Chapitre 4

Analyse distributionnelle des séquences de consonnes

Ce chapitre a fait l'objet de deux communications orales dont une dans un congrès avec comité de lecture assorti d'une publication dans des actes :

XXII^{èmes} Journées d'Etude sur la Parole, 15-19 Juin 1998, Martigny, Suisse.

Journée 'Langage et Lexique', 6 Novembre 1999, Institut des Sciences Cognitives, Lyon, France.

ANALYSE DISTRIBUTIONNELLE DES SEQUENCES DE CONSONNES

Le chapitre précédent nous a conduit à mettre en évidence deux approches de la syllabation : la première consiste à considérer la syllabe comme une structure hiérarchique dont la localisation des frontières reposerait essentiellement sur des paramètres indépendants des caractéristiques des phonèmes impliqués (OOP, MOP, OT). Cette approche attribue une importance considérable aux fonctions d'attaque, de noyau et de coda syllabiques que peuvent prendre les phonèmes. Le second type de propositions repose sur les caractéristiques propres des phonèmes en jeu (principe de sonorité notamment) ; le concept de similarité invoqué pour rendre compte des contraintes phonotactiques dans la racine verbale de l'arabe pourrait aussi -à notre avis- constituer un modèle de la syllabation en envisageant que les séquences de deux phonèmes tendent à être le moins similaires possibles à l'intérieur d'une syllabe ; deux consonnes partageant des traits communs tendraient alors à être prononcées dans deux syllabes différentes. Parmi les approches de la première catégorie, la plupart ne s'appliquent que si les contraintes phonotactiques de la langue ne sont pas transgressées. Il existe donc un lien étroit entre contraintes phonotactiques et syllabation. Nous ne disposons cependant pas d'une définition précise et communément admise de la légalité phonotactique. Deux catégories de travaux nous ont conduit à opposer des conceptions alternatives de cette notion : certains auteurs adoptent une conception 'binaire' de la légalité, toute séquence de phonèmes étant légale ou pas. Dans le cadre

de cette conception, il est admis qu'une séquence constitue une suite légale si elle est attestée dans la langue en début de syllabe. Nous avons mentionné le problème que posent certains groupes de consonnes cités par Dell (1995) comme *légaux mais déviants* malgré leur existence en début de mot (donc en début de syllabe). Le second type de propositions concernant la légalité phonotactique d'une séquence consiste à l'envisager en termes de *similarité structurée* des segments entre eux. Plus deux segments seraient similaires, moins ils auraient tendance à être juxtaposés à l'intérieur d'une même syllabe en raison de leur plus grande illégalité phonotactique. Or cette similarité des segments semble être inversement corrélée avec leur fréquence d'utilisation dans la langue. C'est notamment le cas dans la morphologie de l'arabe. Si ce lien peut être mis en évidence dans le cadre de la syllabation des phonèmes, on peut s'attendre à ce que les séquences très peu similaires (consonne-voyelle par exemple) soient beaucoup utilisées alors que les séquences les plus similaires (une suite de deux occlusives par exemple) seraient peu utilisées. Cette dissociation entre deux conceptions de la légalité ainsi qu'un lien étroit entre légalité phonotactique et syllabation nous conduit à poser la question d'une définition opérationnelle de la légalité des groupes de consonnes en attaque syllabique, c'est à dire de leur taux de bonne formation (*well-formedness*), en envisageant deux alternatives : on peut considérer d'une part comme légal *tout groupe attesté dans la langue en attaque syllabique*. Selon la démonstration effectuée par Dell (1995), il est possible d'assimiler l'ensemble des groupes de consonnes attestés en début de mot à un sous ensemble des groupes de consonnes qui constituent une attaque de syllabe bien formée. La distribution des groupes de consonnes attestés en début de mot nous permettrait donc d'estimer celle des groupes constituant des attaques syllabiques bien formées. Le modèle proposé par Frisch et al. (soumis) consiste quant à lui à définir la légalité phonotactique de suites de deux phonèmes en termes de similarité. Cette similarité s'exprimerait en outre par des différences de fréquence d'apparition des séquences dans la langue. D'autres travaux (Laks, 1995; Vroomen et al., 1998) montrent par ailleurs qu'une structuration syllabique de séquences linéaires de phonèmes peut être développée par des modèles connexionnistes. Partant de l'observation que ce type de modèle se fonde nécessairement sur des informations probabilistes disponibles au cours de l'apprentissage, il est possible d'affirmer que cette structuration syllabique n'est en mesure d'émerger que grâce à l'utilisation de processus et d'informations purement stochastiques. Si l'on s'affranchit de contraintes morphologiques et accentuelles qui influencent fortement la syllabation, il est alors envisageable de considérer que contraintes syllabiques et phonotactiques s'expriment de façon similaire par la fréquence d'occurrence des suites de segments dans la langue.

Nous avons donc décidé de conduire une analyse distributionnelle des séquences de phonèmes dans la langue et plus particulièrement des groupes de deux consonnes afin d'évaluer la validité de ces interprétations. Si l'on est en mesure de mettre en évidence un lien entre fréquence et syllabation (ou légalité phonotactique), nous serons conduit à mener une réflexion plus approfondie sur la contribution des régularités phonologiques dans les processus de segmentation de la parole (McQueen, 1998 ; Vroomen & de Gelder, 1999). En effet, si la structure phonologique des séquences de consonnes est liée à la fréquence d'occurrence dans la langue, il sera nécessaire de préciser quels processus sont à l'œuvre dans l'émergence des effets observés.

1. Présentation du corpus

Le corpus utilisé pour l'analyse distributionnelle des séquences de consonnes en français est une base de données lexicale informatisée couramment utilisée en psycholinguistique : BRULEX (Content et al., 1990). Nous avons aussi utilisé ce corpus pour les diverses recherches lexicales effectuées lors de la constitution du matériel expérimental de ce travail de thèse. C'est également l'un des corpora analysé par Dell (1995) pour son recensement des groupes de consonnes attestés dans la langue française.

1.1. Caractéristiques de la base de données et méthode d'analyse

BRULEX (Content et al., 1990) est une base de données informatique de mots isolés. Chaque entrée correspond à un mot. Cette base de données recense notamment, pour 35746 mots de la langue française, leur orthographe, leur transcription phonémique et pour la plupart d'entre eux (qui sont au nombre de 26413) leur fréquence d'apparition dans un corpus écrit. Il est ainsi possible, avec cet outil, de faire des recherches lexicales à partir de critères multiples au nombre desquels les structures orthographique et phonémique ou la fréquence des mots mais également de dériver -avec les algorithmes adéquats- d'autres catégories d'informations comme le nombre de compétiteurs des mots, la fréquence des diphones qui les constituent, etc. Nos analyses ont été dans tous les cas effectuées sur les champs de la base de données correspondant à la transcription phonémique des mots.

L'ensemble des analyses présentées ici a été réalisé à l'aide du langage de programmation AWK[31] (Aho, Kernighan, & Weinberger, 1988). Ce langage permet de consulter très simplement des fichiers au format texte et de faire toutes sortes d'opérations sur ces fichiers. La version de BRULEX (Content et al., 1990) que nous avons utilisée est donc une version texte de la base de données originale (qui est fournie dans un format binaire propriétaire).

1.2. Problèmes posés par le choix de ce corpus

Le choix du corpus BRULEX (Content et al., 1990) entraîne un certain nombre de difficultés pour l'interprétation ultérieure des données obtenues. Deux problèmes nous semblent essentiels dans le cadre d'une analyse de la fréquence des groupes de consonnes. Le premier a trait au type de transcription fourni dans la base de données alors que le second est lié au choix d'une base de données lexicale ne permettant pas, par définition, de réaliser un comptage de tous les groupes de consonnes qu'il est possible de rencontrer dans les énoncés linguistiques produits par les locuteurs, c'est à dire dans des phrases.

1.2.1. Transcription des sons

La transcription des séquences de sons de chaque mot qui est fournie dans BRULEX (Content et al., 1990) peut poser des problèmes pour l'interprétation des résultats obtenus. En effet, on peut dire qu'elle ne correspond ni à une transcription phonologique ou phonémique (correspondant à ce que l'on a appelé dans le Chapitre 3 la représentation sous-jacente) ni à une transcription phonétique qui pourrait refléter la manière dont le mot est effectivement produit par un locuteur natif. Par exemple, cette transcription distingue les /o/ ouvert ([ɔ]) et fermé ([o]). Cette opposition n'est pas distinctive en français ; les formes [ɔ] et [o] constituent en effet des allophones du même phonème. Le choix de retranscrire le /o/ de cette manière correspondrait donc à une transcription de sa forme phonétique de surface. De même, la règle de chute du schwa est appliquée pour la transcription. Ainsi, le mot 'paquetage' est retranscrit /paktaʒ/, ce qui implique que le schwa sous-jacent ait été supprimé. Ce paramètre est particulièrement important pour l'analyse des groupes de consonnes attestés dans la langue puisqu'il permet d'observer l'occurrence de groupes de consonnes qui ne seraient pas détectés avec une analyse de la forme orthographique ou phonologique des mots. Par contre, la règle de dévoisement des

[31] Le langage AWK est un langage interprété. Les scripts écrits avec ce langage (cf. Annexes pour des exemples) sont 'interprétés' par un logiciel dédié. Celui-ci existe pour diverses catégories d'ordinateurs et de systèmes ➥

occlusives lorsqu'elles précèdent une constrictive non-voisée n'est pas toujours appliquée. Ainsi, bien que 'médecin' subisse la chute de schwa entre /d/ et /s/, le /d/ n'est pas retranscrit en [t], ce qui donne la transcription /mɛdsɛ̃/ et non pas /mɛtsɛ̃/ comme on aurait pu s'y attendre (cf. Chapitre 3, section 1.1.1.2). Ce choix correspond, contrairement à celui effectué pour la transcription du /o/ et la prise en compte de la chute de schwa, à une représentation plutôt sous-jacente du phonème correspondant à la lettre 'd' dans le mot 'médecin'. Par contre, le mot 'absolu' est retranscrit [apsoly] en conformité avec cette règle de dévoisement, fournissant ainsi une représentation de la forme phonétique de surface. Cette position intermédiaire, de même qu'une certaine incohérence dans le choix de la représentation adoptée entre transcription phonologique et phonétique peut poser des problèmes, notamment pour une analyse des groupes de consonnes dans la langue française. En effet, il est clair que l'analyse du champ correspondant au mot 'médecin' fournira une information tronquée sur les groupes de consonnes car la représentation sous-jacente de ce mot ne contient pas de groupe consonantique. Du fait de la prise en considération de la chute de schwa, on est en mesure d'enregistrer la présence d'un groupe consonantique. Mais, si la représentation phonémique qui en est fournie dans la base de données en contient un, il n'est pas celui qui est effectivement réalisé par un locuteur natif. La transcription proposée dans cette base de données pour le mot 'médecin' ne correspond donc ni à la représentation sous-jacente, ni à la forme de surface.

Malgré ce problème, nous avons choisi d'utiliser BRULEX (Content et al., 1990) car c'était la seule base de données lexicale informatisée de la langue française qui nous permettait, du fait de la présence d'une transcription 'phonético-phonologique' des mots, de conduire l'analyse que nous souhaitions réaliser avec un outil informatique. Nous avons choisi d'intégrer dans notre analyse les groupes qui transgressent cette règle d'assimilation du trait de voisement en estimant que, même si leur occurrence peut refléter une observation incorrecte pour les groupes de consonnes pris dans leur individualité, ils fourniraient une information essentielle sur la fréquence d'occurrence des *catégories* de groupes de consonnes (occlusive-fricative ou occlusive-occlusive par exemple). Nous cherchions en effet à mettre en évidence une différence de fréquence entre des groupes tautosyllabiques (comme les occlusive-liquide) et hétérosyllabiques (comme les occlusive-fricative). Or ce problème de dévoisement de la consonne initiale concerne essentiellement les occlusive-fricative et les occlusive-occlusive, catégories dont nous souhaitions montrer la plus faible fréquence d'occurrence dans la langue

d'exploitation (une version libre de l'interpréteur AWK, gawk, peut être récupérée sur le site de la *Free Software Foundation* (`http://www.gnu.org`).

par rapport aux occlusives-liquides. Par conséquent, il nous a semblé important de tenir compte de la présence de ces groupes dans l'analyse puisque, transcription erronée ou pas du trait de voisement, leur appartenance à une classe phonétique n'est pas modifiée par la transcription de ce trait. Nous supposons par ailleurs que la quantité relativement importante de mots présents dans ce lexique devrait permettre d'obtenir des données statistiquement fiables malgré la présence de certains défauts dans la constitution de la base de données ; en effet, un comptage rapide des groupes de consonnes transcrits en [+ voisé][- voisé] aboutit à une quantité relativement faible. Dans BRULEX (Content et al., 1990), on recense ainsi 64 mots contenant une séquence de constrictives occlusives ou fricatives dont la forme est [+ voisé][- voisé] alors que la base de données contient au total 18608 mots intégrant un groupe de consonnes quel qu'il soit. Par ailleurs, il apparaît qu'une majorité de ces 64 mots sont des mots composés (par exemple 'protège-cahier'), ce qui justifie la transcription adoptée puisque la règle de dévoisement peut ne pas s'appliquer dans cette situation.

1.2.2. Base de données de mots isolés

Il est par ailleurs essentiel de garder à l'esprit que l'analyse des groupes de consonnes que nous allons proposer ici est en partie biaisée par le choix du corpus de langue que nous avons choisi. Cette base de données est constituée de mots isolés pour lesquels on dispose d'une information concernant la fréquence d'usage dans un corpus de langue. Il nous est cependant impossible d'identifier dans cette base de données des groupes de consonnes qui seraient générés par la juxtaposition de mots de la langue. Ainsi, dans la séquence 'une petite voiture' (/ynpətitvwatyr/), l'association de 'petite' et 'voiture' donne lieu à la prononciation du groupe de consonnes /tv/. Ce groupe apparaît dans un seul mot de la base de données sélectionnée ('tête-vin'). Il est probable qu'une base de données de phrases aurait permis d'observer certains groupes de consonne en quantité plus importante, voire de faire émerger des groupes qui sont absents du corpus choisi. Un travail précédent nous conduit cependant à affirmer que ceci ne changerait probablement pas beaucoup le rapport des fréquences entre les divers groupes de consonnes étudiés. En effet, Malécot (1974) a conduit une analyse assez similaire à la nôtre sur un corpus de français parlé obtenu à partir d'enregistrements de conversations. Les données présentées ne sont pas assez détaillées pour nous permettre de les utiliser dans notre travail mais elles ont l'avantage de comparer la fréquence d'utilisation de chaque groupe de consonnes d'une part à l'intérieur des mots utilisés par les locuteurs et d'autre part aux frontières entre les mots. Cette comparaison permet à Malécot (1974) de mettre en évidence des fréquences moyennes très similaires pour les groupes recensés dans les mots utilisés et pour les groupes qui sont générés

par la juxtaposition de ces mots dans les énoncés produits. Ainsi, la fréquence des groupes de consonnes recensés par Malécot (1974) reflète selon lui une tendance à utiliser prioritairement *à la frontière entre les mots* de la langue des groupes de consonnes qui sont plus fréquents *à l'intérieur des mots* de la langue. On peut donc s'appuyer sur ces données pour affirmer que les données présentées ici sont assez proches de celles qui auraient été obtenues avec un corpus de phrases.

2. Probabilité d'occurrence indépendante de la position

La première partie de l'analyse distributionnelle a consisté à estimer la fréquence des divers groupes de consonnes dans un lexique français sans prendre en compte leur position d'apparition dans les mots. Trois types de données ont été dérivées de cette analyse. La première catégorie de résultats fournit des informations sur le nombre de mots contenant un groupe de consonnes donné. Les deux autres ensembles de données pondèrent la fréquence de ces séquences de phonèmes par la fréquence des mots dans lesquels ils apparaissent (probabilité pondérée) ou par la fréquence de la première consonne (probabilité transitionnelle). Notre objectif est de montrer que des effets interprétables en termes de segmentation du signal de parole fondée sur les connaissances phonologiques des auditeurs mais reposant sur des processus cognitifs tout à fait différents pourraient expliquer les données obtenues par McQueen (1998) et Vroomen & De Gelder (1999). Si les données statistiques mettent en évidence un lien entre légalité / tautosyllabicité et fréquence, cette analyse nous mettra en position de réinterpréter les données présentées. En effet, si la syllabation des séquences de deux phonèmes peut se refléter dans leur fréquence d'apparition, il est alors possible de donner une interprétation alternative des données obtenues (en termes de segmentation probabiliste notamment, cf. section 4 de ce chapitre pour une discussion plus approfondie). C'est dans le but d'estimer l'impact de la fréquence d'occurrence des séquences de consonnes dans l'environnement linguistique sur les processus perceptifs que nous avons décidé de ne pas nous limiter à une information concernant le nombre de mots dans lesquels chaque séquence phonémique est attestée. En effet, si des processus purement probabilistes peuvent expliquer les effets obtenus, il est essentiel d'obtenir une approximation correcte de la fréquence avec laquelle une séquence de phonèmes est entendue dans les situations de communication naturelles. Or certains mots sont utilisés beaucoup plus fréquemment que d'autres. Il est donc essentiel, pour estimer la fréquence d'apparition d'une forme dans la langue, de pondérer le nombre de mots dans lesquels elle apparaît par leur fréquence d'usage. Le choix de réaliser une analyse des probabilités

transitionnelles d'apparition est quant à lui dicté par des propositions récentes concernant les processus d'acquisition du langage ; propositions consistant à affirmer que le système cognitif de l'enfant mettrait en place une analyse des probabilités transitionnelles des séquences pour découvrir les frontières entre les mots avant d'avoir acquis un lexique (Saffran, Aslin, & Newport, 1996; Brent, 1996) et que l'adulte pourrait continuer d'avoir recours à ce type de procédures (Saffran, Newport et al., 1996; Brent & Cartwright, 1996; Brent, 1997) pour prédire la localisation des frontières de mots une fois ce lexique constitué. La notion de probabilités transitionnelles de même que sa distinction d'avec un simple calcul de fréquence est présentée dans la section 2.1.3. Elle consiste à tenir compte de la fréquence de la première consonne pour estimer la probabilité d'apparition de la seconde consonne.

2.1. Méthode d'analyse

Les analyses statistiques ont été effectuées sur ordinateur à l'aide du langage de programmation AWK (Aho et al., 1988). Pour extraire les ensembles de données présentés dans cette section, le script 'lisait' un fichier contenant la liste des séquences de phonèmes à rechercher dans le lexique (la base de données). Une fois cette liste recensée, il parcourait les champs de la base de données correspondant à la transcription phonémique et recherchait les occurrences de chacune des séquences qui lui étaient fournies en entrée sans se préoccuper de leur position dans les mots. Pour chaque séquence rencontrée dans la base, il incrémentait deux variables : l'une correspondait au nombre de mots contenant cette séquence, l'autre correspondait à la fréquence d'usage des mots qui est fournie dans BRULEX (Content et al., 1990). A la fin de l'analyse, le script sauvegardait un fichier texte qui contenait la liste des séquences de phonèmes recherchées et pour chacune d'entre elles le nombre de mots contenant cette séquence assorti de leur fréquence cumulée. Parallèlement à chaque analyse, nous avons également recensé les occurrences de consonnes prévocaliques afin d'estimer la fréquence des suites CV et de la comparer à celle des suites CC (le script adéquat pour le recensement du nombre de mots constitués de ces séquences est reproduit en Annexe 4, p.IV).

2.1.1. Fréquence d'occurrence

A partir des informations fournies par le script d'analyse (celui-ci est reproduit en Annexe 2, p.III), nous disposons directement d'une première série de données. Cette information est appelée ici *fréquence d'occurrence* des groupes de consonnes. Elle correspond au nombre de mots dans lesquels ces groupes apparaissent.

2.1.2. Probabilités pondérées

En second lieu, nous présentons les résultats d'une analyse de la fréquence de ces groupes en la pondérant par la fréquence des mots dans lesquels ils apparaissent. Nous appelons cet indice une *probabilité d'occurrence* dans la langue. Ces données fournissent une estimation plus fiable de la fréquence d'apparition des séquences dans les situations de communication linguistique puisqu'elles prennent en considération la fréquence d'usage des mots dans lesquels les groupes de consonnes sont prononcés. On peut en effet envisager qu'un groupe de consonnes soit relativement rare mais qu'il soit utilisé dans des mots qui, quant à eux, sont très souvent utilisés dans la langue. Ce groupe rare aurait alors, en réalité, une fréquence d'apparition assez élevée dans la langue malgré le nombre limité de mots qui le contiennent. Cette analyse a été limitée aux mots pour lesquels il existe, dans BRULEX (Content et al., 1990), une information sur la fréquence d'usage. Un second script a donc été utilisé ici qui se restreignait à analyser les entrées de la base de données pour lesquelles une information sur la fréquence du mot était fournie (le script correspondant est reproduit en Annexe 3, p.IV). La pondération a été réalisée en multipliant le nombre de mots dans lesquels le groupe apparaît par la fréquence cumulée de l'ensemble de ces mots. Du fait des différences de valeur importantes entre les résultats obtenus pour les groupes rares et ceux obtenus pour les groupes très fréquents, le résultat de cette multiplication à été transformé en valeurs logarithmiques décimales selon la formule suivante :

$$10 * \log_{10}(\text{Nombre\_de\_mots} * \text{Fréquence\_cumulée})$$

La multiplication par 10 a simplement pour objet de générer des valeurs approximées entières. Cette transformation logarithmique nous semble faciliter la visualisation de la distribution des probabilités moyennes d'occurrence des divers groupes de consonnes étudiés car elle permet de présenter sur une échelle de dimension restreinte les données correspondant aux séquences rares aussi bien que celles liées aux séquences fréquentes.

2.1.3. Probabilités transitionnelles

Finalement, le troisième type de données qui sont présentées ici prend en compte la fréquence de la première consonne du groupe pour estimer les probabilités transitionnelles des séquences qui nous intéressent. Dans l'estimation de la fréquence d'une séquence XY, une valeur élevée peut parfois s'expliquer essentiellement par une haute fréquence d'occurrence de la forme X. Si X est très fréquent, il apparaît souvent dans le corpus ; la forme Y a alors beaucoup de chances d'apparaître en conjonction avec la forme X. Cependant, cette valeur élevée de la

fréquence d'apparition de XY est en partie déterminée par la fréquence de X. Cette fréquence de la suite XY peut donc, paradoxalement, n'être pas très informative.

Dans le cadre de la théorie de l'information (Shannon, 1948), l'informativité est en effet une notion très spécifique. Une forme est informative si sa présence est en mesure d'apporter une quantité d'information plus importante que celle dont elle est intrinsèquement constituée. Notamment, une forme qui permet de prédire la présence d'autres formes est très informative car elle est en mesure de fournir des informations permettant par exemple d'identifier un signal malgré une dégradation de celui-ci liée au médium de transmission. Une séquence XY fréquente n'est pas nécessairement informative. Si X est fréquent, il peut être suivi de n'importe quelle forme (aussi bien Y que Z). Par contre, si la présence de X permet de prédire avec un niveau de certitude élevé la présence de Y mais pas celle de Z, alors on peut dire que X est informatif. C'est dans ce cas précis que la suite XY aura un taux de probabilité transitionnelle élevé. Supposons deux séquences AB et CD. AB est fréquente alors que CD est rare. Il est possible que ces deux séquences ne soient pas plus informatives l'une que l'autre. Si A est très fréquent et que C est très rare, la présence de C peut éventuellement permettre, autant que celle de A, de prédire la forme suivante. C'est par exemple le cas si, malgré une fréquence d'occurrence très restreinte de C, la suite CD est relativement fréquente par rapport aux autres suites contenant C, c'est à dire si la présence de C détermine avec une certitude élevée celle de D. Si l'on se réfère à la théorie de l'information introduite par Shannon (1948), ce n'est donc pas la fréquence en soi qui est informative mais le rapport entre fréquence de la forme globale et celle de l'une des deux formes.

Nous avons insisté, dans le Chapitre 2, sur le caractère sériel (au moins en partie) du traitement appliqué par le système cognitif sur ce signal. Il nous semble donc logique de conduire cette analyse en termes de probabilités transitionnelles en prenant comme référent du calcul des probabilités transitionnelles la première consonne du groupe. Dans le cadre d'une comparaison des groupes de consonnes entre eux, cette distinction entre fréquence et probabilité transitionnelle est importante si l'on compare des groupes de consonnes qui commencent par une consonne différente. En effet, tant que l'on compare des groupes de consonnes qui partagent un élément commun, une estimation de la probabilité transitionnelle des séquences n'apporte rien de plus qu'une analyse de la fréquence de ces séquences dans le corpus puisque la fréquence de l'élément initial est la même. Par contre, lorsque l'on comparera des groupes de consonnes qui ne partagent pas la même consonne initiale, il sera important de recourir à cet indice de probabilité transitionnelle pour estimer leur différence en termes probabilistes.

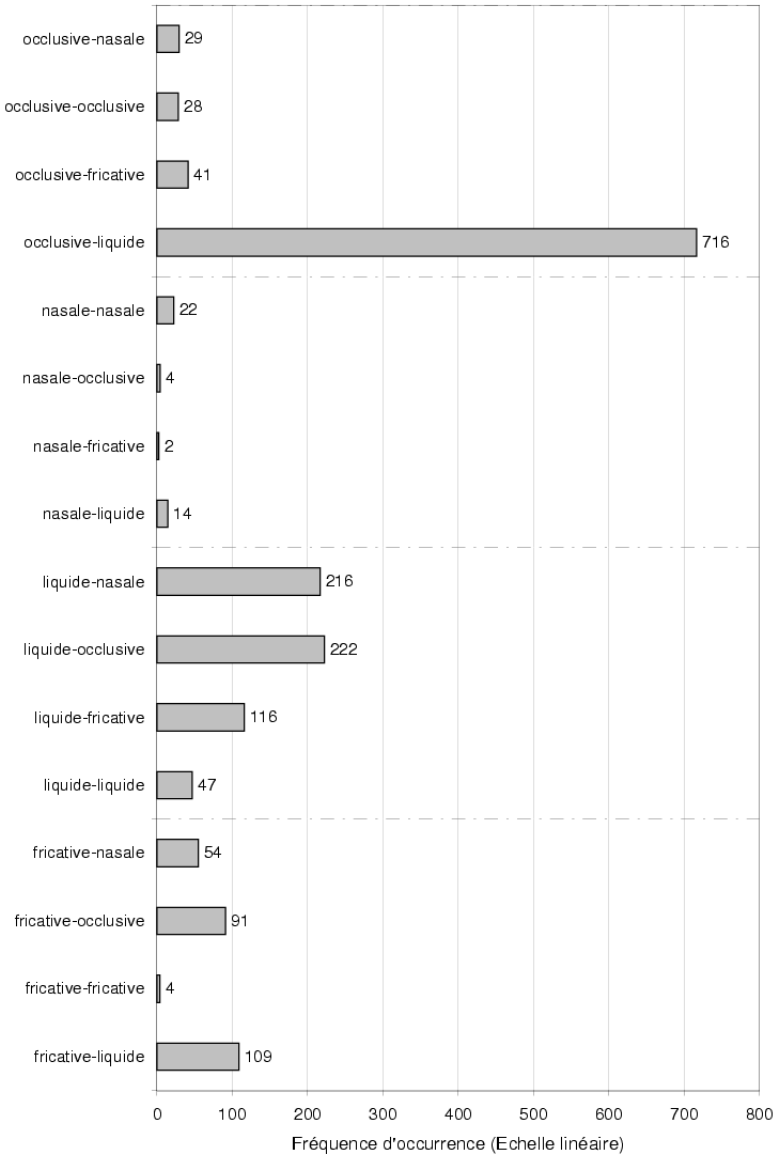

Figure 18 : Nombre moyen d'occurrences des groupes de consonnes en fonction de leur structure phonémique (mode d'articulation). Données calculées sur le corpus BRULEX (Content et al., 1990).

2.2. Résultats

Nous présentons en premier lieu les moyennes obtenues par catégorie de groupe en fonction du mode d'articulation de chaque consonne (occlusive, fricative, nasale, liquide). Nous restreignons notre description à un certain nombre de groupes qui apportent des informations particulièrement intéressantes en ce qui concerne le lien entre syllabation et fréquence. Les données intégrales pour chaque groupe sont présentées en Annexe 5 (p.V). Dans une seconde étape, nous étudions la distribution des fréquences individuelles de chaque groupe en fonction de sa catégorie phonétique.

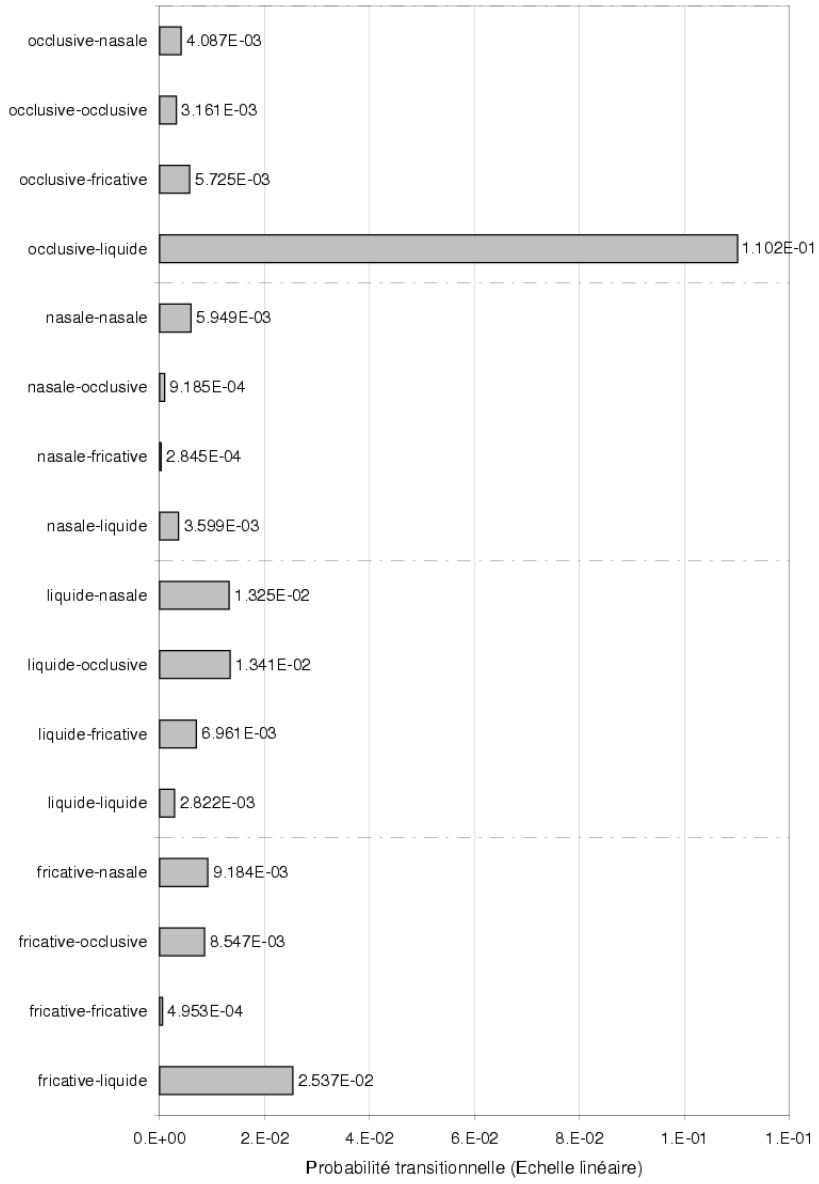

Figure 19 : Probabilité transitionnelle moyenne des groupes de consonnes en fonction de leur structure phonémique (mode d'articulation). Les fréquences du groupe de la consonne initiale sont calculées à partir de la base de données BRULEX (Content et al., 1990).

Une première observation importante est que les trois catégories de calculs que nous avons effectuées sur les groupes de consonnes présentent une forte corrélation positive entre elles. Ainsi, la fréquence des groupes de consonnes (leur nombre d'occurrences) et la fréquence d'usage des mots qui les contiennent sont corrélées positivement ($r = .97$). Cette corrélation met en évidence une tendance, pour les groupes de consonnes fréquents, à être utilisés dans des mots également fréquents. On observe le même phénomène pour le calcul de la corrélation entre fréquence et probabilité transitionnelle ($r = .83$) ; ce qui implique une homogénéité de l'utilisation des consonnes dans les divers groupes recensés. On peut ainsi raisonnablement

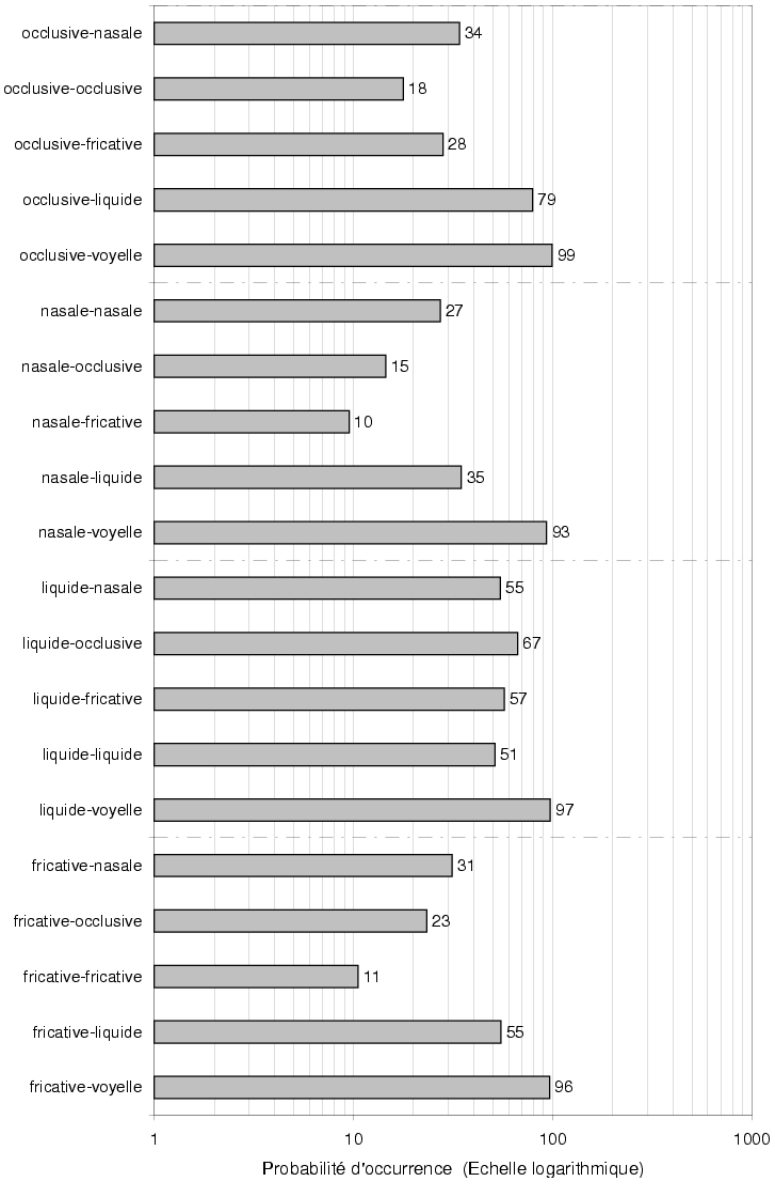

Figure 20 : Probabilité d'occurrence moyenne des groupes de consonnes en fonction de leur structure phonémique (mode d'articulation). Données calculées sur le corpus BRULEX (Content et al., 1990).

affirmer que l'indice de probabilité transitionnelle n'apporte pas beaucoup plus d'information que celui de la fréquence d'occurrence dans le lexique, du moins pour l'étude des fréquences moyennes par catégorie phonétique. La valeur du coefficient de corrélation est cependant moins élevée qu'entre fréquence d'occurrence et fréquence d'usage des mots, ce qui indique qu'une quantité plus importante de groupes de consonnes donne lieu à une non-correspondance entre fréquence et probabilité transitionnelle qu'entre fréquence et probabilité d'occurrence.

2.2.1. Moyennes

La première étape de notre analyse consiste à étudier la fréquence moyenne (selon les 3 indices décrits précédemment) des groupes de consonnes en fonction de leurs caractéristiques phonétiques de mode d'articulation. Les graphiques suivants illustrent les différences de fréquence moyenne entre catégories. Dans le premier (Figure 18) sont présentées les données moyennes obtenues pour le comptage du nombre de mots dans lesquels chaque groupe consonantique est recensé. Nous ne présentons pas dans ce graphique les données obtenues pour les séquences CV (Consonne-Voyelle). Cette omission nous permet de mieux visualiser les différences de fréquence entre groupes de consonnes, ceux-ci étant toujours nettement plus rares que les séquences CV.

Les deux autres graphiques reproduisent respectivement les probabilités transitionnelles (Figure 19) et les probabilités d'occurrence (fréquence du groupe pondérée par la fréquence d'usage des mots, Figure 20). Ce dernier permet de présenter, outre les résultats observés pour les groupes de consonnes, les données obtenues pour les séquences Consonne-Voyelle. Le choix d'une transformation logarithmique nous permet en effet de représenter simultanément les données correspondant aux séquences très rares et très fréquentes. On remarquera ici le cas particulier des groupes fricative-nasale et fricative-occlusive dont la position sur l'échelle du nombre d'occurrences (Figure 18) est l'inverse de celle observée sur l'échelle des probabilités pondérées (Figure 20) ; ceci met en évidence, malgré la forte corrélation observée entre nombre d'occurrences et fréquence d'usage des mots, la possibilité que cette fréquence d'usage puisse contribuer à la probabilité d'occurrence des suites de phonèmes dans la langue. Les données de probabilité transitionnelle ne diffèrent par contre pas des données de fréquence d'occurrence. Cette équivalence des résultats observés est probablement liée à l'équiprobabilité des phonèmes individuels dans la langue. Les résultats présentés seront donc désormais fondés sur les données de *probabilité d'occurrence* (fréquence d'occurrence pondérée par la fréquence d'usage des mots) car ils permettent à la fois de prendre en compte la productivité des groupes de consonnes dans le lexique et de comparer les groupes de consonnes avec les suites Consonne-Voyelle.

Du fait du nombre considérable de catégories à prendre en compte dans une comparaison statistique des moyennes observées en fonction du mode d'articulation des phonèmes constituant ces groupes de consonnes, nous avons conduit cette analyse à l'aide de tests statistiques *post-hoc*. La valeur élevée du nombre de degrés de liberté de la comparaison globale, ce nombre de degrés de liberté étant lié à la quantité d'observations, nous a conduit à sélectionner un test statistique relativement conservateur afin de limiter le nombre de comparaisons qui feraient émerger une différence significative (Winer, 1971). Nous avons donc utilisé le test de Scheffé

pour comparer les moyennes entre elles. Nous focalisons notre analyse sur les groupes constitués d'une fricative ou d'une occlusive à l'initiale car ils nous semblent constituer deux catégories particulièrement propices à l'étude du lien entre syllabation et fréquence et qu'ils fournissent par ailleurs une quantité de données plus substantielle que les autres types de groupes consonantiques. Certains des membres de ces catégories peuvent en outre être raisonnablement considérés comme des attaques de syllabe bien formées. C'est en particulier le cas de la plupart des occlusive-liquide et fricative-liquide. Les groupes à initiale liquide ou nasale nous semblent plutôt correspondre à des coda syllabiques (/rk/ comme dans /bark/) ou même à des séquences comportant une frontière syllabique médiane (/nk/ comme dans /mankɛ̃/, la seule exception nous semblant être l'emprunt 'round'). Or la description de codas complexes (contenant plusieurs consonnes) n'est peut-être pas justifiée (Dell, 1995). Ce dernier propose au contraire que les groupes de consonnes décrits comme des codas complexes constituent plutôt une juxtaposition de deux positions syllabiques différentes, la consonne initiale du groupe ayant le statut effectif de coda alors que la seconde correspondrait à une attaque de syllabe (laquelle comporterait un noyau vide). La problématique que nous avons choisi d'aborder dans ce travail se fonde principalement sur des groupes de consonnes qui peuvent ou non être regroupés à l'attaque syllabique. Nous cherchons donc à déterminer ce qu'est une attaque de syllabe bien formée (ou légale). En restreignant notre analyse à deux catégories de groupes parmi lesquelles une part non négligeable (mais pas l'ensemble) peut être caractérisée comme une attaque de syllabe bien formée, nous serons en mesure d'utiliser ces données fréquentielles pour réanalyser les données comportementales auxquelles nous nous intéressons.

Les séquences C[32]-liquide présentent une probabilité moyenne d'occurrence plus élevée (79 pour les groupes à initiale occlusive ; 55 pour ceux à initiale fricative) que les autres catégories (respectivement 18, 28 et 34 pour les groupes à initiale occlusive ; 11, 23, 31 pour les groupes à initiale fricative). Les séquences C-voyelle présentent quant à elles des probabilités d'occurrence plus importantes que l'ensemble des autres catégories (respectivement 99 et 96 pour les occlusive-voyelle et les fricative-voyelle). L'analyse statistique que nous avons conduite met en évidence, comme il était possible de s'y attendre au vu des graphiques ci-dessus, des différences significatives de probabilité d'occurrence entre les différents groupes de consonnes. Le Tableau 3 présente les seuils de probabilité obtenus avec le test de Scheffé pour la comparaison des séquences commençant par une occlusive (Tableau 3a) et par une fricative (Tableau 3b). Nous avons tracé dans chacun de ces tableaux un rectangle indiquant la frontière

[32] C étant ici mis pour fricative ou occlusive.

entre les séquences qui sont essentiellement tautosyllabiques (C-voyelle, C-liquide) et celles qui seraient plutôt hétérosyllabiques (C-fricative, C-occlusive, C-nasale). A l'intérieur de ce rectangle, on trouve les seuils de probabilité obtenus en comparant des ensembles de séquences en général tautosyllabiques d'une part et hétérosyllabiques de l'autre. Les seuils de probabilité situés à l'extérieur de ce rectangle correspondent quant à eux à des comparaisons de séquences à l'intérieur de l'ensemble considéré comme tautosyllabique (à gauche) ou hétérosyllabique (en dessous du rectangle).

Tableau 3 : Seuils de probabilité des tests de Scheffé appliqués à la comparaison de probabilités d'occurrence des groupes de consonnes à initiale occlusive (a) ou fricative (b). Les comparaisons sont effectuées par catégorie de groupe en fonction du mode d'articulation des phonèmes. Les seuils de probabilité statistiquement significatifs sont retranscrits en caractères gras et italique. Les intitulés des lignes et des colonnes correspondent au second phonème de la séquence.

a/ Occlusive initiale	liquide	fricative	occlusive	nasale
voyelle	0.516	***0.000***	***0.000***	***0.000***
liquide		***0.000***	***0.000***	***0.000***
fricative			0.416	0.929
occlusive				0.171

b/ Fricative initiale	liquide	fricative	occlusive	nasale
voyelle	***0.010***	***0.000***	***0.000***	***0.000***
liquide		***0.000***	***0.002***	0.090
fricative			0.204	***0.037***
occlusive				0.814

Conformément à nos prédictions, il semble exister un lien entre la structure phonologique des groupes de consonnes et leur fréquence d'occurrence dans la langue. Ce lien est particulièrement prégnant si l'on observe les différences de fréquence dans la catégorie des groupes commençant par une occlusive. Il est clair que la plupart des groupes occlusive-liquide, de même que toutes les séquences occlusive-voyelle, sont produits en attaque syllabique et que, pour la plupart, aucune frontière syllabique / phonotactique n'est insérée entre les deux consonnes de ces groupes. Or ceux-ci sont en moyenne beaucoup plus fréquents que toutes les autres catégories de groupes consonantiques commençant par une occlusive. On observe par ailleurs une nette dissociation entre les deux catégories que nous considérons comme essentiellement tautosyllabiques (occlusive-liquide et occlusive-voyelle) et celles dont nous jugeons qu'elles correspondent plutôt à des groupes hétérosyllabiques (occlusive-fricative, occlusive-occlusive, occlusive-nasale). En effet, tous les seuils de probabilité du test de Scheffé sont inférieurs à 0.01

à l'intérieur du rectangle délimitant la frontière entre ces deux ensembles alors qu'aucune des comparaisons intra-catégorie n'est significative. Le test de Scheffé a également été conduit sur les données de fréquence d'occurrence et de probabilité transitionnelle. A l'exception des données de fréquence obtenues pour les groupes à initiale occlusive, il n'émerge que peu d'effets significatifs de ces comparaisons (cf. Annexe 6 ; p.XII). Pour la plupart des comparaisons, seule émerge une distinction entre les suites CV et CC[33].

Les données statistiques obtenues pour les groupes à initiale fricative (Tableau 3b) fournissent des résultats moins clairs concernant le lien entre tautosyllabicité et fréquence dans la langue. Il n'apparaît pas de dissociation nette entre les séquences fricative-voyelle et fricative-liquide d'une part et les 3 autres catégories d'autre part. Les suites fricative-liquide sont significativement moins fréquentes que les suites fricative-voyelle. Ceci ne contredit pas notre hypothèse puisque tautosyllabicité et fréquence pourraient tout à fait être liées sans induire une absence de différence fréquentielle intra-catégorie. On observe cependant, à l'intérieur du rectangle représentant la distinction entre groupes supposés tautosyllabiques et groupes supposés hétérosyllabiques, des comparaisons dont le test de Scheffé ne permet pas d'affirmer qu'elles correspondent à des séquences différant significativement en termes de fréquence d'occurrence. Il est probable que cette distinction moins nette entre les deux ensembles définis *a priori* soit liée à une plus grande variabilité dans la syllabation des groupes de consonnes commençant par une fricative que dans celle des groupes à initiale occlusive. Nous avons mentionné le caractère illégal des deux séquences /tl/ et /dl/ qui, malgré leur prise en compte dans les données des occlusive-liquide ne peuvent pas être tautosyllabiques. L'ensemble des groupes à initiale fricative contient certainement une quantité plus importante de groupes ne pouvant pas être classés simplement en fonction de leur mode d'articulation. Les groupes fricative-liquide comme /vr/ ou /fl/ constituent des attaques syllabiques bien formées. On trouve par exemple des suites fricative-liquide dans 'vrai', 'frelon', 'flibustier'… On trouve par contre dans cette catégorie des séquences qui pourraient être décrites comme hétérosyllabiques. C'est par exemple le cas de /sl/ qui, bien qu'il puisse constituer une attaque syllabique bien formée (comme dans 'slalom')

[33] Il est probable que ce soit lié à des différences de fréquence trop importantes entre les séquences CV et CC lorsque l'on utilise une échelle linéaire. En effet, la plupart des tests statistiques paramétriques a pour principe de comparer la taille de chaque effet en utilisant comme mesure de l'*erreur* la variance globale des mesures effectuées. Dans le cas d'une échelle linéaire des fréquences (ou des probabilités transitionnelles), Les suites CV ont une valeur considérablement plus importante que les suites CC sur cette échelle. Elles accroissent donc la variance à un tel point que le test n'est plus en mesure de détecter que des différences moyennes très importantes. Le choix d'une échelle logarithmique pour la présentation des probabilités d'occurrence a certainement permis de réduire la variance introduite par les séquences CV et de faire émerger, de fait, ces différences entre catégories de groupes consonantiques.

pourrait être considéré comme hétérosyllabique en position intervocalique (dans ‘islam’ par exemple). On peut certainement observer ce type de variabilité parmi les autres catégories de groupes. Le cas des fricative-occlusive est particulièrement frappant. Les groupes /s/ + occlusive apparaissent régulièrement en début de mot ; ce qui n’est pas le cas des groupes /f/ + occlusive. Les premiers pourraient constituer des attaques de syllabe bien formées. Il est par conséquent probable que les catégories comparées pour les groupes à initiale occlusive présentent une variabilité plus limitée que celles correspondant aux groupes à initiale fricative. Cette variabilité plus restreinte pourrait stabiliser les résultats statistiques obtenus. Malgré cette relative incohérence des données statistiques obtenues pour les groupes à initiale fricative, on retrouve dans le tableau des seuils de probabilité du test de Scheffé une certaine tendance à grouper les séquences tautosyllabiques et hétérosyllabiques ensemble : parmi les 6 comparaisons effectuées entre les catégories constituant ces deux ensembles, cinq sont statistiquement significatives. Cette proportion nous semble fournir un assez bon indicateur du lien que nous cherchons à mettre en évidence. L’analyse des groupes à initiale fricative contribue donc également, bien que dans une moindre mesure, à l’affirmation d’un lien entre syllabation et fréquence.

Ces données moyennes ne fournissent cependant pas une explication satisfaisante du lien qu’il est possible de décrire entre fréquence et syllabation. En effet, la question qui nous semble essentielle ici est de déterminer si, alors que fréquence et tautosyllabicité sont effectivement corrélées, il est possible de dissocier ces deux variables ou si, au contraire, elles constituent deux expressions d’un seul et même phénomène : une régularité phonologique observable déterminée par les contraintes inhérentes à la langue et dont ces deux manifestations seraient indissociables. Il est déjà possible de prédire, à partir des données statistiques fournies par le test de Scheffé, que tel n’est pas le cas. En effet, on observe pour les séquences à initiale fricative des incohérences dans la distribution des seuils de significativité qui ne permettent pas entièrement de conclure à une correspondance terme à terme entre tautosyllabicité et fréquence. Cette incohérence pourrait en partie s’expliquer par l’intégration, dans chaque ensemble (tautosyllabique vs. hétérosyllabique), de groupes de consonnes qui appartiennent sans aucun doute à l’ensemble opposé. On peut cependant affirmer que, même si nous avions regroupé dans le cadre de cette analyse des groupes ‘purs’ dont il serait impossible de dire qu’ils devraient être classés dans l’ensemble opposé, nous aurions pu observer un certain flou dans les résultats statistiques. La première raison est liée au caractère conservateur du test de Scheffé qui présente plutôt une

tendance à ne pas rejeter l'hypothèse nulle[34]. Il reste néanmoins que certaines comparaisons font émerger une différence significative à laquelle nous ne nous attendions pas. Ceci peut en partie être lié aux multiples comparaisons qui sont effectuées dans le cadre d'un test post-hoc, cette quantité importante aboutissant au risque d'observer par hasard des effets significatifs même si les tests *post-hoc* sont justement conçus pour limiter ces risques. Une seconde explication -laquelle pourrait rendre compte de l'émergence d'effets significatifs inattendus sans reposer sur des considérations purement statistiques- est que, peut-être, fréquence et tautosyllabicité constituent des phénomènes corrélés mais sans réelle source d'explication commune. Notamment, il est possible que la tendance à utiliser plus fréquemment des groupes tautosyllabiques dans les mots de la langue ne soit que partiellement déterminée par les contraintes phonologiques de la langue ; d'autres paramètres pouvant alors intervenir (comme le hasard). Ces deux observables pourraient alors constituer des phénomènes corrélés mais en partie dissociables. Si c'est le cas, il est alors certainement possible de mettre en évidence des recouvrements dans les distributions de probabilité d'occurrence catégorisée en fonction des types de groupes de consonnes ; recouvrements qui nous permettraient de séparer tautosyllabicité et fréquence et fourniraient par conséquent les moyens d'une étude plus précise de la contribution respective des régularités phonologiques et de la fréquence des groupes de phonèmes dans les résultats comportementaux obtenus McQueen (1998) et par Vroomen & De Gelder (1999).

2.2.2. Distributions

Si la différence de fréquence observée est directement liée à la tautosyllabicité des groupes de consonnes et que ces deux paramètres sont indissociables, alors il sera impossible de déterminer si des processus faisant intervenir des représentations phonologiques sont effectivement à l'œuvre dans l'émergence des effets observés par McQueen (1998) et Vroomen & De Gelder (1999) ou si ces effets pourraient au contraire être déterminés par des processus impliquant le recours à des calculs probabilistes tels que ceux proposés par Saffran, Newport, & Aslin (1996) ou par Brent & Cartwright (1996). Puisque la fréquence des groupes de consonnes dans la langue est fortement liée à leur syllabation, il est essentiel de déterminer si ces deux variables sont séparables ou si, au contraire, elles correspondent à deux facettes d'un phénomène unique. Si fréquence et tautosyllabicité sont liées mais pas confondues, leurs distributions devraient se chevaucher. La mise en évidence de cette orthogonalité des dimensions fournirait

[34] On notera que pour les groupes à initiale fricative, le seul seuil non-significatif dans le rectangle dissociant séquences tautosyllabiques et hétérosyllabiques pourrait en ce sens être considéré comme marginal (p=.090).

alors un outil permettant de tester indépendamment les effets respectifs de la fréquence et de la structuration syllabique dans les processus cognitifs de segmentation lexicale. Afin de répondre à cette question, nous présentons les graphiques correspondant aux distributions de probabilité d'occurrence des groupes de consonnes en les classant par catégorie phonétique de mode d'articulation. La Figure 21 illustre les distributions des divers groupes de consonnes commençant par une occlusive (/b/, /d/, /g/, /p/, /t/, /k/). Les données correspondant aux groupes à initiale fricative (/v/, /z/, /ʒ/, /f/, /s/, /ʃ/) sont présentées dans la Figure 22.

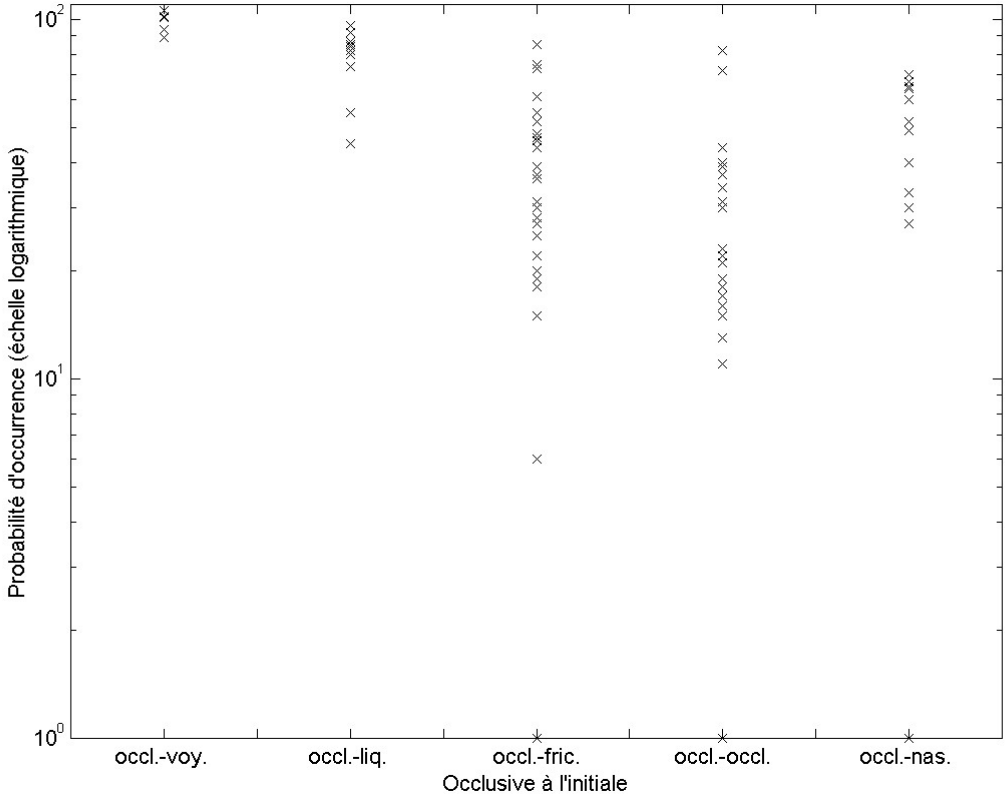

Figure 21 : Distribution des probabilités d'occurrence pour les groupes de consonnes à initiale occlusive.

On peut voir, malgré la tendance d'une fréquence plus élevée des groupes de consonnes tautosyllabiques mise en évidence dans la section précédente, que ces deux variables ne sont pas confondues. Que l'on observe aussi bien le graphique correspondant aux groupes à initiale occlusive que celui des distributions correspondant aux groupes à initiale fricative, les distributions de probabilité d'occurrence se recoupent ; ce qui signifie que deux groupes de consonnes ayant des caractéristiques phonétiques différentes peuvent présenter une fréquence similaire. Notamment, un groupe ayant tendance à être tautosyllabique peut être aussi fréquent qu'un groupe hétérosyllabique. De même, deux groupes de caractéristiques phonétiques similaires peuvent présenter des différences de fréquence de grande amplitude, qu'ils soient

hétérosyllabiques ou pas. On remarque que cette observation vaut aussi bien pour les groupes à initiale occlusive que pour ceux à initiale fricative. Si l'on admet que ces groupes peuvent se scinder en deux catégories : les groupes fricative-liquide et occlusive-liquide appartenant pour la plupart à l'ensemble des groupes tautosyllabiques et les autres types de groupes (C-fricative, C-occlusive, et C-nasale) se prononçant essentiellement de manière hétérosyllabique, il existe -malgré des différences importantes de fréquence moyenne- des paires présentant une probabilité d'occurrence identique mais n'appartenant pas au même ensemble de groupes consonantiques.

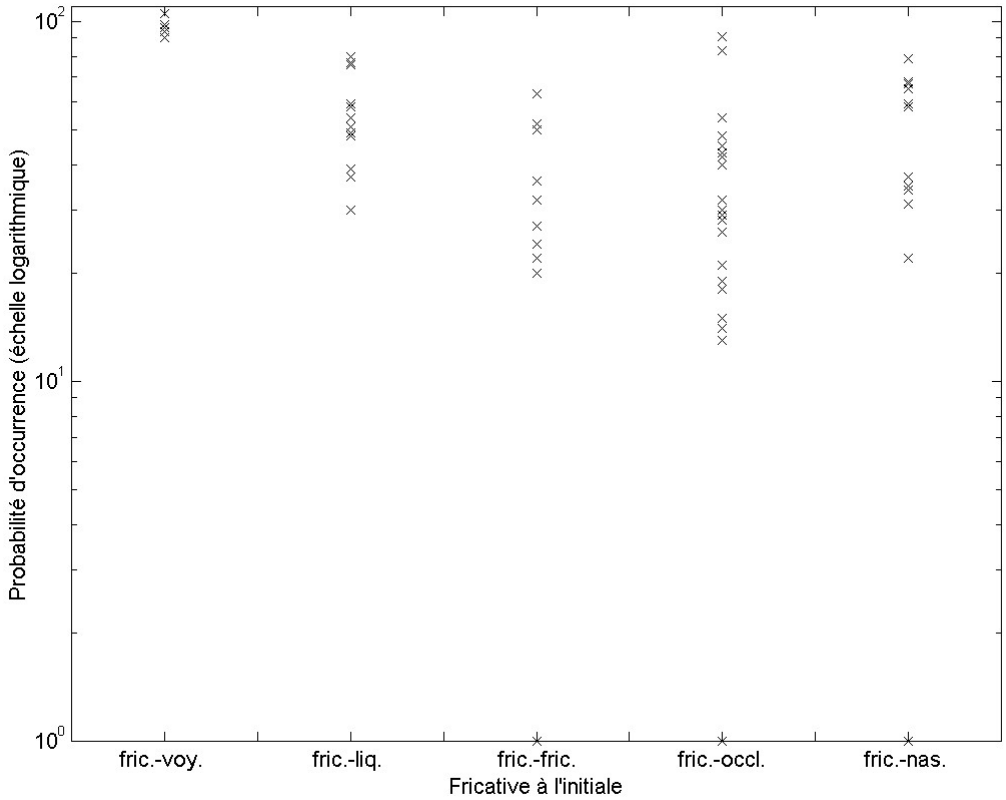

Figure 22 : Distribution des probabilités d'occurrence pour les groupes de consonnes à initiale fricative.

2.3. Discussion

Si fréquence et tautosyllabicité sont liées, elles ne sont cependant pas confondues. Il est par conséquent concevable de dissocier ces deux paramètres dans une étude du rôle de la structuration syllabique ou des contraintes phonotactiques dans les processus cognitifs de segmentation de la parole en mots. En effet, nous avons observé que les groupes typiquement prononcés à l'attaque de syllabe ont une tendance à apparaître plus fréquemment que les autres dans les mots de la langue. Ce résultat met en évidence une relation entre un observable (la

tautosyllabicité) dont la cause supposée est le respect de contraintes phonologiques intégrées au système linguistique propre à la langue et un autre observable (la fréquence) qui pourrait se manifester indépendamment de contraintes linguistiques. Au contraire, la fréquence d'utilisation des séquences de consonnes quelle que soit leur position dans les mots n'est pas indépendante des principes de syllabation déterminant la structuration des consonnes dans la chaîne parlée. Nous verrons, dans la section 4 de ce chapitre que cette observation permet de réanalyser les données présentées par McQueen (1998) et Vroomen & De Gelder (1999) en proposant quelques interprétations alternatives. La section 2.2.2 nous a cependant permis de mettre en évidence une dissociation entre tautosyllabicité et fréquence. Cette dissociation apparaît avec l'observation des distributions de probabilité d'occurrence en fonction de la catégorie de mode d'articulation. Malgré la correspondance observée entre probabilité moyenne d'occurrence et caractère tauto- ou hétérosyllabique, l'analyse précédente permet d'observer un recouvrement des distributions de probabilité d'occurrence relevées pour les diverses catégories de groupes de consonnes. Cette observation vaut aussi bien pour les groupes à initiale fricative que pour ceux à initiale occlusive. Nous sommes donc en mesure d'affirmer que, si fréquence d'occurrence et tautosyllabicité des groupes de consonnes sont liées, elles ne sont pas confondues. De fait, il est légitime d'affirmer que la fréquence d'occurrence d'un groupe de consonnes est un observable qui n'est pas intégralement déterminé par les contraintes linguistiques de la langue. Il devrait donc être possible de dissocier les effets respectifs de la fréquence et de la structure phonologique (qu'elle soit liée à la syllabicité ou à la légalité phonotactique) dans des études portant sur les processus impliqués dans la segmentation de la parole en mots.

3. Probabilité d'occurrence en début de mot

Il est cependant nécessaire, du fait de cette absence de correspondance stricte entre tautosyllabicité et fréquence, de choisir un critère opérationnel qui nous permettra de catégoriser les groupes selon des critères purement phonologiques (phonotactiquement légal vs. illégal ou tautosyllabique vs. hétérosyllabique) et de distinguer celui-ci d'un critère probabiliste sur lequel serait fondée une catégorisation en classes fréquentielles. Pour cela, il nous a semblé intéressant de conduire une analyse similaire à la précédente mais qui serait restreinte à estimer la fréquence des groupes de consonnes en début de mot. Afin de décrire chaque groupe de consonne en termes de structuration syllabique, il nous faut adopter une définition opérationnelle de ce qu'est une attaque de syllabe bien formée. Or nous avons vu dans la description des principes de syllabation aussi bien que dans la section consacrée à la question de la légalité phonotactique que

différentes conceptions de ces notions peuvent être rencontrées. Dans les travaux réalisés en phonologie, on considère en général que toute séquence attestée dans la langue en une position donnée (par exemple en début de mot) est phonotactiquement légale dans cette position (ou constitue une attaque syllabique bien formée). Si l'on se réfère au travail de Dell (1995), le fait qu'un seul exemplaire soit attesté dans la langue en début de mot conduit à le considérer comme une attaque de syllabe bien formée. Il nous semble cependant que ce critère n'est pas assez discriminant. En effet, si l'on recense l'ensemble des groupes attestés en position initiale de mot, on s'aperçoit qu'une quantité considérable de groupes de consonnes peut apparaître en position initiale dans au moins un mot de la langue.

Tableau 4 : Exemples de groupes de consonnes attestés en début de mot dans la base de données BRULEX (Content et al., 1990) mais qui seraient certainement hétérosyllabiques en position intervocalique.

Groupes consonantiques	Nombre d'occurrences	Notes	Exemples
/ps/	44	37 de la famille 'psy'	psychologie, pseudonyme
/dʒ/	9	emprunts récents	jazz
/ts/	7	emprunts récents	tsar, tsigane
/ft/	4	même famille	phtisie
/ʃn/	3	même famille	schnock
/mn/	2	même famille	mnémonique

Nous avons vu que la fréquence d'occurrence quelle que soit la position dans les mots n'est pas non plus un critère discriminant permettant de dissocier les séquences tautosyllabiques et hétérosyllabiques puisque les distributions de probabilité d'occurrence se recouvrent largement. La notion de déviance introduite par Dell (1995) pourrait être utile si l'on était en mesure de définir ce qu'est un groupe déviant. En effet, il suffirait de considérer uniquement les groupes attestés dans la langue mais non-déviants comme tautosyllabiques et les autres comme hétérosyllabiques pour constituer nos deux catégories. Cependant, l'absence de critère permettant d'affirmer qu'un groupe est 'légal mais déviant' rend difficile le choix de cette solution. En outre, si l'on admet comme critère de bonne forme en attaque syllabique le fait qu'une séquence soit attestée en position initiale de mot dans la langue, on est en mesure de recenser -même pour les groupes dits non-déviants (Dell, 1995)- un certain nombre de groupes qui se prononceront de manière hétérosyllabique en position intervocalique. Le Tableau 4 présente une sélection de groupes de consonnes qui sont attestés dans la langue mais doivent plutôt être considérés comme hétérosyllabiques. Bien que la plupart de ces mots constituent des

emprunts récents ou soient utilisés assez rarement dans la langue, ils doivent être pris en compte afin de déterminer un critère de distinction entre groupes tautosyllabiques et hétérosyllabiques. Nous faisons l'hypothèse que la combinaison de critères probabilistes et positionnels (le fait qu'un groupe de phonèmes soit prononcé en début de mot) permettra de dissocier clairement ces deux types de groupes de consonnes.

Nous ne recherchons pas ici un critère qui nous permettrait de déterminer de manière binaire -pour autant que ce soit possible- quels groupes constituent des attaques syllabiques *bien formées* ou pas (c'est à dire quel groupe peut être prononcé en attaque syllabique et être considéré comme une *bonne forme* phonologique par un locuteur natif). Il est donc important de distinguer l'objectif que nous nous fixons de celui du phonologue qui se pose des problèmes tout à fait différents. Nous avons insisté, dans le cadre du Chapitre 3, sur les rapports et les distinctions qui peuvent exister entre légalité phonotactique et syllabation. Il est probable que la légalité phonotactique puisse se décrire en réalité comme un continuum allant du 'tout à fait légal' au 'totalement illégal' en attaque de syllabe, une quantité considérable de séquences phonémiques se situant entre ces deux extrêmes. Ce phénomène pourrait être lié à la variabilité qui est couramment observée concernant la syllabation des séquences de phonèmes en position intervocalique et aux difficultés rencontrées en phonologie pour décrire un modèle non-falsifié de ces procédures de syllabation. Notre objectif est d'aboutir à un critère opérationnel qui serait différent de la fréquence d'occurrence calculée indépendamment de la position dans les mots et qui permettrait de catégoriser les groupes de consonnes du français dans les classes tautosyllabique ou hétérosyllabique sur la base d'une mesure observable dans la langue. Cet indice pourrait s'apparenter à un *observable phonologique* et nous permettrait alors de distinguer indices probabilistes et phonologiques dans la constitution de nos expériences. Ce que nous souhaitons obtenir est donc une mesure *observable dans la langue* du taux de bonne formation des groupes de consonnes en attaque syllabique.

3.1. Méthodes d'analyse

Les mêmes catégories d'informations que celles de la section précédente peuvent être extraites pour une analyse de la fréquence des groupes de consonnes en début de mot. Nous nommons les indices dérivés ici des *mesures de fréquence positionnelles*. Par contraste, les données obtenues dans l'analyse précédente seront appelées des *mesures de fréquence brutes*. Seules les probabilités transitionnelles n'ont pas été estimées ici. En effet, il nous semble que si le système cognitif est en mesure d'utiliser des calculs probabilistes dans le cadre des processus

de traitement de la parole, il serait peu efficace de faire reposer ces calculs sur des représentations élaborées (les mots, les syllabes, …). L'objectif que nous poursuivons ici est de proposer *un indice mesurable du taux de bonne formation en attaque syllabique*. Nous n'envisageons pas que ce type d'information (même s'il repose sur une mesure de fréquence) soit utilisé par le système cognitif dans le traitement qu'il effectue sur les stimuli qui lui parviennent. Il serait beaucoup plus économique d'utiliser une information fréquentielle pure. La méthode d'analyse est la même que dans la section précédente. Les scripts AWK ont été modifiés afin de restreindre la recherche des groupes de consonnes (et des séquences Consonne-Voyelle) aux débuts de mots (cf. Annexe 7, Annexe 8 et Annexe 9, pp.XIII-XIV).

3.1.1. Fréquence d'occurrence

A partir des informations fournies par le script d'analyse, nous disposons directement d'une première série de données : la *fréquence d'occurrence* des groupes de consonnes en début de mot, c'est à dire le nombre de mots dans lesquels chaque groupe apparaît en position initiale dans la base de données.

3.1.2. Probabilités pondérées

En second lieu, nous présentons les résultats d'une analyse de la fréquence de ces groupes en position initiale en la pondérant par la fréquence des mots dans lesquels ils apparaissent. De même que dans la section précédente, nous appelons cet indice une *probabilité d'occurrence* (en début de mot) dans la langue. Cette analyse a été limitée aux mots pour lesquels il existe, dans BRULEX (Content et al., 1990), une information sur la fréquence d'usage. La même formule de transformation des données en valeurs logarithmiques a été appliquée pour dériver les valeurs de probabilité d'occurrence en position initiale de mot :

$$10 * \log_{10}(\text{Nombre\_de\_mots} * \text{Fréquence\_cumulée})$$

Cette formule permet à nouveau de visualiser les fréquences des différentes séquences sur une échelle facilitant leur comparaison.

3.2. Résultats

Nous présentons en premier lieu les moyennes obtenues par catégorie de groupe en fonction du mode d'articulation de chaque consonne (occlusive, fricative, nasale, liquide). Nous restreignons notre description aux groupes à initiale occlusive ou fricative. Les données intégrales pour chaque groupe sont présentées en Annexe 10 (p.XV). Dans une seconde étape,

nous étudions la distribution des fréquences individuelles de chaque groupe en fonction de sa catégorie phonétique. De même que dans l'analyse précédente, on observe des corrélations importantes entre les divers indices dérivés de cette analyse. Ainsi, le nombre d'occurrences des groupes de consonnes et la fréquence des mots porteurs sont très fortement corrélés (r = .97). Il est donc à nouveau possible d'affirmer que les groupes de consonnes fréquents en position initiale de mot ont tendance à apparaître dans des mots également fréquents.

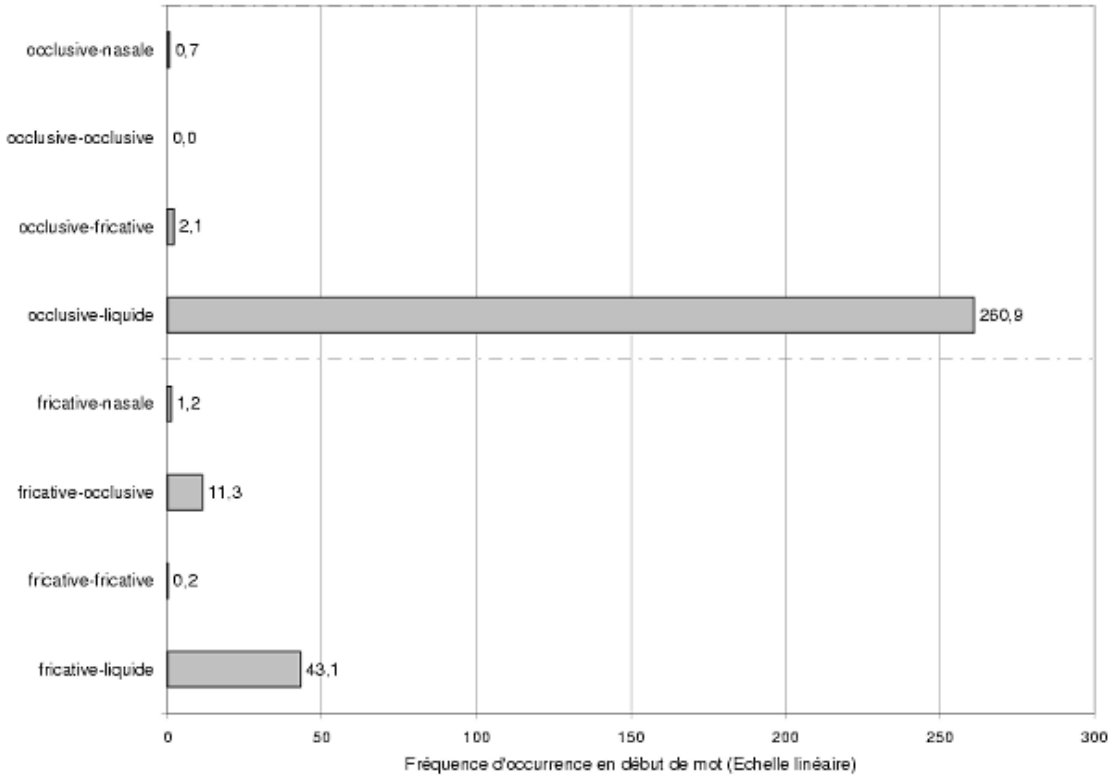

Figure 23 : Nombre moyen d'occurrences en début de mot. Groupes de consonnes classés en fonction de leur mode d'articulation.

3.2.1. Moyennes

La première étape de notre analyse consiste à étudier la fréquence moyenne (selon les 2 indices décrits précédemment) des groupes de consonnes en fonction de leurs caractéristiques phonétiques de mode d'articulation. Les graphiques suivants illustrent les différences de fréquence moyenne en position initiale de mot en fonction de la catégorie phonétique de mode d'articulation. Dans le premier (Figure 23) sont présentées les données moyennes obtenues pour le comptage du nombre de mots dans lesquels chaque groupe consonantique est recensé en position initiale. Nous ne présentons pas dans ce graphique les données obtenues pour les séquences CV (Consonne-Voyelle). Le second graphique reproduit les données de probabilité

d'occurrence en début de mot (fréquence du groupe en position initiale pondérée par la fréquence d'usage des mots, Figure 24). Dans ce graphique nous présentons, outre les résultats observés pour les groupes de consonnes, les données obtenues pour les séquences Consonne-Voyelle.

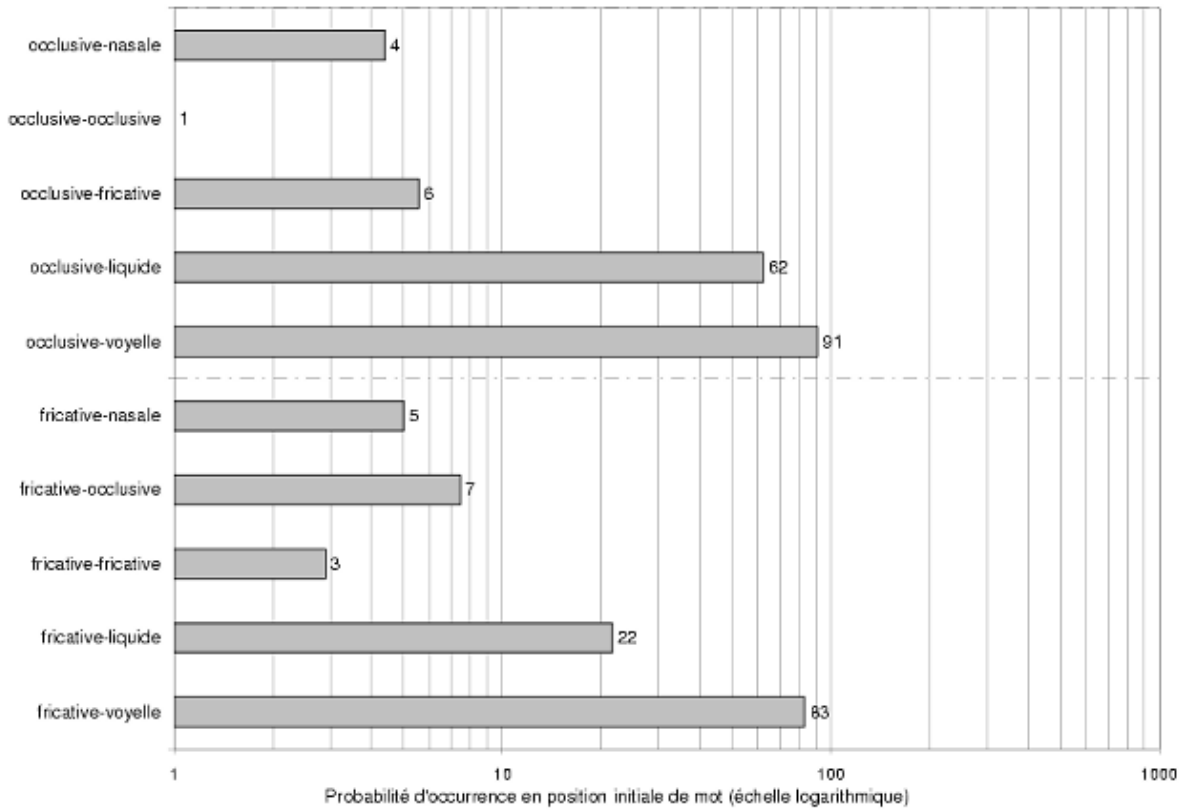

Figure 24 : Probabilité moyenne d'occurrence en début de mot. Groupes de consonnes classés en fonction de leur mode d'articulation.

On observe à nouveau une tendance des groupes tautosyllabiques à être plus fréquents que les groupes hétérosyllabiques. Les séquences C-liquide présentent une probabilité moyenne d'occurrence plus élevée (62 pour les groupes à initiale occlusive et 22 pour ceux à initiale fricative) que les autres catégories (respectivement 1, 4 et 6 pour les groupes à initiale occlusive et 3, 5, 7 pour les groupes à initiale fricative). Les séquences C-voyelle présentent quant à elles des probabilités d'occurrence plus importantes que l'ensemble des autres catégories (respectivement 91 et 83 pour les occlusive-voyelle et les fricative-voyelle). Les mêmes comparaisons *post-hoc* que dans l'analyse précédente ont été effectuées (test de Scheffé) en se restreignant aux groupes constitués d'une fricative ou d'une occlusive à l'initiale et en comparant, pour chaque ensemble, les données obtenues en fonction de la classe phonétique du second phonème. Les seuils de probabilité des diverses comparaisons effectuées à l'aide du test de Scheffé sont présentées dans le Tableau 5.

Cette analyse fournit des résultats tout à fait semblables à ceux qui ont été présentés dans l'étude des fréquences que nous avons conduite sans prendre en compte la position dans les mots. On observe notamment que, dans le cas des groupes à initiale occlusive, l'ensemble des comparaisons effectuées entre groupes *a priori* tautosyllabiques et hétérosyllabiques (l'intérieur du rectangle) permet de conclure à des différences significatives de fréquence. Il émerge en outre une différence significative entre séquences occlusive-voyelle et occlusive-liquide, différence qui avait déjà été observée dans les comparaisons de fréquence d'occurrence effectuées dans la première analyse. En ce qui concerne les groupes à initiale fricative, on observe également une certaine variabilité des seuils de probabilité obtenus. De même que dans la précédente analyse, ces groupes présentent une plus grande variabilité dans la significativité des différences relevées entre séquences tautosyllabiques et hétérosyllabiques. Les mêmes explications peuvent rendre compte de ce phénomène (variabilité plus importante entre structure syllabique déterminée a priori sur la base de leur mode d'articulation et structure syllabique effective en position intervocalique.

Tableau 5 : Seuils de probabilité des tests de Scheffé appliqués à la comparaison de probabilités d'occurrence des groupes de consonnes à initiale occlusive (a) ou fricative (b) apparaissant en position initiale de mot. Les comparaisons sont effectuées par catégorie de groupe en fonction du mode d'articulation des phonèmes. Les seuils de probabilité statistiquement significatifs sont retranscrits en caractères gras et italique. Les intitulés des lignes et des colonnes correspondent au second phonème de la séquence.

a/ Occlusive initiale	liquide	fricative	occlusive	nasale
voyelle	***0.000***	***0.000***	***0.000***	***0.000***
liquide		***0.000***	***0.000***	***0.000***
fricative			0.695	0.999
occlusive				0.935

b/ Fricative initiale	liquide	fricative	occlusive	nasale
voyelle	***0.010***	***0.000***	***0.000***	***0.000***
liquide		***0.016***	0.129	0.098
fricative			0.823	0.995
occlusive				0.990

Dans l'analyse précédente, nous avons vu qu'à ce lien entre fréquence et tautosyllabicité ne correspondait pas nécessairement une distinction nette en termes de fréquence individuelle des groupes de consonnes. Ainsi l'on observe, dans le cadre de l'étude des distributions de fréquence quelle que soit la position des groupes de consonnes dans les mots, des recouvrements

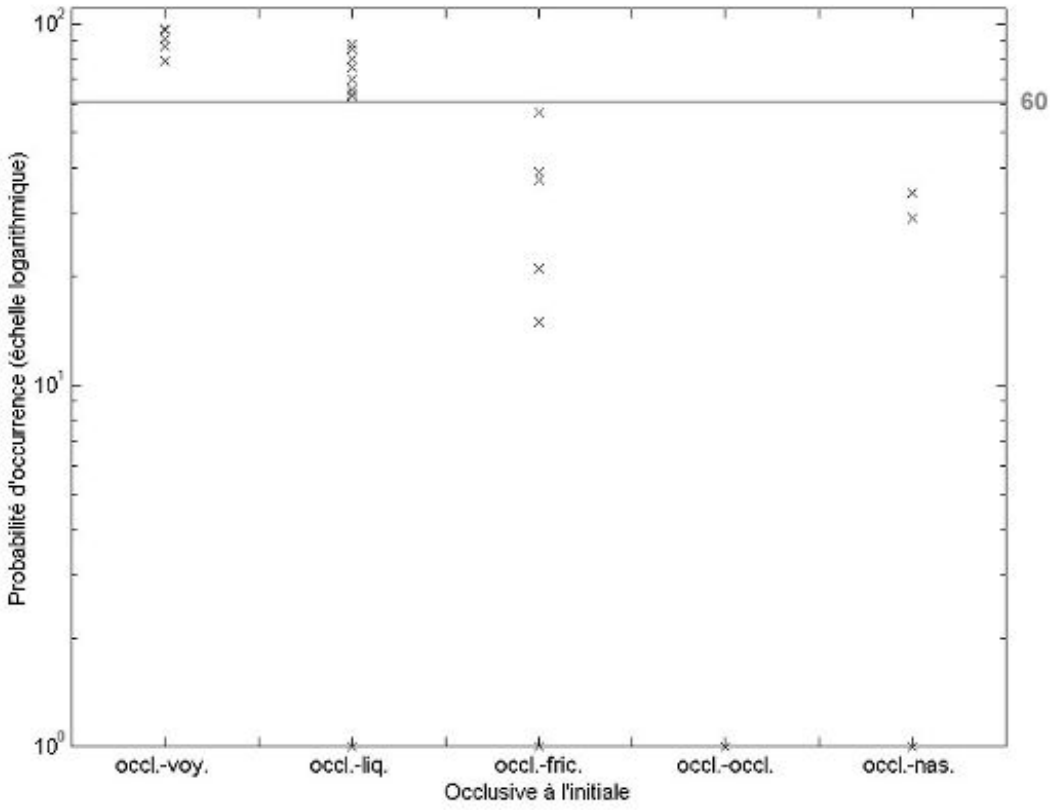

Figure 25 : Distribution des probabilités d'occurrence en début de mot pour les groupes de consonnes à initiale occlusive.

importants des distributions supposées désigner d'une part les groupes tautosyllabiques et, d'autre part, les groupes hétérosyllabiques. Le lien entre tautosyllabicité et fréquence n'étant pas discriminant, nous souhaitons obtenir un indice de bonne forme en position initiale de syllabe qui pourrait à la fois être observable dans un corpus de la langue et fournir une procédure de classification qui soit la plus discriminante possible. Il nous semble par conséquent intéressant d'étudier les distributions de probabilité d'occurrence en début de mot. Nous faisons l'hypothèse que, même si les données moyennes de probabilité d'occurrence fournissent des informations tout à fait semblables à celles qui ont été dérivées de la première analyse distributionnelle, la restriction aux débuts de mots pourrait fournir une illustration totalement différente de la distinction entre groupes tautosyllabiques et hétérosyllabiques si l'on s'intéresse aux distributions des valeurs mesurées. Notamment, nous supposons que cet indice devrait donner lieu à des recouvrements beaucoup moins importants que celui obtenu sans tenir compte de la position des groupes de consonnes dans les mots de la langue. Si c'est effectivement le cas, nous pourrions utiliser cet indice de fréquence en début de mot pour dissocier opérationnellement les groupes de consonnes en fonction de leurs caractéristiques phonologiques.

3.2.2. Distributions

Nous présentons ici les graphiques correspondant aux distributions de probabilité d'occurrence des groupes de consonnes en les classant par catégorie phonétique de mode d'articulation. La Figure 25 illustre les distributions des divers groupes de consonnes commençant par une occlusive (/b/, /d/, /g/, /p/, /t/, /k/). Les données correspondant aux groupes à initiale fricative (/v/, /z/, /ʒ/, /f/, /s/, /ʃ/) sont présentées dans la Figure 26.

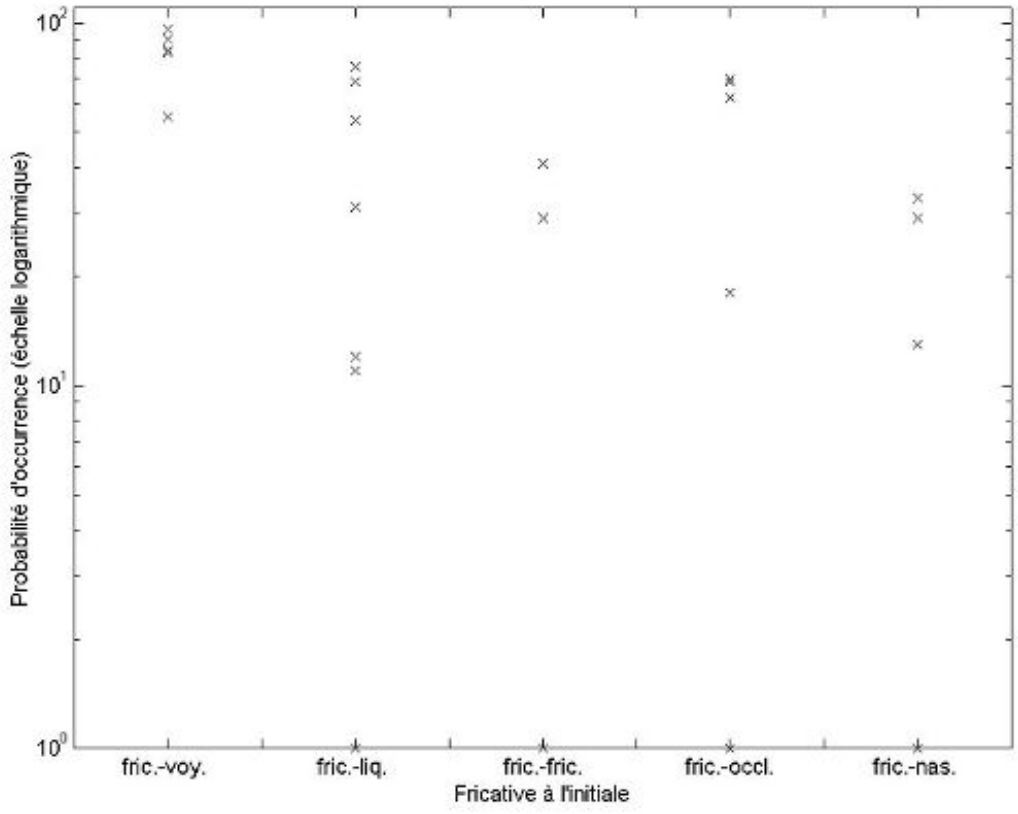

Figure 26 : Distribution des probabilités d'occurrence en position initiale de mot pour les groupes de consonnes à initiale fricative.

Dans le graphique correspondant aux groupes à initiale occlusive, on observe des distributions de probabilité d'occurrence tout à fait différentes de celles obtenues dans la précédente analyse. Contrairement à la distribution des valeurs obtenues sans tenir compte de la position des groupes dans les mots, la restriction du calcul de la fréquence aux groupes apparaissant en position initiale de mot fait émerger une distinction nette entre groupes tautosyllabiques et hétérosyllabiques. Nous avons tracé, dans la Figure 25, une droite horizontale indiquant cette frontière (la valeur discriminante étant approximativement à 60). Les séquences occlusive-voyelle et occlusive-liquide ont presque toutes une probabilité d'occurrence supérieure

à cette valeur (à l'exception des occlusive-liquide coronales /tl/ et /dl/) alors que toutes les autres séquences ont sans exception une valeur de probabilité d'occurrence inférieure à cette limite.

L'analyse des distributions de probabilité d'occurrence des groupes à initiale fricative est nettement plus délicate. On observe des distributions très similaires à celles qui ont été obtenues dans l'analyse globale des probabilités d'occurrence. Cette similarité, de même que la forme très différente des distributions en position initiale de mot selon que l'on s'intéresse aux groupes à initiale occlusive ou fricative, nous conduit à réitérer la remarque d'une variabilité beaucoup plus importante des structures syllabiques possibles dans la catégorie des groupes à initiale fricative. Cette observation est certainement liée à des paramètres phonétiques ou phonologiques. En l'occurrence, il est probable que la classification choisie en termes de mode d'articulation n'est pas la mieux appropriée à une distinction entre groupes tautosyllabiques et hétérosyllabiques.

3.3. Discussion

Que l'on s'intéresse aux probabilités d'occurrence dans la langue quelle que soit la position des séquences dans les mots ou au contraire à celles qui peuvent être mesurées en position initiale de mot, on observe un lien non négligeable entre la fréquence des suites de phonèmes et le statut d'*attaque syllabique bien formée*. Ce lien se manifeste par des différences de fréquence importantes entre diverses catégories de groupes consonantiques catégorisées sur la base du mode d'articulation des phonèmes qui les constituent. Nous avons néanmoins observé que ce lien pouvait n'être pas très stable. Notamment, les données obtenues pour les groupes à initiale fricative ne permettent pas de dissocier clairement, sur la base de la fréquence, les diverses catégories de groupes comparées dans notre analyse. On peut cependant légitimement considérer que la classification adoptée *a priori* n'est pas optimale pour séparer, parmi l'ensemble des groupes à initiale fricative, ceux qui constituent des attaques de syllabe bien formées et les autres, lesquels seraient alors nécessairement hétérosyllabiques en position intervocalique. Il est probable qu'une classification mieux adaptée aurait conduit à des résultats plus convaincants. L'analyse des données concernant les groupes à initiale occlusive est beaucoup plus pertinente. Il semble que la classification adoptée (à l'exception des deux groupes occlusive-liquide coronaux /tl/ et /dl/) soit bien adaptée à la distinction entre groupes tautosyllabiques et groupes hétérosyllabiques parmi les séquences à initiale occlusive. On observe notamment que cette classification permet de dissocier assez clairement deux catégories de groupes de consonnes à l'aide de tests statistiques. Les diverses comparaisons effectuées à l'aide du test de Scheffé permettent en effet de distinguer très clairement les séquences C-voyelle

et C-liquide des 3 autres catégories. Or nous avons supposé dès le départ que ces deux premières catégories devraient en général constituer des attaques syllabiques bien formées. Les 3 autres étaient à notre avis constituées de groupes ne constituant pas réellement des attaques syllabiques légales. La plupart des éléments de ces catégories pouvaient en réalité correspondre à ce que Dell (1995) désigne comme groupes légaux mais déviants. Les données dérivées de l'analyse des groupes à initiale occlusive fournissent donc un argument de poids pour affirmer l'existence d'un lien entre la légalité d'une séquence en attaque de syllabe et sa fréquence dans la langue.

Si ces deux paramètres sont intimement liés, ils ne sont pas confondus. C'est ce qu'illustre l'étude des distributions de probabilité d'occurrence. Ces distributions présentent des recouvrements importants qui mettent en évidence une certaine dissociation entre fréquence et légalité. De fait, nous avons choisi de rechercher un indice qui permettrait de dissocier clairement deux ensembles de groupes de consonnes sur la base de leur légalité en attaque syllabique et qui serait observable dans la langue. L'analyse distributionnelle conduite sur les débuts de mots permet de discriminer nettement les catégories de groupes de consonnes à partir d'une information statistique. Alors que l'étude des distributions de probabilités brutes ne permet pas de dissocier tautosyllabicité et fréquence, l'étude des distributions de probabilité d'occurrence en début de mot fait émerger pour les groupes qui sont attestés dans l'échantillon une discrimination très nette entre d'une part les groupes occlusive-liquide qui sont tous fréquents en début de mot et les 3 autres catégories (occlusive-fricative, occlusive-occlusive et occlusive-nasale) qui sont tous nettement plus rares. L'absence de recouvrement entre les distributions est un indicateur particulièrement utile du lien avec les contraintes phonologiques de la langue. Ainsi, alors que la fréquence brute peut être assimilée à un observable corrélatif mais dissociable des régularités phonologiques, la fréquence d'occurrence en début de mot refléterait fidèlement ces régularités et constituerait une mesure pertinente de la distinction entre séquences légales et illégales en attaque syllabique.

4. Réanalyse des données comportementales

L'analyse distributionnelle conduite sur la base de données BRULEX (Content et al., 1990) fournit les fondements d'une critique raisonnée des données comportementales avancées comme des preuves du recours, de la part des locuteurs natifs, à des processus de segmentation de la parole en mots qui reposeraient en partie sur des connaissances concernant les régularités phonologiques de leur langue. Trois ensembles d'interprétations peuvent en réalité être proposées. Nous procédons en premier lieu à un rapide rappel des données obtenues. Après avoir

décrit à nouveau l'interprétation favorisée par ces auteurs, nous proposons 3 interprétations permettant de prédire ces effets. Aucune d'entre elles ne nécessite le recours à des processus faisant intervenir des connaissances sur les régularités de la langue.

4.1. Rappel des données

Les données expérimentales présentées par (McQueen, 1998) et par Vroomen & De Gelder (1999) ont été mises en évidence à l'aide de deux paradigmes expérimentaux différents : le *word-spotting* et la détection de phonèmes. Les tâches ayant déjà été décrites dans le Chapitre 2 (Sections 2.2.3.2 et 2.2.3.3), nous passons immédiatement au rappel des diverses conditions expérimentales comparées et aux interprétations fournies par les auteurs respectifs de ces études.

4.1.1. Word-spotting

McQueen (1998) a étudié le rôle des contraintes phonotactiques dans les processus de segmentation de la parole en mots avec des locuteurs de langue maternelle néerlandaise. Il reprend la tâche de *word-spotting* introduite par Cutler & Norris (1988) en manipulant le statut du groupe de consonnes médian. Les mots à détecter apparaissent soit en position initiale soit en position finale du non-mot. La légalité phonotactique du groupe consonantique médian est manipulée, celui-ci étant phonotactiquement légal (/vr/, /dr/) ou illégal (*/mr/, */nr/). Lorsque le mot à détecter est en position initiale, le groupe légal donne lieu à un non-alignement de la frontière phonotactique avec la frontière lexicale. Si le mot à détecter est *pil* ('pilule'), le stimulus /pɪlvrem/ -qui contient le groupe consonantique médian légal /vr/- donne lieu à une correspondance entre segmentation phonotactique /pil.vrem/ et segmentation lexicale /pil.vrem/. Lorsque le groupe médian est illégal (par exemple */mr/ dans /pɪlmrem/), il n'y a par contre plus correspondance entre segmentations phonotactique /pilm.rem/ et lexicale /pil.mrem/. Au contraire, si le mot à détecter est en position finale, on observe une relation inverse entre légalité du groupe médian et alignement des frontières. Pour une détection du mot *rok* ('jupe'), la séquence /fimrɒk/ (dans laquelle */mr/ est illégal) donne lieu à un alignement de ces frontières -avec une segmentation phonotactique /fim.rɒk/- alors qu'un groupe de consonnes illégal (par exemple /dr/ dans /fidrɒk/) induit un non-alignement des découpages phonotactique (/fi.drɒk/) et lexical (/fid.rɒk/).

Tableau 6 : Statut du groupe de consonnes médian dans l'expérience de McQueen (1998) en fonction de l'alignement entre frontières phonotactique et lexicale.

	non-alignement	**alignement**
Position initiale	illégal /pɪlmrem/	légal /pɪlvrem/
Position finale	légal /fidrɒk/	illégal /fimrɒk/

Le Tableau 6 reprend les informations présentées dans le Tableau 2 du Chapitre 2 et résume l'agencement du lien entre légalité et alignement dans cette expérience. McQueen (1998) observe qu'une absence d'alignement entre les frontières phonotactique et lexicale donne lieu à des taux d'erreur significativement plus importants que la condition d'alignement. Les mots sont détectés plus facilement lorsque frontière phonotactique et lexicale sont alignées que lorsqu'elles sont discordantes.

4.1.2. Détection de phonèmes

Vroomen & De Gelder (1999) présentent à des locuteurs néerlandais des phrases dans lesquelles les participants doivent détecter un phonème-cible. Selon les conditions, le phonème-cible peut être prononcé en fin de syllabe comme dans la séquence :

'de.boo**t**.die.ge.zon.ke.nis'

dans laquelle la cible est le phonème /t/. Dans cette situation, on observe qu'il y a correspondance entre la frontière syllabique qui sépare /but/ et /di/ et la frontière lexicale qui sépare *boot* ('bateau') et *die* ('qui'). Dans l'autre condition expérimentale, le phonème-cible apparaît en position d'attaque syllabique comme dans l'énoncé :

'de.boo.**t**is.ge.zon.ken'

On observe ici une discordance entre la frontière syllabique (laquelle sépare /bu/ de /tiz/) et la frontière lexicale qui se situe après *boot* ('bateau'). Dans la première condition, la consonne cible (une occlusive) est suivie d'une consonne également occlusive qui induit un alignement des frontières syllabique et lexicale. Dans la condition de non-alignement des frontières, cette consonne cible est au contraire suivie d'une voyelle. Les auteurs observent également un effet de la relation entre frontière syllabique et frontière lexicale. Cet effet se manifeste cependant ici dans les temps de réaction mais pas dans les taux d'erreur : les latences de détection de phonème

sont plus courtes lorsque les frontières syllabique et lexicale sont alignées (c'est à dire lorsque le phonème-cible est en position de coda syllabique) que lorsqu'elles ne le sont pas (phonème en attaque syllabique).

4.1.3. Interprétations

Les auteurs aboutissent à des conclusions similaires à l'issue de l'analyse des données expérimentales. Dans les deux tâches, l'identification du mot pertinent serait influencée par le découpage syllabique ou phonotactique de la chaîne de phonèmes. Lorsque ce découpage concorde avec la segmentation lexicale adéquate, les processus de reconnaissance lexicale seraient facilités, ceci se manifestant par un raccourcissement des délais de réalisation de la tâche ou par un accroissement des taux de réponse correcte. De fait, ces données constitueraient une mise en évidence du recours à des connaissances sur les régularités phonologiques de la langue dans les processus de segmentation lexicale qui sont mis en œuvre au cours du traitement du signal de parole. Nous avons cependant mis en évidence un lien entre tendance des séquences de phonèmes à se regrouper à l'attaque syllabique et fréquence d'occurrence dans la langue.

4.2. Interprétations concurrentes

Il nous semble possible, sur la base du lien observé entre tautosyllabicité et fréquence, de proposer au moins 3 classes d'interprétations de ces effets. Chacune de ces trois interprétations est directement dérivée de la correspondance observée entre légalité phonologique des séquences de phonèmes et fréquence d'occurrence de ces mêmes suites dans les mots de la langue. Nous avons montré dans l'analyse distributionnelle qui a été conduite sur la base de données BRULEX (Content et al., 1990) que, dans un lexique français, les groupes légaux sont utilisés moins fréquemment dans les mots de la langue que les groupes illégaux. Il est probable, si ce lien est effectivement une conséquence des phénomènes de régularité phonologique, que les mêmes résultats seraient obtenus en néerlandais. Dans l'expérience de McQueen (1998), les groupes identifiés comme légaux sont des séquences occlusive-liquide ou fricative liquide (parmi lesquelles /dr/, /vr/, /fl/…) alors que les groupes illégaux sont des séquences nasale-liquide (les deux suites /mr/ et /nr/) ou la suite /tl/. La même classification pourrait en fait être adoptée en français. Ainsi, les suites illégales utilisées par McQueen (1998) n'apparaissent jamais au début des mots de la langue française alors que les suites occlusive-liquide non-coronales et fricative-liquide sont courantes dans cette position. Or nous avons mis en évidence des différences de fréquence considérables entre ces deux catégories de groupes. La même critique peut s'appliquer

aux données de Vroomen & De Gelder (1999). En français, les séquences CV sont nettement plus fréquentes que les suites de consonnes hétérosyllabiques. Il semble donc y avoir une confusion, dans les variables manipulées par ces auteurs, entre légalité / tautosyllabicité et fréquence d'occurrence dans la langue. Cette observation nous conduit à proposer trois interprétations alternatives des données expérimentales présentées.

4.2.1. Un phénomène de sélection lexicale sérielle ?

La première interprétation que nous proposons est inspirée du modèle COHORT (Marslen-Wilson & Welsh, 1978; Marslen-Wilson, 1987) de sélection lexicale que nous avons décrit dans le Chapitre 2. Selon ce modèle, le processus de reconnaissance des mots dans la modalité auditive génère l'activation d'un ensemble de candidats à partir des premières périodes acoustiques du stimulus. Les informations acoustiques ultérieures donnent lieu à une sélection lexicale consistant à supprimer de l'ensemble initial de mots activés (*la cohorte*) les candidats qui ne sont plus appariés avec l'entrée acoustique. Nous avons décrit la possibilité introduite par ce modèle de reconnaître un mot avant même sa fin acoustique. Nous avons également cité plusieurs études computationnelles et comportementales qui mettent en évidence la possibilité qu'un mot soit reconnu après sa fin acoustique s'il existe dans le lexique d'autres mots au début desquels le mot à identifier est enchâssé (Luce, 1986; Frauenfelder & Peeters, 1990; Grosjean, 1985). Ces deux phénomènes sont à mettre en rapport avec les données auxquelles nous nous intéressons ici. Dans les expériences qui ont permis de mettre en évidence un rôle des régularités phonologiques dans les processus de segmentation lexicale, on a contrôlé les caractéristiques des groupes de consonnes comparés en fonction de leur regroupement éventuel à l'attaque de syllabe. Cette distinction entre séquences de phonèmes regroupées à l'attaque syllabique et séquences nécessitant l'insertion d'une frontière phonologique entre les deux phonèmes correspond dans la langue à une distinction en termes de fréquence d'occurrence. Les groupes de phonèmes qui se regroupent à l'attaque de syllabe sont plus fréquemment utilisés dans les mots de la langue. Or, si les séquences légales sont plus souvent utilisées dans les mots de la langue, il est légitime de prédire qu'une séquence CVC donnée aura plus de chances de constituer le début d'un mot pour lequel la consonne finale de la séquence CVC et le phonème suivant (consonne ou voyelle) constitueront une suite fréquente (donc une attaque de syllabe bien formée) qu'une suite rare (illégale ou hétérosyllabique). Ainsi, le monosyllabe 'vague' constitue le début de 14 mots dans la langue. Pour 2 d'entre eux la consonne /m/ fait suite à la séquence /vag/, créant ainsi une séquence hétérosyllabique /gm/. Par contre, les 12 autres mots font suivre /vag/ par une voyelle, induisant alors la consonne /g/ à se retrouver à l'attaque de la syllabe suivante. Une recherche

rapide dans la base de données BRULEX (Content et al., 1990) conduit à cette même observation pour nombre de séquences de 3 phonèmes ayant cette structure CVC. Certaines suites donnent lieu à un déséquilibre encore plus important entre les rattachements possibles de la consonne finale. Une recherche des mots possibles commençant par /bat/ fait ressortir 36 mots au début desquels cette séquence est enchâssée. Trois d'entre eux font suivre le /t/ d'un /l/ ou d'un /m/, donnant ainsi lieu à une séquence hétérosyllabique. L'ensemble des autres mots possibles fait suivre le /t/ d'un /r/ ou d'une voyelle. Lorsque l'on traite une suite CVC dans une tâche d'identification de mot, il est par conséquent très probable que le phonème suivant sera regroupé à l'attaque de syllabe avec la consonne finale de la séquence CVC. Cette probabilité est valable aussi bien dans une tâche de détection de phonème (qui induit souvent le recours à l'identification du mot porteur) que dans une tâche de *word-spotting* (qui l'induit nécessairement). Supposons par exemple que l'on souhaite conduire une expérience avec le paradigme de *word-spotting*, expérience dans laquelle on chercherait à mettre en évidence le rôle des régularités phonologiques dans les processus de segmentation de la parole en mots. L'une des cibles que les participants devront détecter est le mot 'bague'. Dans la condition d'alignement entre segmentation phonologique et segmentation lexicale, la consonne /g/ est suivie d'un /n/. Dans l'autre condition expérimentale, on fait suivre cette même consonne du phonème /a/. Lorsque l'on a traité la suite /vag/, il reste dans le lexique un ensemble de candidats lexicaux constitués du mot 'bague' en position initiale. Ces mots sont au nombre de 14 dans la base de données BRULEX (Content et al., 1990). Il n'est donc pas possible d'identifier le mot 'bague' à partir de la sélection des candidats lexicaux présents dans la cohorte puisqu'ils sont trop nombreux pour cela. Huit d'entre eux commencent par /baga/ alors qu'un seul commence par /bagn/. Une fois le phonème suivant traité et utilisé pour procéder à la suppression des candidats qui ne sont plus appariés avec l'entrée sensorielle, la quantité de candidats maintenus dans la cohorte en fonction de la condition expérimentale subit un déséquilibre. Dans la condition de légalité phonotactique-tautosyllabicité, il faudra sélectionner le mot adéquat 'bague' parmi une quantité plus importante de candidats (Les huit mots commençant par /baga/ plus le mot 'bague') que dans la condition d'illégalité phonotactique-hétérosyllabicité dans laquelle seul le mot 'baguenauder' sera maintenu dans la cohorte. Bien évidemment, nous avons spécifiquement choisi cet exemple afin de mettre en évidence une possibilité de déséquilibre lexical dans la constitution d'un matériel destiné à mettre en évidence des effets non-lexicaux. Il est cependant nécessaire d'admettre que si les phonèmes qui font suite à la consonne finale du mot cible (dans la tâche de *word-spotting*) ou du

mot porteur (dans la tâche de détection de phonèmes) sont choisis au hasard, la probabilité de choisir des phonèmes donnant lieu à une quantité plus importante de candidats lexicaux dans la condition de légalité phonotactique que dans la condition d'illégalité phonotactique est grande. A moins d'être averti de ce risque et de contrôler ce paramètre, on risque d'observer des effets lexicaux qui pourraient passer pour des effets phonologiques.

4.2.2. Segmentation probabiliste

Plusieurs études récentes ont conduit à affirmer que le système de traitement du langage serait en mesure d'utiliser des informations probabilistes pour traiter les informations de l'environnement linguistique (Brent, 1996 ; Brent & Cartwright, 1996). Ces procédures statistiques reposeraient sur des mécanismes généraux de traitement (Aslin, Saffran, & Newport, 1998; Saffran, Johnson, Aslin, & Newport, 1999) mais pourraient être appliquées à des processus dédiés au traitement du langage. Il est possible d'observer le recours à des informations probabilistes dès les premiers mois (Saffran, Aslin et al., 1996). Dans le cadre de l'acquisition du langage, ces mécanismes auraient pour fonction de déclencher la constitution d'un lexique initial qui permettrait par la suite de développer des processus de reconnaissance des mots fondés par exemple sur les compétitions entre candidats lexicaux (McClelland & Elman, 1986). A l'âge adulte, ils continueraient d'être utilisés et permettraient de prédire les frontières entre les mots (Saffran, Newport et al., 1996 ; Brent, 1997). Toute séquence de phonèmes rencontrée dans la chaîne de parole serait codée en termes de probabilité transitionnelle d'apparition. Une séquence très fréquente conduirait le système à supposer que cette séquence fait partie d'un mot unique. Il aurait alors tendance à la regrouper afin de chercher dans son lexique des candidats contenant cette séquence. Une séquence très rare ayant peu de chances d'exister effectivement dans un mot de la langue, les phonèmes la constituant seraient au contraire considérés comme faisant partie de deux mots différents et le système aurait tendance à chercher des séquences de mots en insérant une frontière lexicale entre les phonèmes de la séquence rare. Du fait du lien que nous avons mis en évidence entre fréquence et légalité, il est en réalité difficile d'affirmer que les données avancées comme des preuves du recours à des connaissances sur les régularités de la langue ne constituent pas au contraire un reflet de l'utilisation de calculs probabilistes consistant à séparer les séquences de phonèmes. Des travaux portant sur le traitement langagier chez le jeune enfant et le nourrisson ont conduit à affirmer que ces deux catégories d'informations sont disponibles pour les processus de traitement de la parole. Dès les premiers mois de la vie, les enfants auraient intégré des connaissances concernant la légalité phonotactique des séquences de phonèmes dans leur langue maternelle (Jusczyk, Luce, & Charles-Luce, 1994; Friederici & Wessels, 1993). Ils présentent par ailleurs une tendance à considérer comme plus familières des

séquences fréquentes que des séquences rares dans leur langue (Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993). En fait, il n'est pas sûr que ces deux catégories de processus ne constituent pas deux facettes d'un même phénomène. En effet, puisque fréquence et légalité phonotactique sont liées, il est possible que les comportements observés chez l'enfant lorsqu'il entend des suites de phonèmes phonotactiquement illégales soient déterminés par des différences de fréquence plus que par des différences de légalité phonotactique impliquant le recours à une base de connaissances linguistiques. A l'inverse, on peut tout aussi bien faire l'hypothèse que les effets probabilistes observés consistant à trouver plus familière une séquence fréquente dans la langue pourraient être déterminés par des représentations linguistiques qui permettraient à l'enfant de considérer des séquences fréquentes comme *plus légales* que des séquences rares.

Nous sommes confrontés au même dilemme dans l'étude des processus de segmentation lexicale chez l'adulte. Si l'on met en évidence un effet de la légalité phonotactique ou de la structure syllabique sur les temps de réaction observés, on peut légitimement se demander si les effets interprétés en termes phonologiques ne sont pas tout simplement déterminés par des représentations de type probabiliste. Le système de traitement de la parole pourrait ainsi localiser des groupes de phonèmes rares et faire l'hypothèse d'une frontière lexicale entre les phonèmes constituant des séquences rares dans la langue. Il ne serait donc pas nécessaire, même si les processus semblent très similaires, d'avoir recours à un modèle dans lequel les locuteurs feraient appel à des connaissances sur les régularités linguistiques de leur langue.

4.2.3. Fréquence et compétitions lexicales

Les deux interprétations précédentes font appel à deux types de processus différents. Selon la première, la tendance à répondre plus lentement lorsque l'on entend une séquence pour laquelle il existe un non-alignement entre frontière phonotactique / syllabique et frontière lexicale pourrait s'expliquer par le maintien d'une quantité plus importante de candidats lexicaux dans la cohorte (Marslen-Wilson, 1987; Marslen-Wilson & Welsh, 1978) lorsque le phonème qui suit la consonne finale du mot est en cours de traitement. Ce phénomène implique un processus de traitement séquentiel consistant à sélectionner progressivement les candidats les mieux appariés avec l'entrée acoustique. Cette interprétation suppose que l'effet observé serait localisé à un niveau lexical de traitement. Une seconde interprétation alternative des données obtenues consiste à proposer que cette différence de statut des séquences de phonèmes manipulées dans les expériences impliquerait une segmentation fondée sur des différences de fréquence des séquences plus que sur des différences de statut phonologique. L'effet observé serait alors également interprétable en termes pré-lexicaux : la présence d'une séquence rare (de même que

celle d'une séquence illégale ou hétérosyllabique) conduirait à supposer une frontière lexicale entre les phonèmes qui la constituent.

Une troisième interprétation nous semble envisageable. Dans le cadre d'un modèle de compétitions lexicales comme TRACE (McClelland & Elman, 1986) ou SHORTLIST (Norris, 1994), les candidats lexicaux activés n'ont pas besoin d'être alignés avec le début du mot comme c'est le cas dans COHORT (Marslen-Wilson, 1987; Marslen-Wilson & Welsh, 1978). On parle d'alignement exhaustif des activations lexicales. Dans ces modèles, des processus d'activation interactive entre les divers candidats activés sont implémentés (cf. Chapitre 2). Le traitement effectué sur l'entrée acoustique n'est donc pas intégralement sériel. Il n'est par conséquent pas totalement nécessaire de coder les phonèmes les uns après les autres et de sélectionner progressivement le candidat adéquat en respectant l'axe temporel de prononciation des phonèmes. L'interprétation lexicale proposée dans la section 4.2.1 ne serait donc pas envisageable avec cette classe de modèles. Par contre, si une séquence de phonèmes est fréquente ceci signifie qu'elle apparaît dans une quantité importante de mots de la langue. L'occurrence d'une séquence fréquente conduirait donc à provoquer l'activation d'un grand nombre de candidats lexicaux, quelle que soit la position de la séquence dans les mots. La procédure de compétition entre les candidats activés impliquerait par conséquent une quantité d'unités lexicales beaucoup plus importante que dans le cas de l'occurrence d'une séquence de

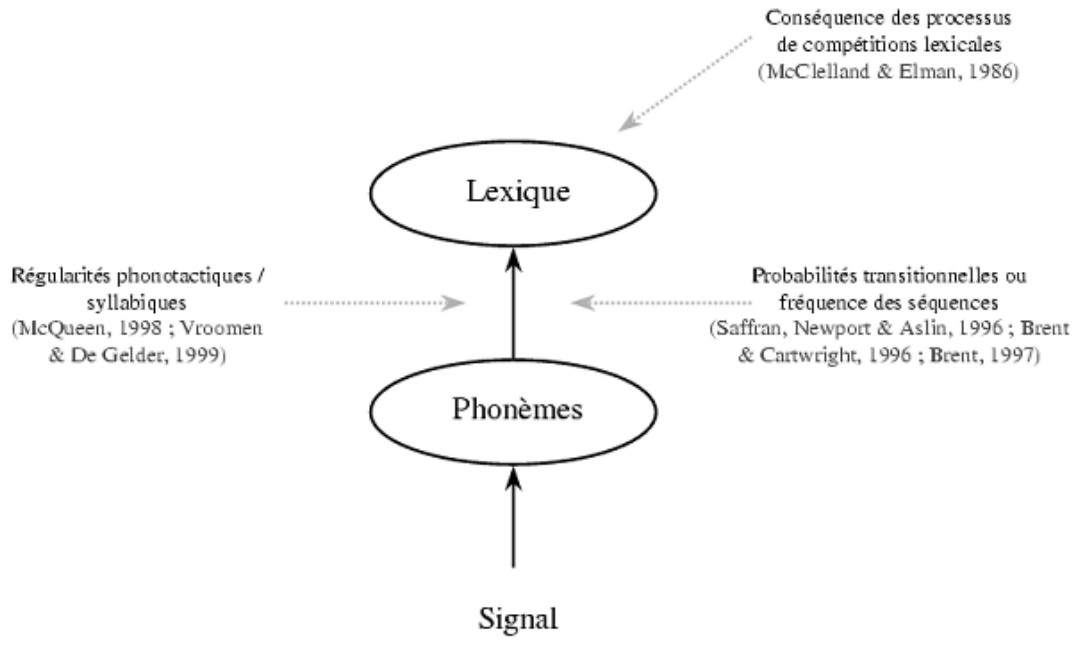

Figure 27 : Représentation graphique des interprétations reposant sur la fréquence des groupes de consonnes (effet prélexical pour une segmentation probabiliste ; effet lexical pour les phénomènes de compétitions lexicales.

phonèmes rare. Il faudrait alors plus de temps au réseau pour aboutir à une stabilisation des niveaux d'activation et, de fait, à la sélection du candidat approprié.

Cette dernière interprétation doit déjà être mise à l'épreuve des données qui sont présentées par McQueen (1998). En effet, lorsqu'il étudie l'effet de l'alignement entre frontières phonotactique et lexicale avec le mot en position initiale, la condition d'alignement (la plus facile) correspond à une séquence phonotactiquement légale (qui, en raison du nombre de mots de la langue dans lesquels elle apparaît, générerait l'activation d'une importante quantité de candidats lexicaux). Or, si notre interprétation était valide, on devrait observer l'effet inverse puisque la séquence légale devrait alors rendre la tâche plus difficile et donner lieu à des taux d'erreur plus importants. La même remarque peut-être faite pour l'interprétation en termes de sélection lexicale sérielle. On notera néanmoins que les effets obtenus par McQueen (1998) ne semblent pas totalement liées au contexte phonotactique mais pourraient s'expliquer par des caractéristiques acoustiques des stimuli ; lesquels seraient prononcés différemment en contexte légal et illégal. Dans une seconde expérience, McQueen (1998) a extrait des stimuli utilisés dans la tâche de *word-spotting* le segment acoustique correspondant au mot et a conduit une tâche de décision lexicale avec ce matériel en étudiant à nouveau l'effet de l'alignement. Si l'effet original est réellement lié au contexte phonotactique, il devrait alors disparaître puisqu'aucun contexte n'est adjoint aux stimuli expérimentaux. C'est effectivement le cas pour les mots qui ont été prononcés en position finale de non-mot. L'effet d'alignement disparaît lorsqu'on réalise une tâche de décision lexicale sur la partie qui correspond au mot. Par contre, l'effet d'alignement se maintient avec les mots qui ont été prononcés à l'origine en position initiale. Il est donc probable que l'effet d'alignement obtenu dans la tâche de *word-spotting* était dans cette situation la conséquence de différences propres aux caractéristiques intrinsèques de la partie lexicale et n'était pas lié au contexte phonotactique. McQueen (1998) conduit une analyse de covariance (ANCOVA) sur les données de *word-spotting* avec comme covariable les taux d'erreurs obtenus dans la tâche de décision lexicale et observe que l'effet d'alignement observé en *word-spotting* se maintient. Il en conclut logiquement que l'effet observé dans la tâche de décision lexicale ne suffit pas à expliquer intégralement l'effet obtenu en *word-spotting*. L'interprétation initiale serait donc quand même valide. En réalité, la tâche de *word-spotting* est beaucoup plus difficile à effectuer que celle de décision lexicale. Lorsque l'on conduit une expérience avec la tâche de *word-spotting*, on observe des taux d'erreurs et des temps de réaction moyens beaucoup plus importants qu'en décision lexicale. Les participants fournissent aussi des appréciations sur la tâche qui montrent qu'elle est particulièrement difficile. Il suffit d'observer les taux d'erreurs moyens produits par les participants de l'expérience de McQueen (1998) qui

vont de 20% (alignement) à 60% (non-alignement) pour s'en persuader. Dans une tâche de décision lexicale, on admettrait difficilement un taux d'erreur supérieur à 10% ! Il nous semble illégitime d'introduire comme covariable d'une ANCOVA des données similaires (taux d'erreur) ayant été obtenues avec une tâche qui s'avère beaucoup plus facile. En effet, si l'on compare les mêmes effets dans une tâche très facile et dans une tâche relativement difficile, on peut s'attendre à ce que les effets observés dans la première soient plus faibles que dans la seconde. On peut également s'attendre à obtenir une variabilité beaucoup plus importante dans la tâche difficile que dans la tâche facile. Or une ANCOVA consiste à évaluer la part de variance d'une variable que ne peut pas expliquer une seconde variable (la covariable). Si la non-correspondance des données est liée à une différence dans la difficulté intrinsèque des tâches et pas à l'intervention de deux catégories de facteurs dont l'un n'intervient pas dans la seconde tâche, il est impossible d'utiliser les données de l'ANCOVA pour tenter d'interpréter les données obtenues dans la première tâche.

Les données observées pour les mots en position initiale ne sont donc pas assez fiables pour permettre de réfuter ces hypothèses lexicales et il est nécessaire de répliquer les expériences afin de confronter les interprétations proposées par McQueen (1998) et par Vroomen & De Gelder (1999) à des données expérimentales permettant de contrôler les diverses variables qui peuvent intervenir dans l'émergence de ces effets.

Résumé

Une analyse distributionnelle a été conduite sur la base de données lexicales informatisée de la langue française BRULEX (Content et al., 1990). Cette analyse permet de mettre en évidence un lien étroit entre légalité phonotactique et fréquence d'occurrence des séquences de phonèmes dans la langue. Cette observation nous a conduit à réanalyser les données comportementales présentées par McQueen (1998) et par Vroomen & De Gelder (1999) en mettant en évidence une confusion entre le statut phonologique des types de séquences comparés et la fréquence des séquences dans la langue. Ce lien entre structure phonologique des groupes de consonnes et fréquence d'occurrence conduit à proposer 3 interprétations différentes des données expérimentales. Aucune de ces interprétations ne nécessite d'avoir recours à des connaissances sur la structure phonologique de la langue. Il est donc nécessaire d'approfondir l'étude du rôle des régularités phonologiques de la langue dans les processus de segmentation de la parole en mots afin

d’approfondir la compréhension de ces effets et des processus qui les sous-tendent.