UNIVERSITÉ LUMIÈRE LYON 2 – UNIVERSITÉ DE LYON

HABILITATION À DIRIGER DES RECHERCHES

DE L'IDENTIFICATION DES LANGUES À LA COMPLEXITÉ PHONOLOGIQUE

VOLUME I – SYNTHÈSE & PERSPECTIVES

Synthèse des travaux et programme de recherche présentés par

FRANÇOIS PELLEGRINO

Le 22 janvier 2009 à Lyon

Jury :

Pr. Björn LINDBLOM, Universités du Texas et de Stockholm	(rapporteur externe)
Pr. Ian MADDIESON, Université du Nouveau-Mexique, USA	(rapporteur externe)
Dr. René CARRÉ, DRCE CNRS, Lyon	(rapporteur interne)
Pr. Régine ANDRÉ-OBRECHT, Université de Toulouse	(examinateur)
Pr. Jean-Paul HATON, Université de Nancy, Institut Universitaire de France	(examinateur)
Pr. Jean-Marie HOMBERT, DRCE CNRS, Lyon	(examinateur)

À mes parents

Robert Delaunay, Rythme n°1, Centre Pompidou, Paris (1937)

TABLE DES MATIÈRES – Volume I

LISTE DES FIGURES

LISTE DES TABLEAUX

PRÉAMBULE

Je me livre aujourd'hui à l'exercice de la rédaction d'une habilitation à diriger des recherches. Ce manuscrit reflète donc la démarche scientifique que j'ai adoptée depuis mes travaux de thèse, sans pour autant être exhaustif. Les résultats présentés ici correspondent à des travaux initiés sous la direction de Régine André-Obrecht à l'Institut de Recherche en Informatique de Toulouse et poursuivis depuis au sein du laboratoire Dynamique Du Langage, en collaboration avec de nombreuses personnes[1]. La plupart sont d'anciens doctorants à l'encadrement desquels j'ai participé : Melissa Barkat-Defradas, Ioana Vasilescu, Jérôme Farinas, Jean-Luc Rouas et Emmanuel Ferragne, auxquels il faut ajouter Rym Hamdi et Jalaleddin Al-Tamimi, quoique de manière plus distante, et mes collègues Egidio Marsico, René Carré et Christophe Coupé pour les travaux plus récents portant sur la complexité. D'autres travaux, menés en particulier avec Fanny Meunier sur la compréhension de la parole dégradée, ne seront qu'effleurés au fil de ce document.

Malgré les nombreux collègues évoqués ci-dessus et ceux que j'ai oublié de citer, toute erreur ou imprécision présente dans ce manuscrit relève de mon entière et seule responsabilité.

[1] À titre indicatif, la liste de mes publications et communications fournie en Annexe A fait apparaître une cinquantaine de collègues co-auteurs.

INTRODUCTION

« Nous sommes tous des ignorants,
mais nous n'ignorons pas les mêmes choses »

La démarche scientifique présentée dans ce document trouve ses racines dans le travail de thèse que j'avais entrepris à l'Institut de Recherche en Informatique de Toulouse à l'automne 1995. Ce travail portait sur le développement d'une approche originale pour l'Identification Automatique des Langues (IAL) basée sur l'utilisation de modèles phonétiques non supervisés. Au cours de cette thèse, j'ai eu l'occasion de porter un regard naïf sur les travaux contemporains de collègues linguistes portant sur la typologie de systèmes vocaliques, soit des langues (thèse de Nathalie Vallée, soutenue un an auparavant) soit des dialectes (thèse de Melissa Barkat, alors en cours). La saisissante dichotomie entre d'une part, le « fatras » de segments vocaliques extraits automatiquement d'un important corpus de parole spontanée multi-locuteur et l'élégante synthèse proposée par la description du système phonologique d'autre part, suscita mon intérêt tout autant que ma perplexité : où était donc la *réalité objective*[2] dans tout cela ? Paradoxalement, et plutôt que de me tourner vers le débat discret/continu alimentant l'interface phonétique/phonologie, je me suis alors intéressé à la réalité *subjective* par le biais de l'identification perceptuelle des langues, en prenant part à l'encadrement de Melissa Barkat et de Ioana Vasilescu. En effet, il nous a semblé que l'étude de la perception de langues, en particulier celles non comprises par des auditeurs, était prometteuse pour identifier ce qui est pertinent dans le flux complexe de la parole, ou en d'autres termes, pour identifier la nature et la localisation de l'information dans le signal. L'écoute d'extraits prononcés dans une langue totalement inconnue est un bon exemple d'écoute de *parole* sans accès direct et complet au *langage*. Pour autant, les sons entendus ont une substance langagière et il est difficile de savoir exactement comment le système neurocognitif des auditeurs les traite. Ce questionnement peut être mis en miroir de la perception des langues sifflées, puisqu'on a là affaire à une perception de *langage* sans *parole* au sens strict, ou en tout cas de parole très fortement dégradée (cf. les travaux de Julien Meyer, e.g. Meyer, 2007). Cette problématique pose la question de la limite entre un stimulus auditif non langagier et un stimulus langagier. Par exemple, la parole préenregistrée reproduite avec une inversion temporelle est-elle un stimulus auditif qui active, dans une certaine mesure, des traitements cognitifs spécifiques au langage ou non ? Sous l'impulsion de Jean-Marie Hombert, plusieurs membres du

[2] J'emprunte l'expression de réalité objective (ou *objective reality*) à John J. Ohala (Ohala, 2007).

laboratoire Dynamique Du Langage ont porté leur attention sur ce *no man's land* entre langage et « non-langage », parole et « non-parole » et je contribue moi-même à ces recherches visant à éclairer le fonctionnement neurocognitif de la perception langagière depuis 2001. Pour des raisons de cohérences internes et de longueur, ces travaux ne seront cependant pas développés dans ce manuscrit.

Parallèlement et suite aux travaux de thèse de Franck Ramus (Ramus, 1999) et aux expériences perceptives menées avec Melissa Barkat (e.g. Barkat, Ohala & Pellegrino, 1999), la possibilité d'exploiter une approche basée sur le rythme pour l'identification automatique des langues m'avait paru particulièrement excitante et prometteuse. Excitante parce que la modélisation multilingue et automatique de caractéristiques suprasegmentales de la parole relevait alors du défi et prometteuse parce que ce qui est saillant pour des nourrissons doit bien être extractible du signal acoustique ! Ces recherches, menées avec mes collègues toulousains (Régine André-Obrecht et Jérôme Farinas, puis Jean-Luc Rouas) allaient se révéler particulièrement fécondes et elles m'incitèrent à m'interroger sur la nature même du rythme de la parole, dont le statut oscille dans la littérature entre celui de primitive biologique et simple « effet de bord » de principes phonologiques spécifiques aux langues. La quête d'invariance isochronique, dont les années 1970 et 1980 furent le théâtre, avait en particulier abouti à une remise en cause des catégories rythmiques traditionnelles tout en leur reconnaissant une certaine saillance perceptuelle, cultivant ainsi une dualité – toujours d'actualité – qui se nourrit plus de variabilité que d'invariance.

L'étude de cette variabilité m'amena par la suite à considérer le débit de parole comme étant partie intégrante de l'organisation rythmique des langues. Dès lors, j'ai eu l'intuition, potentiellement fausse, que le rythme pouvait être étudié sous l'angle de l'organisation informationnelle de la parole, tant comme régulateur du débit d'information linguistique que comme indice de répartition de l'information le long du flux de parole. Cette approche nous a récemment amenés, Egidio Marsico, Christophe Coupé et moi-même, à nous intéresser à la quantification du débit d'information phonologique dans des enregistrements produits en plusieurs langues. En faisant appel à des mesures d'entropie de type shannonienne, nous avons émis l'hypothèse de l'existence d'une régulation du débit d'information dans les langues du monde, liée potentiellement à la perception ou à la capacité de la mémoire de travail des locuteurs des différentes langues et sur la structuration des systèmes phonologiques. En effet, ce type de régulation est généralement révélateur d'un principe de moindre effort à l'œuvre au sein de systèmes dynamiques complexes. Ce principe, s'il existe, est alors potentiellement pertinent pour notre compréhension des contraintes et de degrés de liberté régissant les systèmes et leur évolution. Aussi, nous avons adopté les cadres des systèmes dynamiques complexes et de la théorie de l'information pour tenter d'ouvrir de nouvelles perspectives quant à l'analyse typologique des langues du monde, en termes de systèmes phonologiques d'une part et de structuration temporelle de l'information d'autre part.

Ce manuscrit discute les principaux résultats obtenus durant cette période sur les problématiques introduites ci-dessus. Après avoir explicité plus avant les

motivations qui sous-tendent l'ensemble de ces recherches dans la prochaine partie, je décrirai les recherches menées sur le thème de l'identification des langues et des dialectes dans la partie 2. La partie 3, introduite par une discussion sur la relation entretenue par les notions de complexité et d'information en phonologie, explicite les travaux que nous avons entrepris sur la structure et la dynamique des systèmes phonologiques ainsi que sur un hypothétique principe de régulation du débit d'information linguistique. Tout au long de ces deux parties, les perspectives ouvertes par les travaux réalisés sont également mentionnées et discutées. De fait, ce document constitue autant un bilan qu'un programme de recherche pour les années à venir.

Les annexes sont regroupées dans un second tome. L'annexe A présente une liste thématique complète de mes publications et communications et l'annexe B compose un curriculum vitæ détaillé. L'annexe C regroupe enfin des reproductions des articles ou chapitres d'ouvrage fourni en appui du document de synthèse.

1. MOTIVATIONS

« Les paroles seules comptent. Le reste est bavardage »

(attribué à Eugène Ionesco)

L'identification des langues n'est pas un thème central du traitement de la parole ou de la linguistique. Pourtant, comme nous le verrons dans un premier temps, il se révèle pertinent pour de nombreux enjeux, tant applicatifs que scientifiques (§ 1.1). Par ailleurs, nous expliquerons dans un second temps en quoi ce domaine éclaire une problématique bien plus vaste qui touche à la réflexion même sur la nature du phénomène de parole et qui empruntera aux sciences de la complexité et de l'information (§ 1.2).

1.1. L'IDENTIFICATION DES LANGUES : UNE INTERFACE RICHE ET OUVERTE

L'identification automatique des langues est née au début des années 1970 avec le soutien du *Department of Defense* américain. Parallèlement et à peu près à la même époque, des linguistes se sont intéressés à la faculté humaine à identifier des langues (Gilbert & Ohala, 1981). Ainsi, chercheurs en Traitement Automatique de la Parole (TAP) et linguistes, tout en se focalisant sur des enjeux différents, s'attachaient à déterminer les caractéristiques qui fondent l'unité d'une langue parlée et à définir les frontières existant entre les langues. En TAP, ces premiers travaux jetteront les bases des systèmes actuels, mais le manque de données enregistrées et les performances encore modestes des modèles statistiques ne permettront guère de dépasser le stade exploratoire. En linguistique, les expériences menées permettront de mettre en avant la pertinence à la fois des caractéristiques segmentales et suprasegmentales des langues, sans même évoquer le niveau lexical lorsqu'il est compréhensible aux sujets.

Le schéma dichotomique que j'esquisse ne doit pas masquer que, dès cette époque, certaines connexions entre les deux disciplines sont bien actives, en particulier en phonétique, champ disciplinaire situé historiquement à l'interface du TAP et de la linguistique. À n'en pas douter, les avancées obtenues au cours des années 1980 pour les systèmes de reconnaissance de la parole doivent également à ces phonéticiens, même si approches expertes et statistiques n'ont pas toujours fait bon ménage, probablement pour cause de confusion de leurs rôles respectifs. Jusqu'au début des années 1990, chaque discipline affinera sa connaissance des performances des systèmes (artificiels ou humains) d'identification à défaut de réellement en comprendre les mécanismes : Ces travaux ont abouti, sur le plan de la modélisation, à la conception de plusieurs systèmes informatiques tirant profit de la diffusion libre du corpus OGI MLTS proposant pour la première fois plusieurs heures d'enregistrement de parole dans une dizaine de langues. Durant la même

période, linguistes et psycholinguistes se sont de nouveau intéressés à l'identification des langues *via* des approches inspirées de la typologie linguistique (e.g. Maidment, 1983 ; Hombert & Maddieson, 1998) ou pour étudier des enjeux cognitifs (e.g. Nazzi, Bertoncini & Mehler, 1998)

Par la suite, la deuxième moitié des années 90 marque une stabilisation des performances des systèmes d'IAL après les progrès rapides enregistrés avec la distribution du corpus OGI MLTS. Les campagnes d'évaluation organisées par le NIST en 1995 et 1996 entérine la prééminence d'un état de l'art basé sur un modèle principalement phonotactique, mais les performances plafonnent bientôt. De plus, le système n'est intellectuellement que partiellement satisfaisant, tant il laisse d'indices *a priori* importants et discriminants de côté. C'est dans ce cadre que j'ai débuté mes travaux sur l'IAL fin 1995 à l'Institut de Recherche en Informatique de Toulouse[3]. Le système proposé durant ce travail de doctorat était basé sur la modélisation différenciée des systèmes vocaliques et consonantiques à partir d'un algorithme non supervisé de détection de voyelles. Ce travail avait été inspiré par des considérations typologiques sur les systèmes phonologiques, et il constituait une première ébauche de travail interdisciplinaire. Au fil du temps, j'ai adapté cette démarche, initialement applicative, de manière à me concentrer sur des problématiques à l'interface de l'ingénierie des langues et de la linguistique, puis des sciences cognitives.

Avant d'aller plus loin, on peut mentionner que l'identification des langues et des dialectes est au cœur d'une problématique scientifique plus vaste traitant des notions de variabilité, d'invariance, de catégorisation et d'effet d'échelle. La reconnaissance automatique de la parole (RAP) cherche à identifier les mots prononcés dans une langue donnée, quel qu'en soit le locuteur. La reconnaissance automatique du locuteur (RAL), prise dans un sens large, cherche à identifier la personne ayant parlé, quelle que soit la langue employée et les phrases prononcées. Identifier la langue consiste à prendre une décision en neutralisant à la fois l'effet du contenu lexico-syntaxique et les caractéristiques propres au locuteur. Ainsi, alors que RAP et RAL se focalisent respectivement sur le contenu linguistique d'un énoncé ou sur les caractéristiques spécifiques à la personne ayant produit ce dernier, l'IAL cherche à identifier l'identité de la langue employée, ou plus exactement dans quelle mesure l'énoncé produit est compatible avec les modèles établis pour différentes langues. L'identification des dialectes franchit quant à elle une étape supplémentaire en nécessitant de nuancer la représentation de la structure de la langue par une dose importante de variabilité intermédiaire entre langue et idiolecte individuel. Il me semble intéressant de noter que la RAL s'intéresse à des individus par définition discrets et donc pour lesquels le problème est bien posé : *in fine*, un énoncé *est* ou *n'est pas* produit par une personne donnée. La difficulté ne vient donc pas de la définition du problème mais de l'écho lacunaire et finalement peu caractéristique d'une personne que représente un enregistrement audio. À l'inverse, en

[3] Cette thèse se déroula dans le cadre du projet DGA/DRET n°95/118 *Discrimination multilingue automatique* mené en partenariat avec le laboratoire Dynamique Du Langage (Lyon), l'Institut de la Communication Parlée (Grenoble) et l'Institut de Linguistique et de Phonétique Générale et Appliquée (Paris).

identification des langues et des dialectes, on se heurte à la sempiternelle question de l'identité de la langue ; celle-ci n'est pas une et indivisible, mais elle représente un espace de conventions et de variations caractérisant une communauté de locuteurs. Par ailleurs, la notion de frontière entre langues est particulièrement ambiguë et à « géométrie variable », en fonction des experts consultés et des parlers considérés. En cela, le problème de l'IAL est mal posé, et la notion même de frontières est floue et fortement dépendante du niveau d'analyse considéré.

Au cours de ces années, ma réflexion sur l'IAL a évolué, en particulier par la prise de conscience des enjeux très vastes liés à ce domaine. Partant de considérations initiales applicatives (décrites au paragraphe 1.1.1), je me suis ensuite interrogé sur certains enjeux linguistiques des travaux que nous entreprenions (paragraphe 1.1.2) puis assez naturellement, la prise en compte de la perception humaine m'a amené à considérer que l'identification des langues offrait un angle original pour contribuer à la recherche sur de vastes enjeux cognitifs qui seront développés au paragraphe 1.2.

1.1.1. Enjeux applicatifs

C'est devenu un lieu commun de dire que l'époque actuelle est une ère de communication multilingue, que ce soit entre êtres humains (au sein des grandes mégalopoles ou par téléphone interposé), ou entre humains et machines (domaine des Interfaces Homme-Machine ou IHM). Ce constat implique le développement d'applications capables de gérer plusieurs langues et/ou d'identifier une langue parmi d'autres, qu'il s'agisse d'une tâche d'assistance au dialogue humain (DH), d'IHM ou encore de traitement automatique (indexation) de grandes masses de données audio ou audiovisuelles. La plupart de ces applications peuvent par ailleurs se décliner en termes d'enjeux stratégiques, liés à l'activité militaire d'un pays et il s'agit là de l'une des principales motivations des bailleurs de fonds de ce domaine, depuis son origine. À l'inverse, on envisage depuis peu l'usage de l'IAL dans le cadre de l'apprentissage des langues étrangères, offrant ainsi un nouveau cadre applicatif prometteur.

✦ *Assistance au dialogue humain*

L'exemple le plus célèbre de situation de DH multilingue, cité dans Muthusamy, Barnard & Cole (1994), est assez significatif : aux États-Unis, les numéros des services d'urgence sont centralisés et accessibles en appelant le 911. La nécessité de pouvoir traiter des appels en plusieurs langues est depuis longtemps une réalité dans ce pays multiethnique, à tel point qu'il est fait appel à un service d'interprètes humains chargés de traiter les appels arrivants[4]. Lorsque l'appelant

[4] Les sociétés *TeleInterpreters* ou *Language Line Services* sont parmi les prestataires actuels. Ils indiquent sur leurs sites Internet respectifs la possibilité d'être en contact en quelques secondes avec un interprète, et ce pour plus de 170 langues (sources, consultées le 03/10/2007 :

http://www.teleinterpreters.com/government_services.aspx et http://www.languageline.com/page/industry_government/).

n'indique pas le nom de sa langue, l'aiguillage des appels vers l'interprète adéquat repose sur l'identification de celle-ci et il demeure manuel : le standardiste aiguille l'appelant vers une personne spécialement entraînée à l'identification des langues qui cherche alors à reconnaître la langue en question avant de rediriger la personne vers un interprète compétent. Ainsi l'appel peut transiter par plusieurs interlocuteurs pendant un temps assez long, comme le rapportait Muthusamy dès 1994, alors même que l'urgence de la situation exigerait la plus grande rapidité. L'objectif d'un système d'IAL serait idéalement d'identifier la langue parlée (si elle est connue du système) ou éventuellement de fournir une liste de langues possibles. Cette seconde solution, qui pourrait se baser sur l'identification de traits caractéristiques, soit de certaines familles de langues, soit de zones géographiques, n'a, à notre connaissance, pas été explorée.

✦ Interfaces Homme-Machine

L'IAL a un rôle croissant à jouer au sein des IHM pour permettre leur utilisation dans des pays plurilingues ou dans un cadre international. Si l'on prend l'exemple de l'Espagne, un système de dictée vocale ou de réservation de billets de train par téléphone doit au moins identifier et gérer le castillan, le catalan, et le basque. Ces contraintes linguistiques sont courantes à travers le monde, même si elles nous sont peu familières en France où une seule langue officielle est reconnue[5]. On retrouve également ce type de contraintes pour d'autres applications comme, par exemple, les bornes interactives de renseignement. Dans ce cadre précis, des systèmes opérationnels reconnaissant une dizaine de langues permettraient déjà de répondre à une part importante des demandes, même si certains pays (l'Inde en particulier, avec près d'une vingtaine de langues officielles) ou certaines ères supranationales (par exemple l'union européenne et ses 23 langues officielles) dépassent ce cadre-là.

✦ Indexation multilingue

Le thème de l'indexation automatique de documents multimédia ou sonores est désormais au cœur des enjeux en ingénierie de la parole et de l'image. Tout comme en reconnaissance de la parole, cette problématique est amenée à se développer en IAL pour répondre à une demande croissante. L'objectif poursuivi peut être de détecter les changements de langues dans une bande sonore ou encore de réaliser l'identification des idiomes présents : on peut envisager le filtrage en temps réel des émissions diffusées par satellite ou Internet pour présélectionner les canaux diffusant des émissions dans une langue cible. Qu'il s'agisse d'un traitement à la source visant à la génération de métadonnées additionnelles sur le programme en cours, ou d'un service offert à l'usager au niveau de la réception, on peut s'attendre à voir ce type de systèmes se développer.

---

[5] Alors même que la Délégation Générale à la Langue française et aux Langues de France dénombre 86 langues dites « de France ».

Enjeux stratégiques

Le domaine militaire et policier constitue un autre domaine d'application particulièrement sensible. Le *US Department of Defense* est à l'origine des premières recherches menées en IAL au sein des laboratoires *Texas Instruments* au début des années 70. Depuis, de nombreux autres projets ont vu le jour, aux Etats-Unis comme en Europe. Cet intérêt est pleinement motivé par l'éventail des applications entrevues par les instances de sécurité, que ce soit dans le cadre de la communication ou du renseignement militaire.

Durant la guerre du Kosovo en 1999, plusieurs antennes médicales mobiles disposaient d'un système d'assistance médicale multilingue rudimentaire permettant de diagnostiquer rapidement les troubles dont souffraient les patients, aussi bien d'origine serbe qu'albanaise (Hunt *et al.*, 1999). Par contre, l'identification de la langue parlée par le patient était réalisée par un interprète humain, ce qui n'est réellement possible qu'à faible échelle (peu de langues). Ce type d'usage des technologies de la parole sur des terrains opérationnels (qu'il s'agisse d'ailleurs d'interventions militaires ou humanitaires) ira probablement en se généralisant, nécessitant par là même la mise au point de technologies adaptables rapidement à de nouvelles langues ou dialectes.

Le renseignement militaire est, pour sa part, demandeur de systèmes qui recouvrent l'ensemble des enjeux cités précédemment. De la détection de langues cibles sur des canaux de communication à l'identification de la langue et du dialecte maternels d'une personne, les défis sont nombreux et les enjeux, en particulier éthiques, s'apparentent à ceux de l'utilisation criminalistique de la reconnaissance du locuteur et de ses dérives. Outre le fait que les performances opérationnelles requises dépassent celles des systèmes actuels, il s'agit là d'un domaine dans lequel science et société se doivent de dialoguer.

Enseignement des langues étrangères

Lorsque l'on s'exprime dans une langue étrangère, on conserve souvent un accent qui n'est pas le fruit du hasard. Il emprunte beaucoup aux structures linguistiques de sa langue maternelle, du choix et de la réalisation phonétique des phonèmes jusqu'aux constructions syntaxiques, en passant par toutes les dimensions suprasegmentales de rythme, accentuation et intonation. Cette production non native peut donc être évaluée selon chacune de ces dimensions en termes de distance à la langue maternelle L1 et à la langue seconde L2. Ainsi, si l'on dispose de modèles phonétiques, phonotactiques, rythmiques, intonatifs de ces langues L1 et L2, on peut évaluer la production d'un apprenant, identifier ses forces et ses faiblesses, et proposer des exercices en conséquence, tout en assurant un suivi des progrès réalisés. Pour qu'un tel système soit efficace, il est nécessaire que les distances (statistiques ou pas) qu'il calcule aient une réalité perceptive, c'est-à-dire que distance perceptive et distance paramétrique calculée soient corrélées. Ce type d'application ne relève pas à proprement parler de l'IAL, mais il peut exploiter de tels outils et rejoint les préoccupations des linguistes sur un certain nombre d'enjeux, en particulier liés à la notion de distance perceptive.

Même s'il existe à ce jour des logiciels d'apprentissage assisté par ordinateur calculant des distances (en particulier intonatives) entre une cible et la production d'un apprenant (parfois disponibles sur des consoles de jeux vidéo portables), nous n'avons pas connaissance de l'existence de systèmes plus élaborés issus du domaine de l'IAL. Ce domaine reste donc largement en friche.

1.1.2. Enjeux linguistiques

L'objectif des systèmes applicatifs d'IAL est d'atteindre les performances les plus élevées, parfois au détriment de l'analyse fine du comportement des algorithmes employés. Les linguistes intéressés par l'identification des langues se focalisent plutôt sur l'évaluation de la robustesse des indices potentiels, le type d'erreurs produites par ces systèmes, les relations graduelles reflétées par les distances estimées entre les langues ou encore la corrélation entre ces distances, dites automatiques, et les distances perçues par des sujets humains. À cela s'ajoute à notre avis un domaine encore sous-exploité, à savoir l'utilisation des approches automatiques pour mener des recherches typologiques. Le constat exposé dans cette section sera malheureusement que cette interface entre IAL et linguistique est fortement sous-exploitée, en France et ailleurs.

✦ *La robustesse des indices*

Pour être robuste au sein d'un système d'IAL, un indice doit conjuguer plusieurs qualités : il doit tout d'abord être statistiquement assez fréquent car un modèle automatique ne sera performant que sur les phénomènes qu'il peut correctement estimer. L'estimation proprement dite consiste généralement à caractériser l'indice par sa valeur moyenne et sa variabilité dans un espace paramétrique multidimensionnel, quitte à faire appel à des sous-entités si le phénomène est complexe. La modélisation d'une unité phonétique par un modèle de mélange de lois gaussiennes, par exemple, entre dans ce cadre. Deuxièmement, cet indice doit être discriminant, et donc présenter des caractéristiques (moyenne et variabilité) différentes d'une langue à l'autre. Troisièmement, il est nécessaire que la variabilité intra-langue soit moins importante que la variabilité interlangue. Dans le cas contraire, l'indice n'est pas adapté à la tâche visée.

Au-delà de l'efficacité même des modèles automatiques, cette évaluation de la robustesse des différents indices potentiels est du plus grand intérêt linguistique. En effet, beaucoup des travaux entrepris ces dix dernières années à l'interface entre phonétique et phonologie portent sur l'étude de la variabilité (intra-individuelle, inter-individuelle, inter-dialectale, etc.) et sur son revers, l'invariance. Malheureusement, l'analyse acoustique fine requise pour ce genre d'étude (en termes d'analyse formantique, d'étiquetage et d'alignement phonétique, par exemple) limite fortement la taille des corpus considérés, ou nécessite un travail

considérable[6]. Une autre approche consisterait alors à exploiter des modèles automatiques pour tester des hypothèses linguistiques. À titre d'exemple, on pourrait exploiter un système automatique basé sur une reconnaissance phonétique, extraire certains segments d'intérêt, par exemple les voyelles dites focales dans la théorie quantique (Stevens, 1972 ; 1989), et tester individuellement leur apport à l'identification des langues en parallèle d'une étude de leur variabilité à différentes échelles (locuteur, dialecte, langue). On peut également envisager de procéder, à partir des données paramétrées, à des regroupements non supervisés (approches de *clustering*) à la recherche des indices les plus spécifiques à chaque langue ou à l'inverse les plus translinguistiques (e.g. Boula de Mareüil, Corredor-Ardoy & Adda-Decker, 1999). En fait, la liste des retombées linguistiques que l'on peut attendre d'une exploitation des modèles d'IAL est longue (on peut encore imaginer pour différentes classes phonétiques, d'évaluer la pertinence en identification des trames les plus stables ou les plus transitoires, pour alimenter le débat crucial sur la nature statique ou dynamique des cibles phonologiques (e.g. Carré, Divenyi & Pellegrino, 2007 ; Lindblom, Mauk & Moon, 2006).

✦ La notion de distance linguistique

Si le paragraphe précédent s'aventurait dans une exploitation à une échelle fine des modèles d'IAL, une exploitation de leurs résultats à une échelle plus globale semble également porteuse de connaissance. Transformer les résultats d'un système automatique (qu'il s'agisse de matrices de confusion entre langues ou des listes des vraisemblances des modèles de langues eux-mêmes) en matrice de proximité peut en effet nous amener à interroger la notion de distance linguistique. Comment, en effet, définir ce concept ? La réponse est complexe et fortement multidimensionnelle. Qu'il s'agisse de la lexicostatistique, basée sur un codage en cognats ou de distances acoustiques statistiques, du type de la divergence de Kullback-Leiber entre des modèles multigaussiens, l'éventail des possibilités est large. On peut cependant estimer que toute représentation multidimensionnelle d'une matrice de proximité entre langues est porteuse d'information. Malheureusement, la conception même des systèmes d'IAL (faisant appel à plusieurs types d'information), ne donne pas nécessairement les clefs permettant l'interprétation des résultats.

À titre d'exemple, la Figure 1 représente sous forme de dendrogramme la proximité entre les langues de l'évaluation NIST LRE 2003, dérivée de la matrice de confusion du système d'IAL du MIT (graphique établi d'après Martin & Przybocki, 2003).

[6] Voir par exemple les analyses discriminantes menées en discrimination dialectale par Jalaleddin Al-Tamimi dans sa thèse sur plusieurs milliers de voyelles, afin de tester la pertinence d'indices statiques et dynamiques en termes d'invariance et de variabilité (Al-Tamimi, 2007).

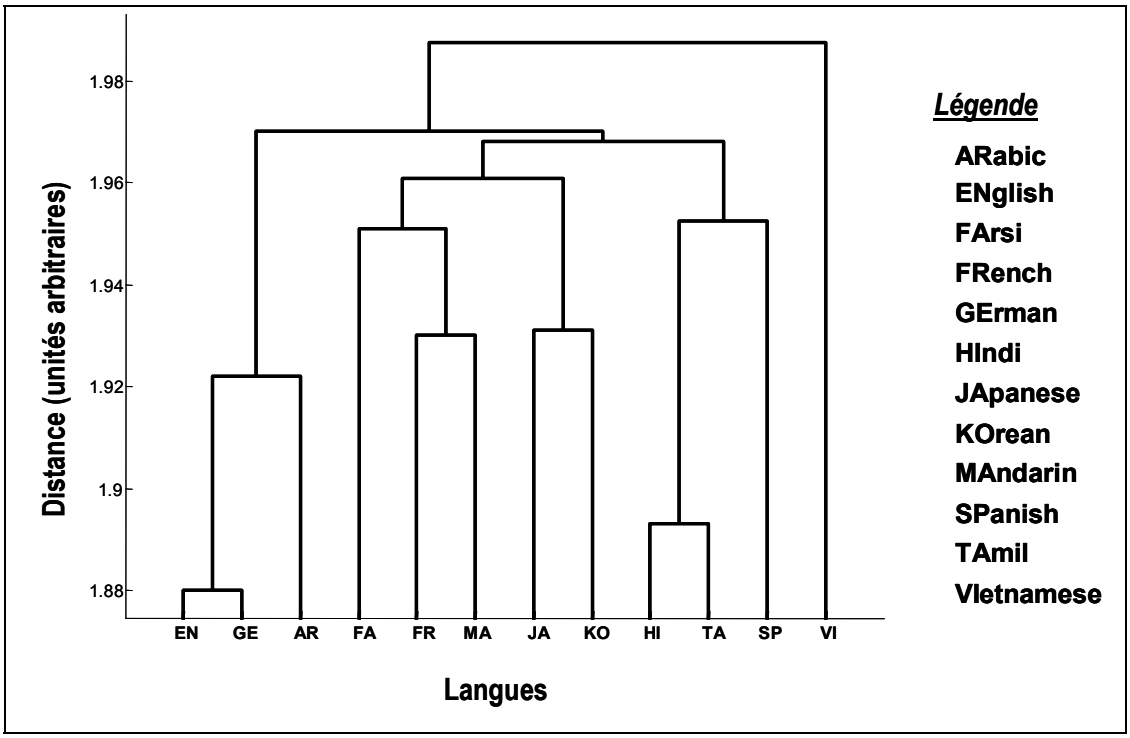


Figure 1 – Dendrogramme dérivé de la matrice de confusion obtenue par le meilleur système du MIT lors de l'évaluation NIST LRE 2003.

Ce système exploitait conjointement une modélisation acoustique, phonétique et phonotactique et l'interprétation des résultats ne peut donc être que spéculative. En particulier, les modèles phonotactiques relèvent à la fois du domaine phonético-phonologiques (structures syllabiques, etc.) et du domaine morpholexical (les mots ou les morphèmes les plus courants étant constitués de chaînes phonétiques brèves). Par conséquent, la proximité entre anglais et allemand d'une part, et hindi et tamil d'autre part peut tout autant refléter une distance phonologique faible qu'une distance lexicale faible, voire un mélange des deux. À cela s'ajoute également que, dès lors qu'une dimension temporelle est intégrée dans les modèles, ceux-ci peuvent intrinsèquement prendre en compte des aspects rythmiques. Enfin, une langue fortement distante des autres dans ce type de représentation peut être tout aussi intéressante à analyser. Dans la figure, on constate en effet que le vietnamien est la langue la mieux individualisée[7] et ce, sans que la raison en soit particulièrement évidente : il s'agit d'une langue à ton (comme le mandarin, également présent) dont le lexique a significativement été emprunté aux parlers chinois. Ainsi, aucune raison évidente ne vient expliquer cette individualisation robuste, en particulier vis-à-vis du mandarin.

En résumé, l'utilisation conjointe par des modélisateurs et des linguistes des modèles employés en IAL est, à notre avis, une source extraordinaire d'information sur les distances entre langues, pour l'heure malheureusement quasi-inexploitée.

[7] Ce qui semble être une tendance partagée par la plupart des systèmes d'IAL de l'évaluation NIST LRE 2003, si l'on en croit les courbes DET établies pour les différentes langues et les taux d'égale erreur atteints (Martin & Przybocki, 2003).

✦ La notion de distance perceptuelle et ses corrélats acoustiques

Historiquement, les linguistes se sont rapidement intéressés à l'identification perceptuelle des langues par des sujets humains, les études sur les primates non humains et d'autres mammifères venant chronologiquement bien après, lorsque l'on cherchera à caractériser les capacités neurocognitives humaines.

Dès les années 1970, ces expériences se sont révélées riches d'information sur la perception de la parole, en particulier sur la saillance perceptuelle de certains indices segmentaux ou suprasegmentaux. Par la suite, les études entreprises sur des nourrissons et des enfants ont également été fructueuses, tant en perception que dans le cadre plus cognitif de l'acquisition du langage chez l'humain. De même, les travaux abordant l'influence de la langue maternelle des sujets sur leur capacité à discriminer des langues ou à juger de leur similarité nous renseignent-ils sur les représentations et les processus de catégorisations phonético-phonologiques.

Il nous semble enfin que la mise en perspective des résultats des modèles automatiques (pour lesquels on pourrait maîtriser les indices exploités) et par des sujets humains dans des tâches similaires d'identification des langues serait réellement éclairante sur la saillance des différents types d'informations présents dans le signal et sur leur exploitation par le modèle et par l'humain. Pour être pertinente, une telle étude devrait cependant évaluer la quantité d'information utilisée et mémorisée par les systèmes automatiques et humains, ce qui semble particulièrement difficile.

✦ Vers une utilisation typologique des modèles ?

La notion de typologie automatique est quasiment inexistante aujourd'hui en linguistique. Elle trouve cependant une implémentation dans les travaux entrepris au sein du projet AUTOTYP (Bickel & Nichols, 2002) sous la forme d'une approche ascendante (*bottom-up*) de la typologie des langues. Le but est d'atteindre une telle typologie des phénomènes linguistiques à partir des descriptions les plus objectives possibles réalisées sur un échantillon de plusieurs dizaines de langues. Cette démarche vise à minimiser le biais dû aux *a priori* théoriques et à faire émerger des catégories consensuelles.

Les modèles automatiques utilisés en IAL n'ont pas la prétention de ne pas avoir de biais. En particulier, il est évident que certains indices sont moins bien estimés que d'autres, voire totalement hors de portée des systèmes. Par contre, ils offrent une qualité essentielle pour comparer des langues : la consistance du traitement réalisé. En effet, les principes des pré-traitements sont généralement établis de manière indépendante des langues considérées et appliqués uniformément avant l'injection du résultat de leur paramétrisation dans des modèles spécifiques aux langues. Ainsi, si le biais existe, il est systématique et son impact est donc estimable sur le processus d'identification. Il est cependant juste de tempérer cet optimisme : en effet, il serait regrettable que les prétraitements introduisent un biais qui occulte totalement un phénomène caractéristique d'une langue (imaginons par exemple un modèle phonétique neutralisant les variations des durées vocaliques,

alors même que l'une des langues exploite une opposition phonologique desdites durées). On ne saurait donc trop insister sur le fait que les linguistes, phonéticiens ou phonologues, peuvent judicieusement attirer l'attention des chercheurs en IAL sur la nature de certains phénomènes pertinents.

L'un des enjeux majeurs de l'utilisation des modèles d'IAL en typologie relève de la validation des classes typologiques et de la classification des langues. Les débats sur la typologie rythmique fournissent un bon exemple, puisque on peut s'intéresser à la congruence entre les prédictions faites par les linguistes et les résultats obtenus par des modèles automatiques. Cette problématique sera abordée en particulier au paragraphe 2.3.2.

Dans l'introduction (page 13), nous mentionnions la dichotomie existant entre la description acoustico-phonétique (espace cepstral) de l'espace vocalique d'une langue et sa description phonologique. La comparaison entre des distances évaluées à partir de l'une et l'autre de ces approches permettrait éventuellement de mieux étudier cette interface : si deux langues, décrites phonologiquement par le même système vocalique, s'avèrent mettre en œuvre des systèmes assez distants au niveau acoustico-phonologique, il y a là matière à une relecture des typologies qui peut être fructueuse.

1.2. UN AUTRE REGARD SUR LA COMMUNICATION PARLÉE

1.2.1. Une fenêtre sur la cognition

Lorsque l'on s'intéresse à la perception humaine, on constate rapidement que les processus complexes mis en jeu ne sont que partiellement identifiés. La perception catégorielle est-elle un processus auditif, phonétique ou phonologique ? Quelles sont les bases biologiques de l'effet d'aimant perceptuel décrit chez l'homme et chez d'autres mammifères ? Comment les catégories phonologiques émergent-elles chez l'individu ? Existe-t-il réellement des *représentations* phonologiques catégorielles, ou les effets observés sont-ils induits par des *tâches* catégorielles, comme le propose Massaro (2001) ? Dans le même ordre d'idée, pourquoi n'aurions-nous pas des représentations continues menant, lorsque la tâche l'exige, à des décisions catégorielles par un processus stochastique ou flou ? *Qu'entend-on* réellement dans un signal de parole : densités spectrales et formants ou leurs variations, etc. ? En quoi le signal de parole est-il différent d'un autre stimulus acoustique ?

Toutes ces questions dépassent largement le cadre de l'identification des langues. Cependant, on peut noter que les processus mis en œuvre lors de l'écoute de langues étrangères (comprises ou pas) ouvrent une fenêtre extraordinaire sur ces mécanismes de perception et doivent être menés en parallèle des recherches multilingues menées avec des locuteurs natifs.

En parallèle de ces aspects perceptifs, il est également intéressant de s'interroger sur la tâche cognitive même d'identification des langues. Comme le rapporte l'UNESCO (2003, p. 13) :

> *« En général, toutefois, les contextes bilingues et multilingues, c'est-à-dire la présence de différences linguistiques au sein du même pays, sont plutôt la norme que l'exception à travers le monde, tant au Nord qu'au Sud. Dans ce contexte, le bilinguisme et le multilinguisme, c'est-à-dire l'emploi de plus d'une langue dans la vie quotidienne, représentent la pratique normale ».*

Ainsi, l'identification de deux langues (ou plus) et le passage plus ou moins régulier de l'une à l'autre est une tâche cognitive automatique et fréquente pour la majorité des êtres humains. Cette situation est bien évidemment l'objet des travaux linguistiques, sociolinguistiques et psycholinguistiques sur le multilinguisme. Si l'on se penche sur la mise en place du bilinguisme chez le bébé au cours des premiers mois, on se rend rapidement compte que l'identification de la langue à partir du signal acoustique revêt une importance critique pour l'acquisition. Là encore, l'identification des indices saillants présents dans le signal peut améliorer notre connaissance de la perception de la parole et, au-delà, de la construction des représentations du langage chez l'enfant.

En résumé, le cadre de l'identification des langues a nourri notre réflexion en fournissant un angle d'éclairage novateur et relativement peu exploré sur des problématiques touchant au fondement même des notions de langage et de parole. Nos recherches entreprises depuis dix ans sont évidemment parcellaires et suivent des directions parfois divergentes, elles restent cependant imprégnées de notre culture initiale en traitement du signal. En effet, elles partagent toutes comme fil conducteur le questionnement sur ce qu'est l'information langagière et sur où elle se trouve dans le signal de parole.

1.2.2. Nature et structure de l'information dans la parole

Nous n'avons absolument pas la prétention de couvrir dans nos recherches toutes les acceptations auxquelles renvoie le terme d'information langagière. En particulier, les aspects liés à la focalisation et pour lesquels il est souvent fait référence à l'idée d'organisation de l'information sont au-delà de nos compétences. Notre propos se limite modestement au niveau sub-morphémique, de la description la plus acoustique d'un signal de parole (principalement dans sa dimension temporelle) jusqu'à l'information véhiculée par le jeu des oppositions phonologiques dans une langue donnée. Sauf mention contraire, il est fait référence dans la suite du document à ce niveau d'information là. Ce questionnement sur nature, structure et distribution de cette information est donc à la base des recherches rapportées en partie 3 de ce manuscrit et il dépasse le cadre de l'identification des langues *stricto sensu*. Nous l'avons abordé en suivant les trois pistes suivantes.

✦ Primitives informationnelles

L'identification des primitives phonético-phonologiques est un enjeu majeur : s'agit-il de traits, de gestes, d'atomes, de phonèmes ou encore de syllabes ? Quelle échelle temporelle est la plus adaptée ? Nous avons décliné ce thème selon deux directions. D'une part, dans le cadre des travaux pilotés par Fanny Meunier sur la compréhension de la parole dégradée, nous avons commencé à explorer la frontière

perméable existant entre les sons du langage et les autres, par le biais de stimuli inversés temporellement. Dans certaines conditions, de tels stimuli garantissent une compréhension parfaite aux auditeurs alors même que leur organisation temporelle est inversée. Par conséquent, il est difficile de considérer que ces stimuli sont totalement à exclure de la sphère langagière, et même dans des conditions où la compréhension est affectée, ils peuvent nous informer grandement sur des processus neurocognitifs à l'œuvre dans le traitement de l'information de la parole. D'autre part, nous avons participé, quoique de façon plus marginale, aux travaux menés par René Carré sur l'identification des primitives phonologiques et en particulier sur leur nature statique ou dynamique.

✦ Primitives et structuration des systèmes phonologiques

Parallèlement à ces recherches menées à partir de l'analyse de signaux de parole réels et des performances de sujets humains en production et perception de la parole, nous avons développé des analyses adaptées à l'utilisation d'inventaires phonologiques (en l'occurrence la base de données UPSID) avec pour objectif de mettre en évidence des régularités structurelles pertinentes pour l'identification des contraintes à l'œuvre dans la structuration et l'évolution des systèmes phonologiques. Cette méthodologie empruntant beaucoup au domaine protéiforme des sciences de la complexité, notre attention a été attirée sur le rôle éventuel de la complexité à l'interface phonético-phonologique. En particulier, ce paradigme offre potentiellement une ouverture pour relier des processus cognitifs (trajectoires individuelles ou spécifiques aux langues en acquisition, compréhension de la variabilité en compréhension de la parole, etc.) aux structures issues de la description des langues. Cette intuition nous a amené à proposer des mesures capables d'évaluer l'étendue de la complexité rencontrée dans UPSID et si sa distribution se révélait pertinente pour la typologie.

✦ Distribution de l'information dans la parole

Le troisième axe est directement hérité des travaux que nous avons menés en IAL. Le rythme, utilisable à la fois par les nourrissons pour distinguer des groupes de langues et par des systèmes automatiques pour différencier des langues elles-mêmes (voir paragraphe 2.3.2) se situe à l'interface du niveau segmental et suprasegmental dans la parole : il est bien évidemment pertinent au niveau suprasegmental et c'est bien là qu'il est généralement classé (Fox, 2000). Il relève également du niveau segmental puisque, contrairement à ce que l'hypothèse de l'isochronie avait prédit, il est influencé par le contenu phonétique des énoncés. Il s'agit en somme d'un phénomène difficile à cerner et complexe à décrire. Par ailleurs, son origine et son rôle mêmes donnent lieu à débat.

C'est dans ce contexte que nous avons étudié la variabilité du rythme au sein des langues ou des parlers, arabes et anglais des îles britanniques et que la notion de débit de parole a attiré notre attention. D'une part, le débit varie au cours des énoncés et d'autre part, son domaine de variation ne semble pas être spécifique à un locuteur, mais bel et bien dépendant de la langue parlée. Ces éléments nous ont

conduit à considérer que le débit fait partie intégrante du rythme des langues, même si évidemment des degrés de liberté sont donnés aux locuteurs. Ce constat suggérait que le débit de parole était un facteur pertinent pour l'évaluation du débit d'information et de ses variations dans les langues du monde. De manière très schématique et dans un cadre théorique où *toutes choses sont égales par ailleurs*, une langue où le débit (phonémique ou syllabique) est 30 % plus lent que dans une autre transmettra nécessairement moins d'informations phonético-phonologiques par seconde. Plus précisément, nous explorerons l'hypothèse que le débit (qui relève d'une dimension syntagmatique) soit en interaction avec le système phonético-phonologique (dans sa dimension paradigmatique en particulier) et que par conséquent, on ne puisse pas réellement rencontrer les deux langues évoquées dans l'exemple ci-dessus, ayant des débits très différents et des systèmes phonologiques identiques. Cette interrogation sur cette éventuelle interaction s'articulera autour des relations entretenues par les notions d'information et de complexité en phonologie depuis près d'un siècle.

2. IDENTIFICATION DES LANGUES

« L'imagination est plus importante que la connaissance. La connaissance est limitée alors que l'imagination englobe le monde entier, stimule le progrès, suscite l'évolution. »

(Albert Einstein)

Cette partie présente les travaux menés sur l'identification des langues en exposant tout d'abord un état de l'art de l'identification *automatique* des langues (§ 2.1). Cet exposé préliminaire permet d'introduire la démarche scientifique que nous avons adoptée de manière à prendre en considération les enjeux ainsi exposés (§ 2.2). La suite de cette partie présentera et discutera alors les travaux entrepris en identification automatique (§ 2.3) puis sur le thème de l'identification perceptuelle (§ 2.4), et enfin par le biais d'une approche intégrant ces deux dimensions sur des données issues tour à tour des aires dialectales de l'arabe et de l'anglais des îles britanniques (§ 2.5).

2.1. ÉTAT DE L'ART

En 2003, la campagne d'évaluation des systèmes d'identification automatique des langues[8] organisée par le *National Institute for Standards and Technology* (NIST) américain visait à mesurer le chemin parcouru depuis la précédente évaluation, menée en 1996. Cette campagne a pleinement joué son rôle, en permettant d'évaluer les progrès effectués par certaines méthodes déjà éprouvées et de mettre en évidence l'apport de nouvelles techniques, en particulier en fusion de modèles. En fait, peu d'équipes de recherche ont participé à la campagne de 2003, mais elle a eu l'impact recherché en étant à l'origine d'un regain d'intérêt très net pour la thématique de l'IAL. Depuis, deux autres campagnes se sont déroulées (en 2005 et 2007), la dernière rassemblant 21 équipes. Nous allons maintenant donner quelques éléments de lecture de la campagne de 2003, à laquelle nous avons participé, en complétant ponctuellement par des informations issues des campagnes ultérieures.

2.1.1. La campagne NIST LRE 2003 : un aperçu

La campagne d'évaluation lancée par le NIST début 2003 s'est tenue selon un calendrier serré puisqu'elle a été annoncée fin 2002 et qu'elle s'est déroulée début 2003. Ces délais particulièrement courts peuvent expliquer le très faible nombre de réponses. En effet, seuls 6 sites (4 américains, 1 australien et 1 européen) ont participé à la campagne, à savoir les :

[8] Cette campagne est connue sous l'acronyme NIST LRE 03 (*Language Recognition Evaluation*).

- *Lincoln Laboratory*, *Massachusetts Institute of Technology*, USA ;
- *Center for Spoken Language Understanding (CSLU)*, *Oregon Graduate Institute*, USA ;
- *Department of Electrical Engineering*, *University of Washington*, USA ;
- R523, *Department of Defense*, USA ;
- *Speech Research Lab*, *Queensland University of Technology*, Australia ;
- Institut de Recherche en Informatique de Toulouse en association avec le laboratoire Dynamique Du Langage, France.

Parmi ces unités, on retrouve deux des laboratoires phares du domaine de l'identification des langues (*Lincoln Lab-MIT* et CSLU-OGI) qui avaient déjà participé à la campagne d'évaluation LID'96, ainsi que plusieurs unités familières des campagnes d'évaluation du NIST relatives à l'identification du locuteur (R523 et SRL). Seules l'université de Washington et le binôme IRIT/DDL participaient là à leur première campagne d'évaluation NIST.

Le protocole d'évaluation était similaire à celui employé en vérification du locuteur puisqu'il s'agissait de vérifier une hypothèse d'identité pour une langue inconnue. Chaque item de test consistait donc en un extrait sonore prononcé dans une langue inconnue accompagné d'une langue hypothèse.

12 langues[9] étaient susceptibles de constituer les hypothèses, et les enregistrements provenaient de 13 langues (les douze mêmes plus une langue intruse – le russe). Les enregistrements, à la fois pour l'apprentissage et le test, étaient très majoritairement extraits de corpus conversationnels téléphoniques (corpus *Callfriend*, encodage 8-*bit*, *mu-law*) auxquels se sont ajoutés quelques données de test provenant d'autres sources (téléphone cellulaire en particulier) de manière à évaluer l'impact du canal d'enregistrement. Les enregistrements étaient structurés en trois sous-ensembles de durées nominales respectives 3, 10 et 30 secondes.

Les systèmes étaient évalués selon une métrique classique de détection prenant en compte de manière équipondérée les probabilités de fausse alarme et de faux rejet, estimées à partir des 1280 enregistrements de test fournis dans chaque condition de durée et tous confrontés aux 12 langues hypothèses[10]. Les performances sont donc exprimées en taux d'égal erreur (ou *Equal Error Rate* – EER) et sous forme de courbes DET (*Detection Error Tradeoff*).

[9] Arabe (égyptien), anglais (américain), farsi, français (canadien), allemand, hindi, japonais, coréen, mandarin, espagnol (d'Amérique Latine), tamil et vietnamien.

[10] Soit un total de 1280 x 3 x 12 = 46 080 tests.

2.1.2. Méthode

La plupart des sites[11] ont mis en pratique un système similaire à celui qui avait obtenu les meilleurs résultats lors de l'évaluation de 1996. Cette architecture, nommée PPRLM (pour *Parallel Phone Recognition followed by Language Modeling*) a été popularisée dans les années 1990 par Zissman (1996). Son principe est présenté sur la Figure 2. Il s'organise séquentiellement autour de trois modules :

- *le module phonétique*

Composé de p décodeurs acoustico-phonétiques (généralement p égale 5 ou 6 décodeurs dépendants chacun d'une langue), ce module a pour objectif de fournir au module suivant plusieurs séquences d'unités phonémiques (prises dans les espaces phonétiques respectifs des p décodeurs) à partir d'une paramétrisation acoustique et d'un système de reconnaissance de type Modèles de Markov Cachés. Il se comporte donc comme un opérateur de projection d'un espace acoustique continu dans p espaces phonémiques discrets. Dans l'exemple illustré Figure 2, quel que soit le nombre de langues à identifier, elles sont toutes projetées dans les espaces phonétiques A, B et C.

- *le module phonotactique*

Constitué de grammaires probabilistes de type n-grammes ($n \leq 5$, plus généralement $n = 3$), ce modèle permet d'évaluer la vraisemblance de chacune des langues cibles en fonction des séquences d'unités fournies par le module précédent. De manière classique, plus la quantité de données disponibles pour l'apprentissage est importante, meilleurs sont les modèles avec n grand, du point de vue de la qualité d'estimation et donc de discrimination.

- *le module de décision*

La présence de p flux parallèles pour chaque langue implique de procéder à une recombinaison des scores de vraisemblance. La décision de valider ou de rejeter l'hypothèse à tester est ensuite prise en appliquant un seuil sur le rapport de vraisemblance entre le modèle de la langue hypothèse et un modèle dit « du monde ». Auparavant, il est nécessaire d'opérer une fusion des vraisemblances issues des différentes voies parallèles pour chaque langue. Toujours dans l'exemple de la Figure 2, cela signifie que le score calculé pour chacune des langues numérotées de 1 à 4 doit être évalué à partir des vraisemblances issues de chacun des trois décodeurs : A1, B1 et C1 pour la langue 1 ; A2, B2 et C2 pour la langue 2 ; etc. Historiquement, la première méthode utilisée a été une simple addition, pondérée ou non, des log-vraisemblances pour chacune des langues, répétant ainsi l'une des entorses habituelles en traitement automatique de la parole puisque

[11] Le système que nous avions utilisé était d'ailleurs de ce type ; par manque de temps de développement, nous avions renoncé à utiliser des modèles plus originaux. Le module phonétique avait été réalisé à l'IRIT, tandis que les modèles phonotactiques étaient implémentés à DDL, tout comme la fusion.

l'indépendance statistique des processus aléatoires donnant A*i*, B*i* et C*i* (*i* étant le numéro de la langue) est loin d'être garantie.

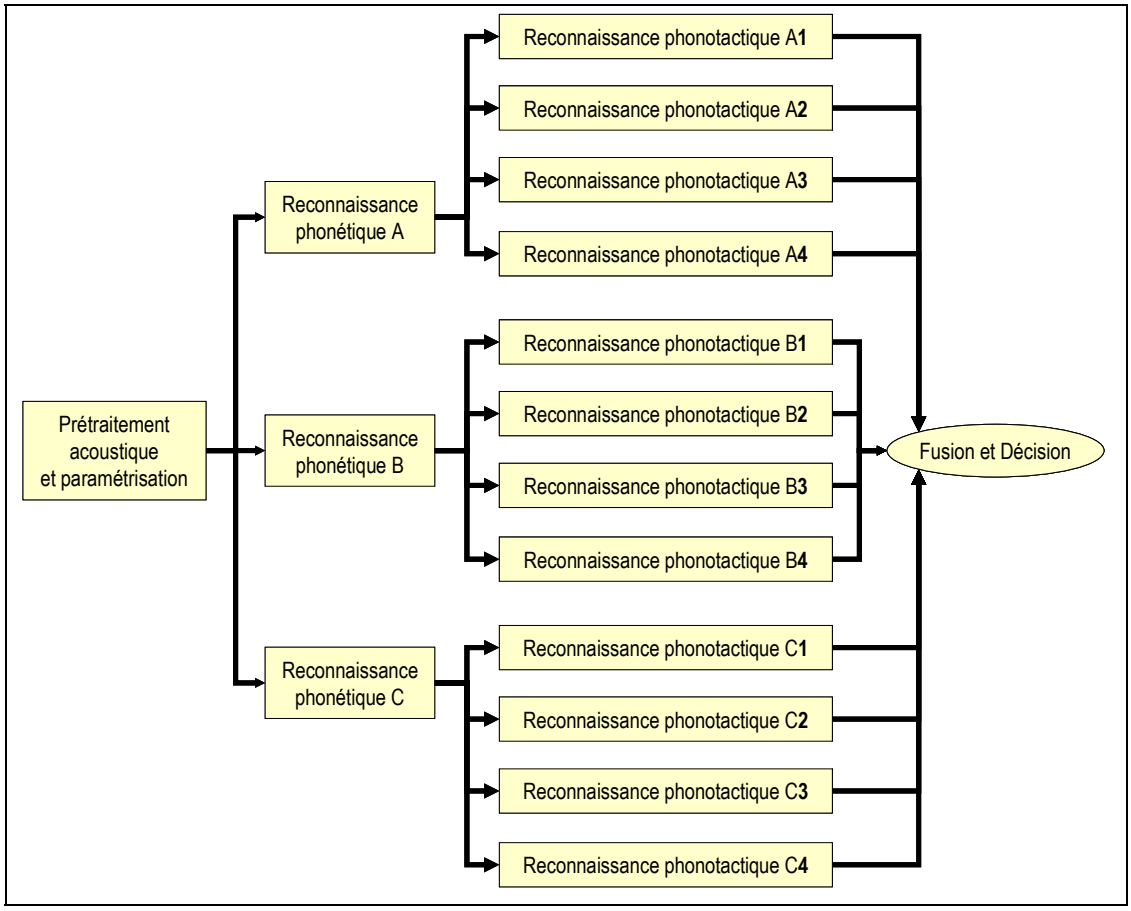


Figure 2 – Schéma d'un système PPRLM de reconnaissance de 4 langues (numérotées de 1 à 4) et utilisant 3 décodeurs acoustico-phonétiques (nommés A, B et C).

2.1.3. Approches composites et performances

La structure parallèle du PPRLM lui garantit une certaine robustesse, puisque les *p* décodeurs acoustico-phonétiques permettent *a priori* une couverture de l'espace phonétique plus large qu'un seul décodeur et que la combinaison de plusieurs systèmes de décision – qu'elle repose sur des méthodes adaptatives ou non – est notoirement robuste. À l'inverse, la structure séquentielle du PPRLM (module phonétique suivi d'un module phonotactique suivi d'un module de décision) constitue à notre avis une faiblesse, au moins dans sa version initiale, puisque d'une part elle ne permet pas l'estimation conjointe des modèles phonétiques et phonotactiques et d'autre part elle n'utilise le décodage phonétique qu'à titre de convertisseur entre l'espace spectro-temporel continu et l'espace – discret et fini – des descripteurs phonétiques. Ainsi, dans la version originale de l'algorithme, les vraisemblances des *p* séquences phonétiques décodées, estimées conditionnellement à la séquence de paramètres acoustiques, ne sont pas prises en compte explicitement pour identifier la langue.

Cette approche a connu plusieurs améliorations ou variantes depuis sa première apparition. Outre l'amélioration des algorithmes de classification

statistique (estimation des paramètres, augmentation du nombre de composantes gaussiennes des modèles, etc.), elles ont principalement porté sur le prétraitement acoustique (détection d'activité vocale, identification du sexe du locuteur, etc.) et sur le module de décision.

À titre indicatif, le système principal proposé en 2003 par le MIT[12] proposait une fusion des scores obtenus par trois approches :

- un modèle PPRLM semblable à celui décrit ci-dessus ;
- un modèle phonétique multi-gaussien (*Gaussian Mixture Model* – GMM) ;
- une approche phonétique discriminante à base de *Support Vector Machines* (SVM).

Le modèle PPRLM du MIT employait six décodeurs phonétiques appris sur des données issues de six langues (anglais, allemand, espagnol, hindi, japonais et mandarin).

Le modèle phonétique multi-gaussien marquait une sorte de retour à des approches envisagées par *Texas Instruments* dans les années 1970 (Leonard, 1980) mais il bénéficiait à la fois des progrès des modèles de mélanges de lois gaussiennes (nombre de composantes, convergence, etc.) et de l'innovation majeure qu'a été l'introduction de la paramétrisation *Shifted Delta Cepstra*[13] (ou SDC, Torres-Carrasquillo, Reynolds & Deller, 2002). Les modèles employés par le MIT étaient composés de 2048 composantes gaussiennes, spécifiques au genre (un modèle homme, un modèle femme), et adaptés à chaque langue à partir d'un modèle dit universel, lui-même estimé à partir des données d'apprentissage de l'ensemble des langues. Cette méthode a l'avantage considérable d'employer des modèles initialisés dans les mêmes conditions, augmentant ainsi leur pouvoir discriminant puisque les différences résultent exclusivement des données spécifiques à chaque langue (et pas d'artefacts liés à l'algorithme d'estimation employé).

L'approche par SVM consistait à projeter les vecteurs d'observation (coefficients SDC) dans un espace paramétrique de dimension supérieure de manière à estimer un hyperplan discriminant pour chaque décision « langue *n* » vs. « langues autres que *n* ». Il s'agit là d'une approche élégante mais assez coûteuse en temps de calcul.

---

[12] Seul les systèmes proposés par le Lincoln Lab du MIT seront abordés dans ces pages parce qu'ils constituent de facto une référence dans ce type de tâche. Cela ne signifie pas pour autant que les autres laboratoires engagés dans ces évaluations ne proposent pas d'approches innovantes et performantes. À titre d'exemple, on peut citer les systèmes proposés par l'université de technologie de Brno, en République tchèque (e.g. Matějka *et al.*, 2007).

[13] Ces paramètres sont intrinsèquement des ΔMFCC. Cependant, à l'instant *t*, le vecteur de SDC est un super-vecteur résultant de la concaténation de *k* blocs de ΔMFCC (généralement calculés sur un empan de 3 trames) décalés de *P* trames. La détermination des valeurs pour *k* et *P* est heuristique et aujourd'hui on utilise souvent des valeurs respectives de 7 et 3. Dans l'esprit, ces paramètres combinent la précision temporelle et fréquentielle de l'analyse MFCC avec la prise en compte d'un empan temporel plus long (équivalent à 7 x 3 = 21 trames de 20 ms).

La décision finale était prise en utilisant un classificateur gaussien, méthode qui a tendance depuis quelques années à supplanter les méthodes non linéaires par réseaux de neurones artificiels. L'idée est de modéliser la distribution des individus de l'ensemble d'apprentissage de chaque langue, non pas dans l'espace de paramètres acoustiques, mais dans l'espace des *scores* obtenus par chacun des individus d'apprentissage avec chacun des modèles. Cet espace de scores était en l'occurrence de dimension 108 (12 x 6 sorties de PPRLM + 12 x 2 sorties des modèles multi-gaussiens + 12 scores SVM). Lors du test, la vraisemblance obtenue par l'échantillon à tester dans les modèles gaussiens des 12 langues cibles était employée sous forme de rapport de vraisemblance pour prendre la décision finale.

Le Tableau 1 indique les résultats obtenus par ces 3 sous-systèmes du MIT et par le système composite dans la tâche NIST LRE 2003.

Tableau 1 – Taux d'égale erreur (EER, %) obtenus par les systèmes du MIT présentés lors de la campagne NIST LRE 03 pour les différentes durées moyennes des enregistrements de test (d'après Singer *et al.*, Eurospeech 2003).

	30s	10s	3s
MIT PPRLM	6,6	14,3	25,5
MIT GMM	4,8	9,8	19,8
MIT SVM	6,1	16,4	28,2
MIT FUSION	2,8	7,8	20,3

En 2004, le LIMSI a également testé son propre système de vérification de la langue dans la tâche NIST 2003, atteignant le même niveau de performances que le MIT, tout en utilisant une méthode nettement moins coûteuse en temps de calcul (Gauvain, Messaoudi & Schwenk, 2004). Ce gain est en particulier lié à l'adaptation faite par le LIMSI de l'approche PPRLM : en utilisant *un ensemble d'hypothèses* phonétiques via des treillis de phones plutôt que *la meilleure hypothèse*, il est possible de procéder à une optimisation globale de la chaîne de modélisation *phonétique* + *phonotactique*, procédure qui a déjà fait ses preuves en RAP et qui se révèle très efficace. Des expériences complémentaires ont de plus montré qu'un faible nombre de décodeurs acoustico-phonétiques (3 en l'occurrence : arabe, anglais américain et espagnol) pouvait se révéler plus efficace que le classique modèle PPRLM à 6 décodeurs.

Depuis, les campagnes NIST LRE 05 et LRE 07 ont permis de tester de nouvelles techniques et d'améliorer celles déjà mises en place en 2003. Lors de la campagne NIST LRE 05, le MIT a proposé un système composite qui tirait partie des trois modèles introduits lors de l'évaluation de 2003, augmentés de deux versions utilisant des treillis pour les GMM et SVM, ainsi que d'un système PPRLM où le modèle phonotactique est de type arbre de décision binaire plutôt que grammaire *n*-gramme (Campbell *et al.*, 2006). Ce dernier système a alors créé la surprise puisqu'il atteignait les meilleures performances pour un système individuel (i.e. ne résultant pas d'une fusion). La campagne 2007 a vu le MIT proposer près d'une dizaine de

systèmes, donnant lieu à une combinatoire encore plus conséquente pour les modèles composites.

Le présent document de synthèse n'est pas le lieu pour mener une étude détaillée des différents sous-systèmes proposés lors de ces évaluations. On peut cependant se reporter à la Figure 3, tirée de Campbell *et al.* (2006) qui illustre l'évolution des performances des systèmes d'identification des langues proposés par le MIT depuis une dizaine d'années.

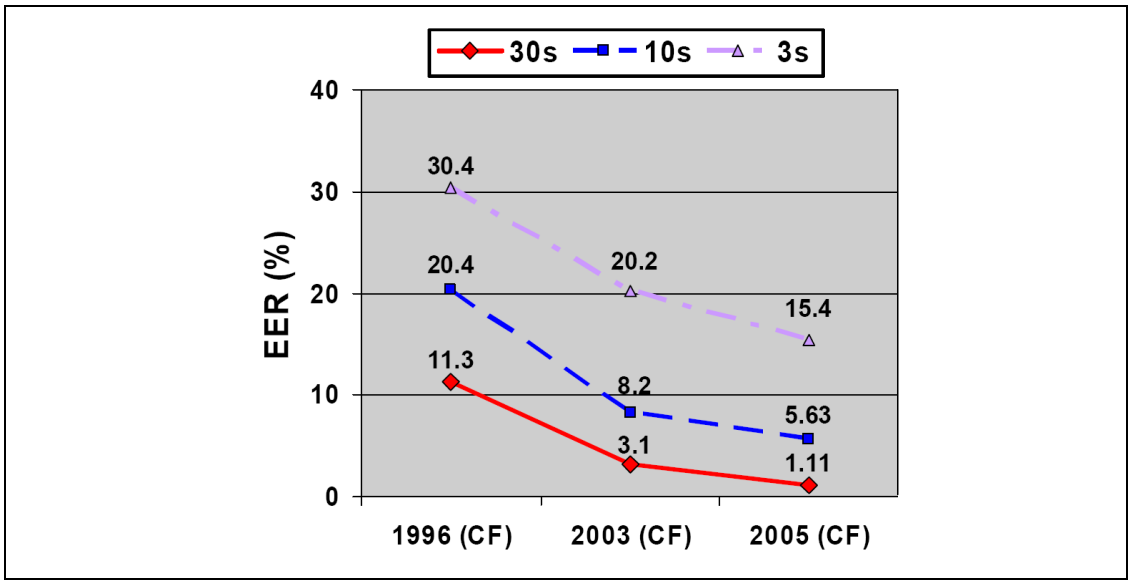


Figure 3 – Évolution des performances obtenues par le système composite du MIT lors des campagnes d'évaluation NIST LRE 96, LRE 03 et LRE 05 (adapté de Campbell *et al.*, 2006).

Cette figure indique clairement qu'une étape décisive a été franchie, en particulier pour des durées supérieures ou égales à 10 secondes, avec des performances de l'ordre de 5 % de taux d'égale erreur, voire 1 % pour les extraits les plus longs. De plus, l'analyse détaillée des résultats montre que les erreurs se produisent souvent pour des fichiers de mauvaise, voire très mauvaise, qualité (en matière de rapport signal/bruit ou d'intelligibilité (Martin & Le, 2006)).

2.2. APPROCHE DÉVELOPPÉE

2.2.1. La fin d'une problématique ?

Ainsi, l'état de l'art actuel en identification de la langue est-il défini par des systèmes développés dans les laboratoires qui sont également *leaders* en reconnaissance automatique de la parole. Cette situation confirme ainsi l'importance fondamentale revêtue à la fois par la qualité des modèles phonétiques utilisés et par le soin scrupuleux accordé à la définition du cadre statistique dans lequel se situe l'étude. J'avais indiqué dans Pellegrino (1998) que le peu de cas fait des caractéristiques phonétiques des langues était non satisfaisant. En effet, à cette

époque, l'architecture PPRLM ne laissait qu'un rôle implicite aux modules phonétiques en ne les intégrant pas au module de décision. Aujourd'hui, la paramétrisation a évolué (les coefficients SDC paraissent ainsi particulièrement pertinents) et la prise en compte de la variabilité phonétique s'améliore également par l'utilisation de mélanges de lois gaussiennes d'ordre important (1024 ou 2048). Du coup, certains des meilleurs systèmes tiennent maintenant compte explicitement des caractéristiques phonétiques des langues par l'intermédiaire de modèles GMM (cas du MIT). Pour autant, le système du LIMSI semble confirmer qu'une modélisation phonétique implicite peut aussi être performante. Cependant, la médaille de la performance présente un revers puisque l'intégration des informations phonétiques et phonotactiques dans un processus d'optimisation global se fait aux dépends de l'analyse linguistique du poids desdites informations, et en particulier, des performances et erreurs spécifiques aux différents modèles. Évidemment, en fonction du type d'application ou du type d'étude linguistique visé, cet aspect peut être considéré comme marginal ou à l'inverse fondamental.

Peut-on pour autant dire aujourd'hui que la problématique de l'IAL est résolue ? À mon sens, les niveaux de performances obtenus sont déjà satisfaisants pour certaines applications commerciales impliquant quelques langues. À l'inverse, dès lors qu'il s'agit de traiter une vingtaine de langues ou plus (pensons aux 23 langues officielles de l'union européenne), des progrès sont encore à accomplir, concernant en particulier la durée de parole nécessaire à l'identification. La comparaison des taux d'égale erreur obtenus en 2003 sur les extraits de 10 et 3 secondes (respectivement 7,9 % et 18,3% pour le système du LIMSI) par rapport à ceux atteints pour 30 secondes (2,7 %) est, à ce titre, révélatrice. D'un point de vue strictement technique, la diminution de la durée de parole disponible pour procéder à l'identification impliquera soit de disposer de modèles plus discriminants, soit de recourir à des informations pertinentes additionnelles[14]. Par ailleurs, il est assez stimulant de constater que les résultats varient de manière très significative en fonction des langues considérées, sans même parler de la prise en compte des variétés dialectales (voir Martin & Le, (2006) pour une discussion de cet effet lors de l'évaluation NIST LRE 05).

D'un point de vue plus large, l'amélioration des performances implique une meilleure prise en compte de la variabilité inter-locuteur et inter-langue et surtout de la relation entre les deux niveaux. Ainsi, il est probable que les performances à venir dépendent de la connaissance qu'on aura des phénomènes linguistiques en jeu et de leur caractère plus ou moins spécifique à certaines langues, et encore plus de notre capacité à en modéliser les effets. À titre d'exemple, il est probable que, tôt ou tard, les débats animant la communauté des linguistes et des psycholinguistes sur la nature des représentations phonologiques (gestes articulatoires, traits, phonèmes,

[14] Évidemment, ces informations additionnelles seront soumises au même compromis entre pertinence en termes de différenciation des langues et robustesse en termes d'évaluation statistique à partir d'échantillons courts (cf. Hombert & Maddieson, 1999). À ce jeu là, les paramètres suprasegmentaux (prosodie, lexique) auront à surmonter le handicap que constitue leur faible nombre au sein d'extraits de trois secondes.

syllabes, etc.) influencent la conception des modèles phonétiques. Force est de reconnaître cependant que l'utilisation de modèles statistiques intégrant « massivement » la variabilité par l'intermédiaire de milliers de composantes gaussiennes en les structurant à partir des données et non à partir de connaissances linguistiques est une méthode tout à fait efficace aujourd'hui. Par conséquent, ce n'est sans doute pas sur ce plan que la linguistique peut obtenir rapidement des résultats significatifs en termes d'amélioration des performances.

Cependant, il me semble que la prise en considération de nouvelles sources d'informations bénéficiera d'une collaboration entre linguistes et spécialistes de l'ingénierie. Dans ce cadre, on peut citer au moins deux principaux champs d'investigation situés à l'opposé sur l'échelle de durée :

- les phénomènes phonétiques fins, liés à des différences d'articulation ou de co-articulation de phonèmes entre les langues ;
- les phénomènes prosodiques et suprasegmentaux.

Les travaux entrepris depuis quelques années sur les différences fines existant entre les langues, que ce soit à DDL (en particulier dans la thèse de Jalaleddin Al-Tamimi, 2007) ou ailleurs, peuvent mettre en évidence des niveaux infra-phonémiques pertinents en identification des langues. La réalisation des consonnes occlusives est un exemple bien connu de variabilité inter-langues. L'utilisation contrastive de phénomènes tels que l'aspiration ou l'éjection joue ainsi un rôle que les modèles statistiques peuvent appréhender pour peu qu'ils soient conçus en conséquence. En parallèle, la recherche d'invariants est une source potentielle de progrès encore peu explorée. Par exemple, le pouvoir discriminant de paramètres issus des équations du locus (Lindblom 1963 ; Sussman, Hoemeke & Ahmed, 1993 ; entre autres) mériterait probablement une étude attentive (cf. également Al-Tamimi (2007:157 et suivantes)). De même, des phénomènes de diphtongaison ou de hiatus peuvent porter des informations pertinentes s'ils sont pris en compte, éventuellement de manière statistique (voir par exemple Hualde & Chitoran, 2003) pour une étude de telles différences dans les langues romanes). Enfin, les différences de performances observées entre sujets humains et modèles statistiques, par exemple pour la détection du voisement, dans des tâches d'identification ou de discrimination phonétiques en environnement bruité mettent également en évidence un terrain sur lequel les modèles peuvent encore progresser (voir par exemple Lippmann, 1997 ; Chang, Shastri & Greenberg, 2001).

Sur le plan prosodique et suprasegmental, l'apport peut être triple. En effet, des modèles prosodiques peuvent se révéler plus robustes que les modèles phonétiques et phonotactiques lorsque les enregistrements sont particulièrement bruités : en s'intéressant à des mesures physiques à relativement long terme, ils peuvent éventuellement dégager une information pertinente d'un signal de faible rapport signal sur bruit. En particulier, les bruits d'origine impulsionnelle, problématiques pour des modèles phonétiques ou phonotactiques, peuvent probablement être « filtrés » par des modèles prosodiques et éviter ainsi la prise en compte d'événements « catastrophiques » dans les modèles phonotactiques. Sur un

autre plan, les modèles prosodiques peuvent être les plus efficaces dans un contexte d'identification dialectale, où les différences peuvent être plus prosodiques que phonétiques. Par exemple, certaines langues chinoises diffèrent principalement dans leur utilisation de la dimension tonale, paramètre peu intégré dans les approches segmentales classiques. De même, la distinction des dialectes japonais de Tokyo et Osaka peut se baser en partie sur cette dimension prosodique (Kamiyama, 2004) puisque seul celui de Tokyo semble être à accent tonal (ou *pitch-accent*). Incontestablement, avant de pouvoir faire leurs preuves, de tels modèles prosodiques doivent être conçus et validés dans le contexte d'utilisation des systèmes d'IAL, à savoir sur des données de parole spontanée. Enfin, les modèles phonotactiques eux-mêmes peuvent probablement bénéficier d'innovations : pourquoi se limiter à prendre en compte uniquement les enchaînements de segments adjacents ? De nombreux phénomènes phonologiques à l'œuvre dans les langues du monde agissent à distance, sous forme de phénomènes d'harmonie. Les plus connus sont évidemment des phénomènes d'harmonie vocalique (exemple du turc ; de l'harmonie de type ±ATR (*Advanced Tongue Root*) dans des langues africaines ; ou même du français (Carré, Bourdeau & Tubach, 1995 ; Fagyal, Nguyen & Boula de Mareüil, 2002)) mais des phénomènes d'harmonies consonantiques sont également bien documentés (par exemple les harmonies de nasalisation dans certaines langues amérindiennes). Par conséquent, la prise en compte en parallèle de modèles phonotactiques adjacents et à distance (de type V_V et C_C) est potentiellement pertinente.

Je me suis efforcé dans les paragraphes précédents de valider l'intérêt d'une approche linguistique en IAL. De manière réciproque, il me semble indispensable de mentionner que les systèmes développés en ingénierie des langues peuvent permettre des avancées nettes en linguistique. En effet, une des limites évidentes de nombreux travaux en linguistique, et en particulier en phonétique, est de travailler sur de petits corpus produits par peu de locuteurs et étiquetés en grande partie à la main. L'analyse linguistique des résultats obtenus par les meilleurs systèmes actuels, aussi imparfaite que soit la modélisation phonétique appliquée, semble donc être un complément indispensable à la description de faibles quantités de données. En premier lieu, nous sommes tous conscients des problèmes de consistance et de variabilité que peut soulever l'étiquetage manuel par plusieurs experts humains. L'utilisation d'un unique outil automatique, aussi imparfait soit-il, garantit quant à lui cette consistance souvent si importante.

En inscrivant l'étude du langage dans une perspective fonctionnelle, il est évident que l'étude statistique des réalisations phonétiques permettra encore une fois de mieux circonscrire la variabilité observée et d'apporter des éléments de réponse à des problématiques connexes, comme l'existence d'effets phonétiques à longue portée mentionnée précédemment ou d'interactions entre caractéristiques spectrales segmentales et prosodie. À ce titre, il me semble très important que la communauté linguistique travaillant sur la langue française s'intéresse aux résultats de la campagne d'évaluation des systèmes de transcription ESTER et tire profit des

alignements phonétiques qui sont diffusés avec les 100 heures de corpus radiophoniques[15]. Outre le bénéfice linguistique attendu, on peut également envisager en retour une amélioration des modélisations utilisées, en particulier sur les aspects phonétiques fins évoqués précédemment et sur des phénomènes morpho-phonologiques, voire syntaxiques.

En résumé, il me semble que l'irrigation réciproque des approches et des problématiques rencontrées en linguistique et en ingénierie des langues est fertile pour les deux champs, plus encore aujourd'hui qu'hier. Je considère que cette interdisciplinarité, construite à partir des meilleures compétences, permet non seulement d'améliorer les performances des systèmes mais également de faire évoluer les problématiques scientifiques et les applications de ces approches. Cependant, il ne faut pas nier les difficultés d'une telle approche qui doit concilier les contraintes méthodologiques issues des deux mondes.

2.2.2. Une approche médiane

Celui qui s'engage sur la voie de l'interdisciplinarité s'expose à des critiques – souvent justifiées. A mon sens, il est nécessaire que cette activité plonge ses racines dans chacun des champs scientifiques concernés plutôt que de s'en couper, sous peine de perdre toute pertinence. Dans le cas du traitement automatique de la parole, un autre écueil est également rencontré. Quel chercheur en traitement automatique de la parole expliquant son travail à des linguistes ne s'est pas entendu répliquer qu'il faisait « n'importe quoi » parce qu'il ne prenait pas en compte explicitement tel ou tel phénomène[16] ? Quel linguiste n'a pas un jour eu l'impression que le chercheur susmentionné le prenait pour un idiot sous prétexte que le phénomène en question était *automatiquement* pris en compte dans les modèles statistiques ?

Plus que la simple anecdote, cette situation révèle une divergence de vue sur l'objet scientifique étudié. Dans ce cadre, on pourrait considérer que le linguiste cultive une vision *analytique* du problème, s'attachant à distinguer le pertinent de l'accessoire, le fond de la forme, la surface de la profondeur ou encore l'invariant de la variabilité et à comprendre cette variabilité. En ce sens, la compréhension du mécanisme sous-jacent est essentielle. Par conséquent, il est impossible d'exploiter un corpus « en aveugle », c'est-à-dire en ne regroupant pas à la main les multiples exemplaires d'un même phénomène et en ne comprenant pas leurs relations. A l'inverse, le chercheur en ingénierie de la parole adoptera une approche beaucoup plus *holistique*, considérant que ce qui l'intéresse en priorité, c'est de disposer du modèle répondant correctement à l'inventaire de situations le plus large possible,

---

[15] Même si la nature des corpus est différente, on peut s'attendre à terme à ce que la mise à disposition du corpus étiqueté ESTER ait un impact pour l'étude du français comparable à la diffusion de *Switchboard* (Godfrey, Holliman & McDaniel, 1992) pour l'anglais (pour ce type d'étude, voir en particulier les travaux de S. Greenberg à ICSI, e.g. Greenberg, Chang & Hitchcock, (2001).

[16] dont le linguiste en question est évidemment le spécialiste mondial !

c'est-à-dire intégrant le maximum de variabilité, de manière à ce que le système ne soit pas pris au dépourvu quand un extrait inconnu lui est présenté.

La situation brossée ici est, là encore, caricaturale et une approche médiane, multi-échelle, est également possible et appliquée par certains scientifiques dans le monde entier. Cette position est parfois inconfortable car elle ne se révèle souvent pas assez efficace en matière de performance pour les uns, et elle manque de précision concernant l'analyse linguistique pour les autres. Néanmoins, elle autorise également de réelles découvertes scientifiques et elle permet d'avoir un point de vue plus complet sur l'objet d'étude complexe que constituent la parole et le langage et en particulier sur la dualité entre invariance et variabilité. Depuis quelques années, une autre approche a également été proposée, visant à relier l'ingénierie de la parole à la cognition. À ma connaissance, cette approche s'est pour l'instant révélée relativement peu efficace dans un domaine comme l'IAL, en particulier parce qu'elle balbutie encore et que son coût de développement est important car elle nécessite la compréhension de mécanismes et de comportements humains complexes[17].

Sans aller jusqu'à revendiquer une approche cognitive sur ce thème, je considère les travaux que j'ai entrepris depuis plusieurs années avec mes collègues comme relevant d'une approche médiane.

2.3. TRAVAUX EN IDENTIFICATION AUTOMATIQUE

La première section ci-dessous rappelle brièvement l'approche que nous avions développée en thèse. Elle est suivie d'une section plus conséquente consacrée à la modélisation du rythme que nous avons développée entre 2000 et 2003 environ. Chacune de ces sections s'achève par une discussion des résultats obtenus et des directions de recherche qui pourraient être poursuivies.

2.3.1. Modélisation segmentale

Document de référence : Annexe C.1

Pellegrino F. & André-Obrecht R. 2000. "Automatic language identification: an alternative approach to phonetic modelling", *Signal Processing*, Vol. 80 issue 7, pp. 1231-1244, July 2000.

À l'époque, plusieurs constatations nous ont guidés vers la conception d'une approche d'IAL basée sur des modèles phonétiques différenciés :

1. La conviction que les indices phonétiques étaient sous-exploités par les modèles de type PPRLM ;

2. La volonté de recourir à des modèles estimés sans données étiquetées manuellement ;

3. l'intuition qu'un espace paramétrique unique pour des classes de sons hétérogènes n'était peut-être pas optimal (prise en compte de la *nature* des données) ;

[17] Certains ne manqueront pas de rappeler que les avions les plus efficaces ne battent pas des ailes. Pour autant, les performances humaines en termes de reconnaissance et de compréhension de la parole me semblent mériter un examen attentif…

4. la volonté de modéliser les classes naturelles phonétiques, de manière à capturer une information systémique sur les oppositions à l'œuvre dans chaque langue (prise en compte de la *structure* des données).

Notre approche a donc visé à étiqueter automatiquement les consonnes et les voyelles dans le signal et à développer des modèles statistiques différenciés et spécifiques aux langues. Nous avons tout d'abord conçu un algorithme non supervisé de localisation des noyaux vocaliques dans le signal acoustique. Cet algorithme exploitait la segmentation statistique du signal *a priori* obtenue à partir de l'algorithme de divergence *Forward-Backward* (André-Obrecht, 1988). Conjointement, un détecteur d'activité vocale qui s'appuyait sur une modélisation gaussienne de l'énergie des segments (et non de trames de longueurs fixes, comme c'était classiquement le cas) était employé pour localiser les pauses dans le signal. Enfin, les pics d'une fonction quantifiant la structure formantique du signal (estimée sur une fenêtre glissante) étaient sélectionnés comme noyaux vocaliques potentiels et validés à l'aide d'un seuillage adaptatif. Ces trois étapes conduisaient pour chaque extrait de parole à l'étiquetage représenté sur la Figure 4.

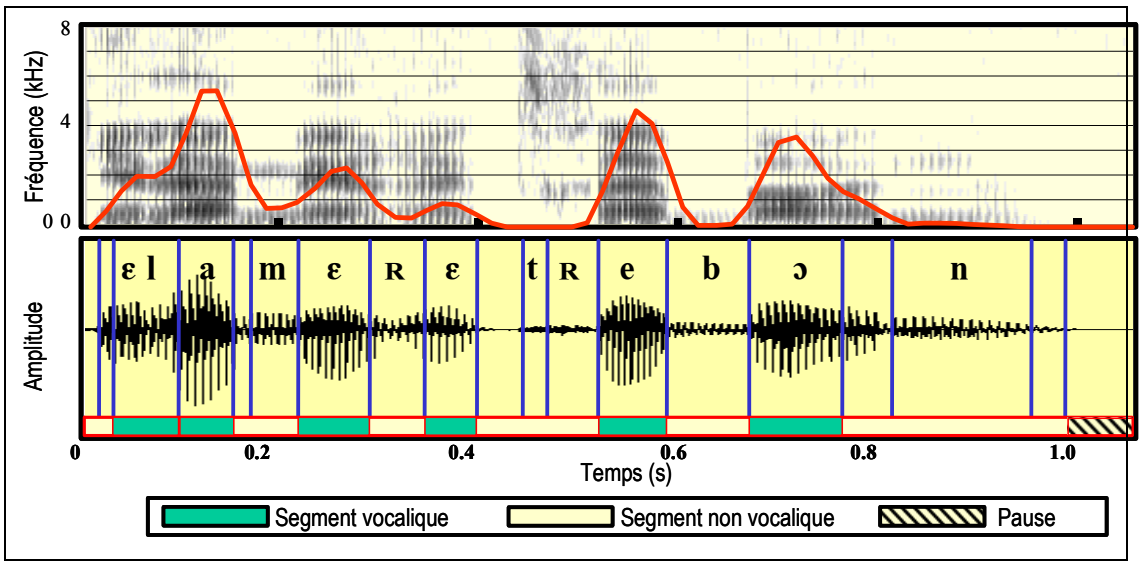


Figure 4 – Signal acoustique et spectrogramme d'un extrait du corpus MULTEXT (locuteur masculin, la phrase prononcée est « … et la mer est très bonne »). La courbe rouge représente le critère de détection de noyaux vocaliques ; les traits verticaux bleus indiquent les frontières segmentales et l'étiquetage final de chaque segment (Pause, Segment non vocalique, Segment vocalique) est codé par les couleurs dans la rangée du bas (pour les détails, voir Pellegrino & André-Obrecht, 2000)

Les résultats obtenus avec cette technique sont sensiblement différents de ce qu'un expert aurait étiqueté. L'exemple présenté ci-dessus fait apparaître les principaux éléments de discordance : d'une part la segmentation utilisée est de nature sub-phonémique et non phonémique : en particulier, des segments courts sont générés dans les phases transitoires (exemple de la transition [a]-[m] dans l'exemple ou de l'atténuation en fin de [n]). Par conséquent, les durées des segments vocaliques correspondent plus aux durées des parties stables des voyelles qu'aux voyelles elles-mêmes. D'autre part, certaines séquences phonétiques ne se prêtent pas systématiquement à une segmentation interne basée sur la détection de

ruptures, en particulier pour cause de coarticulation (exemple de la transition [ɛ]-[l] en début d'extrait). Enfin, il arrive qu'en fin d'énoncés dans des séquences de type consonne liquide ou nasale suivie d'une voyelle haute peu énergétique, le maximum local de la courbe de détection formantique se situe sur le segment consonantique et non vocalique (cas non représenté sur cette figure).

Une évaluation de la qualité de la détection avait été menée et permettait cependant de situer notre algorithme dans la bonne moyenne des détecteurs vocaliques disponibles, en particulier pour un algorithme non optimisé pour une langue donnée (cf. Tableau 2).

Tableau 2 – Comparaison des performances d'algorithmes de détection vocalique. (*) dans cette étude, il s'agit de détection de noyaux syllabiques et non à proprement parler de voyelles (d'après Rouas *et al.*, 2005).

REFERENCE	CORPUS	LANGUE	TAUX D'ERREUR VOCALIQUE
Pfitzinger, Burger & Heid, 1996(*)	PhonDatII (parole lue)	Allemand	12,9 %
	Verbmobil (parole spontanée)	Allemand	21,0 %
Fakotakis, Georgila & Tsopanoglou 1997	TIMIT (parole lue)	Anglais	32,0 %
Pfau and Ruske, 1998	Verbmobil (parole lue)	Allemand	22,7 %
Howitt, 2000	TIMIT (parole lue)	Anglais	29,5 %
Pellegrino, 1998	OGI MLTS (parole spontanée)	Français	19,5 %
		Japonais	16,3 %
		Coréen	28,5 %
		Espagnol	19,2 %
		Vietnamien	31,1 %
		Moyenne	22,9 %

Depuis, d'autres techniques d'analyse non supervisée de la structure formantique ont vu le jour avec des objectifs similaires, mais à notre connaissance, les performances obtenues demeurent du même ordre de grandeur que notre algorithme.

Une fois les segments vocaliques et non vocaliques étiquetés (et les pauses éliminées) une paramétrisation cepstrale était appliquée pour projeter chaque segment dans un espace à 19 dimensions[18]. Deux modèles multigaussiens, l'un dit vocalique et l'autre consonantique, étaient ensuite estimés pour les cinq langues que nous avions à l'époque retenues dans le corpus OGI MLTS.

[18] Huit coefficients cepstraux calculés sur une échelle de Mel (MFCC) statiques + l'énergie ; huit coefficients dynamiques ΔMFCC et la dérivée de l'énergie, ainsi que la durée du segment considéré. Cette paramétrisation est issue des travaux antérieurs de R. André-Obrecht.

Un des inconvénients majeurs des modèles segmentaux (par opposition aux modèles dits centisecondes) est que la quantité de données disponible pour leur estimation est mécaniquement plus réduite : en effet, un segment dure généralement plusieurs trames et l'ensemble d'apprentissage s'en trouve réduit d'autant puisque l'on passe d'un ensemble d'apprentissage de N trames à un ensemble réduit de M segments, avec $M < N$. Par conséquent, le modèle optimal pour ce type d'approche est d'une taille inférieure au modèle centiseconde correspondant. Tout l'enjeu de ce type d'approche est donc d'évaluer si l'apport éventuel de la segmentation (meilleure prise en compte de la cohérence temporelle du signal) compense la perte liée à l'estimation moins précise des paramètres (ou à leur nombre plus réduit). Pour évaluer le nombre de composantes gaussiennes optimal pour chaque langue, nous avions eu recours à un principe de parcimonie de type *Minimum Description Length* ; il s'agissait d'une adaptation de l'algorithme de quantification vectorielle LBG (Linde, Buzo & Gray, 1980) intégrant le critère de Rissanen (1983). Ainsi, pour chaque langue et en fonction de la quantité de données disponibles, nous obtenions un modèle de taille optimale au sens du critère choisi.

In fine, nous avions testé plusieurs configurations : modèles vocaliques segmentaux seuls, modèles consonantiques segmentaux seuls et modèles segmentaux globaux. Il s'était avéré que l'algorithme LBG-Rissanen était pertinent pour déterminer la taille optimale des modèles vocaliques et que la fusion statistique des vraisemblances des modèles vocaliques et globaux permettait d'améliorer les résultats jusqu'à atteindre 91 % d'identification correcte sur les données des locuteurs masculins du corpus de test dit « 45 s » du corpus OGI MLTS (Pellegrino, Farinas & Obrecht, 1999).

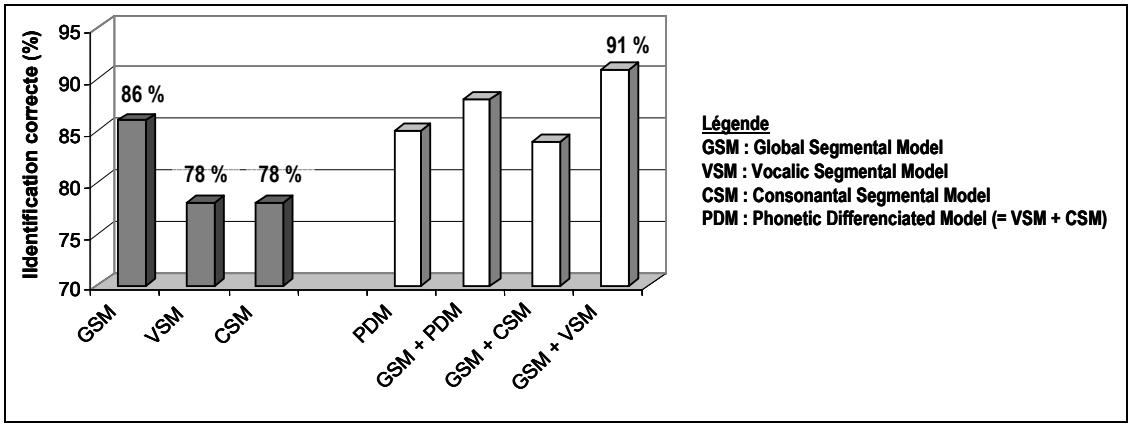


Figure 5 – Taux d'identification correcte obtenus avec les différents modèles développés durant la thèse (d'après Pellegrino, Farinas & André-Obrecht, 1999).

✦ Discussion

L'approche développée, quoique rudimentaire, s'est révélée relativement efficace. Elle aurait probablement mérité d'être approfondie selon plusieurs directions. Tout d'abord, notre travail s'était principalement orienté vers les systèmes vocaliques, même si nous avons par la suite fait une timide tentative de différenciation des classes naturelles consonantiques. La différenciation des systèmes consonantiques et son corollaire, la sélection d'espaces de représentation

paramétriques adaptés aux classes identifiées nous paraît toujours d'actualité. En effet, les coefficients SDC, si efficaces aujourd'hui, reposent sur le choix de plusieurs paramètres et même si l'on se place dans un espace cepstral commun, leurs optima sont potentiellement différents selon les classes considérées. Proposer une paramétrisation adaptative (en fonction de la durée du segment considéré et de sa classe naturelle) peut donc potentiellement être efficace. Ensuite, la fusion d'approches segmentales et centisecondes permettrait une modélisation multi-échelle non supervisée pour un coût dérisoire. En résumé, il nous semble que l'utilisation de modèles de types multigaussiens différenciés avec une sélection des paramètres optimaux à partir des données et fusion d'approches segmentales et centisecondes ouvre une piste potentiellement compétitive en termes d'IAL.

2.3.2. Modélisation du rythme

Documents de référence : Annexes C.2, C.3 et C.4

Pellegrino F. 2008. "Rhythm" in *The Cambridge Encyclopedia of the Language Sciences*. Patrick Colm Hogan (ed.) Cambridge: Cambridge University Press.

Rouas, J-L., Farinas J., Pellegrino F. & André-Obrecht, R. 2005, "Rhythmic Unit Extraction and Modelling for Automatic Language Identification", *Speech Communication*, Vol. 47, issue 4, pp. 436-456

Farinas, J., Rouas, J.L., Pellegrino, F. & André-Obrecht, R. 2005, "Extraction automatique de paramètres prosodiques pour l'identification automatique des langues", *Traitement du Signal*, 22:2

Comme indiqué dans l'exposé de mes motivations (p. 28), le rythme peut être considéré comme une interface entre les niveaux segmental et suprasegmental. Cette dualité suggère ainsi deux orientations de recherche pour son intégration dans des modèles d'IAL. D'une part, on peut substituer des unités rythmiques aux traditionnelles unités segmentales (phonèmes, diphones, etc.) dans les modèles phonétiques et cette direction est celle choisie dans les approches syllabiques ou syllabotactiques développées depuis quelques années, principalement à Berkeley et Nijmegen en reconnaissance automatique de la parole (laboratoire ICSI : Wu, 1998 et *Radboud University* : Hämäläinen *et al.*, 2007) et à Orsay en IAL (laboratoire LIMSI, Zhu & Adda-Decker, 2006). D'autre part, on peut extraire des paramètres rythmiques, liés alors principalement à des durées et des alternances d'unités de durées variables et établir des modèles prosodiques exploitant ces paramètres. C'est dans cette perspective que se situent nos travaux, alors même que peu d'études s'étaient attaquées à ce problème. Thymé-Gobbel et Hutchins (1999) avaient cependant déjà attiré l'attention sur l'importance des paramètres rythmiques en IAL. Elles avaient développé un système basé sur des vraisemblances statistiques estimées à partir d'un grand nombre de paramètres relatifs au rythme et à la distribution temporelle de syllabes détectées à partir de la courbe d'énergie. Au total, 224 paramètres étaient intégrés, avec plus ou moins d'efficacité, dans leur modèle. La validation s'est faite sur les données du corpus OGI MLTS, sur une tâche de discrimination de langues prises deux à deux. La tâche était donc simple, mais il faut reconnaître que défricher le champ de la modélisation prosodique à partir des données de qualité médiocre du corpus MLTS était un pari risqué. Les résultats obtenus s'échelonnaient du niveau du hasard (50 %) à 93 % selon les paires de

langues considérées. Le nombre important de paramètres considérés ne permettait cependant pas d'identifier les primitives rythmiques, rendant l'interprétation des résultats hasardeuse.

✦ Des primitives rythmiques ?

Les travaux de Ramus et ses collègues (Ramus & Mehler, 1999 ; Ramus, Nespor & Mehler, 1999 ; Ramus, 2002), implémentés à cette époque de manière semi-automatique par Dominey & Ramus (2000), ont directement inspiré notre étude. À partir d'un étiquetage *manuel* des séquences consonantiques et vocaliques, un réseau de neurones récurrent était entraîné pour discriminer des langues (prises là encore deux à deux). L'objectif était d'évaluer dans quelle mesure la notion de classes rythmiques était objectivable et partant, d'évaluer si des langues de classes rythmiques distinctes étaient discriminables, par opposition à des langues de même classe rythmique. Les données provenaient du corpus dit RNM « Ramus, Nespor, Mehler » et les modèles discriminèrent correctement des langues de classes rythmiques différentes (78 % d'identification correcte pour la paire japonais-anglais) tandis que le taux obtenu pour des langues de même classe n'était pas significativement différent du hasard. Les travaux de Ramus ont, à la même époque, inspiré une étude à Galves *et al.*, (2002), qui, sans recourir à un étiquetage manuel, obtinrent une partition multidimensionnelle des langues de ce même corpus similaire en calculant des paramètres de sonorité (moyenne et dérivée) qui se révélèrent corrélés aux paramètres définis par Ramus.

Modéliser efficacement le rythme relève du défi permanent, en particulier, de par l'absence de caractérisation consensuelle du phénomène et sa nature multidimensionnelle. Sur ce plan, et en se plaçant dans le cadre applicatif de la synthèse vocale, Zellner-Keller a apporté une contribution remarquable à la délimitation du périmètre du phénomène rythmique (Zellner-Keller and Keller, 2001 ; Zellner-Keller, 2002).

En nous inspirant de sa réflexion et de notre propre expérience, nous avons proposé de définir le rythme comme résultant de l'interaction complexe de trois facteurs portant sur :

- la nature des constituants rythmiques ;
- l'usage facultatif d'alternances entre constituants de proéminence variable ;
- l'existence facultative de régularités dans le regroupement des constituants en unités plus longues.

Dans ce cadre, chacun des facteurs peut mener à une approche partielle de la modélisation du rythme pour l'IAL, et l'on constate effectivement que les recherches menées en amont sur les corrélats acoustiques du rythme se sont également intéressés à ces différents niveaux.

L'approche conçue par Ramus s'intéresse à des propriétés statistiques des durées intervocaliques et vocaliques dans le corpus RNM ; $\%V$ est la quantité vocalique globale (en proportion de la durée totale) d'un énoncé tandis que ΔV (et

respectivement ΔC) est l'écart-type des durées des intervalles vocaliques (resp. intervocaliques) de l'énoncé. L'hypothèse supportant le choix de ces paramètres est que le rythme perçu est lié aux propriétés temporelles des intervalles intervocaliques et vocaliques, considérés comme les primitives pertinentes. Le fait de s'intéresser à des valeurs relatives (ratio de durées ou écarts-types plutôt que durées) présente l'avantage de limiter la dépendance au débit réel de la parole, dans la mesure où l'on considère que son effet sur les durées vocaliques et intervocaliques est homothétique[19]. Depuis, l'expérience a montré à plusieurs reprises que ces corrélats acoustiques étaient pertinents pour distinguer des langues et parfois également des dialectes (voir thèses de Rym Hamdi, (2007) et d'Emmanuel Ferragne, (2008) pour des expériences dans ce cadre). Cependant, ces paramètres statistiques ne permettent pas de prendre en compte des différences d'alternance entre intervalles longs et brefs par exemple, comme illustré Figure 6.

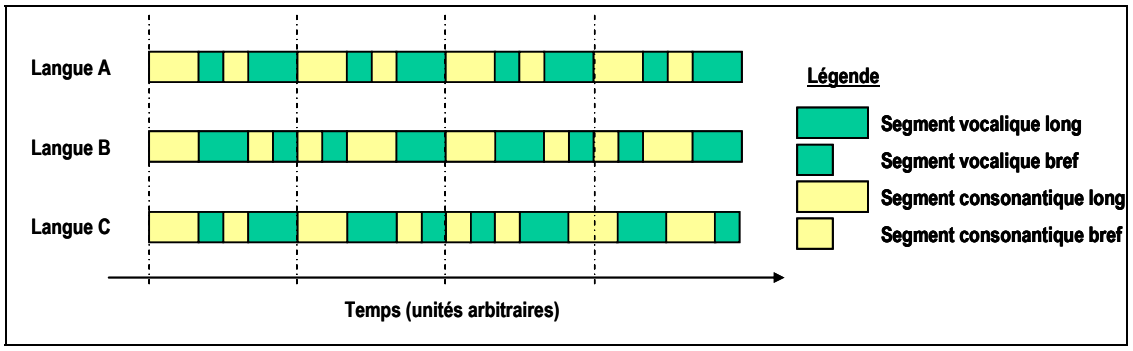


Figure 6 – Représentation schématique de patrons d'alternance entre segments vocaliques et consonantiques dans trois langues imaginaires.

Cette figure représente de manière schématique sous forme de séquences de consonnes et de voyelles des énoncés produits dans trois langues hypothétiques différentes. Chaque énoncé est formé de la même distribution statistique de durées vocaliques et intervocaliques (mêmes nombres d'intervalles consonantiques et vocaliques, longs et brefs, dans chaque langue). Par conséquent, les valeurs de $\%V$, ΔV et ΔC sont identiques dans les trois langues alors même que si l'on considère les séquences temporelles réalisées, certaines régularités permettent de les distinguer : la langue A produit une alternance de composants (consonantiques ou vocaliques, considérés comme deux flux distincts) brefs et longs et les groupes {consonne + voyelle} sont tous isochrones. À plus grande échelle encore, on constate que les groupes {consonne longue + voyelle brève} et {consonne brève + voyelle longue} alternent régulièrement.

La langue B présente un motif inverse puisqu'elle met en jeu une harmonie de durée à l'intérieur des groupes {consonne + voyelle} au sein desquels les deux composantes sont conjointement longues ou brèves. À plus grande échelle les

[19] Cette hypothèse sera discutée plus avant dans le paragraphe consacré spécifiquement au débit. On peut également noter que le coefficient de variation VarCo, introduit par Dellwo (2006) pour répondre à la même problématique, est sans doute plus pertinent par rapport à des variations de débit. Les informations prises en compte ne sont cependant pas de nature fondamentalement différente

groupements de quatre ensembles {consonne + voyelle} se reproduisent régulièrement. La langue C ne présente enfin aucune régularité apparente.

Ces trois schémas, aussi hypothétiques soient-ils, illustrent cependant les limites à ne considérer que la distribution statistique des paramètres de durées vocaliques et intervocaliques. Conscientes de ce défaut important, Grabe & Low (2002) proposèrent alors de s'intéresser aux propriétés temporelles des enchaînements d'intervalles de même nature grâce aux *Pairwise Variability Index* (décliné en *PVIV* et *PVIC* pour les intervalles vocaliques et les intervalles inter-vocaliques), précédemment introduits dans Low, Grabe & Nolan, (2000). En traitant les intervalles (intervocaliques ou vocaliques) deux à deux, elle permettait d'une part de prendre en compte une échelle temporelle plus étendue (de l'ordre du pied) et d'autre part de procéder intrinsèquement à une normalisation des durées par rapport au débit. La définition de ces *PVI* normalisés (ou *nPVI*) est donnée par la formule suivante :

$$nPVI = 100 \times \left[\frac{1}{(m-1)} \sum_{k=1}^{m-1} \left(\frac{2}{(d_k + d_{k+1})} \times |d_k - d_{k+1}| \right) \right]$$

où m représente le nombre total d'intervalles du type considéré (respectivement vocalique ou intervocalique) et d_k la durée de l'intervalle numéro k.

Dans le cas théorique des trois motifs rythmiques présentés ci-dessus, cette approche se révèle partiellement discriminante, même si elle a également pour conséquence de neutraliser certaines différences, potentiellement intéressantes (les langues B et C de la Figure 6 ont des valeurs de nPVIV et nPVIC identiques). Par ailleurs, puisque les valeurs absolues des différences de durées prises deux à deux sont calculées, certaines asymétries d'organisation, telles que des alternances trochaïques *vs.* iambiques, sont invisibles.

Malgré ces différentes limites, d'ailleurs identifiées dès le début par leurs auteurs respectifs, ces deux approches ont joué un rôle majeur dans les études translinguistiques menées sur le rythme depuis la fin des années 1990, en mettant l'accent sur des niveaux de description distincts : Ramus et ses collègues ont démontré la pertinence des caractéristiques moyennes des durées des constituants rythmiques « atomiques » que sont les intervalles intervocaliques et vocaliques, tandis que Grabe et Low ont montré qu'une mesure différentielle calculée à partir de séquences de ces constituants était également porteuse d'information sur l'organisation rythmique, à une échelle plus étendue. De part sa nature, l'espace de description proposé par Grabe et Low peut être vu comme un espace dérivé, au sens mathématique, de l'espace proposé par Ramus. Si l'on rapproche ces deux méthodologies, on dispose donc d'une analyse multi-niveau mêlant informations statiques et dynamiques, même si une certaine redondance a été mise en évidence entre ces niveaux d'informations (voir Hamdi, 2007 et Ferragne, 2008) et si certaines questions restent en suspens. En particulier, il est intriguant de constater que ces approches portent sur des constituants de durées inférieures à ce que l'on appelle généralement des constituants rythmiques, more ou syllabe en particulier. Les espaces de description multidimensionnels proposés par Ramus, Nespor &

Mehler, et Grabe et Low pourraient par conséquent être qualifiés de « sub-atomiques ». Une limitation porte alors sur le traitement indépendant réalisé respectivement sur les intervalles intervocaliques et vocaliques ; en effet, toute interaction entre durées consonantiques et vocaliques est neutralisée. Par exemple, alors que la langue B illustrant cette discussion présente une contrainte d'égalité des durées intervocaliques et vocaliques au sein des groupements {consonnes + voyelles}, la langue C ne présente pas ce type de dépendance. Pourtant les langues B et C présentent les mêmes valeurs de $\%V$, ΔV, ΔC, *nPVIV* et *nPVIC* montrant ainsi que ces mesures peuvent être insensibles à des régularités intervenant à l'intérieur de groupements de consonnes et de voyelles.

✦ Une unité rythmique pseudo-syllabique

Ces constatations, ainsi que la volonté de franchir une étape supplémentaire en extrayant automatiquement les frontières des intervalles considérés (et non à la main, comme dans les approches rythmiques présentées précédemment), nous ont poussé à proposer de modéliser une unité d'une taille proche des syllabes, ou du moins des intervalles {consonnes + voyelles} mentionnés ci-dessus. Ces unités, que nous avons nommées sans grande originalité pseudo-syllabes, sont identifiées à partir des traitements de segmentation et de détection des noyaux vocaliques mis en œuvre précédemment dans le cadre de la modélisation des systèmes vocaliques. À partir des séquences de segments C et V identifiés, des motifs de type C^nV^m (où n et m sont des nombres entiers, éventuellement nuls) sont formés. Si l'on reprend l'exemple donné Figure 4, la séquence de segments CCVVCCVCVCCCVCVCCC génère 6 pseudo-syllabes :
CCVV.CCV.CV.CCCV.CV.CCC. Cette procédure suscite plusieurs remarques ; rappelons tout d'abord que les segments utilisés ne correspondent pas réellement à des phonèmes car ils sont de nature infra-phonémique. Ils sont cependant liés au nombre réel de phonèmes présents et fournissent ainsi un traitement indépendant de la langue, même s'il est évidemment biaisé par rapport à une segmentation syllabique manuelle et que cette limite doit rester à l'esprit dans la suite de cette discussion.

Ensuite, il n'est pas évident que le choix d'un squelette rythmique de type *syllabe ouverte* soit optimal. Même si peu d'études typologiques ont porté sur la distribution des syllabes ouvertes et fermées dans des *corpus* de langues du monde, l'existence de langues présentant une prédominance de syllabes fermées est bien attestée. La seule réelle étude à visée typologique (Rousset, 2004) confirme que, à partir d'une asymétrie dans la distribution d'usage des syllabes CV ou CVC dans leurs *lexiques* respectifs, deux langues vont avoir plutôt tendance à générer des syllabes complexes ouvertes ou, à l'inverse, fermées (par exemple CCV, CCCV plutôt que CCVC, CCVCC, CCCVCC, etc.). Cependant plusieurs arguments nous ont incités à adopter la syllabe ouverte comme squelette. Tout d'abord, la présence universelle de la syllabe CV dans les langues du monde, argumentée par MacNeilage (1998) comme résultant d'un processus phylogénétique, garantit à l'algorithme une certaine universalité. On peut objecter qu'en lui-même, cet argument indique juste que l'algorithme devrait autoriser la segmentation du flux sonore en pseudo-syllabes CV,

mais il n'interdit en rien la prise en compte de syllabes fermées. Malheureusement, si l'on souhaite à la fois autoriser des structures ouvertes et fermées, on se heurte à une difficulté supplémentaire. En effet, la segmentation du flux en syllabes ouvertes et fermées, par exemple .$C_1V_1C_2C_3.C_4V_2.C_5C_6V_3$. nécessite d'une part de placer une frontière syllabique entre des consonnes en position de coda et d'attaque (entre C_3 et C_4 dans l'exemple) et d'autre part d'identifier le cas où des groupes de segments consonantiques constituent des codas ou attaques complexes (cas de l'attaque complexe C_5C_6 dans l'exemple). Certains phonologues ne verraient éventuellement pas là de difficulté particulière, en faisant un usage systématique du principe de l'attaque maximale et du principe de sonorité. Cependant, plusieurs études montrent que ces principes sont difficilement applicables de manière absolue, même dans le contexte d'une seule langue : plusieurs échelles de sonorité sont en concurrence dans la littérature et une fois qu'une hiérarchie est adoptée pour une langue donnée, les exceptions sont relativement courantes et l'existence de consonnes ambisyllabiques (e.g. Content, Dumay & Frauenfelder, 2000) est difficilement compatible avec l'idée d'une segmentation syllabique linéaire (voir également une alternative à la sonorité proposée par J.J. Ohala, par exemple Ohala & Kawasaki-Fukumori, 1997). Dès lors, viser une réelle syllabation dans un contexte indépendant des langues, alors même que des différences liées à certaines classes de sons ont été identifiées (cas des liquides, des semi-voyelles, etc.) relève de la gageure (voir cependant la discussion en fin de section, p. 55).

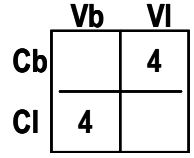
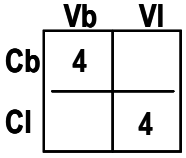
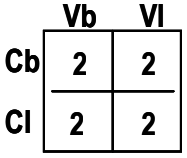
À partir des pseudo-syllabes ainsi définies, nous avons procédé à leur paramétrisation dans un espace multidimensionnel adapté à une modélisation statistique. Les deux premiers paramètres retenus sont D_C et D_V, correspondant aux durées respectives des groupements consonantique et vocalique de la pseudo-syllabe. Ils sont ainsi analogues aux primitives de durées employées par Ramus dans son approche, la différence résidant dans la prise en compte de leur co-occurrence au sein de la pseudo-syllabe. Nous avons enrichi cette description temporelle par la prise en compte de la complexité des intervalles consonantiques estimé par le nombre de segments. Le paramètre n de la pseudo-syllabe est en effet fortement lié à la structure syllabique des langues ou tout au moins à la présence de clusters consonantiques et à leur complexité dans les corpus étudiés. A l'inverse, l'indice m, caractérisant le nombre de segments vocaliques, a finalement été écarté du fait de son très faible pouvoir discriminant : en effet, dans la très grande majorité des cas, il est très faible (inférieur à 5) et sa modélisation statistique n'apporte guère d'information.

✦ Comparaison des approches présentées

Le Tableau 3 présente schématiquement une comparaison des approches mentionnées ci-dessus, basée sur les trois langues hypothétiques A, B et C introduites par la Figure 6. L'approche de Ramus, basée sur la distribution des durées vocaliques et intervocaliques, donne les mêmes indices pour les trois langues. L'approche différentielle de Grabe et Low discrimine quant à elle la langue A des langues B et C, en échouant cependant à distinguer ces dernières entre elles. L'approche pseudo-syllabique, est schématisée par une matrice de contingence des

consonnes (C) et voyelles (V), brèves (b) et longues (l), pour former les quatre types de pseudo-syllabes rencontrées dans ces langues. Ces matrices illustrent que les trois langues diffèrent par leur distribution de chaque type de pseudo-syllabe, permettant ainsi d'envisager de les distinguer automatiquement. Évidemment, cela ne signifie pas nécessairement que l'approche pseudo-syllabique est meilleure, en particulier parce que, telle qu'elle est présentée là, elle fait une hypothèse d'indépendance entre les pseudo-syllabes se succédant. Par conséquent, on peut tout aussi bien envisager des situations où les distributions de pseudo-syllabes sont identiques d'une langue à l'autre alors même que les enchaînements de pseudo-syllabes sont contraints de manière différentiée. L'approche de Grabe et Low, insensible aux frontières entre pseudo-syllabes, intègrerait alors ces contraintes dans l'estimation des PVI et serait alors plus efficace que l'approche pseudo-syllabique.

Tableau 3 – Comparaison des trois approches d'évaluation du rythme sur les langues décrites dans la Figure 6. Vb et Cb schématisent les voyelles et consonnes brèves, tandis que Vl et Cl schématisent leurs contreparties longues

MÉTHODOLOGIE	LANGUE A	LANGUE B	LANGUE C
Ramus	%*V* : 50 Δ*V* : 0,53 Δ*C* : 0,53	%*V* : 50 Δ*V* : 0,53 Δ*C* : 0,53	%*V* : 50 Δ*V* : 0,53 Δ*C* : 0,53
Grabe & Low	nPVIv : 66,7 nPVIc : 66,7	nPVIv : 38,1 nPVIc : 38,1	nPVIv : 38,1 nPVIc : 38,1
Farinas & Pellegrino			

✦ Pseudo-syllabe et identification des langues

Outre la décorrélation entre les indices issus des intervalles vocaliques et intervocaliques, une autre limite commune aux approches de Ramus et de Grabe réside dans l'utilisation de statistiques d'ordre faible : ordre 1 pour Grabe (moyenne des durées différentielles normalisées) et ordre 2 pour Ramus (écart-type des durées des intervalles ; %*V* étant quant à lui calculé sur la totalité de l'énoncé). Pourtant, il est probable que les distributions statistiques des primitives soient relativement complexes. En particulier s'agissant de durées, l'hypothèse d'une distribution normale semble peu étayée, comme cela a depuis été confirmé par Duarde *et al.*, (2001). Par conséquent, nous avons opté pour une modélisation des pseudo-syllabes sous forme de mélanges de lois gaussiennes dans l'espace tridimensionnel (D_C, D_V, n), modèles adaptés à l'estimation de distributions de complexité variable même s'il s'agit là d'une entorse à l'orthodoxie statistique, puisqu'une loi discrète aurait dû être employée pour le paramètre n.

Cette méthode a été validée dans le cadre de l'identification des langues pour de la parole lue. Par la suite, nous l'avons également testée sur des enregistrements

de parole spontanée (corpus OGI MLTS) ainsi que dans le cadre de discrimination dialectale des parlers arabes. La plupart de ces travaux sont présentés en détail dans les thèses de Jérôme Farinas (2002) et Jean-Luc Rouas (2005).

L'utilisation d'un corpus de parole lue multi-locuteur pour la validation initiale a été dictée par un nécessaire compromis entre, d'une part les applications d'identification visées (exploitant des modèles génériques et indépendants du locuteur) et d'autre part la difficulté manifeste à identifier des primitives du rythme robustes (incitant à limiter la variabilité inter-individuelle). Notre choix s'est porté sur le corpus MULTEXT, développé à partir du sous-corpus dit *Few Talker Set* du corpus multilingue EUROM1. Aux cinq langues de l'ensemble initial (allemand, anglais, espagnol, français et italien) se sont ajoutées par la suite le chinois mandarin et le japonais. Ce corpus est malheureusement d'une taille très limitée et pour un certain nombre des expériences, nous avons dû procéder à une approche par validation croisée. Cette méthode permet de disposer d'un ensemble plus important pour l'estimation des paramètres, mais elle limite également la portée des résultats obtenus (le lecteur pourra se reporter aux annexes C.3 et C.4 pour plus de détails).

Le constat principal réside dans la relative efficacité de l'approche pseudo-syllabique. Alors même que l'espace paramétrique employé est extrêmement rudimentaire (avec à peine trois dimensions) et sans commune mesure avec une description cepstrale par exemple, les résultats obtenus (67 ± 8 % d'identification correcte) sont significativement et largement supérieurs au hasard (cf. Tableau 4, expériences menées sans validation croisée). Il s'agissait là de la première approche purement rythmique suffisamment efficace pour envisager de la tester en identification sur un ensemble de langues plutôt qu'en simple discrimination deux à deux, même si les résultats restent en deçà de ceux obtenus par des approches acoustico-phonétiques : à titre indicatif, les résultats de modèles globaux segmentaux similaires à ceux introduits dans la section précédente atteignent 88 ± 5 % d'identification correcte sur la même tâche (Rouas *et al.*, 2005). Parallèlement à ces expériences, Rouas et Farinas ont procédé à une comparaison directe de la modélisation pseudo-syllabique avec les paramètres proposés par Ramus et par Grabe et Low, non plus à partir de l'étiquetage manuel mais à partir de la même segmentation automatique, avec comme résultats respectifs 44 ± 8 % et 37 ± 8 % d'identification correcte. Il semble donc que les paramètres globaux proposés dans ces deux approches s'avèrent relativement peu robustes dès lors que de la variabilité inter-locuteur augmente.

Tableau 4 – Matrice de confusion (en nombres de fichiers de 20 secondes) obtenue avec un modèle pseudo-syllabique (8 composantes gaussiennes par langue). Le taux d'identification global est de 67 ± 8 %.

Modèle / Langue	ANGLAIS	ALLEMAND	MANDARIN	ITALIEN	FRANÇAIS	ESPAGNOL	JAPONAIS
ANGLAIS	16	1	1	1	0	1	0
ALLEMAND	5	14	1	0	0	0	0
MANDARIN	4	3	11	1	0	0	1
ITALIEN	6	1	1	11	0	0	1
FRANÇAIS	0	0	0	0	19	0	0
ESPAGNOL	0	0	0	2	8	6	4
JAPONAIS	2	0	0	2	0	0	16

Si l'on regarde plus en détail les résultats présentés dans le Tableau 4, certains regroupements se dégagent : l'anglais et l'allemand, traditionnellement classés comme langues accentuelles, se distinguent bien du français et de l'espagnol, langues dites syllabiques (une seule confusion se produit entre ces deux groupes, sur 79 fichiers). Le japonais, seule langue moraïque de l'échantillon, est relativement peu confondu avec d'autres langues (80 % d'identification correcte). Les informations les plus intéressantes sont liées au mandarin et à l'italien : le rythme du mandarin a fait l'objet d'un débat assez nourri, mais plusieurs études le rapprochent des langues à tendance accentuelle (Komatsu, Arai & Sugawara, 2004) et dans notre expérience, il est assez largement confondu avec le groupe anglais + allemand, ce qui le place plutôt à proximité des langues à tendance accentuelle. Les résultats obtenus avec l'italien sont quant à eux plus difficiles à expliquer à première vue puisqu'il est confondu dans 30 % des cas avec l'anglais. Deux explications sont *a priori* possibles pour ce résultat : soit il s'agit d'une réalité acoustique, correspondant à une proximité effective entre italien et anglais, soit il s'agit d'un artefact lié aux modèles employés. En l'occurrence, ces modèles tentent de capturer des régularités concernant les distributions de durée des pseudo-syllabes. Par conséquent, ils sont sensibles à tout type de variation de durées non liés à la langue. En particulier, les variations de débits liées au locuteur ont un impact sur la précision des modèles estimés.

Si l'on estime ces variations au sein du corpus (en termes d'écart-type du débit syllabique), on constate en fait qu'elles s'étendent de 0,33 pour le français à 0,64 pour l'italien (voir Tableau 5) et il existe une corrélation négative significative entre ces variations et le taux d'identification correcte par le modèle pseudo-syllabique (corrélation de rang de Spearman ; $\rho = -0{,}77$, $p = 0{,}05$). Il semble donc probable que les différences de performance observées d'une langue à l'autre (et en particulier pour l'italien) soient en partie dues à l'absence de prise en compte des variations de débit dans le modèle proposé.

Tableau 5 – Taux d'identification correcte et débit syllabique (corpus MULTEXT ; valeurs moyennes et écarts-types). Les langues sont ordonnées par taux d'identification décroissants.

	FRANÇAIS	ANGLAIS	JAPONAIS	ALLEMAND	MANDARIN	ITALIEN	ESPAGNOL
DÉBIT MOYEN	6,37	5,39	5,29	5,06	5,05	5,71	6,94
ÉCART-TYPE (DÉBIT)	0,33	0,52	0,51	0,45	0,52	0,64	0,59
TAUX D'IDENTIFICATION	100	80	80	70	55	55	30

✦ *Extensions du modèle rythmique*

Les travaux présentés ci-dessus valident qu'il est possible d'extraire automatiquement des informations rythmiques dans une perspective d'identification des langues. Ils soulèvent cependant de nombreuses questions qui mériteront d'être approfondies.

Tout d'abord, nous avons évoqué plus haut la difficulté à segmenter le flux de parole en pseudo-syllabes fermées. Deux approches sont cependant envisageables pour améliorer ce processus. La première consiste à appliquer une segmentation de chaque intervalle intervocalique en coda et attaque en fonction de la sonorité des segments constituants. Cette sonorité pourrait être évaluée automatiquement en utilisant par exemple l'algorithme proposé par Galves et al, (2002). La frontière syllabique serait alors systématiquement placée en accord avec le principe de sonorité (i.e. devant le segment de sonorité minimale) et, dans le cas où deux minima identiques seraient atteints, le premier serait sélectionné comme début de l'attaque, en accord avec le principe d'attaque maximale. Cette méthode permettrait un traitement systématique et indépendant de la langue, au risque cependant de produire des attaques consonantiques illégales : par exemple, la pseudo-syllabation du mot français spectacle donnerait /spɛ.ktakl/ alors que l'attaque /kt/ est considérée comme illégale en français. Une autre approche consisterait à conjuguer l'approche actuelle (basée sur un squelette C^nV) avec une approche symétrique, basée sur un squelette VC^m, soit avec un unique modèle C^nVC^m où chaque intervalle intervocalique participerait intégralement à deux pseudo-syllabes, soit avec une fusion de deux modèles simples C^nV et VC^m. Ce dernier formalisme se rapproche alors clairement de la notion de demi-syllabe, telle qu'elle est développée dans Clements (1990:303). Le Tableau 6 résume ces différentes propositions en s'appuyant sur l'exemple donné Figure 7. Cette figure représente le résultat de la segmentation de l'énoncé du mot *capter* (locuteur masculin). La séquence segmentale est CCCVCCCCCVCCC une fois les pauses éliminées. Les « x » placés sur certains segments intervocaliques localisent les minima probables de la fonction de sonorité.

Bien évidemment, ces propositions demandent à être évaluées, mais on peut déjà anticiper certaines difficultés liées à l'augmentation du nombre de paramètres des modèles si la quantité de données disponibles pour leur apprentissage reste aussi limitée qu'actuellement.

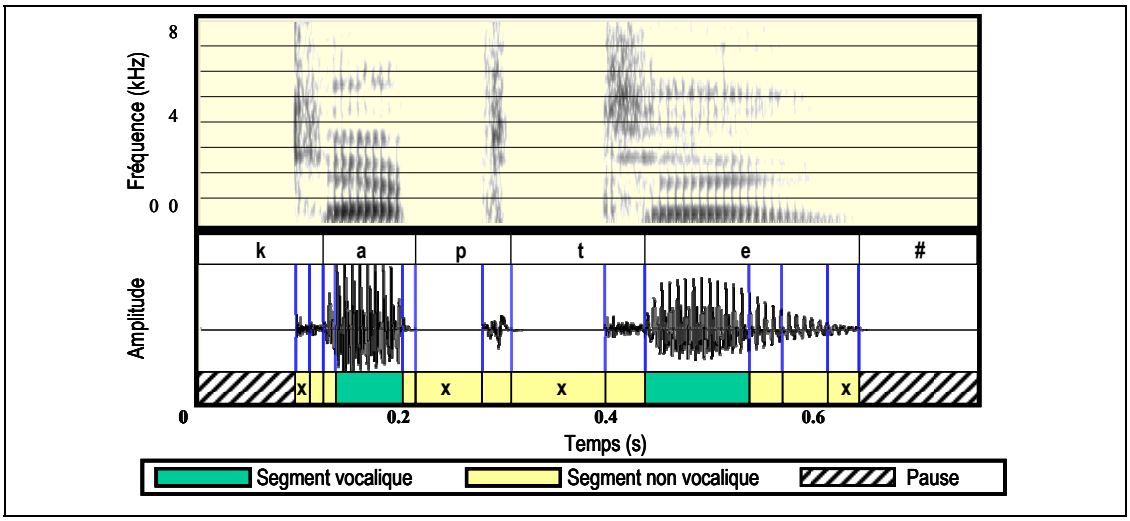


Figure 7 – Exemple de segmentation (mot « capter », locuteur masculin). Les 'x' localisent les minima probables de sonorités.

Tableau 6 – Proposition d'amélioration de la segmentation pseudo-syllabique. Les symboles en gras indiquent les segments redondants.

MÉTHODE PROPOSÉE	PSEUDO-SYLLABES PRISES EN COMPTE
Méthode actuelle	CCCV + CCCCCV + CCC
Segmentation selon sonorité	CCCVC + CCCC + CC + C
Redondance I (CnVCm)	CCC**VCCCCC** + **CCCCCV**CCC
Redondance II (CnV + VCm)	**CCCV.CCCCCV.CCC** + **CCC.VCCCCC.VCCC**

Nous évoquerons maintenant deux extensions supplémentaires. Quelle que soit l'approche de segmentation retenue, la modélisation par mélange de lois gaussiennes utilisée permet de caractériser les relations internes aux pseudo-syllabes mais néglige totalement les caractéristiques ou contraintes pesant sur les enchaînements de pseudo-syllabes. Pourtant ces contraintes peuvent être pertinentes, dès lors que le rythme des langues étudiées présente une structuration à plus grande échelle (organisation en pieds, etc.). Cela incite donc à développer des modèles spécifiques, de type grammaires *n*-grammes par exemple, pour intégrer cette information. Parallèlement la prise en compte d'informations acoustiques liées à l'accentuation et plus généralement à la dimension intonative ou tonale est évidemment prometteuse. Une première approche a consisté à intégrer cette dimension directement au niveau des pseudo-syllabes, en augmentant ainsi la dimension de l'espace de représentation (approche pseudo-syllabique accentuelle). Nous avons ensuite entrepris, à l'initiative de Jean-Luc Rouas, une modélisation intonative multi-échelle par l'intermédiaire de grammaires statistiques estimant les probabilités d'enchaînement de constituants intonatifs élémentaires définis à différentes échelles : segments, pseudo-syllabes et intervalles monotones au sens intonatif (approche adaptée de Adami *et al.*, (2003) et inspirée de Fujisaki (2003)). De nombreuses expériences sont détaillées dans Rouas (2005 ; 2007) et les Tableaux ci-dessous illustrent certaines des conclusions. Le modèle A correspond au modèle multigaussien pseudo-syllabique détaillé ci-dessus. Le modèle B est un modèle également multigaussien où chaque syllabe est décrite dans un espace

prosodique dérivé de mesures de Fo et d'énergie (pour les détails, cf. Rouas, 2005). Le modèle C est composé de deux modèles *n*-multigramme des enchaînements d'intervalles monotones estimés aux niveaux des macro- et micro- variations de Fo et de l'énergie. La fusion est une addition pondérée de vraisemblances issues des modèles considérés ; faute de corpus dédié au développement et à l'ajustement des modèles, les coefficients de pondération sont estimés sur les données d'apprentissage. Étant donnée la faible taille du corpus de test, les intervalles de confiance des performances sont élevés, et seules des tendances peuvent être relevées, les différences entre les scores n'étant que marginalement significatives.

Tableau 7 – Bilan des expériences en identification prosodique des langues (d'après Rouas, 2005). Les performances sont les taux d'identification correcte (avec leurs intervalles de confiance).

IDENTIFICATION PROSODIQUE DES LANGUES (7 LANGUES - MULTEXT)				
CODE	DIMENSION MODÉLISÉE	PRIMITIVES	TYPE DE MODÈLE	PERFORMANCE (%)
A	Pseudo-syllabique (durées)	Durées et complexité	Multigaussien	67 ± 8
B	Pseudo-syllabique (accent)	Fo	n-multigramme	52 ± 8
C	Intonatif (intervalles monotones)	Fo et Énergie	n-multigramme	70 ± 8
Fusion	A + B	-		66 ± 8
Fusion	A + C	-		**75 ± 7**

Le Tableau 7 illustre que l'information intonative est dans une certaine mesure pertinente pour distinguer les langues puisque les modèles B et C obtiennent des scores significativement supérieurs au hasard. Si l'on compare les résultats des fusions A + B et A + C, il semble que l'échelle d'intégration soit importante ; en effet, à l'échelle de la pseudo-syllabe (fusion A + B), on ne constate aucune amélioration par rapport au seul modèle A où les informations intonatives sont absentes. À l'inverse, l'échelle temporelle élaborée à partir du contenu prosodique (par extraction des segments de variation monotone) semble plus adaptée puisqu'elle permet d'atteindre 75 ± 7 % d'identification correcte (en fusion avec le modèle A). Cette intégration de deux échelles, *a priori* distinctes, présente donc un potentiel intéressant pour l'identification des langues.

Ce résultat soulève pourtant une question supplémentaire car il est particulièrement difficile d'évaluer le type d'information que le modèle C intègre : s'agit-il de prosodie lexicale (accent ou ton) ou à l'échelle de la phrase ? En d'autres termes, la différence de performances entre les modèles B et C est-elle due à la différence de pertinence entre les deux échelles ou à la prise en compte d'informations différentes ? Un élément de réponse est apporté par les résultats présentés dans le Tableau 8 et obtenus avec les mêmes approches, mais pour lesquelles on a regroupé les 7 langues en trois groupes en fonction de leurs tendances rythmiques. Un premier groupe rassemble anglais, allemand et mandarin, langues à tendance accentuelle (ou *stress-based*, selon la terminologie proposée par Laver, 1994) ; français, espagnol et italien forment le groupe des langues à tendance syllabique tandis que le japonais est le seul représentant du groupe des langues à tendance moraïque.

Tableau 8 – Bilan des expériences en identification prosodique des groupes de langues (d'après Rouas, 2005). Les performances sont les taux d'identification correcte (avec leurs intervalles de confiance).

IDENTIFICATION PROSODIQUE DE GROUPES DE LANGUES (3 GROUPES - MULTEXT)				
CODE	**DIMENSION MODÉLISÉE**	**PRIMITIVES**	**TYPE DE MODÈLE**	**PERFORMANCE (%)**
A	Pseudo-syllabique (durées)	Durées et complexité	Multigaussien	86 ± 7
B	Pseudo-syllabique (accent)	Fo	n-multigramme	88 ± 6
C	Intonatif (intervalles monotones)	Fo et Énergie	n-multigramme	70 ± 8
Fusion	A + B	-		**91 ± 5**
Fusion	A + C	-		86 ± 7

On constate que le modèle B surpasse le modèle C dans cette tâche, indiquant que les informations prosodiques intégrées au niveau de la pseudo-syllabe (modèle B) sont plus en accord avec la notion de groupes de langues rythmiques que les indices modélisés à plus grande échelle. Cet effet a tendance à se confirmer lors de la fusion des modèles (A + B vs. A + C). Ainsi la fusion A+ B aboutit-elle à la matrice de confusion suivante :

Tableau 9 – Matrice de confusion intergroupe obtenue par la fusion des modèles A et B (d'après Rouas, 2005). Le taux d'identification correct est de 91 ± 5 %.

MODÈLE / LANGUE	« TENDANCE ACCENTUELLE »	« TENDANCE SYLLABIQUE »	« TENDANCE MORAÏQUE »
« TENDANCE ACCENTUELLE »	59	1	-
« TENDANCE SYLLABIQUE »	6	49	4
« TENDANCE MORAÏQUE »	-	1	19

✦ *Conclusion sur la modélisation prosodique*

À l'heure de dresser un bilan, la raison impose de rester particulièrement prudent. En effet, le peu de langues considérées génère un risque relativement important de sur-interpréter des résultats dus en fait au hasard et parallèlement, le faible nombre de fichiers de tests utilisés cause une certaine instabilité des modèles statistiques eux-mêmes (en particulier, moins les données d'apprentissage sont nombreuses, plus le modèle est sensible à la phase d'initiation). Dès lors il semble évident que l'avenir de ces modèles prosodiques ne peut se concevoir qu'en les confrontant à des données plus nombreuses issues par exemple des corpus collectés dans le cadre des évaluations organisées par NIST. Cela implique en particulier d'adapter les paramètres extraits et les modèles développés à la parole spontanée et force est de constater que cette problématique est pour l'instant inexplorée et qu'elle s'annonce extrêmement ardue… Pour ne citer qu'un paramètre clef, la prise en compte du débit relève à elle seule de travaux exploratoires, même si les expériences menées par Wagner et Dellwo sur l'invariance des paramètres rythmiques face aux variations de débit peuvent apporter des informations cruciales sur les relations entretenues par le rythme et le débit. Enfin, il faut également reconnaître que l'existence de langues pour lesquelles les noyaux syllabiques sont des segments non

vocaliques, voire non voisés, propose un défi supplémentaire à ce type d'approche[20], tout comme les langues agglutinantes où des morphèmes vocaliques sont suffixés, générant de longues séquences sans consonne (par exemple le gokana, cf. Hyman, 1983).

2.4. TRAVAUX SUR L'IDENTIFICATION PERCEPTUELLE

Documents de référence : Annexes C.5 et C.6

Barkat-Defradas M., I. Vasilescu & Pellegrino, F. 2003. « Stratégies perceptuelles et identification automatique des langues: application au continuum dialectal arabe », *Revue Parole*, Ed. Univ. de Mons-Hainaut, Belgique, n° 25, pp. 1-44.

Meyer, J., Pellegrino, F., Barkat, M. & Meunier, F. 2003. "The notion of perceptual distance: the case of Afroasiatic languages", *proc. of XVth ICPhS*, Barcelona, Spain.

Si la plupart des travaux menés en identification automatique des langues visent principalement au développement de modèles performants, les objectifs des études menées depuis trente ans sur la capacité humaine à identifier des langues étrangères sont plus variables. Ils s'étendent de la détermination des informations accessibles aux nourrissons dans le signal de parole à l'établissement de performances de référence pour évaluer des systèmes automatiques. En parallèle, le fait de travailler avec des stimuli aussi complexes que la parole et des sujets humains génèrent des difficultés importantes pour interpréter les résultats : quel est l'impact de l'histoire linguistique des sujets ? Quelles informations sont saillantes et extraites par les auditeurs pour effectuer la tâche demandée ? Ces deux questions amènent à s'interroger sur la variabilité du traitement des langues étrangères, mais y répondre n'est pas aisé, et cela a mené certains chercheurs à contrôler du mieux possible ces paramètres complexes, en particulier en utilisant des signaux altérés pour masquer (ou supprimer) certains types d'information au profit d'autres. Un intérêt fondamental de ces travaux réside, à notre avis, dans la fenêtre qu'ils ouvrent sur le traitement cognitif des sons du langage humain, alors même que les informations lexicales et morpho-syntaxiques sont inaccessibles (dans le cas des langues totalement inconnues). En effet, ces protocoles soulèvent souvent des questions extrêmement intéressantes mais malheureusement, ils peinent assez régulièrement à y répondre (paragraphe 2.4.1). Comme d'autres auparavant, nous nous sommes intéressés à ces problématiques, en particulier en travaillant sur les domaines linguistiques délimités que sont les langues romanes et les langues afro-asiatiques (section 2.4.2).

[20] L'utilisation de la fricative /s/ comme noyau de la syllabe /pst/, mentionné par Saussure (Cours de linguistique générale, page 88) n'est qu'un exemple d'un phénomène qui aura largement contribué à la pensée moderne dans le domaine de la phonologie. Voir par exemple les travaux sur le berbère tashelhit (e.g. Ridouane, (2002) et les travaux suivants) ; sur le nuxálk, appelé aussi bella coola (Bagemihl, 1991) ou encore sur le lendu (Demolin, 2002).

2.4.1. Des protocoles expérimentaux pour tester quoi ?

Identifier des langues étrangères n'est pas une activité classique pour de nombreux locuteurs adultes. Certains y verront un moyen d'exercer leur curiosité intellectuelle tandis que d'autres parmi des adultes monolingues n'en percevront pas l'intérêt. Bien évidemment, la situation est toute autre dans des pays multilinguaux où de nombreux langages et dialectes peuvent être parlés, parfois sur des aires géographiques réduites. Ainsi un locuteur polyglotte est-il quotidiennement amené à correctement identifier l'identité de la langue parlée par ses interlocuteurs. On peut éventuellement considérer qu'il ne s'agit plus ici réellement d'identification des langues, dans le sens où les langues sont connues du sujet ; mais à l'inverse, pour des enfants en phase d'apprentissage du langage dans un tel contexte multilingue, il s'agit bien là de discriminer des flux de parole puis d'identifier des langues encore non maîtrisées, en particulier au cours des premières années de vie : pour ces jeunes apprenants, il est de première importance de distinguer efficacement la langue parlée pour acquérir la phonologie, la morphologie, la syntaxe et le lexique correspondant à chacune des langues de l'environnement.

Depuis une trentaine d'années, un nombre relativement important d'études a visé à évaluer les performances des sujets humains en identification des langues, qu'il s'agisse d'adultes ou d'enfants (se reporter à l'annexe C.5 ou à Komatsu, (2007), pour une revue de la question). Au-delà, des expériences ont également attesté dans une certaine mesure l'existence de capacités à discriminer les langues chez d'autres mammifères (primates non humains ; Ramus *et al.*, (2000) et chez les rats plus récemment (Toro, Trobalón & Sebastián-Gallés, 2003). Les caractéristiques segmentales (identités des traits et des phonèmes ; distribution de leurs fréquences d'occurrence), suprasegmentales (enchaînements phonotactiques, rythme et intonation) et évidemment de plus haut niveau (lexique, morpho-syntaxe) sont toutes apparues pertinentes à des degrés divers en fonction des conditions expérimentales (langues, types de stimuli et sujets). Il paraît intéressant de noter que le fait que les enregistrements soient produits selon une séquence temporelle normale ou inversée (*reversed speech*) a un impact sur les performances des sujets, humains ou pas (chez les rats par exemple : Toro, Trobalón & Sebastián-Gallés, 2005). Cet élément suscite des questions fondamentales sur les sons se situant aux frontières du langage humain.

Toute une série d'expériences maintenant célèbres ont prouvé que, dès ses premiers jours, un nourrisson est capable de distinguer entre la langue de sa mère et certaines langues étrangères attestant des différences suprasegmentales (cf. Ramus, (2002) pour une revue). En fonction des protocoles expérimentaux, une prééminence du rythme ou de l'intonation a été observée dans ce type de tâche.

S'il est relativement simple de maîtriser les facteurs influençant les nourrissons dans des expériences de discrimination des langues, il en va autrement pour des adultes au vu du nombre important de paramètres pouvant interagir. Parmi ces facteurs, l'identité de la langue maternelle ainsi que l'histoire linguistique (langues connues, langues familières, etc.) du sujet semblent être critiques bien que difficiles à quantifier. Depuis la fin des années 1960, d'assez nombreuses études ont

tenté de relever ce défi, avec des motivations variées. Les chercheurs en TAP visaient à étalonner des performances humaines pour les mettre en perspective des performances des systèmes d'IAL. De leur côté, les linguistes et les psycholinguistes cherchaient à affiner leur connaissance de la perception humaine et des représentations cognitives des langues. On a considéré plus récemment ce type d'expériences comme un moyen d'étudier le processus de construction du jugement global de similarité perceptive entre langues de manière à déterminer la saillance et le poids des différents indices linguistiques (caractéristiques segmentales et suprasegmentales en particulier). Les premières expériences d'Ohala et Gilbert (1981) ou celles plus récentes de Ramus mêlant parole naturelle et parole synthétique, ont prouvé que le rythme et l'intonation en particulier pouvaient permettre de discriminer ou d'identifier des langues même en l'absence d'indices segmentaux, pour peu que certaines informations relatives aux structures syllabiques des langues soient conservées.

D'un point de vue général, toutes ces expériences ont mis en évidence la capacité notable des sujets humains à identifier des langues étrangères inconnues après une courte phase d'entraînement. Par exemple, Muthusamy, Jain & Cole, (1994) indique que lorsqu'on présente des extraits de 6 secondes issus de 9 langues étrangères à des sujets naïfs, locuteurs d'anglais américain, ils atteignent un score de 54,2 % d'identification correcte, largement supérieur au hasard. Les informations mentionnées par les sujets comme aidant à l'identification se regroupaient en indices segmentaux (mode et lieu d'articulation des consonnes, présence de voyelles nasales, etc.) ; suprasegmentaux (rythme, intonation et tons) et pseudo-lexicaux (récurrence de certains mots, groupes de mots ou « pseudo-mots »). En fait, si l'on regarde les résultats dans le détail, les performances varient très sensiblement, allant de 86,4 % d'identification pour l'espagnol à 26,7 % pour le coréen. À propos des sujets ayant participé à l'expérience, on peut faire le pari qu'en moyenne, même s'ils ne parlaient pas espagnol, ils avaient précédemment été plus exposés à cette langue qu'au coréen. Par conséquent, les taux d'identification eux-mêmes doivent être considérés avec prudence et l'influence de nombreux paramètres restait à quantifier (nombre de langues familières des sujets, durée de l'apprentissage, etc.).

Malheureusement, aucune des études publiées en identification perceptuelle des langues (ou des dialectes) n'est totalement dépourvue de lacunes, qu'il s'agisse des matériaux employés (trop contraints ou à l'inverse trop peu contrôlés), des sujets (populations trop peu homogènes) ou de l'analyse des résultats (nivellement de la variabilité ou surinterprétation des résultats). Par conséquent on peut légitimement s'interroger sur la pertinence de telles études même si on peut espérer que les recherches progressent dans les années à venir, en particulier en identifiant les paramètres principaux intervenant dans ce processus. Cela nécessite cependant un contrôle très strict des matériaux linguistiques utilisés, au risque de s'éloigner encore plus d'une tâche écologique pour les sujets. Sur ce plan, plusieurs études menées en particulier à l'université d'Ohio ont tenté de quantifier les différents effets en jeu de manière ingénieuse. Au cours d'une série d'expériences (Stockmal, Muljani & Bond, 1996; Stockmal, Moates & Bond, 2000; Bond & Stockmal, 2002) cette équipe a étudié plusieurs facteurs sociolinguistiques (origine géographique des

locuteurs, langues connues des sujets, etc.) et linguistiques (caractéristiques rythmiques) importants. Lors des dernières expériences, elles ont mis en place un protocole où les locuteurs étaient bilingues et s'exprimaient dans différentes langues de la famille niger-congo, langues inconnues des sujets du test (Stockmal & Bond, 2003).

En considérant que le contrôle de ces paramètres était fondamental, nous avons également mené des expériences sur l'identification des langues par des sujets humains, de manière à évaluer l'influence de la langue maternelle de sujets en particulier, ou l'impact de la présence (ou absence) de certains traits phonétiques potentiellement saillants. Pour limiter les interférences avec d'autres paramètres, nous nous sommes focalisés dans ces expériences sur des langues issues de familles linguistiques délimitées, et non sur un échantillon de langues disparates.

2.4.2. Du contrôle des protocoles à l'interprétation des résultats

Ces travaux ont été entrepris d'une part dans le cadre de la thèse de Ioana Vasilescu (identification de langues romanes) et d'autre part pendant le stage de DEA de Julien Meyer, en collaboration avec Melissa Barkat-Defradas, Naima Louali et Fanny Meunier.

✦ *Les langues romanes*

Les expériences menées impliquaient cinq langues romanes (espagnol, français, italien, portugais et roumain) ; quatre groupes de sujets de langues maternelles différentes (sujets américains, français, japonais et roumains) et deux protocoles expérimentaux (discrimination de type identique/différent et évaluation perceptive de la similarité entre langues ; à partir d'extraits de 6 secondes en phase de test).

L'objectif de cette étude était d'utiliser un domaine linguistique bien balisé, pour lequel certaines caractéristiques potentiellement saillantes pouvaient être étudiées : en particulier, la variation des systèmes vocaliques des cinq langues à identifier (nombre de timbres mais également présence de voyelles nasales phonologiques en français et portugais) et la présence de gémination en italien.

Les groupes expérimentaux permettaient potentiellement de tester la familiarité entre la langue maternelle et le groupe de langues étudiées (interne : groupes français et roumain ; relativement familier : sujets américains ; très peu familiers : sujets japonais). De plus le contraste entre les groupes français et roumain pouvait également préciser l'influence de la langue maternelle.

Au final, et malgré la relative simplicité des tâches, une assez grande variabilité inter-individuelle a été mise en évidence et l'influence de la familiarité a été prépondérante. Cependant, une fois ce critère majeur établi, une certaine homogénéité des résultats permettait d'émettre l'hypothèse d'une saillance particulière des voyelles dans les tâches demandées. En effet, les différentes projections multidimensionnelles réalisées à partir des résultats mettaient généralement en évidence des regroupements cohérents avec la complexité des systèmes vocaliques (du point de vue du nombre de timbres phonologiques). Les

sujets français ont par ailleurs fait mention dans l'interview post-test de certains « traits » caractéristiques, comme la présence de diphtongues en portugais, de groupes consonantiques [kt] et [ʃt] en roumain, ou des fricatives [θ] et [x] en espagnol (voir Vasilescu, (2001) pour les détails).

Force est cependant de constater que même en contrôlant assez efficacement le matériau linguistique et le protocole expérimental, la portée de ce type d'expériences reste limitée du fait de la profusion d'indices potentiels offerts aux sujets.

✦ Les langues afro-asiatiques

Par la suite, nous avons entrepris une expérience de type discrimination identique/différent sur des langues, là encore issues d'un domaine linguistique délimité, mais présentant une diversité de traits phonétiques et suprasegmentaux plus importante. Dix langues et dialectes ont été initialement utilisées, dont huit de la famille afro-asiatique (amharique, arabe dialectal marocain et jordanien, langues berbères tarifit et touareg, hausa, hébreu et somali) et deux langues « intruses » exhibant certaines similitudes – la présence de consonnes d'arrière – avec les langues précédentes (arménien et turc). Les sujets de l'expérience étaient tous locuteurs natifs de français et l'analyse des résultats a été faite dans le cadre de la théorie de la détection (Macmillan & Creelman, 2005) et par positionnement multidimensionnel (*Multidimensional scaling*, Kruskal, 1978).

La tâche de discrimination des 100 stimuli présentés a été jugée comme très difficile par la majorité des sujets et la moitié d'entre eux (9 sur 18) a obtenu des résultats significativement identiques au hasard. De plus, les très faibles taux d'identification comme identiques obtenus lorsque deux extraits de hausa étaient présentés ensemble nous ont amené à écarter cette langue dans notre tentative d'interprétation[21].

L'hébreu, le somali et le turc étaient les langues les mieux individualisées des autres dans l'expérience, tandis que touareg et jordanien ont posé des problèmes importants aux sujets.

La Figure 8 présente les résultats de la représentation tridimensionnelle des résultats des 9 sujets (sur 18) pour lesquels les résultats étaient significatifs. Sans réellement pouvoir trancher, nous avions formulé plusieurs hypothèses compatibles avec les résultats observés (Meyer *et al.*, 2003). Le jordanien et le touareg, qui n'autorisent pas de groupes consonantiques, sont distincts des autres langues dans l'espace des dimensions 1 et 2. La dimension 2 peut être interprétée comme liée à la présence d'affriquées dans l'inventaire. La dimension 3 semble plus complexe, même si elle individualise bien le somali, seule langue présentant couramment des enchaînements vocaliques VV.

[21] Il est possible que les locuteurs employés pour le hausa aient présenté des différences dialectales très importantes, bien que cette langue ait été la seule langue à tons de l'échantillon.

Là encore, on atteint rapidement les limites de l'interprétation et la difficulté de la tâche ne permet pas d'envisager une analyse fine des processus mis en jeu lors de ces tâches. Le protocole utilisé (interface d'apprentissage réalisée en Flash par Julien Meyer) permettait cependant de comptabiliser le nombre d'écoutes des extraits d'apprentissage auxquelles les sujets avaient procédé : il est intéressant de noter que bien que ce nombre ait varié du simple au double entre les sujets, aucune corrélation n'a été observée entre durée d'apprentissage et performances de discrimination, suggérant que ce processus met en jeu des traitements *online* plutôt que des représentations en mémoire, du moins lorsque les langues en jeu sont parfaitement inconnues et partagent des traits similaires (consonnes d'arrière par exemple).

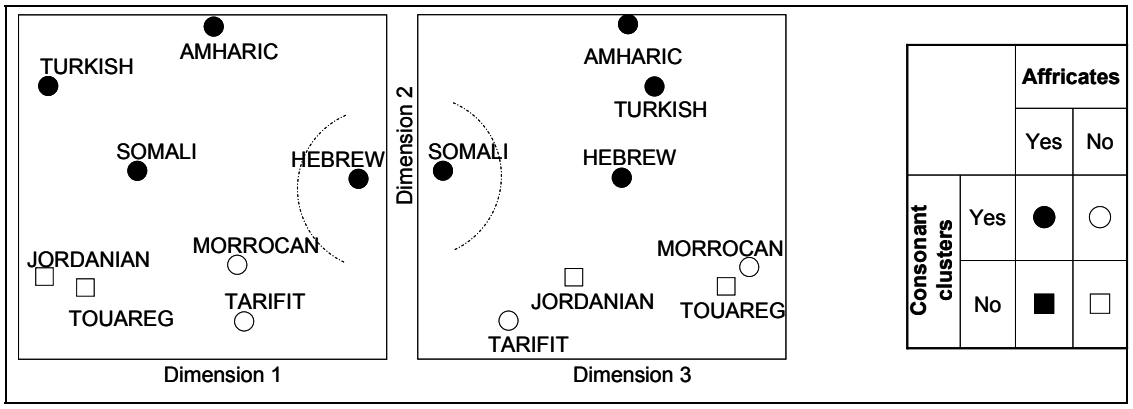


Figure 8 – Représentation multidimensionnelle des résultats des sujets dans la tâche d'identification des langues afro-asiatiques. À gauche, dans le plan des deux premières dimensions principales et au milieu dans le plan constitué des seconde et troisième dimensions. La matrice de contingence de droite fournit la légende des motifs employés (d'après Meyer *et al.*, 2003).

2.4.3. Discussion

Comme indiqué précédemment, les expériences menées depuis trente ans sur l'identification perceptuelle des langues ont éclairé certains facteurs importants du processus, même si notre vision n'en est encore qu'incomplète. Les réflexions suivantes portent sur un certain nombre des points en suspens.

✦ *Les protocoles de test*

Toutes les études dont nous avons connaissance débutent par une phase de familiarisation des sujets avec les langues testées, la durée de cette phase variant d'une étude à l'autre, sans que son influence ne soit clairement identifiée. Ensuite, deux types de protocoles susceptibles de mettre en œuvre des mécanismes différents sont couramment employés : une tâche de détection et une tâche d'identification. Les paragraphes suivants proposent une introduction à la discussion de ces mécanismes, de manière largement spéculative cependant.

1) Dans une tâche de *détection*, le sujet doit prendre une décision de type « identiques » ou « différents » pour deux stimuli sonores A et B de quelques

secondes[22]. Cette situation s'apparente à une tâche de vérification en traitement automatique et on pourrait imaginer qu'elle soit effectuée *online*, sans que les sujets ne fassent appel à des connaissances antérieures ni à des représentations mentales des langues en question : en caricaturant, si une simple distance acoustique entre les extraits est inférieure à un seuil *a priori*, les langues seraient identiques et, dans le cas contraire, différentes. Évidemment, la notion de distance n'est pas si simple et rien ne vient étayer l'hypothèse très improbable de l'existence d'un seuil *a priori* et donc indépendant des langues, même si les post-traitements issus de la théorie de la détection (mesures *d'* et *A'*, Pierce, 1980) font en un certain sens l'hypothèse de l'existence d'un tel seuil, propre à chaque sujet.

Selon nous, ce protocole peut générer deux types de stratégies chez les sujets. Dans la première, l'accent est mis sur les similitudes et les différences entre les extraits eux-mêmes plus que sur leur similitude avec les langues entendues lors de la phase de familiarisation : la décision repose alors sur la distance perceptuelle entre les extraits et sur l'existence ou l'absence d'indices saillants partagés. C'est la stratégie que semblent avoir employée les sujets de notre expérience sur les langues afro-asiatiques. Sur le plan linguistique, ce type de comportement peut permettre d'évaluer la saillance des différentes informations exploitées par les sujets de manière efficace si l'on est en mesure d'évaluer finement quelles informations étaient présentes dans les stimuli. Considérons par exemple, quatre langues combinant deux paramètres indépendants et binaires, disons la présence *vs.* l'absence de tons lexicaux d'une part (T_0 vs. T_1) et la présence *vs.* l'absence de consonnes géminées d'autre part (G_0 vs G_1), toutes choses étant égales par ailleurs. Les quatre langues sont alors définies comme T_0G_0, T_0G_1, T_1G_0 et T_1G_1. En fonction des scores de discrimination, on pourra étudier la saillance relative des deux paramètres et également évaluer dans quelle mesure ces différences sont additives dans le processus de discrimination (les langues T_0G_0 et T_1G_1 sont-elles mieux discriminées que T_0G_0 *vs.* T_0G_1 et T_0G_0 *vs.* T_1G_0 ?)[23].

Une seconde stratégie peut être d'implicitement *identifier* les langues de production des extraits proposés en faisant appel aux *représentations* établies lors de la phase de familiarisation (et au cours de sa propre expérience passée). Bien que l'on se trouve dans une tâche explicite de détection, le sujet réalise alors une tâche d'identification. En pratique, il est probable qu'en fonction des paires de langues testées, les sujets basculent d'une stratégie à l'autre, ce qui complique encore l'analyse de leurs performances. La Figure 9 présente schématiquement un

[22] Plusieurs variantes de cette tâche existent : chaque stimulus peut être constitué soit de deux extraits A et B (tâche identique/différent ou de quantification de la proximité entre eux) soit de trois extraits A, B et X ; dans ce cas la tâche est une tâche d'appariement de X à A ou B.

[23] La difficulté principale réside dans l'assertion *toutes choses étant égales par ailleurs*, particulièrement difficile à réaliser dans le cadre de comparaison de langues. On peut cependant imaginer que des stimuli s'approchant de cette condition peuvent être construits, par exemple dans des langues bantoues ou des langues romanes, avec des tâches d'identification de langues ou d'accents étrangers (e.g. Boula de Marëuil & Vieru-Dimulescu, 2006).

comportement possible des sujets lors d'une tâche de discrimination. A et B représentent les extraits entendus dans un stimulus et L_A et L_B correspondent aux langues de production de ces extraits, appartenant à l'ensemble $\mathcal{L} = \{L_1, ..., L_i, L_N\}$ des langues entendues lors de l'apprentissage. Si les langues L_A et L_B sont directement identifiées comme correspondant aux représentations des langues L*i* et L*j*, le processus se réduit à valider si $i = j$ ou pas, sans procéder spécialement à un calcul de distance (trait rouge sur la figure). Lorsqu'une seule des langue L_A ou L_B est identifiée, on est alors dans une procédure intermédiaire. Par exemple, même si seule L_A est reconnue comme étant L*i*, le sujet est généralement capable de définir un ensemble de langues $\overline{L}_B$ pour lesquelles il est sûr qu'il ne s'agit pas de L_B. Si L_i appartient à $\overline{L}_B$, une décision directe (L_A et L_B sont différentes) peut être prise. Dans le cas contraire, nous sommes dans un réel cas de détection et la similarité des stimuli doit être estimée et comparée à un seuil. Dans le dernier cas de figure, aucune des deux langues n'est identifiée, mais des ensembles $\overline{L}_A$ et $\overline{L}_B$ d'hypothèses écartées sont déterminés. À moins que l'union de ces deux ensembles ne couvre l'ensemble des hypothèses $\mathcal{L}$, excluant ainsi la possibilité que les deux extraits proviennent de la même langue, une tâche de détection est alors opérée.

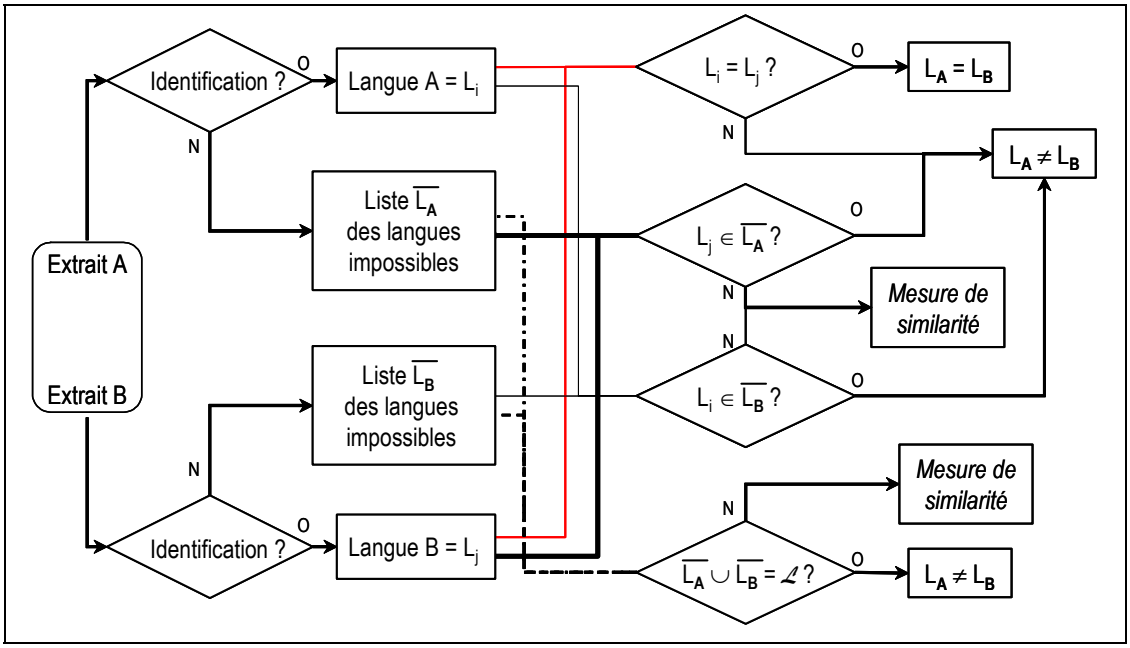


Figure 9 – Schéma d'un modèle de décision lors d'une tâche de discrimination de deux extraits A et B, produits dans des langues parmi un ensemble $\mathcal{L}$ de langues possibles.

Cet algorithme n'est évidemment qu'illustratif et spéculatif et aucune validation expérimentale n'en a à ce jour été recherchée[24]. On peut pourtant envisager des protocoles permettant aux sujets d'indiquer s'ils ont reconnu une des langues proposées lors de chaque test, de manière à évaluer dans quelle proportion il s'agit d'une réelle situation de détection. Il est également possible de présenter le même ensemble de stimuli à deux groupes de sujets, l'un ayant été exposé à une

[24] Une version légèrement différente avait été suggérée dans la thèse de I. Vasilescu.

phase de familiarisation avec les langues testées tandis que le second groupe n'aura aucune connaissance préalable sur ces langues. Ce second groupe ne pourra ainsi pas faire appel à des stratégies d'identification.

2) Le second protocole demande au sujet d'*identifier* la langue du stimulus parmi un certain nombre de langues candidates. Cette tâche fait explicitement référence aux représentations de ces langues établies lors de la phase de familiarisation et il s'agit d'une tâche de reconnaissance en ensemble fermé. Traditionnellement, la matrice de confusion obtenue est considérée comme une matrice de similarité et traitée, soit de manière unidimensionnelle (sous forme de dendrogramme par exemple), soit par projection dans un espace multidimensionnel. La détermination des métriques pertinentes pour ce type de tâche est loin d'être évidente, d'autant plus que les projections employées reposent souvent sur des hypothèses fortes. À titre d'exemple, les algorithmes classiques de projection multidimensionnelle procèdent par étapes, en partant de la paire de stimuli présentant la plus grande distance. Au fur et à mesure de l'ajout des autres stimuli dans la représentation, l'erreur de projection augmente (si l'on reste dans un espace de faibles dimensions). Au final, l'information obtenue porte plus sur les *grandes* distances que sur les *petites*, et l'interprétation des axes est souvent spéculative (comme on l'a vu dans le paragraphe précédent).

L'étude des processus cognitifs intervenant dans ces tâches relève de la psychologie cognitive, et on peut en particulier citer les travaux de R. Nosofsky, (1992) et Harnad, (1999) sur le sujet.

✧ Indices d'identification

Le constat est simple : la saillance perceptive des indices linguistiques est très variable d'un individu à l'autre, même si des tendances fortes sont liées à la langue maternelle et aux autres langues connues des sujets. La littérature regorge d'exemples illustrant cela ; qu'il s'agisse de difficultés de perception des contrastes de durée ou de l'accent par des sujets francophones (e.g. Peperkamp, Dupoux and Sebastián-Gallés, 1999) ; de la perception de voyelles épenthétiques par les sujets japonais (Dupoux *et al.*, 1999) ou encore des difficultés de perception et de catégorisation des consonnes en langues étrangères (e.g. Best, McRoberts and Sithole, 1988), il est clair que beaucoup reste encore à faire pour mieux comprendre les indices pris en compte par les sujets dans ces tâches. De plus, il nous semble indispensable de procéder à des analyses fines des extraits utilisés pour les tests. Il est en effet prévisible que, lorsqu'il s'agit d'enregistrements de quelques secondes, les traits potentiellement pertinents ne soient pas forcément représentés dans l'échantillon. On peut tout à fait imaginer que deux paires d'extraits impliquant les mêmes langues puissent, pour l'une, présenter des différentes très saillantes et pour l'autre, être caractérisées par l'absence de tels contrastes.

✧ Intégration et décision

Si ces expériences convergent pour indiquer les indices potentiels permettant d'identifier certaines langues, la compréhension de leur interaction et du processus

cognitif menant à la décision reste hors d'atteinte. On touche en effet là à un processus où, en fonction de la familiarité des sujets avec les langues testées (contrastes phonologiques perçus ou non ; recouvrement lexical entre les langues connues du sujet et la langue cible ou non) une interaction se produit entre la compréhension monolingue de la parole, des processus plus généraux de perception de la parole, ainsi qu'éventuellement des aspects liés au traitement musical du rythme (par exemple Todd & Brown, 1994 ; Besson & Schön, 2001).

En résumé, je suis convaincu que les expériences portant sur l'identification perceptuelle des langues, ou plus généralement sur la perception des langues et dialectes étrangers, sont porteuses d'informations fondamentales pour notre compréhension des mécanismes de perception. Cependant, ce type d'étude est particulièrement difficile à mener et, en l'absence d'hypothèses claires et précises portant soit sur les indices pertinents, soit sur les processus cognitifs mis en jeu, il me semble illusoire d'en attendre beaucoup. Pour nuancer ce propos, je signale cependant que les études récentes portant sur l'intelligibilité de la parole et la perception des dialectes ou accents non familiers me semblent particulièrement prometteuses (e.g. Clopper & Bradlow, 2008 ; Bradlow & Bent, 2008). En effet, il s'agit là de coupler deux types de manipulation de l'information disponible dans le signal et je suis convaincu que cette direction est pertinente pour améliorer notre compréhension de la nature de l'information dans le signal extraordinairement redondant qu'est la parole.

2.5. APPROCHES DIALECTALES MULTIDIMENSIONNELLES

Document de référence : Annexes C.7 et C.8

Ferragne, E. & Pellegrino, F. 2007. "Automatic dialect identification: a study of British English", in *Speaker Classification II*, LNCS, Müller, C. & Schötz, S. (eds), Springer, pp. 243-257

Barkat-Defradas, M., Hamdi, R., & Pellegrino, F. 2004. « De la caractérisation linguistique à l'identification automatique des dialectes arabes », *actes du Workshop MIDL*, Paris, 29-30 novembre

2.5.1. Les dialectes arabes

Le domaine arabe est particulièrement propice à l'étude de la variabilité dialectale. En effet, son extension sur une zone de plusieurs milliers de kilomètres, bordée schématiquement au nord par la méditerranée et au sud par une zone désertique, permet au premier abord d'identifier un axe d'étude ouest-est privilégié, même si des considérations socio-linguistiques (modes et lieux de vie, contacts de langues, etc.) rendent la réalité des parlers arabes infiniment plus complexe. Ce domaine linguistique a ainsi servi de cadre aux thèses de doctorat de Melissa Barkat-Defradas, (2000), Rym Hamdi, (2007) et Jalaleddin Al-Tamimi (2007), toutes trois préparées à DDL. Dans ce contexte, des aspects descriptifs phonétiques, phonologiques et lexicaux ont été étudiés (variabilité phonétique et phonologique des systèmes vocaliques, corrélats acoustiques du rythme, structures syllabiques au sein du lexique), des expériences de perception ont été menées (catégorisation des voyelles, discrimination et identification des parlers et des zones dialectales, évaluation de la saillance perceptive des différences prosodiques) et des approches

automatiques (systèmes vocaliques, modèles rythmiques et intonatifs) ont été développées[25].

Ces études sont décrites en détail dans chacune des thèses citées, et leur mise en perspective sera probablement développée très prochainement dans l'habilitation à diriger des recherches de Melissa-Barkat-Defradas. En conséquence, cette section apporte juste quelques éléments d'éclairage sur certains des aspects étudiés.

La thèse de Melissa Barkat-Defradas portait principalement sur une dimension de dialectologie phonétique et phonologique, appuyée par plusieurs études de perception visant à mieux caractériser la proximité dialectale entre les différents parlers étudiés, de l'océan atlantique au levantin. La description phonético-phonologique qualitative et quantitative d'une dizaine de parlers arabes constitue sans nul doute une très importante contribution à la dialectologie arabe, que nous avons complétée en utilisant la modélisation des systèmes vocaliques introduite durant ma thèse dans une tâche de discrimination des zones et des dialectes. Le taux d'identification correcte atteignait environ 80 % lorsque seuls les zones Maghreb et Moyen-Orient étaient testées et la prise en compte de la zone intermédiaire, moins bien caractérisée, dégradait ces performances. Dans cette étude, nous nous étions heurtés de front à un clivage majeur entre les approches descriptive et automatique : là où une dizaine de locuteurs (parlant chacun quelques minutes) représentait déjà une quantité de données significatives en description phonétique, cette quantité d'enregistrement ne représente malheureusement qu'un corpus dérisoire si l'on souhaite estimer correctement des modèles statistiques pour l'identification. En l'absence de réelle alternative, nous avions cependant exploité une procédure de validation croisée (de type *leave-one-out*) pour évaluer nos hypothèses sur la pertinence de ces modèles pour discriminer les parlers arabes. Il s'agissait donc là plus d'un modèle-jouet que d'une réelle étude grandeur nature. De manière générale, ce constat vaut également pour les autres modélisations implémentées par la suite.

Dans le cadre de la même thèse, nous avions voulu évaluer la saillance des indices discriminants présents dans le signal pour des locuteurs arabophones et francophones. Un protocole utilisant de la parole naturelle avait permis de mettre en évidence que les locuteurs arabophones étaient très performants pour discriminer les zones ainsi que pour l'identification des parlers eux-mêmes. Lorsqu'il était demandé aux sujets arabophones de préciser les indices qui les avaient aidés, les réalisations vocaliques étaient évoquées, tant en matière de timbres que de réalisation de l'opposition de durée. L'autre élément souvent mentionné était le rythme perçu, en particulier pour distinguer les parlers moyen-orientaux (MO) des parlers maghrébins (MA).

---

[25] Hamdi, Barkat & Pellegrino, (2004) propose un aperçu des travaux menés à cette date. Les travaux menés en identification automatique ne seront pas abordés en détail car ils reprennent en substance les méthodes développées dans le cadre de l'identification automatique des langues par R. André-Obrecht, J. Farinas, J.-L. Rouas et moi-même. Les meilleurs résultats ont été obtenus avec les modèles prosodiques introduits par J.-L. Rouas dans sa thèse (cf. Rouas *et al.*, 2006).

Nous avons alors testé si l'information suprasegmentale seule permettait une discrimination. Pour cela, nous avons procédé à une synthèse d'une onde acoustique dont l'amplitude et la fréquence étaient calquées sur les enregistrements de nos locuteurs. Cette onde variable, était, de plus, discontinue puisque des silences correspondaient aux zones non voisées des enregistrements initiaux. Le son produit n'avait guère plus d'attributs de « parole » et il s'agissait là d'une dégradation considérable par rapport à un enregistrement naturel ou même resynthétisé en procédure « saltanaj » (Ramus & Mehler, 1999). Pourtant, les locuteurs arabophones obtenaient encore des performances légèrement supérieures au hasard dans la tâche de discrimination par zone MA vs. MO (58% d'identification correcte, $p<.005$). Les locuteurs francophones obtenaient en moyenne des performances comparables au hasard, même si certains individus discriminaient très bien les stimuli (Barkat, Ohala & Pellegrino, 1999). Il est probable que ces résultats indiquent des différences de stratégie dans la tâche ; les locuteurs arabophones ont probablement procédé en évaluant intrinsèquement si l'extrait proposé ressemblait à ce qu'aurait donné leur propre parler avec ce type de transformation, tandis que pour les sujets francophones, une stratégie « musicale » était plus probable.

Le fait qu'une importante variation rythmique soit présente au sein du continuum dialectal arabe et le fait que dans une certaine mesure, cette variation soit saillante pour les auditeurs, suggérait là une piste de recherche intéressante qui a été développée par Rym Hamdi dans sa thèse. À partir d'une segmentation manuelle comparable à celle proposée par Franck Ramus (en intervalles vocaliques et intervocaliques), elle a étudié la variation inter-dialectale par projection dans les espaces paramétriques devenus maintenant classiques dans ce type d'étude : %V, ΔC, ΔV, PVIv et PVIc (cf. p. 46 et suivantes pour une discussion de ces paramètres). Parmi les résultats les plus intéressants, on relèvera en particulier la très importante variation inter-individuelle et inter-dialectale mise en regard de la variation observée avec d'autres langues utilisées comme repères (anglais américain de la côte est, catalan et français). Il est ainsi une nouvelle fois démontré que ces paramètres indiquent uniquement des tendances des langues (ou dialectes) plutôt qu'elles ne reflètent une organisation rythmique rigide. Par ailleurs, la distribution des dialectes dans l'espace bidimensionnel défini par ΔV et ΔC est également instructive. L'étude de R. Hamdi met en effet en évidence une distribution où les dialectes de la zone dite intermédiaire ZI (égyptien et tunisien), ne sont en réalité pas situés topologiquement *entre* les zones MA et MO mais en position *hybride* : ils présentent des caractéristiques significativement semblables aux parlers MA pour ΔV et comparables aux parlers MO pour ΔC (Figure 10). Suite à l'étude princeps de Ramus, Nespor & Mehler, (1999), ΔC est considéré comme l'un des principaux corrélats acoustiques du rythme, tandis que ΔV reflèterait une interaction plus complexe entre des aspects de réduction vocalique, de contraste de durée et éventuellement de différences de durées vocaliques intrinsèques ou contextuelles.

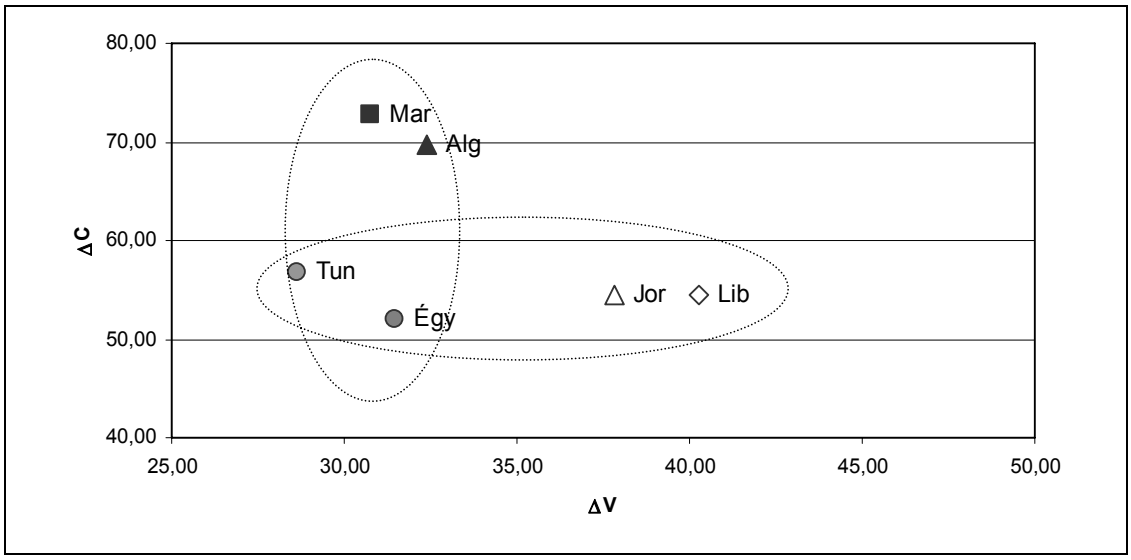


Figure 10 – Projection dans l'espace ΔV-ΔC des valeurs moyennes obtenues pour six dialectes arabes : algérien, égyptien, jordanien, libanais, marocain et tunisien (d'après Hamdi, 2007:229).

Rym Hamdi propose dans sa thèse une discussion approfondie de ces différents aspects dans les dialectes arabes (p. 247 et suivantes) sous le double éclairage phonétique (réalisation de l'opposition de durée, en particulier) et phonologique (structure et occurrences des différents types syllabiques dans un corpus de parole spontanée pour trois des dialectes). Il en résulte en particulier que la proportion de syllabes complexes (i.e. présentant au moins un cluster consonantique) est extrêmement faible en libanais (8,8 % des occurrences) et nettement plus élevée en marocain (33,3 %). La valeur obtenue pour le tunisien est intermédiaire, (15,4 %) mais relativement plus proche du marocain que du libanais, ce qui est en apparente contradiction avec le résultat précédant : les valeurs de ΔC des dialectes de la zone intermédiaire et du Moyen-orient étant proches, on se serait attendu à retrouver cette proximité dans la distribution des syllabes complexes.

En fait, ce paradoxe s'explique probablement si l'on prend en compte le nombre de consonnes présentes dans les clusters complexes. On voit en effet dans le Tableau 10 que les syllabes contenant 3 consonnes ou plus sont deux fois plus fréquentes en marocain qu'en tunisien (21,4 % vs. 9,7 %). Cette différence de proportion peut être la source de la plus grande variabilité ΔC observée en marocain, même si une étude plus poussée semble nécessaire[26].

[26] Il serait particulièrement intéressant d'évaluer avec des données portant sur plus de dialectes et de langues dans quelle mesure la notion de syllabes complexes (définies par la présence de clusters consonantiques), est ou non plus pertinente que le nombre total de consonnes des syllabes comme indice de complexité syllabique. Cette remarque anticipe des aspects qui seront abordés dans la partie 3 de ce document.

Tableau 10 – Distribution des fréquences d'occurrence des types syllabiques dans les trois dialectes du corpus de Hamdi, (2007).

TYPE DE SYLLABE		NBRE DE CONSONNES	MAROCAIN	TUNISIEN	LIBANAIS
SIMPLE	v	0	3,75%	4,29%	3,55%
	cv	1	32,82%	35,40%	43,23%
	cvv	1	4,52%	4,65%	5,23%
	vc	1	0,90%	1,64%	0,28%
	cvc	2	22,74%	30,57%	33,61%
	cvvc	2	1,94%	8,03%	5,32%
COMPLEXE	ccv	2	10,98%	4,01%	2,99%
	ccvv	2	0,90%	1,73%	0,28%
	ccvc	3	10,85%	4,38%	3,73%
	cvcc	3	3,10%	1,82%	0,93%
	cccv	3	2,45%	0,18%	-
	ccvvc	3	1,68%	2,65%	0,84%
	cccvc	3	1,55%	-	-
	cccvv	3	0,52%	-	-
	ccvcc	4	1,03%	0,64%	-
	cccvcc	5	0,26%	-	-
Proportion de syllabes à 3 Consonnes et Plus			**21,44%**	**9,67%**	**5,50%**

La Figure 11 est une extension de la Figure 10 ; les six dialectes arabes, regroupés en trois zones (MA, ZI et MO), sont projetés dans le même espace bidimensionnel et la variabilité intra-dialectale est matérialisée par les barres d'écarts-types. De plus, les données issues des trois langues « repères » sont également projetées (en pointillés). Les données enregistrées pour ces trois langues sont théoriquement plus homogènes que pour l'arabe dialectal, les groupes de sujets présentant globalement une origine commune pour chacune des langues. Contrairement à la figure précédente, les axes ΔV et ΔC sont ici normés à la même échelle, ce qui permet de visualiser sur quelle dimension la variabilité est la plus importante.

Le premier constat porte sur l'importante variabilité observée, quelles que soient la dimension et la langue, à l'exception toutefois du faible écart-type de ΔV pour l'anglais. L'anglais se démarque également par la variabilité observée sur ΔC, largement supérieure à celles des autres langues étudiées. Malheureusement, identifier avec certitude l'origine de ces variations n'est pas aisé. On peut cependant tenter d'évaluer l'effet des variations de débit. Une analyse de Kruskal-Wallis des débits des dialectes arabes de l'étude de Rym Hamdi montrait que ces variations n'étaient pas expliquées par la variable « dialecte » (p. 275 de sa thèse). Pour autant le débit varie d'un locuteur à l'autre, comme le montre le Tableau 11.

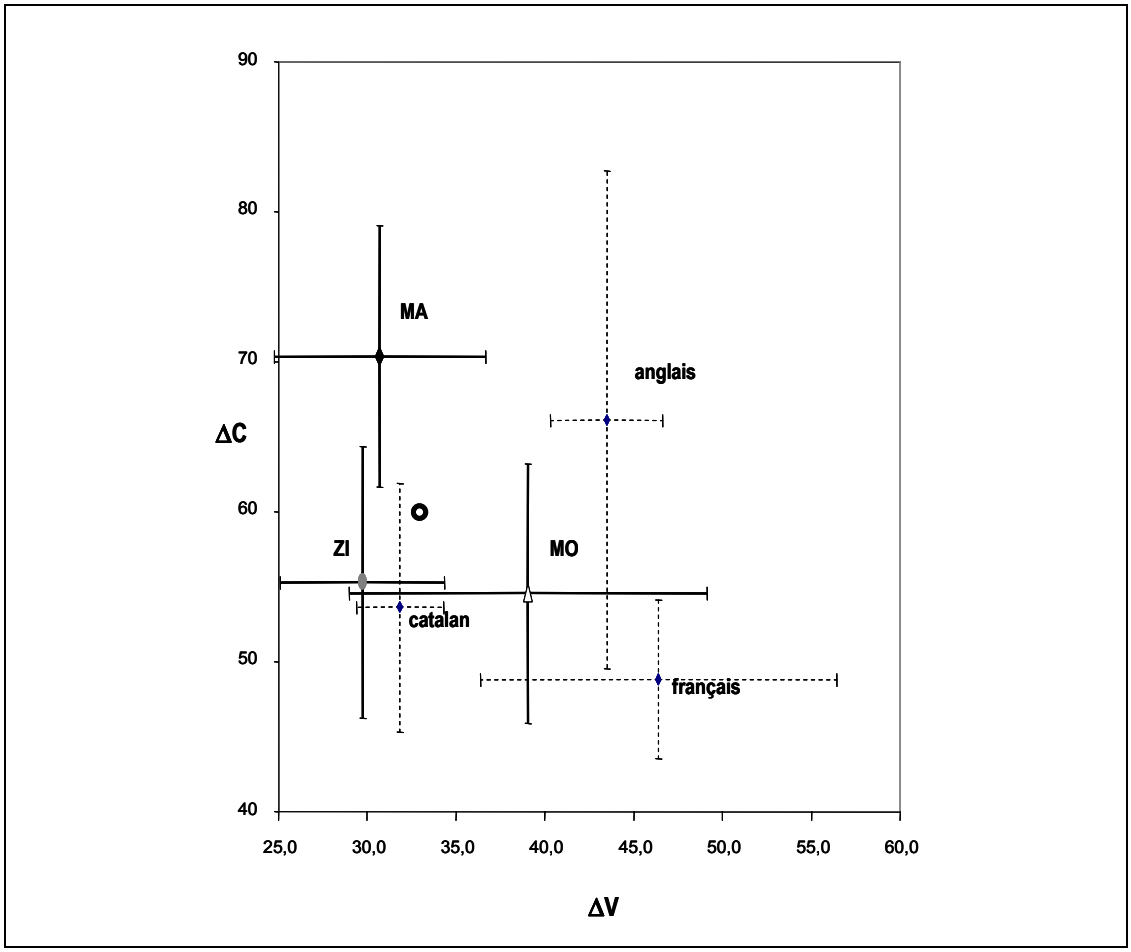


Figure 11 – Projection des dialectes arabes (regroupés par zones) et des langues repères dans l'espace ΔV-ΔC (MA : Maghreb, ZI : zone intermédiaire ; MO Moyen-orient). Les barres d'erreur correspondent à l'écart-type. Le cercle (trait gras) correspond à la moyenne des 6 dialectes arabes (graphique établi à partir des données de Hamdi, 2007).

On y trouve le débit mesuré en syllabes par seconde (valeur moyenne et écart-type) dans chacune des langues ainsi que la durée moyenne des syllabes[27] pour des locuteurs ayant un débit rapide (débit moyen + 1 écart-type) et lent (débit moyen – 1 écart-type). La différence entre ces deux valeurs, exprimée en millisecondes, est un indice de la variation attendue pour les paramètres ΔV et ΔC, même si la proportion affectant respectivement voyelles et consonnes est variable en fonction des langues (e.g. Dellwo & Wagner, 2003). On constate que l'écart mesuré pour l'anglais (328 ms) est largement supérieur aux écarts évalués avec les autres langues, qui varient de 58 ms pour le français à 86 ms pour l'arabe. On peut donc avancer l'hypothèse que l'écart-type très important observé pour ΔC en anglais est probablement dû à d'importantes variations de débits des locuteurs.

[27] La durée moyenne des syllabes est définie ici comme l'inverse du débit syllabique. Par commodité, elle est indiquée en millisecondes.

Tableau 11 – Variations de débits et influence sur la durée moyenne des syllabes (d'après les données de Hamdi, 2007).

LANGUE	DÉBIT (SYL/S)		FOURCHETTE DE VARIATION DES DURÉES SYLLABIQUES (MS)		ÉCART (MS)
	MOYEN μ	ÉCART-TYPE σ	LOCUTEUR RAPIDE (DÉBIT $\mu + \sigma$)	LOCUTEUR LENT (DÉBIT $\mu - \sigma$)	
ARABE	3,99	0,66	215	301	86
FRANÇAIS	3,19	0,29	287	345	58
CATALAN	3,85	0,54	228	302	74
ANGLAIS	3,83	1,85	176	504	328

Quels autres enseignements peut-on tirer de la Figure 11 ? Tout d'abord, la projection des trois zones des dialectes arabes couvre une aire du même ordre de grandeur que les trois autres langues. La position moyenne des dialectes arabes (matérialisée par un cercle noir sur la figure), se trouve approximativement en position intermédiaire entre le catalan et l'anglais, légèrement décalée vers les faibles valeurs de ΔV. Dans la mesure où ΔC serait un corrélat des classes rythmiques, les trois zones dialectales de l'arabe se situent vers des valeurs « plus accentuelles » que le catalan et *a fortiori* le français. Dans le détail, on constate que la zone MA atteint même une valeur moyenne de ΔC supérieure à celle de l'anglais, alors que les valeurs caractéristiques des zones ZI et MA sont très proches de celle du catalan. L'unité rythmique acoustique des dialectes arabes est ainsi largement battue en brèche, tout comme le laissaient présager les résultats de l'expérience de perception dégradée commentée précédemment (Barkat, Ohala & Pellegrino, 1999). Sur l'axe ΔV, les langues et dialectes étudiés se projettent & l'ordre suivant :

$$\Delta V_{ZI} < \Delta V_{MA} < \Delta V_{cat} < \Delta V_{MO} < \Delta V_{ang} < \Delta V_{fra}$$

Cette gradation est purement descriptive dans le sens où l'analyse statistique n'a pas montré de différences significatives sur cette dimension pour certains des groupes considérés (cas de ZI et MA en particulier). Elle permet cependant d'illustrer les facteurs influençant potentiellement cette variable. En particulier, le fait que ΔV soit supérieur dans les dialectes MO plutôt que dans les autres dialectes est parfaitement compatible avec ce que l'on sait de la réalisation de l'opposition phonologique de durée vocalique en arabe (voir la thèse de Rym Hamdi, p. 247 et suivantes). Le catalan, qui réalise un phénomène de réduction vocalique, présente des valeurs relativement faibles, similaires aux parlers arabes MA et ZI tandis que l'anglais et le français montrent des valeurs plus importantes encore que les parlers MO, pour des raisons probablement différentes. Dans le cas de l'anglais, on peut évoquer la présence de diphtongues, conjuguée à celle des corrélats temporels du contraste tendu/relâché. Pour le français, l'explication est moins simple, d'autant plus que la valeur moyenne obtenue est largement supérieure à celle mesurée dans Ramus, Nespor & Mehler, (1999 :272). Cet écart peut être par exemple lié au matériau même, pour peu que les proportions de voyelles intrinsèquement longues (comme les nasales) et brèves (comme des réalisations réduites des schwas) aient été

nettement différentes dans les deux corpus. Une analyse plus poussée nécessiterait cependant un étiquetage phonétique plus fin des enregistrements.

2.5.2. Les dialectes anglais des îles britanniques

Cette section porte sur des travaux entrepris dans le cadre de la thèse d'Emmanuel Ferragne, soutenue en 2008. Leur objectif était de caractériser les dialectes anglais des îles britanniques de manière à proposée une identification semi-automatique, voire automatique. Les données se trouvent au cœur de la démarche ambitieuse mise en place et il était donc impératif d'en disposer en quantité représentative. La commercialisation du corpus ABI (*Accents of the British Isles*) est donc arrivée à point nommé fin 2003. Ce corpus, contenant des enregistrements d'une vingtaine de locuteurs issus de 14 points d'enquête (pour un total de 284 locuteurs) a évidemment permis de mener un travail plus représentatif que ne l'aurait été une étude nécessitant une collecte de données dans le temps d'une thèse. Cependant, il est extrêmement regrettable que la qualité du corpus laisse autant à désirer. Outre le fait que la qualité des enregistrements s'avère très variable, l'absence totale de données socio-linguistiques sur les locuteurs paraît aberrante dans un corpus de ce type. Les locuteurs sont sensés être représentatifs des zones d'enquête, mais rien n'est malheureusement connu sur leur histoire linguistique et même leur âge fait défaut.

Malgré ces limitations, ce corpus a permis de développer une approche dialectologique innovante et multidimensionnelle : la variation du rythme parmi les dialectes[28] a été étudiée (Ferragne & Pellegrino, 2004b) non seulement en termes de durées, mais également en intégrant des PVI d'intensité, sur une idée originale d'E. Ferragne (Ferragne & Pellegrino, 2008) qui s'est révélée particulièrement pertinente[29]. Ces travaux ont mis en évidence une importante variabilité inter-dialectale sous forme de plusieurs continuums (en fonction des paramètres étudiés) aux extrémités desquels se trouvent des dialectes facilement discriminables, alors que les dialectes intermédiaires sont peu individualisables. Une autre piste, basée sur les différences de réalisation de la diphtongaison de l'ensemble lexical FACE a également été menée, quoique de manière plus marginale (Ferragne, 2006 ; Ferragne and Pellegrino, 2004a).

Au-delà, une importante contribution de cette thèse exploite la méthode ACCDIST – introduite par Huckvale (2004) – et fait ainsi écho à la modélisation différenciée mise en place durant ma propre thèse pour l'identification des langues puisque *in fine*, les deux approches visent à la modélisation des systèmes vocaliques. Plus précisément, l'approche par modélisation statistique GMM différenciée visait à prendre en compte implicitement et de manière non supervisée

[28] L'étude menée montre une importante variation entre les dialectes, sans que le pouvoir discriminant des paramètres rythmiques soit très fort.

[29] Cette constatation est conforme à des résultats que nous avions obtenus dans le cadre de l'identification des langues à partir des pseudo-syllabes ; dans ce cadre en effet, la prise en compte de l'intensité normalisée du segment vocalique comme indice de l'accentuation permettait d'améliorer les résultats d'identification (Farinas *et al.*, 2002).

la structure acoustique du système vocalique tandis que l'approche ACCDIST tâche d'appréhender la structure phonético-phonologique des dialectes. En effet, il s'agit là d'une modélisation *explicitement* structurelle puisque seules les relations entre constituants vocaliques sont prises en compte. Par contre, les contraintes sont différentes et l'algorithme est partiellement supervisé puisqu'il nécessite d'étiqueter chaque segment vocalique en termes d'ensemble lexical. L'approche choisie dans la thèse d'E. Ferragne exploite les « passages lus », trois textes courts produits par l'ensemble des locuteurs. Le système développé est ainsi un système d'identification automatique du dialecte très performant puisqu'il dépasse 90% d'identification correcte pour les locuteurs hommes et femmes, parmi 13 dialectes possibles. Ce système est cependant limité à un cadre dépendant du texte : à partir de la transcription des textes, un système d'alignement automatique implémenté sous HTK permet alors d'identifier chaque segment vocalique en termes d'ensemble lexical ; la distance cepstrale entre les différents ensembles lexicaux est alors évaluée et utilisée pour établir un modèle matriciel qui capture la structure du système vocalique de chaque dialecte[30].

D'une certaine manière, alors que l'approche par GMM générait des modèles continus à partir de segments représentés également dans un espace continu, l'approche ACCDIST discrétise les voyelles (passage de l'espace cepstral à un label d'ensemble lexical) avant d'en proposer un élégant modèle structurel. L'utilisation même de la notion d'ensemble lexicaux met l'accent sur le fait que cette méthodologie exploite à la fois des indices acoustiques et des informations de plus haut niveau, liées aux représentations lexicales.

2.6. EN GUISE DE CONCLUSION

Ce chapitre dressait le bilan des quelques dix années de recherche que j'ai menées dans le champ de l'identification des langues et des dialectes. À l'issue de ma thèse, l'opportunité de rejoindre le laboratoire DDL, d'abord en postdoc puis en tant que chercheur CNRS m'a amené vers des problématiques plus linguistiques qu'auparavant. Bien évidemment, cette orientation correspondait à mes aspirations scientifiques mais force est de constater que cette transition m'éloigne des aspects les plus *automatiques*. En particulier, l'échec du renforcement de l'activité modélisation à DDL, faute de recrutement dans la durée, entraîne une perte de vitesse de ce thème à Lyon : les idées et les perspectives ne manquent pas, mais les moyens mis en regard ne sont pas à la hauteur de ce que doit être une recherche d'excellence ; je pense donc que ces aspects sont amenés à devenir plus marginaux dans mes travaux, même si je souhaite néanmoins pouvoir continuer à m'investir sur ces thèmes dans le cadre de réseaux de collaboration.

---

[30] Un schéma didactique explicatif de l'algorithme est proposé par E. Ferragne dans sa thèse et j'y renvoie le lecteur intéressé (Ferragne, 2008 :313-314). De manière plus générale, cette thèse contient également une description fine des systèmes phonético-phonologiques des dialectes du corpus et une riche discussion sur la pertinence de la notion de classification dialectale et sur son extension au cadre de la logique floue. J'encourage donc vivement le lecteur à lire ce manuscrit dans son intégralité.

Toutefois, les travaux entrepris sur la caractérisation du rythme et la modélisation des systèmes vocaliques sont à l'origine des questionnements scientifiques qui sous-tendent la plupart de mes recherches actuelles portant sur les relations entretenues par les notions d'information et de complexité en phonologie d'une part et sur la compréhension de la parole dégradée d'autre part. Comme indiqué dans la première partie, ce dernier point ne sera pas abordé plus avant dans ce document. J'ai cependant la conviction que l'étude de la résistance de la compréhension et de la perception à la dégradation du signal est un excellent moyen d'investiguer les niveaux de traitement, les primitives informationnelles et les interactions éventuelles entre les niveaux. Les premiers résultats que nous avons obtenus dans le cadre des projets pilotés par Fanny Meunier ont ainsi mis en relation des effets neurocognitifs avec des caractéristiques acoustiques et lexicales des stimuli (Hoen *et al.*, 2007). Il est probable que des interactions soient à l'œuvre et le niveau phonético-phonologique mérite évidemment une attention particulière pour identifier les traits, au sens le plus large, qui sont pertinents du point de vue informationnel. En cela, les travaux portant sur la parole dégradée rejoignent les préoccupations développées au chapitre suivant.

3. VERS UNE APPROCHE INFORMATIONNELLE DE LA PAROLE

"Make everything as simple as possible, but not simpler" (attribué à A. Einstein)

3.1. INTRODUCTION

Ma première rencontre avec John J. Ohala à mon arrivée à DDL, fin 1998, fut l'occasion de lui présenter mes travaux de thèse. Dans la discussion, il me demanda alors en substance, si j'avais des éléments permettant de répondre à la question suivante : les voyelles sont-elles acoustiquement plus longues dans des langues présentant un système vocalique de grande taille (i.e. avec de nombreux segments) par rapport à des langues à systèmes vocaliques de taille plus réduite ? Vu d'aujourd'hui, il me semble avoir alors répondu que cela paraissait plausible qu'une plus grande imprévisibilité (paradigmatique) soit compensée en fournissant plus d'information (syntagmatique) à l'auditeur pour éviter les confusions, même si je n'avais jamais véritablement cherché à évaluer de telles différences. Aujourd'hui, cette réponse me paraît largement insuffisante, mais je pense que la question des relations entre complexité[31] et information reste d'actualité et qu'elle m'a attiré dans un engrenage intéressant et fructueux. Peu après cette discussion anecdotique, j'acceptais de participer au projet « Cognitique » P3 (*Production et Perception de la Parole*) mené par René Carré[32] portant sur une étude de la variabilité observée sur la production et la perception des voyelles, et ce à plusieurs échelles (variation en fonction du contexte informationnel, des locuteurs et des dialectes ou langues de ces locuteurs). Mon intervention dans ce projet portait principalement sur des aspects techniques (conception et réalisation des interfaces logicielles pour le recueil des données) mais les questions scientifiques soulevées, sur lesquels je n'étais initialement pas spécialement compétent, ont attisé ma curiosité. Là encore, les notions de complexité et d'information semblaient participer d'une même problématique susceptible d'éclairer les arcanes de la communication parlée. À l'heure où les sciences de la complexité[33] se généralisent, on peut ainsi se poser la

[31] En assimilant la taille du système vocalique à une complexité, j'anticipe nettement sur la discussion qui suivra.

[32] Jean-Sylvain Liénard (LIMSI) et Jean-Marie Hombert (DDL) étaient co-pilotes de ce projet dont René Carré assurait la coordination.

[33] La définition des sciences de la complexité mériterait probablement de s'y arrêter tant le champ est vaste, de la physique des particules aux systèmes sociaux. Il nous semble cependant que cela nous mènerait au-delà de l'objet de ce document.

question de leur pertinence dans le cadre de la linguistique et plus particulièrement de la phonologie. Dans cette tâche, il nous semble important de distinguer dans un premier temps la *notion* de complexité elle-même des sciences de la complexité, vue plus particulièrement comme *l'étude des systèmes dynamiques complexes*. À notre avis, la notion de complexité est pertinente si elle permet de mieux appréhender des phénomènes linguistiques (description ou explication), en particulier de manière translinguistique, ou encore s'il s'avère qu'elle joue un rôle dans l'ontogénie ou la phylogénie du langage et des langues[34]. Les sciences de la complexité fournissent quant à elles une multitude de concepts et d'outils nourris en particulier de recherches en mathématiques, en physique et en ingénierie. On peut donc aborder la question de la complexité sans faire référence à des systèmes dynamiques complexes comme l'illustrent les linguistes de la première moitié du XXe siècle. Vice-versa, certains concepts fondamentaux qui sont associés à ces systèmes dynamiques ne nécessitent pas forcément de s'intéresser à la complexité même. Par exemple, la notion de rétroaction joue un rôle fondamental dans la science des systèmes ou cybernétique, sans faire explicitement référence à la complexité.

Notre intuition est cependant qu'à la fois les notions de complexité et de systèmes dynamiques complexes sont pertinentes pour la compréhension de la communication langagière. En particulier, nous pensons que ces concepts se rejoignent en partie autour de l'idée d'*information linguistique*.

Les notions de complexité et d'information langagières sont présentes depuis près d'un siècle dans les travaux linguistiques et au cours de cette période, ils ont revêtu des formes diverses, tant ils recouvrent de champs possibles. La section suivante fournit quelques repères historiques en phonologie sous l'angle de la (science de) la complexité (§ 3.2), de manière à mettre en perspective nos propres travaux – encore largement inachevés – qui seront eux-mêmes commentés au

---

[34] Le débat sur l'égale complexité des langues du mondes et son corollaire, l'existence de compensation entre des composantes linguistiques "simples" et d'autres "complexes" au sein d'une langue connaît un regain d'intérêt. Ses fondements ont en particulier été exprimés par Hockett (1958:180-181) :

> "Objective measurement is difficult, but impressionistically it would seem that the total grammatical complexity of any language, counting both morphology and syntax, is about the same as that of any other. This is not surprising, since all languages have about equally complex jobs to do, and what is not done morphologically has to be done syntactically. Fox, with a more complex morphology than English, thus ought to have a somewhat simpler syntax; and this is the case. Thus one scale for the comparison of the grammatical systems of different languages is that of average degree of morphological complexity – carrying with it an inverse implication as to degree of syntactical complexity."

Des études récentes ont évalués cette hypothèse dans une perspective typologique (Fenk-Oczlon G. & Fenk, 1999, 2005 ; Shosted, 2006) ou diachronique (Comrie, 1992 ; Dahl, 2004) ; en effet, dans ce schéma, la complexification d'une composante s'accompagne nécessairement de la simplification d'une autre. Cette question était déjà présente dans Hagège & Haudricourt (1978), mais elle devient centrale dans Planck (1998) et elle est de fait abordée dans Trudgill (2004) et les commentaires provoqués par cet article.

paragraphe 3.3. L'objectif n'est pas ici d'établir un état de l'art exhaustif, mais bien d'esquisser certains des travaux antérieurs ayant influencé notre propre recherche.

3.2. COMPLEXITÉ & INFORMATION EN PHONOLOGIE : UN ENRACINEMENT ANCIEN

Dans le Cours de Linguistique Générale (de Saussure, 1916/1996:65) , on trouve la définition suivante du phonème :

> « La délimitation des sons de la chaîne parlée ne peut donc reposer que sur l'impression acoustique ; mais pour leur description, il en va autrement. Elle ne saurait être faite que sur la base de l'acte articulatoire, car les unités acoustiques prises dans leur propre chaîne sont inanalysables. (…) le phonème est la somme des impressions acoustiques et des mouvements articulatoires, de l'unité entendue et de l'unité parlée, l'une conditionnant l'autre : ainsi c'est déjà **une unité complexe**, qui a un pied dans chaque chaîne » (emphase ajoutée, F.P.).

Le phonème est ainsi conçu comme intrinsèquement complexe, de par sa dualité entre perception (l'unité entendue) et production (l'unité parlée). La caractérisation du phonème occupera longtemps l'arène phonologique, tant dans sa nature (physique, psychologique ou fonctionnelle, universelle ou relative, etc.) que dans son rôle dans les différentes théories phonologiques. Même si la notion de complexité n'est pas réellement au cœur de ces débats, elle est parfois présente. À titre d'exemple, Swadesh introduit la notion de critère d'association constante dans la caractérisation des phonèmes, qu'il définit ainsi :

> "*The criterion of constant association*. If a set of phonetic elements only occur together, they constitute a phonemically unitary complex; thus, the stop and the aspiration in English initial *p*. One or both of the phonetic elements may recur in other complexes without affecting the unitary nature of the complexes; in this event, all the phonemes that involve a given phonetic element constitute a phonemic class" (Swadesh, 1934:123).

Cette mention de *complexe phonémique* de nature unitaire bien que constitué d'éléments phonétiques dissociables et réutilisables n'est évidemment pas unique et elle traversa le siècle dernier. Les deux citations présentées ci-dessus illustrent deux directions selon lesquelles la question de la complexité phonético-phonologique a été explorée (comme dualité entre production et perception pour de Saussure et comme dualité entre élément constituant un tout et élément composé pour Swadesh et bien d'autres[35]). Dans les deux cas, cette complexité est plutôt de nature intrinsèque au segment phonétique, même si celui-ci est conçu comme élément d'un système phonologique.

[35] Par exemple, l'introduction d'un formalisme basé sur des éléments unaires dans la phonologie du gouvernement (Kaye, Lowenstamm & Vergnaud, 1990) permet d'évaluer la complexité des phonèmes explicitement à partir des éléments les constituant.

Des approches plus explicitement systémiques ont également été menées, soit dans une perspective structuraliste, soit dans une perspective fonctionnelle. Parmi d'autres, des notions de traits distinctifs, de rendement fonctionnel ou encore de marque s'inscrivent selon nous dans la problématique de la complexité et ces quelques pages visent à montrer en quelque sorte comment de nombreux linguistes du XXe siècle ont fait « de la complexité » sans nécessairement l'expliciter. Les relations entre les notions de *complexité* et d'*information* seront en particulier abordées au cours de ce survol, en rapport avec la vision structurelle dynamique et fonctionnelle que nous promouvons.

✦ Les années 1920-1930

Dès les années 1920 apparaissent des notions relatives à la complexité, qui révolutionneront la linguistique du XX^{e} siècle. En 1929 paraît le premier tome des Travaux du Cercle Linguistique de Prague, manifeste extrêmement riche, en particulier grâce aux contributions de N.S. Troubetzkoy et R. Jakobson. La même année paraît également la thèse de G. K. Zipf, intitulée *Relative Frequency as a Determinant of Phonetic Change* (Zipf, 1929). Ces deux productions scientifiques, aussi différentes soient-elles, sont à notre sens, emblématiques des idées novatrices circulant à l'époque.

On trouve par exemple dans les « thèses présentées au premier congrès de philologues slaves » le programme suivant :

> « Tâches fondamentales de la phonologie synchronique.
>
> 1. Il faut caractériser le système phonologique, c'est-à-dire établir le répertoire des images acoustico-motrices les plus simples et significatives dans une langue donnée (phonèmes), en spécifiant obligatoirement les relations existant entre lesdits phonèmes, c'est-à-dire en traçant le schème de structure de la langue considérée (…).
>
> 2. Il faut déterminer les combinaisons de phonèmes réalisées dans une langue donnée en comparaison avec les combinaisons théoriquement possibles de ces phonèmes, les variations de l'ordre de leur groupement et l'étendue de ces combinaisons.
>
> 3. On doit aussi déterminer le degré d'utilisation et la densité de réalisation des phonèmes en question et des combinaisons de phonèmes d'étendue variée. Il faut également étudier la charge fonctionnelle des divers phonèmes et combinaisons de phonèmes dans une langue donnée (…). »

Ces quelques lignes font référence à la relation entre production et perception (images acoustico-motrices), à la structure de la langue et aux contraintes phonotactiques dans l'étude des systèmes phonologiques, ce qui n'est pas surprenant étant donné leur auteur (R. Jakobson). Il est cependant remarquable de voir également mentionné des éléments relevant de la parcimonie (ou de la complexité algorithmique) : " (…) le répertoire des images acoustico-motrices **les**

plus simples et significatives (…)", ainsi que la notion de charge fonctionnelle[36] des phonèmes et de leurs combinaisons, et enfin le fait qu'il est nécessaire de mettre en regard les combinaisons de phonèmes réellement utilisées et leur proportion d'utilisation vis-à-vis des combinaisons théoriquement possibles. Par ailleurs ces paragraphes relatifs à l'étude synchronique sont précédés d'un fort encouragement à rapprocher phonologies diachronique et synchronique, en mettant en particulier l'accent sur l'aspect systémique et fonctionnel :

> "Si l'on envisage en linguistique synchronique les éléments du système de la langue du point de vue de leurs fonctions, on ne saurait juger non plus les changements subis par la langue sans tenir compte du système qui se trouve affecté par lesdits changements. **Il ne serait pas logique de supposer que les changements linguistiques ne sont que des atteintes destructives s'opérant au hasard et hétérogènes du point de vue du système. Les changements linguistiques visent souvent le système, sa stabilisation, sa reconstruction, etc.** Ainsi l'étude diachronique, non seulement n'exclut pas les notions de système et de fonction, mais, tout au contraire, à ne pas tenir compte de ces notions, elle est incomplète" (emphase ajoutée, F.P.)

On a là en quelques sortes non seulement un programme de phonologie, mais également les prémisses de ce que seront les approches de type systèmes dynamiques complexes après-guerre.

À la même période, G.K. Zipf, en s'appuyant en particulier sur les travaux antérieurs de Jespersen, élabore sa thèse, qui constituera la base d'un chapitre sur la phonologie dans son ouvrage de 1935. Ces travaux seront à l'origine d'un débat houleux, et des attaques violentes lui seront portées, en particulier par Joos (1936) – voir également la réponse de Zipf (1937) – et Troubetzkoy, quoique pour des raisons différentes. L'élément central de la thèse défendue par Zipf se trouve résumé dans la citation suivante :

> "(…) there exists an equilibrium between the magnitude or degree of complexity of a phoneme and the relative frequency of its occurrence" (Zipf, 1935, 49).

Il précise dans la suite qu'il s'agit d'une relation inverse (mais pas nécessairement proportionnelle) et il va plus loin puisqu'il indique que le système phonémique des langues lutte constamment pour maintenir cet équilibre, ce qui est probablement la cause des changements phonétiques (*ibid.*). Tout le chapitre intitulé *The form and behavior of phonemes* argumente vers ces conclusions. Ce chapitre présente ainsi des idées très intéressantes mais, souvent de manière non convaincante, en particulier pour cause de manque de fondements de certains arguments, en particulier pour l'estimation de la complexité des phonèmes. Par contre, Zipf a la conviction que si l'on veut éviter la circularité dans le raisonnement

[36] En français, la terminologie proposée par Martinet (rendement fonctionnel, 1933) sera préférée par la suite à *charge fonctionnelle*. En anglais, Twaddell (1935) parlait de *functional burdening*, mais la littérature retiendra *functional load*. Sur l'origine du concept, voir par exemple King, (1967).

et donc les tautologies, il est nécessaire de le fonder en dehors de la phonologie, en s'appuyant sur la physiologie ou la physique.

Au cœur de ce travail se situe la nécessaire mesure de complexité et Zipf, comme la plupart des phonologues s'étant essayé à l'exercice depuis, se heurtera à des difficultés peut-être insolvables. Son raisonnement s'appuie uniquement sur la production et plus précisément sur la décomposition des phonèmes en séquences de sous-gestes articulatoires. L'idée défendue est que la complexité d'un phonème résulte de l'accumulation de traits nécessitant chacun une dépense d'énergie au niveau de la production. Cette méthodologie sera contestée, aussi bien par Joos qui pointera en particulier que la complexité peut à l'inverse relever plus du contrôle (c'est-à-dire des commandes) que de l'énergie[37], que par Troubetzkoy (voir *infra*).

Pour Zipf cependant, l'existence d'un équilibre complexité/fréquence est incontestable :

> "it is clearly evident that the frequency distribution of phonemes in the steam of speech is by no means completely a matter of random chance but that the relative frequency of occurrence of a phoneme depends to a considerable extend upon its form" (*ibid.* p. 79).

De plus, il considère qu'il y a relation de causalité (un autre sujet de débat acerbe avec Joos) et il argumente que la fréquence exerce une pression sur la complexité dans une partie assez clairement téléologique sur les changements phonétiques. La contrainte invoquée est qu'un système phonologique en évolution tente de maintenir constant sa complexité globale, définie comme la somme des complexités des phonèmes, pondérées par leur fréquence (*ibid.*, p. 93). Plus on s'éloigne de cet équilibre (par changement de fréquence d'un phonème), plus la probabilité de changement phonétique augmente. Selon Zipf, deux types de changement peuvent permettre une régulation, soit en diminuant la complexité (*abbreviatory phonetic change*) ou à l'inverse en l'augmentant (*augmentative phonetic change*) en cas de fréquence trop faible. La direction du changement est

---

[37] Si l'on prend l'exemple de la paire /t, d/ en anglais comme le fait Zipf, la tâche est alors de déterminer lequel des deux phonèmes est le plus complexe. Il conclut que le phonème /d/ est plus complexe, de manière largement ambiguë : "When the stops of this pair are each viewed as a configuration of sequences of sub-gestures, the question of difference in magnitude resolves itself into the following formulation: Is the additional magnitude of complexity represented by *voicing* in the voiced stop *d* sufficient to counterbalance or outweigh a possible greater magnitude of complexity in the voiceless *t* which results from the *fortis* pronunciation and slight *aspiration* in *t*, which are absent in *d*? (...) There happens to be some slight direct evidence which clearly though not conclusively suggests that the total magnitude of complexity of a voiced stop *may* be greater than that of the voiceless stop" (emphase originale, Zipf, 1935:66).

Pour Joos, la cause est entendue et la conclusion totalement opposée : "(...) a voiced stop is easier to manage than a voiceless one, since it does not require cessation of voice after a preceding vowel (...)" (Joos, 1936:207).

Ohala aborde également cette question de la quantification du coût articulatoire dans le cadre de l'assimilation, même si, en un certain sens, il considère que le problème est secondaire puisque la notion de facilité articulatoire n'est pas à même d'expliquer tous les cas attestés (Ohala, 1990: 260).

conditionnée à la fréquence, mais également au contexte d'apparition du phonème, à sa variabilité phonétique[38] ainsi qu'à sa charge fonctionnelle, bien que le terme ne soit pas employé[39].

Le travail de Zipf sur les changements phonétiques est très riche et porteur de nombreuses idées qui seront reprises plus tard. Cependant, force est de constater que l'argumentation présentée est souvent peu convaincante. En bref, Zipf propose des intuitions visionnaires mais à une époque où de nombreux concepts des sciences de la complexité font défaut. En particulier, l'absence de concepts de boucle de rétroaction ou d'auto-organisation amènent à une vision très téléologique où les changements sont décrits de manière fondamentalement séquentielle et déterministe (si un seuil de fréquence est dépassé, alors un changement a lieu, résultant en un nouveau système, etc.).

Dès sa publication, le texte de Zipf essuie des critiques particulièrement appuyées, en particulier de la part de Troubetzkoy qui indique dans la version française des *Grundzüge* (Troubetzkoy, 1938:282) que selon lui « le degré de complexité de l'articulation ne se laisse pas mesurer » et que par conséquent « cette théorie – au moins dans la rédaction indiquée ci-dessus – doit donc être résolument écartée. ».

Ce qui est, selon nous, une force du travail de Zipf relève à l'inverse d'une faiblesse du point de vue de Troubetzkoy qui indique que chercher à expliquer des faits phonologiques par des causes biologiques, est contestable. Par contre, il conçoit qu'une interprétation phonologique des propositions de Zipf puisse faire sens :

> « Dans sa rédaction phonologique cette théorie pourrait se présenter ainsi : "des deux termes d'une opposition privative le terme non marqué apparaît plus souvent dans le discours suivi que le terme marqué". En gros et en bloc cette formule pourrait se trouver juste. Mais on ne peut en aucune façon la considérer comme une règle sans exception ».

De manière intéressante, là où Zipf souhaitait éviter une certaine circularité en ancrant sa réflexion dans des mesures objectives de complexité (articulatoires en l'occurrence), Troubetzkoy ramène le débat dans la sphère phonologique sous la forme – malheureuse à mon sens – d'une tautologie reposant sur la notion de marque. Dans le même court chapitre des *Grundzüge*, intitulé *De la statistique phonologique*, il introduit une distinction entre les fréquences d'apparition d'un

[38] On trouve d'ailleurs dans cet ouvrage toute une discussion sur la caractérisation de la variabilité phonétique (supposée gaussienne lorsqu'elle est non biaisée) et de son conditionnement par le contexte.

[39] Zipf indique que lorsque la fréquence d'un phonème est faible, il est très distinctif, ce qui encourage son hyperarticulation et son renforcement.

élément phonologique au sein d'un corpus et au sein du lexique de la langue considérée[40].

La fréquence au sein d'un corpus pose le problème de la dépendance au corpus (relativement à la taille et au style de composition), illustrée par la comparaison de deux textes brefs composés en allemand (*ibid.*, p 277-279) et pour lesquels il constate que, contrairement aux fréquences des mots et morphèmes, "la fréquence des différents phonèmes paraît être assez indépendante du genre de style du texte". Sur ces derniers, il indique enfin que :

> « (...) s'il n'y a aucun doute que la distinction entre termes d'opposition marqués et non marqués, de même que la distinction entre oppositions neutralisables et non neutralisables, ont une influence sur la fréquence des phonèmes, il est toutefois également clair que ces faits ne suffisent pas à expliquer les rapports de fréquence » (*ibid.*, p. 283).

En bref, et de manière peu surprenante, Troubetzkoy donne à la notion de marque un statut de primitive qui influence la fréquence d'apparition des phonèmes, sans que cette influence ne soit exclusive ou totalement déterministe[41]. Cette position est donc à l'opposé de celle de Zipf pour qui la fréquence est la primitive qui va influencer les changements phonétiques vers une plus grande ou moins grande complexité.

---

[40] Bien des années plus tard, Greenberg tentera lui aussi d'éviter un raisonnement circulaire en se référant explicitement à la complexité, sans toutefois le faire de manière entièrement convaincante :

> "Are there any properties which distinguish favored articulations as a group from their alternatives? There do, as a matter of fact, appear to be several principles at work. [There is one] which accounts for a considerable number of clusters of phonological universals (...) This is the principle that of two sounds that one is favored which is the less complex. The nature of this complexity can be stated in quite precise terms. The more complex sound involves an additional articulatory feature and, correspondingly, an additional acoustic feature which is not present in the less complex sound. This additional feature is often called a "mark" and hence the more complex, less favored alternative is called marked and the less complex, more favored alternative the unmarked. (...) It may be noted that the approach outlined here avoids the circularity for which earlier formulations, such as those of Zipf, were attacked. (...) In the present instance, panhuman preferences were investigated by formulating universals based in the occurrence or non-occurrence of certain types, by text frequency and other evidence, none of which referred to the physical or acoustic nature of the sounds. Afterward, a common physical and acoustic property of the favored alternatives was noted employing evidence independent of that used to establish the universals" (Greenberg, 1969:476-477).

Encore plus tard, cette relation entre fréquence, complexité et marque sera considérée comme acquise : « les phonèmes les plus fréquents sont les moins marqués et doivent donc être les plus simples » (Hagège & Haudricourt, 1978:21).

[41] Dans le même paragraphe, N.S. Troubetzkoy cite l'exemple du français pour lequel la fréquence relative du phonème sonore est supérieure à celle du phonème sourd dans les paires /ʃ-ʒ/ et /f-v/ alors que les phonèmes sourds sont relativement plus fréquents dans les paires /p-b/, /k-g/ et /s-z/.

La prise en compte de la fréquence au sein du lexique est cependant selon Troubetzkoy indispensable à l'étude du rendement fonctionnel (notion uniquement présente dans ce chapitre des *Grundzüge*) d'un élément ou d'une opposition phonologiques au sein de la langue. Il souligne à la suite de V. Mathesius, l'aspect trompeur que peuvent revêtir les fréquences absolues d'apparition des éléments en question et la ferme nécessité de considérer plutôt l'écart entre fréquence observée et fréquence théorique attendue. Dans le même chapitre, Troubetzkoy attire également l'attention sur l'importance du ratio entre le nombre de combinaisons réellement attestées dans le lexique d'une langue et le nombre théorique de combinaisons, faisant ainsi écho au texte de Jakobson paru dans les travaux du Cercle linguistique de Prague de 1929 et cité *supra*. Ainsi, il relève qu'en français, 73 % des mots monosyllabiques de structure CV concevables à partir de l'inventaire phonologique existent dans le lexique, alors que cette proportion n'est que de 31,8 % en allemand. Il attire ainsi l'attention sur le fait que ces différences peuvent permettre de caractériser les langues en fonction de critères que l'on appellera plus tard d'économie et de parcimonie :

> « … il y a des langues "économes" et des langues "prodigues". Dans les langues économes les mots qui ne se distinguent entre eux que par un seul phonème sont très nombreux et le pourcentage de réalisation des combinaisons de phonèmes théoriquement possibles est très élevé. Dans les langues "prodigues" existe la tendance à distinguer les mots les uns des autres par plusieurs procédés phonologiques et à ne réaliser qu'une petite partie des combinaisons de phonèmes théoriquement possibles » (*ibid.*, p. 288).

Troubetzkoy reconnaît cependant que la seule étude du lexique ne suffit pas à conclure sur la fréquence réelle d'apparition de ces mots dans le discours normal, et ce constat l'amène à poser une question fondamentale :

> « Existe-t-il à ce point de vue des règles de valeur générale ou bien les langues diffèrent-elles les unes des autres à cet égard ? On ne peut encore rien dire là-dessus pour le moment, car la statistique phonologique a été encore beaucoup trop peu utilisée. En tout cas, on doit se mettre formellement en garde contre des théories et des conclusions prématurées en ce domaine » (*ibid.*, p. 289).

✦ La charnière des années 1940 puis les années 50-60

Nous avons déjà mentionné R. Jakobson lorsqu'il préconisait en 1929 une convergence entre phonologies synchronique et diachronique, toutes deux au niveau systémique. Il introduisait alors la notion de stabilité systémique de manière assez elliptique. Par la suite, en mettant en perspective l'acquisition du langage, les troubles aphasiques et les caractéristiques typologiques des langues en termes de richesse et de stratification (termes préférés par R. Jakobson à complexité) dans cette même perspective systémique, il suggérait des directions de recherche qui sont encore d'actualité :

> "The stratification of components in a phonemic system proves to be strictly regular. These laws can be explained, however, only by

> considering and demonstrating their inner necessity. All attempts at atomistic interpretation, which necessarily explain only one aspect or a single phenomenon and are therefore never comprehensive, are clearly inadequate. Thus, the phonological laws of child language are not to be mechanically separated from the corresponding evidence of the languages of the world and of aphasia, and *the appearance of single sounds must not be treated in an isolated fashion without regard for their place in the sound system*" (Jakobson, 1941/1968:67 italiques ajoutées FP).

La seconde guerre mondiale marquera l'avènement de nouveaux paradigmes issus de l'ingénierie tournée vers l'effort de guerre. Dans cette tourmente dramatique, une ébullition intellectuelle intense repousse les frontières de la connaissance sur le langage humain, irriguée de recherches en télécommunications, en contrôle de systèmes automatiques, ou encore dans le champ naissant des sciences cognitives dont l'acte fondateur est peut-être la Conférence organisée par la Fondation Macy de 1942 à New York[42]. La charnière de la fin des années 1940 et du début des années 1950 est marquée par la tenue de la 5ème conférence Macy à laquelle participa Jakobson (en 1948) ainsi que par la *Speech Communication Conference* tenue au MIT au printemps 1950. Cette ébullition doit évidemment beaucoup à deux ouvrages majeurs issus des travaux de la guerre : en 1948 paraît *Cybernetics or Control and Communication in the Animal and the Machine* par N. Wiener qui marque l'entrée fracassante du XXe siècle dans l'ère de la rétroaction (*feedback*) et qui modifie profondément la notion de causalité (puisque la sortie d'un système peut en moduler l'entrée, mettant ainsi l'accent sur la question de *l'émergence*). L'année suivante paraît *The Mathematical Theory of Communication*, ouvrage dans lequel C.E. Shannon donne une grille de lecture très mathématique de la communication et dont W. Weaver s'atèle à étendre la portée, sans doute au-delà même de la volonté de son créateur[43].

La vision ingénieriste de la communication parlée donna lieu à plusieurs publications (par exemple celles issues de la conférence de 1950 au MIT, parues dans le numéro 22-6 du *Journal of the Acoustical Society of America*) dont l'apogée se situe probablement avec la publication de la première édition de *On Human Communication: A Review, a Survey, and a Criticism* en 1959 par Colin Cherry, mais nous insisterons plutôt sur ce que les linguistes, et en particulier Hockett, Jakobson et Martinet, feront de ces théories.

Dans sa revue de l'ouvrage de Shannon et Weaver, C.F. Hockett (1953) montre un réel enthousiasme pour la théorie de la communication, et en particulier pour les notions d'entropie et de redondance. Il discute en particulier de l'impact du choix des primitives dans le calcul d'entropie. Il écrit notamment que lorsqu'il établit le système phonologique d'une langue, le linguiste est confronté à des difficultés

---

[42] Au cours de cette conférence, C.E. Shannon présenta d'ailleurs ce qui n'était encore qu'une théorie de la communication en devenir !

[43] C.E. Shannon notera ainsi plus tard : "The basic results of the subject are aimed in a very specific direction, a direction that is not necessarily relevant to such fields as psychology, economics, and other social sciences" (Shannon, 1956).

(identification des allophones, distribution complémentaire, etc.) qui peuvent amener à plusieurs codifications. L'entropie de la parole elle-même étant probablement invariante par rapport au choix du code, il suggère alors de choisir la codification qui maximise l'entropie par phonème donnant en quelque sorte le codage le plus parcimonieux (*ibid.*, p.81-84). Deux importantes sources de difficulté sont par ailleurs identifiées ; tout d'abord, rien n'indique que l'utilisation de primitives morphémiques donnerait un résultat similaire :

> "It is by no means certain that calculation of the entropy of speech in terms of morphemes will give the same results as calculation in terms of phonemes –though it is certain that the computation is vastly more difficult." (*ibid.* p. 86).

D'autre part, il relève que contrairement à la télégraphie par exemple, la communication parlée peut transmettre *simultanément* des informations et il cite l'exemple de l'accentuation portée par les différents timbres vocaliques en anglais. Il suggère alors de traiter la communication comme un processus multilinéaire :

> "An alternative [to the multiplication of symbols according to the different stresses] is to modify Shannon's mathematical machinery so as to take care of a set of several linear sequences of symbols transmitted in parallel, where there are statistical restraints not only among those in the same linear sequence, but also between those in one linear sequence and those in others. I have no idea how complicated the necessary mathematical machinery might be, but I suspect that it would be very complicated indeed" (*ibid.*, p. 84).

Cette notion même de multilinéarité[44] sera ultérieurement approfondie dans Hockett (1955).

L'autre notion qui suscite un grand intérêt de la part de Hockett est celle de redondance, vue non pas comme un manque d'efficacité (en considérant que le canal de transmission ne serait pas employé de manière "optimale") mais à l'inverse comme une indication de l'extraordinaire flexibilité du système à s'adapter aux conditions de bruits les plus diverses (*ibid.*, p. 85). Plus tard, il fera même de cette redondance son premier universel phonologique : "in every language, redundancy, measured in phonological terms, hovers near 50%" (Hockett, 1966).

L'influence de la théorie de Shannon ne se limitera pas à Hockett et elle donnera un éclairage nouveau à de nombreux travaux sur la communication langagière, à commencer par les recherches entreprises par Jakobson sur les traits distinctifs. La structure même de ces traits, et leur caractère binaire en particulier, font écho à la théorie de l'information et Jakobson a lui-même tenté de formaliser ce lien en reliant le codage des systèmes phonologiques en système de traits distinctifs (pertinence et complexité algorithmique) à l'entropie de la parole (Cherry, Halle &

[44] Au sujet de la « machine mathématique » évoquée par Hockett, on peut noter aujourd'hui qu'il décrivait là les grandes lignes de ce qu'allaient être les modèles de Markov cachés parallèles quelques dizaines d'années plus tard…

Jakobson, 1953) ainsi qu'en pointant lui aussi l'importance du principe de redondance (Jakobson & Halle, 1956/2002:20).

Dans un autre contexte, en lisant par exemple le paragraphe consacré par A. Martinet à *l'information* dans ses *Éléments de linguistique générale*, il est évident que le cadre probabiliste employé est directement transcrit de la théorie shannonienne (Martinet, 1960/1973:181-184). Pourtant, à notre connaissance, Martinet ne cite jamais Shannon dans la bibliographie de ses ouvrages, ce qui est pour le moins surprenant. À l'inverse, tout au long de sa carrière, il revendiqua clairement l'influence de Zipf, comme nous allons le voir. Dès ses ouvrages d'avant-guerre (e.g. Martinet, 1933), il mettra en exergue le rôle potentiel du rendement fonctionnel dans l'évolution des langues, tout en se gardant de lui donner une importance exclusive. Tout comme Troubetzkoy avant lui, il poursuit la réflexion sur l'approche la plus correcte à employer pour déduire le rendement fonctionnel à partir de l'analyse des fréquences des oppositions ; s'il met l'accent sur la fréquence d'usage, il met aussi en garde contre ce qui peut être artéfactuel (Martinet, 1955:54-59) : dans sa vision, les paires « émonder-émender » où les deux termes sont caractérisés par de très faibles fréquences d'usage et « blond-blanc » nettement plus fréquents ne doivent probablement pas avoir le même poids dans l'estimation du rendement fonctionnel de l'opposition /ɔ̃-ɑ̃/. Pour autant, la très grande fréquence d'usage de l'article indéfini *un* génère une surreprésentation de la voyelle /œ̃/, tout comme peuvent le faire certaines désinences morphologiques (voir également Blevins, 2004:204-209).

Il affinera ensuite sa conception systémique du langage au sein duquel des forces antagonistes interagissent, apportant ainsi sa contribution à la vision du langage comme système dynamique complexe. Dans l'*Économie des changements phonétiques* (Martinet, 1955), il rend hommage à Zipf et lui sait tout particulièrement gré d'avoir mis l'accent sur un lien éventuel entre fréquence et complexité linguistique. Par contre, il n'adhère pas à la définition de la complexité avancée par Zipf, notant en particulier que des articulations complexes peuvent tout à fait être maintenues dans des items lexicaux fréquents. Là où Zipf voyait la simplification uniquement sous l'angle de la disparition de traits articulatoires lorsque la fréquence augmente, il s'agit selon Martinet d'une diminution du degré de netteté articulatoire, affectant tout le segment et pas seulement un trait (*ibid.* p132-137). Tout en prenant ses distances sur la méthode d'évaluation de la complexité, Martinet fait sien le principe du moindre effort (Zipf, 1949) et reconnaît qu'il est à l'œuvre dans la communication parlée de manière antagoniste avec le besoin de communiquer son message.

Par la suite, Martinet reformulera le postulat de Zipf sous la forme d'une « relation entre la fréquence d'une unité linguistique et sa forme, de telle sorte que toute variation de la fréquence entraîne un changement dans l'aspect phonique » (Martinet, 1962/1969:163) en donnant à l'information le rôle de chaînon logique entre fréquence et complexité : une augmentation de la fréquence d'une unité entraîne une diminution de l'information véhiculée (dans le sens shannonien). La tendance à l'économie exerce alors une pression visant à réduire son coût,

entraînant ainsi des changements dans l'aspect physique de cette unité. Il s'agit là du principe fondamental de l'économie linguistique, qui stipule que « la quantité d'énergie dépensée à des fins linguistiques tendra à être proportionnelle à la quantité d'information que l'on doit transmettre » (Martinet, 1962/1969:167). Martinet reconnaît cependant que les choses ne sont pas si simples, d'abord pour des raisons sociolinguistiques (influence du prestige, existence d'un conservatisme) et surtout car la communication étant majoritairement en contexte bruité, la redondance est nécessaire et donc souhaitable pour garantir sa robustesse.

Au final, Martinet se place dans un cadre fonctionnel fortement influencé par le schéma général de la théorie de la communication, mais en en relevant également les limites. Il suggère en particulier de renoncer aux formulations mathématiques pour deux raisons : d'une part, la quantification de l'énergie – et donc du coût – est très approximative (il est difficile de faire autre choses que de compter des unités : ajouter un phonème coûte plus) et d'autre part la théorie de la communication ne peut pas appréhender les facteurs sociolinguistiques. On peut donc juste essayer de déterminer dans quelles directions la variation de certains facteurs à des chances de provoquer la variation d'autres facteurs. Pour cela, Martinet insiste sur la prise en compte de quatre facteurs : a) le nombre d'unités parmi lesquelles le locuteur à le choix à un point donné de l'énoncé, b) la probabilité de ces unités, c) leurs coûts respectifs et d) l'information qu'elles véhiculent. (Martinet 1960/1973:82, Martinet 1962/1969:167).

Cette section débutait par une citation de Jakobson datant de 1941 et dans laquelle il insistait sur l'importance des approches systémiques. Elle finira par une autre citation du même Jakobson, écrite trente ans plus tard, alors que la vague générativiste est passée par là, et qui marque l'entrée de plein pied dans l'ère moderne des sciences de la complexité :

> "Like any other social modelling system tending to maintain its dynamic equilibrium, language ostensively displays its self-regulating and self-steering properties. Those implicational laws which build the bulk of phonological and grammatical universals and underlie the typology of languages are embedded to a great extent in the internal logic of linguistic structures, and do not necessarily presuppose special 'genetic' instructions" (Jakobson, 1973:48).

Ainsi, Jakobson, recommande-t-il d'étudier à la fois la typologie linguistique et la linguistique historique comme résultant de la nature même des langues, considérées comme des systèmes dynamiques complexes.

✦ L'avènement des sciences de la complexité

Les idées exprimées par Jakobson dans la citation ci-dessus ne traduisent pas une prise de position isolée, et, à partir des années 1970, de nombreuses études du langage ont adopté des paradigmes issus des sciences de la complexité, qu'il s'agisse d'émergence, d'auto-organisation, de régulation, de systèmes multi-agents, etc. Ce mouvement s'accélèrera jusqu'au milieu des années 1990, pour aujourd'hui imprégner l'ensemble des problématiques liées au langage humain. Dresser un

panorama fidèle de ces recherches dépasse largement le cadre de ce chapitre, tant elles ont été protéiformes et foisonnantes (voir par exemple Steels, 2000 ; Brighton, 2003 et Kirby, 2007 pour des synthèses récentes, ou Petitot-Cocorda, 1985 pour une approche plus ancienne construite sur la théorie des catastrophes de René Thom). Nous nous focaliserons plutôt dans la section suivante sur les travaux ayant directement nourri notre réflexion sur les systèmes phonologiques. *A contrario*, plusieurs ouvrages récents qui font la part belle à la notion de complexité ne seront pas discutés du fait qu'ils se focalisent principalement sur des niveaux morpho-syntaxiques. Nous pensons en particulier à Dahl (2004) et Hawkins (2004). Ö. Dahl dresse, dans le second chapitre de *The growth and maintenance of linguistic complexity* (intitulé *Complexity, order, and structure*), un inventaire de dimensions potentiellement pertinentes pour mesurer la complexité des langues et son ouvrage s'ouvre sur un chapitre intitulé *Information and redundancy* qui réintroduit de manière claire ces notions. Dans *Efficiency and Complexity in Grammars*, paru la même année, John A. Hawkins tente d'expliquer les tendances universelles observées typologiquement au sein des grammaires par des contraintes fonctionnelles, insistant en particulier sur les aspects d'efficience. Ces deux ouvrages s'intéressent également au rôle de la complexité dans l'évolution des langues et en cela, ils sont emblématiques d'un mouvement qui prend maintenant de l'ampleur (depuis Comrie, 1992, jusqu'aux contributions rassemblées dans Sampson, Gil & Trudgill, à paraître).

3.3. PISTES DE RECHERCHE EN COURS

Documents de référence : Annexes C.9, C.10 et C.11

Marsico, E., Maddieson, I., Coupé, C. & Pellegrino, F. 2004. "Investigating the "hidden" structure of phonological systems", *Proceedings of the 30th Meeting of the Berkeley Linguistic Society*, 256-267.

Coupé, C., Marsico, E. & Pellegrino, F. À paraître. "Structural complexity of phonological systems", in Pellegrino, F., Marsico, E., Chitoran, I. & Coupé C. (eds), *Approaches to Phonological Complexity*, Mouton de Gruyter, à paraître.

Pellegrino, F. Marsico, E. & Coupé C. En préparation. "A Cross-linguistic investigation of the phonological information rate"

Comme nous venons de le voir, les notions de complexité et d'information irriguent les recherches sur le langage dès le milieu du XXe siècle. Au cours des années 1970 à 1990 s'est ensuite développé le concept plus vaste de « Sciences de la complexité » et on peut émettre le constat apparemment paradoxal que cet essor a en partie occulté les questions antérieures, même si depuis maintenant une dizaine d'années la théorie de l'information et la quantification de la complexité linguistique suscitent un regain d'intérêt, en particulier en phonologie (e.g. Goldsmith, 2000 ; Aylett & Turk, 2004 ; Harris, 2005 ; Hume, 2006 ; Maddieson, 2006 ; Shosted, 2006). De fait, deux approches se revendiquant de la complexité coexistent aujourd'hui.

La première approche, orientée « systèmes dynamiques complexes », hérite de l'étude des systèmes non linéaires et porte principalement sur la modélisation de la production de la parole, de la perception de la parole, ou de leur interaction. Dans ce

cadre, il s'agit surtout d'utiliser et de développer des *outils* et des *modèles* permettant de résoudre des problèmes résistant à des approches plus classiques. Le système étudié est alors considéré comme complexe de par ses propriétés, sans que sa complexité elle-même ne soit nécessairement quantifiée. Les notions d'émergence et d'auto-organisation, tout comme les mécanismes de rétroaction et de régulation nécessaires, sont alors invoqués et mis en œuvre dans des modèles dynamiques ou des simulations multi-agents (voir Demolin & Soquet, 2001 ; Nam & Saltzman, 2003, parmi beaucoup d'autres).

La seconde approche s'intéresse plus directement à la notion de complexité elle-même et elle est imprégnée des recherches menées en typologie phonologique et en diachronie. La complexité d'un système ou d'un phénomène est vue comme graduelle (un système phonologique donné peut être plus complexe qu'un autre) et multidimensionnelle (complexités des systèmes vocaliques ou consonantiques, des systèmes de tons, des processus phonologiques, etc.). De tels travaux postulent, implicitement ou explicitement, que ces complexités sont des facteurs pertinents pour la compréhension du fonctionnement et de la dynamique des langues, sur les plans phylogénétique et/ou ontogénétique. Un exemple d'une telle hypothèse est celle – déjà mentionnée plus haut – d'une compensation entre les composantes de la grammaire, une phonologie « simple » étant contrebalancée par une morphologie ou une syntaxe plus « complexe » et vice-versa[45]. Sans aller nécessairement aussi loin, il nous semble intéressant d'étudier la dynamique des langues sous l'angle de la complexité : lors de l'acquisition précoce, la complexité globale du système linguistique en maturation évolue-t-elle de manière semblable d'une langue à l'autre ? Observe-t-on des diminutions ou augmentations abruptes de la complexité d'un système phonologique au cours de son évolution diachronique ? Cette dernière question peut également être formulée sous l'angle des systèmes dynamiques complexes, en visualisant le langage comme un système dynamique multistable, et un tel changement abrupt comme une transition de type catastrophique entre deux états stables. Un tel formalisme est en particulier développé par Christophe Coupé dans le chapitre 6 de sa thèse (Coupé, 2003).

De manière générale, les réponses à de telles questions ne sont pas triviales et nous en sommes aux balbutiements, en particulier du fait que l'on ne se les posait pas, encore récemment, du moins dans ces termes. Par exemple, lorsque J.J. Ohala pointait qu'un inventaire phonologique particulièrement exotique comme { ɗ k' ts ɬ m r | } n'est pas observé dans les langues ayant un faible nombre de consonnes, il suggérait qu'un principe d'économie est à l'oeuvre au niveau systémique[46] (Ohala, 1980; mais voir aussi, Ohala, à paraître). Par conséquent, on peut en déduire qu'un tel système est trop complexe pour être viable. Mais trop

[45] Cette hypothèse a également été suggérée à l'intérieur même d'une composante linguistique (phonologie par exemple), sans qu'elle ne soit réellement avérée. Ce point est en particulier discuté au travers d'UPSID dans Maddieson (1984:17-21).

[46] La même analyse peut être faite à propos des systèmes vocaliques ; voir par exemple Maddieson, (1984:16) : "The most frequent vowel inventory is /i, e, a, o, u/, not /i, ẽ, a̰, o̰, uˤ/."

complexe par rapport à quoi ? Et comment mesurer cette complexité ? Est-ce une question de nombre total de traits articulatoires, de complexité intrinsèque des phonèmes ou encore de taille et de densité de l'espace phonétique utilisé par la langue en question ? À notre avis, et contrairement à ce que J. Greenberg laissait entendre dans la citation mentionnée dans la note en bas de page 86, mesurer la complexité n'est pas trivial – même si l'on se restreint à un champ très spécifique comme la complexité articulatoire (e.g. Ohala, 1990:260 ; voir aussi Moon & Lindblom, 2003) – et les outils adéquats font encore souvent défaut. À titre d'exception, Lindblom & Maddieson, (1988) ont approfondi cette question et ils ont proposé de répartir l'ensemble des consonnes en trois groupes (consonnes simples, élaborées et complexes) en fonction d'une évaluation phonétique de leur complexité articulatoire. En analysant la distribution des segments dans UPSID (Maddieson, 1984), ils ont ainsi mis en évidence que les langues ont tendance à recruter les consonnes et les voyelles au sein d'un espace phonétique adaptatif, en fonction de la taille globale du système phonologique.

À ce jour, relativement peu de travaux concilient les perspectives *système dynamique* et *complexité* et Björn Lindblom, en collaboration avec des collègues issus de différents champs disciplinaires, est l'un des rares à réellement proposer une approche unifiée, construite à partir de plusieurs axes de ses recherches depuis le début des années 1970 : relation articulatoire-acoustique (Lindblom & Sundberg, 1970) ; simulation des systèmes vocaliques (théories du contraste maximal, suffisant et de la dispersion adaptative, Liljencrants & Lindblom, 1972 ; Lindblom, 1986 ; Lindblom & Engstrand, 1989) ; typologie et organisation des espaces vocaliques et consonantiques (e.g. Lindblom & Maddieson, 1988) rôle de la communication et de l'information déterminant dans l'interaction entre production et perception (*H&H theory*, Lindblom, 1990a) l'ensemble aboutissant à une approche fondamentalement dynamique, dans le sens de la phylogénie et de l'ontogénie (Lindblom, 1998, 1999, et jusqu'aux travaux les plus récents, e.g. Lindblom, 2005 ; Lindblom *et al.*, soumis).

D'autres études sont également pertinentes sur les deux plans, en particulier celles entreprises dans le cadre de la *Frame/Content Theory* (e.g. MacNeilage, 1998 ; MacNeilage & Davis, 2000 ; MacNeilage, Davis & Matyear, 2002 ; Kern & Davis, à paraître) ou dans la lignée des travaux initiés par la phonologie articulatoire (e.g. Browman & Goldstein, 1992 ; Studdert-Kennedy & Goldstein, 2003) et pour lesquels des problématiques d'émergence phylogénétique de la complexité phonologique ou de trajectoire d'acquisition ontogénétique de la complexité peuvent être mises en évidence. Les propositions de M. Studdert-Kennedy (1998 ; 2000), sur la base du *particulate principle* (Abler, 1989), relèvent également de ce cadre. Sur un autre plan, l'approche déductive mise en place par R. Carré lors du développement du modèle DRM (*Distinctive Region Model*) fait un usage évident du principe de parcimonie, puisque la structure émergente des huit régions distinctives du modèle permet d'obtenir un contraste acoustique maximal avec un minimum de déplacement physique et qu'il est fait explicitement référence à un principe d'économie des commandes, dans le sens donné en automatique (Carré & Mrayati, 1988 ; Carré, 2004, 2008).

Chaque langue est un système en évolution, façonné sous l'influence de contraintes de nature externe ou interne (Labov, 2001) et l'hypothèse la plus courante est que ces contraintes agissent de manière antagoniste (e.g. Lindblom, 1990b ; Schwartz *et al.*, 1997), même si une synergie est plus rarement évoquée (cf. par exemple Bradlow, 2002). L'existence de telles contraintes universelles, associée au fait que de nombreux degrés de libertés et de redondance subsistent dans la communication, permet d'envisager un vaste ensemble de systèmes phonologiques fonctionnels. Cette conception a été formulée maintes fois, et on y a vu de longue date un procédé élégant pour concilier l'existence de tendances universelles et la variabilité observée dans les langues du monde :

> « Il n'existe pas de langue où rien ne soit motivé ; quant à en concevoir une où tout le serait, cela serait impossible par définition. Entre les deux limites extrêmes – minimum d'organisation et minimum d'arbitraire – on trouve toutes les variétés possibles. Les divers idiomes renferment toujours des éléments des deux ordres – radicalement arbitraires et relativement motivés – mais dans des proportions très variables, et c'est là un caractère important, qui peut entrer en ligne de compte dans leur classement » (de Saussure, 1916:183).

Pour autant, l'existence de telles contraintes ne résout pas tout et en particulier certains, comme J. Blevins, n'y voient qu'une reformulation d'un processus téléologique déguisé : "In the majority of models in which teleological explanations are suggested, it is the perpetual conflict between effort minimization and contrast maximization which leads to variety and complexity in the world of sounds" (Blevins, 2004:71). Elle explique également que : "Sound change happens, but it does not occur in order to make speech easier to articulate, easier to perceive or easier to transmit; it does not necessarily result in a more symmetrical, more stable or generally improved phonological system; for every case where it happens, there is a parallel case where it does not happen" (Blevins, 2004:45).

En d'autres termes, J. Blevins réfute l'idée que des critères d'efficience (c'est-à-dire prenant en compte à la fois l'efficacité et l'effort nécessaire) aient un pouvoir explicatif des changements phonétiques et elle considère comme apparentées les tentatives d'explications téléologiques et fonctionnelles, même si elle reconnaît que la fonction communicative du langage et l'usage (*language use*) peuvent jouer un rôle mineur dans les changements phonétiques (*ibid*, p. 17). En cela, il nous semble qu'elle restreint assez largement ce que J.J. Ohala définissait comme une vision non téléologique : "(…) a non-teleological view of sound change, that is, that neither speaker nor hearer chooses – consciously or not – to change pronunciation" (Ohala, 1990:266). Selon nous en effet, il n'y a pas d'incompatibilité entre 1. le fait que le locuteur ou l'auditeur fasse un tirage aléatoire d'un son produit/perçu parmi un ensemble de possibles, caractérisé par une fonction de densité de probabilités et 2. le fait que cette fonction de densité de probabilités soit biaisée en réponse à des contraintes liées directement à la production ou à la perception, mais aussi pour des raisons fonctionnelles, et en particulier de niveau d'efficience suffisante (*"Sufficiently good" solution*, selon la terminologie proposée par Lindblom, 1990b:89).

Lorsque l'on cherche à identifier les contraintes universelles pesant sur les systèmes phonologiques, force est de constater qu'elles ne sont pas si universelles que cela, et que chaque tendance universelle est accompagnée de contre-exemples plus ou moins rares dans les langues du monde. Selon nous, cela implique soit que les contraintes responsables de ces tendances se sont relâchées au cours de l'évolution des langues[47], soient qu'elles ont été contrebalancées par des contraintes plus prégnantes. En particulier, les contraintes sociolinguistiques et « écologiques » sont largement susceptibles d'intervenir dans cette régulation : un système phonologique n'évolue pas à partir de rien, mais à partir d'un état antérieur qui détermine ainsi en quelque sorte les conditions initiales de l'évolution. De même, les populations de locuteurs sont en contact (éventuellement interne : plurilinguisme) avec d'autres langues et l'emprunt constitue un processus qui influence l'évolution des systèmes. Comme le souligne I. Maddieson : "It appears that the evidence is heavily in favor of the view that segments are most likely to be borrowed when there are already appropriate segments to promote the adoption of the new segment" (Maddieson, 1986:4).

Les éléments mentionnés dans ces quelques pages pourraient amener au constat pessimiste que le langage, même restreint au seul niveau phonologique, est un système *trop* complexe pour être efficacement étudié : en effet, que nous apprennent les contraintes perceptives ou articulatoires si leur pondération relative fluctue en fonction de l'environnement immédiat des locuteurs (bruit ambiant, mais aussi aspects pragmatiques de la communication) ? Qu'attendre d'un modèle d'évolution fondé sur l'étude des changements phonétiques attestés si la dimension sociolinguistique en est absente ? Selon nous, pourtant, le tableau n'est pas aussi noir et la richesse de la problématique justifie pleinement de multiplier les points de vue. En substance, il est nécessaire de faire progresser nos connaissances du niveau le plus individuel (production et perception de la parole, acquisition du langage) au niveau le plus général (typologie) en passant évidemment par les niveaux de la langue (évolution, vision systémique). Dans cette gigantesque tâche, il est de plus nécessaire de prendre en compte la variation observée aux différents niveaux, non pas comme un épiphénomène mais comme la substance même du caractère dynamique du langage.

Le travail que Christophe Coupé, Egidio Marsico et moi-même avons entrepris ces dernières années s'inscrit dans le vaste champ d'investigation ouvert par les travaux évoqués ci-dessus. Nous n'avons évidemment pas l'ambition de couvrir son intégralité et nous nous consacrons plus humblement à deux aspects, brièvement présentés dans les sections suivantes[48]. Le premier porte sur une vision systémique

[47] Dans le cadre de la théorie Frame/Content, il est par exemple envisageable que lors de l'émergence du langage, la contrainte d'alternance CV stricte ait été très forte et que, les langues se développant et l'information devenant plus redondante et distribuée, cette contrainte se soit relâchée jusqu'à permettre des types de syllabes complexes, voire sans noyau vocalique, mais présentant cependant une modulation acoustique (voir aussi Ohala & Kawasaki-Fukumori, 1997 ; Demolin, 2008 ; Ohala, 2008).

[48] Ces recherches s'inscrivent dans le cadre plus vaste du projet ANR NT05-3_43182 CL² « Complexité, Langage et Langues » (2005-2009 ; coordinateur : François Pellegrino).

des inventaires phonologiques ; l'hypothèse sous-jacente est qu'un système phonologique est, à un instant donné, une convention sociale émergeant à partir de la substance phonétique des échanges entre locuteurs de la langue. En tant que motif émergent, il doit nécessairement refléter certaines des contraintes opérant sur lui ou sur ses constituants. Bien entendu, étant donné la difficulté de la tâche, il est plus raisonnable de s'intéresser à un échantillon de systèmes de manière à faire émerger des régularités plutôt qu'aux systèmes pris individuellement. Cet axe de nos recherches, basé sur la base de données UPSID, se situe donc directement dans la continuité de travaux tels que ceux de Maddieson (e.g. 1984, 2006), Lindblom & Maddieson (1988), Vallée (1994), Vallée *et al.*, (2002), Clements (2003a, 2003b).

Le second axe de recherche poursuivi porte directement sur l'encodage de l'information phonologique lors de la communication parlée, dans le cadre de la théorie de l'information (Shannon & Weaver, 1949). I. Maddieson souligne à juste titre que : "The restrictions on inventory size may therefore not be the theoretical ones relating to message density and channel capacity in language processing. Although such considerations have been the most widely discussed, they are far from the only ones likely to influence the typical language inventory" (Maddieson, 1984:8). Il y a cependant fort à parier que des interactions puissent être mises en évidence entre les dimensions phonologiques paradigmatique et syntagmatique (e.g. Nettle, 1995), et nous nous sommes donc intéressés à la quantification et à l'encodage de l'information phonologique véhiculée dans la communication parlée, considéré comme un médiateur entre la complexité des inventaires phonologiques et les processus de construction des énoncés parlés.

3.3.1. Organisation des systèmes phonologiques

✦ Les données

La base de données UPSID, dans sa version à 451 langues (Maddieson, 1984 ; Maddieson & Precoda, 1990) modifiée ultérieurement par Maddieson et Marsico (non publié), a été utilisée dans les travaux évoqués ici. Cette base a été utilisée par ailleurs dans de nombreux travaux typologiques mais elle a fait également l'objet de plusieurs critiques concernant sa représentativité, les choix de transcriptions pour certains systèmes, etc. (cf. la discussion dans Vallée, 1994:43-44).

Pour le type d'étude que nous menons, deux limitations intrinsèques à la structure d'UPSID sont à garder en mémoire. La première porte sur le fait que la base est une coupe synchronique de 451 trajectoires individuelles puisque « la structure linguistique (…) n'est donc, en fait, qu'un des termes d'une série d'équilibres instables » (Hagège & Haudricourt, 1978:37). La fréquence d'apparition d'un phénomène, qu'il s'agisse d'un trait, d'un segment ou d'un motif plus vaste, dans cet échantillon est donc à interpréter avec précaution. En effet elle résulte d'une interaction entre la stabilité du phénomène (et donc sa propension à se maintenir dans un système) et sa probabilité d'apparition diachronique, comme cela a été souligné par J. Greenberg :

> "The two factors of probability of origin from other states and stability can be considered separately. If a particular phenomenon can arise very frequently and is highly stable once it occurs, it should be universal or near universal. This could be true of front unrounded vowels. If it tends to come into existence often and in various ways, but it stability is low it should be found fairly often but distributed relatively evenly among genetic linguistic stocks. A possible example is vowel nasalization. If a particular property rarely arises but is highly stable when it occurs, it should be fairly frequent on a global basis but be largely confined to a few linguistic stocks, e.g. vowel harmony. If it occurs only rarely and is unstable when it occurs, it should be highly infrequent or non-existent and sporadic in its geographical and genetic distribution, e.g. velar implosives" J. Greenberg (1978:76).

La détermination objective de ce qui est extrêmement fréquent ou extrêmement rare dans la base est peu problématique. À l'inverse, dès qu'il s'agit de fréquences non extrêmes, déterminer si un phénomène est plutôt fréquent ou modérément fréquent, etc. n'est pas trivial. Il n'est donc pas simple de relier la fréquence d'un trait ou d'un motif à une notion de stabilité ou d'attractivité intrinsèque de ce trait ou de ce motif. Cependant, la base peut être un moyen de détecter de tels phénomènes présentant potentiellement des caractéristiques d'attracteurs et une étude comparative des langues où ce trait est avéré et de leur voisinage (historique et aréal) peut éventuellement lever le doute : si la présence du trait fluctue dans le voisinage en question, il est probable qu'il s'agit d'un trait transitoire alors que si l'on observe une forte présence du trait dans le voisinage, il s'agit sans doute d'un attracteur stable.

Le deuxième élément à garder à l'esprit est que la base rassemble des *inventaires* phonologiques et que plusieurs éléments évidemment pertinents font défaut : systèmes de tons et d'accents, contraintes phonotactiques, phénomènes d'harmonie et, en complément, fréquences d'occurrence des segments dans le lexique ou dans des corpus. Les inventaires ne portent en quelque sorte aucune information sur l'usage qui est fait des segments dans la formation des morphèmes et donc dans la langue[49].

Pour autant, UPSID est un matériau de tout premier plan pour explorer sur le plan typologique la réponse des langues aux contraintes pesant sur les systèmes phonologiques, les inventaires constituant un échantillon conçu pour être représentatif des langues du monde[50].

✦ Recherche d'indices sur les contraintes à l'œuvre

> "The new form [i.e. the new pronunciation that yields a potential sound change] gets tested implicitly on a number of dimensions: 'articulatory

---

[49] La nouvelle base de données en cours de conception par I. Maddieson corrige la plupart de ces limites (voir Maddieson, 2007, pour un exemple d'analyse possible).

[50] Les informations contenues dans UPSID pour les langues citées en exemple dans les pages suivantes sont rassemblées dans le Tableau 19.

> ease', 'perceptual adequacy', 'social value' and 'systemic compatibility'. If the change facilitates articulation and perception, carries social prestige and conforms with lexical and phonological structure, its probability of acceptance goes up. If the change violates the criteria, it is likely to be rejected" (Lindblom, 1998:245).

B. Lindblom identifie dans ce passage quatre dimensions particulièrement pertinentes dans l'implémentation des changements phonétiques, et par extension, pour l'évolution et la structure des systèmes phonologiques. Parmi eux, la dimension sociolinguistique (*social value*) est clairement hors de portée d'une étude basée sur UPSID. On peut par contre considérer que des éléments liés aux contraintes articulatoires et perceptuelles (*articulatory ease* et *perceptual adequacy*) peuvent être mis en évidence. Cependant, certaines de ces contraintes s'expriment *de facto* au niveau de la formation des syllabes ou au moins dans la coarticulation de sons (e.g. Maddieson, 1992 ; Oudeyer, 2006 ; Demolin, 2008) et ne sont donc pas directement observables à partir des inventaires. Des hypothèses, telles que celle du MUAF (*maximal use of available features*) ou du contraste maximal restent cependant potentiellement quantifiables. Pourtant, comme il a déjà été mentionné à plusieurs reprises, l'interaction de ces contraintes est extrêmement difficile à prédire à l'échelle d'un système phonologique et de son évolution. Par conséquent, nous considérons que les résultats obtenus (Marsico *et al.*, 2004 ; Coupé, Marsico & Pellegrino, à paraître) relèvent plus de cette interaction que de l'une ou l'autre des deux dimensions. De même, UPSID semble adapté à la mise en lumière d'éventuelles contraintes systémiques (*systemic compatibility*) et nous nous sommes proposés d'étudier plus avant cette dimension, en particulier par une approche structurelle des systèmes phonologiques.

Au final, cela nous a amené à étudier des indices liés :

- aux composants eux-mêmes (traits et segments) ;
- aux interactions deux à deux entre segments ;
- et à la topologie des systèmes phonologiques.

Au niveau méthodologique, nous avons souhaité limiter les biais possibles, en particulier en intégrant le minimum de connaissances *a priori* dans les analyses. Il persiste évidemment un biais lié à la description même des systèmes dans UPSID, mais nous avons également procédé à une évaluation de la robustesse des résultats vis-à-vis de l'ensemble de traits descriptifs adopté (Marsico *et al.*, 2004).

Alors que Lindblom & Maddieson (1988) ont identifié trois sous-ensembles de consonnes, dites simples (groupe I), élaborées (groupe II) et complexes (groupe III) en fonction de la complexité de leur articulation, nous sommes directement partis de la description des segments en traits pour évaluer le caractère basique des traits eux-mêmes, puis par extension, des segments et des systèmes. Un trait basique est défini comme étant indispensable à la formation d'un segment alors qu'un trait est non basique dans le cas contraire. Ainsi, *high*, *front*, *unrounded* et *vowel* sont des traits basiques alors que *long* est un trait non basique puisque si on l'enlève de la

description de /iː/, on obtient un segment attesté dans UPSID (en l'occurrence /i/) contrairement aux autres traits considérés (cf. ci-dessous).

(1) i {high front unrounded vowel}

(2) iː {long high front unrounded vowel}

Une différence majeure par rapport aux groupes I, II et III proposés par Lindblom & Maddieson (1988) est que dans notre étude, aucune distinction de complexité n'est faite selon le lieu d'articulation des consonnes ; ainsi, alors que les consonnes labiodentales, palato-alvéolaires (ou postalvéolaires), rétroflexes, uvulaires et pharyngales sont assignées dans leur étude au groupe II, il s'agit de traits basiques dans notre analyse.

Par extension, un segment est basique s'il est formé uniquement de traits basiques ; il est non basique dans le cas contraire. 315 des 833 segments d'UPSID sont ainsi basiques. De même, un système phonologique est basique s'il est formé exclusivement de segments basiques. Avec cette définition de basicité, et en considérant les systèmes dans leur intégralité (consonnes + voyelles + diphtongues + clicks), nous avons mis en évidence la proportion de segments basiques dans les 451 systèmes (cf. Figure 12). Parmi ceux-ci, 132 systèmes (soit 29 % des langues) sont totalement basiques. Le plus grand système basique est celui du bashkir (11 voyelles + 27 consonnes) et le parauk présente un très haut pourcentage de segments basiques (95 %) malgré sa taille élevée (77 segments). Ses quatre segments non basiques sont des consonnes aspirées (trois occlusives et une affriquée). Il faut cependant noter que cette langue utilise de manière intensive un contraste entre séries *breathy-voiced*, *voiced* et *voiceless* et l'on peut soulever la question de l'existence de variations de complexité intrinsèques aux différents modes laryngés.

Certaines langues présentent une importante proportion de segments non basiques malgré une taille relativement faible ; si l'on se limite aux 377 systèmes ayant jusqu'à 38 segments (taille du système du bashkir), 132 seulement sont basiques (soit 35 % de ces langues) et, en moyenne sur les 245 systèmes restants, 20 % de leurs segments sont non basiques.

Ces résultats suggèrent que, même pour des tailles de systèmes modérées ou faibles, le recours à des traits non basiques est courant dans les langues du monde. Par exemple, le maxacali (10 consonnes + 10 voyelles) fait un usage intensif du trait de nasalité, mais uniquement comme trait secondaire (5 voyelles sont nasales et 4 consonnes sont prénasalisées[51]).

[51] Notons cependant qu'en ce qui concerne les consonnes, la prénasalisation est totalement redondante avec le trait de voisement dans ce système.

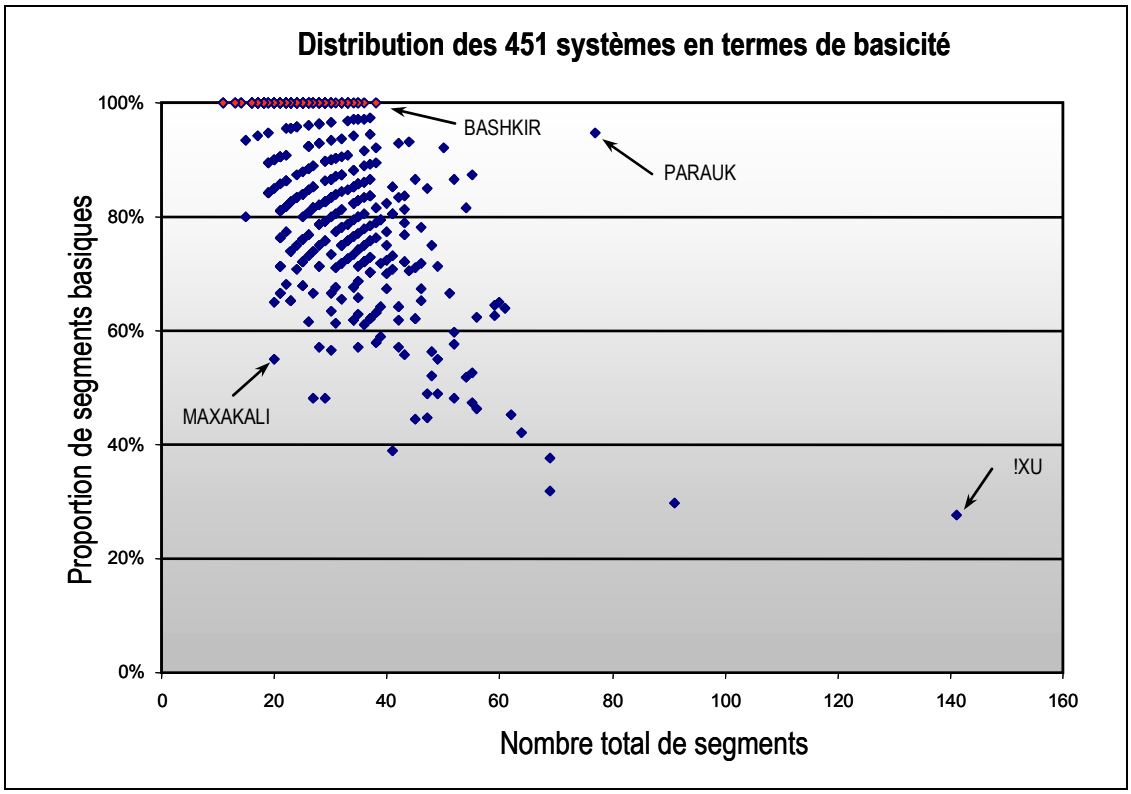


Figure 12 – Proportion de segments basiques pour les 451 langues d'UPSID. Les motifs rouges correspondent aux systèmes totalement basiques et les motifs bleus aux systèmes non basiques. Les flèches indiquent trois systèmes commentés dans le texte.

L'extension de l'espace phonétique par utilisation de segments non basiques est donc observée lorsque la taille de leur système phonologique augmente, conformément au *size principle* (Lindblom & Maddieson, 1988). Cependant, on constate également qu'en l'absence d'une telle contrainte induite par la taille du système, la contrainte supposée d'économie des traits (Clements, 2003a et 2003b) n'est pas stricte au point d'empêcher la formation de systèmes non basiques de taille réduite.

Un autre résultat porte sur les relations existant entre segments basiques et non basiques sous l'angle de la dérivation ; pour chacun des 315 segments basiques, le décompte des segments qu'il peut générer par complexification (ajout de traits non basiques) a été effectué, ce qui permet d'identifier les segments les plus « productifs ».

Pour les voyelles, il s'agit, par ordre décroissant, de /a/, /i/, /o/, /u/, /e/, etc. comme indiqué dans le Tableau 12. On constate que les voyelles les plus productives sont également les plus fréquentes dans les langues du monde (au moins pour les cinq premières). On peut par ailleurs vérifier que les sept voyelles présentes dans plus du quart des langues d'UPSID sont présentes dans cette liste au sept premières places ce qui renforce l'idée de l'existence d'un lien entre la fréquence et le nombre de segments accessibles par dérivation (toutes les voyelles les plus fréquentes ont un fort degré de dérivation et toutes les voyelles ayant un fort degré de dérivation sont fréquentes).

Tableau 12 – Classement des voyelles basiques arrivant aux dix premiers rangs pour le nombre de voyelles dérivées dans UPSID (adapté d'après Marsico *et al*., 2004).

RANG	SEGMENT	NOMBRE DE SEGMENTS DÉRIVÉS	FRÉQUENCE OBSERVÉE DANS UPSID
1	/a/ voiced low central unrounded	12	87 %
2	/i/ voiced high front unrounded	11	87 %
2 ex-aequo	/o/ voiced higher-mid back rounded	11	69 %
4	/u/ voiced high back rounded	9	82 %
4 ex-aequo	/e/ voiced higher-mid front unrounded	9	65 %
6	/ɔ/ voiced lower-mid back rounded	8	36 %
7	/ɛ/ voiced lower-mid front unrounded	6	41 %
8	/ɨ/ voiced high central unrounded	5	14 %
8 ex-aequo	/ɣ/ voiced higher-mid back unrounded	5	4 %
10	/ə/ voiced mid central unrounded	4	17 %
10 ex-aequo	/ɷ/ voiced lowered-high back rounded	4	15 %
10 ex-aequo	/ɯ/ voiced high back unrounded	4	9 %
10 ex-aequo	/ɘ/ voiced higher-mid central unrounded	4	4 %
10 ex-aequo	/ɒ/ voiced low back rounded	4	4%
10 ex-aequo	/ø/ voiced higher-mid front rounded	4	3 %

La nature du lien reste par contre à établir ; en particulier l'hypothèse d'une causalité, dans un sens ou dans l'autre, n'est pas validée à ce stade. On peut en effet imaginer plusieurs explications à cette observation, dans un sens ou dans l'autre. Dans le sens d'une explication du potentiel de dérivation par la fréquence, il est possible que du fait de leur présence dans un grand nombre de langues, probablement depuis très longtemps, ces voyelles soient plus susceptibles d'être l'objet de processus de transphonologisation, générant ainsi au cours de l'évolution des séries de voyelles dérivées, que l'on retrouve alors plus ou moins sporadiquement dans l'échantillon d'UPSID.

En sens inverse, on peut supposer que ces segments étant capable de générer de nombreux segments dérivés, ils soient observés plus fréquemment dans les langues comme attracteurs en cas d'erreur de perception des traits dérivés, un peu à l'image des articulations neutres (*neutral articulations*) envisagées dans Lindblom & Maddieson, (1998).

On peut enfin envisager que cette relation entre dérivation et fréquence résulte d'une cause commune, liée aux caractéristiques intrinsèques des segments considérés. En particulier, on constate que les neuf premières voyelles (par ordre de rang dans le Tableau 12) sont des voyelles périphériques ; elles sont à la fois atteintes par les trajectoires qui constituent les primitives dans la théorie de R. Carré (Carré, sous presse) et susceptibles d'être sélectionnées dans le cadre des théories de la dispersion ou de jouer un rôle d'ancrage dans la perception (voir en particulier Polka & Bohn, (2003) et Schwartz *et al.*, (2005) pour une discussion). Il est cependant intéressant de remarquer que /y/ qui est une voyelle focale dans la théorie de la

dispersion-focalisation (Schwartz *et al.*, 1997) est à la fois peu fréquente (5 % des langues de l'échantillon) et peu « apte » à générer des voyelles dérivées (seules les contreparties longue et nasale sont attestées).

Si l'étude des voyelles n'amène pas d'interprétation triviale, le même type d'analyse sur les consonnes s'avère également intriguant (Tableau 13).

Tableau 13 – Classement des consonnes basiques arrivant aux dix premiers rangs pour le nombre de consonnes dérivées dans UPSID (adapté d'après Marsico *et al.*, 2004).

RANG	SEGMENT	NOMBRE DE SEGMENTS DÉRIVÉS	FRÉQUENCE OBSERVÉE DANS UPSID
1	/k/ voiceless velar stop	17	89 %
2	/tʃ/ voiceless postalveolar sibilant-affricate	14	42 %
3	/t/ voiceless alveolar stop	13	74 %
3 ex-aequo	/q/ voiceless uvular stop	13	12 %
5	/p/ voiceless bilabial stop	11	83 %
5 ex-aequo	/ts/ voiceless alveolar sibilant-affricate	11	24 %
7	/q͡χ/ voiceless uvular non-sibilant-affricate	10	1 %
8	/b/ voiced bilabial stop	9	64 %
8 ex-aequo	/d/ voiced alveolar stop	9	47 %
8 ex-aequo	/s/ voiceless alveolar sibilant-fricative	9	73 %
8 ex-aequo	/χ/ voiceless uvular non-sibilant-fricative	9	10 %

Une première observation porte sur la prédominance des consonnes non voisées dans ce tableau (9 segments sur 11) et sur la présence majoritaire du mode occlusif (6 occlusives, 3 affriquées et 2 fricatives). On constate également que les consonnes présentes se répartissent du lieu bilabial au lieu uvulaire, les consonnes gutturales étant absentes.

Enfin, il apparaît que les consonnes des premiers rangs sont plus productives que les voyelles de même rang. Le segment /k/ est ainsi le segment le plus productif des 315 segments basiques avec 17 segments dérivés (Tableau 14).

Contrairement à ce que l'on observait pour les voyelles, il ne semble pas y avoir de relation claire entre la fréquence du segment dans les langues et sa capacité à générer des segments dérivés. Ainsi, les trois consonnes uvulaires (occlusive, fricative et affriquée non voisées) sont plutôt rares et, inversement, des segments extrêmement fréquents dans les langues d'UPSID sont absents du tableau (par exemple les nasales /m/ et /n/). Si l'on se concentre sur /q͡χ/, il s'avère que cet effet est lié à 3 des 5 langues où ce segment apparaît et pour lesquelles des séries importantes de consonnes complexes sont exploitées aux différents lieux d'articulation (il s'agit de l'archi, de l'avar et du kabardian). Cette stratégie a pour conséquence qu'une consonne basique rare dans les langues du monde peut cependant générer un ensemble important de segments dérivés distincts. Il s'agit évidemment là d'un principe de réutilisation des traits disponibles qui évoque le principe MUAF *Maximal Use of Available Features* (Ohala, 1980, voir cependant les

nuances apportées dans Ohala, à paraître). On peut également envisager des effets systémiques, la présence d'un sous-système régulier dérivé de /q͡χ/ jouant un rôle stabilisateur (voir plus loin pour une discussion sur les aspects systémiques à proprement parler). Si l'un de ces principes est à l'œuvre, on peut l'interpréter du point de vue de la densité de systèmes : les langues auront tendance à utiliser des inventaires denses plutôt qu'épars, et donc à limiter la distance entre segments participant d'une même opposition.

Tableau 14 – Description des 17 segments dérivés de l'occlusive vélaire non voisée /k/ (d'après UPSID, version modifiée par Maddieson & Marsico, non publié).

DESCRIPTION DU SEGMENT	SYMBOLE UPSID
voiceless velar stop	k
voiceless velar stop with-breathy-release	k^{fi}
prenasalized voiceless velar stop	ŋk
prenasalized palatalized voiceless velar stop	$ŋk^{j}$
prenasalized labialized voiceless velar stop	$ŋk^{w}$
prenasalized voiceless aspirated velar stop	$ŋk^{h}$
voiceless preaspirated velar stop	^{h}k
voiceless velar ejective stop	k’
pharyngealized voiceless velar stop	$k^{ʕ}$
long voiceless velar stop	kː
palatalized voiceless velar stop	k^{j}
palatalized voiceless velar ejective stop	k
palatalized voiceless aspirated velar stop	k^{jh}
labialized voiceless velar stop	k^{w}
labialized voiceless velar ejective stop	k^{w}’
long labialized voiceless velar stop	k^{w}ː
labialized voiceless aspirated velar stop	k^{wh}
voiceless aspirated velar stop	k^{h}

Pour chaque langue d'UPSID, nous avons donc calculé pour chacun de ses segments la distance d'édition à son plus proche voisin, calculée en traits et sans autoriser de substitution en dehors des classes d'équivalences établies pour les traits basiques (listées dans le Tableau 15). Chaque classe a été établie automatiquement à partir de l'impossibilité des traits la constituant à entrer simultanément dans la description d'un segment et ces classes recoupent les classes naturelles traditionnelles. Ainsi des segments ne différant que d'un trait basique (mode, lieu voisement, etc.) sont distants d'une unité alors que deux segments A et B, dérivés du même segment basique par ajout d'un trait secondaire pour A et d'un autre trait secondaire pour B, seront distants de deux unités. Des exemples sont proposés dans le Tableau 16. De part la nature largement différenciée des traits de description vocalique et consonantique, le plus proche voisin d'une consonne sera toujours une consonne et de même, le plus proche voisin d'une voyelle sera toujours une voyelle.

Tableau 15 – Classes d'équivalences établies automatiquement à partir des traits descriptifs d'UPSID. Les traits spécifiques aux diphtongues ont été intégrés dans ce tableau.

NOM DE LA CLASSE D'ÉQUIVALENCE	TRAITS DE LA CLASSE D'ÉQUIVALENCE
Vowel frontness	front central back back-front central-front front-central back-central front-back central-back
Vowel height	high higher-mid higher-mid-high high-higher-mid high-low high-lower-mid low lowered-high lowered-high-higher-mid lower-mid lower-mid-high low-high raised-low raised-low-high high-mid mid lower-mid-higher-mid higher-mid-low higher-mid-mid low-higher-mid low-lower-mid mid-high mid-lower-mid lowered-high-high
Rounding	unrounded rounded unrounding rounding
Laryngeal settings	voiceless voiced narrow-voiceless breathy-voiced creaky-voiced
Place of articulation	bilabial alveolar dental labial-velar labiodental postalveolar palatal retroflex uvular velar epiglottal labial-palatal
Manner of articulation	stop approximant implosive nasal non-sibilant-fricative non-sibilant-affricate flap lateral-approximant lateral-fricative sibilant-affricate sibilant-fricative tap trill-or-unspecified affricate-trill fricative-flap fricative-trill lateral-flap

La redondance pour chaque système est calculée comme étant la moyenne des distances entre chaque segment et son plus proche voisin : un système au sein duquel chaque segment est distinct de son plus proche voisin par un unique trait aura une redondance de 1 alors qu'un système utilisant des segments plus distants aura une redondance supérieure.

Tableau 16 – Exemples de calcul de distances entre segments pour l'évaluation de la redondance au sein des systèmes phonologiques

PAIRE DE SEGMENTS	TRAITS PERTINENTS	DISTANCE
/i/ - /iː/	long vs. Ø	1
/i/ - /u/	unrounded vs. rounded front vs. back	2
/iː/ - /uː/	unrounded vs. rounded front vs. back	2
/iː/ - /ũ/	unrounded vs. rounded front vs. back long vs. Ø nasalized vs. Ø	4
/p/ - /d/	bilabial vs. alveolar voiceless vs. voiced	2
/p/ - /q/	bilabial vs. uvular	1
/p/ - /s/	bilabial vs. alveolar stop vs. sibilant-fricative	2
/pʷ/ - /qʷˤʰ/	bilabial vs. uvular pharyngealized vs. Ø aspirated vs. Ø	3

La Figure 13 illustre ce processus pour trois systèmes phonologiques hypothétiques (le système B est, en fait, attesté puisqu'il s'agit de l'inventaire de la variété de rotokas décrite dans UPSID). Le système A est parfaitement « régulier » ; il présente deux séries d'occlusives contrastées en voisement et chaque segment est

séparé de ses plus proches voisins par 1 trait. Sa redondance est donc de 1. Le système B, présente une série d'occlusives non voisées, et à chaque lieu d'articulation correspond une contrepartie voisée, qui présente cependant des variations de modes. La redondance du système est de (1 + 1 + 1 + 1 + 2 + 2)/6 = 1,33. Le système C présente plusieurs particularités ; du fait de la prédominance du lieu d'articulation alvéolaire, certains segments ne diffèrent que par le mode d'articulation (lieu et voisement étant identiques). Par contre les consonnes non alvéolaires sont distantes de leur plus proche voisin de 2 traits pour la nasale voisée bilabiale /m/ (mode + lieu) et de 3 traits pour l'occlusive éjective vélaire non voisée /k'/ (mode + lieu + trait éjectif). De plus elle présente la particularité de ne pas constituer un graphe connexe puisque les consonnes voisées et non voisées forment deux groupes non connectés. Sa redondance est particulièrement élevée : (1 + 1 + 1 + 1 + 1 + 2 + 3)/7 = 1,43. On définit ainsi une mesure liée à la densité des systèmes, c'est-à-dire à l'usage fait de l'espace phonétique « accessible » avec l'ensemble des traits présents dans chaque système.

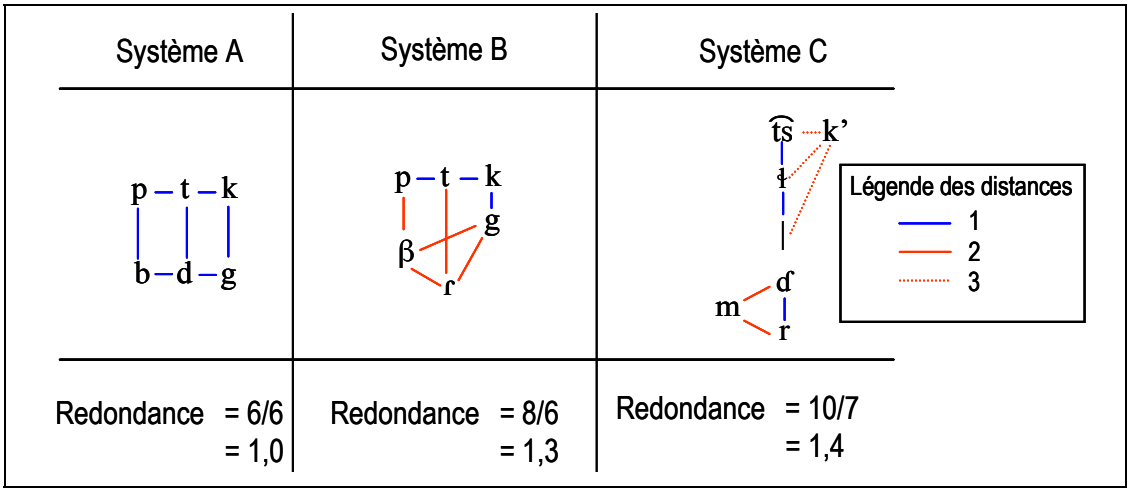


Figure 13 – Exemples de calcul de redondance sur des inventaires consonantiques théoriques. Les traits de couleur relient chaque segment à son ou ses plus proche(s) voisin(s). La couleur et le type de trait codent la distance entre les segments reliés (de 1 à 3 sur les exemples proposés).

La Figure 14 présente la distribution de la redondance dans la base UPSID. La redondance moyenne est de 1,06 et la redondance médiane est quant à elle de 1,04. 111 systèmes ont une redondance de 1 (valeur minimale possible). À titre indicatif, les redondances calculées pour les trois langues caucasiques (archi, avar et kabardian) évoquées précédemment et utilisant des séries importantes d'affriquées uvulaires non voisées, sont particulièrement faibles (respectivement 1, 1,02 et 1,02) de manière parfaitement compatible avec le MUAF.

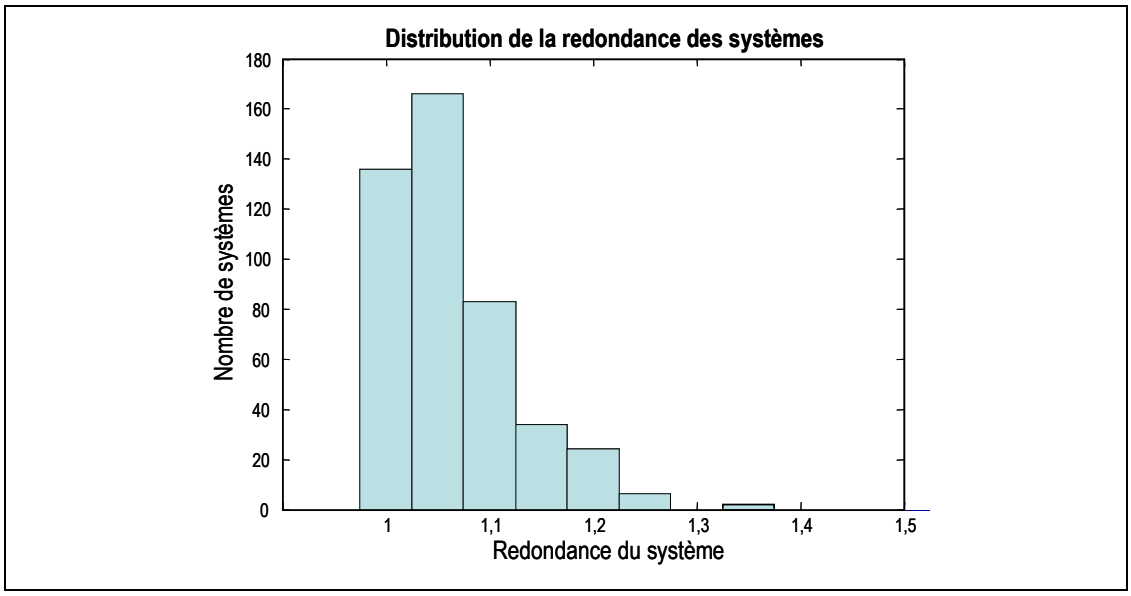


Figure 14 – Distribution de la redondance des 451 systèmes phonologiques d'UPSID. La redondance moyenne est de 1,06.

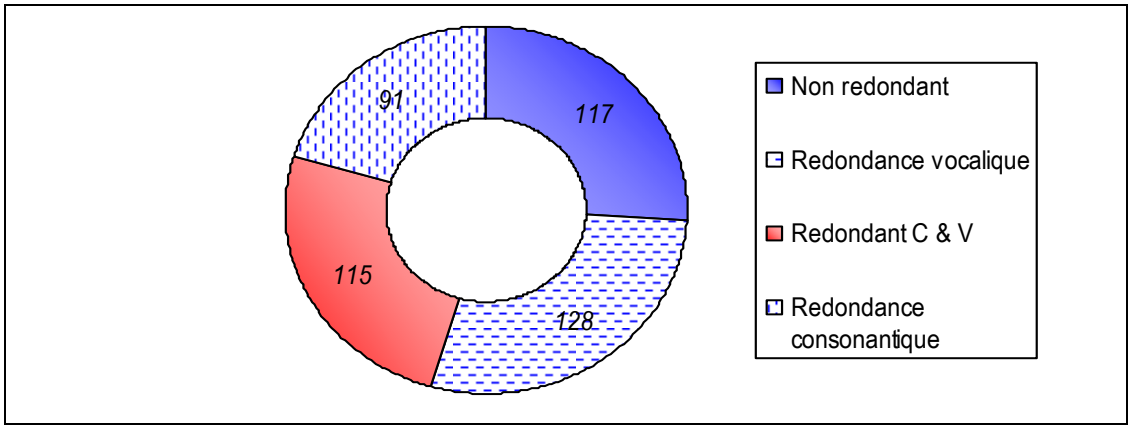


Figure 15 – Répartition des systèmes d'UPSID en fonction de la redondance observée indépendamment pour leurs systèmes vocaliques et consonantiques. Les diphtongues sont exclues de cette analyse (d'après Marsico *et al.*, 2004).

Seules deux langues ont une redondance supérieure à 1,3 (soit 30 % de redondance par rapport à un système minimalement redondant) : le piraha (1,36 pour un système à 11 segments) et le gadsup (1,33 pour 15 segments). Dans les deux cas, cette forte redondance est générée au niveau du système vocalique[52]. Cette observation soulève la question d'une éventuelle relation entre redondance vocalique et redondance consonantique pour chaque langue mais on constate qu'elles ne sont pas corrélées (r^2 = 0,01 ; corrélation de Pearson) et qu'elles semblent indépendantes comme le montre la Figure 15, même si un biais non

[52] Pour le piraha, il s'agit d'un système vocalique /i a o/ à 3 niveaux d'aperture pour 3 timbres ; la structure du système vocalique du gadsup est également relativement complexe : /i eː ɜ ạː u oː/. Il s'agit d'un système à 3 voyelles brèves et 3 voyelles longues pour lesquelles les oppositions reposent également sur des changements de timbres : /i~eː/, /ɜ~ạː/ et /u~oː/.

significatif est observé vers une occurrence plus forte de redondance vocalique que consonantique (test de Fisher ; p=0,45).

Avec la méthodologie sommairement décrite ci-dessus, il apparaît qu'un principe d'économie semble à l'œuvre dans les systèmes phonologiques, avec pour effet de privilégier des systèmes phonologiques denses, où les oppositions phonologiques minimales reposent sur une distance d'un unique trait plutôt que sur plusieurs dimensions. Il est cependant possible que cet effet soit en partie artefactuel. Considérons par exemple le pseudo-système C, reproduit de la Figure 13 ; sa redondance élevée vient uniquement de la présence de /m/ et /k'/ qui forment des sortes de segments « singletons » dans le système. Il aurait ainsi suffi que les mêmes traits aient été réutilisés pour former deux segments directement voisins (c'est-à-dire distants d'un unique trait) pour que la redondance diminue, sans pour autant garantir un système dense et faisant un usage maximal (ou au moins consistant) des traits disponibles. La Figure 16 illustre ce biais en représentant le système C précédant et un autre système théorique D, tout aussi exotique, mais pour lequel le calcul de redondance est de 1. En résumé, pour peu que le système étudié mette uniquement en jeu des oppositions minimales, il est considéré comme non redondant, même s'il est constitué de sous-ensembles de segments disjoints et que sa densité dans l'espace phonétique qu'il définit est faible[53]. Celle limitation nous a amené par la suite à considérer des approches de quantification de la complexité structurelle des inventaires, à partir de la théorie des graphes.

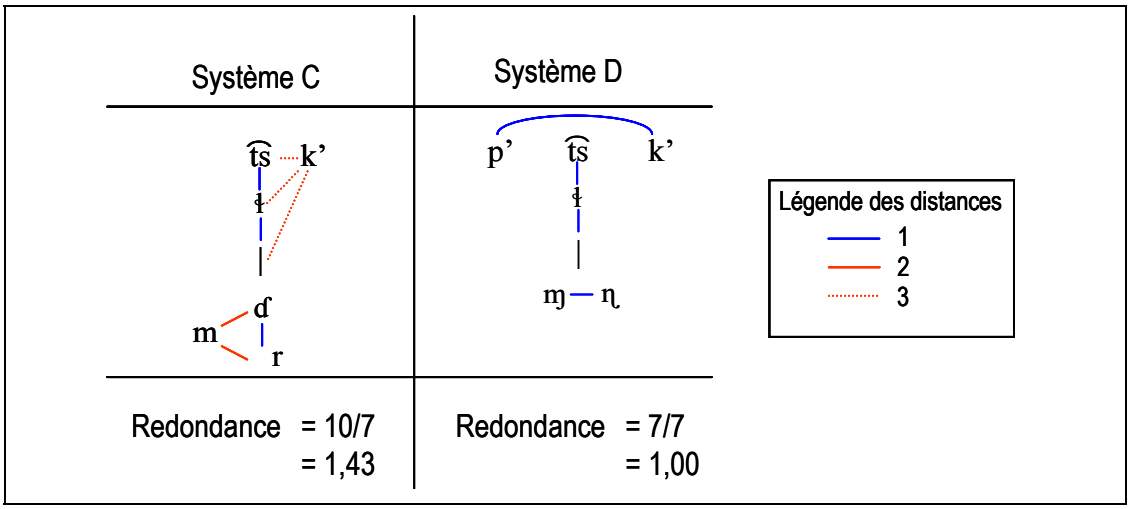


Figure 16 – Comparaison de la redondance pour deux systèmes théoriques atypiques.

Une alternative à la mesure de redondance à partir des distances entre segments consiste à quantifier la complexité structurelle des systèmes vocaliques et

[53] Ce constat fait écho à une remarque de Lindblom où il indique que le critère d'efficience articuloire-perceptuel ("the balance between articulatory and perceptual constraints in structuring phonetic inventories") ne suffit pas à expliquer pourquoi le MUAF s'applique : "[this balance] offers no guarantee that the derivation (…) will make *consistent use* of one secondary mechanism before moving to the next. In fact, for large systems, [it] predicts a collection of 'assorted bonbons' (…)" (Lindblom, 1998:250 ; italiques originales).

consonantiques pour essayer de prendre en compte l'usage consistant – ou non – des traits au sein des systèmes. C'est ce que nous avons proposé dans Coupé, Marsico & Pellegrino, (à paraître) où le lecteur pourra trouver la description de l'algorithme de construction des graphes à partir des inventaires que nous avons développé, la mesure de complexité elle-même étant la mesure de complexité *offdiagonal* (Claussen, 2004) qui a l'avantage important de ne pas être sensible à la taille du graphe mais à la présence de motifs réguliers. Par contre, l'algorithme est développé pour des graphes non valués, c'est-à-dire des réseaux où les arêtes sont toutes de même valeur, alors même que dans notre approche elles correspondent à des distances basées sur les traits constituant les segments. Dans sa version actuelle, notre algorithme ne tient donc pas compte de ces distances. Cette approche n'ayant pas encore donné de résultats particulièrement pertinents, elle ne sera évoquée que rapidement.

La Figure 17 illustre le comportement de l'algorithme pour quatre exemples de systèmes vocaliques simples. Le système à cinq voyelles le plus courant dans UPSID /i a e o u/ a une complexité structurelle de 0,64 alors qu'un système où les deux voyelles d'aperture moyenne seraient longues (schéma 2) a une complexité légèrement supérieure, tout en ayant le même nombre de voyelles. Le système présenté au schéma 3 présente un nombre plus important de voyelles (7), mais il est cependant caractérisé par la même complexité, alors qu'un système, à 7 voyelles également, mais de structure plus complexe atteint une valeur assez nettement supérieure (schéma 4).

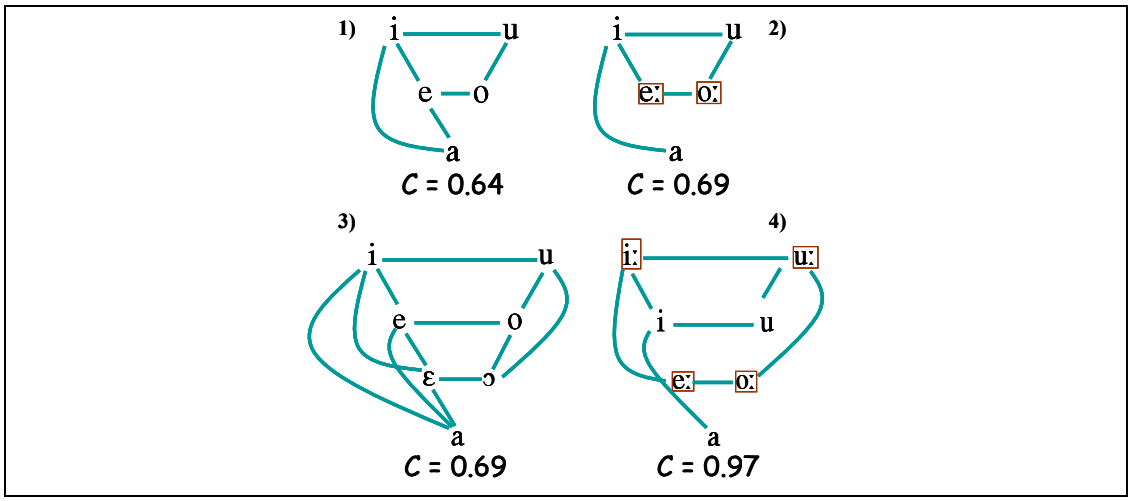


Figure 17 – Exemples de graphes établis pour 4 systèmes vocaliques de taille 5 (graphes 1 et 2) et 7 (graphes 3 et 4). La valeur de complexité *offdiagonal* C est indiquée pour chaque exemple (d'après Coupé, Marsico & Pellegrino, à paraître).

Si l'on s'intéresse à l'ensemble des systèmes vocaliques d'UPSID (Figure 18), une faible corrélation positive mais significative est observée entre la taille du système et sa complexité *offdiagonal* (r^2 = 0,16 ; p= 0,00) sans qu'il soit possible de dire s'il s'agit d'une complexification réelle lorsque la taille du système augmente ou d'un léger artéfact. Plusieurs langues à 4 voyelles ont des systèmes structurellement complexes (on peut citer par exemple l'alawa : /ɪ e a ʊ̟/). Le système vocalique le plus complexe est celui du kashmiri, qui s'avère également être le plus grand et qui,

surtout, met en jeu des traits secondaires de manière partiellement irrégulières (8 voyelles orales, 6 voyelles longues, 6 voyelles nasales, 6 voyelles nasales longues, 1 voyelle nasale vélarisée, le tout avec certaines différences de timbres vocaliques). Le kolokumaijo est par contre un exemple de système vocalique relativement grand (18 voyelles) mais de complexité modérée ; il met en œuvre une symétrie parfaite entre un sous-système oral à 9 timbres périphériques et le sous-système nasal correspondant : /i ɪ e ɛ a ɔ o ʊ u ĩ ɪ̃ ẽ ɛ̃ ã ɔ̃ õ ʊ̃ ũ/. Le naxi quant à lui présente une taille de système vocalique relativement faible mais une complexité importante : /i y ɛ ə̤ æ ɑ o ɯ ɯ̤/.

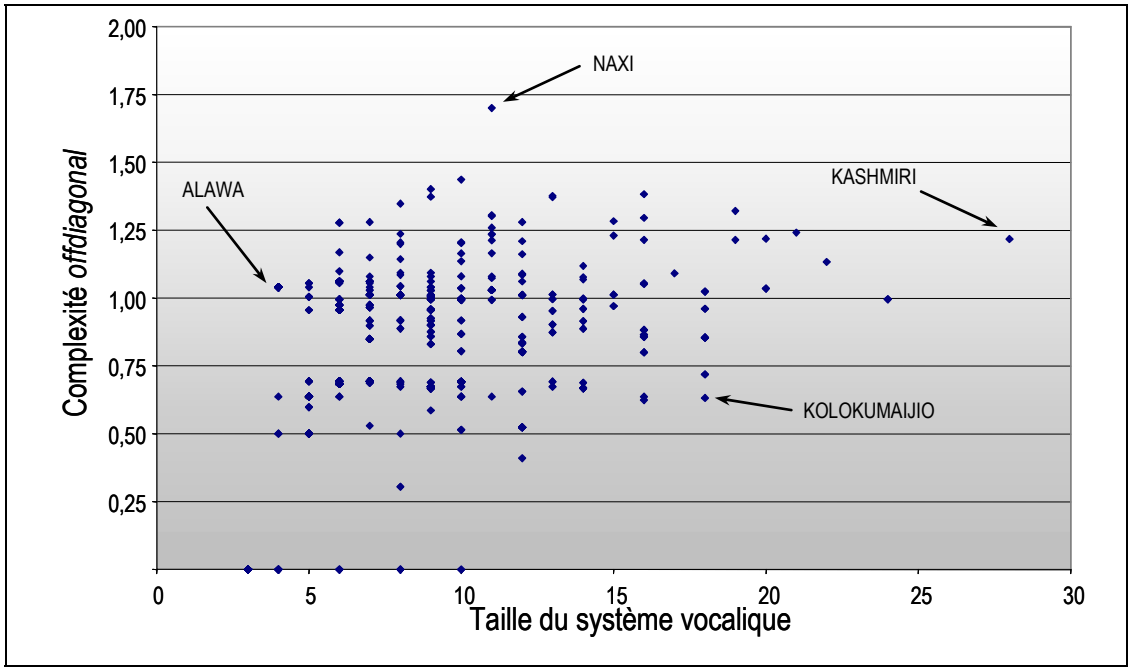


Figure 18 – Distribution de la complexité *offdiagonal* des systèmes vocaliques des langues d'UPSID en fonction de la taille de ces systèmes. Les langues mentionnées dans la discussion sont indiquées sur la figure.

Le même type d'analyse peut également être mené sur les systèmes consonantiques. Pour comparer les résultats avec d'autres approches de mesure de complexité récemment proposées, nous avons subdivisé les langues d'UPSID en 6 groupes conformes à ceux proposés dans Maddieson (2006) et le Tableau 17 résume les résultats obtenus. Lorsqu'on observe tout d'abord les résultats globaux (à l'échelle des 451 langues d'UPSID), on constate que la complexité consonantique moyenne est plus du double de celle observée pour les systèmes vocaliques qui semblent donc plus réguliers. En complément, une analyse de corrélation entre complexités vocalique et consonantique permet d'écarter l'hypothèse d'une quelconque corrélation (r^2 = 0,0006), en accord avec d'autres évaluations menées dans Maddieson (2006). Lorsque l'on s'intéresse aux résultats par zone, une MANOVA a mis en évidence des différences significatives entre les complexités moyennes des systèmes vocaliques et des systèmes consonantiques des différentes

zones, malgré une très importante variation à l'intérieur de chaque zone. Les analyses *post hoc* font ressortir un continuum de complexité plutôt que des groupes très distincts. Le groupe *Australia and New Guinea* se caractérise cependant par les plus faibles complexités vocalique et consonantique.

Tableau 17 – Valeurs moyennes et écarts-types des complexités structurelles des systèmes vocaliques et consonantiques des langues d'UPSID regroupées par grandes zones géographiques et génétiques, ordonnées par complexité consonantique moyenne croissante (d'après Coupé, Marsico et Pellegrino, à paraître).

ZONE	NOMBRE DE LANGUES	COMPLEXITÉ *OFFDIAGONAL* DU SYSTÈME VOCALIQUE VALEUR MOYENNE (ÉCART-TYPE)	COMPLEXITÉ *OFFDIAGONAL* DU SYSTÈME CONSONANTIQUE VALEUR MOYENNE (ÉCART-TYPE)
Australia and New Guinea	64	0,65 (0,39)	1,45 (0,34)
South and Central America	66	0,81 (0,29)	1,51 (0,24)
East and South-East Asia	108	0,87 (0,26)	1,61 (0,29)
North America	68	0,73 (0,37)	1,67 (0,28)
Europe, South and West Asia	71	0,90 (0,26)	1,83 (0,29)
Africa	74	0,79 (0,25)	1,84 (0,30)
UPSID	**451**	**0,79 (0,31)**	**1,67 (0,33)**

Les résultats présentés ci-dessus et basés sur la complexité structurelle des graphes, se révèlent finalement assez décevants. En effet, ils n'apportent guère d'information nouvelle, même si la méthodologie a l'avantage de proposer une quantification permettant d'ordonner les systèmes phonologiques par complexité structurelle croissante. Cette hiérarchisation pourra peut-être se révéler pertinente lorsqu'il s'agira de corréler entre elles plusieurs mesures de complexité comme celles proposées dans Maddieson (à paraître). L'auteur y propose d'évaluer la complexité de trois inventaires consonantiques en sommant les complexités des 22 segments les constituant, cette complexité segmentale correspondant elle-même au groupe d'appartenance du segment (I, II ou III), dans la terminologie proposée dans Lindblom & Maddieson, (1988). Un développement possible est donc d'utiliser une telle mesure sur l'ensemble des langues d'UPSID et de la mettre en perspective avec la complexité *offdiagonal*. D'ores et déjà, une très légère corrélation négative existe entre la redondance introduite précédemment et la complexité *offdiagonal* pour les 451 systèmes d'UPSID (r^2 = 0,10 ; p = 0,00). Le lien entre les deux variables est cependant ténu puisque 90 % de la variance observée pour l'une n'est pas expliquée par l'autre. On trouve ainsi des exemples de complexité (consonantique et vocalique) relativement forte avec une redondance minimale (cas de l'alladian, avec une complexité *offdiagonal* de 1,70 et une redondance de 1,00) ou, à l'inverse, une redondance relativement forte pour une complexité faible (cas du rotokas, avec une complexité *offdiagonal* de 0,32 et une redondance de 1,27).

✦ *Une ébauche de simulation stochastique d'évolution*

Dans Coupé, Marsico & Pellegrino (à paraître), nous avons également étudié une autre piste de recherche visant à établir une simulation de l'évolution des systèmes avec un modèle probabiliste, également à partir d'UPSID.

La procédure adoptée visait à caractériser chaque système par une mesure de cohérence, tenant compte des fréquences d'occurrence des segments dans UPSID et des fréquences de co-occurrences de paires de segments puis à implémenter une simulation stochastique d'évolution. Cette méthodologie nous a été inspirée en discutant avec Pablo Jensen du Laboratoire d'Économie des Transports (LET, UMR 5593) qui avait adopté une démarche apparentée pour l'étude des interactions de proximité entre des commerces en ville (Jensen, 2006). Une différence méthodologique importante avec nos études sur UPSID réside dans le fait qu'ici, la fréquence d'occurrence observée pour les segments dans UPSID est utilisée en entrée de la simulation, et non en variable émergente. Il est clair que ce raisonnement présente une certaine circularité et que le modèle relève donc du domaine de la simulation et non de l'explication.

Dans notre approche, la cohérence d'un système est définie comme résultant :

- de l'efficience intrinsèque *E* des segments le constituant, que ce soit pour des raisons articulatoires, perceptuelles, développementales, etc. ;
- des forces d'attraction et de répulsion au sein du système, approximées par l'interaction *I* entre les segments le constituant, pris deux à deux.

Faute de mieux, la fréquence d'apparition du segment dans les langues d'UPSID a été prise comme mesure de son efficience intrinsèque, après application d'une transformation non linéaire issue du test statistique de Fisher (extension du test du χ^2). Un segment apparaissant dans plus de 50 % des langues aura une efficience positive alors qu'elle sera négative dans le cas contraire. Un segment apparaissant dans 50 % des langues du monde est considéré comme « neutre » à cette étape.

Dans un second temps, des matrices de contingence dans UPSID sont établies pour toutes les paires de segments d'UPSID. Ainsi, pour la paire /a - ã/, 82 langues ont à la fois /a/ et /ã/, 58 n'ont aucun des deux segments, et 311 langues ont l'un ou l'autre, avec une forte asymétrie dans ce cas :

	/ã/ présent	*/ã/ absent*
/a/ présent	*82*	*310*
/a/ absent	*1*	*58*

À partir de ces matrices et par le biais de tests statistiques de Fisher, une force d'interaction est calculée pour chacune des paires. Dans l'exemple ci-dessus, la cooccurrence des événements « /ã/ présent » et « /a/ absent » est très rare par rapport à ce que l'on attendrait si les deux événements avaient été indépendants. On

quantifie ainsi les tendances universelles implicationnelles sous forme de forces d'attraction et de répulsion[54] : en l'occurrence /ã/ attire très fortement /a/.

Une méthodologie similaire a été suggérée par Clements (2003b) de manière plus restreinte, comme évaluation de l'hypothèse d'économie des traits. Notre approche est cependant plus exhaustive dans le sens où nous nous intéressons aux systèmes dans leur intégralité et non à une unique classe (les occlusives en l'occurrence dans Clements, 2003b).

La Figure 19 schématise les interactions significatives les plus fortes mises en évidence au sein des systèmes vocaliques. Les traits pleins correspondent aux interactions « attractrices », c'est-à-dire les paires de segments qui ont tendance à être présents ensemble plus que ce que leurs fréquences individuelles laissent supposer. À l'inverse, les traits pointillés correspondent aux forces « répulsives » pour lesquelles les deux segments considérés ont tendance à ne pas être présents simultanément dans une langue. La première observation porte sur ce que l'on peut appeler une tendance à régulariser le système : dans le cas des nasales, il s'agit d'une tendance à présenter simultanément les trois voyelles /ĩ ã ũ/ ; dans le cas des voyelles orales, cela se traduit par une tendance à la symétrie, en particulier pour les apertures *high* et *lowered-high*.

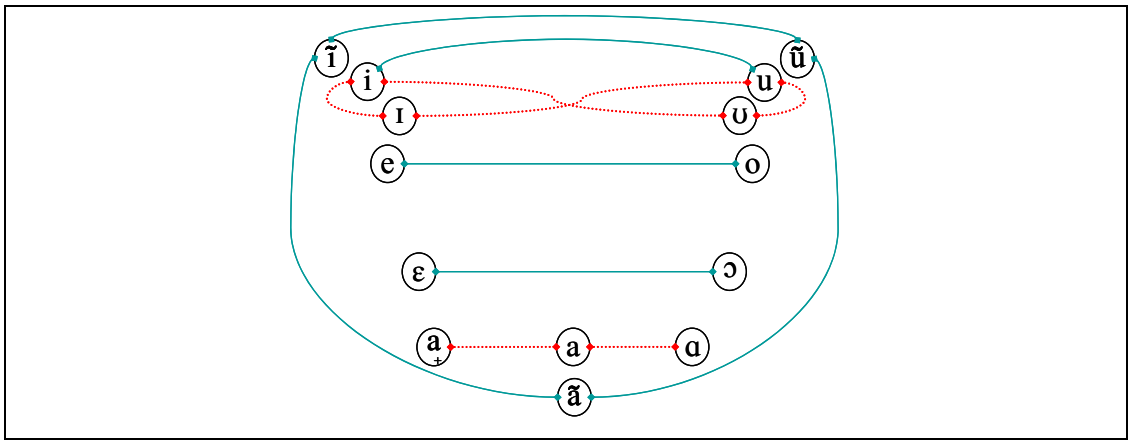


Figure 19 – Schéma des interactions significatives les plus fortes mises en évidence par le test de Fisher. Les traits pleins correspondent aux forces d'attraction et les traits pointillés aux forces de répulsion.

La mesure de cohérence *C* définie au niveau de chaque système est la somme des efficiences intrinsèques *E* associées à chaque segment et de toutes les interactions *I* entre les segments pris deux à deux. Une des originalités de l'approche réside dans la prise en compte, non seulement de la présence de certains segments

[54] En réalité, la procédure est plus complexe et toutes les interactions potentiellement significatives ne sont pas détectées à cause de la distribution des segments eux-mêmes. Par exemple, /i/, /a/ et /u/ sont si fréquents que le test de Fisher ne permet pas de déterminer s'ils interagissent de manière significative entre eux. De même, les interactions mettant en jeu un segment extrêmement rare peuvent ne pas être détectées (pour les détails, se reporter à Coupé, Marsico & Pellegrino, à paraître en 2009).

mais également de l'absence d'autres segments dans le système. À partir de la Figure 19, on peut intuitivement déduire que le système /i a u/ est « meilleur » que le système /i a u ɪ/ puisque/ɪ/ interagit négativement avec /i/ et avec /u/. Pour tenir compte des interactions positives et négatives, nous avons donc décrit le système /a i u/ comme /a i u !ɪ/ où « ! » indique l'absence du segment considéré. Le calcul de l'énergie prend alors en compte les valeurs intrinsèques des segments présents mais également des « anti-segments » absents et de même en matière d'interaction :

$$C(/\text{a i u}/) = E(\text{a}) + E(\text{i}) + E(\text{u}) + \ldots$$
$$I(\text{a} - \text{i}) + I(\text{a} - \text{u}) + I(\text{i} - \text{u}) \ldots$$
$$I(\text{a} - !\text{ɪ}) + \ldots$$

En fait, tous les « anti-segments » absents et toutes les interactions sont prises en compte, et la cohérence de chacun des systèmes est la somme de 833 efficiences intrinsèques (correspondant aux 833 segments ou anti-segments rencontrés dans UPSID) et de 833*832/2 = 346 528 interactions, quelle que soit la taille du système.

Nous avons mis en œuvre cet algorithme sur les systèmes vocaliques. La Figure 20 représente en ordonnée les énergies minimales, maximales et moyennes calculées pour les différentes tailles de systèmes vocaliques d'UPSID (en abscisse). Lorsque les trois courbes sont confondues en un unique point, cela correspond à une taille de système ayant un unique représentant dans la base de données : 17, 21, 22, 24 et 28 voyelles.

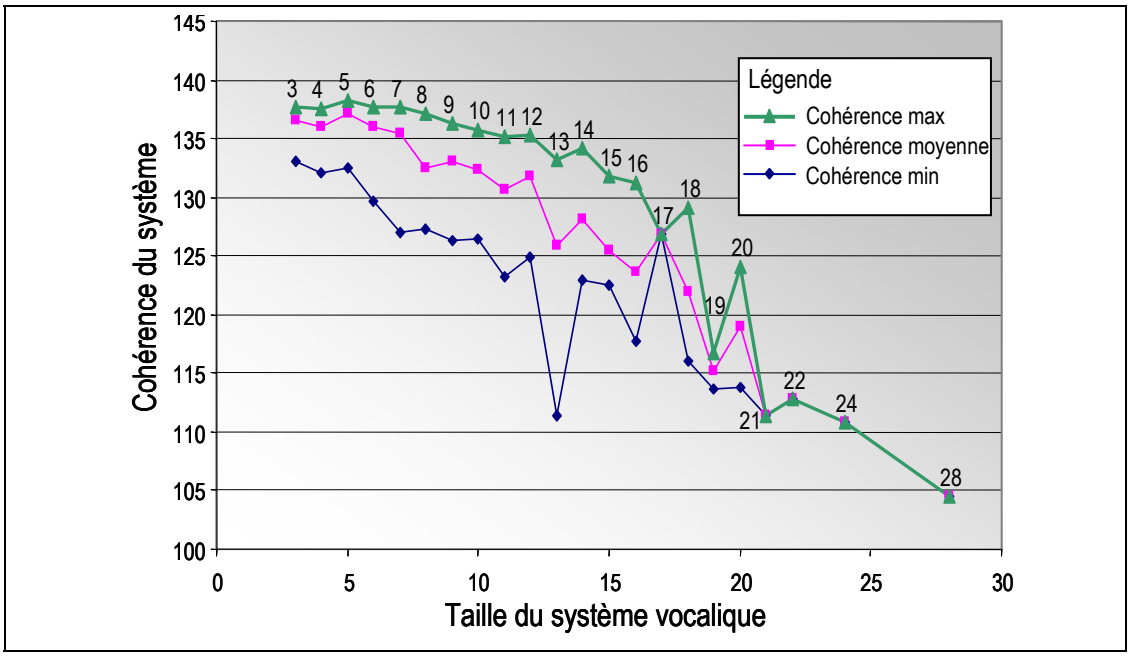


Figure 20 – Distribution des cohérences des systèmes vocaliques des 451 langues en fonction de la taille du système. Les trois courbes verte, magenta et bleue correspondent respectivement aux valeurs maximale, moyenne et minimale calculées pour une taille de système donnée (rappelée pour la courbe verte). N.B. La cohérence est une mesure sans unité.

On observe que la cohérence moyenne a tendance à décroître lorsque la taille du système augmente. La courbe verte, correspondant aux systèmes de cohérences

maximales pour une taille donnée, atteint cependant des valeurs relativement élevées jusqu'à environ 15 segments. La dégradation observée pour des systèmes plus grands peut être artefactuelle et liée à la taille réduite de l'échantillon de langues concernées puisqu'à peine 26 langues, soit moins de 6 % d'UPSID, présentent plus de 15 segments vocaliques. Cette dégradation est néanmoins compatible avec le fait que plus un système est grand, plus il utilise des segments rarement observés et dont l'efficience intrinsèque est donc faible.

On remarque par ailleurs qu'un système à 13 voyelles présente une cohérence particulièrement basse ; il s'agit du dinka qui peut être qualifié de système atypique puisqu'il est décrit avec 7 voyelles de type *breathy-voiced* et 6 voyelles de type *creaky-voiced*. Ainsi, aucune voyelle présentant un mode de phonation modal n'est présente dans le système, réduisant considérablement son efficience intrinsèque et ce même si la force interactionnelle du système n'est pas particulièrement faible[55].

Par la suite, nous avons utilisé cette mesure de cohérence comme critère de *fitness* dans une simulation d'évolution. L'hypothèse initiale est que la mesure définie ci-dessus permet de capturer les trois contraintes non sociolinguistiques proposées par B. Lindblom (cf. citation de la page 99) et qu'un changement la renforçant est plus probable qu'un changement la dégradant. L'objectif de la simulation est d'évaluer la stabilité des systèmes, c'est-à-dire leur propension à rester inchangés au cours du temps.

Ces quelques lignes expliquent le mode de fonctionnement général de notre simulation :

Étape 1 : les systèmes vocaliques d'UPSID sont pris comme systèmes de départ. Pour chaque système, 100 tirages aléatoires sont effectués avec à chaque fois 2 % de chance qu'une modification ait lieu, par changement d'un faible nombre de segments. Ces systèmes sont supposés représenter des changements possibles en un pas d'évolution.

Étape 2 : les cohérences globales des nouveaux systèmes sont comparées à celle du système initial, et cette distribution est normalisée de façon à obtenir un jeu de probabilités (entre 0 et 1) pour chaque système dérivé.

Étape 3 : Un système parmi les 100 est choisi de façon aléatoire en suivant la distribution de probabilités établie à l'étape 2. Les systèmes conduisant à un renforcement de cohérence sont donc les plus probables, mais un système amenant une dégradation de la cohérence peut également être tiré au sort.

Cette procédure (étapes 1 à 3) est réitérée 500 fois, donnant pour chaque système d'UPSID 500 hypothèses d'évolution. La stabilité est alors estimée en mesurant le pourcentage de ces évolutions potentielles qui au final maintiennent le système dans son état, sans évoluer. Plus la cohérence d'un système est élevée vis-à-vis du voisinage exploré dans le processus d'évolution (par changement aléatoire

[55] Le caractère atypique du système du dinka s'exprime également en matière de quantité vocalique puisqu'il s'agirait d'un des rares systèmes à trois degrés d'opposition de durée V, VV et VVV (Remijsen & Gilley, 2008).

d'un petit nombre de segments), plus sa stabilité est élevée. À la limite, un système pour lequel aucune des 500 hypothèses d'évolution ne générerait de changement serait stable à 100 %.

La Figure 21 représente la stabilité des systèmes avec les mêmes conventions que sur la figure précédente. La première constatation est qu'avec la simulation très schématique mise en place, certains systèmes sont extrêmement stables ; par exemple, le système /i a e o u/ est stable à 99,5 %. De plus, certains systèmes vocaliques de taille élevée sont très stables (exemple de l'ewe, système à 14 voyelles atteignant une stabilité de 93,3 %) ou relativement stables (comme le kolokumaijo, système à 18 voyelles dépassant 60 % de stabilité).

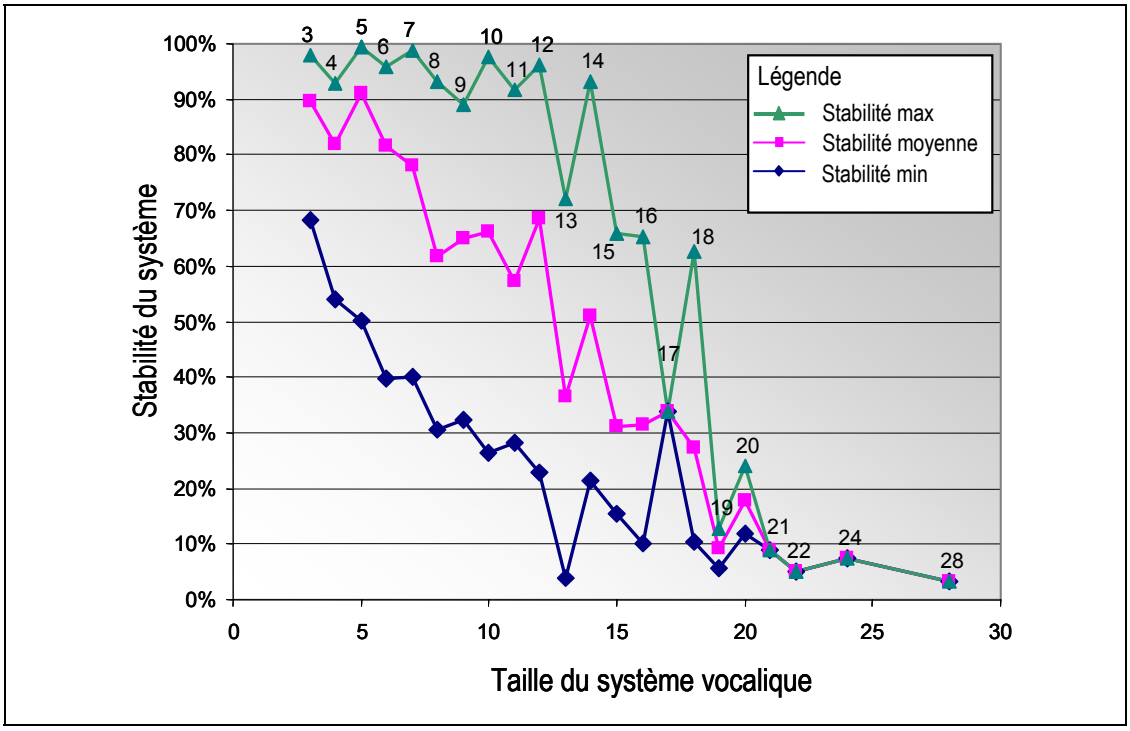


Figure 21 – Distribution des stabilités des systèmes vocaliques des 451 langues en fonction de la taille du système. Les trois courbes verte, magenta et bleue correspondent respectivement aux valeurs maximale, moyenne et minimale calculées pour une taille de système donnée (rappelée pour la courbe verte).

Il est également intéressant de noter qu'un changement de mode se produit au niveau des systèmes à 9 voyelles ; si l'on considère les systèmes les plus stables (courbe verte), pour les systèmes de taille inférieure, les systèmes de taille impaire (3, 5 et 7) sont plus stables que ceux de taille paire (4, 6 et 8). À l'inverse, pour les systèmes plus grands, les tailles de systèmes paires se révèlent plus stables (12, 14, 16 ,18 et 20). On retrouve donc là le changement d'organisation rapporté dans la littérature (par exemple Vallée, 1994:96), avec des systèmes composés d'un nombre impair de voyelles primaires (en deçà de 9 segments) et de deux séries l'une primaire et l'autre secondaire, contrastées par un trait secondaire (au-delà de 9 segments). Le kolokumaijo illustre bien cette tendance puisqu'il se compose de 9 voyelles orales et de leurs 9 contreparties nasales.

Il est plus difficile d'émettre des observations relatives à la stabilité moyenne des systèmes ; en effet, le fait que le nombre de systèmes diffère pour chaque taille implique des variations importantes de l'intervalle de confiance de la moyenne et il ne serait pas prudent d'analyser en détail la forme de la courbe obtenue. Tout au plus peut on remarquer que la variation de stabilité pour une taille de système donnée peut être très importante ; par exemple, pour une taille de 10 voyelles, la stabilité varie de 97,5 % (le saliba, composé de /i a e o u/ et de leurs contreparties nasales) à 26,5 % pour le hamer, dont le système vocalique est /i e ɑ o u ɪˤ eˤ æˤ ɔˤ ʊˤ/. Il s'agit d'un système présentant deux séries vocaliques, l'une primaire, l'autre pharyngalisée, mais également des différences de timbre entre ces séries, ce qui pénalise sa stabilité par rapport à un système où les deux séries seraient en miroir. On retrouve enfin le dinka avec une très faible stabilité (4 %) qui semble confirmer la situation transitoire de ce système.

✦ Discussion

Dans cette section, nous avons présenté plusieurs approches basées sur l'exploitation d'UPSID et visant à contribuer à notre connaissance des principes d'organisation des systèmes phonologiques en termes d'utilisation maximale des traits disponibles et d'économie des traits mais également pour évaluer si des critères structurels intervenaient dans leur organisation. Nous avons tout d'abord introduit une formalisation de la basicité des segments, élaborée à partir de leur description en traits phonétiques. Ce formalisme est logiquement lié à l'ensemble des traits employés dans UPSID et il correspond seulement partiellement à une hiérarchie de complexité. Par exemple, le fait que les traits *breathy-voiced* et *creaky-voiced* alternent avec *voiced* et *voiceless* dans la description des voyelles implique qu'ils sont considérés tous quatre comme basiques alors que leurs complexités intrinsèques ne semblent pas équivalentes. Ce constat permet de mettre l'accent sur le fait que les traits basiques que nous utilisons définissent les caractéristiques indispensables à l'existence des segments, dans le sens où un lieu et un mode d'articulation ainsi qu'un mode de phonation sont nécessaires à la définition des segments consonantiques. En revanche, la complexité intrinsèque liée à l'une ou l'autre des alternatives (e.g. lieu bilabial vs. vélaire ou uvulaire) n'est pas prise en compte, contrairement à l'approche proposée dans Lindblom & Maddieson (1988).

Lorsque l'on se reporte au niveau des inventaires phonologiques, on observe que 71 % des langues ont recours à des segments non basiques, même lorsque la taille du système est faible. Si l'on se limite à l'examen des tailles pour lesquelles des systèmes basiques sont attestés (soit jusqu'à 38 segments), là encore la majorité des systèmes (65 %) sont non basiques, à hauteur de 20 % de leurs segments en moyenne. Ces éléments semblent indiquer que si un principe d'économie est à l'œuvre, il n'est pas impératif au point d'empêcher le recrutement de traits non basiques par les langues. Cette conclusion doit cependant être nuancée, en particulier à la lumière de Ohala (à paraître) qui attire l'attention sur le fait que des traits peuvent être disponibles (au sens du MUAF), sans pour autant être

phonologisés et donc présents dans les inventaires des langues. Pour évaluer cette hypothèse, nous nous sommes concentrés sur les langues dont l'inventaire fait état d'un unique trait non basique, soit 147 langues. La répartition des traits employés a été dressée et nous les avons rassemblés en trois groupes (voir Tableau 18) en nous intéressant aux processus ayant pu les faire émerger diachroniquement. Ces groupes sont *Coarticulation*, *Articulatory* et *Timing* et ils présentent malheureusement une part d'arbitraire puisque certains traits relèvent potentiellement de plusieurs catégories. *In fine*, ils relèvent tous d'une synchronisation de gestes articulatoires, à l'exception peut-être du trait *ejective*. La Figure 22 illustre le résultat de ce regroupement.

Tableau 18 – Distribution des traits non basiques utilisés dans les 147 langues ayant un unique trait non basique dans leur inventaire. (Source : UPSID, Maddieson & Marsico, non publié).

TRAIT	GROUPE	NOMBRE
nasalized	Coarticulation	37
aspirated	Timing	27
prenasalized	Coarticulation	17
ejective	Articulatory	16
labialized	Coarticulation	15
long (vowel)	Timing	14
palatalized	Coarticulation	5
long (consonant)	Timing	3
overshort	Timing	3
velarized	Coarticulation	3
pharyngealized	Coarticulation	2
retracted	Articulatory	1
advanced	Articulatory	1
nasal release	Articulatory	1
lateral release	Articulatory	1
with breathy release	Articulatory	1
TOTAL		**147**

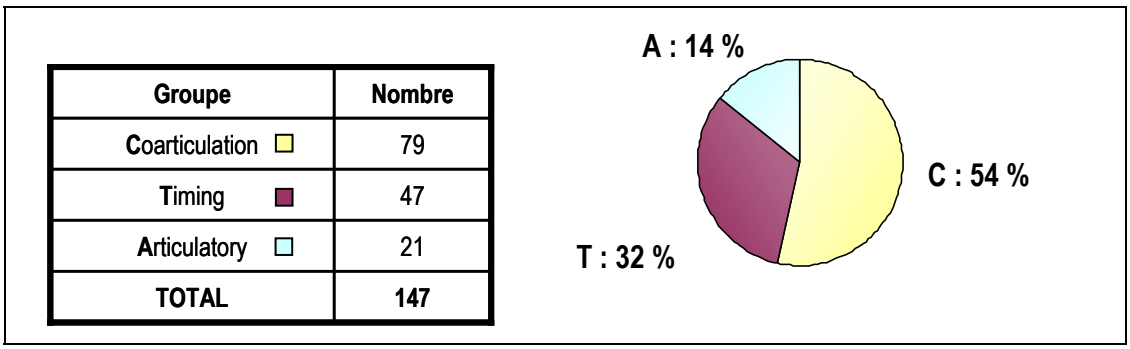


Figure 22 – Répartition des groupes de traits non basiques au sein des 147 langues ayant un unique trait non basique dans leur inventaire. (Source : UPSID, Maddieson & Marsico, non publié).

On constate que la grande majorité des traits employés par ces langues ne reposent pas sur un geste articulatoire supplémentaire mais sur des modifications

temporelles (assimilation ou utilisation d'un contraste de durée). Il est donc tout à fait possible que ces traits aient préexistés au sein du système phonétique avant d'être phonologisés, ce qui est compatible avec la reformulation du MUAF proposée dans Ohala (à paraître).

Le MUAF, dans sa version initiale, faisait principalement référence à l'utilisation faite par les langues des traits disponibles et non à la nature de ces traits. Si l'on regarde plus précisément l'usage que les 147 langues évoquées ci-dessus font du trait non basique qu'elles ont dans leur inventaire, on s'aperçoit que ce trait entre dans la description de 4,20 segments en moyenne et le mode de la distribution est de 4 (cf. Figure 23, histogramme de gauche). 17 langues (soit 11,6 % de l'échantillon) ont un unique segment non basique.

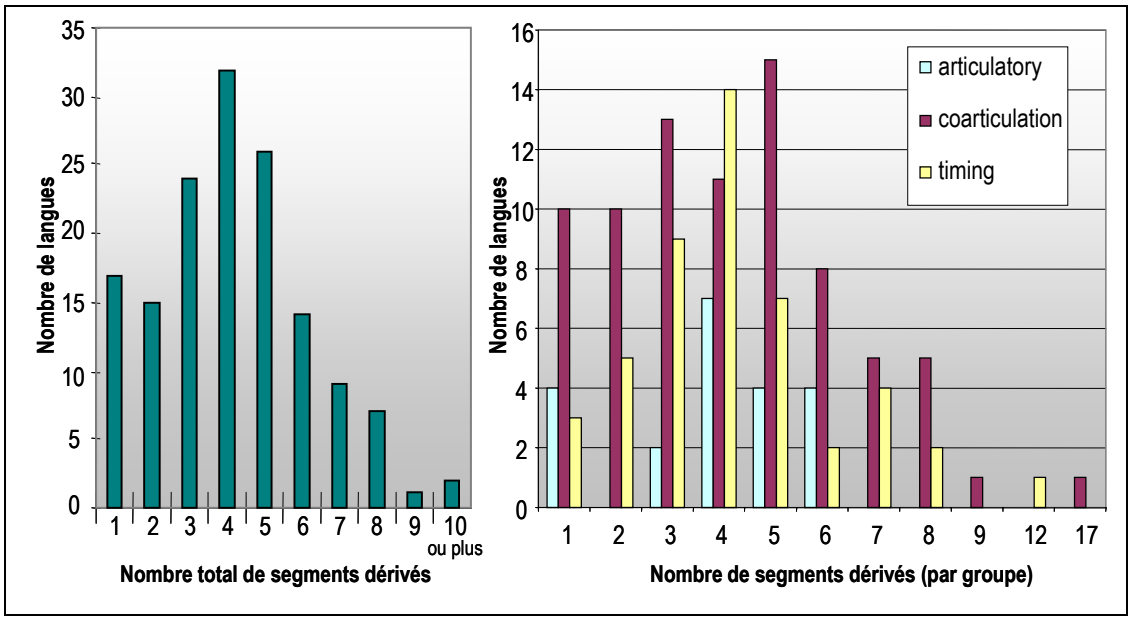


Figure 23 – Distribution du nombre de segments non basiques dans les 147 langues ayant un unique trait non basique : à gauche, distribution globale, à droite distribution par groupe. (Source : UPSID, Maddieson & Marsico, non publié).

Deux langues ont plus de 10 segments dérivés ; il s'agit du wolof (12 voyelles longues) et du saami (17 consonnes palatisées). À partir de cette distribution, il n'est cependant pas possible de déduire directement si les langues font un usage maximal du trait non basique puisque le nombre maximal théorique de segments « dérivables » dépend au moins du trait lui-même et des segments basiques présents dans le système susceptibles de porter ce trait. Néanmoins, la Figure 23 (histogramme de droite) donne une première idée en représentant le même type de distribution pour chacun des trois groupes de traits identifiés précédemment. Les valeurs moyennes ne montrent pas de différences significatives (Test de Kruskal-Wallis, $\chi^2(2,144) = 0,11$; $p = 0,94$). On constate cependant que pour le groupe *articulatory*, aucune langue ne génère plus de 6 segments par intégration du trait non basique. Pour le trait *ejective* par exemple, une étude cas par cas des 16 langues concernées montre une assez grande diversité de situation puisque une proportion variable (mais le plus souvent forte) des occlusives et affriquées non voisées donne lieu à une série éjective, alors que les fricatives éjectives sont seulement présentes

dans deux langues parmi les 16 (koma et socotri). L'utilisation du trait *ejective* n'est donc pas réellement maximale, mais elle est forte au sein de la série (occlusive ou affriquée) où ce trait a fait son apparition.

Nous nous sommes par la suite intéressés à la structure des systèmes phonologiques en introduisant les notions de redondance et de complexité structurelle *offdiagonal*. La redondance fait référence à la distance basée sur les traits intervenant dans les oppositions phonologiques de la langue. Elle s'avère très faible avec une distance moyenne entre deux segments voisins au sein des systèmes de l'ordre de 1,06 trait et alors que des redondances fortes (supérieures à 1,3) sont quasiment absentes puisque seulement rencontrées pour deux langues ayant un nombre de traits relativement important pour un nombre de segments faible.

La complexité *offdiagonal* est intrinsèquement une mesure structurelle et sa pertinence n'est pas clairement établie. L'objectif était de prendre en compte des motifs saillants dans les réseaux phonologiques (par exemple, sous-systèmes vocaliques oral et nasal en miroir, etc.) mais, faute d'une réelle possibilité d'interprétation des valeurs obtenues, le résultat est décevant. De plus, même si des variations aréales sont suggérées par les valeurs moyennes de complexité (cf. Tableau 17), l'amplitude de variation démontrée par les écarts-types rend toute analyse hasardeuse. Il est toutefois possible d'envisager des améliorations dans deux directions : la première consiste à intégrer dans le calcul une complexité intrinsèque des segments, de manière à être plus proche de la réalité des systèmes. La seconde visera à prendre en compte des graphes valués en affectant des valeurs aux arêtes des réseaux. Ces valeurs pourront être par exemple les distances en traits entre les phonèmes constituant les sommets ou encore les forces d'interaction déterminées dans le cadre de l'élaboration du modèle d'évolution stochastique esquissé par la suite.

Ce modèle, basé sur les fréquences d'occurrences des segments et de co-occurrences de paire de segments dans les systèmes vocaliques d'UPSID a permis de mettre en évidence plusieurs éléments intéressants. Tout d'abord, alors même qu'il prend uniquement en compte des interactions locales, il fait apparaître des propriétés globales des systèmes réels comme la préférence pour des systèmes de taille impaire en dessous de 9 segments, paire au-delà (cf. Figure 21). Par ailleurs, les mesures de stabilité permettent d'illustrer un élément important : même en faisant abstraction des facteurs sociolinguistiques, l'évolution des systèmes phonologiques n'est pas régie uniquement par des principes généraux mais elle dépend également du voisinage du système de départ. Le modèle suggère que lorsque un changement potentiel dégrade la cohérence du système, il a relativement moins de chance d'être instancié (cf. la notion de *systemic compatibility* (Lindblom, 1998:245)). Il s'agit là évidemment d'une première esquisse de modèle d'évolution, mais elle nous semble pouvoir être développée plus avant. En particulier, le mécanisme de détermination du voisinage des systèmes demande à être affiné, et il sera intéressant de regarder pour certains systèmes les évolutions proposées et leurs probabilités associées. La prise en compte d'autres sources de données, comme

celles intégrées dans la base UNIDIA[56], développée à DDL, est une source importante de progrès, soit pour affiner le modèle, soit pour le valider *a posteriori*. Enfin, la question de la nature des primitives phonologiques reste entière. Il est tout à fait envisageable en effet qu'un modèle basé sur des interactions définies au niveau des traits soit plus efficace. Alternativement, la description des systèmes en gestes articulatoires peut se révéler pertinente et parcimonieuse (Proctor, 2007). Enfin, certaines expériences menées par René Carré (Carré, Pellegrino & Divenyi, 2007 ; voir aussi Al Tamimi, 2007 et Carré, à paraître) mettent l'accent sur la description des segments vocaliques dans un espace dynamique dérivé de l'espace formantique par rapport au temps. Ces approches, aussi diverses soient-elles, mettent toutes en avant le fait que l'on ne pourra probablement pas développer notre compréhension des systèmes phonologiques des langues du monde si l'on n'intègre pas plus avant la dimension temporelle dans les analyses développées. Selon nous, la notion d'information telle qu'introduite par Shannon peut jouer le rôle de médiateur entre les aspects systémiques et la dimension temporelle de la communication parlée, comme nous allons maintenant l'évoquer pour conclure ce manuscrit.

Tableau 19 – Liste des langues d'UPSID citées dans la section 3.3.1. Pour chaque langue sont indiqués : sa classification linguistique, son inventaire phonologique (voyelles, consonnes et éventuellement diphtongues) ainsi que la ou les page(s) où elle est citée (Source UPSID, Maddieson et Marsico, non publié).

NOM DE LA LANGUE	CLASSIFICATION LINGUISTIQUE	INVENTAIRE PHONOLOGIQUE	PAGES
ALAWA	Australian, Maran	ɪ a e ɯ̟ m t ŋ r j n̪ nd̪ nd ɳɖ mb l̪ n ɳ l ɹ p t̪ k l̪ ŋg ʈ w	109
ALLADIAN	Niger-Kordofanian, Kwa	o ɪ i ɛ ɔ ɛ̃ ã e ɔ̃ u a ɟ s k p b d g c g͡b v k͡p ʒ f m n j z ɲ ɽ l t w ʃ h ɥ	111
ARCHI	Caucasian	oˤ iˤ a e o u eˤ aˤ i uˤ ʔ t͡s z ṕ pː tː kː pʰ tʰ kʰ g q͡χ ḱ t͡sʰ t͡ʃː t͡sː t͡ʃ t́ χˤ t͡ʃʰ t͡ɬʰ χˤː b ħ χ q͡χˤː ʁˤ q͡χː t͡ɬ q͡χ sː h χː ʁ ʕ s ʃ q͡χˤ d q͡χˤ t͡ʃʷ zʷ sʷ ɬː q͡χʷˤ m rr j ɣ kʷː n ɬ q͡χʷ t͡sʷ kʷ gʷ dʷ kʷʰ ʃː w χʷˤ q͡χʷˤ t͡ɬʷʰ l χʷ tʷʰ χʷː χʷˤː ʁʷˤ q͡χʷ sʷː ʃʷː t͡ɬʷ ʁʷ ʃʷ t͡ʃʷʰ t͡sʷʰ ɬʷː ɬʷ ʒʷ ʒ	103, 106
AVAR	Caucasian	o i u a e q͡χː ḱ q͡χː t́ pʼ ʡ g tʰ t͡ʃʰ ʒ b t͡ʃː kʰ k͡xː t͡sʰ k͡xː t͡sː t͡ʃ t͡s t͡sː pʰ t͡ʃː x ːχː ɬː t͡ɬː sː ħ f s ʃ χ ʃː t͡ɬː w z l h d j rr n m ɬ	103, 106
BASHKIR	Ural-Altaic, Turkic	æ i ø ɑ ɤ ɛ ɔ o ə y ʊ ʒ t͡ʃ v t ʔ f b h s z ʃ θ̪ t͡s x ʁ d k n ŋ l r g p m q j w	100

[56] UNIDIA est une base de données en cours de constitution à DDL sous la responsabilité de Mahé Ben Hamed. Elle référence les changements phonétiques attestés dans les langues du monde. À ce jour, environ 3700 changements observés dans environ 190 langues sont codifiés dans la base de données (Ben Hamed, 2007 ; UNIDIA, 2008).

NOM DE LA LANGUE	CLASSIFICATION LINGUISTIQUE	INVENTAIRE PHONOLOGIQUE	PAGES
DINKA	Nilo-Saharan, East-Sudanic, Nilotic, Dinka-Nuer	a e ɛ u i o a ɔ o i ɔ ɛ e p g t c k b d ɟ m w n n̪ d̪ j r ŋ ɲ t̪ l	115, 117
EWE	Niger-Kordofanian, Kwa	ɛ ɪ̃ i u e ẽ ã ɔ ɔ̃ õ ũ o a ɛ̃ g f d̪ k^{h} k͡p g͡b t͡s̪ d͡z̪ z β p^{h} ɸ s t^{h} t̪ b ɦ ʕ m n ŋ j v ɰ l ɲ	116
GADSUP	Papuan, Trans-New-Guinea	ɜ i u eː oː a̟ː j d m p t β ʔ k n	107
HAMER	Afro-Asiatic, Omotic	ɔˤ æˤ u o ɪˤ ʊˤ e eˤ i ɑ ɓ p d z ɗ c ɟ g t͡s b s t ɠ m ʃ h ŋ n ɲ w ɽ l j k ɠ	117
KABARDIAN	Caucasian	oː ɨ ɜ iː uː eː a̟ː kˠ gˠ kʷʰ kʲ gʲ tʲ tʰ qʲ b f ʃ pʲ v d̪ qʷ s̪ z̪ q͡χʷ q͡χ t͡s̪ d͡z̪ kʲʰ ɣ̪ʲ f̪ ð̪ ʁʷ ʁ χʷ χ ɣʲ r xʲ ʔ ɦ ʃ̪ ɦ θ̪ ʒ̪ xʷ pʰ ʕ ɬʲ w j ɬʲ t͡s̪ n̪ m ʔʷ	103, 106
KASHMIRI	Indo-European, Indic	ɘ e ɜ̃ u ɪ̃ ẽ ã ũ aː ɜ ɒ ɨ i õ ɔː ɔ̃ ɔ̃ː uː ɨː ɘː ĩˠ iː ãː ũː ɨ̃ː ɛ̃ː o ɛː kʰ pʰ p t̪ tʰ t̪ d̪ ʈʰ t͡s ʃʲ k g t͡sʰ ɖ d͡ʒʲ n b w t̪͡ʃʲʰ m l ɦ s d͡z j r ʈ͡ʃʲ	109
KOLOKUMAIJIO	Niger-Kordofanian, Kwa	õ ɪ̃ ɛ̃ o ʊ̃ ʊ ɔ i a ɔ̃ u ɛ ɪ̃ ẽ ũ ã e ɪ d v f p t k b g l w j ɽ ŋ n m z s g͡bʷ k͡pʷ	110, 116
KOMA	Nilo-Saharan, Koman	ə a i e ɔ u ɛ o b ɓ m w t' d ɗ s p l t r p' ʃ j k k' ŋ n h s' ʔ g	120
MAXACALI	South-American, Macro-Ge	ã e i a o ẽ ɯ̃ õ ɪ̃ ɯ mb nd p k t̪͡ʃ n̪d̪͡ʒ h ʔ ŋg t	100
NAXI	Sino-Tibetan, Lolo-Burmese	ɛ ɑ ɚ y æ o ɯ̟ ə ɯ u i t͡ʂʰ t ɟ͡ʑ kʰ b d g mb pʰ k nd c͡çʰ t͡sʰ c͡ç t͡ʂ ŋg p ɲ t͡s ɲɟ͡ʑ ɣ v ʂ f tʰ n̪d̪͡z̪ s ɳɖ͡ʐ̩ n m χ ɕ d͡z̪ ɖ ʑ ʐ̩ z̪ l	110
PARAUK	Austro-Asiatic, Palaungic	ɤ ɯ i e ɛ a a o u e ɯ i ɤ o ɔ ɛ ɔ u d k f tʰ kʰ b ʃ g d̪ g p t̪͡ʃʰ d̪͡ʒ d̪͡ʒ t t̪͡ʃ n pʰ b m ʒ m l n̪ ŋ n n̪ ŋ l v l v l̪ ʒ̪ ɯi ui ua ai aɯ au iu ia ɔi ɔi ɯi iɛ au iɛ ua oi iu io ai ɤi ui ɤi oi aɯ ia	100
PIRAHA	South-American, Paezan	a o i h s k p b ʔ g t	107
ROTOKAS	Papuan, East-Papuan	u e a o i t p g β ɾ k	105, 111
SALIBA	South-American, Equatorial	e o i ũ ẽ ã õ a u ɪ̃ kʷ k t ɸ ʔ b d gʷ g d͡ʒ p t̪͡ʃ r l x ɽ n m s β h n̪	117
SOCOTRI	Afro-Asiatic, Semitic	u ə i a o e ʃ' b t' k g ʔ f s s' ʃ k' t ʒ ɦ ɬ d ʕ rr h ɬ m w j l rr z n l	120

3.3.2. Complexité & information[57]

> "In a small [consonantal] system, the average information carried by each consonant is smaller than in a large system. Hence, smaller paradigms entail less competition among units and make them, in relative terms, more predictable. Greater predictability implies reduced demands on perceptual distinctiveness. Such demands could, in principle, be met by random selection of phonetic values from a universal set, but, instead, language prefers to exhaust its choices among the 'less complex' segments before it recruits 'more complex' possibilities for additional distinctions" (Lindblom, 1998:247-249).

Si l'on met en perspective cette citation avec la référence faite à la conversation avec J.J. Ohala qui ouvre ce chapitre (p. 79), on a matière à s'interroger sur le lien entre, d'une part, la complexité du système phonologique et d'autre part, l'organisation temporelle de l'information véhiculée par les unités phonologiques dans la parole. Comme cela a été largement évoqué dans la section 3.2, le lien entre information et complexité a suscité des réflexions en phonologie depuis plus de cinquante ans, même si l'intérêt qui lui a été consacré a été fluctuant. La relation entre prédictibilité et effort d'articulation est en particulier largement étudiée (Hay, Pierrehumbert & Beckman, 2001, Bybee, 2006 ; Zhao & Jurafsky, 2007 parmi d'autres) faisant ainsi écho aux prédictions de la *H&H theory*. Pourtant, la manière dont l'information phonologique se répartit tout au long du signal acoustique lors de la communication parlée reste encore assez mystérieuse, même si l'on a parfois cherché à quantifier cette information (e.g. van Son & Pols, 2003).

Nous partons du constat que la complexité phonologique des langues du monde, quelle que soit la méthode envisagée à l'heure actuelle, met en évidence une très large variabilité (e.g. Maddieson, 2006 ; Shosted, 2006 ainsi que nos propres travaux présentés dans ce chapitre). Pourtant, si l'on considère que complexité et information sont liées, cela semble en contradiction avec le fait que les langues du monde sont fonctionnellement équivalentes : on peut bien sûr envisager que les besoins en matière de lexique soient différents d'une langue à l'autre, mais il est beaucoup plus difficile d'argumenter que les capacités élémentaires de la communication humaine diffèrent d'un groupe de locuteurs d'une langue à un autre. Une première question est alors d'identifier le domaine dans lequel cette équivalence se réalise. S'il s'agit de la complexité, cela implique que la perception que l'on en a aujourd'hui est fausse ou en tout cas insuffisante pour l'appréhender dans sa globalité, ce qui est vraisemblable. L'hypothèse d'une compensation de complexité entre les composantes d'une langue n'a donc pas été invalidée. Cependant, la diversité de complexité observée le long des multiples dimensions linguistiques (taille des inventaires phonologiques, nombre de segments composant

[57] Ces travaux sont encore au stade de l'élaboration et sont non publiés. Des versions antérieures ont été présentées oralement (Pellegrino, Coupé & Marsico, 2007 ; Pellegrino ; 2007) et un manuscrit est en cours de préparation (Pellegrino, Marsico & Coupé, en préparation) fourni en annexe C.11.

les syllabes, nombre de cas, etc.) est vaste. On peut donc considérer que les contraintes de complexité offrent de nombreux degrés de liberté et que l'équivalence entre langues sera difficile à prouver à ce niveau grammatical, tant l'étendue des systèmes possibles est vaste.

Nous faisons plutôt l'hypothèse que c'est au niveau du débit d'information que l'équivalence se situe, soit sous forme d'un débit d'information préférentiel quelle que soit la langue, soit sous la forme d'une double contrainte de débit d'information minimal pour que la langue soit fonctionnelle et maximal pour que le système cognitif puisse traiter le flux parlé en temps réel.

Si cela se vérifie, la taille d'un paradigme[58] grammatical n'est qu'une partie de l'équation, et il est nécessaire de prendre en compte également la dimension temporelle. Cette dimension temporelle est composée d'une part de l'agencement séquentiel des unités ou des primitives phonologiques et d'autre part de la dimension physique de la chaîne parlée (durée des unités et débit de parole). Toutes choses étant égales par ailleurs, une phrase articulée deux fois, d'abord à un débit articulatoire donné, puis à un débit 30 % plus rapide, si elle demeure intelligible, véhicule la même quantité d'information sémantique, mais à deux débits d'information différents. Cette formulation soulève néanmoins de nombreuses questions puisque les variations de débit n'ont pas un effet uniforme (e.g. Dellwo & Wagner, 2003) et qu'en particulier les questions de prédictibilité – liées à l'information véhiculée – interviennent. De plus, il est probablement nécessaire de distinguer les variations de débit d'information intrinsèques à la langue de celles liées au style du locuteur : « le style est largement une histoire de densité d'information » (Martinet, 1969:187).

Notre objectif est double ; il s'agit premièrement de déterminer des unités susceptibles d'être utilisées pour quantifier l'information linguistique et dans un deuxième temps, d'étudier l'agencement de ces unités dans la chaîne parlée. Cette étude est faite dans une perspective translinguistique de manière à valider l'hypothèse d'une interaction en matière d'information entre les dimensions syntagmatique et paradigmatique. Nous ne sommes évidemment pas les premiers à formuler ce type d'hypothèse (voir par exemple Hockett, 1953:90 ; Karlgren, 1961 ou plus récemment, S. Greenberg, & Fosler-Lussier, 2000 ; Locke, 2008) mais il nous semble qu'aujourd'hui, les corpus de parole et les bases de données disponibles permettent d'aller plus avant dans l'expérimentation. La démarche adoptée se déroule en trois phases :

- Étape 1 : détermination d'unités potentiellement pertinentes et quantification de leur information shannonienne ;

[58] Nous utilisons les termes de dimension paradigmatique et de dimension syntagmatique dans un sens très large : le premier fait référence aux unités linguistiques d'une langue (phonèmes, syllabes, morphèmes, etc.) et le second à l'agencement séquentiel et temporel de ces unités (les phonèmes au sein des syllabes, les syllabes au sein des mots et les morphèmes au sein des propositions).

- Étape 2 : évaluation de ces unités comme mesures d'information linguistique : si elles sont pertinentes, plus elles porteront d'information dans une langue donnée, moins elles seront nombreuses pour un contenu sémantique donné ;
- Étape 3 : étude d'une éventuelle interaction entre l'information portée par ces unités et leur utilisation temporelle, menant à une régulation du débit d'information.

✦ Détermination des unités d'information

Dans le cadre de la théorie shannonienne de l'information, nous considérons une langue L comme une source d'énoncés, définis comme des séquences de symboles, eux-mêmes choisis dans un inventaire de taille finie. L'information apportée en moyenne par chaque symbole est alors quantifiée par l'entropie H_L de la source L, connaissant N le cardinal de l'inventaire de symboles et $\{ p_i \}$, $\forall\ 1 \leq i \leq N_L$, les probabilités d'apparition de chaque symbole i. Si les symboles sont indépendants, c'est-à-dire si leur probabilité d'apparition à l'instant t ne dépend pas des symboles apparus précédemment dans la séquence, H_L est défini par :

$$H_L = -\sum_{i=1}^{N_L} p_i . \log_2(p_i)$$

Pour mettre en œuvre cette quantification, il est donc nécessaire :

- De pouvoir segmenter la séquence linguistique en unités – ou symboles – discrets sur un axe unidimensionnel ;
- de connaître l'inventaire de dimension finie de ces unités ou symboles et d'être capable d'estimer leurs probabilités d'apparition dans la langue ;
- de respecter l'indépendance conditionnelle entre les unités au sein des séquences ou phrases.

Notre étude est orientée vers la quantification de l'information dans la communication parlée, et non dans le codage écrit de la langue ; le codage orthographique n'est donc pas adapté et les principales unités potentiellement utilisables sont les segments, les syllabes, les morphèmes et les mots. Les traits phonétiques ou les gestes articulatoires, même s'ils présentent des caractéristiques intéressantes comme primitives phonologiques, seraient plus difficiles à utiliser du fait de leur caractère multidimensionnel (constellations de gestes ou matrices de traits).

Pour les mots, il est possible de dresser un inventaire approximatif et d'estimer les fréquences d'apparition à partir de grands corpus ou de lexiques, même si chacune de ces deux sources introduit des biais potentiels (surévaluation de certaines fréquences liées à quelques morphèmes très fréquents ou à l'inverse minimisation de l'incidence des morphèmes grammaticaux). Dans le cas des segments et des syllabes, on peut dresser un inventaire précis et non plus approximatif, avec les mêmes biais cependant. Le cas des morphèmes est plus

complexe, en particulier du fait de l'existence de morphèmes nuls et de morphèmes porte-manteau. L'hypothèse d'indépendance entre unités est particulièrement contraignante et son respect strict semble hors de portée : les modèles de langage en RAP se fondent sur les probabilités *n*-grammes des mots, preuve de leur interdépendance ; les morphèmes ne sont généralement pas indépendants et, enfin, des contraintes syllabotactiques et phonotactiques sont à l'œuvre dans toute langue. De plus, plus l'unité choisie sera longue (mots *vs*. phonèmes par exemple), plus la taille du corpus utilisé pour évaluer les fréquences doit être importante. Nous nous sommes donc orientés vers le niveau phonologique, et plus précisément le niveau syllabique de préférence aux segments phonétiques. En effet, les contraintes phonotactiques sont fortes dans chaque langue, et elles sont rarement violées. À l'inverse, les règles d'enchaînement dites syllabotactiques relèvent plus de régularités morphologiques et sont en nombre plus limité. Un autre avantage du niveau syllabique est que, même si une forte réduction opère en parole spontanée au niveau segmental, il est plus rare qu'il y ait omission complète d'une syllabe entière. Cela se produit néanmoins dans environ 1 % des cas dans le corpus Switchboard, à comparer aux 22 % d'omission rapportés au niveau segmental (S. Greenberg, 1999). De même d'autres limitations doivent également être gardées en mémoire ; les phénomènes de resyllabation ou d'ambisyllabicité introduisent un biais, en particulier à l'oral et d'autant plus que la plupart des bases de données permettant d'établir les fréquences syllabiques se basent sur des corpus écrits.

Dans notre étude en cours, les entropies syllabiques ont été estimées pour les sept langues du corpus MULTEXT déjà évoquées au chapitre précédent : allemand, anglais, espagnol, français, italien, japonais et mandarin. Le calcul repose sur les inventaires syllabiques et les fréquences relatives d'apparition des syllabes établis à partir de bases de données conçues soit dans un but linguistique, soit psycholinguistique, à partir de corpus écrits. Sauf mention contraire dans la suite, une méthode de *bootstrapping* (algorithme implémenté sous Matlab) a été appliquée pour obtenir une évaluation robuste des statistiques calculées, en particulier pour l'entropie.

Le Tableau 20 donne pour chaque langue, la source utilisée, la taille de l'inventaire syllabique et la taille du corpus ayant servi à son établissement, ainsi que l'entropie syllabique calculée par nos soins (pour les détails, voir Pellegrino, Marsico & Coupé, en préparation). Étant donné la charge fonctionnelle des tons en mandarin (Surendran & Levow, 2004), nous avons distingué chaque syllabe en fonction du ton porté, obtenant ainsi un inventaire de 1 191 syllabes.

Bien que nous ne disposions que d'un faible échantillon d'à peine sept langues, l'entropie syllabique varie nettement, de 6,02 pour le japonais à 9,14 pour l'anglais, soit 50 % de plus. La corrélation entre la taille de l'inventaire et l'entropie n'est pas significative, même si une tendance est observée (r^2 = 0,51 ; p = 0,07), alors même que la limite théorique que peut atteindre l'entropie est strictement déterminée par la taille de l'inventaire : elle est égale à $\log_2(N_L)$ et elle est atteinte dans le cas où la distribution des probabilités de syllabes est uniforme (Shannon & Weaver, 1949). Cela signifie que, outre le fait bien établi que la taille de l'inventaire syllabique varie

d'une langue à l'autre, la part relative de l'entropie syllabique et de ce que la théorie de l'information nomme redondance varie également d'une langue à l'autre. Le terme de redondance fait référence au fait que le canal de transmission n'est pas utilisé au maximum de sa capacité puisque, pour un débit constant (en nombre de symboles par seconde), un débit d'information supérieur pourrait être atteint.

Tableau 20 – Calcul de l'entropie syllabique : Description des données sources et valeurs d'entropie syllabique estimées. Les valeurs fournies entre parenthèses correspondent aux intervalles de confiance établis par *bootstrapping*.

LANGUE (CODE)	SOURCE	NBRE DE SYLLABES DIFFÉRENTES N_L	NBRE TOTAL DE SYLLABES (EN MILLIONS)	ENTROPIE SYLLABIQUE H_L
Anglais (EN)	WebCelex (Baayen, Piepenbrock & Rijn, 1993)	7 931	1,0	9,14 (± 0,007)
Français (FR)	Lexique3 (New *et al.*, 2004)	5 685	1,3	8,43 (± 0,008)
Allemand (GE)	WebCelex (Baayen, Piepenbrock & Rijn, 1993)	4 207	0,8	8,27 (± 0,008)
Italien (IT)	(Pone, 2005)	2 719	27,0	7,67 (± 0,008)
Japonais (JA)	(Tamaoka & Makioka, 2004)	416	575,7	6,03 (± 0,010)
Mandarin (MA)	(Peng, 2005)	1 191 (tons inclus)	138,0	8,58 (± 0,006)
Espagnol (SP)	(Pone, 2005)	1 593	0,9	7,75 (± 0,008)

✦ Comparaison translinguistique de l'information syllabique

La mesure d'entropie établie ci-dessus n'a pas de valeur linguistique intrinsèque. En particulier, on peut s'interroger sur son lien éventuel avec l'information linguistique ou sémantique véhiculée par un énoncé. Pour cela, il est nécessaire de calculer l'information linguistique moyenne portée par les syllabes dans différentes langues, puis d'évaluer si cette information est corrélée à l'entropie syllabique calculée précédemment.

L'idéal serait de disposer d'un corpus multilingue où le contenu sémantique a été contrôlé de manière très stricte. En suivant cette piste, Fenk-Oczlon & Fenk (1999) ont traduit un ensemble de 22 propositions simples (« le sang est rouge » ; « le soleil brille », etc.) en 34 langues pour réaliser une étude quantitative translinguistique. Selon nous, cette méthodologie présente deux inconvénients majeurs. Le premier est que, pour minimiser l'impact éventuel de chaque mot du lexique pris individuellement, il est préférable d'utiliser des énoncés longs plutôt que courts. Le second est qu'avec des phrases aussi brèves, on s'éloigne de ce que serait une élocution normale et cela empêche d'étudier par exemple les variations de débit au cours de l'énonciation.

Notre étude s'appuie sur une sous-partie du corpus MULTEXT constituée de vingt textes de cinq phrases sémantiquement connectées. Certains sont des textes narratifs, tandis que d'autres sont écrits comme des requêtes téléphoniques.

L'ensemble de ces textes avait été traduit dans le cadre du projet EUROM dans les différentes langues sans qu'il s'agisse de traduction mot à mot. Comme le montrent les exemples du Tableau 21, certaines traductions sont très proches, mais il reste cependant des différences. À condition que le biais de traduction ne soit pas systématique, le fait de travailler avec vingt textes devrait permettre de dégager une tendance moyenne qui corresponde à la part de variation due à la langue (et non aux textes eux-mêmes).

Tableau 21 – Exemples de deux textes du corpus MULTEXT, dans leurs versions française, anglaise et espagnole.

	TEXTE NARRATIF (N° P8)	**REQUÊTE (N°O2)**
VERSION FRANÇAISE	Hier soir, j'ai ouvert la porte d'entrée pour laisser sortir le chat. La nuit était si belle que je suis descendu dans la rue prendre le frais. J'avais à peine fait quelque pas que j'ai entendu la porte claquer derrière moi. J'ai réalisé, tout d'un coup, que j'étais enfermé dehors. Le comble c'est que je me suis fait arrêter alors que j'essayais de forcer ma propre porte !	Passez-moi les réclamations, s'il vous plaît. On est venu réparer le tuyau d'arrivée d'eau, devant chez moi, et ça n'a pas tenu : ma cave est inondée. Quand j'ai téléphoné, on m'a répondu que toutes les équipes de dépannage étaient occupées pendant les deux semaines qui viennent. On peut vraiment pas faire confiance au Service des Eaux. Si j'ai bien compris, en attendant, ma cave va me servir de piscine.
VERSION ANGLAISE	Last night I opened the front door to let the cat out. It was such a beautiful evening that I wandered down the garden for a breath of fresh air. Then I heard a click as the door closed behind me. I realised I'd locked myself out. To cap it all, I was arrested while I was trying to force the door open!	Please put me through to the complaints department. The repair to the water main outside my house was unsuccessful, and my cellar's flooded. Your Water Services Department was singularly unsympathetic. All their repair teams are apparently booked out for the next two weeks. Am I supposed to use the cellar as a swimming pool till then?
VERSION ESPAGNOLE	Anoche, abrí la puerta del jardín para sacar al gato. Hacía una noche tan buena que pensé en dar un paseo y respirar el aire fresco. De repente, se me cerró la puerta. Me quedé en la calle, sin llaves. Para rematarlo, me arrestaron cuando trataba de forzar la puerta para entrar.	Por favor, póngame con el departamento de reclamaciones. Con la reparación de la cañería exterior de mi casa me han hecho una chapuza y se ha inundado el sótano. Su Departamento de Servicios al Cliente no me ha hecho ningún caso. Me han dicho que todos los fontaneros están ocupados durante las dos próximas semanas. ?`Se supone que hasta que vengan tengo que utilizar el sótano como piscina?

Pour chaque texte et dans chaque langue, plusieurs paramètres ont été calculés par des locuteurs natifs ou très compétents : nombre de syllabes phonologiques, nombre de mots, durée moyenne d'énonciation par les locuteurs. Une difficulté supplémentaire vient du fait que les textes, dans une langue donnée, peuvent contenir un nombre de syllabes très variable (par exemple de 62 à 104 pour l'anglais et de 72 à 123 pour le japonais). Calculer des tendances centrales (moyenne ou médiane), alors que ces valeurs sont uniquement déterminées par le contenu textuel (et non par la langue) n'aurait donc que peu de sens. Pour surmonter cette difficulté, nous avons utilisé une langue externe au corpus comme référence, avec l'assistance d'Eric Castelli, du laboratoire MICA (Hanoï, Vietnam). Il a fait procéder à la traduction en vietnamien du corpus de vingt textes et l'a fait enregistrer par des locuteurs natifs. Ces données ont servi de référence pour normaliser les durées des énoncés (en secondes) et leur longueur (en nombres de syllabes). Ainsi, la longueur

de chaque texte a été divisée, pour chaque langue, par la longueur du texte équivalent en vietnamien. On obtient ainsi des ratios sans dimension qui ne sont plus liés au choix du texte source.

La densité d'information linguistique de la langue *L* (*Information Density* ID_L) est ensuite définie comme le ratio du nombre de syllabes en vietnamien divisé par le nombre de syllabes dans la langue *L*. Il s'agit donc d'une densité *syllabique* d'information linguistique. Une langue où les textes nécessitent deux fois plus de syllabes qu'en vietnamien aura une densité d'information par syllabe de 0,5 (par rapport au vietnamien), tandis qu'une langue où le même texte contient deux fois moins de syllabes aura une densité d'information de 2 (toujours par rapport au vietnamien). Une fois ces valeurs calculées individuellement pour chacun des vingt textes, une valeur moyenne de ID_L est calculée par langue en utilisant une méthode de *bootstrapping* pour pallier à la faible taille de l'échantillon. Le Tableau 22 récapitule pour chaque langue les caractéristiques du corpus et la densité d'information calculée.

Tableau 22 – Corpus de parole (basés sur MULTEXT, Campione & Véronis, 1998) utilisés pour la comparaison inter-langue. Les valeurs entre parenthèses définissent l'intervalle de confiance de ID_L, établi par *bootstrapping*. (*) pour le vietnamien, chacun des quatre locuteurs a répété le texte deux fois.

LANGUE	SOURCE	NBRE DE LOCUTEURS	DURÉE TOTALE	DENSITÉ D'INFORMATION SYLLABIQUE ID_L	DURÉE NORMALISÉE D_L
EN	Multext	10	18 min.	0,91 (± 3 ‰)	0,95 (± 3 ‰)
FR	Multext	6	14 min.	0,73 (± 2 ‰)	1,01 (± 3 ‰)
GE	Multext	10	27 min.	0,82 (± 2 ‰)	1,10 (± 3 ‰)
IT	Multext	10	18 min.	0,70 (± 2 ‰)	1,01 (± 4 ‰)
JA	(Kitazawa *et al.*, 2002)	5	33 min.	0,49 (± 1 ‰)	1,35 (± 3 ‰)
MA	(Komatsu, Arai & Sugarawa, 2004)	9	23 min.	0,94 (± 1 ‰)	1,10 (± 2 ‰)
SP	Multext	8	17 min.	0,63 (± 2 ‰)	1,08 (± 3 ‰)
VI	Mis à disposition par E. Castelli, MICA	4(*)	38 min.	1 (référence)	1 (référence)
TOTAL		**62**	**3h 08 min.**	**-**	

La première constatation est qu'aucune des langues de l'échantillon ne présente une densité syllabique d'information supérieure au vietnamien. Le mandarin est la langue qui s'en rapproche le plus (0,94) ce qui suggère des stratégies d'encodage proches. Ce résultat est tout à fait compatible avec leur proximité lexicale et morpho-syntaxique, même si, contrairement au vietnamien, le mandarin n'est pas une langue mono-syllabique au sens strict. Le japonais, avec une valeur de densité de 0,49, est la langue s'éloignant le plus de la référence vietnamienne. Cela peut éventuellement mettre en évidence que le niveau syllabique n'est pas le plus adapté, soit parce que les morae sont plus pertinentes, soit parce que pour cette langue à la morphologie concaténative, les relations syllabotactiques revêtent une importance majeure.

En matière de durée, le français et l'italien ont des caractéristiques très proches, tandis que le japonais présente en moyenne des durées 35 % supérieures à celles observées en vietnamien. L'anglais est la seule langue à présenter une durée moyenne plus brève que le vietnamien. Le fait même que les durées moyennes diffèrent d'une langue à l'autre répond en partie à notre question initiale en invalidant l'hypothèse d'un débit d'information constant à travers les langues : en effet, si tel était le cas – la quantité d'information étant supposée identique dans chacune des langues étudiées – on obtiendrait des durées identiques. Ce constat n'invalide cependant pas l'hypothèse déjà évoquée selon laquelle il existerait une interaction entre des contraintes de débit d'information minimal (pour des raisons fonctionnelles) et maximal (pour des raisons cognitives).

Nous avons donc à notre disposition d'une part une mesure de la densité syllabique d'information ID_L, obtenue à partir du corpus Multext et d'autre part une mesure d'entropie syllabique H_L calculée à partir de bases de données indépendantes. La Figure 24 représente la relation entre ces deux mesures, qui montre une corrélation positive élevée significative (test de Spearman, $\rho = 0{,}89$; $p < 0{,}05$). Visuellement, il semble possible que le japonais soit responsable de cette corrélation ; si on écarte cette langue, la corrélation n'est en effet plus significative pour les six langues restantes, mais une tendance est observée ($\rho = 0{,}83$; $p = 0{,}06$). On atteint évidemment là la limite de ce qui peut être fait avec un nombre de langues si réduit et une extension sera donc nécessaire pour étudier de manière plus robuste cette relation.

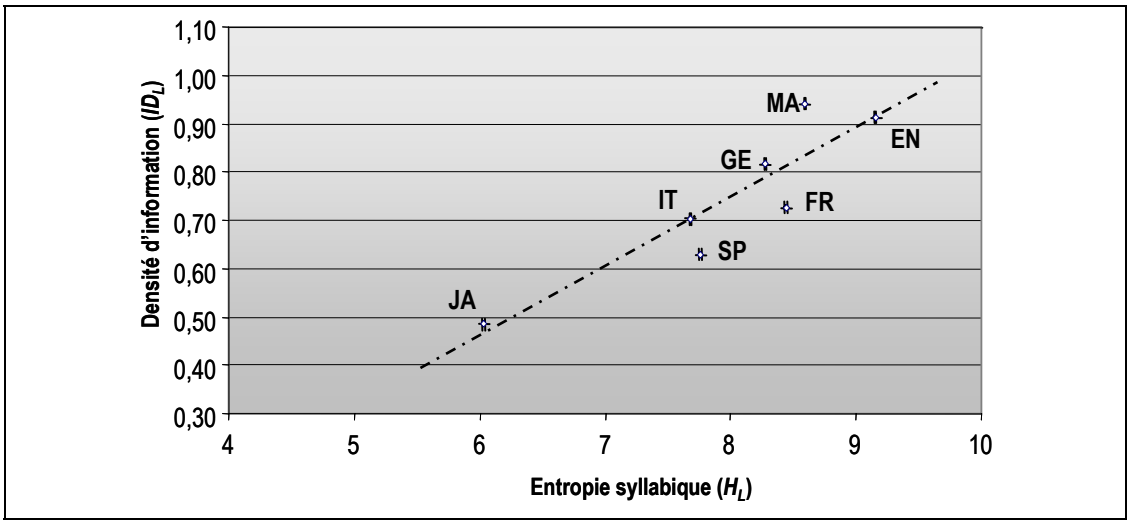


Figure 24 – Relation entre l'entropie syllabique H_L (en abscisse) et la densité d'information ID_L (en ordonnée). Les barres correspondent aux intervalles de confiance des estimations ; la ligne pointillée est la meilleure régression linéaire au sens des moindres carrés.

✦ Lien entre complexité syllabique et information syllabique

L'entropie syllabique, mesurée selon l'axe paradigmatique, semble donc mesurer une quantité d'information moyenne portée par les syllabes, telle qu'on peut l'évaluer selon un axe syntagmatique. Nous allons maintenant chercher à déterminer dans quelle mesure cette entropie est liée à la complexité syllabique.

Dans les rares études ayant abordé le sujet, la complexité d'une syllabe est évaluée par le nombre de segments la composant, augmenté éventuellement d'une unité lorsque un ton est également porté (voir Maddieson, 2006 pour une étude typologique, ou Service, 1998 et Mueller, Seymour, Kieras & Meyer, 2003, dans une perspective psycholinguistique). Il est ainsi possible d'évaluer une longueur moyenne de la syllabe, à partir des longueurs des syllabes de la langue. Si l'on ne tient pas compte de la distribution de fréquence des syllabes mais uniquement de l'inventaire, on parle de longueur moyenne des types syllabiques, alors que si la longueur de chaque syllabe est pondérée par sa fréquence d'apparition, on parle de longueur moyenne d'occurrence.

Dans le cas du mandarin, seule langue tonale de notre échantillon, on peut distinguer d'une part une longueur moyenne, correspondant au nombre de segments phonétiques et d'autre part un nombre de constituants, correspondant au nombre de segments augmenté d'un élément d'information (et de complexité) tonal. Les phénomènes tonaux sont complexes en eux-mêmes et il est certain que cette approximation est extrêmement grossière, ne serait-ce que du fait de l'application du phénomène de sandhi. Par conséquent, la valeur ainsi calculée est une borne supérieure, chaque syllabe portant *au maximum* un ton.

Tableau 23 – Évaluation de la complexité moyenne des syllabes (en nombre de constituants). Pour le mandarin, les valeurs « segments + ton » correspondent à l'ajout d'une unité pour prendre en compte le ton porté. (Voir le texte pour les détails ; mêmes sources que dans le Tableau 20)

LANGUE	COMPLEXITÉ SYLLABIQUE (TYPE)	COMPLEXITÉ SYLLABIQUE (OCCURRENCE)
EN	3,70	2,48
FR	3,50	2,21
GE	3,70	2,68
IT	3,50	2,30
JA	2,65	1,93
MA (segments) MA (segments + ton)	2,89 3,89	2,63 3,63
SP	3,30	2,40

La complexité la moins importante, à la fois par type et par occurrence, est observée en japonais, tandis que les valeurs les plus importantes concernent l'anglais et l'allemand. Si l'on prend toutefois en compte le ton, le mandarin présente la complexité la plus élevée. En moyenne, la différence entre les mesures de complexité par type et par occurrence est de l'ordre d'un segment (0,94). On retrouve là un effet connu, à savoir que, même dans les langues pour lesquelles des structures syllabiques complexes sont présentes dans le lexique, on observe une tendance à utiliser préférentiellement des syllabes simples à l'oral (e.g. S. Greenberg, 1999).

Tableau 24 – Coefficients de corrélation (ρ de Spearman) établi entre les statistiques d'information (IDL et HL) et de complexité des syllabes (par type et par occurrence). Les

valeurs entre parenthèses correspondent aux mesures sans prise en compte du ton pour le Mandarin. (** : p < 0,01 ; * : p < 0,05 ; n.s. non significatif).

	COMPLEXITÉ SYLLABIQUE (TYPE)	COMPLEXITÉ SYLLABIQUE (OCCURRENCE)
ID_L	0,98 ** (0,44 n.s.)	0,82 * (0,75 n.s.)
H_L	0,80 * (0,47 n.s.)	0,64 n.s. (0,57 n.s.)

Le Tableau 24 présente les corrélations (test de Spearman) établies entre les deux statistiques d'information (ID_L et H_L) et les indices de complexité des syllabes (par type et par occurrence). Le mandarin a un impact majeur sur ces statistiques et dans le cas où le ton n'est pas intégré à la complexité syllabique, aucune corrélation significative n'est mise en évidence entre information et complexité. À l'inverse, si le ton est intégré, on trouve une relation extrêmement forte entre ID_L et la complexité moyenne par type syllabique ($\rho = 0{,}98$; $p < 0{,}01$), comme illustré Figure 25.

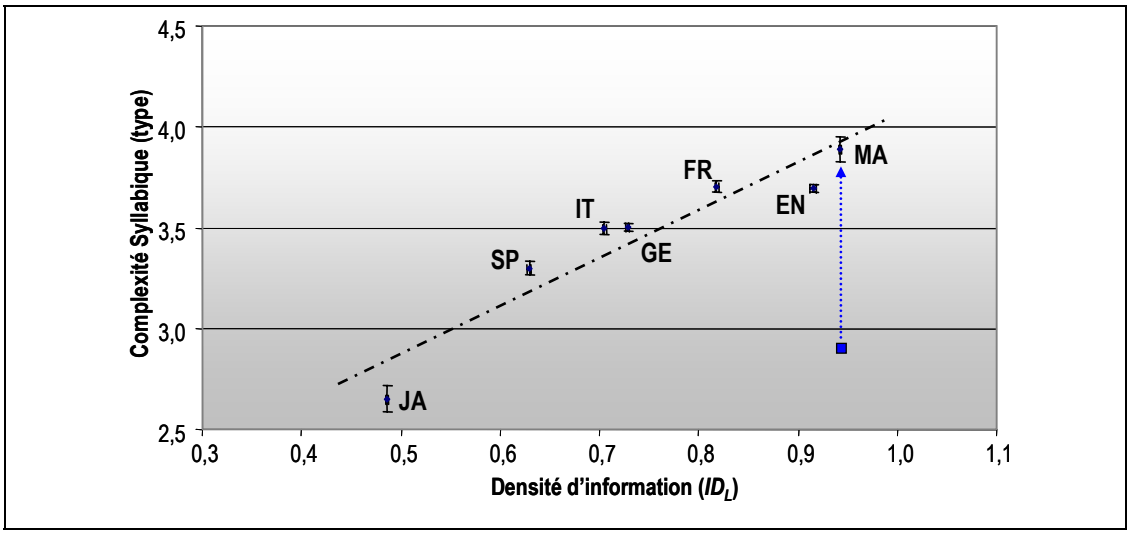


Figure 25 – Relation entre la densité d'information (en abscisse) et la complexité syllabique calculée par type, en prenant le ton en compte pour le mandarin. Le carré et la flèche bleus indiquent la position du mandarin sans prise en compte du ton. La ligne pointillée est la meilleure régression linéaire au sens des moindres carrés.

✦ *Une régulation du débit d'information ?*

Le débit temporel d'information est défini par le produit de la densité d'information syllabique (ID_L) par le débit syllabique (en nombre de syllabes par seconde). Comme cela a déjà été indiqué plus haut, le fait même que les durées moyennes d'énonciation varient d'une langue à l'autre indique qu'il n'existe pas de contrainte forte sur un débit d'information préféré et constant parmi les langues. L'étude du débit syllabique moyen observé pour chacune des langues du corpus met cependant en évidence des variations inter-linguistiques. Une analyse ANOVA a révélé que l'identité de la langue explique 60,7 % de la variance observée et que certaines différences sont significatives entre des sous-groupes disjoints de langues (mandarin < allemand et anglais < italien et français < japonais et espagnol). Si l'on se limite à cette comparaison d'un corpus de parole lue, enregistré dans des

conditions similaires, l'affirmation selon laquelle les différences de débit dues à la langue est un mythe (Roach, 1999) semble donc invalidée.

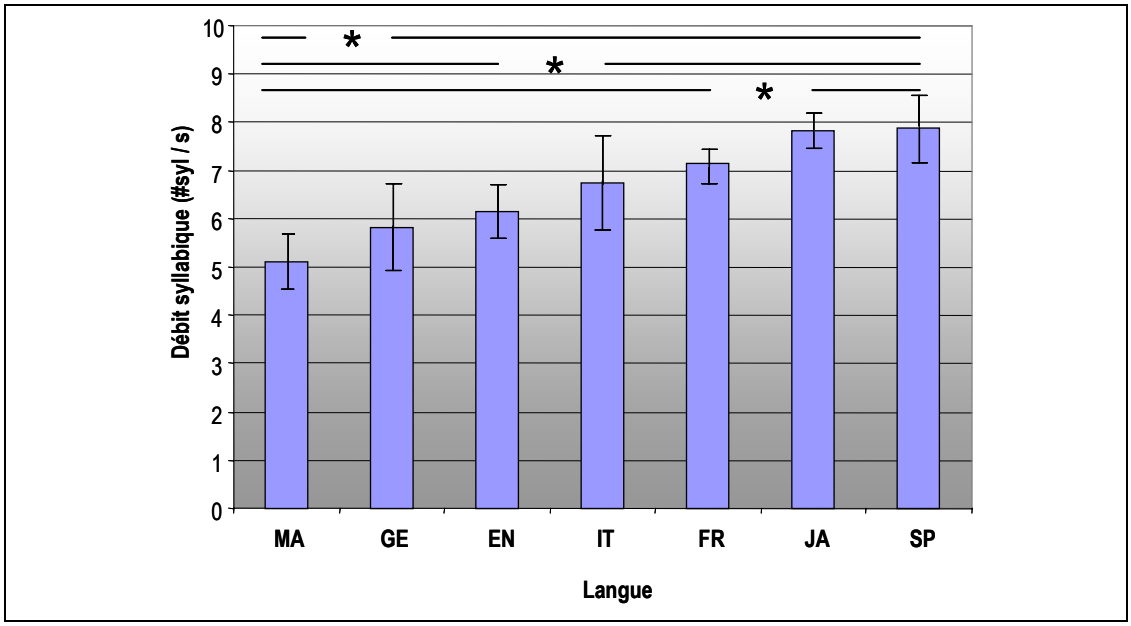


Figure 26 – Débit syllabique moyen observé par langue. Les intervalles représentent les écarts-types et les « * » séparent les groupes pour lesquelles des différences significatives sont observées.

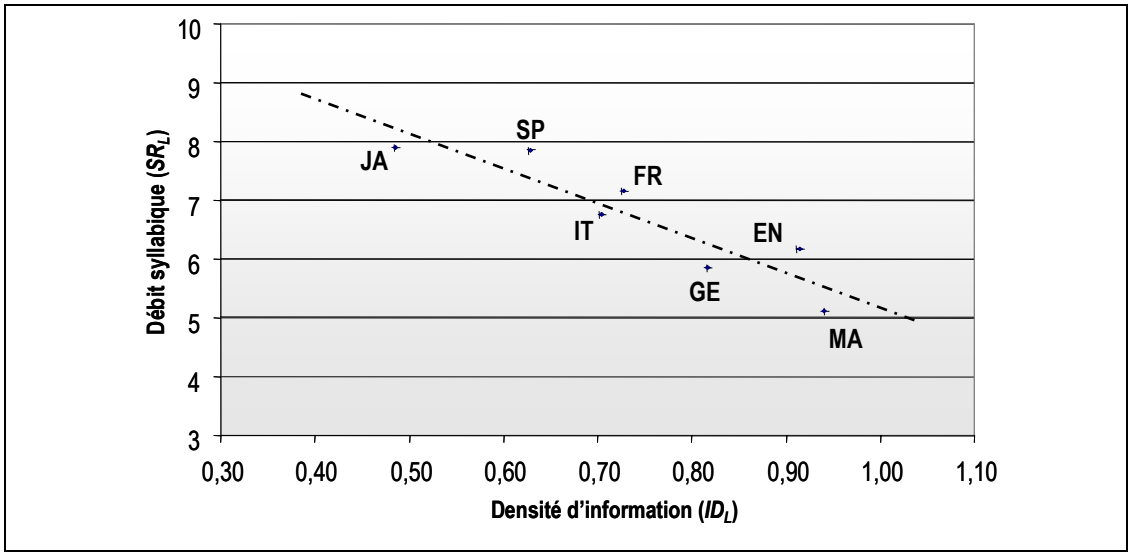


Figure 27 – Interaction entre la densité d'information (abscisse) et le débit syllabique (ordonnée). La ligne pointillée est la meilleure régression linéaire au sens des moindres carrés.

La Figure 27 met en évidence une forte interaction négative très significative entre la densité d'information syllabique et le débit syllabique (ρ = -0,93 ; $p < 0,01$). Dans cet échantillon de langues, on n'observe donc pas le cas, évoqué précédemment comme hypothétique, où une forte densité informationnelle est couplée à un débit rapide.

La Figure 28 permet d'appréhender de manière plus globale la stratégie suivie par chacune des langues pour encoder l'information. Pour des raisons de lisibilité, la

densité d'information syllabique ID_L est multipliée par 10 sur la figure et elle peut donc être affichée sur la même échelle que le débit syllabique. On observe bien l'interaction négative entre densité d'information (barres vertes) et débit syllabique (barres bleues), l'une étant croissante alors que l'autre décroît. La courbe noire représente le débit d'information, défini comme étant l'inverse de la durée normalisée présentée précédemment dans le Tableau 22. L'anglais, qui était la seule langue à présenter une durée moyenne d'énoncé plus faible que le vietnamien est la seule langue ayant un débit d'information supérieur à 1. À l'inverse, le japonais, qui avait les énoncés les plus longs, est caractérisé par un débit d'information de 0,74. On constate que deux langues peuvent atteindre un débit d'information proche (par exemple l'espagnol et l'allemand), tout en ayant des stratégies d'encodage nettement différenciées : l'espagnol est caractérisé par un débit rapide de syllabes peu denses en information alors que l'allemand se base sur un débit plus lent d'environ 25 % mais avec des syllabes plus denses d'environ 30 %.

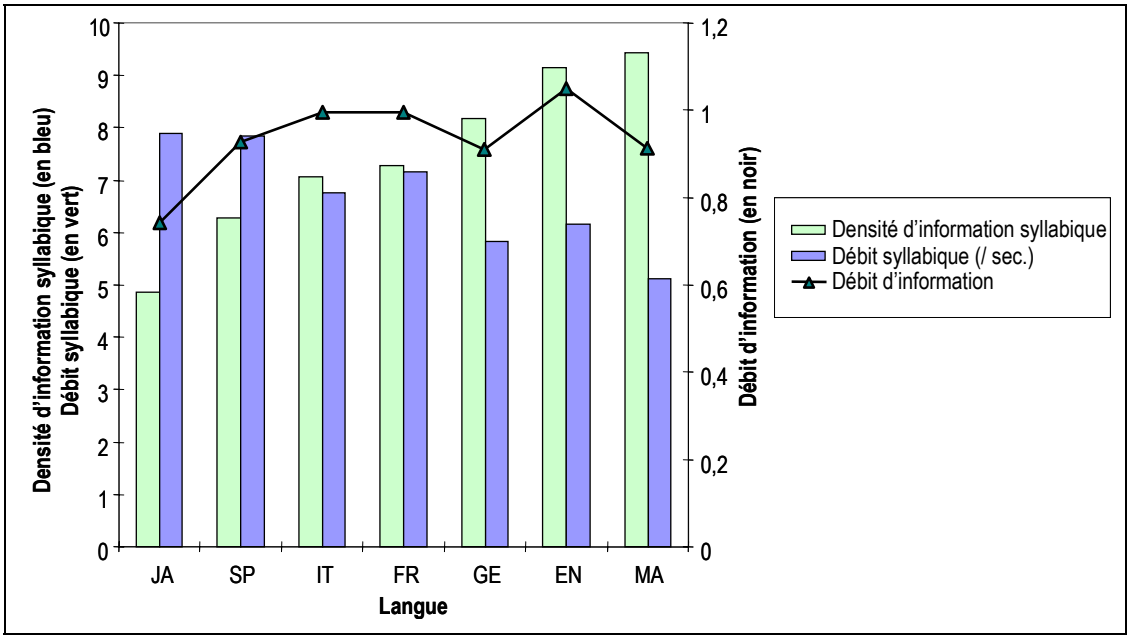


Figure 28 – Comparaison des stratégies linguistiques d'encodage de l'information. Les barres verte et bleue représentent respectivement la densité d'information syllabique ID_L et le débit syllabique (axe de gauche). Pour des raisons de lisibilité, les valeurs d'ID_L sont multipliées par 10. Les triangles noirs matérialisent le débit d'information (axe de droite). Les langues sont ordonnées par ID_L croissante.

✦ *Discussion*

L'étude présentée dans cette section 3.3.2 visait à étudier l'interaction entre l'information phonologique, estimée au niveau syllabique, et le débit de parole. À partir d'un corpus de sept langues (plus le vietnamien, utilisé comme point de référence), nous avons montré d'une part que la densité d'information encodée par syllabe varie d'une langue à l'autre et d'autre part que le débit syllabique varie également.

Les sociolinguistes ont souvent considéré que les variations de débits au sein d'une population étaient régies par des considérations liées au style de parole et

donc au contexte de communication et au statut social (Brown, Giles & Thakerar, 1985 ; Wells, 1982). À l'inverse, l'existence de différences systématiques entre populations parlant des langues différentes est rarement évoquée, à l'exception de l'hypothèse formulée par Trudgill (2004) selon laquelle la parole est produite avec un débit rapide de manière plus fréquente dans de petites communautés isolées. Pour sa part, notre étude met en évidence une variation de débit qui ne semble pas imputable à des raisons d'ordre sociolinguistique.

Dans un second temps, en utilisant des statistiques sur les fréquences de syllabes établies à partir de corpus écrits de grande taille, nous avons suggéré l'existence d'un lien fort entre l'entropie syllabique (qui représente une mesure paradigmatique de l'information shannonienne) et la densité d'information des syllabes (qui représente une mesure syntagmatique de l'information linguistique). Nous avons également soumis l'hypothèse d'une relation quantifiable entre cette densité et la complexité des types syllabiques d'une langue, même si la prise en compte des tons du mandarin rend la prudence particulièrement nécessaire sur ce point. Une forte interaction négative a enfin été mise en évidence entre la densité d'information syllabique et le débit syllabique validant ainsi l'une des hypothèses proposées initialement. À l'heure actuelle, il est difficile d'évaluer si cette interaction relève d'un principe de moindre effort et d'une régulation du débit d'information (sans qu'il soit pour autant strictement contrôlé) ou d'un artefact lié à la complexité syllabique. Dans ce dernier cas, une interprétation crédible serait la suivante : une langue qui disposerait d'un riche inventaire de syllabes complexes présenterait une entropie syllabique élevée. Parallèlement, la complexité de ses syllabes (en nombre de segments) impliquerait une durée d'articulation et donc une durée syllabique moyenne plus élevée. Par conséquent, son débit syllabique moyen serait plus lent.

De manière générale, le type d'investigation que nous menons sur ce thème est particulièrement difficile du fait des multiples corrélations existant entre tailles des inventaires vocaliques, consonantiques, tonaux et syllabiques (voir en particulier Maddieson, 2007). Plusieurs pistes peuvent cependant être envisagées pour mieux comprendre les mécanismes en jeu. La première est évidemment d'ajouter d'autres langues au corpus, ce qui nécessite à la fois des enregistrements comparables à ceux de MULTEXT et des statistiques de fréquences syllabiques. Si l'on souhaite étendre l'étude à un nombre conséquent de langues, cela implique de se focaliser en premier lieu sur des langues pour lesquelles des corpus de textes écrits (ou de transcriptions orales) sont disponibles. La diversification du corpus est cependant incontournable pour obtenir des résultats robustes, confirmant ou infirmant les premières tendances dégagées ici. L'intégration de langues issues d'aires et de familles linguistiques diverses paraît donc indispensable. En complément, l'étude de langues proches mais présentant des différences de densité syllabique d'information, comme par exemple les langues chinoises ou les langues romanes, peut également s'avérer très utile pour affiner l'interprétation des résultats, éventuellement dans une perspective diachronique. Il serait également intéressant d'étudier plusieurs langues présentant des inventaires syllabiques proches mais avec par exemple une langue tonale et une langue non tonale, pour étudier les variations éventuelles d'entropie et de débit syllabiques. Pour rassembler de telles données, la difficulté la plus grande

réside probablement dans la nécessité d'estimer l'entropie syllabique à partir de grands corpus. L'alternative qui consiste à évaluer cette mesure à partir de lexiques ou de dictionnaires mérite également d'être testée, même s'il est probable que le biais introduit (lié en particulier à des différences morphologiques) rende toute comparaison inter-langue difficile ou peu fiable.

Une autre piste que nous n'avons pas encore explorée porte sur la variation de débit et de densité d'information au long d'un énoncé. En effet, nous avons pour l'instant raisonné en termes de valeurs moyennes car nous visions à établir des tendances générales dans les langues étudiées. Cependant, la principale caractéristique de la parole repose sur sa variabilité et il est très probable qu'une part très importante de l'information langagière réside dans cette variation. L'entropie syllabique est une valeur globale, mais il est évident qu'il y a des fluctuations de l'information apportée par les différentes syllabes composant un message, tout comme il y a des fluctuations de débit. Il sera donc particulièrement intéressant d'évaluer si ces fluctuations montrent des régularités, qu'il s'agisse d'une contrainte attentionnelle, comme suggéré par de Jong (2002), ou de ce qu'Ohala appelle le contraste syntagmatique (Ohala, 1995, 2008).

Ce dernier point soulève plusieurs questions sur la nature de l'interaction entre débit et densité d'information. S'il s'agit de l'artefact évoqué précédemment, pour que la densité d'information augmente, il faut nécessairement que la durée physique des syllabes augmente. Cela n'empêche cependant pas de tester si la courbe obtenue présente des caractéristiques pertinentes au niveau neurocognitif, en termes de traitement de l'information. S'il ne s'agit pas d'un artefact, son origine est peut-être à chercher dans des principes généraux des systèmes complexes, dans lesquels l'existence de contraintes contradictoires mène à de tels équilibres dynamiques caractéristiques de principes du moindre effort. Dans ce cas précis, il s'agirait d'un équilibre entre le débit minimal d'information à transmettre (pour des raisons fonctionnelles) et le débit maximal (pour des raisons cognitives de mémorisation et de traitement de l'information). Dans Dellwo, Ferragne & Pellegrino, (2006), nous avons mené une expérience qui apporte un éclairage complémentaire à ces résultats. Des auditeurs francophones devaient qualifier en termes de débit sur une échelle graduée et étiquetée de « très lent » à « très rapide » des extraits entendus en allemand, en anglais et en français. Les extraits eux-mêmes avaient été produits à différents débits par les locuteurs, dans le cadre du projet BonnTempo (Dellwo *et al.*, 2004). De manière remarquablement consistante, les sujets identifiaient très facilement les extraits produits à débit normal quelle que soit la langue, alors même que le débit objectif syllabique variait de 5,36 syl/s en allemand à 6,80 syl/s en français. À partir de ce résultat, on peut émettre au moins deux hypothèses : la première consiste à dire que, même dans une langue étrangère, on perçoit une augmentation de l'effort vocal produit par un locuteur lorsque son débit s'éloigne de son débit normal (hypothèse orientée « production ») ; la seconde interprétation tendrait à considérer que d'une langue à l'autre, le débit normal transmet un flux d'information spectrale comparable (hypothèse orientée « perception »). Quelle que soit l'interprétation qui s'avèrera correcte, il est évident

que cette expérience demande à être répliquée avec d'autres langues, à la fois pour les stimuli et comme langues maternelles des sujets.

La discussion ci-dessus m'amène également à mentionner des résultats portant sur des caractéristiques neurocognitives générales du traitement de l'information par le cerveau humain. Des expériences récentes ont, en effet, mis en évidence l'influence de phénomènes de verrouillage de phase et de synchronisation de fréquences entre des oscillations corticales et des caractéristiques spectro-temporelles de la parole (enveloppe spectrale et débit) sur la compréhension de la parole (Ahissar *et al.*, 2001). Il semblerait donc intéressant de reproduire ce même type d'expérience dans une perspective translinguistique de manière à mieux cerner les facteurs influençant les performances en compréhension.

Sur un autre plan, la notion de mémoire de travail fait référence à des structures et des processus impliqués dans le stockage temporaire et le traitement neurocognitif de l'information. L'un des modèles les plus influents intègre un sous-système nommé boucle phonologique (Baddeley, 2000 ; Baddeley & Hitch, 1974) qui maintiendrait une quantité limitée d'information verbale disponible pour la mémoire de travail durant un empan temporel donné. Un modèle mathématique intégrant une atténuation temporelle a été proposé (Schweickert & Boruff, 1986) et plusieurs facteurs pouvant influencer la capacité de la boucle phonologique ont été évoqués. En particulier, la durée d'articulation, la similarité phonologique, la longueur phonologique (c'est-à-dire le nombre de syllabes des items) sont souvent mentionnés (Baddeley, 2000; Mueller *et al.*, 2003; Service, 1998). De manière assez surprenante, la langue maternelle des sujets n'est pas considérée en tant que telle, comme un facteur potentiel. Des tâches effectuées en anglais américain oral et en langue des signes américaine ont pourtant montré des différences, mais sans que l'on puisse les imputer de manière certaine au changement de langue ou de modalité (Boutla, Supalla, Newport & Bavelier, 2004; Bavelier, Newport, Hall, Supalla & Boutla, 2006).

Pourtant, puisque des variations significatives de débit de parole sont observées d'une langue à l'autre, des expériences multilingues permettraient probablement de déterminer si une partie des résultats provient de la langue et non de capacités cognitives générales. Plus précisément, si l'on suppose que l'empan de la boucle phonologique est purement temporel, il devrait être possible de mettre en évidence qu'en fonction de la langue du sujet, le nombre de syllabes gardé en mémoire varie. À l'inverse, si la boucle phonologique maintient un nombre limité d'informations en mémoire (sans que la durée physique ne soit décisive), on peut mettre en évidence d'autres facteurs en manipulant la complexité des syllabes et la langue. En conclusion, il nous semble que des expériences translinguistiques seraient particulièrement pertinentes pour évaluer si la capacité de la mémoire de travail se mesure en termes de syllabes, de segments, de quantité d'information ou de durée physique.

De telles expériences sont cependant particulièrement difficiles à préparer du fait de la multiplicité des facteurs à contrôler pour obtenir des résultats directement comparables d'une langue à l'autre. Tout comme le traitement automatique de la

parole s'est attaqué depuis dix ans au « casse-tête » du traitement multilingue, il nous semble néanmoins qu'un des enjeux majeurs de l'étude neurocognitive du langage sera dans les prochaines années de développer ce type d'approches multilingues.

4. CONCLUSION

© Gary Larson, the far Side

Ce document a présenté un panorama des activités de recherche que j'ai menées depuis dix ans en identification des langues et dans l'étude de la complexité phonologique. Ces deux champs d'investigation s'articulent autour de plusieurs fils conducteurs que sont les dimensions translinguistique, temporelle et systémique et ils portent tous deux plus sur la communication parlée comme objet scientifique que sur des volets strictement linguistique, cognitif ou de traitement automatique du signal.

La dimension translinguistique de ce travail est évidemment explicite. Un tel positionnement à l'articulation des caractéristiques cognitives partagées par l'humanité et des spécificités liées aux différentes langues parlées me semble indispensable pour patiemment démêler l'écheveau particulièrement intriqué que

constitue la communication langagière[59]. La dimension temporelle de mes recherches a pris proportionnellement plus d'importance au fur et à mesure de ces années. À l'origine, elle apparaissait marginalement sous la forme de la durée vocalique en identification des langues, puis, lorsque la modélisation du rythme est devenue une piste prometteuse (en 2001), elle est devenue plus présente, à la fois dans la recherche d'invariants rythmiques spécifiques aux langues et en attirant mon attention sur les variations elles-mêmes, qu'il s'agisse de variations de débits entre langues ou au sein d'un même énoncé. Il me semble qu'il y a là matière à de nombreuses recherches futures, en particulier dans une perspective informationnelle. Enfin, la dimension systémique, initiée par la modélisation acoustique globale des systèmes vocaliques en identification des langues et poursuivie par les approches de mesures de complexité appliquées aux inventaires phonologiques fournit selon moi un cadre d'études particulièrement pertinent à l'interface phonétique/phonologie. Si l'on souhaite jargonner, on peut considérer qu'il s'agit là d'une sorte de « néo-structuralisme dynamique » mais je pense plus correct de placer cette approche systémique dans le cadre des Sciences de la Complexité, avec comme objectif de concilier les niveaux microscopique et macroscopique, les caractéristiques acoustico-phonétiques et les interactions entre segments au sein des systèmes phonologiques. Là encore, il me semble que les portes entrouvertes méritent d'être explorées plus avant.

Plusieurs autres aspects auraient éventuellement pu être développés dans cette synthèse. En particulier, tout le travail que nous avons entrepris à DDL depuis 2002 sur la perception de la parole dégradée (par inversion temporelle ou par ajout de bruit de type *cocktail party*) aurait mérité que l'on s'y attarde car il soulève des questions particulièrement pertinentes pour la problématique de l'identification des primitives de la communication parlée. Cependant, de par leur dimension très interdisciplinaire (impliquant outre de la phonétique, du traitement du signal, des mesures fonctionnelles du système auditif, de l'électrophysiologie, de l'imagerie cérébrale fonctionnelle, de la psycholinguistique, etc.), ces recherches reposent sur une équipe relativement importante et mon mémoire d'habilitation n'est pas le support adéquat pour un tel travail collaboratif. Le lecteur peut cependant se reporter à la liste thématique de publications données en Annexe A.

Toutefois, je mentionnerai l'un des aspects directement en lien avec les travaux présentés dans ce document. Pour cela, je m'appuierai sur des résultats obtenus avec des sujets normo-entendants qui devaient retranscrire ce qu'ils entendaient à partir de stimuli ayant subi une dégradation spectro-temporelle forte puisque une partie du signal avait été retournée temporellement (*reversed speech*). La première expérience menée à DDL avec ce protocole visait à transposer en français une expérience initialement menée avec des sujets américains, écoutant de l'anglais américain (Greenberg & Arai, 2001). La Figure 29 présente les courbes

[59] J'emploie ici le terme de *communication langagière* car je suis également convaincu que les langues des signes ou encore la transposition sifflée de langues parlées sont tout aussi instructifs pour notre compréhension des mécanismes en jeu que la communication parlée *stricto senso*.

d'intelligibilité obtenues dans les deux langues en fonction de la taille de la fenêtre d'inversion (Meunier *et al.*, 2002). Outre la remarquable similitude de forme observée pour les deux langues, on note un décalage temporel assez important entre les courbes. Ainsi, le taux de 80 % d'intelligibilité est atteint pour une taille de fenêtre de 40 ms en anglais et 55 ms en français. Cet écart s'accentue lorsque la dégradation augmente et le taux de 50 % est atteint pour des fenêtres respectives de 66 et 90 ms. Ainsi, l'intelligibilité du français semble préservée de manière plus robuste à des inversions portant sur des fenêtres plus longues qu'en anglais. Si parallèlement on prend en compte les débits syllabiques moyens calculés dans MULTEXT (cf. Figure 26), on obtient à titre indicatif une durée moyenne de syllabe de 161 ms en anglais et 141 ms en français. Si l'on calcule alors les tailles de fenêtre listées ci-dessus en proportion de la longueur de syllabe moyenne dans chaque langue (et non plus en durées physiques), on constate que si l'on inverse jusqu'à 40 % de la durée moyenne de syllabe en français, l'intelligibilité reste élevée (80 % ou plus) alors que pour l'anglais, une inversion de l'ordre de 25 % seulement de la durée syllabique moyenne suffit à descendre en dessous de ce taux de 80 % d'intelligibilité. Il n'est pas évident d'interpréter ces résultats, mais ils suggèrent que les locuteurs natifs de l'anglais et du français ne portent pas leur attention sur des unités de même durée. Il est par contre impossible à ce stade de déterminer s'il s'agit de différences de primitives phonologiques, de représentations au sein du lexique mental ou plus simplement de différences de durée acoustique des segments phonétiques eux-mêmes.

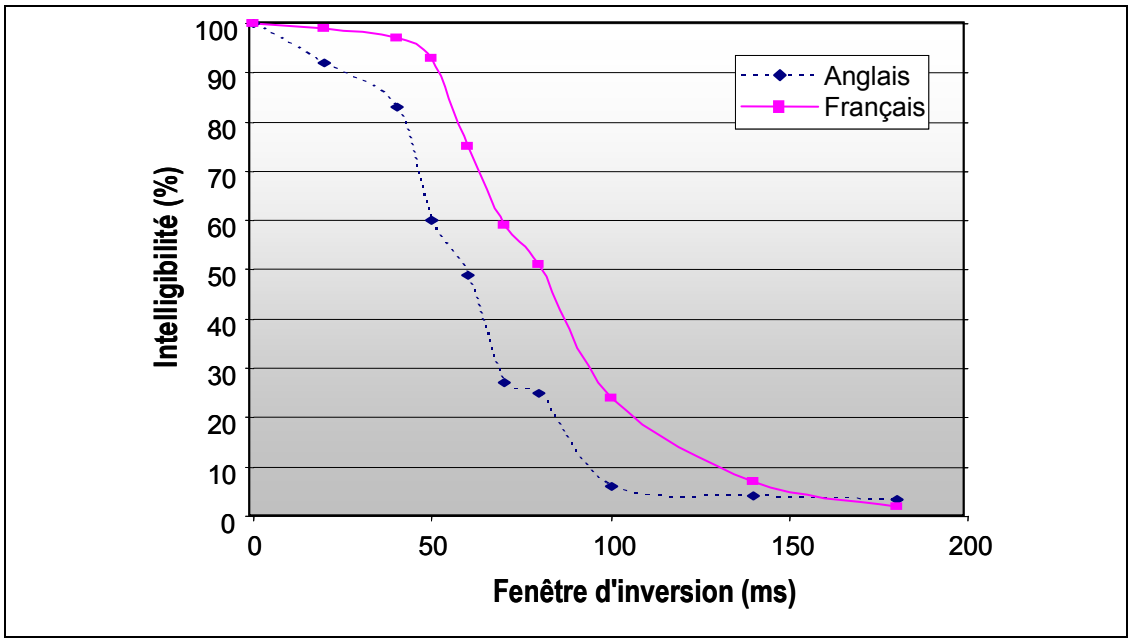


Figure 29 – Intelligibilité préservée en fonction de la taille de la fenêtre d'inversion temporelle en anglais (courbe pointillée, losanges bleus) et en français (courbe pleine, carrés magenta). D'après Meunier *et al.*, (2002).

Depuis, nous avons mené d'autres expériences dans la lignée de celle évoquée ci-dessus, mais la dimension translinguistique n'a plus été intégrée du fait de la complexité des protocoles mis en jeu et du nombre de conditions expérimentales

testées. Il me semble cependant que le champ de l'écoute de parole inversée mériterait d'être exploré dans cette direction.

Pour conclure, je souhaiterai évoquer brièvement l'approche globale dans laquelle se situent nos recherches, et plus particulièrement celles portant sur la complexité et l'information phonologique, développés avec Egidio Marsico et Christophe Coupé. La plupart des travaux dont nous revendiquons l'héritage se situent dans une perspective hypothético-déductive (initiée par Björn Lindblom et René Carré), et/ou une perspective de phonétique expérimentale (incarnée par John J. Ohala et Ian Maddieson en particulier), mais l'approche principale que nous avons utilisée ne relève pas clairement de l'une ou l'autre de ces méthodologies. En effet, nous avons employé la base de données UPSID comme médiateur entre ces deux dimensions, avec comme objectif de faire émerger des motifs réguliers (les tendances universelles). Il s'agit donc là d'apporter un éclairage complémentaire, en espérant évidemment qu'il se révèle pertinent pour confirmer ou infirmer des hypothèses intéressantes de la littérature (principe du MUAF, etc.) et éventuellement pour attirer l'attention sur des points d'intérêt potentiellement négligés jusqu'à présent. Ainsi, ce qui peut sembler être un paradoxe représente en fait une approche agnostique de la communication langagière, qui s'alimente cependant de la tradition hypothético-déductive, comme lorsque nous formulons et testons l'hypothèse d'une régulation du débit d'information. Finalement, cette démarche nous semble prendre en compte les apports multiples de nos prédécesseurs de manière constructive, représentant en cela une approche scientifique objective et fructueuse.

5. RÉFÉRENCES

Abler,W. 1989, “On the particulate principle of self-diversifying systems”, *Journal of Social and Biological Structures,* 12, 1-13.

Adami, A., Mihaescu, R., Reynolds, D.A. & Godfrey, J. 2003. “Modeling Prosodic Dynamics for Speaker Recognition”, *proc. IEEE ICASSP*, Hong Kong, China.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. 2001. “Speech comprehension is correlated with temporal response patterns recorded from auditory”, *Proceedings of the National Academy of Sciences,* 96:23, 13367-13372.

Al-Tamimi J. 2007. *Indices Dynamiques et perception des voyelles : Étude Translinguistique en Arabe Dialectal et en Français*, Thèse de doctorat, Université Lumière Lyon2, Lyon, France, 465 p.

André-Obrecht, R. 1988. “A New Statistical Approach for Automatic Speech Segmentation”, *IEEE Trans. on ASSP*, 36:1, 29-40.

Auer, P. 1993. “Is a rhythm-based typology possible? A study of the role of prosody in phonological typology” *KontRI Working Paper 21*, Universität Hamburg: Hamburg.

Aylett, M.P. & Turk A. 2004. “The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships between Redundancy, Prosodic Prominence, and Duration in Spontaneous Speech”, *Language and Speech*, 47:1, 31-56

Baayen, H., Piepenbrock, R., & Rijn, H. V. 1993. *The CELEX lexical database*, Linguistic Data Consortium: Philadelphia, University of Pennsylvania.

Baddeley, A. D. 2000. “The episodic buffer: a new component of working memory?”, *Trends in Cognitive Science,* 4, 417-423.

Baddeley, A. D., & Hitch, G. J. 1974. “Working memory”, in G. H. Bower (Ed.), *The psychology of learning and motivation: advances in research and theory* (Vol. 8), Academic Press: New York, 47-89.

Bagemihl, B. (1991). "Syllable structure in Bella Coola". *Linguistic Inquiry 22*, 589-646.

Barkat M. 2000. *Détermination d’indices acoustiques robustes pour l’identification automatique des parlers arabes*, Thèse de doctorat, Université Lumière Lyon2, Lyon, France, 300 p.

Barkat, M., Ohala, J.J. & Pellegrino, F. 1999. “Prosody as a Distinctive Feature for the Discrimination of Arabic Dialects”, *proc. of Eurospeech’99*, Budapest, Hungary.

Barkat-Defradas, M., Hamdi, R., & Pellegrino, F. 2004. « De la caractérisation linguistique à l’identification automatique des dialectes arabes », *actes du Workshop MIDL*, Paris, 29-30 novembre

Barkat-Defradas, M., Vasilescu I. & Pellegrino, F. 2003. « Stratégies perceptuelles et identification automatique des langues: application au continuum dialectal arabe », *Revue PArole*, n° 25, 1-44.

Bavelier, D., Newport, E. L., Hall, M. L., Supalla, T., & Boutla, M. 2006. “Persistent difference in short-term memory span between sign and speech”, *Psychological Science,* 17:12, 1090-1092.

Ben Hamed, M. 2007. “UNIDIA, a database for diachronic universals”, *18th International Conference on Historical Linguistics*, Montréal, August 2007.

Besson, M. & Schön D. 2001. “Comparison between language and music”, in R. Zatorre & I. Peretz, (eds), *The biological foundations of music*, Annals of The New York Academy of Sciences, Vol. 930.

Best, C.T., McRoberts, G.W. & Sithole, N.M. 1988. “Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants”, *Journal of Experimental Psychology: Human Perception and Performance*, 14:3, 345-360.

Bickel, B. & Nichols, J. 2002. “Autotypologizing databases and their use in fieldwork”, *proc. of Int. LREC Workshop on Resources and Tools in Field Linguistics*. Las Palmas, Spain.

Blevins, J. 2004. *Evolutionary Phonology. The emergence of sound patterns*. Cambridge University Press: Cambridge.

Bond, ZS & Stockmal, V. 2002. “Distinguishing samples of spoken Korean from rhythmic and regional competitors”, *Language Sciences*, 24:2, 175-185.

Boula de Marëuil, P., Corredor-Ardoy, C. and Adda-Decker, M. 1999. “Multi-lingual automatic phoneme clustering “, *proc. of XIVth ICPhS*, San Francisco, USA.

Boula De Marëuil, P. & Vieru-Dimulescu, B. 2006. “The contribution of prosody to the perception of foreign accent”, *Phonetica*, 63:4, 247-267.

Boutla, M., Supalla, T., Newport, E. L., & Bavelier, D. 2004. “Short-term memory span: insights from sign language”, *Nature Neuroscience,* 7:9, 997-1002.

Bradlow, A.R. 2002. “Confluent talker- and listener-oriented forces in clear speech production”, in C. Gussenhoven & N. Warner (eds), *Papers in Laboratory Phonology 7*, Mouton de Gruyter: Berlin.

Bradlow, A. R. & Bent, T. 2008. “Perceptual adaptation to non-native speech”, *Cognition*, 106:2, 707-729.

Brighton, H. 2003. *Simplicity as a driving force in linguistic evolution*, Ph.D. University of Edinburgh, Edinburgh, United Kingdom, 224.

Browman, C.P. & Goldstein, L. 1992. “Articulatory phonology: An overview”, *Phonetica*, 49: 155-180

Brown, B. L., Giles, H., & Thakebar, J. N. 1985. “Speaker evaluation as a function of speech rate, accent and context”, *Language and Communication,* 5:3, 207-220.

Bybee, J.L. 2006. *Frequency of Use and the Organization of Language*, Oxford University Press: New York, 365 p.

Campbell, W., Gleason, T., Navratil, J., Reynolds, D., Shen, W., Singer, E., & Torres-Carrasquillo P. 2006. “Advanced Language Recognition using Cepstra and Phonotactics: MITLL System Performance on the NIST 2005 Language Recognition Evaluation”, *proc. of workshop Odyssey 2006*, San Juan, Puerto Rico.

Campione, E., & Véronis, J. 1998. “A multilingual prosodic database”, *proc. of 5th ICSLP*, Sydney, Australia.

Carré, R. 2004. “From an acoustic tube to speech production”, *Speech Communication*, 42:2, 227-240.

Carré, R. 2008. “Dynamic properties of an acoustic tube: Prediction of vowel systems”, *Speech Communication*, Corrected Proof, Available online 7 June 2008.

Carré, R. à paraître. “ Signal dynamics in the production and perception of vowels”, in Pellegrino F. *et al.* (eds), *Approaches to Phonological Complexity*.

Carré, R., Bourdeau M. & Tubach J.-P. 1995. “Vowel-Vowel Production: The Distinctive Region Model (DRM) and Vowel Harmony”, *Phonetica*, 52, 205-214.

Carré, R. & Mrayati, M. 1990. “Articulatory-acoustic-phonetic relations and modelling, regions and modes”, in Hardcastle W.J. & Marchal A. (eds), *Speech Production and Speech Modelling*, Kluwer Academic Publishers: Dordrecht.

Carré, R., Pellegrino, F. & Divenyi, P., 2007, "Speech Dynamics: epistemological aspects", *proc. of XVIth ICPhS*, Saarbrücken, Germany.

Cercle linguistique de Prague, 1929, « Thèses présentées au Premier congrès de philologues slaves », *Travaux du Cercle linguistique de Prague 1*, 5-29, consulté sur Internet le 01/05/2008 à l’adresse http://www2.unil.ch/slav/ling/textes/theses29.html

Chang, S., Shastri, L. & Greenberg, S. 2001. “Robust phonetic feature extraction under a wide range of noise backgrounds and signal-to-noise ratios”, *proc. of Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis*, Aalborg, Denmark.

Cherry, C. 1959. *On human communication: a review, a survey, and a criticism*, The MIT Press: Cambridge, 337 p.

Cherry, C., Halle, M. & R. Jakobson. 1953. “Toward the Logical Description of Languages in Their Phonemic Aspect”, *Language*, 29:1, 34-46.

Claussen, J.C. 2004. “Offdiagonal Complexity: A computationally quick complexity measure for graphs and networks”, *q-bio.MN/0410024*

Clements, G.N. 1990. “The role of the sonority cycle in core syllabification”, in Kingston J. & Beckman M.E. (eds), *Papers in Laboratory Phonology I*, Cambridge University Press: Cambridge, 283-333

Clements, G.N. 2003a. “Testing feature economy”, *proc. of XVth ICPhS*, Barcelona, Spain.

Clements, G.N. 2003b. “Feature economy as a phonological universal”, *proc. of XVth ICPhS*, Barcelona, Spain.

Clopper, C. G. & Bradlow, A. R. In press. “Perception of dialect variation in noise: Intelligibility and classification”, *Language and Speech*.

Comrie B. 1992. “Before Complexity”, in Hawkins J.A. & Gell-Mann M. (eds), *The evolution of human languages*, SFI Studies in the Sciences of Complexity, Vol. X, Addison-Wesley: Redwood City, 193-211.

Content, A., Dumay, N., & Frauenfelder, U.H. 2000. “The role of syllable structure in lexical segmentation in French”, *proc. of the Workshop on Spoken Word Access Processes*, Nijmegen, The Netherlands.

Coupé, C. 2003. *De l’origine du langage à l’origine des langues : Modélisations de l’émergence et de l’évolution des systèmes linguistiques*, Thèse de doctorat, Université Lumière Lyon2, Lyon, France, 390 p.

Dahl, Ö. 2004. *The growth and maintenance of linguistic complexity*, John Benjamins: Amsterdam.

Dellwo, V. 2006. “Rhythm and Speech Rate: A Variation Coefficient for deltaC”, in Karnowski,P. & Szigeti,I. (eds), *Language and language-processing*, Peter Lang: Frankfurt am Main, 231-241.

Dellwo, V., Ferragne, E. & Pellegrino, F. 2006. “The perception of intended speech rate in English, French, and German by French speakers”, *proc. of Speech Prosody '06*, Dresden, Germany.

Dellwo, V., Steiner, I., Aschenberner, B., Dancovicová, J., & Wagner, P. 2004. “BonnTempo-Corpus and BonnTempo-Tools: A database for the study of speech rhythm and rate”, *proc. of the 8th ICSLP*, Jeju, South Korea.

Dellwo, V. & Wagner, P. 2003. “Relations between language rhythm and speech rate”, *proc. of XVth ICPhS*, Barcelona, Spain.

Demolin, D. 2002. “The search for primitives in phonology and the explanation of sound patterns: The contribution of fieldwork studies”, in C. Gussenhoven & N. Warner (eds), *Papers in Laboratory Phonology 7*, Mouton de Gruyter: Berlin.

Demolin, D. 2007. “Phonological universals and the control and regulation of speech production”, in M.-J. Sole, P.S. Beddor & M. Ohala (eds), *Experimental Approaches to Phonology*, Oxford University Press: New York

Demolin, D. 2008. “The frame/content theory and the emergence of consonants”, in B.L. Davis & K. Zadjo (eds), *The syllable in speech production*, Lawrence Erlbaum Associates: New York.

Demolin, D. & Socquet, A. 2001. “The role of self-organisation in the emergence of phonological systems”, *Evolution of Communication*, 3:1.

Dominey, P. F., & Ramus, F. 2000. “Neural Network Processing of Natural Language: I. Sensitivity to Serial, Temporal and Abstract Structure in the Infant”, *Language and Cognitive Processes*, *15:*1, 87-127.

Duarte, D, Galves, A., Lopes N. & Maronna, R. 2001. “The statistical analysis of acoustic correlates of speech rhythm”, proc. of W*orkshop on Rhythmic patterns, parameter setting and language change*, ZiF, Bielefeld University, Germany.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. 1999. “Epenthetic vowels in Japanese: A perceptual illusion?”, *Journal of Experimental Psychology: Human Perception and Performance*, 25:6, 1568-1578.

Fagyal, Z., Nguyen, N., Boula De Marëuil, P. 2002. “From dilation to coarticulation : is there vowel harmony in French ?”, *Studies in the Linguistic Sciences*, 32:1, 1-21.

Fakotakis, N., Georgila, K. & Tsopanoglou, A. 1997. “A Continuous HMM Text-Independent Speaker Recognition System Based on Vowel Spotting”, *proc. of Eurospeech’97*, Rhodes, Greece.

Farinas, J. 2002. *Une modélisation automatique du rythme pour l'identification des langues*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France, 130 p.

Farinas, J., Pellegrino, F., Rouas, J.L. & André-Obrecht, R. 2002. “Merging Segmental And Rhythmic Features For Automatic Language Identification”, *proc. of IEEE ICASSP 2002*, Orlando, USA.

Farinas, J., Rouas, J.L., Pellegrino, F. & André-Obrecht, R. 2005. « Extraction automatique de paramètres prosodiques pour l’identification automatique des langues », *Traitement du Signal*, 22:2.

Fenk-Oczlon G. & Fenk A. 1999, “Cognition, quantitative linguistics, and systemic typology”, *Linguistic Typology*, 3:2, 151-177.

Fenk-Oczlon G. & Fenk A. 2005, “Crosslinguistic correlations between size of syllables, number of cases, and adposition order”, in G. Fenk-Oczlon & Ch. Winkler (eds), *Sprache und Natürlichkeit, Gedenkband für Willi Mayerthaler*, Tübingen.

Ferragne, E. 2006. « Diphtongaison et identification automatique dans les dialectes de l'anglais britannique », *actes du 13e colloque d'anglais oral*, Villetaneuse, France.

Ferragne E. 2008. *Étude phonétique des dialectes modernes de l’anglais des îles britanniques : vers l’identification automatique du dialecte*, Thèse de doctorat, Université Lumière Lyon2, Lyon, France, 408 p.

Ferragne, E. & Pellegrino, F. 2004a. “Diphthongization as a cue for the automatic identification of British English dialects”, *proc. of 148th meeting of the Acoustical Society of America*, San Diego, USA.

Ferragne, E. & Pellegrino, F. 2004b. "Rhythm in read British English: interdialect variability", *proc. Interspeech ICSLP 2004*, Jeju Island, Korea.

Ferragne, E. & Pellegrino, F. 2008. « Le rythme dans les dialectes de l'anglais : une affaire d'intensité ? », *actes des XXVIIèmes Journées d’études sur la Parole*, Avignon, France.

Fox, A. 2000. *Prosodic features and Prosodic Structure*, Oxford University Press: Oxford.

Fujisaki, H. 2003. “Prosody, Information and Modeling - with Emphasis on Tonal Features of Speech”, *proc. of ISCA Workshop on Spoken Language Processing*, Mumbai, India.

Galves, A., Garcia J., Duarte D. & Galves C. 2002. “ Sonority as a Basis for Rhythmic Class Discrimination ”, *proc. of 1st Speech Prosody conference*, Aix-en-Provence, France.

Gauvain, J.-L., Messaoudi A. & Schwenk H. 2004. “Language recognition using phone lattices”, *proc. of ICSLP’04*, Jeju, Korea.

Godfrey, J.J., Holliman, E.C., & McDaniel, J. 1992. “SWITCHBOARD: Telephone speech corpus for research and development”, *proc. IEEE ICASSP’92*, San Francisco, USA.

Goldsmith, J.A. 2000. “On information theory, entropy, and phonology in the 20th century”, *Folia Linguistica*, 34:1-2, 85-100.

Grabe, E. & Low, E.L. 2002. “Durational Variability in Speech and the Rhythm Class Hypothesis”, in C. Gussenhoven & N. Warner (eds), *Papers in Laboratory Phonology 7*, Mouton de Gruyter: Berlin.

Greenberg J.H. 1969. “Language Universals: A Research Frontier”, *Science*, 166, 24 October 1969, 473-478

Greenberg, J.H. 1978. “Diachrony, synchrony, and language universals”, in J. H. Greenberg, C. A. Ferguson & E. A. Moravcsik (eds.), *Universals of human language*, vol. 1, Stanford University Press: Stanford, 61-93.

Greenberg, S. 1999. “Speaking in a shorthand – A syllable-centric perspective for understanding pronunciation variation”, *Speech Communication*, 29, 159-176.

Greenberg, S. & Arai, T. 2001. “The relation between speech intelligibility and the complex modulation spectrum”, *procs. of the 7th Eurospeech Conference*, Aalborg, 473-476.

Greenberg, S. & Fosler-Lussier, E. 2000. “the uninvited guest: Information’s role in guiding the production of spontaneous speech”, *proc. of the CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modeling*, Kloster Seeon, Germany.

Greenberg, S., Carvey, H.M. & Hitchcock, L. 2002. “The relation of stress accent to pronunciation variation in spontaneous American English discourse”, *proc. of 1st Speech Prosody conference*, Aix-en-Provence, France.

Greenberg, S. & Chang, S. 2001 “Phonetic Dissection of the Switchboard Corpus Automatic Speech Recognition Systems”, *proc. of NIST Workshop on Large Vocabulary Continuous Speech Recognition (LVCSR)*, Linthicum Heights, USA.

Hagège, C. & Haudricourt, A. 1978. *La phonologie panchronique*. PUF: Paris.

Hämäläinen, A., Boves, L., de Veth, J. & ten Bosch L. 2007. “On the Utility of Syllable-Based Acoustic Models for Pronunciation Variation Modelling”, *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007.

Hamdi, R., 2007, *La variation rythmique dans les dialectes arabes*, Thèse de doctorat, Université Lumière Lyon2, Lyon, France, 325 p.

Hamdi, R., Barkat, M. & Pellegrino, F. 2004. « De la caractérisation linguistique à l’identification automatique des dialectes arabes », *actes du Workshop MIDL*, Paris, France.

Harnad, S. 1999. “The Symbol Grounding Problem”, *Arxiv preprint cs.AI/9906002*.

Harris, J. 2005. “Vowel reduction as information loss”, in Carr,P., Durand J. & Ewen, C. J. (eds), *Headhood, elements, specification and contrastivity*, Benjamins:Amsterdam, 119-132.

Hawkins, J.A. 2004. *Efficiency and Complexity in Grammars*, Oxford University Press:Oxford

Hay, J. Pierrehumbert, J. & Beckman M. 2001. “Speech perception, well-formedness, and the statistics of the lexicon”, in J. Local, R. Ogden & R. Temple (eds), *Papers in laboratory phonology VI*, Cambridge University Press: Cambridge.

Hockett C.F., 1953. “*Review: The Mathematical Theory of Communication* by Claude L. Shannon; Warren Weaver”, *Language*, 29:1, 69-93.

Hockett, C.F. 1955. *A manual of phonology*, Waverly Press: Baltimore.

Hockett, C.F. 1958. *A Course in Modern Linguistics*, The MacMillan Company: New York.

Hockett, C.F. 1966. “The problem of universals in language”, in J.H. Greenberg (ed), *Universals of Language*, 2nd Edition, The MIT Press: Cambridge.

Hoen, M., Meunier, F., Grataloup, C., Pellegrino, F., Grimault, N., Perrin, F., Perrot, X. & Collet, L. 2007. “Phonetic and lexical interferences in informational masking during speech-in-speech comprehension”, *Speech Communication*, 49:12, 905-916.

Hombert, J.-M. & Maddieson, I. 1998. “A Linguistic Approach to Automatic Language Recognition”, *UCLA Working Papers in Phonetics*, 96, 119-125.

Hombert, J.-M. & Maddieson, I. 1999. “The Use of 'Rare' Segments for Language Identification”, *proc. of Eurospeech’99*, Budapest, Hungary.

Howitt, A.W. 2000. “Vowel Landmark Detection”, *proc of 6th ICSLP*, Beijing, China.

Hualde, J.I. & Chitoran, I. 2003. “Explaining the distribution of hiatus in Spanish and Romanian”, *proc. of XVth ICPhS*, Barcelona, Spain.

Huckvale, M. 2004. “ACCDIST: a Metric for Comparing Speakers' Accents”, *proc. Interspeech ICSLP 2004*, Jeju Island, Korea.

Hume E. 2006. “Language Specific and Universal Markedness: An Information-theoretic Approach”, *Linguistic Society of America Annual Meeting. Colloquium on Information Theory & Phonology*, Albuquerque, NM.

Hunt, M., Bamberg, P., Tucker, J. & Anderson, S. 1999. “A Military Operational Automatic Interpreting System”, *proc. ESCA-NATO Workshop Multilingual Interoperability in Speech Technology*, Leusden.

Hyman L. 1983. “Are there syllables in Gokana?” in J. Kaye et al. (eds), *Current approaches to African linguistics*, vol. 2, Foris: Dordrecht, 171–179.

Jakobson R. 1941/1968. *Child Language Aphasia and Phonological Universals*, Mouton: La Hague, 101 p.

Jakobson, R. 1973. *Main trends in the Science of Language*, Main trends in the social Sciences Series. Harper & Row: New-York, 76 p.

Jakobson, R. & Halle, M. 1956/2002. *Fundamentals of Language*, Mouton de Gruyter:Berlin, 96 p.

Jensen, P. 2006. “Network-based predictions of retail store commercial categories and optimal locations”, *Phys. Rev. E*, 74, 035101.

Jong, K. de . 2002. “Attention modulation and the formal properties of stress systems”, *Chicago Linguistic Society* 36, 71-91.

Joos, M. 1936. “Review: The Psycho-Biology of Language by George K. Zipf”, *Language*, 12:3, 196-210.

Kamiyama, T. 2004. “Tokyo and Osaka Japanese: is it possible to distinguish them by prosody alone?”, *Proc. of Workshop MIDL*, Paris, France.

Karlgren, H., 1961, “Speech Rate and Information Theory”, *proc. of 4th ICPhS*, 671-677

Kaye, J., Lowenstamm, J., & Vergnaud, J.-R. 1990. “Constituent Structure and Government Phonology”, *Phonology*, 7, 193-231.

Kern, S. & Davis, B. à paraître. “Emergent complexity in early vocal acquisition: Cross-linguistic comparisons of canonical babbling”, in Pellegrino, F. *et al.* (eds), *Approaches to phonological complexity*.

King, R.D. 1967. “Functional Load and Sound Change”, *Language*, Vol. 43, No. 4 (Dec., 1967), 831-852.

Kirby, S. 2007. “The evolution of language”, in Dunbar R. & Barrett L. (eds), *Oxford Handbook of Evolutionary Psychology*, Oxford University Press:Oxford, 669-681.

Kitazawa, S., Kitamura, T., Mochizuki, K. & Itoh, K. 2002. “Periodicity of Japanese Accent in Continuous Speech”, *proc. of 1st Speech Prosody Conference*, Aix en Provence, 435-438.

Komatsu, M. 2007. “Reviewing Human Language Identification”, in Müller, C. & Schötz, S. (eds), *Speaker Classification II*/2, Lecture Notes in Computer Science, Springer-Verlag: Heidelberg, 206-228.

Komatsu, M., Arai, T. & Sugawara, T. 2004. “Perceptual discrimination of prosodic types”, *proc. of 2nd Speech Prosody conference*, Nara, Japan.

Kruskal, J.B. & Wish, M. 1978. *Multidimensional Scaling*, Sage University Paper series on Quantitative Application in the Social Sciences, 07-011, Sage Publications: Beverly Hills.

Labov, W. 2001. *Principles of linguistic change : social factors* (Vol. 2), Blackwell: Oxford.

Laver, J. 1994. *Principles of Phonetics*. Cambridge University Press: Cambridge.

Leonard, R. G. 1980. *Language Recognition Test and Evaluation*, Technical Report RADC-TR-80-83, RADC/Texas Instruments Inc:Dallas.

Lindblom, B. 1963. *On vowel reduction*, Report #29, The Royal Institute of Technology, Speech Transmission Laboratory: Stockholm.

Lindblom, B. 1986. “Phonetic Universals in Vowel Systems”, in Ohaha J.J. & Jaeger J.J. (eds), *Experimental Phonology*, Academic Press: Orlando.

Lindblom, B. 1998. “Systemic constraints and adaptive change in the formation of sound structure”, in Hurford J.R., Studdert-Kennedy M. & Knight C. (eds), *Approaches to the evolution of language*, Cambridge University Press: Cambridge.

Lindblom, B. 1999. “Emergent phonology”, *proc. of the Berkeley Linguistic Society*, 25.

Lindblom, B. 1990a. “Explaining phonetic variation: a sketch of the H&H theory”, in Hardcastle W.J. & Marchal A. (eds), *Speech Production and Speech Modelling,* Kluwer Academic Publishers: Dordrecht.

Lindblom, B. 1990b. “Models of phonetic variation and selection”, *PERILUS*, University of Stockholm, XI, 65-100

Lindblom, B. 2005. “Deducing language from non-language”, *Workshop on Phonological Systems and Complex Adaptive Systems*, Lyon, France, July 2005.

Lindblom, B., Diehl, R., Park, S.-H. & Salvi G. soumis. “Sound systems are shaped by their users: the recombination of phonetic substance”, submitted to Clements G.N. & Ridouane R. (eds), *Where do features come from? The nature and sources of phonological primitives*.

Lindblom, B. & Engstrand, O. 1988. “In what sense is speech quantal? A commentary on Stevens K.N. (1989)”, *Journal of Phonetics*, 17.

Lindblom, B. & Maddieson, I. 1988. “Phonetic universals in consonant Systems”, in Hyman L.M. & Li C.N. (eds), *Language, Speech, and Mind*, Routledge: New York.

Lindblom, B., Mauk, C. and Moon, S.J. 2006. “Dynamic Specification in the Production of Speech and Sign”, in P.L. Divenyi, S. Greenberg & G. Meyer (eds), *Dynamics of Speech Production and Perception*, NATO Science Series: Life and Behavioural Sciences, Vol. 374, Ios Press: Amsterdam.

Lindblom, B. & Sundberg, J. 1970. “Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement”, *JASA*, 50, 1166-1179.

Linde, Y., Buzo, A. & Gray, R. 1980. “An Algorithm for Vector Quantizer Design”, *IEEE Trans. on Communication*, 28:1, 84-95.

Lippmann, R.P. 1997. “Speech recognition by machines and humans”, *Speech Communication*, 22:1, 1-15.

Locke, J.L. 2008. “Cost and Complexity: Selection for speech and language”, *Journal of Theoretical Biology*, 251:4, 640-652.

Low, E-L., Grabe, E. & Nolan F. 2000. “Quantitative characterizations of speech rhythm: syllable-timing in Singapore English”, *Language & Speech*, 43, 377-401.

Macmillan, N.A. & Creelman, C.D. 2005. *Detection Theory: A User's Guide*, Lawrence Erlbaum Assoc Inc: Mahwah.

MacNeilage, P.F. 1998. “The frame/content theory of evolution of speech production”, *Brain and Behavioral Sciences*, 21, 499-546.

MacNeilage P.F. & Davis B.L. 2000. “On the Origin of Internal Structure of Word Forms”, *Science,* 288:5465, 527 – 531.

MacNeilage P.F., Davis B.L. & Matyear C.L. 2002. “Acquisition of Serial Complexity in Speech Production: A Comparison of Phonetic and Phonological Approaches to First Word Production”, *Phonetica*, 59, 75-107.

Maddieson, I. 1984. *Patterns of Sounds*, Cambridge University Press:Cambridge.

Maddieson, I. 1986. “Borrowed Sounds”, in J.A. Fishman *et al.* (eds), *The Fergusonian Impact: In Honor of Charles A. Ferguson on the Occasion of His 65th Birthday (Contributions to the Sociology of Language)*, Walter de Gruyter: Berlin

Maddieson, I. 1992. “The structure of segment sequences”, *UCLA Working Papers in Phonetics*, 83.

Maddieson, I. 2006. “Correlating phonological complexity: data and validation”, *Linguistic Typology*, 10:1, 106-123.

Maddieson, I. 2007. “Issues of phonological complexity: Statistical analysis of the relationship between syllable structures, segment inventories and tone contrasts”, in M.-J. Sole, P.S. Beddor & M. Ohala (eds), *Experimental Approaches to Phonology*, Oxford University Press: New York, 93-103.

Maddieson, I. à paraître. “Calculating phonological complexity”, in Pellegrino F. *et al.* (eds), *Approaches to Phonological Complexity*.

Maddieson, I. & Precoda, K. 1990. “Updating UPSID”, *UCLA Working Papers in Phonetics*, 74, 104-111.

Maidment, J.A. 1983. “Language recognition and prosody: further evidence”, in *Speech, hearing and language: Work in progress*, University College London, 1, 133-141.

Martin, A.F. & Przybocki, M.A. 2003. “NIST 2003 Language Recognition Evaluation”, *proc. of Interspeech 2003*, Geneva, Swizerland.

Martin, A.F., & Le, A.N. 2006. “The Current State of Language Recognition: NIST 2005 Evaluation Results”, *proc. of workshop Odyssey 2006*, San Juan, Puerto Rico.

Martinet, A. 1933. « Remarques sur le système phonologique du français », *Bulletin de la Société de ling3uistique de Paris*, 33, 191-202.

Martinet, A. 1955. *Économie des changements phonétiques. Traité de phonologie diachronique*, Francke: Berne, 396 p.

Martinet, A. 1960/1973. *Éléments de linguistique générale*. Armand Colin: Paris, 223 p.

Martinet, A. 1962/1969. *Langue et fonction. Une théorie fonctionnelle du langage*, (1962 : 1ère édition en anglais ; 1969 : 1ère édition en français). Denoël: Paris, 199 p.

Massaro, D.W. 2001. “Speech Perception” in N.M. Smelser & P.B. Baltes (Eds) & W. Kintsch (Section Ed.), *International Encyclopedia of social and Behavioral Sciences*, Elsevier:Amsterdam, pp. 14870-14875.

Matějka, P., Burget, L., Glembek, O., Schwarz, P., Hubeika, V., Fapšo, M., Mikolov, T. & Plchot, O. 2007. “BUT system description for NIST LRE 2007”, *proc. NIST Language Recognition Evaluation Workshop*, Orlando, USA.

Meyer, J., 2007, "Acoustic Strategy and Typology of Whistled Languages; Phonetic Comparison and Perceptual Cues of Whistled Vowels", *Journal of the International Phonetic Association*, 37.

Meyer, J., Pellegrino, F., Barkat, M. & Meunier, F. 2003. “The notion of perceptual distance: the case of Afroasiatic languages”, *proc. of XVth ICPhS*, Barcelona, Spain.

Meunier, F., Cenier, T., Barkat, M. & Magrin-Chagnolleau, I. 2002. « Mesure d’intelligibilité de segments de parole à l’envers en français », actes des XXIVèmes JEP, Nancy, 117-120.

Moon, S.-J. & Lindblom, B. 2003. “Two experiments on oxygen consumption during speech production: vocal effort and speaking tempo”, *proc. of XVth ICPhS*, Barcelona, Spain.

Mueller, S. T., Seymour, T. L., Kieras, D. E., & Meyer, D. E. 2003. ”Theoretical implications of articulatory duration, phonological similarity and phonological complexity in verbal working memory”, *Journal of Experimental Psychology, Learning, Memory, and Cognition,* 29:6, 1353-1380.

Muthusamy, Y.K., Barnard, E., Cole, R.A. 1994. “Automatic Language Identification: A Review/Tutorial”, *IEEE Signal Processing Magazine*, Vol. 11, N°4, p.33-41.

Muthusamy, Y.K., Jain, N., Cole, R.A. 1994. “Perceptual benchmarks for automatic language identification”, *proc. of IEEE ICASSP*, Adelaide, Australia, 333-336.

Nam, H. & Saltzman E. 2003. “A competitive, coupled oscillator of syllable structure”, *proc. of XVth ICPhS*, Barcelona, Spain.

Nazzi, T., Bertoncini, J. & Mehler, J. 1998. “Language discrimination by newborns: towards an understanding of the role of rhythm”, *Journal of Experimental Psychology: Human Perception and Performance*, 24:3, 756-766.

Nettle, D. 1995. “Segmental inventory size, word length, and communicative efficiency”. *Linguistics,* 33, 359-367.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. 2004. “Lexique 2 : a new French lexical database”, *Behavior Research Methods, Instruments, & Computers,* 36:3, 516-524.

Nguyen, N., & Fagyal, Z. 2003. “Acoustic aspects of vowel harmony in French”, *proc. of XVth ICPhS*, Barcelona, Spain.

Nosofsky, R.M. 1992. “Similarity Scaling and Cognitive Process Models”, *Annual Reviews in Psychology*, 43:1, 25-53.

Ohala, J.J. 1980. “Moderator’s summary of symposium on ‘Phonetic universals in phonological systems and their explanation”, *proc. of 9th ICPhS*, Institute of Phonetics:Copenhagen

Ohala, J.J. 1990 “The phonetics and phonology of aspects of assimilation”, in J. Kingston & M. Beckman (eds), *Papers in Laboratory Phonology I: Between the grammar and the physics of speech*, Cambridge University Press: Cambridge, 258-265.

Ohala, J.J. 1995. “Speech perception is hearing sounds, not tongues”, JASA, 99, 1718-25.

Ohala, J.J. 2007 “An interpretive history of phonological science”, conférence donnée lors des *Premières Journées des Sciences de la Parole*, 30-31 mars 2003, Charleroi, Belgique.

Ohala, J.J. 2008. “The Emergent Syllable”, in B.L. Davis & K. Zadjo (eds), *The syllable in speech production*, Lawrence Erlbaum Associates: New York.

Ohala, J.J. à paraître. “Languages’ sound inventories: the devil in the details”, in Pellegrino F. *et al*. (eds), *Approaches to Phonological Complexity*.

Ohala, J.J. & Gilbert, J.B. 1981. “Listeners' ability to identify languages by their prosody”, in P. Leon & M. Rossi (eds.), *Problèmes de prosodie, Vol. II: Expérimentations, modèles et fonctions*, Studia Phonetica 18, Didier: Ottawa, 123-131.

Ohala, J.J. & Kawasaki-Fukumori, H. 1997. “Alternatives to the sonority hierarchy for explaining segmental sequential constraints”, in S. Eliasson & E. H. Jahr (eds.), *Language And Its Ecology: Essays In Memory Of Einar Haugen*, Trends in Linguistics. Studies and Monographs, Vol. 100. Mouton de Gruyter: Berlin, 343-365.

Oudeyer, P.-Y. 2006 *Self-Organization in the Evolution of Speech*, Studies in the Evolution of Language, Oxford University Press: Oxford.

Pellegrino, F. 1998. *Une approche phonétique en identification des langues : la modélisation acoustique des systèmes vocaliques*, Thèse de doctorat, Université Paul Sabatier, Toulouse, France, 224 p.

Pellegrino, F. 2007. « Théorie de l'information et corpus multilingue : (encore) une approche de la complexité phonético-phonologique », Séminaire du laboratoire Parole et Langage, Aix en Provence, février 2007.

Pellegrino, F. 2008. “Rhythm”, in P.C. Hogan (ed), *The Cambridge Encyclopedia of the Language Sciences*, Cambridge University Press: Cambridge.

Pellegrino, F. & André-Obrecht, R. 2000. “Automatic Language Identification: An Alternative Approach to Phonetic Modeling”, *Signal Processing*, 80:7, 1231–1244.

Pellegrino, F. Coupé, C. & Marsico, E. 2007. “An information theory-based approach to the balance of complexity between phonetics, phonology and morphosyntax”, *Annual Meeting of the Linguistic Society of America*, Anaheim, CA, USA, January 2007.

Pellegrino, F., Farinas, J. & André-Obrecht, R. 1999. “Vowel System Modeling: A Complement to Phonetic Modeling in Language Identification”, *proc. of Workshop MIST: Multilingual Interoperability in Speech Technology*, Leusden, The Netherlands.

Pellegrino, F. Marsico E., Chitoran, I. & Coupé C. (eds). à paraître. *Approaches to Phonological Complexity*, Mouton de Gruyter: Berlin, 380 p.

Peng, G. 2005. “Temporal and tonal aspects of Chinese syllables: a corpus-based comparative study of Mandarin and Cantonese”, *Journal of Chinese Linguistics,* 34:1, 134-154.

Peperkamp, S., E. Dupoux & N. Sebastián-Gallés. 1999. “Perception of stress by French, Spanish, and bilingual subjects”, *Proc. of EuroSpeech ’99,* Budapest, Hungary.

Petitot-Cocorda, J. 1985. *Morphogenèse du sens - Volume 1. Pour un schématisme de la structure*, PUF:Paris.

Pfau, T. & Ruske G. 1998. “Estimating the speaking rate by vowel detection”, *proc. of IEEE ICASSP’98*, Seattle, USA.

Pfitzinger, H.R., Burger, S. & Heid, S. 1996. “Syllable detection in read and spontaneous speech”, *proc. of ICSLP’96*, Philadelphia, USA.

Plank, F. 1998. “The co-variation of phonology with morphology and syntax: a hopeful history”, *Linguistic Typology, 2*:2, 195-230.

Polka, L. & Bohn, O.-S. 2003. “Asymmetries in vowel perception”, *Speech Communication*, 41, 221–231.

Pone, M. 2005. *Studio e realizzazione di una tastiera software pseudo-sillabica per utendi disabili*. PhD Dissertation, Università degli Studi di Genova, Genova.

Proctor, M.I. 2007. “The organization of phonological inventories an articulatory approach”, *proc. of XVIth ICPhS*, Saarbrücken, Germany.

Ramus, F. 1999. *Rythme des langues et acquisition du langage*. Thèse de doctorat, EHESS, Paris, France.

Ramus, F. 2002. “Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues”, *Annual Review of Language Acquisition, 2*, 85-115.

Ramus, F., Hauser, M. D., Miller, C., Morris, D., & Mehler, J. 2000. “Language discrimination by human newborns and by cotton-top tamarin monkeys”, *Science, 288*, 349-351.

Ramus, F. & Mehler, J. 1999. “Language identification with suprasegmental cues: A study based on speech resynthesis”, *JASA, 105*:1, 512-521.

Ramus, F., Nespor, M. & Mehler, J. 1999. „Correlates of linguistic rhythm in the speech signal”, *Cognition, 73:*3, 265-292.

Remijsen, B. & Gilley, L. 2008. “Why are three-level vowel length systems rare? Insights from Dinka (Luanyjang dialect)”, *Journal of Phonetics* 36:2, 318-344.

Ridouane, R. 2002. “Words without vowels: Phonetic and phonological evidence from Tashlihyt Berber”. *ZAS Papers in Linguistics*, *28*, 93 - 110.

Rissanen, J. 1983. “A universal prior for integers and estimation by minimum description length”, *Annals of Statistics*, 11:2, 416-431.

Roach, P. 1999. “Some languages are spoken more quickly than others”, in L. Bauer & P. Trudgill (Eds.), *Language myths*, Penguin: London, 150-158.

Rouas, J.-L. 2005. *Caractérisation et identification automatique des langues*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France, 253 p.

Rouas, J.-L. 2007. “Automatic prosodic variations modelling for language and dialect discrimination”, *IEEE Trans. on ASLP*, 15:6, 1904-1911.

Rouas, J.-L., Barkat-Defradas, M., Pellegrino, F. & Hamdi-Sultan, R. 2006. « Identification automatique des parlers arabes par la prosodie ». *actes des XXVIèmes Journées d'Etude sur la Parole*, Dinard, France.

Rouas, J.L., Farinas, J., Pellegrino, F. & André-Obrecht, R. 2005. “Rhythmic Unit Extraction and Modelling for Automatic Language Identification”, *Speech Communication*, 47:4, pp. 436-456.

Rousset I. 2004. *Structures syllabiques et lexicales des langues du monde*, Thèse de doctorat, Université Grenoble III, Grenoble, France, 204 p.

Sampson, G., Gil, D. & Trudgill, P, (eds). à paraître. *Language Complexity as an Evolving Variable*, Studies in the Evolution of Language 13, Oxford University Press:Oxford. To appear in February 2009.

Saussure, F. de. 1916/1996. *Cours de linguistique générale*, Payot:Paris (1916 : 1ère édition ; édition de 1996), 520 p.

Schwartz, J.-L., Abry C., Boë, L.-J., Ménard, L. & Vallée N. 2005. “Asymmetries in vowel perception, in the context of the Dispersion-Focalisation Theory”, *Speech Communication*, 45:4, 425-434.

Schwartz, J.-L., Boë, L.-J., Vallée, N. & Abry, C. 1997. “The dispersion-focalization theory of vowel systems”, *Journal of Phonetics*, 25, 255-286.

Schweickert, R., & Boruff, B. 1986. “Short-term memory capacity: magic number or magic spell”, *Journal of Experimental Psychology, Learning, Memory, and Cognition,* 12, 419-425.

Service, E. 1998. “The effect of word-length on immediate recall depends on phonological complexity, not articulation duration”, *Quaterly Journal of Experimental Psychology, 51 Section A*, 283-304.

Shannon, C.E. 1956. “The bandwagon”, *IRE Transactions on Information Theory*, 2:1, 3.

Shannon, C.E. & Weaver W. 1949. *The Mathematical Theory of Communication*, University of Illinois Press, Urbana.

Shosted, R. K. 2006. “Correlating complexity: A typological approach”, *Linguistic Typology, 10:*1, 1-40.

Singer, E., Torres-Carrasquillo, P.A., Gleason, T.P., Campbell, W.M. & Reynolds, D.A. 2003. “Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification”, *proc. of Eurospeech’03*, Geneva, Switzerland, 1345-1348.

Steels, L. 2000. “Language as a complex adaptive system”, *Lecture Notes in Computer Science*, vol. 1917, Springer:Berlin, 17-26

Stevens, K .N . 1989. “On the quantal nature of speech”, *Journal of Phonetics,* 17 , 3-45.

Stevens, K.N. 1972. “The quantal nature of speech: Evidence from articulatory-acoustic data”, in E.E. Davis Jr . & P.B. Denes (eds), *Human communication: a unified v*i*ew*, McGraw-Hill:New-York , 51-66.

Stockmal, V. & Bond, Z.S. 2003. “Same talker, different language: a replication”, in J. Mugane (ed), *Linguistic Typology and Representation of African Languages*, Africa World Press: Lawrenceville.

Stockmal, V., Moates, D.R. & Bond, Z.S. 2000. “Same talker, different language”, *Applied Psycholinguistics*, 21:3, 383-393.

Stockmal, V., Muljani, D. & Bond, Z. 1996. “Perceptual features of unknown foreign languages as revealed by multi-dimensional scaling”, *proc. of 4th ICSLP*, Philadelphia, USA.

Studdert-Kennedy, M. 1998. “The particulate origins of language generativity: from syllable to gesture”, in Hurford, J.R., Studdert-Kennedy, M. & Knight, C. (eds.), *Approaches to the Evolution of Language*, Cambridge University Press: Cambridge, 202-221.

Studdert-Kennedy, M. 2000. “Evolutionary implications of the particulate principle: Imitation and the dissociation of phonetic form from semantic function”, in Knight, C., Studdert-Kennedy, M., and Hurford, J.R. (eds.), *The Evolutionary Emergence of Language*, Cambridge University Press: Cambridge, 161-176.

Studdert-Kennedy, M. & Goldstein, L. 2003. ”Launching language: The gestural origin of discrete infinity”, in Christiansen M.H. & Kirby S. (eds), *Language Evolution: The States of the Art*. Oxford University Press:Oxford.

Surendran, D. & Levow, G.-A. 2004. “The functional load of tone in Mandarin is as high as that of vowels”, *proc. of Speech Prosody 2004*, Nara, Japan, 99-102.

Sussman H.M., Hoemeke K.A. & Ahmed F.S. 1993. “A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation”, *JASA*, 94-3, 1256-1268

Swadesh M. 1934. “The Phonemic Principle Source”, *Language*, 10:2, 117-129.

Tamaoka, K., & Makioka, S. 2004. “Frequency of occurrence for units of phonemes, morae, and syllables appearing in a lexical corpus of a Japanese newspaper”, *Behavior Research Method, Instruments, & Computers, 36*:3, 531-547.

Thymé-Gobbel, A., & Hutchins, S. E. 1999. “Prosodic features in automatic language identification reflect language typology”, *proc. of XIVth ICPhS’99*, San Francisco, USA.

Todd, N. P. & Brown, G. J. 1994. “A computational model of prosody perception”, *proc. of ICSLP’94*, Yokohama, Japan.

Toro, J.M., Trobalón, J.B. & Sebastián-Gallés, N. 2003. “The use of prosodic cues in language discrimination tasks by rats”, *Animal Cognition*, 6, 131-136.

Toro, J.M., Trobalón, J.B., & Sebastián-Galles, N. 2005. “The effects of backward speech and speaker variability in language discrimination by rats”, *Journal of Experimental Psychology:Animal Behavior Processes*, 31, 95-100.

Torres-Carrasquillo, P.A., Reynolds, D.A., & Deller Jr, J.R. 2002. “Language identification using Gaussian mixture model tokenization”, *proc. of IEEE ICASSP’02*, Orlando, USA.

Troubetzkoy, N.S. 1938/1970. *Principes de phonologie*, Klincksieck: Paris, 396 p.

Trudgill, P. 2004. “Linguistic and social typology: The Austronesian migrations and phoneme inventories”, *Linguistic Typology,* 8:3, 305-320.

Twaddell, W.F., 1935. “On Defining the Phoneme”, *Language*, 11:1, pp. 5-62.

UNESCO, 2003. *L’éducation dans un monde multilingue*, Document cadre, UNESCO:Paris.

UNIDIA, 2008, Base de données d’Universaux Diachroniques, consultée le 31/07/2008 à l’adresse : http://www.diadm.ish-lyon.cnrs.fr/unidia

van Son, R.J.J.H. & Pols, L.C.W. 2003. “Information Structure and Efficiency in Speech Production”, *proc. of Eurospeech 2003*, Geneva, Switzerland.

Vallée, N.. 1994. *Systèmes vocaliques : de la typologie aux prédictions*. Thèse de doctorat, Université Stendhal, Grenoble, France, 305p.

Vallée, N., L.-J. Boë, J.-L. Schwartz, P. Badin, & C. Abry. 2002. The Weight of Phonetic Substance in the Structure of Sound Inventories. *ZAS Papers in Linguistics* 28: 145-168.

Vasilescu, I. 2001. *Contribution à l’identification automatique des langues romanes*. Thèse de doctorat, Université Lumière Lyon2, Lyon, France, 246 p.

Wells, J. C. 1982. *Accents of English I*. Cambridge University Press: Cambridge.

Wiener, N. 1948. *Cybernetics, or control and communication in the animal and the machine*, The MIT Press: Cambridge, 598 p.

Wu, S.-L. 1998. *Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition*, Ph.D. Thesis, UC Berkeley, USA (ICSI Technical Report TR-98-014).

Zellner Keller, B. 2002. “Revisiting the Status of Speech Rhythm”, *proc. of the 1st Speech Prosody conference*, Aix-en-Provence, France.

Zellner Keller, B. & Keller, E. 2001. “Representing Speech Rhythm”, in Keller, E., Bailly, G., Monaghan, A., Terken, J. and Huckvale, M. (eds), *Improvements in Speech Synthesis*, John Wiley: Chichester.

Zhao, Y. & Jurafsky, D. 2007. “the effect of lexical frequency on tone production”, *proc. of XVIth ICPhS*, Saarbrücken, Germany.

Zhu, D., & Adda-Decker, M. 2006. “Language identification using lattice-based phonotactic and syllabotactic approaches”, *proc of IEEE Workshop Odyssey*, San Juan, Puerto Rico.

Zipf, G.K. 1929. “Relative Frequency as a Determinant of Phonetic Change”, *Harvard Studies in Classical Philology*, Vol. 40, (1929), 1-95.

Zipf, G.K. 1935/1965. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*, MIT Press:Cambridge, 336 p.

Zipf, G.K. 1937. “Statistical Methods and Dynamic Philology”, *Language*, 13:1, 60-70.

Zipf, G.K. 1949. *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology,* Addison-Wesley Press: Cambridge.

Zissman, M. 1996. “Comparison of four approaches to automatic language identification of telephone speech”, *IEEE Trans. on SAP*, 4:1.