

# PHONOTACTIQUE PREDICTIVE ET ALIGNEMENT AUTOMATIQUE : APPLICATION AU CORPUS MARSEC ET PERSPECTIVES

Cyril Auran, Caroline Bouzon

auran@lpl.univ-aix.fr; bouzon@lpl.univ-aix.fr

## Résumé

*Cet article présente la méthodologie employée lors de la constitution du corpus aligné (phonèmes, constituants syllabiques, syllabes, mots, pieds accentuels et unités intonatives) Aix-MARSEC. Après avoir défini les concepts d'alignement et de granularité, cet article détaille, dans sa partie centrale, les trois phases menant à la transcription phonétique alignée du corpus : phonétisation automatique brute par dictionnaire, optimisation par règles d'élision et alignement automatique par "force Viterbi" de la transcription optimisée. Après évaluation de la phonétisation optimisée et de l'alignement final, des perspectives d'amélioration de ces deux composantes sont proposées.*

Mots-clés : phonétisation, règles d'élision, contraintes phonotactiques, alignement, corpus.

## Abstract

*This paper presents the methodology used during the generation of Aix-MARSEC, a multi-level aligned corpus (phonemes, sub-syllabic components, syllables, words, stress feet and intonation units). After defining the concepts of alignment and granularity, the main part of this paper details the three phases leading to the aligned phonetic transcription of the corpus: raw dictionary-based grapheme-to-phoneme conversion, optimisation by phoneme elision rules and automatic forced Viterbi alignment of the optimised transcription. After the evaluation of both the optimised transcription and the final alignment, perspectives of improvement of these two components are suggested.*

Keywords : grapheme-to-phoneme conversion, elision-prediction rules, phonotactic constraints, alignment, corpus.

## Introduction

La recherche linguistique contemporaine, quelles que soient ses orientations théoriques (cognitive, pathologique, linguistique texto-centrée, etc.), semble s'orienter vers une prise en compte plus étendue d'exemples avérés, que ce soit pour la constitution de stimuli naturels

dans le cadre de tests psycholinguistiques ou pour des analyses formelles à proprement parler (Cornish, 1999). Dans cette optique, le travail sur corpus constitue donc une base fondamentale. Les études portant sur l'oral ont ainsi grand intérêt à s'appuyer sur de grands corpus de parole, dont le style répondra aux exigences de l'approche choisie.

Pour certains courants de la phonétique et de la phonologie plus particulièrement, de grandes bases de données alignées à différents niveaux de constituance représentent un objet d'étude précieux. Rares sont cependant, quelle que soit la langue étudiée, les corpus qui offrent un alignement au niveau du phonème, du constituant syllabique ou même de la syllabe. Ceci peut s'expliquer par le coût important que représente la phase d'alignement tant en niveau d'expertise requis qu'en durée de traitement manuel ; on estime en effet qu'une minute de parole nécessite environ seize heures de travail expert pour une transcription et un alignement au niveau du phonème (Di Cristo et Hirst, 1997). On pourrait alors objecter que des traitements existent qui accomplissent une telle tâche de manière totalement automatique et en un temps nettement plus court (le rapport est de l'ordre de 1 pour 660 dans le cas du corpus MARSEC que nous utilisons dans ce travail). Une évaluation de tels procédés de phonétisation-alignement montre cependant que la transcription finale ainsi obtenue peut parfois diverger de manière significative d'une transcription manuelle experte.

Cet article présente une méthodologie alternative destinée à la phonétisation et à l'alignement de grands corpus oraux, notamment d'anglais britannique, que nous avons appliquée au corpus MARSEC dans le cadre du projet Aix-MARSEC. La première partie approfondit le concept d'alignement ainsi que la notion de finesse ou granularité d'alignement. La deuxième partie de ce travail décrit les différentes phases d'évolution du corpus sur lequel se fonde cette étude, du corpus SEC d'origine au corpus final Aix-MARSEC en passant par le corpus MARSEC. La troisième partie détaille la phase de phonétisation à proprement parler ; on présentera ainsi les différentes approches envisageables dans cette perspective avant de préciser les principes, les traitements et les problèmes spécifiques à la méthode de phonétisation par dictionnaires choisie. La quatrième partie présente et évalue la phase suivante d'optimisation de la transcription phonématique à l'aide de règles d'émission de phonèmes<sup>1</sup>. Plus précisément, cette démarche étant fondée sur la définition majoritairement phonotactique de contextes pertinents, nous définirons la "phonotactique prédictive" comme l'application de ce type de règles dans le cadre de la phonétisation optimisée de parole continue. La cinquième partie porte sur deux types de méthodes d'alignement automatique

---

<sup>1</sup> On pourra se référer à Kipp *et al.* (1996) pour la description d'une méthodologie similaire (bien qu'appliquée à l'allemand) à celle présentée dans les troisième et quatrième parties de ce travail.

disponibles pour le traitement Aix-MARSEC et nous amène à évaluer la précision de la transcription alignée ainsi obtenue. La sixième et dernière partie de ce travail évoque les perspectives de traitement envisagées dans le but d'améliorer encore la qualité de la transcription et de l'alignement ; on mentionnera notamment la formulation de nouvelles règles d'élimination de phonèmes et l'utilisation d'algorithmes itératifs de traitement automatique.

### **1. Précisions terminologiques : granularité d'alignement**

Le concept d'alignement est récurrent dans nombre d'analyses dans les sciences du langage, sans pour autant être défini de manière consensuelle et explicite. C'est pour cette raison qu'il nous semble important d'approfondir notre conception de cette notion dans ce domaine.

Fondamentalement, nous considérerons que l'alignement consiste à établir des relations associant des ensembles d'unités linguistiques à des ensembles d'unités métalinguistiques (Culioli, 1990). Cette conception est à mettre en parallèle avec le concept d'annotation linguistique tel que le définissent Bird & Liberman (2001):

*« 'Linguistic annotation' covers any descriptive or analytic notations applied to raw language data. The basic data may be in the form of time functions — audio, video and/or physiological recordings — or it may be textual. » (2001 : 23)*

La deuxième partie de cette définition insiste plus particulièrement sur l'importance du format des données linguistiques. En conséquence, nous postulerons que nous avons affaire à une annotation avec alignement (nous dirons « un alignement ») lorsque les données linguistiques sont disponibles sous la forme d'enregistrements et que l'annotation est de type  $f(t)$  (fonction du temps) ; dans le cas contraire, nous sommes alors confrontés à une annotation sans alignement (ou « annotation simple »).

De manière évidente, les données linguistiques peuvent être analysées à différents niveaux de leur structure postulée. On observe ce phénomène pour des disciplines assimilables au domaine général de la phonétique/phonologie (groupe A ci-dessous), tout autant que pour des disciplines non spécifiques à l'aspect oral de la parole et du langage (qualifions-les « linguistiques et pragmatiques »), réunies dans le groupe B ci-dessous :

**GROUPE A :**

- une analyse acoustique (cf. Rabiner, 1984), mais aussi peut-être phonétique/phonologique articulatoire (cf. Browman & Goldstein, 1989), pourra mettre en jeu des unités infra-segmentales (respectivement de type «état » ou «geste articulatoire ») ;
- une analyse phonétique/phonologique segmentale fera appel à des unités de types « phone » ou « phonème » ;
- une analyse prosodique pourra s'appuyer sur des unités aussi bien infra-segmentales (comme dans le cas d'un accent mélodique analysé comme la succession de deux tons sur un segment vocalique) que supra-segmentales, comme par exemple les unités rythmiques élémentaires ou complexes (de type « syllabe », « more », « accent », « pied accentuel », « unité rythmique étroite », « anacrouse », *etc.*) ;

#### GROUPE B :

- une analyse morphosyntaxique aura recours à des unités élémentaires de type « morphème », regroupées en unités complexes de type « syntagme / proposition » ;
- une analyse discursive pragmatique s'appuiera par exemple (cf. le Modèle Genevois d'Analyse du Discours, Roulet *et al.*, 1985) sur des éléments fondamentaux de type « acte », regroupés en unités de rang supérieur, de type « intervention », « échange », *etc.*

Certains niveaux de l'analyse, bien entendu, autorisent une certaine forme de récursivité dans les processus de structuration des unités constituantes : une unité de rang  $n$  pourra alors contenir non seulement des unités de rang  $n-1$ , mais aussi de rang  $n$  ; on pensera notamment aux cas classiques de récursivité en syntaxe, mais aussi à des analyses (telles que Hyman *et al.*, 1987 ou plus récemment dans les travaux de Di Cristo) mettant en cause le principe de hiérarchie stricte ou « *Strict Layer Hypothesis* » en prosodie (cf. Selkirk, 1984 et surtout Pierrehumbert & Beckman, 1988 qui l'axiomatisent). On se fondera cependant sur cette relation générale d'inclusion entre les unités linguistiques pour définir le concept de granularité en linguistique : des analyses portant sur des unités hiérarchiquement « basses » seront ainsi caractérisées par une granularité plus fine que des analyses mettant en jeu des unités de rang hiérarchique plus élevé.

On peut donc finalement définir l'alignement en linguistique comme la mise en correspondance sous la forme de fonctions du temps de séquences linguistiques enregistrées avec des séquences métalinguistiques, et ce avec une granularité variable.

Le point terminologique que nous venons d'effectuer nous permet de préciser nos besoins spécifiques de phonéticiens britannicistes : notre intérêt porte sur des corpus de parole britannique alignée présentant idéalement une granularité phonématique et, de manière optionnelle, plusieurs autres granularités d'alignement, dont le nombre et la nature permettraient diverses analyses tant dans le domaine de la phonétique que dans d'autres domaines linguistiques (notamment en syntaxe et en analyse du discours). Un tour d'horizon des corpus disponibles (Bouzon *et al.* 2002) permet de constater que rares sont les corpus disponibles (à coût réduit) qui correspondent à ces exigences. En conséquence, le chapitre suivant présente le corpus MARSEC, base de l'élaboration du projet Aix-MARSEC, qui répond précisément au cahier de charges lancé plus haut ...

## **2. De SEC à MARSEC**

### **2.1. Les origines : *Spoken English Corpus***

Le SEC (« *Spoken English Corpus* ») est un corpus d'anglais britannique standard contemporain, d'une durée totale de plus de cinq heures de parole naturelle, contenant approximativement 55.000 mots répartis dans 411 fichiers représentant onze styles de parole différents. Les onze catégories sont les suivantes :

*Groupe A.* Commentaires

*Groupe B.* Bulletin d'informations

*Groupe C.* Parole publique de type I

*Groupe D.* Parole publique de type II

*Groupe E.* Emissions religieuses

*Groupe F.* Reportages

*Groupe G.* Fiction

*Groupe H.* Poésie

*Groupe J.* Dialogues

*Groupe K.* Propagande

*Groupe M.* Divers

Ces différents styles présentent un certain éventail de catégories de discours et peuvent être exploités séparément lors d'études expérimentales. En effet, ce corpus peut être d'une part exploité dans sa totalité en tant qu'échantillon de parole naturelle ; on peut d'autre part observer et comparer un ou plusieurs styles spécifiques. Le corpus SEC correspond à des

enregistrements venant des archives de la BBC. Le type de parole est naturel dans le sens où les enregistrements ont été effectués dans le but de communiquer leur contenu. Le corpus rassemble 17 femmes et 36 hommes soit un total de 53 locuteurs. Divers chercheurs appartenant à l'université de Lancaster et au groupe IBM sont à l'origine du projet SEC, notamment G. Knowles, P. Alderson, B. Williams et L. Taylor.

Le corpus possède un certain nombre d'informations à différents niveaux. Outre le signal sonore, le corpus est transcrit orthographiquement (version ponctuée et version non-ponctuée), étiqueté morphosyntaxiquement grâce au système CLAWS (Garside, 1987) et annoté prosodiquement par G. Knowles et B. Williams à l'aide d'un ensemble de quatorze marques tonétiques («*tonetic stress marks*»). Les symboles utilisés par les deux transcrip-teurs sont résumés dans la figure 1.

Chaque syllabe accentuée est précédée d'un accent tonétique indiquant le mouvement de la fréquence fondamentale ; ce mouvement débute sur la syllabe annotée et continue jusqu'à la syllabe accentuée ou la frontière d'unité intonative suivante (Roach, 1994). Afin de tester la fiabilité de l'annotation prosodique, 9% de la totalité des fichiers sont alignés par les deux transcrip-teurs, ce chevauchement révélant une certaine homogénéité des annotations. Les mots sont regroupés en unités intonatives (UI) mineures et majeures.

—	Low level
	Minor tone-group boundary [Double bar indicates major tone-group boundary]
—	High Level
↓	Down Arrow
↑	Up Arrow
^	High rise fall
v	High fall rise
∩	High fall
∪	High rise
∩	Low fall
∪	Low rise
∧	Low rise fall
∨	Low fall rise
o	Level (circle)

**Figure 1**

*Marques prosodiques utilisés dans SEC*

## 2.2. Le corpus MARSEC

Le SEC fut ensuite adapté dans le but de le rendre exploitable de manière informatique, il devient alors le corpus MARSEC («*MAchine Readable Spoken English Corpus*»). Les modifications apportées au SEC portent tout d'abord sur les marques prosodiques ; les symboles utilisés dans le SEC posent le problème fondamental de portabilité, notamment lors d'utilisation de logiciels de traitement du signal. Ces quatorze marques sont homogénéisées et modifiées afin de comporter uniquement des symboles ASCII, ceux-ci sont présentés dans la figure 2. Ces symboles sont utilisés selon le même fonctionnement que celui du SEC mais permettent une utilisation informatisée facilitée.

L'ajout fondamental apporté à MARSEC est l'alignement temporel du signal sonore au niveau du mot. En effet, la totalité des mots transcrits orthographiquement sont alignés temporellement avec le signal sonore. Cet alignement se présente sous la forme de fichiers (format texte) avec une suite de balises temporelles marquant le début et la fin de chaque mot.

<i>_</i>	<i>low level</i>	,	<i>low rise</i>
<i>~</i>	<i>high level</i>	'	<i>low fall</i>
<i>&lt;</i>	<i>step-down</i>	,\	<i>(low rise-fall – not used)</i>
<i>&gt;</i>	<i>step-up</i>	\,	<i>low fall-rise</i>
<i>/'</i>	<i>(high) rise-fall</i>	*	<i>stressed but unaccented</i>
<i>'/</i>	<i>high fall-rise</i>		<i>minor intonation unit boundary</i>
<i>/</i>	<i>high rise</i>		<i>major intonation unit boundary</i>
<i>\</i>	<i>high fall</i>		

**Figure 2**

*Symboles ASCII utilisés dans MARSEC*

## 2.3. Homogénéisation de MARSEC

L'homogénéisation du corpus correspond à la phase précédant les traitements automatiques du corpus dans le cadre du projet Aix-MARSEC développé au Laboratoire Parole et Langage. La première étape consiste à vérifier manuellement la correspondance exacte entre les fichiers sonores et les fichiers d'étiquettes en mot<sup>2</sup>. Ce tri est nécessaire notamment en raison des 9% de chevauchement correspondant au même signal sonore.

---

<sup>2</sup> Trois fichiers du corpus sont abandonnés pour cause de défectuosité du signal ou de manque d'étiquettes.

La deuxième étape de cette homogénéisation consiste à vérifier manuellement l'alignement entre la totalité des 55.000 étiquettes en mots et le signal sonore correspondant. Pour cela, les fichiers d'étiquettes (format texte) sont convertis en format TextGrid, à savoir le format d'étiquettes utilisé dans le logiciel Praat (Boersma et Weenink, 1996). En effet, Praat permet un contrôle auditif et visuel simultané du signal pour la vérification de l'alignement. Tout décalage de plus de 50 millisecondes entre le signal et les étiquettes de mot est corrigé manuellement dans le but d'obtenir des données fiables et utilisables ultérieurement. Des modifications sont apportées dans environ 20% des fichiers, ces différents fichiers étant sauvegardés avec une extension supplémentaire afin de pouvoir les identifier aisément.

### **3. Phonétisation**

La phase de transcription phonématique du signal sonore est une étape fondamentale pour de nombreuses études expérimentales en phonétique ; ce type d'« annotation simple » est également nécessaire à l'obtention des niveaux supérieurs de représentation que sont les constituants syllabiques et la syllabe, déterminée en fonction du Principe d'Attaque Maximale (Pulgram, 1970), eux-mêmes regroupés en pieds accentuels selon le modèle d'Abercrombie (1967). Même si cette phase de regroupement en constituants syllabiques puis en syllabes de l'ensemble des phonèmes du corpus ne fait pas l'objet d'une description détaillée dans cet article, il est important de souligner qu'elle est réalisée de manière entièrement automatique à l'aide de scripts en langage Perl. Nous allons à présent nous pencher sur la problématique des méthodes de phonétisation avant de présenter de manière précise les principes appliqués au corpus Aix-MARSEC.

#### **3.1. Différentes approches**

La transcription phonétique d'un corpus oral peut revêtir des formes très diverses en fonction des besoins des utilisateurs, des hypothèses de leur cadre théorique ou encore des données disponibles. On remarquera notamment qu'une prise de position dans le cadre d'une phonologie/phonétique articulatoire induira l'utilisation de méthodes spécifiques (Damper, 2001 : chapitre 8) qui dépassent le cadre de cet article. Les méthodes destinées à fournir la transcription phonémique d'un corpus oral (étape qui correspond à ce que nous avons défini plus haut comme une « annotation simple ») sont elles aussi nombreuses et diverses. On peut cependant regrouper ces méthodes en deux grandes catégories selon qu'une annotation orthographique du corpus est disponible ou pas.

Dans l'éventualité où seuls les enregistrements sont disponibles, la tâche correspond à un processus de reconnaissance de la parole. C'est alors typiquement à une méthode stochastique que l'on va avoir recours : chaque phonème est en général modélisé à l'aide d'un HMM (« Hidden Markov Model » ou « modèle de Markov caché »). Dans le cadre d'un système de « décodage acoustico-phonétique », les séquences possibles de phonèmes ne sont pas contraintes : on ne prend pas en compte le contexte afin de définir une probabilité d'apparition d'un segment donné ; dans le cadre d'un véritable système de reconnaissance de la parole, les séquences de phonèmes sont en général conditionnées par leur contexte (on utilise typiquement des « n-grammes » afin de sélectionner les séquences dont la probabilité d'apparition est la plus élevée).

Lorsqu'une annotation orthographique existe pour le corpus, la tâche consiste à générer, à partir de celle-ci, la suite de phonèmes correspondant le mieux au signal annoté. On est alors dans le cadre de la phonétisation d'un texte orthographique, aussi appelée « conversion graphème-phonème » (« *Grapheme-(to)-Phoneme Conversion* » ou « *G2P* » dans la littérature de langue anglaise) et qui constitue l'une des étapes fondamentales de tout système de synthèse de la parole à partir du texte (Damper *et al.*, 1999). Plusieurs méthodes sont là aussi disponibles ; sans entrer dans le détail de l'inventaire de ces méthodes, nous noterons que l'on peut les classer en deux grandes catégories selon :

- qu'elles font appel principalement à des règles phonologiques produites de manière non automatique (« *rule-based systems* », cf. McIlroy, 1973 pour l'un des premiers systèmes ou Divay & Vitale, 1997 pour l'un des plus récents) ou, au contraire,
- qu'elles s'appuient sur différentes méthodes automatiques fondées sur l'exploitation directe de données (« *data-driven systems* ») : on pensera dans cette catégorie à des algorithmes de prononciation par analogie (implicite ou explicite, cf. Damper & Eastmond, 1997), à NETspeak (un réseau de neurones de type « perceptron multi-couche », cf. Rumelhart *et al.*, 1986) ou encore à IB1-IG (fondé sur une méthode statistique de classification automatique, cf. Daelemans *et al.*, 1997 et Van Den Bosch, 1997).

Tous ces systèmes ont en commun un fonctionnement fondé sur la recherche d'entrée dans un lexique/dictionnaire phonétisé.

Le système de phonétisation utilisé pour le corpus Aix-MARSEC, vers lequel nous allons à présent nous tourner, appartient à la première de ces catégories : en effet, fondé sur la recherche d'entrées dans un lexique, il est ensuite complété par l'utilisation de règles

phonotactiques destinées à simuler certains des phénomènes de production spécifiques à la parole continue. La démarche adoptée est d'inspiration clairement linguistique, caractérisée par un équilibre entre portabilité (lexique auquel viennent s'ajouter certaines règles phonotactiques et contraintes non spécifiques à la langue) et applicabilité spécifique à l'anglais britannique oral (ensemble de contraintes et de règles phonotactiques spécifiques à la langue).

### **3.2. Principes de phonétisation Aix-MARSEC**

De manière plus précise, le fonctionnement global du système de phonétisation utilisé dans ce travail consiste à rechercher automatiquement chacun des mots du corpus (à partir de l'alignement orthographique) dans un dictionnaire électronique de prononciation à l'aide de scripts Perl. Le dictionnaire utilisé est l'*Advanced Learners' Dictionary* (publié par *Oxford University Press*) qui contient un nombre total d'environ 71.000 mots.

Lors de la conversion graphème-phonème, une série d'environ 700 mots présents dans le corpus n'avait aucune correspondance dans le dictionnaire ; il s'agit ici principalement de noms propres associés à des personnalités ou à des lieux. Un dictionnaire de formes complémentaires a ainsi été créé dans lequel se trouve la transcription manuelle de la totalité de ces mots à partir du dictionnaire de prononciation de Wells (1990). Par conséquent, pour chaque mot du corpus absent du dictionnaire de prononciation principal, on cherche son entrée dans ce deuxième dictionnaire de mots complémentaires.

Ce premier système de conversion graphème-phonème permet d'obtenir une transcription phonologique de surface puisque le dictionnaire liste des formes de citation. Or, la spécificité de la parole naturelle réside dans un décalage entre les réalisations phonétiques des locuteurs et les formes de citation. Par exemple, les formes présentes dans le dictionnaire ne tiennent pas compte de la réduction inhérente à la parole naturelle (ainsi, la conjonction *and* est uniformément transcrite /ænd/). Dans le but d'améliorer la correspondance entre le signal et la transcription obtenue automatiquement à partir des dictionnaires, s'ensuit un traitement spécifique des formes réduites. Un troisième dictionnaire, élaboré à partir de O'Connor (1967) et Faure (1975), composé de l'ensemble des mots anglais qui possèdent à la fois une forme pleine et une forme réduite est alors utilisé avec en entrée cette liste de mots suivie de la transcription de leur forme réduite respective. Lors de la phase de phonétisation, le choix entre la forme pleine (dans le dictionnaire principal) ou la forme réduite (dans le dictionnaire de formes réduites) se fait en fonction de la présence ou absence d'une marque prosodique sur

le mot en question : s'il est précédé d'une marque prosodique, il sera transcrit avec sa forme pleine alors que si aucune marque prosodique ne précède ce mot, il sera transcrit avec sa forme réduite. Par exemple, «*' and* » est transcrit /ænd/ (présence de la marque prosodique '*high fall-rise*') et «*and* » (aucune marque) est transcrit /ənd/.

Pour résumer, la phonétisation de la totalité du corpus s'effectue grâce à l'utilisation de trois dictionnaires différents : le dictionnaire principal, le dictionnaire des mots complémentaires transcrits manuellement et le dictionnaire des formes réduites. Notons que ces trois dictionnaires sont séparés pour deux raisons ; dans le cas des formes réduites, il est évident que les formes pleines et les formes réduites doivent être séparées afin que le script sache quelle forme utiliser en fonction de la présence ou absence d'une marque prosodique. En ce qui concerne les formes complémentaires, elles sont regroupées dans un dictionnaire isolé plutôt qu'ajoutées au dictionnaire principal dans le but de pouvoir faire évoluer ce système de phonétisation à d'autres corpus tout en gardant un dictionnaire spécifique à MARSEC<sup>3</sup>.

### 3.3. Traitements spécifiques

Lors de la phase de phonétisation, certaines formes nécessitent un traitement spécifique, notamment la réalisation du morphème 's' du pluriel et de la troisième personne du singulier, ainsi que le morphème 'd' du prétérit régulier et du participe passé régulier en fonction du contexte phonémique, ou plus exactement du voisement de la consonne précédente.

De plus, la conversion graphème-phonème ne permet pas de traiter les génitifs, les contractions (de type *I'm*), les abréviations, les chiffres et les dates. En effet, ceux-ci ne sont pas présents en entrées des différents dictionnaires (il serait trop coûteux de les ajouter manuellement) et l'application de notre système à d'autres données poserait les mêmes problèmes. Pour y remédier, un ensemble de fonctions en langage Perl permet de traiter ces différentes formes en les décomposant en formes présentes dans le dictionnaire principal. Les génitifs sont décomposés en 'mot + forme du génitif' ce qui permet de rechercher le mot dans le dictionnaire puis de transcrire le génitif en fonction du contexte phonémique (/s/ après une consonne non-voisée, /z/ après une consonne voisée et /ɪz/ après /szʒ/). De la même manière, les formes contractées sont décomposées en 'pronom + contraction' pour ensuite être transcrites en tant que deux formes différentes regroupées. Dans le cas de *I'm* par exemple, la forme contractée est décomposée en *I + 'm* ainsi présentes dans le dictionnaire.

---

<sup>3</sup> Ainsi, une version mise à jour du dictionnaire *Advanced Learners' Dictionary* pourrait être intégrée sans pour autant perdre nos modifications.

En ce qui concerne les abréviations absentes du dictionnaire général, on remarque deux types différents : les abréviations alphabétiques (composées uniquement de lettres) et les abréviations que nous qualifions de ‘mixtes’ (lettres et chiffres). Les premières sont décomposées en lettres comme par exemple BBC qui devient B + B + C, ces lettres étant des entrées du dictionnaire principal. Les abréviations mixtes, mêlant des lettres et des nombres comme par exemple dans les codes postaux anglais (“YO1 1ET”), subissent le même type de traitement de décomposition en lettres + nombres (convertis en mots) puis de conversion graphème-phonème.

Les chiffres sont convertis en mots orthographiques pour ensuite être recherchés dans le dictionnaire principal. Ce traitement des chiffres a toutefois posé le problème des dates puisqu’il est difficile de distinguer de manière automatique un chiffre d’une date, leurs réalisations étant fondamentalement différentes. Par exemple, le chiffre 1975 peut tantôt être considéré comme une date et être réalisé « *nineteen seventy five* » ou comme un chiffre et être réalisé « *one thousand nine hundred and seventy five* ». La solution adoptée est de considérer que tous les chiffres entre 1000 et 2000 ont plus de chance d’être des dates plutôt que des nombres ; ils sont alors convertis comme tels par notre système.

### **3.4. Problèmes spécifiques**

Deux problèmes se sont posés lors de la phonétisation du corpus. Le premier concerne le traitement des dates soulevé dans la section précédente. En effet, les nombres entre 1000 et 2000 sont considérés arbitrairement comme des dates mais ce choix ne nous garantit pas un traitement correct de ces cas. Cette solution, permettant de limiter les erreurs, est pour l’instant conservée mais demande à être plus amplement étudiée.

Le deuxième problème se posant lors de ce traitement porte sur les doublons ; par doublons, nous entendons les formes ayant deux entrées différentes dans le dictionnaire principal et donc deux réalisations possibles. Le mot « *object* » par exemple possède deux entrées dans le dictionnaire : le verbe prononcé /əb'dʒekt/ et le substantif /'ɒbdʒɪkt/. Aucune solution automatique n’est pour l’instant appliquée. Plusieurs possibilités s’offrent à nous concernant ce problème :

— la première possibilité de solution consisterait à nous appuyer sur le tagging lexical effectué à l’aide de CLAWS par les responsables du projet SEC ; l’information lexicale ainsi récupérée permettrait alors la sélection de la phonétisation adéquate dans le cas d’homographes de catégories lexicales différentes (Nom vs. Verbe) ;

— la deuxième possibilité de traitement des doublons pourrait quant à elle s'appliquer y compris dans le cas d'homographes appartenant à la même catégorie lexicale (comme dans le cas des substantifs « *wind* » pouvant être réalisés /waɪnd/ ou /wɪnd/). La solution consisterait alors à fournir au système d'alignement la totalité des possibilités de phonétisation, la plus optimale étant automatiquement retenue.

On peut finalement imaginer une solution mixte qui consisterait à s'appuyer de manière préférentielle sur le tagging (solution 1), plus robuste, et à recourir à la comparaison des phonétisations concurrentes (solution 2) lorsque la solution 1 s'avère non pertinente. Cette voie constitue une piste de recherche dont la description et l'évaluation feront l'objet de publications futures.

## **4. Optimisation par règles d'élision**

### **4.1. Justification**

La méthode utilisée pour la phonétisation du corpus Aix-MARSEC est, nous l'avons vu, fondée principalement sur un algorithme d'extraction de lexique. Ce procédé, qui a indubitablement l'avantage de la rapidité et de la portabilité, comporte cependant plusieurs inconvénients ; nous nous attacherons particulièrement au problème de l'« abstraction phonologique » de la phonétisation extraite du lexique par l'algorithme. En effet, la phonétisation récupérée dans le lexique est fondée sur la forme de citation de l'unité graphique phonétisée ; la transcription phonétique ainsi produite correspond à une prononciation canonique qui ne tient aucun compte des nombreux phénomènes de réduction vocalique, d'assimilation régressive et progressive, d'élision, d'épenthèse, de métathèse, *etc.* qui constituent certaines des caractéristiques les plus typiques de la parole continue. L'utilisation d'un algorithme d'optimisation de la transcription phonémique par règles d'élision permet alors de simuler la réalité de la parole continue de manière plus fidèle, et constitue donc un élément de réponse à ce problème. De telles règles doivent nécessairement voir leur application restreinte par des contraintes (phonotactiques, morphosyntaxiques et autres) dont nous allons à présent donner le détail en ce qui concerne le protocole d'optimisation appliqué au corpus Aix-MARSEC.

### **4.2. Condition d'application des règles**

Dans l'optique d'améliorer le système de phonétisation, nous avons élaboré une série de seize règles d'élision de phonèmes. On peut répartir ces règles en deux catégories selon qu'elles s'appuient ou non sur des contraintes phonotactiques :

- Règles non-phonotactiques de deux types : élision des phonèmes dont la durée prédite (cf. *infra*) est inférieure ou égale à 5ms ; élision fondée sur la nature morphologique des unités (cas de *and*, *he*, *he's*, *he'll*, *he'd*, *him*, *his* et *her*).
- Règles phonotactiques : elles précisent l'inventaire des contextes spécifiques "autorisant" l'élision d'un phonème donné. C'est dans cette perspective que nous proposons d'utiliser l'expression "phonotactique prédictive" que nous définirons comme l'application de règles phonotactiques en vue de l'optimisation d'une phonétisation brute dans le cadre de la parole continue.

Les règles phonotactiques ne sont pas appliquées au signal de manière brute mais requièrent un certain nombre de conditions. Ces conditions sont de trois ordres : intonatif, temporel et phonotactique.

Concernant la contrainte intonative, la condition d'application porte sur la présence ou absence de marque prosodique. En effet, on suppose que dans certains cas les mots sujets aux élisions mais précédés d'une marque prosodique ne sont pas réduits du fait même de la présence d'un contour mélodique sur ce mot. Ces cas seront précisés au cours de la description des règles.

Dans la catégorie des contraintes temporelles, la première condition consiste en un seuil minimal, fixé pour quatre phonèmes, au-dessus duquel le phonème ne peut être supprimé ; le seuil des phonèmes /t/, /d/ et /θ/ est de 55ms et celui de /θ/ de 110ms. Ces seuils sont établis en fonction de la liste des durées minimales (tous contextes confondus) de Klatt (1979) et confirmés lors des observations des données. La seconde condition est relative à ce que l'on appelle le «facteur z» en référence à l'approche de Campbell (1992). Comme le montre la formule 1 ci-dessous, la méthode consiste, à partir d'une part de la durée d'un mot donné du corpus et, d'autre part, de la somme des moyennes et des écarts types de chacun des phonèmes qui le composent, à calculer un coefficient de modification de durée segmentale (ou «facteur z») pour le mot. Cette méthode, qui correspond à l'utilisation de la transformée z utilisée de manière classique en statistiques pour la réduction d'une variable centrée, est fondée sur un «principe d'élasticité » étendu à l'échelle du mot (et non plus de la syllabe dans Campbell, 1992).

$$Duree\_Mot = \sum_{i=1}^{Nb\_phonemes} (m_{ph\alpha} + z * sd_{ph\alpha}) \quad (1)$$

$$z = \frac{Duree\_Mot - \sum_{i=1}^{Nb\_phonemes} (m_{ph\alpha})}{\sum_{i=1}^{Nb\_phonemes} (sd_{ph\alpha})} \quad (2)$$

### Formules 1 et 2

*Facteur Z exprimé en fonction de la durée du mot  
et de la moyenne et de l'écart type de chaque phonème*

Le calcul du facteur z (formule 2) consiste à soustraire à la durée du mot la somme de la durée moyenne de chacun des phonèmes qui le composent (numérateur de la formule) ; cette valeur est ensuite divisée par la somme des écarts types de ces mêmes phonèmes (dénominateur). Un facteur z négatif dénote alors une durée observée plus petite que la somme des durées moyennes et donc une probabilité de voir se réaliser certains phénomènes d'élision caractéristiques de la parole continue.

La catégorie des contraintes phonotactiques est obtenue après observation d'une partie des données du corpus et à partir des études de Jones (1991), Wells (1990) et Gimson (réédité par Cruttenden, 1997). Les règles sont par conséquent établies en fonction des élisions systématiques observées dans le signal ainsi que par les règles données dans les études citées ci-dessus et avérées dans le signal. Le détail de ces contraintes se trouve dans la description des règles d'élision présentée dans la section suivante.

### 4.3. Règles d'élision

Pour des raisons d'écriture du script Perl gérant les phénomènes d'élision de phonèmes, seize règles (expressions régulières) ont été formulées. Ces seize règles peuvent être regroupées en onze principes morpho-phonologiques. Nous présentons dans le reste de cette section ces onze principes en les explicitant et en fournissant des exemples pour chaque cas pertinent. Ces principes sont appliqués à condition que le facteur z du mot soit inférieur à zéro, qu'il n'y ait aucune marque prosodique (selon les principes) et que le seuil soit respecté pour les phonèmes concernés. Notons que dans la représentation de chacun des principes, la forme '#' symbolise une frontière de mot et '-' en exposant symbolise l'exclusion du ou des phonème(s) suivant(s).

- *Principe n°0* : élision de tout phonème dont la durée prédite est inférieure ou égale à 5ms.
- *Principe n°1*: élision du phonème [d] dans **and**

La forme de la conjonction *and* concernée est nécessairement la forme réduite /ənd/ puisque, suivant notre système de phonétisation, la forme pleine est précédée d'une marque prosodique. Cette forme est souvent réduite à /ən/ qu'elle soit suivie d'un mot commençant par une voyelle ou une consonne.

- *Principe n°2*: élision de [h] dans les formes **he, he'd, he'll, he's, his, him, et her**

En parole continue, la fricative [h] dans les pronoms et/ou contractions énumérés ci-dessus est souvent élidée ; toutefois, cette consonne est supprimée dans la transcription à condition qu'aucune marque prosodique ne précède le mot en question, dans ce cas, on imagine que le pronom fortement accentué sera réalisé avec sa forme pleine, sans élision du /h/.

- *Principe n°3* : élision de [t] ou de [d] dans le contexte **{[t][d]} # {[t][d]}**

Lorsqu'un mot se termine avec un [t] ou un [d] et que le mot suivant commence par un [t] ou un [d], l'alvéolaire finale est souvent supprimée, cette règle ayant comme condition le facteur *z* et le seuil fixé à 55ms. Ce principe s'applique aux énoncés tels que *I've got to go* qui sera réalisé *gɒtə/* ou *the red dragon* réalisé *ðərədrægən/* en parole continue. Il existe une restriction à ce principe : outre la prise en compte du seuil minimal, il ne s'applique pas lorsque le [d] correspond au morphème du prétérit ou du participe passé -ed réalisé /ɪd/. Ainsi, dans *an unexpected turn*, le /d/ n'est pas supprimé. Les deux consonnes alvéolaires ne sont pas produites distinctement, à savoir par deux réalisations articulatoires complètes, mais la durée de la tenue du [t] ou du [d] restant nettement au-dessus de la moyenne reflète la présence des deux consonnes. Pour l'alignement automatique, il est nécessaire de rendre compte de ces phénomènes puisque le système ne pourrait trouver les deux consonnes dans le signal. Nous perdons l'information sur cette tenue plus marquée, mais nous pourrions la récupérer grâce à l'allongement de l'alvéolaire concernée.

- *Principe n°4* : élision de [t] et [d] dans le contexte **C<sub>1</sub> + {[t][d]} # C<sub>2</sub> - {[h][j]}**

Si [t] ou [d] en position finale de mot est précédé d'une consonne (quelle qu'elle soit) et suivi d'un mot commençant par une consonne autre que [h] ou [j], alors il est supprimé. Ce principe

concerne les énoncés du type *you mustn't lose* réalisé /mʌsn ɪu:z/ et est caractérisée par deux restrictions :

1) [t] ou [d] doit être en position finale de mot

2) C<sub>2</sub> peut être n'importe quelle consonne sauf [h] ou [j] : Gimson (réédité par Cruttenden, 1997) précise qu'un [t] ou [d] final suivi d'un [j] est généralement réalisé par une affriquée, comme dans *helped you* réalisé /heɪptʃə/. Cette réalisation est tout à fait fréquente, mais nous avons également rencontré des cas où le [j] était simplement dévoisé sous l'effet du caractère non-voisé de la plosive précédente. De ce fait, cette affriqation n'est pas généralisée à la totalité du corpus.

- *Principe n°5* : élision de [p] ou de [k] dans le contexte **nasale homorganique +** **{[p][k]}** (#) C - {[r][l][j]}

Ce principe supprime le phonème [p] ou [k] dans les groupes consonantiques dans lesquels sa position est homorganique avec celle de la nasale ([m] ou [ŋ]) précédente. Il traite les mots tels que *glimpse* produit *gɪlɪms/*, mais également l'ajout du morphème du pluriel, de la troisième personne du singulier, du prétérit et du participe passé dans les séquences /mps/ et /mpt/ : on a par exemple *camp*s /kæəm/s/, *jump*s /dʒʌm/s/ et *jump*ed /dʒʌmt/. L'élision peut apparaître à l'intérieur d'un mot mais également au-delà d'une frontière de mot comme dans *they jump silently*. Concernant l'élision du [k], ce principe permet de traiter les élisions dans des mots tels que *thanks, thanked* mais également dans les énoncés du type *thank Peter* /θæŋ pi:tə/. Dans ces groupes consonantiques composés de trois consonnes, la consonne centrale a tendance à être supprimée.

Ce principe est également restreint la nature de la consonne suivant la plosive : il s'applique pour toutes les consonnes sauf [r - l - j] pour éviter de faire l'élision dans des cas comme *computers* ou *wrinkle* où le /p/ ou le [k] serait alors supprimé.

- *Principe n°6* : élision de [l] dans le contexte **[ɔ:] + [l]** (#) C

Ce principe s'applique à la fois à l'intérieur des mots et au-delà des frontières de mot. Ainsi, il rend compte de l'élision de /l/ dans les mots tels que *always, already, although, all right* et *almanac* (mots cités par Gimson).

- *Principe n°7* : élision du phonème [θ] dans le contexte **C + [θ]** (#) [s]

Il s'applique à l'intérieur des mots, comme par exemple dans *months*, *twelfths* et *fifths* (Gimson), mais également au-delà des frontières de mot, comme par exemple dans *the fifth soldier*. Ce principe a comme condition le facteur  $z$  mais également le seuil du phonème [θ].

- *Principe n°8* : élision de la plosive en contexte [s|z] + {[p|b][t|d][k|g]} (#) [s|z]

Deux traitements différents sous-tendent ce principe : dans le cas d'une de ces trois séquences, il y a élision de la plosive, on se retrouve avec deux [s] adjacents, le premier est alors éliminé. Dans l'énoncé *tourists* ou *the tourist seems*, la séquence /sts/ sera tout d'abord réduite à /ss/, puis un traitement ultérieur supprimera le premier /s/ et réduira cette séquence à /s/. Le principe n°8 est également étendu à l'élision de la plosive dans le contexte [z] + {[b][d][g]} (#) + [z] bien qu'aucun cas ne soit présent dans le corpus.

- *Principe n°9* : élision du schwa dans [ə] + {[l][r]} (#) + **voyelle réduite** {[ɪ][ə]}

Ce principe s'applique en fonction du seuil fixé pour le schwa et possède une restriction : il ne s'applique pas dans le cas de la séquence [r] + [ə] + [l] + voyelle réduite, principalement parce que

- 1) le [r] ne peut être final en anglais britannique standard et
- 2) [rl] ne représente pas une attaque licite.

Dans *necessarily*, il n'y a pas élision de /ə/ et le mot est réalisé /nesəsərəlɪ/. Globalement, ce principe permet de rendre compte des mots tels que *camera* /kæmərə/, *Reverend* /revrənd/ et *library* /laɪbrəri/.

Dans les cas avec un double [l] ou un double [r] après application du principe d'élision, il y a effacement du premier phonème ; dans *honorary* /ɒnərəri/, il y a élision du premier schwa, puis du deuxième, et simplification de /ɒnrri/ en /ɒnrɪ/. Notons que le principe s'applique également lorsqu'il s'agit du -r- de liaison comme dans *after a while* qui peut être réalisé /ɑ:ftəwaɪl/ en parole rapide.

- *Principe n°10* : élision du schwa dans le contexte # [kən] ('syll (syll<sub>[0...n]</sub>)) #

Ce principe traite les élisions de /ə/ dans les mots tels que *con'front* ([kən] suivi d'une syllabe accentuée terminale) et *con'stituency* ([kən] suivi d'une syllabe accentuée non terminale), ainsi que l'auxiliaire de modalité *can* non précédé d'une marque prosodique (*I can* [kn] *really believe*). Un seuil de 55ms est utilisé pour le schwa.

- *Principe n°11* : élision du schwa dans le contexte  $\{[k][p]\} + [ə] + [n] \#$

Le seuil du schwa est ici aussi fixé à 55ms et ce principe s'applique uniquement dans les syllabes en position finale de mot, celles-ci étant nécessairement inaccentuées. Il traite les mots tels que *open*, *thicken*. Jones (1991) précise qu'il n'y a pas d'élision après ces deux plosives ; toutefois, lors de notre observation de certains extraits du corpus, les mots tels que *happen(ed)* étaient réalisés avec un [ŋ] syllabique et donc avec suppression du [ə]. Nous avons donc décidé d'étendre l'élision du schwa à tous les types de consonnes précédant un [n] final.

#### 4.4. Évaluation des règles d'élision

L'application des seize règles d'élision de phonèmes conduit à la suppression de 4083 phonèmes dans la totalité du corpus. Le nombre de phonèmes élidés représente ainsi approximativement 2% des 199.770 phonèmes qui constituent la phonétisation brute du corpus Aix-MARSEC. L'évaluation de la qualité prédictive de ces règles a été effectuée de manière manuelle sur un échantillon de dix fichiers du corpus et est quantifiée à l'aide des mesures « rappel », « précision », « silence », « bruit » et « F-mesure » (cf. tableau 1), couramment employées en recherche documentaire (Van Rijsbergen, 1979).

MESURES	
RAPPEL	50,51 %
PRECISION	74,44 %
SILENCE	49,49 %
BRUIT	25,56 %
F-MESURE	60,18 %

**Tableau 1**

*Mesures d'évaluation de l'algorithme de prédiction des élisions*

Le rappel quantifie le rapport des éléments pertinents récupérés sur le total des éléments pertinents. Dans notre cas, cette mesure représente la proportion d'élisions prédites sur la totalité des élisions rencontrées. Un taux de rappel de 100 % signifierait que toutes les élisions rencontrées lors de l'évaluation ont été prédites par nos règles. Un rappel de l'ordre de 50 % signifie donc que notre algorithme prédit de manière correcte la moitié des élisions effectivement réalisées par les locuteurs dans le corpus. La mesure complémentaire du rappel

est le silence, qui représente la proportion d'élisions non prédites sur la totalité des élisions rencontrées

La précision mesure le rapport des éléments pertinents récupérés sur le total des éléments récupérés ; sa mesure complémentaire, le bruit, est le rapport du nombre d'éléments récupérés à tort sur le nombre d'éléments récupérés. Dans notre cas, la précision quantifie le nombre d'élisions prédites de manière correcte sur le nombre d'élisions prédites. Un taux de précision de 100 % correspondrait à l'absence totale de prédiction erronée d'élision. Une précision de 74,44 % indique donc dans notre cas que près des trois quarts des élisions prédites par notre algorithme ont effectivement été réalisées par les locuteurs du corpus.

Il est nécessaire d'insister sur l'importance de la F-mesure dans le cadre d'une évaluation. Dans un cas extrême, on peut en effet atteindre un rappel de 100 % en prédisant que tous les phénomènes rencontrés sont pertinents : cela reviendrait pour nous à éluder la totalité des phonèmes du corpus ... La précision, cependant diminuerait de manière proportionnelle car la plupart des élisions prédites le seraient à tort. La performance d'un système est optimale lorsque ce dernier obtient le couple de valeurs (rappel, précision) le plus élevé ; cette prise en compte simultanée du rappel et de la précision d'un système est reflétée par la F-mesure qui correspond à la moyenne harmonique des deux taux. Notre algorithme bénéficie d'une F-mesure de l'ordre de 60 % qui, sans caractériser un système optimal, démontre la qualité de la démarche de phonotactique prédictive adoptée.

L'annotation simple phonématique obtenue bénéficie d'un taux de fiabilité de 94,79 % qui, comme nous le développerons dans la sixième section de cet article, pourrait être encore amélioré par l'augmentation du taux de rappel.

## **5. Alignement du corpus Aix-MARSEC**

L'une des caractéristiques qui font du corpus Aix-MARSEC une base de donnée particulièrement intéressante pour toute recherche en phonétique/phonologie anglaise est liée à la disponibilité d'un alignement phonématique. Celui-ci constitue la base fondamentale sur laquelle s'appuient les alignements des autres niveaux de l'analyse linguistique (syllabe, pied, unité rythmique, mot, unité intonative). Les sections suivantes vont donc présenter brièvement les différentes méthodes qui se sont offertes à nous pour l'alignement des phonèmes d'Aix-MARSEC avant de fournir une évaluation détaillée de la qualité de cet alignement.

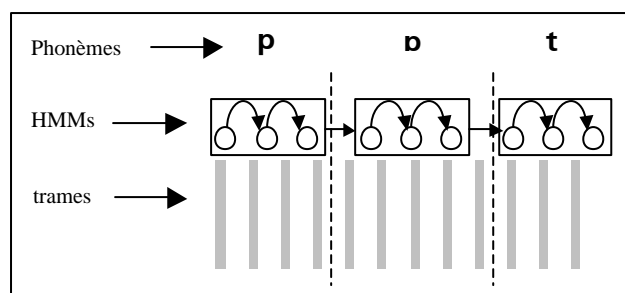
### **5.1. Méthodes d'alignement**

Une fois le corpus phonétisé, deux méthodes principales d'alignement sont disponibles :

- La première méthode consiste à utiliser un algorithme Viterbi classique (Viterbi, 1967) qui calcule la séquence optimale d'états dans un modèle de Markov caché (*HMM*) étant donnée la séquence d'observations que constitue notre annotation phonétique. Cette méthode dite de « force Viterbi » consiste alors à attribuer les trames temporelles pertinentes aux phonèmes transcrits.
- La seconde méthode fait appel à la technique de programmation dynamique (« *Dynamic Time Warping* » ou « *DTW* ») traditionnellement utilisée pour le transfert d'un jeu d'étiquettes d'un enregistrement à un autre (Di Cristo & Hirst, 1997). Dans cette perspective, la tâche consiste à effectuer un calcul de la distance spectrale entre un signal de synthèse produit à partir de l'annotation phonématique et le signal d'origine.

La première version de l'alignement du corpus, que nous présentons dans cet article, est fondée sur une implémentation de la première méthode, l'implémentation de la seconde (*DTW*) étant prévue lors de phases d'alignement ultérieures que nous mentionnerons dans la sixième partie de ce travail.

De manière plus précise, l'alignement du corpus Aix-MARSEC a été réalisé par Christophe Lévy et Pascal Nocéra du Laboratoire d'Informatique d'Avignon. La méthode employée a consisté à modéliser chaque phonème à l'aide d'un modèle de Markov caché (*HMM*) « gauche-droite » composé de trois états émetteurs (Rabiner, 1984) et entraîné, pour des raisons de disponibilité, sur le corpus TIMIT. Chaque état est représenté par un modèle de mélange de gaussiennes (« *Gaussian Mixture Model* ») à 8 composantes et des matrices de covariance diagonales. Le signal sonore est quant à lui représenté à l'aide de douze coefficients cepstraux (*MFCC*) auxquels viennent s'ajouter un coefficient d'énergie ainsi que les coefficients delta et delta-delta pour un vecteur total de 39 coefficients par trame de signal. L'algorithme Viterbi est ensuite utilisé pour attribuer la ou les trame(s) temporelle(s) pertinente(s) à chaque état émetteur, comme le représente la figure 3 ci-après.



### Figure 3

*Attribution par l'algorithme Viterbi des trames temporelles aux états émetteurs des HMMs*

#### 5.2. Evaluation

Tout alignement, de manière évidente, présente un intérêt dès lors qu'il est relativement fiable. Le seuil de fiabilité retenu dépend bien entendu de l'exploitation de l'alignement concerné. Notre tâche concernant l'évaluation de l'alignement phonématique du corpus Aix-MARSEC consiste alors à fournir une quantification des décalages observés entre les données automatiques et les données manuelles, et ce à différents seuils. Notre intérêt résidant principalement dans des études appartenant au domaine phonétique, nous fournissons ci-après les résultats correspondants à des seuils de 5 ms à 64 ms, les valeurs retenues correspondant à celles généralement observées dans ce type d'étude (Di Cristo & Hirst, 1997).

L'évaluation des « erreurs » d'alignement dans le corpus a impliqué la comparaison de 4 fichiers d'environ une minute de parole alignés manuellement avec l'alignement automatique de ces mêmes fichiers, fourni par la méthode décrite plus haut. La mesure des décalages a été effectuée de manière automatique à l'aide de scripts en langage Perl, et peut être résumée à l'aide du tableau suivant :

Seuil	% de décalages inférieurs au seuil
64 ms	93.25 %
32 ms	82.02 %
20 ms	68.37 %
16 ms	59.97 %
15 ms	57.40 %
10 ms	42.43 %
5 ms	23.72 %

**Tableau 1**

*Evaluation de l'alignement automatique à différents seuils*

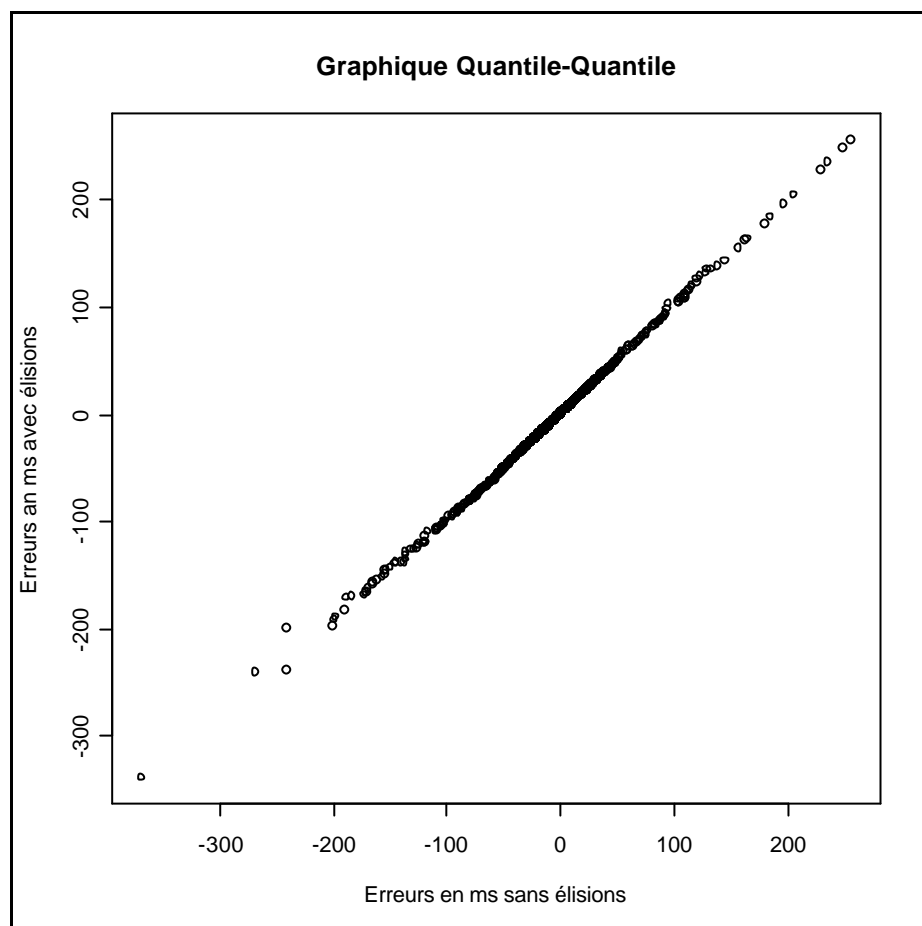
On voit que l'alignement obtenu de manière automatique est fiable à près de 70 % pour un seuil d'acceptabilité de 20 ms. Ce résultat, sans toutefois constituer un score remarquable, est cependant de l'ordre des 77 % présentés par Dalsgaard *et al.*(1991) pour deux minutes de parole anglaise lue extraite du corpus EUROM0.

Ces résultats quantifient un produit dont il nous semble important de dissocier les deux composantes :

- la **composante « phonétisation »** dont la finalité est l'obtention d'une suite de phonèmes correspondant de manière optimale avec la production effective des locuteurs ;
- la **composante « alignement »** dont la finalité est la mise en correspondance temporelle des étiquettes de phonèmes avec les portions de signal correspondantes.

La qualité globale de l'alignement final dépend donc de la qualité de chacune de ces composantes. On pourra alors considérer qu'il y a amélioration de l'alignement final si l'une des deux composantes voit sa qualité croître sans que la qualité de l'autre composante ne décroisse. Dans le cas qui nous intéresse ici, la composante « phonétisation » comporte deux phases : phonétisation brute puis optimisation par règles d'élosion. Nous avons vu (section 4.4) que la phase d'optimisation permet une amélioration de la qualité de la composante « phonétisation » ; il nous semble alors légitime de nous questionner sur l'impact de la phase d'optimisation sur la composante « alignement ». Pour que la qualité globale de l'alignement final soit effectivement améliorée par l'optimisation de la composante « phonétisation », il faut que la composante « alignement » ne soit pas pénalisée par cette optimisation.

Dans cette perspective, nous avons comparé les décalages d'alignement pour les versions respectivement élidée et non élidée de notre phonétisation. Comme le montrent la figure 4 ci-après, les distributions observées ne semblent pas différer de manière significative.

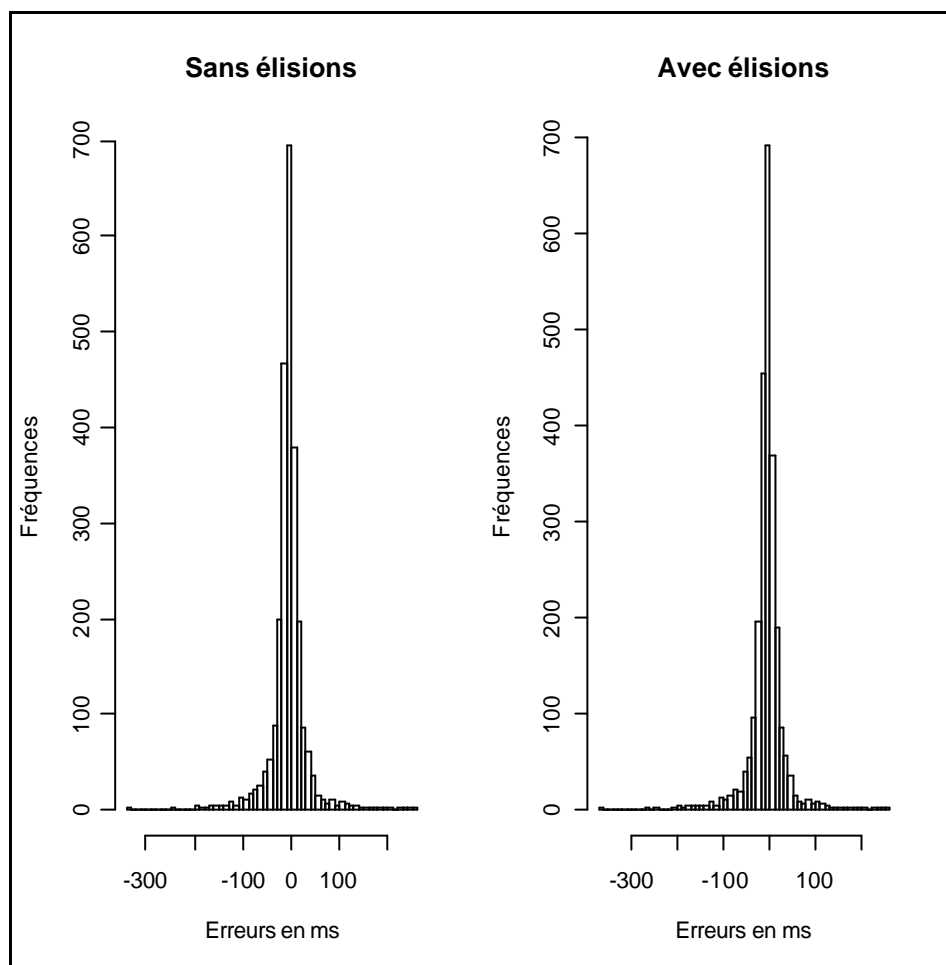


**Figure 4**

*Graphique quantile-quantile des distributions des erreurs pour les alignements fondés sur les phonétisations sans et avec élisions*

Une analyse visuelle du graphique 4 semble indiquer une absence de différence significative entre les deux conditions analysées. Cette observation des distributions, cependant, notamment en ce qui concerne la significativité des résultats suggérés, doit être corroborée par des tests statistiques formels.

On remarquera cependant, après une rapide observation de la figure 5 ci-après, que les distributions observées divergent de manière significative avec des distributions normales ; cette divergence est confirmée par les indices d'aplatissement (ou « *kurtosis* ») et de dissymétrie (ou « *skewness* ») donnés dans le tableau 2.



**Figure 5**

*Histogrammes des distributions des erreurs pour les alignements fondés sur les phonétisations sans et avec élisions*

<b>Indices</b>	<b>Sans élisions</b>	<b>Avec élisions</b>	<b>Dist. normale</b>
<b>Aplatissement</b>	13,07	14,64	1
<b>Dissymétrie</b>	-0,29	-0,57	0

**Tableau 2**

*Aplatissement et dissymétrie pour les distributions des erreurs, mis en rapport avec une distribution normale*

Cette divergence de la normale, notamment en ce qui concerne les forts coefficients d'aplatissement observés (caractéristiques d'une forte concentration autour de la moyenne), rend inapplicables d'une part une classique analyse de variance ou un test de Student pour

tester l'hypothèse de l'égalité des moyennes, et, d'autre part, le test de Fisher pour tester l'égalité des variances, donnée de dispersion qui n'a de sens que dans le cadre d'une distribution normale. Nous avons donc eu recours à des tests qui n'exigent pas la normalité des distributions et avons testé les hypothèses d'égalité des moyennes et d'absence de divergence entre les distributions au moyen du logiciel/environnement de programmation R (R Project for Statistical Computing).

Nous avons tout d'abord utilisé le test de somme ordonnée de Wilcoxon (avec correction de continuité) afin de tester l'hypothèse de l'égalité des moyennes. Fournissant un p-valeur de 0,7757, ce test confirme l'égalité des moyennes des erreurs d'alignement, que la phonétisation se soit appuyée sur une version brute de la phonétisation (sans élisions) ou sur une version optimisée (avec élisions).

Le test de Kolmogorov-Smirnov pour deux échantillons a finalement été employé afin de tester l'hypothèse de l'absence de divergence entre les deux distributions. Avec une p-valeur de 1 (arrondie à la seizième décimale), ce test confirme sans ambiguïté l'absence de différence significative entre les deux distributions, et ce malgré des nombres d'éléments (nécessairement) différents.

Nous pouvons donc conclure d'après ces évaluations quantitatives que la distribution des erreurs n'est pas significativement différente selon que la phonétisation est optimisée (règles d'élisions) ou pas ; cela signifie que l'application de nos règles d'élision à la phonétisation brute du corpus ne génère aucun biais durant la phase d'alignement automatique. La composante « phonétisation » voit donc sa qualité croître alors que la qualité de la composante « alignement » reste constante. L'application des règles d'élision à notre phonétisation brute permet ainsi une amélioration de la qualité globale de l'alignement final, dont la fiabilité est de l'ordre de 70% pour un seuil de 20 ms.

## **6. Perspectives de traitement de l'alignement phonématique**

Suite à l'application de nos algorithmes de conversion graphème-phonème (génération d'une forme phonologique de surface, puis optimisation par règles d'élision), le corpus Aix-MARSEC bénéficie d'un alignement phonématique satisfaisant qui pourrait cependant voir la qualité de ses composantes « phonétisation » et « alignement » améliorée par diverses modifications. Concernant la composante « phonétisation », nous allons principalement présenter certaines des évolutions envisagées au niveau des règles phonotactiques avant de décrire un nouveau protocole automatique itératif. L'amélioration de la qualité de la

composante « alignement », finalement, sera envisagée dans le cadre de l'application itérative d'une nouvelle méthode d'alignement automatique, ainsi que dans le cadre de l'utilisation de modèles phonétiques d'anglais britannique. En effet, pour des raisons de disponibilité, les modèles phonétiques utilisés, entraînés sur le corpus TIMIT, sont spécifiques à l'anglais américain ; ceci explique en grande partie la qualité de l'alignement actuel, pénalisée par l'emploi de modèles parfois inadaptés.

### **6.1. Nouvelles règles**

Dans l'optique d'une amélioration de la phonétisation, on pourra plus particulièrement approfondir notre analyse de la mesure de rappel, qui correspond au taux le plus faible (cf. section 4.4) du couple (rappel, précision). En effet, seule la moitié des élisions effectives est prédite par notre algorithme. Cette mesure est à mettre en relation avec le faible nombre d'élisions prédites (2% du corpus) et appelle les remarques suivantes :

- Certains phénomènes d'élisions récurrents mais non traités par des règles d'élision ont été identifiés dans le corpus. La génération de nouvelles règles phonotactiques prenant en compte ces observations est donc envisagée, laissant espérer une augmentation significative du rappel du système de phonétisation. A titre d'exemple, on notera que la prise en compte de l'élision du *ð*/ initial de l'article défini «*the* », qui représente 22 % du silence, permettrait une telle amélioration.
- Les contraintes appliquées sur les règles ont entraîné à tort le blocage de 19,59 % des élisions prédites par les règles phonotactiques. Un affinement de ces contraintes permettrait donc de diminuer cette proportion et ainsi d'augmenter le taux de rappel du système.

### **6.2. Optimisation de la phonétisation par le système d'alignement**

Etant donné la valeur minimale de 10 ms de la fenêtre utilisée par le système d'alignement automatique, il est envisageable d'optimiser la phonétisation du corpus par la suppression des phonèmes dont la durée est fixée à ce seuil inférieur. En effet, lors de la phase d'alignement automatique, tout phonème présent dans la phonétisation mais non détecté par l'aligneur est automatiquement réduit à cette durée minimale. On peut donc faire l'hypothèse que la suppression de ces phonèmes non détectés constituerait une approximation plus fine de la production effective des locuteurs. L'application itérative et conditionnée (nouvelles règles phonotactiques) de cette phase d'optimisation est ainsi envisagée et fera l'objet de travaux ultérieurs.

### **6.3. Alignement automatique : nouveau protocole itératif**

L'amélioration du composant « correspondance temporelle » de l'alignement est envisagée sous la forme de l'application itérative du système « *DTW* » (cf. section 5.1). En effet, le calcul de distance spectrale effectué par ce système lors de sa première utilisation permet un premier alignement qui pourra ensuite servir de base à la génération d'un second signal de synthèse. Ce procédé peut être appliqué de manière itérative (Di Cristo & Hirst, 1997) jusqu'à obtention d'un alignement au moins localement optimal, permettant ainsi la génération d'un alignement dont il sera intéressant de comparer la précision temporelle avec celle obtenue à l'aide de la première méthode (HMMs et Viterbi).

### **Conclusion**

Cet article a présenté le projet Aix-MARSEC en tant qu'objet scientifique double. En effet, après un rapide approfondissement des concepts d'alignement et de granularité, nous avons détaillé à la fois la méthodologie et le produit final que constitue Aix-MARSEC. Sur le plan méthodologique, nous avons tenté de démontrer l'intérêt d'une approche automatique fondée sur des règles linguistiques d'élimination de phonèmes dans le cadre de la phonétisation étroite d'un grand corpus oral. Cette approche, dont la fiabilité avoisine les 95%, est caractérisée par une modularité (dictionnaires et règles phonotactiques spécifiques ; algorithmes généraux) qui autorise son application à d'autres langues et à d'autres styles de parole.

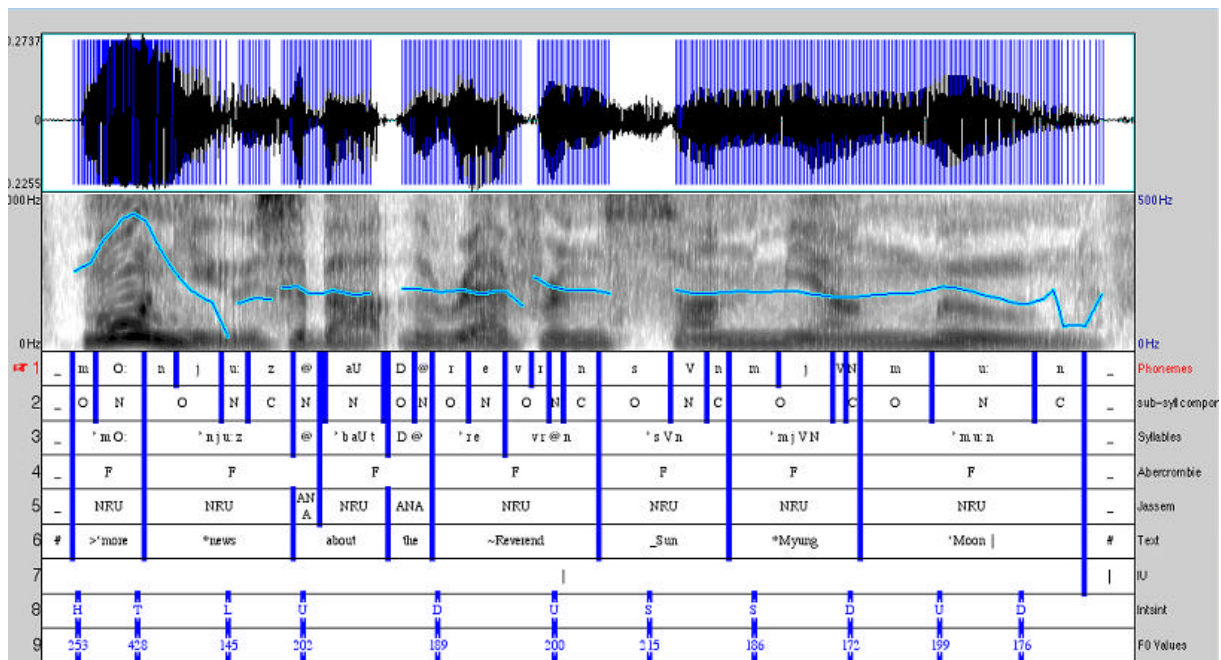
Le "produit" final Aix-MARSEC représente quant à lui une base de données de parole britannique continue unique. Composé de 195.687 phonèmes, regroupés en 88.794 syllabes qui composent elles-mêmes 54.083 mots pour un total de 5 heures 30 de parole, le corpus Aix-MARSEC est caractérisé par un alignement à ces différents niveaux de granularité ainsi qu'aux niveaux du constituant syllabique, du pied accentuel et de l'unité intonative (cf. grille en annexe pour illustration).

Ces caractéristiques, auxquelles viennent s'ajouter un codage et un alignement automatique de l'intonation à l'aide des algorithmes MOMEL et INTSINT (cf. grille en annexe), font de ce corpus une base privilégiée pour des études portant entre autres sur l'effet de la structuration rythmique sur l'organisation temporelle de l'anglais, ou encore sur l'interprétation pragmatique des phénomènes anaphoriques en relation avec la prosodie (thèses en cours au Laboratoire Parole et Langage). Les outils et le corpus Aix-MARSEC seront disponibles sous peu gratuitement sur la page du *English Prosody Group of Aix* du LPL ([www.lpl.univ-](http://www.lpl.univ-)

aix.fr/~EPGA/), permettant ainsi une large diffusion et, nous l'espérons, un large retour qui contribuera à l'amélioration et à l'enrichissement du projet.

## Annexe

Illustration de l'alignement multiple du corpus Aix-MARSEC dans l'environnement Praat.



## Références

- ABERCROMBIE, D. (1967). *Elements of General Phonetics*. Edinburgh : Edinburgh University Press.
- BIRD, S. & LIBERMAN, M. (2001). A formal framework for linguistic annotation. *Speech Communication* 33(1,2), pp. 23-60.
- BOERSMA, P. & WEENINK, D. (1996). Praat, a System for doing Phonetics by Computer, version 3.4. *Institute of Phonetic Sciences of the University of Amsterdam, Report 132*.
- BOUZON, C. & AURAN, C. (2002). Aix-MARSEC : une proposition de traitement automatique de corpus d'anglais britannique oral. *Journée corpus : les corpus oraux en anglais et en français, constitution et exploitation*. Toulouse, 15 novembre 2002.
- BROWMAN, C. & GOLDSTEIN, L. (1989). Articulatory gestures as phonological units. *Phonology*, 6, pp. 201-252.
- CAMPBELL, N. (1992). *Multi-level Timing in Speech*. PhD Thesis, University of Sussex.

- CORNISH, F. (1999). *Anaphora, Discourse and Understanding. Evidence from English and French*. Oxford : Clarendon Press.
- CRUTTENDEN, A. (1997). *Gimson's Pronunciation of English. Fifth edition*. England: Arnold.
- CULIOLI, A. (1990). La Linguistique : de l'empirique au formel. In Culioli, A. (éd.) : *Pour une Linguistique de l'énonciation. Opérations et représentations. Tome 1*. Paris : Ophrys, pp. 9-46.
- DAELEMANS, W., VAN DEN BOSCH, A. & WEIJTERS, T. (1997). IGTrees: Using trees for compression and classification in lazy learning algorithms. *Artificial Intelligence Review 11*, pp. 407-423.
- DALSGAARD, P., ANDERSEN, O & BARRY, W. (1991). Multi-lingual alignment using acoustic-phonetic features derived by neural-network technique. *ICASSP 91*, pp. 197-200.
- DAMPER, R.I. (ed.) (2001). *Data-Driven Techniques in Speech Synthesis*. Kluwer Academic Publishers.
- DAMPER, R.I. & EASTMOND, J.F.G. (1997). Pronunciation by analogy: impact of implementational choices on performance. *Language and Speech 40(1)*, pp. 1-23.
- DAMPER, R.I., MARCHAND, Y., ANDERSON, M.J. & GUSTAFSON, K. (1999). Evaluating the pronunciation component of text-to-speech systems for English: a performance comparison of different approaches. *Computer Speech and Language, Vol. 13 No. 2*, pp. 155-176.
- DI CRISTO, P. & HIRST, D.J. (1997). Un procédé d'alignement automatique de transcriptions phonétiques sans apprentissage préalable. *4° Congrès Français d'Acoustique*, 1, Marseille, 14-18 avril, France : SFA, Teknea.
- DIVAY, M. & VITALE, A.J. (1997). Algorithms for grapheme-phoneme translation for English and French: Applications for databases searches and speech synthesis. *Computational Linguistics 23*, pp. 495-523.
- GARSDIE, R. (1987). The CLAWS word-tagging system. in Garside, R., Leech, G., Sampson, G. (eds.), *The Computational Analysis of English : a Corpus Based Approach*, Longman, pp. 30-41.
- GRABE, E. & POST, B. (2002). Intonational Variation in English. in Bel, B. and Marlien, I. (eds), *Proceedings of the Speech Prosody 2002 Conference*, 11-13 April 2002, Aix-en-Provence: Laboratoire Parole et Langage, pp. 343-346.
- GREENBAUM, S. (1996). *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

- HIRST, D.J. & ESPESSER, R. (1993). Automatic Modelling of Fundamental Frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix-en-Provence*, 15, pp. 75-85.
- HIRST, D., DI CRISTO, A. & ESPESSER, R. (2000). Levels of Representation and Levels of Analysis for the Description of Intonation Systems. In Horne, M. (éd.): *Prosody : Theory and Experiment. Text, Speech and Language Technology, 14*. Kluwer Academic Publishers, pp. 51-87.
- HYMAN, L., KATAMBA, F. & WALUSIMBI, L. (1987). Luganda and the strict layer hypothesis. *Phonology Yearbook 4*, pp. 87-108.
- IViE : disponible à partir de <http://www.phon.ox.ac.uk/~esther/ivyweb/index.html>
- JONES, D. (1991). *English Pronouncing Dictionary*. London : Longman.
- KIPP, A., WESENICK, M.-B. & SCHIEL, F. (1996). Automatic detection and segmentation of pronunciation variants in German speech corpora. *Proceedings of ICSLP 96, 4<sup>o</sup> Int. Conf. on Spoken Lang.* USA : University of Delaware & Alfred I. du Pont Institute, pp.106-109.
- McILROY, M. (1973). Synthetic English Speech by Rule. *Bell Telephone Laboratories Memo*.
- NELSON, G. *et al.* (2002). *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.
- PIERREHUMBERT, J. & BECKMAN, M. (1988). *Japanese tone structure*. Cambridge, MA: MIT Press.
- PULGRAM, E. (1970). *Syllable, Word, Nexus, Cursus*. The Hague : Mouton.
- R PROJECT FOR STATISTICAL COMPUTING, disponible à l'adresse suivante: <http://www.r-project.org>
- RABINER L.R. (1984). A tutorial on hidden Markov Models and selected applications in speech recognition. *IEEE transactions on Speech Audio Processing, vol. 2*.
- ROACH, P. (1994). Conversion between prosodic transcription systems: "Standard British" and ToBI. *Speech Communication, 15*, pp. 91-99.
- ROULET, E. ET AL. (1985). *L'articulation du discours en français contemporain*. Berne : Lang.
- RUMELHART, D.E., HINTON, G.E. & WILLIAMS, R. (1986). Learning representations by back-propagating errors. *Nature 323*, pp. 533-536.
- SELKIRK, E. (1984). *Phonology and syntax: the relation between sound and structure*. Cambridge, MA : MIT Press.

- VAN DEN BOSCH, A. (1997). *Learning to Pronounce Written Words: A Study in Inductive Language Learning*. Thèse de Doctorat, Université de Maastricht, Pays-Bas.
- VAN RIJSBERGEN, C.J. (1979). *Information Retrieval, 2nd edition*. Glasgow : University of Glasgow.
- VITERBI, A. (1967). Error bounds for convolutional Codes and an asymptotically optimum decoding algorithm, *IEEE Transactions on Information Theory*, vol. 2, pp. 260-269.
- WELLS, J.C. (1990). *Pronunciation Dictionary*. London : Longman.