entre le nuage de point centré autour de [u] et un nouveau nuage centré autour de [i], en utilisant les données projetées sur l'axe discriminant précédemment calculé, la distance ainsi obtenue sera nulle (si on simplifie en imaginant que [u] et [o] ont la même valeur de F_2, et que [u] et [i] on la même valeur de F_1), ce qui n'est pas représentatif de la distance séparant les deux nuages de points dans l'espace original à 2 dimensions.

Il faut donc garder à l'esprit les implications de prendre une mesure de distance sur des données projetées. Dans cet exemple, la deuxième mesure peut se formuler comme la quantité de différence entre [u] et [o] présente entre les nuages de points [u] et [i], ce qui revient, dans une interprétation phonétique, à la différence d'antériorité qui sépare les voyelles. Il est donc moins surprenant de constater qu'elle est nulle pour les nuages représentant [u] et [i] et maximale pour les nuages représentant [u] et [o]. Si l'axe discrimi-

nant est dans cet exemple facilement interprétable car identifié avec l'axe F_1, il n'en est pas de même avec les coefficients DCT qui décrivent la forme du spectre. Il a cependant été construit de façon à décrire une séparation maximale entre les deux locuteurs à partir de paramètres acoustiques, et il est donc intéressant de s'en servir comme référence. Notons que les paramètres décrivant cet axe sont bien évidemment propres à chaque segment critique, cette nécessité étant illustrée par l'exemple caricatural ci-dessus, mais ils sont aussi, pour chaque segment, propres à chaque dyade : rien n'empêche que les dimensions acoustiques qui discriminent le mieux les réalisations des /ti/ pour deux membres d'une dyade soient différents pour deux membres d'une autre dyade, en raison par exemple de différence de bruit de frication trouvé chez les différents locuteurs, et qui peut se manifester entre autres par des différences dans le premier coefficient DCT, qui encode la composante linéaire du spectre, ou pente spectrale.

En fixant un axe discriminant de référence, on fixe un type de convergence. Le cas envisagé jusqu'ici suppose que les deux groupes de référence sont les deux locuteurs membres de la dyade. Cela évalue le degré de convergence entre les deux locuteurs, sur une échelle de distance qui est propre à chaque dyade. On peut voir cette analyse comme l'évaluation d'une convergence « vocale » entre les locuteurs : les cibles sont spécifiques à chaque locuteur, et sont dérivées directement de paramètres acoustiques. Une interprétation phonétique de ces cibles ne coule pas de source, et peut se faire de façon indirecte en examinant les contributions de chaque dimension acoustique à la variable canonique. Ces analyses sont présentées à la section 5.3.1.

Un autre type de convergence qu'il est intéressant d'examiner est un rapprochement non pas mesuré directement entre les réalisations des interlocuteurs, mais entre les réalisations de chaque locuteur et celles du groupe dont fait partie son interlocuteur. Les deux classes sont alors constituées, pour chaque segment critique évalué, par les réalisations de l'ensemble des locuteurs du français méridional d'une part et du français standard d'autre part. Les réalisations des deux locuteurs sont ensuite transférées dans cet espace. Par comparaison avec une évaluation de la convergence qui construit l'axe discriminant directement à partir des réalisations des locuteurs, cette analyse nous permettra de répondre à la question suivante, concernant la cible de la convergence : la cible est-elle constituée par les réalisations de l'interlocuteur, indépendamment du caractère prototypique de celles-ci par référence à la variété parlée, ou au contraire est-elle constituée par les cibles caractéristiques de la variété parlée, telles que mesurées sur plusieurs locuteurs ? Dans ce dernier cas, on peut plus facilement imaginer un mécanisme qui impliquerait une évaluation de la variété parlée du locuteur (de façon consciente ou inconsciente). Dans le formalisme des modèles à exemplaire, cela reviendrait

à l'activation de poids associés aux exemplaires qui soient également associés à la variété perçue. Ces analyses font l'objet de la section 5.3.2.

Finalement, un autre type de convergence similaire au précédent peut être évalué en prenant comme classes de référence les réalisations prototypiques des deux variétés parlées telles qu'elles sont décrites dans la littérature. Pour le cas de la voyelle moyenne antérieure (segment critique M1), la première classe est constituée des réalisations [e] de tous les locuteurs, et la deuxième classe des réalisations [ɛ] de tous les locuteurs. Cette analyse formalise davantage l'axe discriminant, qui est déterminé en ajustant les paramètres qui discriminent au mieux les réalisations typiques des deux variétés. Pour la voyelle moyenne, on peut considérer que l'axe discriminant représente un continuum [e]–[ɛ]. Notons que cette analyse fait usage d'une information supplémentaire par rapport aux deux analyses précédentes, qui sont les résultats d'alignements obtenus au chapitre précédent. La différence avec l'analyse précédente est que celle-ci évalue une convergence vers la variété opposée telle qu'elle est formalisée dans les descriptions phonétiques des deux variétés, par rapport à une convergence vers la variété opposée telle qu'elle est parlée par les membres de cette variété. Cette analyse est présentée dans la section 5.3.3.

5.2.3 Modèles mixtes

Les modèles mixtes sont une technique récente de modèles statistiques qui permettent de traiter des designs à mesures répétées (voir par ex. Baayen, 2008). Les facteurs à niveaux non répétables, comme le locuteur ou le mot et qui sont tirés aléatoirement d'une grande population, sont modélisés par une variable aléatoire de moyenne nulle et de variance inconnue (estimée par le modèle). Les facteurs à effet fixes sont les facteurs à niveaux répétables, qui sont les variables indépendantes dont on veut tester l'influence dans le design, et sont modélisés au moyen des contrastes, comme dans les modèles linéaires traditionnels. Nous avons utilisé le package `lme4` du logiciel `R`.

5.3 Résultats

Nous présentons dans cette section les résultats des analyses statistiques qui permettent d'évaluer les différents types de convergence présentées ci-dessus : convergence vers l'interlocuteur (section 5.3.1), convergence vers le groupe de l'interlocuteur (section 5.3.2), et convergence vers l'accent parlé par l'interlocuteur (section 5.3.3). Pour chaque analyse, nous comparons les résultats obtenus avec une seconde analyse qui écarte certaines données qui peuvent potentiellement limiter l'observation d'un rapprochement. Ces données ont

été identifiés expérimentalement au chapitre précédent. Les segments critiques C2, M3, C4, N4 et M4, sont les moins informatifs dans l'attribution d'une variété aux locuteurs (voir section 4.3.3) ; ils ont par conséquent moins de chance de constituer une dimension sur laquelle les locuteurs peuvent se rapprocher. Par ailleurs, l'appartenance à la variété standard ou méridionale des locuteurs m01, s02, m03, m06 et s12 a été mal évaluée par le classifieur, ce qui limite potentiellement l'observation d'un rapprochement entre les réalisations de ces locuteurs avec celles de leurs partenaires. L'ensemble des réalisations des dyades a donc été également écarté de ce sous-ensemble.

5.3.1 Convergence vers l'interlocuteur

Les classes utilisées pour la construction de l'axe discriminant de référence sont les deux locuteurs membres de la dyade : pour chaque segment et chaque dyade, l'axe qui passe par les centres de gravité des nuages de points associés aux deux locuteurs est calculé sur les données du pre-test, et les données des phases suivantes (jeu 1, jeu 2, jeu 3 et post-test) sont transférées dans cet espace. La figure 5.1 montre une diminution de la distance séparant les réalisations des deux locuteurs au cours des 3 phases de l'interaction (3 panneaux centraux), qui est plus précisément le fait d'un rapprochement des réalisations du locuteur m04 vers celles du locuteur s04. Ce rapprochement culmine à la troisième et dernière phase de l'interaction, pour laquelle l'écart interquartile est identique pour les deux locuteurs, ce qui signifie que leurs productions sont indistinguables, dans l'espace qui différencie le mieux les locuteurs. Il est aussi intéressant de noter que la séparation des nuages de points s'écarte à nouveau dans la phase de post-test, lorsque les sujets sont enregistrés séparément juste après l'interaction, mais cette distance n'atteint pas, par construction, la valeur obtenue dans le pre-test.

Nous nous sommes intéressés à évaluer la constance de ces effets sur l'ensemble des segments pour chaque dyade. Nous avons donc testé l'existence d'une relation entre le score discriminant et le temps, pris pour la phase d'interaction. Plus précisément, les temps d'occurrence des segments critiques ont été normalisés à une plage de 0 à 3, décrivant les 3 phases de l'interaction : les temps de la phase jeu1 s'étendent de 0 à 1 ; ceux de la phase jeu2 de 1 à 2, et ceux de la phase jeu3 de 2 à 3. Cette transformation permet de comparer les différentes phases de l'interaction pour l'ensemble des dyades, certaines d'entre elles passant plus de temps que d'autres à résoudre la tâche interactive. Une observation individuelle des schémas a suggéré pour certains d'entre eux un rapprochement maximal pour la deuxième phase de l'interaction, et non systématiquement pour la troisième comme présenté dans la figure 5.1, ce qui nous a conduit à tester une relation sur différentes plages

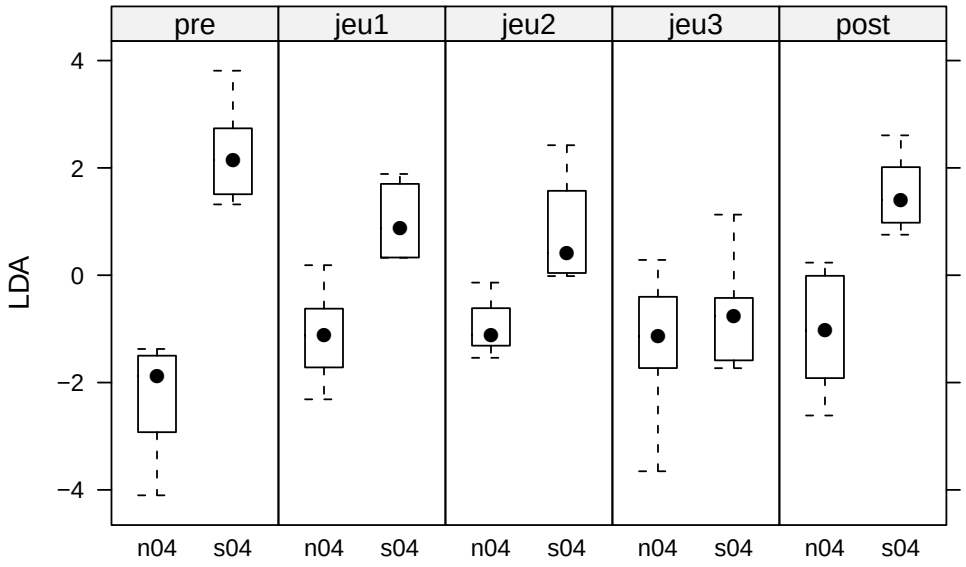

FIGURE 5.1 – Boxplots des scores discriminants associés aux réalisations du segment N4 pour les deux locuteurs de la dyade d04 au cours des 5 phases d'enregistrement : pre-test, jeu 1, jeu 2, jeu 3, post-test. Les phases d'interaction sont les 3 phases du milieu.

temporelles : chaque phase prise séparément (jeu1, jeu2, jeu3), la plage temporelle décrivant les deux premiers jeux, (jeu1–jeu2) les deux dernier jeux (jeu2–jeu3), et la totalité de l'interaction (jeu1–jeu2–jeu3).

La figure 5.3 montre les scores sur la variable canonique pour la dyade d01, pendant les deux premiers jeux. La droite de régression est tracée pour chaque locuteur séparément, et pour chaque segment. On s'intéresse ici en particulier à l'interaction entre les facteurs **locuteur** et **temps** : si celle-ci est nulle, c'est que la distance entre les nuages de points associés aux deux locuteurs n'évolue pas au fil de l'interaction. Si elle est positive, c'est que la distance augmente et si elle est négative, la distance diminue et on peut conclure à une convergence.

Nous avons ajusté à ces données de la dyade d01 pendant les deux premiers jeux un modèle mixte qui autorise un intercept et une pente aléatoire par segment, ainsi qu'un intercept aléatoire par mot contenant les segments. L'équation du modèle est `lda` $\sim$ `tnorm * spkr + (tnorm | seg) + (tnorm | mot)`. Les effets aléatoires introduits pour les segments et les mots permettent de normaliser les décalages entre les niveaux de ces facteurs. La variable canonique est en effet calculée indépendamment pour chaque segment, et chaque mot comporte un contexte phonétique différent. Ces facteurs présentent bien des niveaux non répétables : ceux-ci sont tirés d'une grande population et l'ajout d'un nouveau niveau implique la prise en compte d'une moyenne et d'une variance inconnue associée à ce niveau.

Les estimations des effets fixes du modèle pour la dyade d01 pour les deux premiers jeux sont donnés ci-dessous :

```
                Estimate Std..Error t.value
(Intercept)      -2.9163     0.6669  -4.373
tnorm             0.9403     0.5216   1.803
spkrs01           4.1435     0.5627   7.364
tnorm:spkrs01    -1.3063     0.4450  -2.935
```

Par construction de l'axe discriminant, les scores sont centrés autour de 0, mais pour cette dyade l'intercept estimé est significativement négatif [2]. La moyenne des scores associés aux réalisations du locuteur m01 est significativement supérieure à ceux du locuteur s01 ce qui est également attendu, car par construction l'analyse discriminante cherche l'axe qui différencie maximalement les réalisations des deux locuteurs. Ce qui est en revanche remarquable

[2]L'implémentation actuelle des modèles mixtes dans le package `lme4` ne fournit pas de p-value associée aux estimations des effets fixes, notamment en raison de l'incertitude à déterminer le nombre de degrés de libertés. On peut néanmoins estimer la significativité d'un effet en observant sa t-value, et conclure à la significativité si cette valeur dépasse 2 en valeur absolue. (voir Baayen, 2008, p. 269)

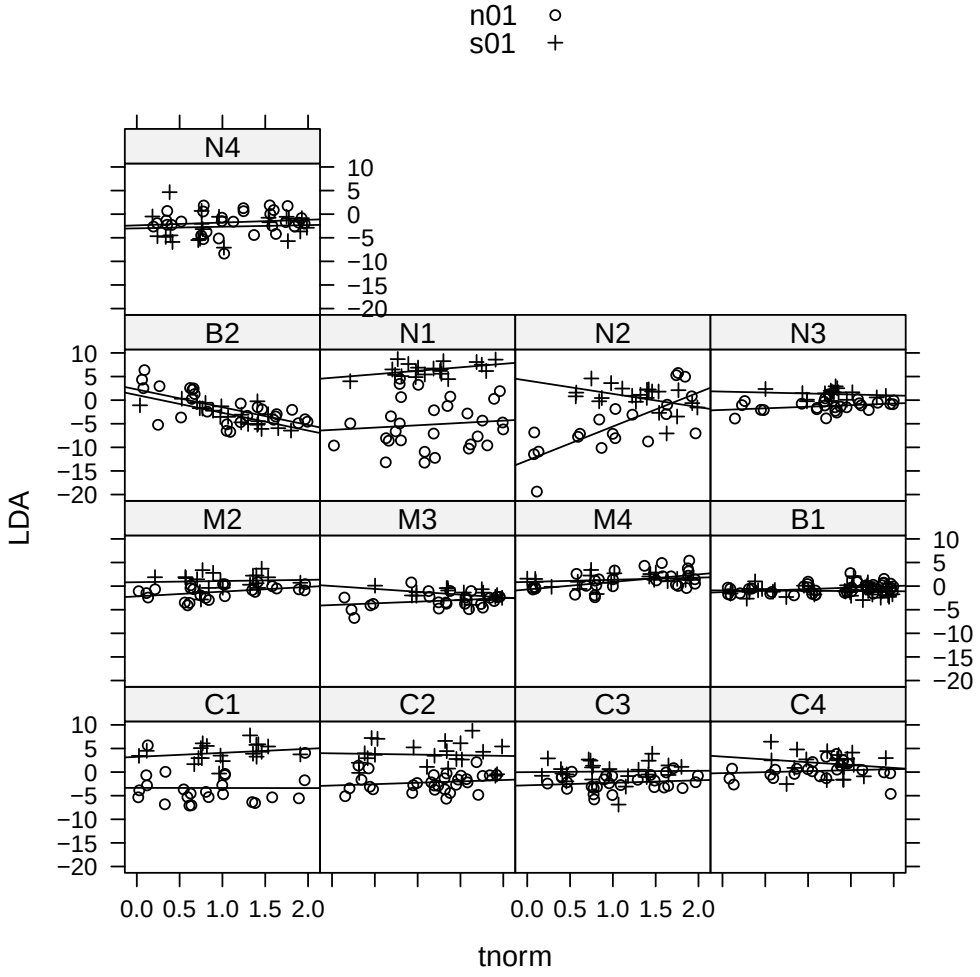

FIGURE 5.2 – Score des locuteurs de la dyade d01 pour chaque segment en fonction du temps, pour les deux premiers jeux. La droite de régression calculée pour chaque locuteur est représentée.

est l'interaction entre le facteur **temps** et le facteur **locuteur**, donné par la dernière ligne des estimations des effets fixes, et qui est significativement négative ($t = -2.935$). Cette interaction correspond à un rapprochement des réalisations des locuteurs au fil du temps, du début du jeu 1 à la fin du jeu 2. Les droites de régression obtenues par combinaison des coefficients du modèle et des effets aléatoires sont présentées dans la figure 5.3. Le rapprochement se vérifie par l'écart de pente entre les droites de régression associées aux deux locuteurs.

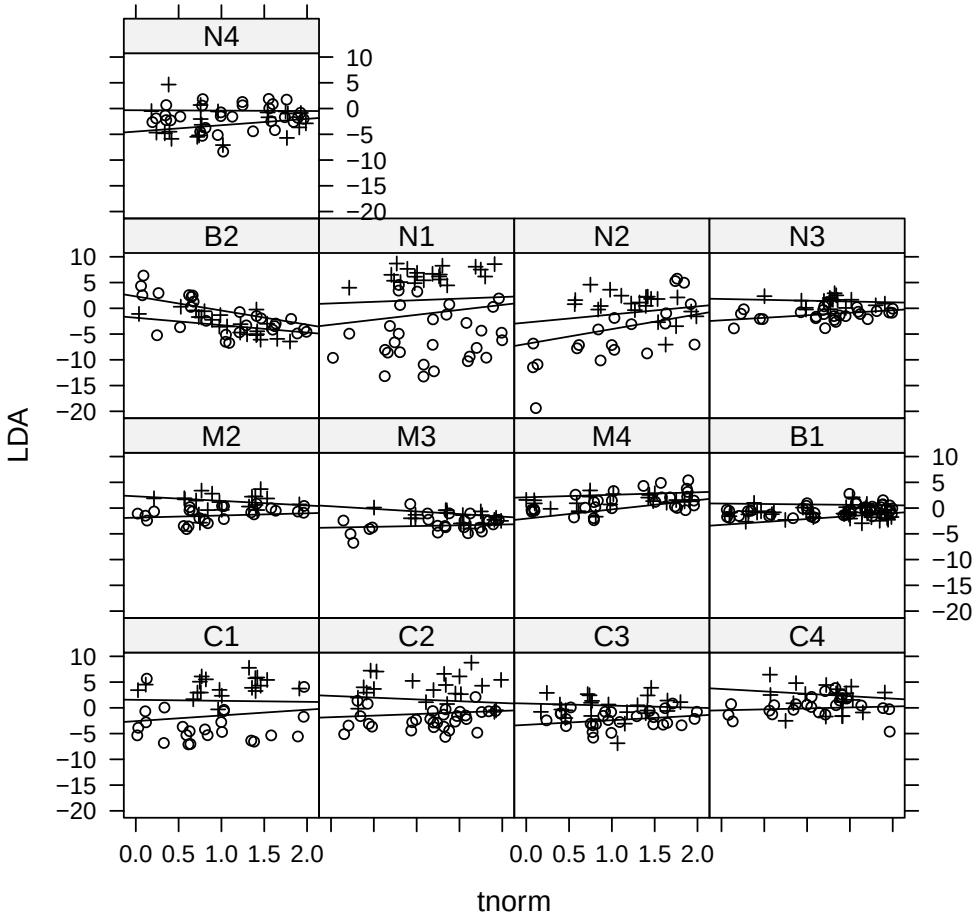

FIGURE 5.3 – Droites de régression associées au modèle mixte pour la dyade d01, pour l'ensemble des segments critiques et sur les deux premiers jeux.

Ce modèle a été calculé pour chaque combinaison de dyade et de phase temporelle. La t-value de l'interaction entre les facteurs **temps** et **locuteur** pour chaque modèle est reportée dans le tableau 5.1. On retrouve la valeur présentée ci-dessus dans la première ligne, à la quatrième colonne.

	jeu1	jeu2	jeu3	jeu1–jeu2	jeu2–jeu3	jeu1–jeu2–jeu3
d01	1.002	-3.223	0.124	-2.935	0.695	-0.729
d02	-1.135	-0.955	-0.42	1.465	1.216	3.275
d03	0.877	-0.964	0.301	0.119	-0.328	0.665
d04	-1.339	0.463	0.325	-1.001	-0.022	-1.164
d05	0.347	0.593	0.958	-0.405	1.512	0.517
d06	0.568	1.092	0.607	0.544	-0.547	-0.3
d07	-0.115	-0.036	0.552	0.589	-1.43	-0.994
d08	0.987	0.364	-0.17	2.745	0.827	3.399
d09	-0.344	-0.536	-0.297	-1.436	1.329	0.699
d10	1.505	-1.347	-0.112	0.263	-1.378	-0.608
d11	2.367	-1.036	1.023	-2.489	2.784	-1.206
d12	-1.293	-0.07	-0.212	-0.749	1.558	1.012
moyenne	0.286	-0.471	0.223	-0.274	0.518	0.38
t-value	0.848	-1.415	1.59	-0.594	1.402	0.833
p-value	0.414	0.185	0.14	0.565	0.188	0.423

TABLE 5.1 – t-values de l'interaction entre les facteurs **temps** et **locuteur** du modèle mixte calculé pour chaque combinaison de dyade et de plage temporelle. Une valeur négative indique un rapprochement des réalisations, une valeur positive un éloignement. Les 3 dernières lignes présentent les résultats d'un t-test effectué pour chaque plage temporelle.

Ce tableau montre qu'un rapprochement significatif des réalisations n'est observé que pour 3 combinaisons de dyade et de plage temporelle : la dyade d01 durant le jeu 2 ($t = -3.223$), la dyade d01 pendant les deux premiers jeux ($t = -2.935$) et la dyade d11 pendant les deux premiers jeux ($t = -2.489$). A l'inverse, des valeurs d'interaction positives et supérieures à 2 sont observées pour 3 dyades, à savoir les dyades d02, d08 et d11, et différentes plages temporelles. Un t-test a été conduit pour chaque plage temporelle, pour évaluer les tendances des valeurs d'interaction sur l'ensemble des dyades, pour ces différentes plages. Comme le montrent les p-values, aucune tendance claire n'émerge de ces analyses, les valeurs moyennes se distribuant autour de 0. Cela signifie que lorsqu'on considère les tendances sur l'ensemble des dyades et l'ensemble des segments, on ne peut conclure ni à un rapprochement

ni à un éloignement consistant des réalisations entre les membres des dyades.

Le tableau 5.2 montre les t-values de l'interaction entre les facteurs **temps** et **locuteur** calculées après avoir écarté les données limitant potentiellement l'observation d'un rapprochement, identifiées au chapitre précédent.

	jeu1	jeu2	jeu3	jeu1–jeu2	jeu2–jeu3	jeu1–jeu2–jeu3
d04	-1.703	0.321	0.554	-1.185	-0.485	-1.529
d05	0.122	-0.694	0.483	-1.59	0.002	-1.212
d07	0.705	0.001	-0.283	1.238	-1.692	-0.324
d08	0.406	1.07	-0.389	1.724	0.606	2.26
d09	-0.557	-0.531	-0.383	-2.435	1.3	-0.551
d10	1.704	-1.684	-0.399	-0.236	-0.675	-0.56
d11	0.591	0.756	2.118	-3.167	4.009	-1.569
moyenne	0.181	-0.109	0.243	-0.807	0.438	-0.498
t-value	0.446	-0.305	0.693	-1.173	0.629	-1.003
p-value	0.671	0.771	0.514	0.285	0.552	0.355

TABLE 5.2 – t-values de l'interaction entre les facteurs **temps** et **locuteur** du modèle mixte calculé pour chaque combinaison de dyade et de plage temporelle, en écartant les données potentiellement problématiques. Une valeur négative indique un rapprochement des réalisations, une valeur positive un éloignement. Les 3 dernières lignes présentent les résultats d'un t-test effectué pour chaque plage temporelle.

Une t-value négative et supérieure à 2 en valeur absolue, indiquant un rapprochement des réalisations, est observée pendant les deux premiers jeux pour les dyades d09 (t=-2.435) et d11 (t=-3.167). Un éloignement des réalisations est constaté pour la dyade d11 durant le jeu 3 ($t = 2.118$) et durant les deux derniers jeux ($t = 4.009$), ainsi que pour la dyade d08, sur la totalité de l'interaction (t=2.26). A nouveau, en observant les valeurs des interactions calculées séparément pour chaque dyade, il n'y a pas de tendance globale vers la convergence ou la divergence, sur les différentes combinaisons des phases temporelles.

Il est intéressant de voir l'évolution des scores discriminant en regroupant les données non par dyade mais par segment. L'équation du modèle devient `lda ~ tnorm * accent + (tnorm | dyade) + (tnorm | mot)`. Ce modèle évalue les relations existant entre les scores discriminants d'une part et le temps normalisé et les locuteurs, regroupés par accent, d'autre part. Un effet aléatoire est autorisé par dyade et par mot. Les t-values associées à l'interaction entre les facteurs **temps** et **accent** sont présentées dans le

tableau 5.3.

	jeu1	jeu2	jeu3	jeu1–jeu2	jeu2–jeu3	jeu1–jeu2–jeu3
C1	-0.152	-1.604	1.347	0.485	-1.833	-0.165
C2	0.481	0.111	-0.891	1.238	-0.203	1.219
C3	-0.189	2.044	-0.534	0.411	1.12	0.661
C4	0.541	0.631	1.319	-0.779	2.041	0.723
M2	1.131	-0.887	0.755	0.831	-0.318	0.744
M3	0.253	-1.298	-0.133	-0.541	-0.183	-0.111
M4	1.45	-0.458	-0.61	0.649	-0.12	0.469
B1	-0.008	-1.962	-1.218	-2.615	0.212	-0.879
B2	0.244	0.815	-1.455	-2.037	1.399	0.206
N1	0.699	-1.568	-0.918	0.867	-1.087	1.301
N2	-1.068	-1.623	0.067	-3.827	0.447	-3.154
N3	1.533	-0.172	-0.242	-0.145	0.471	0.438
N4	-0.872	2.006	2.287	1.529	1.816	2.425
moyenne	0.311	-0.305	-0.017	-0.303	0.289	0.298
t-value	1.412	-0.807	-0.056	-0.673	0.94	0.821
p-value	0.183	0.435	0.957	0.513	0.366	0.428

TABLE 5.3 – t-values de l'interaction entre les facteurs **temps** et **accent** du modèle mixte calculé pour chaque combinaison de segment et de plage temporelle. Une valeur négative indique un rapprochement des réalisations, une valeur positive un éloignement. Les 3 dernières lignes présentent les résultats d'un t-test effectué pour chaque plage temporelle.

Les segments B1, B2 et N2 présentent des t-values négatives et supérieures à 2 en valeur absolue pour la phase décrivant les deux premiers jeux ($t = -2.615$, -2.037 et -3.827 resp.). C'est aussi le cas du segment N2 pour la totalité de l'interaction ($t = -3.154$). Les segments qui présentent des t-values significativement positives concernent principalement les phases jeu2 et suivantes : pour le segment C3, $t = 2.044$ pour le jeu 2, pour C4, $t = 2.041$ pour les deux derniers jeux. Le segment N4 présente une t-value positive et supérieure à 2 pour le jeu 2 ($t = 2.006$), le jeu 3 ($t = 2.287$) et pour la totalité des jeux ($t = 2.425$). Cette tendance à un rapprochement des réalisations au début de l'interaction et à un éloignement vers la fin de l'interaction ne se vérifie cependant pas globalement sur l'ensemble des segments, comme le montrent les p-values associées au t-test effectué pour chacune des phases temporelles.

Le tableau 5.4 montre les t-values de l'interaction entre les facteurs **temps**

et **accent** calculées par le modèle après avoir écarté les données potentiellement problématiques, identifiées au chapitre précédent.

	jeu1	jeu2	jeu3	jeu1–jeu2	jeu2–jeu3	jeu1–jeu2–jeu3
C1	0.226	-0.66	-0.15	-0.219	-1.956	-1.579
C3	0.239	1.931	-1.007	0.508	1.216	1.224
M2	0.568	-1.623	1.187	-0.438	-0.009	0.089
B1	0.895	-0.071	-0.616	-0.756	-0.138	-0.622
B2	-1.134	3.23	-1.297	-0.585	1.081	-0.12
N1	0.646	-0.119	-0.555	-1.039	1.048	0.523
N2	-2.176	-0.791	0.411	-3.387	0.797	-2.645
N3	1.053	-1.872	-0.837	-0.137	-1.849	-1.233
moyenne	0.04	0.003	-0.358	-0.757	0.024	-0.545
t-value	0.1	0.005	-1.241	-1.846	0.052	-1.241
p-value	0.923	0.996	0.254	0.107	0.96	0.255

TABLE 5.4 – t-values de l'interaction entre les facteurs **temps** et **accent** du modèle mixte calculé pour chaque combinaison de segment et de plage temporelle, en écartant les données potentiellement problématiques. Une valeur négative indique un rapprochement des réalisations, une valeur positive un éloignement. Les 3 dernières lignes présentent les résultats d'un t-test effectué pour chaque plage temporelle.

Les t-values de l'interaction significativement négatives concernent le segment N2, pour les phases jeu1 ($t = -2.176$), les deux premiers jeux ($t = -3.387$), et la totalité de l'interaction ($t = -2.645$). Seul le segment B2 présente une t-value de l'interaction positive et supérieure à 2, pour la phase jeu2 ($t = 3.23$).

Un t-test par phase temporelle sur l'ensemble des segments montre une légère tendance à un rapprochement des réalisations durant les deux premiers jeux pour l'ensemble des dyades (après avoir écartées celles dont un des locuteurs avait été mal classé par le classifieur), montré par une moyenne négative ($M = -0.757$), avec un indice de confiance de 90% seulement ($p = 0.107$).

Finalement, un modèle global a été ajusté sur l'ensemble des données, en regroupant les dyades et les segments. Ce modèle global combine les résultats obtenus par les modèles précédents. Il autorise une pente aléatoire par segment et par locuteur. L'équation du modèle est `lda` $\sim$ `tnorm * accent + (tnorm | seg) + (tnorm | spkr) + (tnorm | mot)`. Les résultats de ce modèle sont présentés dans le tableau 5.5, qui montre la t-value de l'interac-

tion entre le facteur **temps** et **accent** qui est une indication de l'évolution comparée des scores des locuteurs au cours du temps.

	jeu1	jeu2	jeu3	jeu1–jeu2	jeu2–jeu3	jeu1–jeu2–jeu3
Modèle I	0.846	-0.166	-0.506	-1.035	0.874	-0.246
Modèle II	-0.002	-0.697	-0.614	-1.565	-0.243	-1.553

TABLE 5.5 – t-values de l'interaction entre les facteurs `temps` et `accent`. Modèle I : modèle sur la totalité des dyades et segments. Modèle II : modèle sur le sous-ensemble de données écartant les dyades et segments potentiellement problématiques.

Une tendance vers la convergence, indiquée par une valeur d'interaction négative, est observée pour le modèle II. Cette convergence s'observe pour les deux premiers jeux ($t = -1.565$), et pour l'ensemble de l'interaction ($t = -1.553$).

5.3.2 Convergence vers le groupe de l'interlocuteur

Nous présentons à présent les résultats de l'analyse de la convergence entre les réalisations des locuteurs et les réalisations du groupe dont fait partie son interlocuteur (voir p. 117). Le principe de l'analyse est identique à celle présentée dans la section précédente, si ce n'est que l'analyse discriminante prend comme classes de référence non les réalisations des deux locuteurs membres de la dyade dont on examine le rapprochement, mais les réalisations de l'ensemble des locuteurs des deux variétés.

Un modèle mixte a été ajusté pour les réalisations de chaque dyade, pour évaluer la relation existant entre les scores discriminants d'une part, et le temps normalisé et les locuteurs d'autre part. Un effet aléatoire est autorisé pour les facteurs **segment** et **mot**. L'équation du modèle est `lda` $\sim$ `tnorm * spkr + (tnorm | seg) + (tnorm | mot)`.

Le tableau 5.6 présente les t-values associées à l'interaction entre les facteurs **temps** et **locuteur**, qui indique la direction et la magnitude de l'évolution de la distance séparant les réalisations des locuteurs.

	jeu1	jeu2	jeu3	jeu1–jeu2	jeu2–jeu3	jeu1–jeu2–jeu3
d01	0.721	-3.994	2.088	-3.703	0.373	-1.418
d02	-1.186	-0.87	-0.573	-0.054	0.994	1.711
d03	-2.014	-0.846	-0.718	-1.162	0.441	0.214
d04	-0.696	-0.648	-0.365	1.536	-1.062	1.177
d05	-0.024	0.948	-1.264	-0.49	-0.2	-1.332
d06	-0.536	-0.165	1.404	-1.267	0.874	-0.549
d07	0.647	0.625	-0.162	1.107	-0.958	-0.136
d08	0.068	1.462	1.227	1.358	0.989	1.531
d09	-2.858	-3.961	0.118	-1.127	-2.484	-1.376
d10	1.793	-0.067	-0.981	1.307	-0.766	0.574
d11	0.752	-0.633	0.762	-1.506	0.58	-1.39
d12	0.528	0.431	-2.623	-0.305	-0.366	-0.244
moyenne	-0.234	-0.643	-0.091	-0.359	-0.132	-0.103
t-value	-0.62	-1.294	-0.24	-0.806	-0.435	-0.308
p-value	0.548	0.222	0.815	0.437	0.672	0.764

TABLE 5.6 – t-values de l'interaction entre les facteurs **temps** et **locuteur** du modèle mixte calculé pour chaque combinaison de dyade et de plage temporelle. Une valeur négative indique un rapprochement des réalisations, une valeur positive un éloignement. Les 3 dernières lignes présentent les résultats d'un t-test effectué pour chaque plage temporelle.

Des t-values négatives et supérieures à 2 en valeur absolue sont observées pour quatre dyades, pour diverses plages temporelles. Il s'agit des dyades d01 pour les plages jeu2, jeu1 – jeu2, d03 pour la plage jeu1, d09 pour les plages jeu1, jeu2 et jeu2–jeu3, et d12 pour la plage jeu3. Seule la dyade d01 pour la plage jeu3 présente une valeur positive et supérieure à 2. Cette tendance à la convergence se retrouve dans les valeurs moyennes calculées pour chaque plage temporelle pour toutes les dyades, mais n'atteint pas la significativité statistique comme les p-values associées au t-test effectué sur ces séries de valeurs.

Le tableau 5.7 montre les t-values de l'interaction entre les facteurs **temps** et **locuteur** calculées après avoir écarté les données limitant potentiellement l'observation d'un rapprochement, identifiées au chapitre précédent.

	jeu1	jeu2	jeu3	jeu1–jeu2	jeu2–jeu3	jeu1–jeu2–jeu3
d04	-0.88	-0.743	-0.378	1.762	-0.126	2.563
d05	0.884	1.405	-1.421	-1.283	0.026	-2.216
d07	0.537	-0.055	-0.539	-0.085	0.712	0.355
d08	0.423	1.244	0.32	1.286	-0.124	0.573
d09	-3.59	-1.32	-0.702	-3.04	-0.401	-2.091
d10	1.988	-1.779	-1.318	0.54	-1.371	-0.109
d11	-1.114	1.069	1.35	-3.026	2.541	-1.362
moyenne	-0.25	-0.026	-0.384	-0.549	0.18	-0.327
t-value	-0.366	-0.052	-1.054	-0.743	0.392	-0.507
p-value	0.727	0.96	0.333	0.486	0.708	0.63

TABLE 5.7 – t-values de l'interaction entre les facteurs **temps** et **locuteur** du modèle mixte calculé pour chaque combinaison de dyade et de plage temporelle, en écartant les données potentiellement problématiques. Une valeur négative indique un rapprochement des réalisations, une valeur positive un éloignement. Les 3 dernières lignes présentent les résultats d'un t-test effectué pour chaque plage temporelle.

Après avoir écarté les réalisations associées aux dyades et aux segments potentiellement problématiques, trois dyades présentent des valeurs d'interaction significativement négatives, indiquant une convergence des réalisations. Il s'agit des dyades d05 pour l'ensemble de l'interaction, d09 pour les plages jeu1, jeu1–jeu2, et la totalité de l'interaction, et d11 pour la plage jeu1–jeu2. Les t-values positives et supérieures à 2 sont observées pour la dyade d04 pour l'ensemble de l'interaction, et la dyade d11 pour la plage jeu2–jeu3. A nouveau une tendance globale vers une convergence se manifeste dans les

moyennes, mais n'est pas vérifiée par un t-test. Pour la dyade d11, on remarque à nouveau un patron de convergence au début de l'interaction, et une divergence vers la fin de celle-ci.

Nous présentons à présent les résultats de l'évolution des scores discriminants en regroupant les données par segment. L'équation du modèle est `lda ~ tnorm * accent + (tnorm | dyade) + (tnorm | mot)`. Ce modèle évalue les relations existant entre les scores discriminants d'une part et le temps normalisé et les locuteurs, regroupés par accent, d'autre part. Un effet aléatoire est autorisé par dyade et par mot. Les t-values associées à l'interaction entre les facteurs **temps** et **accent** sont présentées dans le tableau 5.8.

	jeu1	jeu2	jeu3	jeu1–jeu2	jeu2–jeu3	jeu1–jeu2–jeu3
C1	1.565	-1.625	0.657	1.099	-0.525	1.018
C2	-0.51	-0.702	-0.5	2.397	-2.078	1.256
C3	-0.209	0.05	-0.813	1.12	0.266	1.69
C4	0.4	0.686	0.924	1.884	-0.991	-0.043
M2	0.263	-0.858	-1.271	1.401	-1.569	0.539
M3	-0.636	-1.316	-0.134	-0.33	1.031	1.408
M4	0.439	-0.652	-0.641	0.665	0.144	1.216
B1	-0.354	0.966	0.026	-0.557	0.038	-0.744
B2	-2.063	-0.423	-1.453	-2.616	-1.328	-3.242
N1	-0.976	-0.68	-1.125	0.435	-1.069	-0.124
N2	-1.957	-0.749	-1.45	-3.761	0.265	-2.495
N3	-0.256	-1.456	-1.163	-0.998	-0.08	-0.01
N4	-1.177	-0.76	-0.415	-1.172	0.763	0.407
moyenne	-0.421	-0.578	-0.566	-0.033	-0.395	0.067
t-value	-1.524	-2.734	-2.657	-0.067	-1.506	0.163
p-value	0.153	0.018	0.021	0.947	0.158	0.873

TABLE 5.8 – t-values de l'interaction entre les facteurs **temps** et **accent** du modèle mixte calculé pour chaque combinaison de segment et de plage temporelle. Une valeur négative indique un rapprochement des réalisations, une valeur positive un éloignement. Les 3 dernières lignes présentent les résultats d'un t-test effectué pour chaque plage temporelle.

Des t-values indiquant une convergence sont observées pour trois segments, sur diverses plages temporelles. Il s'agit des segments C2 pour la plage jeu2–jeu3, B2 pour les plages jeu1, jeu1–jeu2 et la totalité de l'interaction, et N2 pour la plage jeu1–jeu2 et la totalité de l'interaction. Une t-value positive et supérieure à 2 est observée pour le segment C2 pour la plage jeu1–jeu2.

La moyenne des t-values pour les segments est globalement négative sur les différentes plages temporelles. Cette moyenne est significativement différente de 0 pour les phases jeu2 ($p = 0.018$) et jeu3 ($p = 0.021$) ce qui indique un rapprochement consistant sur l'ensemble des segments pour ces phases.

Le tableau 5.9 montre les t-values de l'interaction entre les facteurs **temps** et **accent** calculées par le modèle après avoir écarté les données potentiellement problématiques, identifiées au chapitre précédent.

	jeu1	jeu2	jeu3	jeu1–jeu2	jeu2–jeu3	jeu1–jeu2–jeu3
C1	1.212	0.124	1	1.006	0.245	1.024
C3	-2.433	2.057	-1.276	1.337	0.718	1.578
M2	0.846	-1.157	-1.161	0.777	-1.184	0.129
B1	-0.2	1.935	-0.239	-1.225	2.367	0.103
B2	-1.112	2.082	-1.876	-2.562	0.118	-3.436
N1	-1.603	-0.761	-0.878	-0.391	-1.064	-1.036
N2	0.188	0.096	-1.679	-2.327	0.99	-1.133
N3	1.402	-1.938	-0.074	0.721	-1.476	0.316
moyenne	-0.212	0.305	-0.773	-0.333	0.089	-0.307
t-value	-0.431	0.549	-2.292	-0.61	0.194	-0.556
p-value	0.679	0.6	0.056	0.561	0.851	0.596

TABLE 5.9 – t-values de l'interaction entre les facteurs **temps** et **accent** du modèle mixte calculé pour chaque combinaison de segment et de plage temporelle, en écartant les données potentiellement problématiques. Une valeur négative indique un rapprochement des réalisations, une valeur positive un éloignement. Les 3 dernières lignes présentent les résultats d'un t-test effectué pour chaque plage temporelle.

Des t-values indiquant une convergence sont observées pour trois segments, sur diverses plages temporelles. Il s'agit des segments C3 pour la plage jeu1, B2 pour les plages jeu1–jeu2 et la totalité de l'interaction, et N2 pour la plage jeu1–jeu2. Une t-value positive et supérieure à 2 est observée pour les segments C3 pour la plage jeu2, B1 pour la plage jeu2–jeu3, et B2 pour la plage jeu2.

Une tendance globale à la convergence sur l'ensemble des segments est observée pour la phase jeu3, pour laquelle la p-value associée au t-test vaut 0.056.

Finalement, le tableau 5.10 montre les résultats d'un modèle ajusté sur l'ensemble des données, en regroupant les dyades et les segments. Ce modèle global combine les résultats obtenus par les modèles précédents. Il autorise

une pente aléatoire par segment et par locuteur, et a pour équation `lda` ∼ `tnorm * accent + (tnorm | seg) + (tnorm | spkr) + (tnorm | mot)`. Le tableau présente la t-value de l'interaction entre le facteur **temps** et **accent** qui est une indication de l'évolution comparée des scores des locuteurs au cours du temps.

	jeu1	jeu2	jeu3	jeu1–jeu2	jeu2–jeu3	jeu1–jeu2–jeu3
Modèle I	-0.374	-0.702	-0.728	-0.7	-0.182	-0.411
Modèle II	0.246	-0.315	-1.754	-1.224	-0.025	-1.369

TABLE 5.10 – t-values de l'interaction entre les facteurs `temps` et `accent`. Modèle I : modèle sur la totalité des dyades et segments. Modèle II : modèle sur le sous-ensemble de données écartant les dyades et segments potentiellement problématiques.

Une tendance vers la convergence, indiquée par une valeur négative, est observée pour le modèle II, pour les phases jeu3 ($t = -1.754$), jeu1–jeu2 ($t = -1.224$) et pour la totalité de l'interaction ($t = -1.369$).

5.3.3 Convergence vers l'accent de l'interlocuteur

Nous présentons à présent les résultats de l'analyse de la convergence entre les productions des locuteurs de la dyade, évaluée comme un rapprochement entre les productions de chaque locuteur vers les productions prototypiques de la variété parlée par son interlocuteur. L'analyse discriminante prend comme classes de référence, pour chaque segment critique, les productions de tous les locuteurs du corpus ayant été catégorisées par l'alignement automatique comme une réalisation représentative de la variété standard ou méridionale (voir p. 118).

Un modèle mixte a été ajusté pour les réalisations de chaque dyade, pour évaluer la relation existant entre les scores discriminants d'une part, et le temps normalisé et les locuteurs d'autre part. Un effet aléatoire est autorisé pour les facteurs **segment** et **mot**. L'équation du modèle est `lda ~ tnorm * spkr + (tnorm | seg) + (tnorm | mot)`.

Le tableau 5.11 présente les t-values associées à l'interaction entre les facteurs **temps** et **locuteur**, qui indique la direction et la magnitude de l'évolution de la distance séparant les réalisations des locuteurs.

Des t-values négatives et supérieures à 2 en valeur absolue sont observées pour trois dyades, pour diverses plages temporelles. Il s'agit des dyades d01 pour les plages jeu2, jeu1–jeu2, d03 pour la plage jeu3, et d09 pour la totalité de l'interaction. Deux dyades présentent des valeurs positives et supérieures à 2. Il s'agit de la dyade d10 pour la plage jeu1, et de la dyade d11 pour la plage jeu2–jeu3. Bien que sur l'ensemble des dyades la moyenne des t-values soit légèrement négative sauf pour le premier jeu, cette tendance ne se vérifie pas dans un t-test de ces séries.

Le tableau 5.12 montre les t-values de l'interaction entre les facteurs **temps** et **locuteur** calculées après avoir écarté les données limitant potentiellement l'observation d'un rapprochement, identifiées au chapitre précédent.

Aucune t-value indiquant un rapprochement significatif des réalisations n'est observée pour cette analyse. En revanche, trois dyades présentent des valeurs positives et supérieures à 2. Il s'agit des dyades d04 pour la totalité de l'interaction, d10 pour la plage jeu1, et d11 pour la plage jeu2–jeu3.

Pour cet ensemble de données, on observe sur la totalité de l'interaction un éloignement progressif des réalisations qui approche la significativité ($p = 0.068$).

Les résultats qui suivent concernent l'évolution des scores discriminants en effectuant un regroupement par segment. L'équation du modèle est `lda ~ tnorm * accent + (tnorm | dyade) + (tnorm | mot)`. Ce modèle évalue les relations existant entre les scores discriminants d'une part et le temps nor-

	jeu1	jeu2	jeu3	jeu1–jeu2	jeu2–jeu3	jeu1–jeu2–jeu3
d01	0.205	-2.698	1.145	-2.224	-0.826	-1.962
d02	-0.032	-0.674	-0.149	-0.676	0.617	0.338
d03	0.248	-0.465	-2.475	-0.985	0.166	-0.011
d04	0.037	0.726	1.394	1.173	-0.834	0.009
d05	0.929	0.876	-0.119	0.633	0.413	0.324
d06	1.149	-0.573	0.952	-0.362	-0.301	-0.841
d07	0.987	-1.288	-1.763	0.462	-0.129	0.986
d08	-0.105	0.651	-0.532	1.838	-1.849	0.045
d09	-1.961	-1.232	1.674	-1.694	-1.579	-3.036
d10	3.081	0.023	-0.32	1.413	0.687	1.574
d11	-0.696	1.345	1.353	-1.103	2.209	0.341
d12	1.624	0.521	-1.639	1.013	-0.315	0.838
moyenne	0.455	-0.232	-0.04	-0.043	-0.145	-0.116
t-value	1.26	-0.698	-0.1	-0.112	-0.459	-0.314
p-value	0.234	0.5	0.922	0.913	0.655	0.759

TABLE 5.11 – t-values de l'interaction entre les facteurs **temps** et **locuteur** du modèle mixte calculé pour chaque combinaison de dyade et de plage temporelle. Une valeur négative indique un rapprochement des réalisations, une valeur positive un éloignement. Les 3 dernières lignes présentent les résultats d'un t-test effectué pour chaque plage temporelle.

	jeu1	jeu2	jeu3	jeu1–jeu2	jeu2–jeu3	jeu1–jeu2–jeu3
d04	-1.586	-0.53	0.918	1.316	0.598	2.295
d05	1.19	1.072	-0.886	-0.013	0.96	0.12
d07	-0.163	0.198	0.237	-0.303	1.74	1.344
d08	0.082	0.118	-1.237	1.134	-0.604	1.133
d09	-1.775	-1.766	0.596	-1.715	-0.366	-0.865
d10	2.436	-1.117	-0.255	1.075	0.414	1.682
d11	0.277	1.081	1.252	-1.485	2.879	0.499
moyenne	0.066	-0.135	0.089	0.001	0.803	0.887
t-value	0.118	-0.333	0.255	0.002	1.759	2.222
p-value	0.91	0.75	0.807	0.998	0.129	0.068

TABLE 5.12 – *t*-values de l'interaction entre les facteurs **temps** et **locuteur** du modèle mixte calculé pour chaque combinaison de dyade et de plage temporelle, en écartant les données potentiellement problématiques. Une valeur négative indique un rapprochement des réalisations, une valeur positive un éloignement. Les 3 dernières lignes présentent les résultats d'un *t*-test effectué pour chaque plage temporelle.

malisé et les locuteurs, regroupés par accent, d'autre part. Un effet aléatoire est autorisé par dyade et par mot. Les *t*-values associées à l'interaction entre les facteurs **temps** et **accent** sont présentées dans le tableau 5.13.

Des *t*-values indiquant une convergence sont observées pour trois segments, sur diverses plages temporelles. Il s'agit des segments C3 pour la plage jeu2, B2 pour les plages jeu1–jeu2 et la totalité de l'interaction, et N4 pour la plage jeu2–jeu3. Une *t*-value positive et supérieure à 2 est observée pour les segments B1 pour la plage jeu2, B2 pour la plage jeu2–jeu3, et N4 pour la plage jeu1–jeu2.

Aucune tendance claire ne se dégage sur l'ensemble des dyades pour chaque plage temporelle, comme le montrent les résultats du *t*-test effectué sur ces séries de valeurs.

Le tableau 5.14 montre les *t*-values de l'interaction entre les facteurs **temps** et **accent** calculées par le modèle après avoir écarté les données limitant potentiellement problématiques, identifiées au chapitre précédent.

Des *t*-values indiquant une convergence sont observées pour trois segments, sur diverses plages temporelles. Il s'agit des segments C3 pour la plage jeu2, B2 pour les plages jeu1–jeu2 et la totalité de l'interaction, et N2 pour la plage jeu1–jeu2. Une *t*-value positive et supérieure à 2 est observée pour les segments C1 pour la totalité de l'interaction, B1 pour la plage jeu2–jeu3, et

	jeu1	jeu2	jeu3	jeu1–jeu2	jeu2–jeu3	jeu1–jeu2–jeu3
C1	1.714	-1.089	0.183	0.844	0.455	1.747
C2	1.417	1.243	0.766	-0.656	0.901	-0.763
C3	-0.315	-2.879	-0.91	0.229	-0.955	1.027
C4	0.937	0.853	0.941	0.323	-0.002	-0.303
M2	0.231	-0.778	-1.215	1.907	-1.732	0.937
M3	1.514	0.291	0.934	1.149	-0.878	-0.772
M4	-0.089	-0.676	-0.609	-0.12	-0.064	0.264
B1	-0.849	2.1	0.239	-0.629	1.951	0.282
B2	-0.86	0.678	-0.019	-4.929	3.035	-2.36
N1	0.531	-1.118	0.379	1.076	-1.396	0.33
N2	1.432	-0.945	-1.133	-1.825	1.125	-0.363
N3	-0.266	-1.116	-1.319	-0.542	0.118	0.396
N4	0.39	0.259	0.302	3.95	-2.69	1.639
moyenne	0.445	-0.244	-0.112	0.06	-0.01	0.159
t-value	1.784	-0.675	-0.49	0.104	-0.024	0.516
p-value	0.1	0.512	0.633	0.919	0.982	0.615

TABLE 5.13 – t-values de l'interaction entre les facteurs **temps** et **accent** du modèle mixte calculé pour chaque combinaison de segment et de plage temporelle. Une valeur négative indique un rapprochement des réalisations, une valeur positive un éloignement. Les 3 dernières lignes présentent les résultats d'un t-test effectué pour chaque plage temporelle.

	jeu1	jeu2	jeu3	jeu1–jeu2	jeu2–jeu3	jeu1–jeu2–jeu3
C1	1.399	0.63	0.252	0.897	1.175	2.145
C3	-1.678	-3.32	0.095	-1.738	0.395	-0.054
M2	0.593	-0.819	-1.166	1.475	-0.874	1.31
B1	-0.106	1.771	-0.947	0.113	2.336	1.768
B2	0.382	1.266	-0.218	-3.182	2.609	-2.124
N1	-1.1	-0.57	1.032	-0.392	0.121	-0.248
N2	0.835	0.128	-1.261	-2.012	1.306	-0.668
N3	0.87	-0.911	-0.745	1.103	-0.511	1.053
moyenne	0.149	-0.228	-0.37	-0.467	0.82	0.398
t-value	0.401	-0.407	-1.305	-0.786	1.835	0.787
p-value	0.7	0.696	0.233	0.458	0.109	0.457

TABLE 5.14 – t-values de l'interaction entre les facteurs **temps** et **accent** du modèle mixte calculé pour chaque combinaison de segment et de plage temporelle, en écartant les données potentiellement problématiques. Une valeur négative indique un rapprochement des réalisations, une valeur positive un éloignement. Les 3 dernières lignes présentent les résultats d'un t-test effectué pour chaque plage temporelle.

B2 pour la plage jeu2–jeu3.

A nouveau, aucune tendance claire ne se dégage sur l'ensemble des dyades pour chaque plage temporelle, comme le montrent les résultats du t-test effectué sur ces séries de valeurs.

Le tableau 5.15 montre les résultats d'un modèle ajusté sur l'ensemble des données, en regroupant les dyades et les segments. Ce modèle global combine les résultats obtenus par les modèles précédents. Il autorise une pente aléatoire par segment et par locuteur, et a pour équation `lda` $\sim$ `tnorm * accent + (tnorm | seg) + (tnorm | spkr) + (tnorm | mot)`. Le tableau présente la t-value de l'interaction entre le facteur **temps** et **accent** qui est une indication de l'évolution comparée des scores des locuteurs au cours du temps.

Le modèle I, qui est calculé sur l'ensemble des données, met en évidence une divergence des réalisations des locuteurs pour le premier jeu ($t = 2.655$).

	jeu1	jeu2	jeu3	jeu1–jeu2	jeu2–jeu3	jeu1–jeu2–jeu3
Modèle I	2.655	-1.031	-0.001	0.103	-0.743	-0.252
Modèle II	1.13	-0.528	-0.008	-0.276	1.067	0.893

TABLE 5.15 – t-values de l'interaction entre les facteurs `temps` et `accent`. Modèle I : modèle sur la totalité des dyades et segments. Modèle II : modèle sur le sous-ensemble de données écartant les dyades et segments potentiellement problématiques.

5.4 Discussion

Les résultats sur les trois types de convergence présentent à la fois des régularités d'une analyse à l'autre mais aussi des différences, qu'il est intéressant d'examiner plus en détail.

Les analyses suivant les trois types de convergence ont mis en évidence un rapprochement significatif pour la dyade d01 au cours des deux premiers jeux et au cours du jeu 2. De la même façon, un rapprochement est observé pour la dyade d09 pour les trois types d'analyses : au cours des deux premiers jeux pour l'analyse de la convergence vers le locuteur (voir tableau 5.2), au cours de trois plages temporelles pour l'analyse de la convergence vers le groupe dont fait partie l'interlocuteur (voir par ex. tableau 5.6) ou sur la totalité de l'interaction pour l'analyse de la convergence vers l'accent de l'interlocuteur (voir tableau 5.11). Dans aucune de ces trois analyses cette dyade n'a présenté une t-value indiquant une divergence. Pour les dyades d06, d07, en revanche, les t-values n'excèdent dans aucune analyse 2 en valeur absolue, ce qui signifie qu'aucune tendance ne se manifeste dans l'évolution comparée des réalisations, que ce soit un rapprochement ou un éloignement. Pour d'autres dyades (d04 ou d08 par exemple), seuls des éloignements significatifs des réalisations sont observés. Un autre cas intéressant est la dyade d11, pour laquelle sur les trois types d'analyses on observe un rapprochement des réalisations au cours des deux premiers jeux et/ou un éloignement au cours des deux derniers jeux.

On retrouve des régularités similaires d'une analyse à l'autre lorsque l'on considère les regroupements par segments. Par exemple, au segment N2 n'est associé que des t-values négatives lorsqu'elle est supérieure à 2 en valeur absolue sur les trois types d'analyses, et le segment B2 donne lieu en général sur l'ensemble des dyades à une convergence au début de l'interaction et à une divergence vers la fin de l'interaction. D'autres segments, comme M2, N1 ou N3 ne donnent pas lieu à l'observation de changements sur aucune des analyses.

Ces régularités établies pour certaines dyades et segments permettent d'une part de valider les effets obtenus individuellement pour chaque analyse, mais d'autre part de tirer des conclusions concernant ces dyades et ces segments. On peut ainsi affirmer que pour les dyades d01 et d09, la convergence « vocale » établie par la première analyse correspond à une convergence sur la dimension de l'accent, que celle-ci soit décrite directement par les réalisations des locuteurs de ces variétés, ou par les alignements en catégories phonétiques représentatives de ces variétés. Cette cohérence des résultats pour ces dyades permet également de valider indirectement la pertinence des segments choisis pour la description des variétés considérées. De la même façon, la stabilité

des effets de convergence observés pour les segments B2 et N2 sur l'ensemble des dyades permet de conclure à une diminution de la distance acoustique existant entre les locuteurs avant qu'ils ne s'engagent dans une interaction, et que cette distance est bien caractéristique de la différence phonétique entre la variété méridionale et standard du français parlé, qu'elle soit décrite par les réalisations des locuteurs de ces variétés, ou par les alignements en catégories phonétiques représentatives de ces variétés. Cette stabilité permet également de valider indirectement l'appartenance des locuteurs à l'une ou l'autre des variétés telle qu'elle a été évaluée par le questionnaire.

Ainsi, lorsque l'on examine individuellement les t-values associées à l'interaction entre les facteurs **temps** et **locuteur** indiquant l'évolution de la distance entre les réalisations acoustiques des locuteurs au cours du temps, on remarque une consistance globale des patrons sur les trois analyses. Les modèles évaluant une relation globale sur l'ensemble des dyades et des segments révèlent quant à eux une différence entre l'analyse de la convergence vers l'interlocuteur et le groupe associé à l'interlocuteur d'une part, et l'analyse de la convergence vers l'accent associé à l'interlocuteur d'autre part. En effet, si les deux premières analyses montrent globalement une tendance vers une convergence des réalisations (sur la totalité de l'interaction et sur les deux premiers jeux pour les deux analyses ; mais aussi sur le jeu 3 pour l'analyse de la convergence vers le groupe associé à l'interlocuteur, voir tableaux 5.5 et 5.10), l'analyse construite à partir des classes de référence composée par les réalisations représentatives des deux variétés parlées met en évidence une divergence des réalisations pour le jeu 1 (voir tableau 5.15). Remarquons d'abord que les résultats de la convergence établis pour les deux premières analyses ont été obtenus par les modèles qui considèrent uniquement un sous-ensemble des productions des locuteurs. Ce sous-ensemble a été obtenu en laissant de côté les dyades et les segments critiques pouvant potentiellement limiter une convergence de se produire, dont la liste a été établie indépendamment au chapitre précédent d'après les résultats de la classification automatique des locuteurs dans l'une ou l'autre des deux variétés parlées. L'observation d'une convergence pour ce sous-ensemble et non pour la totalité des données valide ainsi cette sélection (même si la dyade écartée d01 présente une convergence dans les trois analyses individuelles regroupées par dyade, par exemple). Elle pointe également vers un comportement global, pour des locuteurs de variété du français différentes engagés dans une interaction conversationnelle, à une convergence des réalisations phonétiques de certains segments qui illustrent les différences entre ces deux variétés. Ce comportement général est consistant avec les résultats obtenus par Delvaux et Soquet (2007) dans un cadre expérimental non interactif, dans lequel les locuteurs manifestaient sur un ensemble de segments phonétiques, décrivant

des différences de régiolectes ou non, un rapprochement systématique vers des cibles auditives auxquelles ils étaient exposés. D'après nos résultats, seul un sous ensemble des segments supportent une convergence. Il est possible que la salience perceptive individuelle des segments en tant que marqueurs dialectaux soit un facteur déterminant pour permettre à une convergence de se produire. Labov (2001, p. 78) propose trois degrés de salience perceptive (indicateurs, marqueurs, stéréotypes), mais il n'est pas clairement établi quels sont les causes qui sont à l'origine de l'existence de cette différenciation. Kerswill et Williams (2002) ont avancé que le caractère discret ou continu des caractéristiques phonétiques de la variable phonologique pourrait être un facteur, ainsi que sa fréquence relative ou sa proéminence prosodique. De plus, il y a une grande variabilité dans l'évaluation perceptive de l'information indexicale par les auditeurs qui dépend de l'individu, des dialectes et des situations Foulkes *et al.* (2010).

La mise en évidence d'une divergence dans l'analyse qui prend comme classe de référence les réalisations représentatives des deux variétés est plus difficilement interprétable. On peut cependant observer que ces résultats ont été obtenus sur la totalité des données du corpus, y compris les réalisations correspondant aux dyades et aux segments qui peuvent potentiellement limiter une convergence de se produire. Pour les deux autres types d'analyse, cet ensemble total de données ne donnait cependant pas lieu à l'observation d'une divergence. Une raison pour laquelle les résultats pourraient être plus incertains pour cette analyse est que cette dernière utilise une information supplémentaire, externe à la structure des données, qui est l'association des réalisations des segments critiques avec des catégories phonétiques, obtenue par la procédure d'alignement forcé en variantes de prononciation. Cette information est donc dans une certaine mesure indépendante de l'information acoustique caractérisant les segments dans l'analyse discriminante : cette dernière est obtenue en calculant les coefficients DCT en un point temporel des segments, tandis que la première dépend des modèles acoustiques utilisés par la procédure d'alignement. Il se pourrait donc que l'ajout de cette information externe introduise du bruit dans l'initialisation des classes de référence pour l'analyse discriminante. Mais dans ce cas d'indépendance parfaite entre l'information acoustique utilisée par l'analyse discriminante et l'information catégorielle phonétique dérivée des modèles acoustiques, on devrait trouver une absence d'effet dans l'évolution des réalisations dans l'espace de projection. Or, cette analyse met bien en évidence un effet mais qui est dans la direction opposée aux deux autres analyses de la convergence. Ce résultat est d'autant plus surprenant que lorsque des effets de divergence étaient observés, ils concernaient le plus souvent la deuxième partie de l'interaction, (jeu2 ou jeu2–jeu3), et moins souvent le premier jeu.

Enfin, un résultat important mis en évidence par les différentes analyses est la variabilité inter-dyades et inter-segments des patrons de convergence. En effet, même si les deux premières analyses ont conclu à une convergence globale sur l'ensemble des dyades et des segments (sur le sous ensemble des données non problématiques), des différences systématiques existent entre les différentes dyades et les différents segments, et ceci de façon cohérente sur les trois types d'analyses. Pour certaines dyades, des patrons de convergence ont été systématiquement observés (par exemple d01 ou d09), pour d'autres (d04 ou d08 par exemple), ce sont principalement des patrons de divergence qui ont été observés, et pour la dyade d11, une convergence a été observée au début de l'interaction et une divergence sur les deux derniers jeux. Les segments manifestent des régularités similaires. Une telle variabilité inter-individuelle permet de nuancer le caractère automatique prêté aux phénomènes d'adaptation relevés dans la littérature (par ex. Delvaux et Soquet, 2007). Dans le cadre des modèles à exemplaires, ces patrons s'expliquent par un mécanisme par défaut qui est la convergence, c'est-à-dire la modification progressive des représentations mentales associées aux variables phonologiques produites en interaction, sous l'effet de l'exposition répétée à celles-ci. Il faut rajouter des facteurs de plus haut niveau, les *poids* dans le formalisme exemplariste, qui modulent ce comportement par défaut. Ces facteurs peuvent par exemple décrire l'implication attentionnelle des locuteurs dans la tâche, ou être liée à des facteurs sociaux, comme l'identification des locuteurs à un groupe social. Nous avions introduit dans le design de notre expérience un facteur de compétence sociale, la désirabilité sociale, pour évaluer une influence éventuelle sur les patrons de convergence (voir section 3.5). Plus précisément, les locuteurs étaient appariés selon leur position sur l'échelle de désirabilité sociale. Ceci permettait d'une part d'assurer un équilibre dans les échanges, mais d'autre part de tester l'hypothèse d'une convergence plus marquée chez les dyades dont les membres présentaient un score de désirabilité sociale élevé que chez les dyades dont le score était situé vers le bas de l'échelle. Cependant, une examination des schémas de convergence individuels n'a pas révélé une telle répartition des patrons de convergence. En revanche, des tendances ont été observées dans la symétrie des rapprochements entre les locuteurs, qui semblait être liée à la *différence de score* entre les membres de la dyade, les locuteurs de score plus élevé manifestant plus de convergence que leur partenaire. L'appariement des locuteurs ayant précisément tenté de minimiser cette différence de score de désirabilité sociale, ce facteur a par la suite été écarté des analyses. De nouveaux enregistrements employant un appariement différent sur ce facteur seraient nécessaire pour valider ces observations préliminaires.

Chapitre 6

Conclusion

Nous nous sommes intéressés dans ce travail à la caractérisation des représentations mentales des sons de la parole, à savoir leur forme et leur évolution. Nous avons choisi la variation phonologique régionale comme niveau privilégié d'analyse étant donné la pertinence qu'il revêt pour les locuteurs et le degré de granularité d'analyse qu'il permet. Le choix de l'étude de la parole en interaction s'est imposé à nous, comme cadre nécessaire à un accès non biaisé aux représentations mentales.

Nous avons développé une tâche interactive pour aborder cette problématique dans laquelle deux participants, locuteurs des variétés du français standard et méridional respectivement, étaient amenés à produire de façon répétée des mot-cibles qui illustraient les principales différences phonétiques et phonologiques entre les deux variétés. A partir d'outils de reconnaissance automatique de la parole, nous avons ensuite développé une méthode destinée à identifier les productions des variables phonologiques sous étude et de localiser l'information acoustique pertinente pour caractériser ces réalisations.

Une procédure bayesienne de classification a été utilisée pour caractériser automatiquement la variété parlée des locuteurs sur la base de l'alignement des réalisations des variables phonologiques en variantes de prononciation. Cette procédure a permis de mettre en évidence l'importance relative des variables phonologiques dans la distinction des variétés, en montrant par exemple le statut particulier de la variable phonologique /di/ dans la différenciation des deux variétés, un trait peu remarqué dans la littérature sociophonétique. Un maintien de l'opposition /ɛ̃/–/œ̃/ a également été observé dans les deux variétés, ainsi qu'un changement en chaîne des voyelles nasales. Elle a aussi mis en évidence une relative stabilité temporelle des formes sonores associées aux variables phonologiques tout au long de l'interaction, évaluée au niveau catégoriel des variantes de prononciation.

Enfin, une analyse des réalisations à un niveau acoustique plus détaillé, sub-phonémique, a permis de dégager des tendances de modification progressive de la forme prononcée des variables phonologiques au cours de l'interaction. Cette modification a été évaluée en projetant les réalisations des locuteurs sur un axe illustrant soit les différences acoustiques entre les locuteurs engagés dans l'interaction, soit entre les deux variétés. Un rapprochement a été observé lorsque les axes étaient construits en utilisant l'information d'appartenance à la variété parlée des locuteurs, mais pas lorsqu'ils étaient construits à partir de la caractérisation en variantes de prononciation représentatives des deux variétés, telle qu'évaluée par l'apprentissage de modèles acoustiques.

L'utilisation hiérarchique de dimensions phonologiques pour la caractérisation des variétés régionales, en permettant une classification correcte de 79% des 24 locuteurs pour un ensemble réduit de 19 variables phonologiques, pourrait être étendue à un nombre plus important de locuteurs et de dimensions phonologiques, et ainsi fournir une caractérisation segmentale de la variété parlée à d'autres échelles que la variété régionale. Par exemple, il semblerait possible de caractériser avec précision les caractéristiques phonétiques associées à des groupes de locuteurs plus spécifiques, comme le groupe de pairs, le cercle familial, etc. Il serait également intéressant de vérifier s'il existe des corrélats perceptifs aux différents degrés de pouvoir discriminant des variables phonologiques mis en évidence. En effet, les indices potentiellement utilisés par les auditeurs dans l'établissement de différents degrés de salience perceptive ne sont à ce jour pas clairement établis, et en particulier on ne sait pas s'ils sont indépendants de la langue ou de la communauté de locuteurs (voir par ex. Docherty, 2007; Foulkes *et al.*, 2010).

Les résultats de convergence obtenus apportent deux éléments importants. Tout d'abord, le fait que ces résultats caractérisent l'ensemble des locuteurs et des variables phonologiques pointe vers une modification systématique et automatique des représentations mentales associées aux sons de la parole en interaction conversationnelle.

L'empan temporel considéré dans notre étude, d'une vingtaine de minutes, apparaît relativement court en comparaison avec la quantité de parole à laquelle sont quotidiennement exposés les locuteurs. Cette durée pourrait constituer une raison pour laquelle les modifications graduelles des réalisations des variables phonologiques n'atteignent pas une magnitude suffisante pour provoquer un changement dans l'attribution des variantes de prononciation par les procédures d'alignement forcé. Cependant, des modifications graduelles de la forme sonore des sons de la parole ont été observés à presque toutes les échelles temporelles, dans des études utilisant divers paradigmes expérimentaux. Par exemple, Goldinger (1998) a montré dans une tâche de

shadowing une différence de similarité dans la forme prononcée des mots par rapport à la cible, en fonction de la fréquence lexicale, du nombre de répétitions au cours de la tâche ou du délai de présentation de la cible. Delvaux et Soquet (2007) ont montré, au cours d'une tâche dans laquelle des locuteurs prononcaient des mot-cibles dans des phrases porteuses en alternance avec un locuteur pré-enregistré, que les caractéristiques spectrales des réalisations de segments appartenant aux mot-cibles se rapprochaient automatiquement des caractéristiques spectrales des mêmes segments produits par le locuteur pré-enregistré. L'empan temporel de l'exposition était similaire à notre étude (une trentaine de minutes pour la phase de test). A une plus grande échelle temporelle, des modifications graduelles des caractéristiques acoustiques associées à des catégories phonétiques ont également été mises en évidence. Par exemple, Sancier et Fowler (1997) ont trouvé que la durée de "Voice Onset Time" (VOT) des plosives non voisées d'une locutrice se modifiait après des séjours de plusieurs mois au Brésil et aux Etats-Unis. La durée du VOT était plus longue après exposition à l'anglais américain et plus courte après l'exposition au portugais brésilien. La direction du changement correspond à la différence de durée du VOT qui existe entre ces deux langues, les plosives non voisées ayant une durée plus longue en anglais américain qu'en portugais brésilien. Harrington (2006) a quant à lui découvert une évolution de la qualité de voyelle [ɪ ː] dans la parole de la Reine d'Angleterre dans ses discours annuels adressés à la nation, sur une période s'étendant sur quarante ans. L'évolution de la qualité de voyelle, caractérisée par un rapprochement des valeurs de F_1 et F_2 vers celles caractéristiques du [i ː] et une curvature plus importante de la trajectoire de F_2, était en accord avec un changement phonétique ayant pris place pour la variété RP (Received Pronunciation) de l'anglais sur la même période.

Ces exemples de modification graduelle des formes sonores des sons de la parole à un niveau détaillé illustrent donc une possibilité d'évolution des représentations mentales à des échelles temporelles très différentes. Ils montrent en particulier une plasticité des représentations mentales, et pointent vers un possible mécanisme automatique de modification des représentations mentales en direction des cibles auxquelles le locuteur est exposé au fil de ses interactions.

L'existence de ce mécanisme appelle lui-même une autre question, celles des causes qui peuvent être invoquées pour que celui-ci se mette en place, ou qu'il recoive différents niveaux de magnitude suivant les individus. Une hypothèse est que ce mécanisme par défaut est *inhibé* par des facteurs de plus haut niveau, et limite ainsi son application. C'est par exemple l'hypothèse avancée par Dijksterhuis et Bargh (2001), qui mettent en avant un lien automatique entre perception et action, dû au fait que les représentations mentales asso-

ciées à la perception et au comportement sont partagées. Une conséquence de ces représentations partagées est l'imitation, qui serait un comportement par défaut, mais en pratique inhibé par une série de facteurs individuels, comme l'expérience négative associées à des imitations passées, le conflit avec les buts poursuivis ou encore des traits de caractère individuels comme l'attention portée sur soi. Nous avions apparié les 2 membres de chaque dyade dans notre étude suivant la similarité de leur score sur une mesure de compétence sociale individuelle, la désirabilité sociale (Crowne et Marlowe, 1960), qui s'était montrée par exemple être corrélée avec la magnitude de la convergence d'intensité vocale vers l'interlocuteur (Natale, 1975). Cependant, les patrons de convergence que nous avons observés sur les réalisations acoustiques ont suggéré que ce n'est pas la valeur absolue de cette mesure qui détermine si les réalisations de deux locuteurs engagés dans une interaction conversationnelle se rapprochent au fil du temps, mais que ce serait plutôt la *différence* de désirabilité sociale entre les deux membres de la dyade qui serait un prédicteur plus robuste d'un rapprochement des réalisations. Ce point mérite bien évidemment d'être vérifié par un design expérimental qui apparie spécifiquement les membres de chaque dyade en maximisant la différence de cette mesure avec leur partenaire. Plus généralement, l'étude systématique de facteurs d'ordre psychologique et/ou social semble prometteuse pour expliciter les patrons de variation inter-individuelle associés avec les comportements imitatifs (voir par ex. Babel, 2009, pour une étude récente concernant la dimension phonétique).

Finalement, une des contributions de ce travail pourrait concerner les mutations actuelles que l'on voit s'opérer dans l'utilisation de la parole. En effet, celle-ci se trouve de plus en plus souvent dématérialisée, désolidarisée de ses conditions d'occurrence premières, l'interaction sociale. On est dans notre vie quotidienne de plus en plus confrontés à des systèmes vocaux interactifs, dont les performances s'améliorent continuellement. Cependant, ils sont bien souvent jugés comme peu utiles et difficiles à utiliser par les utilisateurs, et cette performance ne dépend plus des capacités de reconnaissance ou de synthèse de la parole (Gravano et Hirschberg, 2010). Une des raisons à cela apparaît tenir au fait que la dimension interactionnelle n'est que peu intégrée dans ces systèmes, qui sont essentiellement unidirectionnels. Or, bien plus que d'être un moyen de véhiculer une simple séquence de symboles constituant un message linguistique univoque, la parole est en premier lieu utilisée par les locuteurs comme un outil pour « naviguer dans leurs interactions sociales quotidiennes » (Ford et Couper-Kuhlen, 2004). Il semblerait ainsi que de tels systèmes bénéficieraient de l'intégration des caractéristiques qui décrivent les spécificités de la parole en interaction, y compris les phénomènes d'ajustement vers l'interlocuteur qu'elle peut présenter. Par exemple, on pourrait

imaginer qu'un système qui intègre les backchannels de l'utilisateur pour modifier sa propre réponse, ou ajuste certains paramètres de synthèse vocale pour correspondre à la parole son interlocuteur (comme le débit, la hauteur de voix moyenne ou l'accent), facilite les échanges avec celui-ci. En plus de l'amélioration significative de ces systèmes, ces avancées permettraient de mieux comprendre les mécanismes qui sous-tendent les échanges communicationnels entre humains. C'est dans ces directions que nous voudrions orienter nos recherches futures.

Appendice

Questionnaire 1

Nom : ..... Prénom : .....

Age : ..... Sexe : M ☐ F ☐ Classe : .....

Lieu de naissance : .....

Lieu de résidence actuel (Commune, ou Arrondissement pour Marseille) : .....

Depuis quand vivez-vous en Région PACA ? (Mois, Année) : .....

Avez-vous vécu dans d'autres régions ou à l'étranger pendant plus de 6 mois ? Si oui, précisez (Lieu et Période):

.....

.....

.....

Langue maternelle : .....

Parlez-vous d'autres langues ? Si oui, lesquelles ? .....

Des personnes vous ont-elles déjà dit que vous avez un accent ? (régional ou étranger) :

Oui, souvent ☐ Oui, parfois ☐ Non, jamais ☐

Père :
Lieu d'origine: ..... Profession : .....

Langues parlées (étrangères et/ou régionales) : .....

Mère :
Lieu d'origine: ..... Profession : .....

Langues parlées (étrangères et/ou régionales) : .....

Y a-t-il dans votre entourage d'autres personnes (grands-parents, nourrice,…) qui parlent plusieurs langues (étrangères ou régionales) ? Si oui, précisez :

.....

.....

FIGURE 6.1 – Questionnaire 1, destiné à évaluer la variété de français parlé. Adapté du questionnaire utilisé dans le projet PFC (Durand *et al.*, 2003a).

Questionnaire 2

Liste 1

Lisez attentivement chaque phrase, et décidez si l'affirmation est *Vraie* ou *Fausse* selon qu'elle s'applique à vous personnellement. Certaines situations sont imaginaires, mais essayez de répondre ce que vous feriez si cela vous arrivait.

		V	F
1.	Avant de voter, j'examine minutieusement les programmes de tous les candidats	☐	☐
2.	Je n'hésite jamais à me donner du mal pour aider quelqu'un en difficulté	☐	☐
3.	Il m'est parfois difficile de poursuivre mon travail si on ne m'encourage pas	☐	☐
4.	Je n'ai jamais détesté quelqu'un profondément	☐	☐
5.	Il m'est arrivé de temps en temps à douter de mes capacités à réussir dans la vie	☐	☐
6.	J'ai parfois du ressentiment lorsque je n'obtiens pas ce que je veux	☐	☐
7.	Je suis toujours attentif(ve) à ma façon de m'habiller	☐	☐
8.	Mes manières à table sont aussi bonnes à la maison qu'au restaurant	☐	☐
9.	Si je pouvais aller voir un film sans payer et être sûr(e) qu'on ne me voie pas je le ferai certainement	☐	☐
10.	A quelques occasions, j'ai renoncé à faire quelque chose parce que je sous-estimais mes capacités	☐	☐
11.	J'aime bien faire des ragots par moments	☐	☐
12.	Il y a eu des moments où j'ai eu envie de me rebeller contre des personnes qui représentaient l'autorité même si je savais qu'elles avaient raison	☐	☐
13.	Peu importe à qui je parle, je sais toujours être à l'écoute	☐	☐
14.	Je me rappelle d'avoir 'fait le malade' pour me sortir d'une situation	☐	☐
15.	Il y a eu des occasions où j'ai profité de quelqu'un	☐	☐
16.	Je suis toujours prêt à admettre lorsque je fais une erreur	☐	☐
17.	J'essaie toujours de mettre en pratique ce que je prêche	☐	☐
18.	Je ne trouve pas cela particulièrement difficile de m'entendre avec des personnes "grande gueule" et antipathiques	☐	☐
19.	J'essaie parfois d'être quitte plutôt que de pardonner et oublier	☐	☐
20.	Lorsque j'ignore quelque chose ça ne me dérange pas du tout de l'admettre	☐	☐
21.	Je suis toujours courtois, même avec des personnes désagréables	☐	☐
22.	Par moments j'ai vraiment insisté pour obtenir les choses telles que je les voulais	☐	☐
23.	Il y a eu des occasions où j'avais envie de tout casser	☐	☐
24.	Je ne pourrais pas imaginer laisser quelqu'un d'autre être puni pour mes fautes	☐	☐
25.	Je n'ai jamais de rancune lorsqu'on me demande de rendre un service en retour	☐	☐
26.	Je n'ai jamais été irrité lorsque des personnes exprimaient des idées très différentes des miennes	☐	☐
27.	Je ne fais jamais un long voyage sans vérifier la sécurité de ma voiture	☐	☐

1/2

FIGURE 6.2 – Questionnaire 2 (p. 1/2), destiné à évaluer différentes mesures de compétence sociale. Liste 1 : désirabilité sociale (Crowne et Marlowe, 1960). Liste 2 : Self-monitoring Lennox et Wolfe (1984).

28. Il y a eu des moments où j'étais assez jaloux de la chance des autres ..... ☐ ☐
29. Je n'ai presque jamais ressenti le besoin de passer un savon à quelqu'un ..... ☐ ☐
30. Je suis parfois irrité par les personnes qui me demandent un service ..... ☐ ☐
31. Je n'ai jamais senti que j'ai été puni sans raison ..... ☐ ☐
32. Je pense parfois, lorsqu'il arrive un malheur à des personnes, qu'ils n'ont que ce qu'ils méritent ..... ☐ ☐
33. Je n'ai jamais délibérément dit quelque chose qui puisse blesser quelqu'un ..... ☐ ☐

Lisez attentivement chaque phrase, et donnez votre réponse selon l'échelle suivante :

0 : Certainement, toujours faux
1 : Généralement faux
2 : Faux d'une certaine façon, mais avec exceptions
3 : Vrai d'une certaine façon, mais avec exceptions
4 : Généralement vrai
5 : Certainement, toujours vrai

Liste 2

		0	1	2	3	4	5
1.	Dans des situations sociales, j'ai la capacité de changer mon comportement si je sens que c'est nécessaire	☐	☐	☐	☐	☐	☐
2.	Je suis souvent capable de déchiffrer correctement les vraies émotions des gens dans leurs yeux	☐	☐	☐	☐	☐	☐
3.	J'ai la capacité de contrôler la façon dont j'apparais aux autres, en fonction de l'impression que je veux leur donner	☐	☐	☐	☐	☐	☐
4.	Dans une conversation, je suis sensible au moindre changement dans l'expression faciale de la personne avec qui je parle	☐	☐	☐	☐	☐	☐
5.	Mes pouvoirs d'intuition sont plutôt bons lorsqu'il s'agit de comprendre les émotions et les motivations des autres	☐	☐	☐	☐	☐	☐
6.	Je suis en général capable de savoir lorsque les autres pensent qu'une blague est de mauvais goût, même s'ils rient de façon convaincante	☐	☐	☐	☐	☐	☐
7.	Quand je sens que l'image que je donne ne marche pas, je peux facilement la changer en quelque chose qui marche	☐	☐	☐	☐	☐	☐
8.	Je suis d'habitude capable de savoir lorsque j'ai dit quelque chose de déplacé en le lisant dans les yeux de mon interlocuteur	☐	☐	☐	☐	☐	☐
9.	J'ai des difficultés à changer mon comportement pour m'adapter à différentes personnes et à différentes situations	☐	☐	☐	☐	☐	☐
10.	Je me suis rendu compte que je peux ajuster mon comportement pour m'adapter à n'importe quelle situation dans laquelle je me trouve	☐	☐	☐	☐	☐	☐
11.	Si quelqu'un me ment, je le sais d'habitude tout de suite à partir de sa façon de s'exprimer	☐	☐	☐	☐	☐	☐
12.	Même lorsque ce pourrait être dans mon avantage, j'ai des difficultés à donner une bonne image de moi-même	☐	☐	☐	☐	☐	☐
13.	Une fois que je sais ce que demande la situation, il m'est facile de réguler mes actions en conséquence	☐	☐	☐	☐	☐	☐

FIGURE 6.3 – Questionnaire 2 (p. 2/2).

DP	SC	Liste 1	Liste 2	Liste 3	Liste 4	Liste 5
Schw.	S1	Correfère	Ollevinté	Convetumeau	Audelippe	
		Danterauque	Sorrediveau	Etullepaure	Erebisch	
	S2	Botonne	Lodolle	Attilaume	Dundomme	Chambourre
		Correfère	Longraule	Etullepaure	Itinause	Longaule
		Sambaule	Santère	Tungaube		Malogue
	S3	Danterauque	Aduraque		Audelippe	Udurausse
					Bunglauche	
	S4	Outimil	Edinel	Chomeub	Moducam	Jeanbril
		Stimen	Punquel	Linkiem	Roskouz	Vinstig
	S5	Adinac	Contor	Dévoc	Erebisch	Agator
		Vinquior	Eturec	Inducar	Lotop	Itunis
				Sunlic		
Post.	B1	Botonne	Brozio	Chomeub	Lotop	Samoté
		Noturé	Clotien	Lodini	Moducam	Vodino
			Lodolle		Noturais	
	B2	Correfère	Bordula	Cortainté	Roskouz	Fordunquais
		Vicolfi	Sorrediveau	Rolphonsi	Taimbortan	Gordité
Moy.	M1	Noturé	Lanturé	Auranssié	Blandré	Gordité
		Sondité	Ollevinté	Cortainté	Mandibé	Samoté
	M2	Bazintais	Guintinais	Pandurais	Chontumais	Fordunquais
		Lundurais	Juntivais	Santinais	Noturais	Mondurais
	M3	Botonne	Contor	Dévoc	Dundomme	Malogue
		Vinquior	Lodolle		Lotop	Agator
	M4	Danterauque	Aduraque	Attilaume	Bunglauche	Longaule
		Sambaule	Longraule	Etullepaure	Itinause	Udurausse
				Tungaube		
Cor.	C1	Outimil	Guintinais	Attilaume	Gatimon	Chuntica
		Stimen	Juntivais	Santinais	Itinause	Ortibeau
						Vinstig
	C2	Matuca	Eturec	Convetumeau	Chontumais	Itunis
		Noturé	Lanturé	Etullepaure	Noturais	Vintudan
	C3	Adinac	Edinel	Erdimon	Mandibé	Gordité
		Sondité	Sorrediveau	Lodini	Taindiron	Vodino
	C4	Dunduco	Adurauque	Inducar	Fondula	Mondurais
		Lundurais	Bordula	Pandurais	Moducam	Udurausse
Nas.	N1	Bazintais	Guintinais	Cortainté	Taimbortan	Vinstig
		Vinquior	Ollevinté	Linkiem	Taindiron	Vintudan
	N2	Dunduco	Juntivais	Sunlic	Bunglauche	Chuntica
		Lundurais	Punquel	Tungaube	Dundomme	Fordunquais
	N3	Danterauque	Lanturé	Auranssié	Blandré	Chambourre
		Sambaule	Santère	Pandurais	Mandibé	Jeanbril
				Santinais		
	N4	Léonstan	Contor	Convetumeau	Chontumais	Longaule
		Sondité	Longraule	Rolphonsi	Fondula	Mondurais

TABLE 6.1 – Cinq listes de noms générés. Chaque liste comporte 16 noms. Lorsqu'un nom contient plusieurs segments critiques, il apparaît autant de fois dans la colonne. La forme orthographique du segment critique est soulignée. Sa prononciation attendue est listée dans la table 3.1.

Symbole	IPA	Précisions
i	i	
y	y	
u	u	
e	e	
E	ɛ	
2	ø	
9	œ	
o	o	
O	ɔ	
a	a	
°	ə	
5	ɛ̃	
1	œ̃	
§	ɔ̃	
@	ɑ̃	
j	j	
8	ɥ	
w	w	
p	p	
b	b	
m	m	
f	f	
v	v	
t	t	
d	d	
n	n	
s	s	
z	z	
l	l	
S	ʃ	
Z	ʒ	
k	k	
g	g	
G	ŋ	
R	ʁ	
.		N'importe quel son
V	iyueɛøœɔaɛ̃œ̃ɔ̃ɑ̃	Voyelles sauf [ə]
W	iyueɛøœaɛ̃œ̃ɔ̃ɑ̃	Voyelles sauf [ɔɔə]
X	iyuøœɔaɛ̃œ̃ɔ̃ɑ̃	Voyelles sauf [eɛə]
C	jɥwpbmfvtdnszlʃʒkgŋʁ	Consonnes et semi-consonnes
T	pftsʃkʁ	Consonnes non voisées
D	bmvdnzlʒg	Consonnes voisées

TABLE 6.2 – Codage phonétique utilisé pour l'écriture des patrons. A chaque symbole est associé un son ou une classe de son du français, dont la transcription est donnée dans l'alphabet phonétique international (IPA).

Glossaire

Axe discriminant Droite passant par les centres de gravité des deux classes de référence de l'analyse discriminante linéaire.

DCT *Discrete Cosine Transform.* Transformée en Cosinus Discrète. Transformation qui décompose une portion de signal en un ensemble de coefficients, qui sont les amplitudes d'ondes cosinus de demi-période et de fréquence croissante.

Dyade Paire de deux locuteurs engagés dans une interaction conversationnelle.

Dimension phonologique Ensemble de variables phonologiques décrivant une différence phonétique et/ou phonologique entre les variétés français standard et français méridional.

Modèle acoustique Objet statistique décrivant la réalisation moyenne d'un segment phonétique observée dans un corpus d'apprentissage.

MFCC *Mel-Frequency Cepstral Coefficients.* Coefficients représentant le spectre de puissance à court terme d'un son, basé sur une transformée en cosinus du logarithme d'un spectre sur l'échelle de fréquence non-linéaire en Mel. Ils sont traditionnellement utilisés en reconnaissance automatique de la parole en raison de leur robustesse.

Score discriminant Valeur associée à un segment critique après projection des coefficients DCT le décrivant sur l'axe discriminant.

Segment critique Variable phonologique insérée dans un mot cible.

Variable phonologique Séquence d'un ou plusieurs segments phonétiques décrivant une différence de réalisation entre les variétés français standard et français méridional.

Variante de prononciation Séquence de modèles acoustiques incluses dans un dictionnaire de pronconciation utilisé pour l'alignement automatique.

Word-spotting Localisation de mot-cibles dans le signal de parole.

Bibliographie

Abercrombie, D. 1967, *Elements of General Phonetics*, Aldine, Chicago.

Adank, P., R. Smits et R. van Hout. 2004, «A comparison of vowel normalization procedures for language variation research», *The Journal of the Acoustical Society of America*, vol. 116, p. 3099–3107.

Adda-Decker, M. et L. Lamel. 1999, «Pronunciation variants across system configuration, language and speaking style», *Speech Communication*, vol. 29, n° 2, p. 83–98.

Anderson, A. H., M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thomson et R. Weinert. 1991, «The HCRC Map Task corpus», *Language and Speech*, vol. 34, n° 4, p. 351–366.

Armstrong, N. 2002, «Nivellement et standardisation en anglais et en français», *Langage et société*, vol. 102, p. 5–32.

Armstrong, N. et M. Jamin. 2002, «Le français des banlieues : uniformity and discontinuity in the French of the Hexagon», dans *French in and out of France : Language Policies, Intercultural Antagonisms and Dialogues*, édité par K. Sahli, Peter Lang, Bern, p. 107–136.

Atkinson, J. et J. Heritage. 2006, «Jefferson's transcript notation», dans *The Discourse Reader*, édité par A. Jaworski et N. Coupland, Routledge, London, p. 158–165.

Aubanel, V. et N. Nguyen. 2010, «Automatic recognition of regional phonological variation in conversational interaction», *Speech Communication*, vol. 52, p. 577–586.

Auvinen, P. 2009, *Achievement of Intersubjectivity in Airline Cockpit Interaction*, Thèse de doctorat, University of Tampere, Tampere, Finlande.

Baayen, R. H. 2008, *Analyzing linguistic data. A practical introduction to statistics*, Cambridge University Press, Cambridge, 400 p..

Babel, M. 2009, *Phonetic and Social Selectivity in Speech Accommodation*, Thèse de doctorat, UC Berkeley, Berkely, CA, USA.

Bailly, G. et A. Lelong. 2010, «Speech dominoes and phonetic convergence», dans *Proceedings of Interspeech*, Makuhari, Japan, 26-30 Septembre, p. à paraître.

Bates, R. A., M. Ostendorf et R. A. Wright. 2007, «Symbolic phonetic features for modeling of pronunciation variation», *Speech Communication*, vol. 49, n° 2, p. 83–97.

Bertrand, R., P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde et S. Rauzy. 2008, «Le CID – Corpus of Interactional Data : Annotation et exploitation multimodale de parole conversationnelle», *Traitement Automatique des Langues*, vol. 49, n° 3, p. 105–134.

Binisti, N. et M. Gasquet-Cyrus. 2003, «Les accents de Marseille», *Cahiers du français contemporain*, vol. 8, p. 107–129.

Blanche-Benveniste, C. et A. Chervel. 1978, *L'orthographe*, 3e éd., F. Maspéro, Paris, 260 p..

Bloomfield, L. 1933, *Language*, Holt, New-York.

Boughton, Z. 2005, «Accent levelling and accent localisation in northern French : Comparing Nancy and Rennes», *French Language Studies*, vol. 15, p. 235–256.

Boughton, Z. 2007, «Ce que prononcer veut dire : The social value of variable phonolgy in French», *Nottingham French Studies*, vol. 46, n° 2, p. 7–22.

Boula de Mareüil, P., M. Adda-Decker et C. Woehrling. 2007, «Analysis of oral and nasal vowel realisation in northern and southern french varieties», dans *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)*, Saarbrücken, 6-10 Août, p. 2221–2224.

Boula de Mareüil, P., B. Vieru-Dimulescu, C. Woehrling et M. Adda-Decker. 2008, «Accents étrangers et régionaux en français : Caractérisation et identification», *Traitement Automatique des Langues*, vol. 49, n° 3, p. 135–163.

Bradlow, A., L. Nygaard et D. Pisoni. 1999, «Effects of talker, rate, and amplitude variation on recognition memory for spoken words», *Perception and Psychophysics*, vol. 61, n° 2, p. 206–219.

Bradlow, A. R., R. E. Baker, A. Choi, M. Kim et K. J. Van Engen. 2007, «The Wildcat corpus of native- and foreign-accented English», *The Journal of the Acoustical Society of America*, vol. 121, n° 5, Pt. 2, p. 3072.

Branigan, H., M. Pickering et A. Cleland. 2000, «Syntactic co-ordination in dialogue», *Cognition*, vol. 75, n° 2, p. B13–B25.

Brunellière, A., S. Dufour, N. Nguyen et U. H. Frauenfelder. 2009, «Behavioral and electrophysiological evidence for the impact of regional variation on phoneme perception», *Cognition*, vol. 111, n° 3, p. 390–396.

Bürki, A., C. Gendrot, G. Gravier, G. Linarès et C. Fougeron. 2008, «Alignement automatique et analyse phonétique : comparaison de différents systèmes pour l'analyse du schwa», *Traitement Automatique des Langues*, vol. 49, n° 3, p. 165–197.

Bybee, J. 2002, «Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change», *Language Variation and Change*, vol. 14, n° 3, p. 261–290.

Bybee, J. et J. McClelland. 2005, «Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition», *The Linguistic Review*, vol. 22, n° 2-4, p. 381–410.

Caplan, P. 1987, *The cultural construction of sexuality*, Routledge, London.

Caramazza, A. et G. H. Yeni-Komshian. 1974, «Voice onset time in two French dialects», *Journal of Phonetics*, vol. 2, p. 239–245.

Carton, F., M. Rossi, D. Autesserre et P. Léon. 1983, *Les accents des Français*, Hachette, Paris.

Chen, M. Y. 2000, «Nasal detection module for a knowledge-based speech recognition system», dans *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, vol. IV, Beijing, China, p. 636–639.

Chomsky, N. 1965, *Aspects of the theory of syntax*, MIT Press, Cambridge, MA.

Church, B. et D. Schacter. 1994, «Perceptual specificity of auditory priming : Implicit memory for voice intonation and fundamental frequency», *Journal of Experimental Psychology : Learning, Memory, and Cognition*, vol. 20, n° 3, p. 521–533.

Clopper, C. G. et A. R. Bradlow. 2008, «Perception of dialect variation in noise : Intelligibility and classification», *Language and Speech*, vol. 51, n° 3, p. 175–198.

Clopper, C. G. et J. B. Pierrehumbert. 2008, «Effects of semantic predictability and regional dialect on vowel space reduction», *The Journal of the Acoustical Society of America*, vol. 124, n° 3, p. 1682–8.

Connine, C. M. 2004, «It's not what you hear but how often you hear it : On the neglected role of phonological variant frequency in auditory word recognition», *Psychonomic Bulletin & Review*, vol. 11, n° 6, p. 1084–1089.

Conrey, B., G. F. Potts et N. A. Niedzielski. 2005, «Effects of dialect on merger perception : ERP and behavioral correlates», *Brain and Language*, vol. 95, p. 435–449.

Coquillon, A. 2005, *Caractérisation prosodique du parler de la région marseillaise*, Thèse de doctorat, Université de Provence, Aix-en-Provence, France.

Coveney, A. 2001, *The Sounds of Contemporary French : Articulation and Diversity*, Elm Bank Publications, Exeter, UK.

Crowne, D. P. et D. Marlowe. 1960, «A new scale of social desirability independent of psychopathology», *Journal of Consulting Psychology*, vol. 24, p. 349–354.

Cutler, A., R. Smits et N. Cooper. 2005, «Vowel perception : Effects of non-native language vs. non-native dialect», *Speech Communication*, vol. 47, n° 1-2, p. 32–42.

Delvaux, V. et A. Soquet. 2007, «The influence of ambient speech on adult speech productions through unintentional imitation», *Phonetica*, vol. 64, n° 2-3, p. 145–173.

Deshmukh, N., A. Ganapathiraju et J. Picone. 1999, «Hierarchical search for large-vocabulary conversational speech recognition», *IEEE Signal Processing magazine*, p. 84–107.

Dijksterhuis, A. et J. A. Bargh. 2001, «The perception-behavior expressway : Automatic effects of social perception on social behavior», *Collection*, vol. 33, p. 1 – 40.

Docherty, G. 2007, «Speech in its natural habitat : Accounting for social factors in phonetic variability», dans *Papers in Laboratory Phonology IX*, édité par J. Cole et J. I. Hualde, Mouton de Gruyter, Berlin, p. 1–35.

Docherty, G. et P. Foulkes. 1999, «Instrumental phonetics and phonological varlatlon : Case studies from Newcastle upon Tyne and Derby», dans *Urban Voices : Accent Studies in the British Isles*, édité par P. Foulkes et G. Docherty, Arnold, London, p. 47–71.

Doehler, S. P. 2010, «Conceptual changes and methodological challenges : on language, learning and documenting learning from a conversation analytic perspective on SLA», dans *Conceptualising Learning in Applied Linguistics*, édité par P. Seedhouse, S. T. Walsh et C. Jenks, Palgrave Macmillan, p. 105–127.

Dufour, S., N. Nguyen et U. H. Frauenfelder. 2007, «The perception of phonemic contrasts in a non-native dialect», *Journal of the Acoustical Society of America Express Letters*, vol. 121, p. 131–136.

Durand, J. 1988, «Les phénomènes de nasalité en français du Midi : phonologie de dépendance et sous-spécification», *Recherches Linguistiques de Vincennes*, vol. 17, p. 29–54.

Durand, J. 1990, *Generative and Non-Linear Phonology*, Longman, London.

Durand, J., B. Laks et C. Lyche. 2003a, «Le projet « Phonologie du français contemporain » (PFC)», *La Tribune Internationale des Langues Vivantes*, vol. 33, p. 3–9.

Durand, J., B. Laks et C. Lyche. 2003b, «Linguistique et variation : quelques réflexions sur la variation phonologique», dans *Corpus et variation en phonologie du français : méthodes et analyses*, édité par E. Delais-Roussarie et J. Durand, Presses Universitaires du Mirail, Toulouse, p. 11–88.

Durand, J. et C. Lyche. 2004, «Structure et variation dans quelques systèmes vocaliques du français : l'enquête « Phonologie du français contemporain (PFC) »», dans *Variation et Francophonie*, édité par A. Coveney et C. Sanders, L'Harmattan, Paris, p. 217–240.

Evans, B. G. et P. Iverson. 2004, «Vowel normalization for accent : An investigation of perceptual plasticity in young adults», *The Journal of the Acoustical Society of America*, vol. 115, p. 352–361.

Eychenne, J. 2006, *Aspects de la phonologie du schwa dans le français contemporain*, Thèse de doctorat, Université de Toulouse-Le Mirail, Toulouse, France.

Fagyal, Z. 2006, «Phonetics and phonology», dans *French : A Linguistic Introduction*, édité par Z. Fagyal, D. Kibbee et F. Jenkins, Cambridge University Press, Cambridge, p. 17–78.

Fagyal, Z. 2010, *Accents de banlieue : aspects prosodiques du français populaire en contact avec les langues de l'immigration*, L'Harmattan, Paris.

Fagyal, Z., S. Hassa et F. Ngom. 2002, «L'opposition [e]-[ɛ] en syllabes ouvertes de fin de mot en français parisien : étude acoustique préliminaire», dans *Journées d'Etudes sur la Parole*, vol. XXIV, Nancy, p. 165–168.

Fant, G. 1960, *Acoustic Theory of Speech Production*, Mouton, The Hague.

Floccia, C., J. Goslin, F. Girard et G. Konopczynski. 2006, «Does a regional accent perturb speech processing ?», *Journal of Experimental Psychology : Human Perception and Performance*, vol. 32, n° 5, p. 1276–1293.

Fonagy, I. 1989, «Le français change de visage ?», *Revue Romane*, vol. 24, p. 225–254.

Ford, C. E. et E. Couper-Kuhlen. 2004, «Conversation and phonetics», dans *Sound Patterns in Interaction : Cross-Linguistic Studies from Conversation*, édité par E. Couper-Kuhlen et C. E. Ford, Benjamins, Amsterdam, p. 3–25.

Forrest, K., G. Weismer, P. Milenkovic et R. N. Dougall. 1988, «Statistical analysis of word-initial voiceless obstruents : Preliminary data», *Journal of the Acoustical Society of America*, vol. 84, p. 115–124.

Foulkes, P. et G. Docherty. 1999, *Urban Voices : Accent Studies in the British Isles*, Arnold, London.

Foulkes, P. et G. Docherty. 2006, «The social life of phonetics and phonology», *Journal of Phonetics*, vol. 34, p. 409–438.

Foulkes, P., J. Scobbie et D. Watt. 2010, «Sociophonetics», dans *The Handbook of Phonetic Sciences, 2nd edition*, édité par W. J. Hardcastle, J. Laver et F. E. Gibbon, Blackwell, Oxford, p. 1–53.

Gadet, F. 2003, *La variation sociale en français*, Ophrys, Gap, 1–67 p..

Garrod, S. et A. Anderson. 1987, «Saying what you mean in dialogue : A study in conceptual and semantic co-ordination», *Cognition*, vol. 27, n° 2, p. 181–218.

Gaskell, M. G. 2003, «Modelling regressive and progressive effects of assimilation in speech perception», *Journal of Phonetics*, vol. 31, p. 447–463.

Goldinger, S. D. 1996, «Words and Voices : Episodic Traces in Spoken Word Identification and Recognition Memory», *Journal of Experimental Psychology : Learning, Memory, and Cognition*, vol. 22, n° 5, p. 1166–1183.

Goldinger, S. D. 1997, «Words and voices : Perception and production in an episodic lexicon», dans *Talker variability in speech processing*, édité par K. Johnson et J. W. Mullennix, Academic Press, San Diego, p. 33–66.

Goldinger, S. D. 1998, «Echoes of echoes ? An episodic theory of lexical access», *Psychological Review*, vol. 105, n° 2, p. 251–279.

Gravano, A. et J. Hirschberg. 2010, «Turn-taking cues in task-oriented dialogue», *Computer Speech & Language*, p. sous presse.

Gregory, S. W. et S. Webster. 1996, «A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions», *Journal of Personality and Social Psychology*, vol. 70, n° 6, p. 1231–40.

Hansen, A. B. 2001, «Les changements actuels des voyelles nasales du français parisien : confusions ou changements en chaîne ?», *La Linguistique*, vol. 37, n° 2, p. 33–47.

Harrington, J. 2006, «An acoustic analysis of 'happy-tensing' in the Queen's Christmas broadcasts», *Journal of Phonetics*, vol. 34, p. 439–457.

Harrington, J. 2010, «Acoustic Phonetics», dans *The Handbook of Phonetic Sciences, 2nd edition*, édité par W. J. Hardcastle, J. Laver et F. E. Gibbon, Blackwell, Oxford, UK, p. 81–129.

Hawkins, S. 2003, «Roles and representations of systematic fine phonetic detail in speech understanding», *Journal of Phonetics*, vol. 31, n° 3-4, p. 373–405.

Hawkins, S. 2010, «Phonetic variation as communicative system : Perception of the particular and the abstract», *Laboratory Phonology X : Variation*.

Hay, J. et K. Drager. 2007, «Sociophonetics», *Annual Review of Anthropology*, vol. 36, p. 89–103.

Hay, J., A. Nolan et K. Drager. 2006, «From Fush to Feesh : Exemplar Priming in Speech Perception», *The Linguistic Review*, vol. 23, n° 3, p. 351–379.

Hay, J., A. Walker, G. Docherty et R.-J. Whitcombe. 2009, «Talking about Old Events using Old /t/s», dans *NWAV*, vol. 37, Houston, Texas, p. 1–1.

Heller, K. A. 2008, *Efficient Bayesian Methods for Clustering*, Thèse de doctorat, University of College London, London, UK.

Heritage, J. 1984, *Garfinkel and Ethnomethodology*, Polity Press, New York.

Hintzman, D. L. 1986, «"schema abstraction" in a multiple-trace memory model», *Psychological Review*, vol. 93, n° 4, p. 411–428.

Hosom, J.-P. 2008, «Speaker-independent phoneme alignment using transition-dependent states», *Speech Communication*, p. 1–17.

Hultzén, L., J. Allen et M. S. Miron. 1964, *Tables of Transitional Frequencies of English Phonemes*, University of Illinois Press, Urbana, IL.

Jamin, M., C. Trimaille et M. Gasquet-Cyrus. 2006, «De la convergence dans la divergence : le cas des quartiers pluri-ethniques en France», *French Language Studies*, vol. 16, p. 335–356.

Jefferson, G. 2004, «Glossary of transcript with an introduction», dans *Conversation Analysis. Studies from the first generation*, édité par G. H. Lerner, Benjamins, Amsterdam, p. 13–31.

Johnson, K. 1997, «Speech perception without speaker normalization», dans *Talker variability in speech processing*, édité par K. Johnson et J. W. Mullennix, Academic Press, San Diego, p. 145–165.

Johnson, K. 2006, «Resonance in an exemplar-based lexicon : The emergence of social identity and phonology», *Journal of Phonetics*, vol. 34, n° 4, p. 485–499.

Jongman, A., R. Wayland et S. Wong. 2000, «Acoustic characteristics of English fricatives», *The Journal of the Acoustical Society of America*, vol. 108, n° 3, p. 1252.

Joos, M. 1948, «Acoustic Phonetics», *Language*, vol. 24, p. 1–140.

Jurafsky, D. 2003, «Probabilistic modeling in psycholinguistics : Linguistic comprehension and production», dans *Probabilistic Linguistics*, édité par R. Bod, J. Hay et S. Jannedy, MIT Press, Cambridge, p. 39–95.

Kelly, J. et J. Local. 1989, *Doing Phonology*, Manchester University Press, Manchester.

Kerswill, P. 2003, «Dialect levelling and geographical diffusion in British English», dans *Social dialectology. In honour of Peter Trudgill*, édité par D. Britain et J. Cheshire, Benjamins, Amsterdam, p. 223–243.

Kerswill, P. et A. Williams. 2002, «"salience" as an explanatory factor in language change : Evidence from dialect levelling in urban england», dans *Language change. The interplay of internal, external and extra-linguistic factors*, édité par M. Jones et E. Esch, Mouton de Gruyter, Berlin, p. 81–110.

Kraljic, T., S. Brennan et A. G. Samuel. 2008, «Accommodating variation : Dialects, idiolects, and speech processing», *Cognition*, vol. 107, n° 1, p. 54–81.

Kuhl, P., J. E. Andruski, L. A. Chistovich, E. Kozhevnikova, V. Ryskina, E. Stolyarova, U. Sundberg et F. Lacerda. 1997, «Cross-language analysis of phonetic units in language addressed to infants», *Science*, vol. 277, n° 5326, p. 684.

Labov. 1966, «The Social Stratification of English in New York City», cahier de recherche, Center for Applied Linguistics, Washington, D. C.

Labov, W. 2001, *Principles of Linguistic Change : Social Factors*, Blackwell, Oxford.

Ladefoged, P. et D. E. Broadbent. 1957, «Information conveyed by vowels», *The Journal of the Acoustical Society of America*, vol. 29, n° 1, p. 98–104.

Lahiri, A. et H. Reetz. 2002, «Underspecified recognition», *Laboratory Phonology 7*, vol. Mouton, n° Berlin, p. 637–675.

Lennox, R. D. et R. N. Wolfe. 1984, «Revision of the self-monitoring scale», *Journal of Personality and Social Psychology*, vol. 46, n° 6, p. 1349–1364.

Liu, S. A. 1996, «Landmark detection for distinctive feature-based speech recognition», *The Journal of the Acoustical Society of America*, vol. 100, p. 3417–3430.

Local, J. 2003, «Variable domains and variable relevance : interpreting phonetic exponents», *Journal of Phonetics*, vol. 31, n° 3-4, p. 321–339.

Malécot, A. et P. Lindsay. 1976, «The neutralization of /ɛ̃/ – /œ̃/ in French», *Phonetica*, vol. 33, p. 45–61.

Malsheen, B. J. 1980, «Two hypotheses for phonetic clarification in the speech of mothers to children», dans *Child phonology (Vol. 2). Perception*, édité par G. H. Yeni-Komshian, J. F. Kavanagh et C. A. Ferguson, Academic Press, New York, p. 173–184.

Marlowe, D. et D. P. Crowne. 1961, «Social desirability and responses to perceived situational demands», *Journal of Consulting Psychology*, vol. 25, p. 100–115.

Martinet, A. 1945, *La Prononciation du Français Contemporain*, 2e éd., Droz, Geneva.

Martinet, A. 1958, «C'est jeuli, le Mareuc !», *Romance Philology*, vol. 11, p. 345–355.

Mondada, L. 2003, «Working with video : how surgeons produce video records of their action», *Visual Studies*, vol. 18, n° 1, p. 58–72.

Mondada, L. 2006, «Interactions en situations professionnelles et institutionnelles : de l'analyse détaillée aux retombées pratiques», *Revue française de linguistique appliquée*, vol. 11, p. 5–16.

Natale, M. 1975, «Convergence of mean vocal intensity in dyadic communication as a function of social desirability», *Journal of Personality and Social Psychology*, vol. 32, n° 5, p. 790–804.

New, B., C. Pallier, M. Brysbaert et L. Ferrand. 2004, «Lexique 2 : A new French lexical database», *Behavior Research Methods, Instruments, & Computers*, vol. 36, n° 3, p. 516–524.

Ney, H. et S. Ortmanns. 1999, «Dynamic programming search for continuous speech recognition», *IEEE Signal Processing magazine*, p. 64–83.

Nguyen, N. 2010, «Representations of speech sound patterns in the speaker's brain : Insights from perception studies», *Handbook of Laboratory Phonology*.

Nosofsky, R. M. 1986, «Attention, similarity, and the identification-categorization relationship», *Journal of Experimental Psychology : General*, vol. 115, n° 1, p. 39–57.

Nye, P. W. et C. A. Fowler. 2003, «Shadowing latency and imitation : the effect of familiarity with the phonetic patterning of English», *Journal of Phonetics*, vol. 31, n° 1, p. 63–79.

Ogden, R. 2004, «Non-modal voice quality and turn-taking in Finnish», dans *Sound Patterns in Interaction : Cross-Linguistic Studies from Conversation*, édité par E. Couper-Kuhlen et C. E. Ford, Benjamins, Amsterdam, p. 29–62.

Pardo, J. S. 2006, «On phonetic convergence during conversational interaction», *The Journal of the Acoustical Society of America*, vol. 119, p. 2382–2393.

Pickering, M. J. et S. Garrod. 2004, «Toward a mechanistic psychology of dialogue», *Behavioral and Brain Sciences*, vol. 27, n° 02, p. 169–190.

Pierrehumbert, J. B. 2001, «Exemplar dynamics : Word frequency, lenition and contrast», dans *Frequency effects and the emergence of linguistic structure*, édité par J. Bybee et P. Hopper, John Benjamins, Amsterdam, p. 137–157.

Pierrehumbert, J. B. 2002, «Word-specific phonetics», dans *Papers in Laboratory Phonology VII : Phonology and Phonetics*, édité par C. Gussenhoven et N. Warner, Mouton de Gruyter, Berlin, p. 101–140.

Pierrehumbert, J. B. 2003, «Phonetic diversity, statistical learning, and acquisition of phonology», *Language and Speech*, vol. 46, n° 2-3, p. 115–154.

Pierrehumbert, J. B. 2006, «The next toolkit», *Journal of Phonetics*, vol. 34, p. 516–530.

Rabiner, L. R. 1989, «A tutorial on Hidden Markov Models and selected applications in speech recognition», *Proceedings of the IEEE*, vol. 77, n° 2, p. 257–286.

Racine, I. 2008, *Les effets de l'effacement du schwa sur la production de la parole en français*, Thèse de doctorat, Université de Genève, Genève.

Reed, B. S. 2004, «Turn-final intonation in English», dans *Sound Patterns in Interaction : Cross-Linguistic Studies from Conversation*, édité par E. Couper-Kuhlen et C. E. Ford, Benjamins, Amsterdam, p. 97–118.

Sancier, M. L. et C. A. Fowler. 1997, «Gestural drift in a bilingual speaker of Brazilian Portuguese and English», *Journal of Phonetics*, vol. 25, p. 421–436.

Shannon, C. E. 1948, «A mathematical theory of communication», *The Bell System Technical Journal*, vol. 27, p. 379–423, 623–656.

Snow, C. E. 1995, «Issues in the study of input : Finetuning, universality, individual and developmental differences, and necessary causes», dans *The handbook of child language*, édité par P. Fletcher et B. MacWhinney, Blackwell, Oxford, p. 180–193.

Snyder, M. 1974, «Self-monitoring of expressive behavior», *Journal of Personality and Social Psychology*, vol. 30, n° 4, p. 526–537.

Stevens, K. N. 2002, «Toward a model for lexical access based on acoustic landmarks and distinctive features», *The Journal of the Acoustical Society of America*, vol. 111, p. 1872–1891.

Strand, E. A. 2000, *Gender Stereotype Effects in Speech Processing*, Thèse de doctorat, Ohio State University, Columbus, OH, USA.

Strik, H. et C. Cucchiarini. 1999, «Modeling pronunciation variation for ASR : A survey of the literature», *Speech Communication*, vol. 29, n° 2-4, p. 225–246.

Sumner, M. et A. G. Samuel. 2009, «The effect of experience on the perception and representation of dialect variants», *Journal of Memory and Language*, vol. 60, p. 487–501.

Tabain, M. 2001, «Variability in fricative production and spectra : implications for the hyper- and hypo- and quantal theories of speech production», *Language and Speech*, vol. 44, n° 57-94.

Tanaka, H. 2004, «Prosody for marking transition-relevance places in Japanese conversation : The case ofturns unmarked by utterance-final objects», dans *Sound Patterns in Interaction : Cross-Linguistic Studies from Conversation*, édité par E. Couper-Kuhlen et C. E. Ford, Benjamins, Amsterdam, p. 63–96.

Tannen, D. 1989, *Talking Voices : Repetition, Dialogue and Imagery in Conversational Discourse*, Cambridge University Press, Cambridge, UK.

Trimaille, C. 2008, «Consonnes dentales palatalisées/affriquées en français contemporain : indicateurs, marqueurs et/ou variantes en développement ?», dans *AFLS Conference*, Oxford, UK, 3-5 September.

VanRullen, T., P. Blache, C. Portes, S. Rauzy, J. Maeyhieux, M. Guénot, J. Balfourier et E. Bellengier. 2005, «Une plateforme pour l'acquisition, la maintenance et la validation de ressources lexicales», dans *Proceedings of Traitement Automatique des Langues Naturelles*, Dourdan, France, 6-10 Juin.

Walker, D. C. 1984, *The Pronunciation of Canadian French*, University of Ottawa Press, Ottawa, 186 p..

Walter, H. 1976, *La dynamique des phonèmes dans le lexique français contemporain*, France-Expansion, Paris.

Walter, H. 1982, *Enquête phonologique et variétés régionales du français*, Presses Universitaires de France, Paris.

Walter, H. 1988, *Le français dans tous les sens*, Robert Laffont, Paris.

Woehrling, C. 2009, *Accents régionaux en français : Perception, analyse et modélisation à partir de grands corpus*, Thèse de doctorat, Université Paris-Sud, Paris, France.

Woehrling, C. et P. Boula de Mareüil. 2006, «Identification d'accents régionaux en français : perception et analyse», *Parole*, vol. 37, p. 25–65.

Yaeger-Dror, M. et W. Kemp. 1992, «Lexical classes in Montreal French : The case of (ɛ :)», *Language and Speech*, vol. 35, n° 3, p. 251–293.

Young, S., G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev et P. Woodland. 2006, *The HTK Book*, 3e éd., Cambridge University Engeneering Department, Cambridge, 368 p..