De l'analyse syntaxique à la synthèse de la parole dans le système FipsVox : phonétisation et génération de la prosodie

Jean-Philippe Goldman[1]
Laboratoire d'Analyse et de Technologie du Langage
Université de Genève
<Jean-Philippe.Goldman@lettres.unige.ch>

1. Introduction

Dans cet article, nous présentons le système FipsVox, un synthétiseur de parole développé au LATL (Université de Genève). Cette application aborde le domaine de la synthèse de la parole selon une approche originale puisqu'elle met un analyseur syntaxique au cœur du long processus consistant à transformer un texte en un signal sonore intelligible et naturel. Nous nous efforcerons dans cette contribution de montrer les avantages qu'apporte une analyse syntaxique détaillée par rapport à des approches dans lesquelles le traitement linguistique est moins développé.

Comme la plupart des systèmes complets de synthèse de la parole, le traitement des informations s'articule en plusieurs modules organisés en cascade. En premier lieu, un **traitement linguistique** tente de déterminer les informations syntaxiques et grammaticales nécessaires aux deux étapes suivantes : la phonétisation et la génération de la prosodie. Plus précisément, Fips, l'analyseur syntaxique mis en œuvre assure un double rôle : l'identification des unités lexicales et la détermination des structures des phrases.

La **phonétisation** consiste à traduire les mots orthographiques en une suite de phonèmes sachant que l'on peut rencontrer de nombreux mots inconnus comme les noms propres, les néologismes, les sigles, les abréviations et autres nombres avec ou sans unité de mesure. De plus, il faut procéder à de nombreux ajustements phonologiques propres au

[1] L'auteur est financé par le projet CTI n° 4607 et le projet plurifacultaire « Prosodie » de l'Université de Genève. Il tient à remercier Céline Courtin, Arnaud Gaudinat, et Eric Wehrli.

français (liaison, élision, dénasalisation). Durant la **génération de la prosodie**, les deux paramètres prosodiques principaux sont déterminés : le rythme (de manière plus concrète, la durée de chaque phonème et des pauses) et l'intonation (c'est-à-dire une courbe mélodique la plus naturelle et cohérente possible). De manière interne, on passe d'abord par une représentation phonologique intermédiaire dans laquelle les mots sont rassemblés en groupes accentuels puis en groupes prosodiques.

La chaîne phonétique et les informations rythmiques et mélodiques sont les paramètres nécessaires au très connu codeur de parole MBROLA[2] (Dutoit 2001), pour générer le signal acoustique final. Ce dernier module opère par concaténation de diphones, c'est-à-dire qu'il construit le signal de parole en mettant bout à bout des paires de phonèmes pré-enregistrées. Puis le système utilise un algorithme de traitement du signal pour ajuster la parole selon les paramètres prosodiques précédemment calculés.

L'architecture générale du système FipsVox est très modulaire mais les informations transmises entre les modules en cascade sont très détaillées, notamment la structure syntaxique calculée par l'analyseur Fips. Cette contribution décrit les trois premiers modules entièrement développés au LATL.

2. Analyseur syntaxique

L'analyseur syntaxique Fips effectue une double tâche : d'une part, il segmente le texte d'entrée en phrases, elles-mêmes découpées en unités lexicales, et d'autre part, il cherche à déterminer la structuration la plus appropriée (selon des critères grammaticaux, psycholinguistiques et statistiques) de chacune des phrases (Wehrli 1997).

Ces deux fonctions sont importantes du point de vue de la synthèse de la voix. En effet, l'identification correcte des unités lexicales constitue un prérequis fondamental pour la phonétisation. Il s'agit, en particulier, de lever toutes les ambiguïtés liées aux homographes hétérophones (mots de même orthographe mais prononcés différemment, eg. *les poules du couvent couvent*). La structure syntaxique, quant à elle, joue un rôle important non seulement dans la phonétisation d'une phrase, mais plus encore dans le calcul prosodique qui lui est appliqué.

De manière très schématique, l'analyseur syntaxique Fips incorpore deux composantes, l'analyse lexicale et l'analyse syntaxique, chacune responsable d'une des deux tâches décrites ci-dessus. L'analyse lexicale

[2] Développé à la Faculté Polytechnique de Mons (Belgique).

s'appuie sur un lexique d'environ 35'000 formes de base (lexèmes). Dans les cas simples, l'analyse lexicale identifie un élément du fichier d'entrée à un lexème du lexique. Dans des cas plus complexes, l'identification se fait après une analyse morphologique, ou après le regroupement de plusieurs mots simples, comme dans l'identification de mots composés (*au fur et à mesure*, *ordre du jour*).

L'analyse syntaxique tente (elle n'y parvient pas toujours !) d'associer à un fragment du texte d'entrée une structure de phrase. Le processus d'analyse est de type ascendant (*bottom up*), puisqu'il part des unités lexicales, identifiées par l'analyse lexicale, qu'il tente de regrouper en syntagmes de plus en plus grands, jusqu'au syntagme phrase couvrant l'intégralité d'une phrase orthographique. Pour prendre un exemple très simple, une phrase comme (1), se voit assigner la structure (2) par un processus dont les étapes principales sont énumérées en (3) a-g. :

(1) Le petit éléphant joue.

(2) [TP [DP le [NP [AP petit] éléphant]] joue [VP e]]

(3) a. le → DP
b. petit → AP
c. éléphant → NP
d. AP + NP → NP (combine AP et NP en attachant AP comme spécificateur de NP)
e. DP + NP → DP (combine DP et NP en attachant NP comme complément du déterminant)
f. joue → TP-vp (projection complexe d'un TP avec VP vide comme complément)
g. DP + TP → TP (combine DP et TP en attachant DP comme spécificateur de TP)

On le voit, les règles de la syntaxe sont essentiellement de deux types : projection et combinaison. Les projections concernent principalement les éléments lexicaux. On distingue les projections simples des projections complexes. La projection simple constitue le cas normal, non-marqué et correspond à la projection d'un constituant syntaxique de catégorie XP sur la base d'une unité lexicale de type X. Les étapes a, b et c ci-dessus sont des exemples de projections simples.

Les projections complexes sont des règles qui projettent une structure plus riche, plus complexe, à partir d'un élément lexical. En français, cela ne concerne guère que le verbe conjugué, qui donne lieu à une projection d'une structure TP -- avec le verbe conjugué comme tête -- dominant un complément VP dont la tête est vide, comme dans (3 f). Cela correspond à l'analyse linguistique qui postule un mouvement du verbe conjugué, de VP vers TP.

Les règles de combinaison se rapportent toujours à deux constituants **A** et **B** adjacents et correspondent soit à un attachement de **A** comme sous-

constituant gauche de **B**, soit à un attachement de **B** comme sous-constituant gauche de **A**. Pour des raisons historiques, on appelle souvent le premier cas *attachement de spécificateur* et le deuxième *attachement de complément.* Dans notre exemple, (3d) est un cas d'attachement de l'adjectif comme sous-constituant gauche du groupe nominal (attachement de AP comme spécificateur de NP). De même, (3g) correspond à l'attachement du DP *le petit éléphant* comme sous-constituant gauche du syntagme TP (attachement de DP comme spécificateur de TP), autrement dit, attachement d'un groupe sujet.

Par contre, (3e) est un cas d'attachement à droite, plus spécifiquement, il s'agit de l'attachement du NP *petit éléphant* comme complément du DP dont la tête est le déterminant le.

Les règles de combinaison s'accompagnent de conditions et d'actions. Les conditions vérifient des traits de sélection, d'accord ou autres traits lexicaux associés aux constituants **A** et **B**. A titre d'exemple, l'attachement d'un groupe sujet (cf. 3g) est soumis à la condition d'accord en nombre et en personne des deux constituants DP et TP. L'attachement d'un groupe nominal comme complément d'un déterminant (cf. 3e) est contraint, d'une part, par les propriétés sélectionnelles du déterminant et, d'autre part, par l'accord en nombre et en genre du déterminant et du groupe nominal. Les actions sont des processus liés à la création du constituant résultant de l'application d'une règle. Elles portent sur la spécification du type d'attachement (attachement à droite vs. attachement à gauche) et sur la modification de propriétés sélectionnelles ou d'accord (p. ex. unification de traits d'accord des constituants **A** et **B**). Ainsi, les actions associées à la règle d'attachement d'un groupe nominal à un déterminant (3e) spécifient (i) que **B** s'attache comme complément de **A**, (ii) que le trait sélectionnel de **A** est satisfait et (iii) que les traits d'accord du constituant résultant de l'attachement sont l'unification des traits d'accord de **A** et de **B**.

3. Phonétisation

L'étape de phonétisation (ou conversion graphème-phonème) consiste à transformer le texte orthographique à prononcer en une suite de phonèmes. Elle a lieu après l'analyse syntaxique car elle nécessite un calcul préalable d'informations grammaticales. Cette tâche est particulièrement délicate pour le français où l'on ne trouve pas une relation quasi-biunivoque entre les graphèmes et les phonèmes, contrairement à des langues plus transparentes comme l'espagnol ou l'italien. Par exemple, le phonème /o/ (sans distinction de variétés ouvertes ou fermées) peut être la réalisation orale des suites de lettres 'o','eau' et 'au', sans compter les finales de mots

comme ‘aux’,’ault’, ‘eaux’, ‘ot’, ‘ots’, ‘os’, ’oc’, ’ocs’, ‘aulx’,… Inversement, la lettre ‘x’ peut être silencieuse (en fin de mot par exemple), prononcée ‘ks’ (*Mexico*) ‘gz’ (en début de mot ou dans *exemple*), ‘s’ (*dix* ou dans une forme assimilée de *express* ou *extra*) ou encore ‘z’ comme consonne de liaison (*dix abeilles*).

Avant de décrire les méthodes existantes et celle que nous avons retenue, notons qu’il est possible de trouver à partir de ces quelques exemples des règles simples de phonétisation. Mais, si de nombreux mots sont très transparents quant à leur prononciation et nécessitent moins de traitement ou de règles particulières, d’autres sont plus exceptionnels et, l’effort à fournir en matière de recherche et d’implémentation pour une phonétisation correcte de ces mots rares ou non transparents, peut parfois s’avérer considérable au regard de l’amélioration général du système. D’un point de vue technique, comme la fréquence d’utilisation des mots peut varier énormément, des contraintes de mémoire ou de rapidité de traitement peuvent justifier de privilégier certains mots plutôt que d’autres.

Différentes techniques de phonétisation existent : certaines utilisent des données à connaissance explicite comme un lexique phonétique ou des règles de phonétisation, d’autres passent d’abord par une étape d’apprentissage (probabiliste, classificatoire ou selon une méthode dite par analogie) à partir de corpus phonétisés et alignés (voir Boula de Mareuil 1997 pour une revue détaillée).

Notre système combine un lexique principal et un jeu de règles de phonétisation pour les mots hors-lexiques. Ce choix est justifié parce qu’un système basé uniquement sur un lexique est insuffisant en raison de la diversité des objets linguistiques possibles dans un texte et qui ne peuvent tous être stockés (nombres, heure, date, noms propres, néologismes, abréviations, acronymes,…). A l’inverse un système uniquement basé sur des règles voit le nombre de ses règles croître exponentiellement à mesure que l’on veut phonétiser correctement des mots non-transparents (abrév.), qui nécessitent une information implicite supplémentaire (comme justement *abrév*. pour *abréviation*), ou exceptionnels (*Brooglie*, *arcs-en-ciel, grands-oncles*). Ces exceptions sont entrées sous forme de règles qui ne sont utilisées quasiment que pour ces mots. Certains systèmes de phonétisation basés principalement sur un jeu de règles optent pour de petits lexiques d’exceptions plutôt que de surcharger les jeux de règles. Ces cas particuliers des mots rares sont assez nombreux et on voit ici que lorsque une règle est spécifique à un seul mot, approches par règles et par lexiques ne sont pas si différentes.

L'autre argument justifiant notre approche pour la phonétisation est qu'un lexique détaillé existe déjà et est utilisé pour l'analyse syntaxique.

Finalement, on pourra noter qu'un lexique permet de stocker des informations supplémentaires pour des raffinements ultérieurs concernant le caractère facultatif d'un schwa dans un mot (comparez *sam(e)di, zib(e)line, principal(e)ment* avec *m(e)rise, ch(e)vreuil*), la consonne latente de liaison, la fréquence d'un mot, la propriété de troncature finale pour des mots comme *plus, tous, dix* (voir ci-après).

De manière générale, en matière de phonétisation, il vaut mieux toujours prononcer quelque chose que d'ignorer une partie du texte. D'un autre côté, il est bon de rester prudent dans certains traitements que l'on complexifie rapidement à mesure que les cas particuliers surviennent et qu'on tente de résoudre par des règles supplémentaires. Par exemple, la phonétisation systématique de chiffres romains sans précaution aucune peut engendrer des erreurs plus grossières, comme *MCM* par exemple qui serait phonétisé comme *1900* au lieu d'être épelé puisqu'il s'agit du sigle d'une chaîne de télévision (voir 3.3.1 pour plus d'exemples) ou 1° qui peut être prononcé *primo* ou *un degré*.

3.1. Lexique

Notre lexique contient environ 200.000 formes lexicales avec leur transcription phonétique en plus de toutes les informations grammaticales utilisées lors de l'analyse syntaxique (sous-catégorisation grammaticale, nombre, genre, cas, …). Ce n'est donc pas un lexique d'exceptions mais la source d'information principale pour la phonétisation. Néanmoins, des contraintes de mémoire nous ont suggéré la création d'un lexique phonétique réduit dans lequel figurent uniquement les formes lexicales non-récupérables par le système de phonétisation à base de règles. Pour à peu près 95% des mots, il y a équivalence entre la forme phonétique stockée dans le lexique et la forme phonétique restituée par les règles.

Plusieurs informations phonétiques supplémentaires existent déjà ou sont en cours de développement :

- la consonne latente de liaison, généralement prévisible sur la base de la finale orthographique mais certaines exceptions doivent être notées comme pour les préposition *hors*, *vers*, *envers* et *selon* après lesquelles la liaison est interdite (*vers|un arbre* vs. *dans_un arbre*)
- un trait définissant les mots pouvant être soumis à une troncature phonologique selon la configuration syntaxique et le contexte de liaison

comme *tous*, *plus* dont le s final peut être prononcé /s/ (*j'en veux plus*), /z/ (*plus il mange, plus il a faim*) ou silencieux (*je n'en veux plus*)

- les formes phonétiques alternatives. On distingue des variations sur les consonnes finales (anana(s), ani(s), cassi(s), suspe(ct), verdic(t), hui(t)), les semi-voyelles (tri(j)èdre, qu(w)adri- , obliq(u)ité), les affriquées concernant la plupart du temps pour les mots d'emprunt (adda(d)gio, win(t)ch, (d)jeep, mana(d)ger) et plus rarement des consonnes internes au mot (dom(p)ter, prom(p)titude) ;
- le statut particulier de certains phonèmes comme :
  - le schwa pour lequel l'élision ou le maintien est aussi un effet lexical et pas seulement un effet de la fréquence de mot, du contexte phonétique ou du style et du débit de parole (comparez *belette* et *fenêtre*)
  - le h aspiré (noté ') qui bloque la liaison en début de mot comme dans *hache*, *hibou* mais aussi devant *whisky* et *ouistiti* et pour lequel il n'existe pas de règle simple en fonction de la racine ('*holisme* et *holistique*) ou des premières lettres du mot (*hiératique* et '*hiérarchique*)
  - les géminées et les consonnes ambilsyllabiques (su(rr)éservation, vi(ll)a, i(mm)euble)
  - les différents degrés d'assimilation d'un mot d'emprunt vers le français (*New-York* / ( ) /, *milk-shake* / ( ) /)
- l'accent lexical est également noté dans les lexiques italien et anglais.

L'intérêt principal de l'analyse syntaxique associée à un lexique est le traitement aisé des **homographes hétérophones**, c'est-à-dire des mots graphiquement identiques mais prononcés différemment. Les trois terminaisons de mots ambigus les plus fréquentes sont *–ent* (prononcé /ã/ou //), *-er* (/e/ ou / /), *-tions* (/ / ou / /) :

Les amis du *président président*.
Il vient de *poster* le *poster*.
Nous *notions* des *notions* importantes.

Dans ces trois cas, la catégorie grammaticale désambiguïse immédiatement la prononciation. C'est aussi le cas du mot 'est' qui peut être un verbe (/ /) ou un nom ('/ /'). Ce cas est une erreur classique et comme la fréquence de *est* en tant que verbe est très élevée, cette erreur est très fâcheuse et nuit considérablement à l'intelligibilité de la parole synthétique.

Certaines paires d'homographes hétérophones appartiennent à la même catégorie grammaticale et se distinguent par d'autres traits comme le nombre :

- en tant que verbe : Il *pressent*. Ils *pressent*.
- en tant que nom : Elle a coupé les *fils* de la veste de mon *fils*.

Notons que le mot *fils* (dans le sens du lien de parenté) est invariable et représente un cas pur d'ambiguïté sémantique insoluble avec un analyseur syntaxique.

3.2. Règles

La phonétisation des items lexicaux hors lexique est assurée par un système plus générique qu'un lexique et qui fonctionne à base de règles de phonétisation. Ces règles comportent plusieurs parties : la séquence graphémique à convertir, la séquence phonétique résultante, les contextes orthographiques gauche et droit ainsi que le contexte syntaxique (catégorie grammaticale) du mot à phonétiser. Ces 3 contextes sont optionnels dans les règles mais, s'ils existent,ils doivent être respectés pour l'application de la règle. Ces règles se présentent sous la forme

- 'cd' + G + 'cd' (cs) => P

où:

- cg : contexte gauche
- G : séquence graphémique
- cd : contexte droit
- cs : contexte syntaxique (catégorie grammaticale)
- P séquence phonétique

Les contextes gauches et droits peuvent être :

- vides
- graphémiques (un ou plusieurs graphèmes)
- une marque de début de mot ou de fin de mot
- un identificateur défini préalablement et faisant référence à une suite de contextes possibles (par exemple, Vei représente toutes les voyelles e,é,è,ë,i,ï devant lesquelles la lettre c se prononce généralement /s/ au lieu de /k/)

Les contextes peuvent être également une séquence de plusieurs contextes. Ce qui rend la combinatoire très puissante.

Les identificateurs sont souvent des groupes de consonnes ou de voyelles, ou encore des affixes (micro-, radio-, -ment, -ware)

Voici quelques exemples de **règles** :

- *ph* se prononce /f/ quelque soit le contexte (*éléphant*)

  R1. ph => /f/

- *c* se prononce /s/ devant des voyelles du type i,e,… (*ciel*, *cette*,…), /k/ devant les voyelles de type a,o,u (***ca**pric**o**rne*,…). Mais bien d'autres utilisations de la lettre c doivent être prises en compte (ch, sc, chr, c en fin de mot,…)

  R2. c + 'Vie'=> /s/

  R3. c + 'Vaou' => /k/

- Dans les mots en *–ent*, la finale est /ã/ pour les noms (*bâtiment*) et adjectif (*prudent*) mais silencieuse pour les verbes à la 3ème personne du pluriel (*mangent*). Il existe dans notre lexique 651 noms en –ent (tous en /ã/ sauf *management* et *va-et-vient*), verbes (seuls *consent*, *dément*, *ment*, *sent*, *repent*, *pressent* (de *pressentir* et non de *presser*) sont en /ã/), 1659 adverbes et 120 adjectifs (tous en /ã/). Parmi ces mots, il en existe pour lesquels la confusion est possible (affluent, président, couvent, ferment, équivalent…) et la phonétisation rendue correcte en précisant simplement la catégorie grammaticale (R4 et R5). On prononce ici /ã/ par défaut (R5) et on ne prononce rien en cas de verbe (R4). Le cas est similaire pour les mots finissant en *–tions* (*notions*)

  R4. ent + '#' => //(verbe)

  R5. ent + '#' => /ã/

- *b* se prononce /b/ par défaut (R6) (ou peut rester silencieux comme dans *plomb*) mais perd son trait de voisement et devient /p/ devant s (R7) sauf dans *abstract* /abstrakt/ qui est un mot emprunté à l'anglais, dans les racines *subsid-* (*subside*, *subsidiaire*,…) et *subsist-* (*subsister*, *subsistance*,…) pour lesquelles le *s* se voise en /z/ (R8) ainsi que dans *subsonique* et *subsaharien* dans lesquels *sub* est un préfixe.

  R6. b => /b/

  R7. b + 's' => /p/

  R8. 'su' + bs + 'i' => /bz/

  R9. '#' + 'su' + b => /b/

  R10. '#' + abstract + ('s' #)|'#' => /abstrakt/

L'exception *abstract* peut être lexicalisée ou entrée sous forme de règle ad hoc (R10) en précisant des marques de début et fin de mots pour éviter le voisement du /b/ dans l'expression latine *in abstracto* ou dans *abstraction*, qui lui est bien francophone. Il faut affiner le contexte droit grâce à une disjonction pour phonétiser *abstracts* sous l'hypothèse que le *s* final ne se prononce pas en français contrairement à l'anglais.

Le risque d'oublier de futurs néologismes possibles issus de la préfixation de *abstract* (*sous-abstract*) étant minimes, nous faisons l'hypothèse que dans des néologismes, on prononcera /b/ par défaut ou en cas de préfixe sub- (R9) mais /s/ devant un s.

Notons ici la proximité des approches 'règles' et 'lexique' car en nous attelant à la bonne prononciation de quelques mots, nous inférons des

règles très spécifiques, quasiment une règle par mot, ce qui n'est pas moins coûteux qu'une entrée dans le lexique en terme de mémoire et temps de traitement.

L'algorithme de phonétisation parcourt la chaîne graphémique de gauche à droite. Tant qu'il y a des graphèmes à convertir, la règle sélectionnée est la *première* qui convertit *la séquence graphémique la plus longue* et pour laquelle les contextes (graphémiques et syntaxiques) sont vérifiés. L'ordre des règles est donc important.

Le moteur de phonétisation est un transducteur générique utilisé également pour des conversions d'alphabets phonétiques (Musillo 2001).

3.3. Traitement des mots inconnus

Les mots dits inconnus sont tous les mots qui ne figurent pas dans le lexique qui doivent être phonétisés par règles. La plupart nécessitent un prétraitement comme l'écriture en toutes lettres d'un nombre, la détection du mode de prononciation d'un sigle (lu/épelé), la restitution de la forme complète d'une abréviation ou la détection de l'origine linguistique d'un nom propre. Parmi ces mots inconnus, on distingue :

3.3.1. Les nombres

Selon le type de texte étudié, une dépêche résumant une journée boursière ou un roman classique, la proportion de nombres écrits en chiffres arabes peut varier énormément. Il est en général admis que les dates (12 janvier 1998), les heures (12 heures 30), les numéros de page et d'immeuble et les grands nombres soient écrits sous forme de chiffres. Mais on trouve de nombreux objets linguistiques comportant un chiffre : nombres avec unité de mesure (220V, $12.50), pourcentages (1,2%), fractions (1/8), numéros de téléphone (++41.22.705.73.32), ainsi que les nombreuses manières de les écrire (12:30, 12h30), et les variations sur la présence d'espaces entre les chiffres et les signes de ponctuation.

Il faut également prendre en compte le séparateur décimal (généralement la virgule en français et le point en anglais, mais on trouve aussi ce dernier en français) et le séparateur de milliers (l'espace, le point ou l'apostrophe). Ainsi un nombre peut être écrit : 1.350.000,20 ou 2 500 000 ou encore 2'500'000. D'autre part, si un zéro est placé en début de nombre comme dans un indicatif téléphonique (022), nous avons choisi de le prononcer car sa présence n'est pas anodine.

Ajoutons à cela l'existence des **chiffres romains** (Jean XXIII, le XIXe arrondissement) pour lesquels une conversion préalable est nécessaire. Les signes autorisés sont I(1), V(5), X(10), L(50), C(100), D(500) et M(1000)

et la quantité représentée se calcule comme suit : ajouter les chiffres romains qui sont placés à la droite d'un chiffre qui lui est supérieur ou égal et retrancher les chiffres placés à gauche d'un chiffre qui lui est supérieur. Des contraintes supplémentaires stipulent que la soustraction ne peut s'opérer qu'entre deux chiffres de plus de deux rangs d'écart (*IL,IC,ID,IM,XD,XM) et qu'on ne peut soustraire V,L et D.

Une précaution doit être prise concernant certains chiffres pouvant être confondus avec :

- des sigles comme MCM (chaîne de télévision), XL (taille vestimentaire), MDC (parti politique), CIC (groupe bancaire), CD, LCI, CV , XM, DL, CDD, CDI, MC, MMX, MM
- des mots comme Le (cinquantième ou article défini), Les, Ce, Ces, Me, Mes, Xe, De, Des, Ie
- des abréviations : V (volts)

Finalement, pour ce qui concerne les **nombres ordinaux**, plusieurs formes sont d'usage: 1^{er},I^{er}, 1^{re}, I^{re}, 1^{ers} , I^{ers}, 1^{res}, 2^{e}, II^{e}, 2^{es}, $XVIII^{e}$, mais des formes attestées comme $XVIII^{ième}$, $XVIII^{ème}$ ou encore $XVIII^{me}$ sont admises et correctement phonétisées par notre système. Les formes 1° (primo), 2° (secundo), 3° (tertio) ne sont pas considérées car elles peuvent être confondues avec l'abréviation de 1,2 et 3 degrés.

Enfin le traitement de la barre oblique (/) est délicat car si les formes 1/4, 2/3, 1/2, 1/25000, 1/50000 sont communément lues comme une fraction, il est possible de les confondre avec deux nombres simples.

L'algorithme traitant l'ensemble de ces formes numérales tente de détecter en premier lieu un nombre comportant des séparateurs de milliers (avec pour contrainte de trouver des paquets de trois chiffres et que le séparateur de milliers ne doit pas être le même que le séparateur décimal), un séparateur décimal, au risque de confondre avec un numéro de téléphone formé de la même manière (ex : 70.573.533). Dans le cas du séparateur de milliers, les paquets de chiffres sont prononcés comme un seul nombre. Dans tous les autres cas, on tronçonne la lecture selon les séparateurs comme pour les numéros de téléphone, les scores (3-4), les dates (12/01/2001).

La conversion des nombres de chiffres en lettres se fait selon un algorithme original et appliqué à cinq langues (allemand, français, italien, anglais et espagnol). Pour chaque langue, le système connaît le nom des nombres de 1 à 19 (de 1 à 29 pour l'espagnol), des dizaines (et des centaines pour l'espagnol), des puissances de 10 (cent, mille, million, milliard, billion), des caractères ou mots de liaison autorisés (espace, tiret et 'et') et le suffixe indiquant le trait ordinal (-ième). Une approche

récursive traite chaque nombre et constitue progressivement l'écriture littérale du nombre, qui sera phonétisée par la suite par règles comme un mot hors-lexique quelconque. Les spécificités de chaque langue font l'objet de quelques règles d'exceptions. Les caractères de ponctuation et séparateurs de milliers ne sont pas inscrits dans la forme littérale. Seul le séparateur décimal sera prononcé (virgule ou point).

Les unités de mesures associées à un nombre sont considérées comme des abréviations et sont abordées dans la rubrique 3.3.3. Leur proximité d'un nombre permet de les distinguer de certaines autres abréviations. Par exemple, dans '220V', V n'est certainement pas le chiffre romain ni l'abréviation d'un prénom (abréviation de discrétion).

Notons enfin que la restitution des formes belges et suisses de 70 (septante), 80 (huitante uniquement en Suisse) et 90 (nonante) est paramétrable dans les options du système.

Toutes ces formes numérales sont détectées durant l'analyse par des automates à états finis définis pour les nombres, les heures et les dates (Goldman 1998).

3.3.2. Les sigles et les acronymes

Le sigle et l'acronyme se distinguent par leur mode de prononciation. Le sigle est épelé et l'acronyme est lu. Au besoin, lors de la création de l'acronyme, quelques lettres sont maintenues pour rendre la séquence prononçable : *Ra.D.A.R*, *BENELUX*. Ils sont de plus en plus utilisés dans le langage moderne pour désigner une institution, une administration, une organisation (*ONU* qui peut être lu ou épelé, *OTAN*), un pays (*RFA*,*USA*), ou même un objet courant (*VTT*). Certains ont même tendance à se lexicaliser complètement et à perdre leur casse majuscule (*abc*, *ovnis*) ce qui rend plus difficile leur détection dans le texte. On voit même apparaître des dérivés comme *VTTistes*, *vététistes*, *onusiens* mais à l'exception du premier exemple, tous seront considérés comme des noms communs pour ce qui concerne la phonétisation.

L'usage veut que le sigle soit en lettres majuscules séparées par des points, mais leur nombre grandissant dans les textes, les points peuvent disparaître. On voit aussi des sigles avec seulement la première lettre en majuscule : *P.c.*, ou tout en minuscule *w.-c.* (car issu d'un nom commun).

L'acronyme est écrit sans point, tout en majuscule ou avec l'initiale en majuscule en cas de nom propre et en minuscule pour les noms communs.

Le prétraitement précédant leur phonétisation comporte deux étapes : la **détection** et la **décision du mode de prononciation**. La détection se base

sur des heuristiques simples de typographie (majuscules, avec ou sans point entre les lettres). L'analyse syntaxique filtre préalablement cette détection en autorisant les sigles uniquement en position de nom. Certains sigles et acronymes peuvent être présents dans le lexique ou faire partie de listes d'exceptions. La décision du mode de prononciation fait l'objet de nombreuses études spécifiques et s'appuie sur des principes phonologiques et phonotactiques, mais aussi sur des heuristiques plus simples, visant à classer les sigles à prononcer dans des catégories selon leur représentation en CV (consonne-voyelle). Par exemple, les sigles de 3 lettres de la forme CVC (CAC) ou VCV (*UFO*,*IRA*) sont généralement lus, et les autres épelés.

Notons que la lecture des sigles ne suit pas tout à fait les règles phonotactiques qui semblent établies en français dans la mesure où l'on peut rencontrer de nouveaux groupes de consonnes comme /fn/ présent dans FNAC ou AFNOR mais dans aucun mot du français.

Une fois le mode de prononciation décidé, l'acronyme sera lu par le système de phonétisation par règles et le sigle sera épelé par des règles ad hoc. Une amélioration envisagée consiste à prendre en charge les sigles mixtes (FGOLF) et les sigles à répétition de lettres (SSII lu comme SS2I) et à connecter ce module de phonétisation des sigles et acronymes à la gestion de la liaison (SSII / /, IEEE / /).

3.3.3. Les abréviations

L'abréviation désigne un mot ou une suite de mots réduits à l'une ou plusieurs de ses lettres (M. pour Monsieur, n° pour numéro, Cie pour compagnie). Elle se distingue de la réduction qui donne naissance à un mot (signifiant) nouveau aussi bien pour l'oral que pour l'écrit (comme *accu, ciné, cinéma, pneu, photo, métro* qui sont des apocopes, et *bus* ou *scope* qui sont des aphérèses), et du diminutif, qui est un mot nouveau et distinct comme signifiant et comme signifié (*fillette* pour *fille*). Les unités de mesures monétaires, scientifiques ou autres, accompagnant des nombres sont également considérées comme abréviations (12V, CHF 12.-, 12m03s, 12 min.)

Le but étant de gagner de la place et du temps, l'abréviation est usuellement réduite:
- à sa lettre initiale (M. pour Monsieur, le XIXème s. pour siècle, c-à-d. pour c'est-à-dire) ou à son début. S'il n'est pas réduit à sa lettre initiale, ce sera soit pour ne pas couper un digramme consonantique initial (Ch. pour Charles), soit pour des abréviations moins courantes (chap., hab.), soit quand l'initiale est une voyelle (ib.=ibidem). L'usage est de terminer le mot

abrégé à son début par une consonne. De plus, dans les deux cas, l'abréviation se termine par un point.
- à son début et sa fin (Mlle, Mgr, Me, Dr) et n'est usuellement pas terminé par un point.

Mais l'usage pouvant être moins strict, de nombreuses variantes attestées existent : pas de point (c-à-d), pas de tiret (*càd*), les lettres devant être placées au dessus de la ligne (XIX^{ème}) ne le sont pas (XIXème), pas d'espace entre le nombre et l'unité de mesure, que celle ci soit à gauche ou à droite du nombre.

La casse est importante et plus respectée que pour les sigles. En effet, *CM* peut désigner un cours moyen dans les classes primaires françaises ou un cours magistral dans le domaine universitaire, alors que *cm* désignera invariablement le centimètre. Contrairement au sigle et à l'acronyme, la lecture de l'abréviation implique la restitution d'une information non présente dans le texte (*cm* suivant un nombre ne sera jamais épelé mais prononcé comme *centimètre*) En fonction de la nature de l'abréviation, de sa fréquence d'utilisation et de la rigueur de l'auteur du texte, le point sera présent ou non.

Il existe plusieurs listes d'abréviations, certaines sont communes, d'autres divergent. En effet, des abréviations spécifiques peuvent être propres à un domaine terminologique particulier (terminologique, médical, pharmaceutique, chimique). Néanmoins de nombreuses abréviations courantes ne sont pas détournées vers d'autres significations dans ces domaines typiques et peuvent être intégrées au système.

3.3.4. Les symboles non alpha-numériques (% , ! ? «)

Les symboles non-alphanumériques ainsi que les caractères de ponctuation sont prononcés ou non selon le contexte dans lequel ils sont utilisés. Il est possible de distinguer :

- les adresses électroniques sur l'internet (http://www.unige.ch) ou sur les messageries e-mail (goldman@unige.ch) ;
- les références de documents WT/IFSC/W/6/Rev.1 ;
- les symboles suivants : & (et), § (paragraphe), = (égale), + (plus, phonétisé /plys/), £ (livre sterling), $ (dollar), % (pour cent), ° (degré ; ainsi que °C et °F).

3.3.5. Les noms propres et mots étrangers

Les noms propres sont souvent source d'erreurs de phonétisation et sont l'objet d'études importantes (projet européen ONOMASTICA). La

prononciation des noms propres est transparente pour les noms les plus courants, c'est-à-dire conforme à la prononciation des noms communs, mais peut être sujette à des variations. La preuve en est qu'il n'est pas rare qu'une personne rectifie la mauvaise prononciation de son nom par un tiers ou qu'on demande confirmation de la manière dont on prononce un nom, à la personne porteuse de ce nom. Outre les idiosyncrasies (l'exemple le plus connu étant *de Broglie* prononcé / /), ces variations sont principalement :

- des habitudes régionales (entre autre la prononciation des –x, -z, -s finaux: Chamonix, Ovronnaz, Saint-Gaudens) ;
- des assimilations diverses pour les noms propres d'origine linguistique étrangère. Par exemple *New-York* peut être prononcé / /, / / ou encore / /. L'assimilation peut être due à la fréquence du mot dans la langue courante, ou à la volonté d'une personne de (faire) prononcer son nom fidèlement à sa langue d'origine ou au contraire de s'intégrer à la prononciation francophone usuelle, ou simplement à l'ignorance des locuteurs francophones quant à la prononciation des mots d'origine linguistique étrangère.

Le traitement de noms propres est donc soumis à une tâche préalable consistant à deviner l'**origine linguistique** des noms propres. Par exemple, un nom à consonance germanique comme Pfeiffenberger verra sa finale prononcée en / / et non / / comme on le ferait pour Berger (dans un texte ou un contexte francophone uniquement) ou Duberger (quelque soit le contexte). Cet exemple justifie le prétraitement plutôt que le simple ajout de règles de phonétisation ad hoc car il montre des prononciations ambiguës de terminaisons de mots.

La détermination de l'origine linguistique peut se faire par des stratégies à connaissances explicites comme la présence d'un préfixe ou d'un suffixe caractéristique et non ambiguë d'une langue. Par exemple, *Pfeiffen-* sonne germanophone et déclenchera la prononciation correcte de *–berger*. Plus généralement, on trouve des systèmes reposant soit sur des règles (Carlson 1989), (Belhoula 1993), soit sur des techniques à base d'analogie, soit sur des modèles statistiques (Vitale 1991).

Nos propres travaux nous ont conduit à l'élaboration d'un mécanisme basé sur deux de ces stratégies : des règles portant sur les caractères graphiques, les préfixes et les suffixes discriminants d'une part, et une analyse statistique utilisant des n-grammes.

La première méthode permet d'attribuer une étiquette d'origine linguistique à environ 40% des noms grâce à des listes d'affixes propres

aux langues ou groupes de langues considérées. Par exemple, voici la liste des suffixes anglophones discriminants : *black, brigh, bridge, broad, brown, east, good, green, light, south ,stone ,weat ,white ,wood.* La méthode statistique émet une hypothèse sur les noms inclassables par la première méthode par un apprentissage préalable des suites de 2,3 et 4 lettres les plus fréquentes. Une étude interne portant sur un corpus de 2600 noms propres, montre que ces deux approches combinées déterminent correctement l'origine linguistique à plus de 85% (Ndiaye 2001).

Dans le cadre notre phonétisation par règles, cette information sur l'origine linguistique active un jeu de règles supplémentaires propres à l'assimilation d'un nom d'origine linguistique étrangère vers une prononciation francophone. Nous nous cantonnons à une assimilation assez brutale puisqu'elle n'autorise pas de phonèmes non francophones. Selon l'étude de (Mengon 2001) qui adopte une approche plus phonologique de l'assimilation des mots d'emprunts anglophones et germanophones, les différences principales sont en autres le raccourcissement des voyelles longues (*tea* /ti:/>/ti/), la présence modérée de diphtongues (*road* /r«Ud/>/ /) et le déplacement de l'accent lexical vers la fin du mot. Ces règles d'assimilation seront autant de règles de phonétisation qui sont ajoutées momentanément au jeu de règles principal pour le français. Elles sont prioritaires sur les règles habituelles.

- ea+# => i
- 'C'+oa+'C' => 'o'

Ajoutons des règles de phonétisation qui diffèrent de la prononciation usuelle du français comme par exemple les affriquées en début de mot (*James*, *change*) en général correctement prononcés par les francophones :

- #+j+'V'=> / /
- #+j+'V'=> / /

Des améliorations possibles sont déjà envisagées, notamment la possibilité de prononcer un nom propre étranger selon un degré d'assimilation réglable. Pour ce faire, nous devons envisager le classement des règles supplémentaires propres à une langue en plusieurs niveaux. Une limitation se fait déjà sentir quant aux phonèmes possibles dans une langue étant donné le caractère figé des bases de diphones du codeur de parole utilisé, MBROLA.

3.4. Phonétisation de *plus*, *tous* et quelques nombres

Il existe un certain nombre de mots pour lesquels la réalisation phonétique peut varier en fonction du contexte sémantique et syntaxique. C'est le cas de ***plus*** (*J'en veux plus* vs. *je n'en veux plus*), ***tous*** (*ils sont tous partis* vs. *tous les enfants*) ainsi que de certains nombres comme ***cinq, six, huit, neuf,***

dix, vingt (*Il y en a cinq. Cinq enfants. Cinq gamins. Cinq cents bambins*). On peut noter que dans les différents exemples, la consonne finale se fait entendre ou non (/ply/ vs. /plys/). De plus, la liaison peut provoquer une troisième variante phonétique du même mot (*Je ne veux plus avoir d'ennuis*) dans laquelle la consonne finale 's' se vocalise en /z/ : /plyz/.

Cette variation phonologique très particulière apparaît dans la prononciation de peu de mots mais la fréquence d'utilisation de ces mots est suffisamment élevée pour que l'on s'y attarde.

Dans les systèmes de phonétisation existants, la prédiction de la prononciation de *plus* est uniquement contextuelle (*plus* est en finale de phrase, suivi d'un mot commençant par une consonne ou suivi d'un mot commençant par une voyelle), ce qui tend à sous-estimer le nombre de cas où ce mot conserve le /s/ final. Dans « *plus que la moitié »* , *plus* est un déterminant quantifieur partitionnel alors que dans « *il s'est plus ennuyé » (qu'amusé)*, *plus* a un lecture adverbiale modificateur de verbe. Nous pensons qu'une analyse syntaxique est appropriée pour traiter ces exemples puisque la configuration syntaxique est complètement déterminante pour la prononciation de *plus* (Goldman 1999).

La phonétisation de *tous* est encore plus simple car sa catégorie suffit pour le prononcer correctement : en tant que déterminant-quantifieur (comme dans *tous les enfants, tous azimuts*) *tous* est prononcé /tu/ ; en tant que pronom (*Tous arriveront, ils sont tous partis, Je leur ai dit à tous*), le /s/ final est conservé.

Pour les nombres comme *cinq, six, huit, dix et vingt*, la configuration syntaxique en plus du contexte phonétique (précédent une consonne, une voyelle, une pause) permet, comme dans le cas de *plus*, de prédire avec justesse la prononciation de la consonne finale. Comparez : *six francs / six de ces hommes*. Dans ces deux exemples, *six* précède une consonne mais se prononce différemment selon qu'il précède un nom ou une préposition.

3.5. Le schwa et son élision

Le schwa, appelé également e-muet ou e-caduc, est un phonème instable pouvant disparaître dans certaines conditions. Ce schwa peut se trouver en début de mot (*f**e**nêtre*), en milieu de mot (*sam**e**di*) ou en fin de mot (*j**e** m**e** l**e** demande*). Une séquence de schwas, comme dans le dernier exemple, peut donner plusieurs prononciations possibles (*j'me le d'mande, je m'le d'mande, j'me l'demande*,…) ainsi que dans le mot *genevois* (*gen'vois* ou *g'nevois*). Les liens entre l'élision du schwa et la linguistique étant limités, nous nous contentons de décrire brièvement la problématique, les études portant sur ce sujet et le traitement du schwa dans notre système.

Les facteurs influençant la chute, ou amuïssement, du schwa sont d'ordre phonétique (contexte phonétique), articulatoire (selon Malécot 1955, la chute du schwa dépend de la différence d'aperture entre les consonnes accolées), prosodique (accentuation du mot), linguistique (nature du mot, sa fréquence – comparez *cerise* et *merise*) et phonostylistiques (conversation spontanée ou lecture oralisée, langage familier ou soutenu, débit de parole, accent régional). En plus de ces domaines, la chute du schwa soulève d'autres questions en traitement automatique du langage ainsi qu'en psycholinguistique comme la représentation lexicale du schwa et l'influence sur l'accès au lexique mental (Racine 2001), (Boula de Mareuil 1997), (Dard 2001).

Les études les plus connues visant à modéliser ce phénomène phonologique sont celles de Grammont dès 1894, puis celles de Malécot qui reprennent les premières. Le résultat principal est l'établissement de la règles dite des 3 consonnes, qui consiste à interdire la chute du schwa si celle-ci provoque une séquence de 3 consonnes (sa**m'd**i mais *mer**cr'd**i) à moins que la troisième ne soit une liquide ou une semi-voyelle. De plus si la deuxième consonne est une liquide, l'élision est également interdite (ex : *donnerions* et *Richelieu*). De nombreux cas particuliers existent et demandent des traitements spécifiques comme les paires minimales de mots (le schwa de *belette* ne s'élide pas à cause de l'existence de *blette,* tout comme *peluche* et *pluche* malgré leur éthymologie commune), le h aspiré qui bloque l'élision (**l'hibou*) sauf dans les nombres (*quarant'huit*), le 'de' nobiliaire qui n'est jamais élidé (*Madame de Fontenay*)…

Dans le système FipsVox, en plus de la règle de trois consonnes, sont implémentées quelques règles phonotactiques interdisant la création de groupes de consonnes peu fréquents ou le redoublement de la même consonne. De plus, l'épenthèse d'un schwa peut survenir en finale de mot comme dans *La boucle brille.*

3.6. Liaison

La liaison est un autre phénomène phonologique caractéristique du français. Elle apparaît lorsqu'un mot se terminant par une consonne latente est suivi d'un mot débutant par une voyelle ou un h muet. Cette consonne est alors prononcée et enchaînée à la voyelle du mot suivant. L'adjectif *petit* par exemple est prononcé / / dans la séquence *petit ami.* La liaison peut entraîner un changement de phonème par exemple lorsque la consonne orthographique finale est un *d* : la liaison se réalise alors avec / / (par exemple dans *grand ami*).

La liaison est apparue avec l'évolution de la langue notamment la disparition de la prononciation des consonnes finales puis des voyelles post-accents. On distingue les liaisons obligatoires, les liaisons facultatives et les liaisons interdites[3].

Dans le cadre de la synthèse de la parole, le traitement de la liaison est primordial car la non-réalisation d'une liaison obligatoire ou la réalisation d'une liaison interdite peut nuire considérablement à la compréhension. De plus, le contexte immédiat du lieur et du lié peut ne pas suffire :

- L'homme qui est avec vous a oublié son livre
- L'homme vous a oublié

Les facteurs principaux influençant la liaison sont la cohésion syntaxique et la fréquence des deux mots mis en jeu. Ils déterminent les liaisons obligatoires et interdites. L'approche par l'analyse syntaxique est encore une fois justifiée.

La réalisation des liaisons facultatives dépend d'autres facteurs comme la structure phonologico-prosodique qui dépend elle-même en partie de la structure syntaxique, comme également les propriétés morphologiques et lexicales (nature de la consonne de liaison, nombre de syllabes). À cela, il faut ajouter des facteurs extralinguistiques comme le style de parole, la classe socio-économique du locuteur, l'âge, le sexe et la variation régionale.

De plus certains ajustements phonologiques supplémentaires sont à considérer: la dénasalisation (*un bon ami* se prononce / / et non / /) et l'ouverture du /é/ (*le dernier homme* se prononce / / et non / /).

Notre système se base sur des informations catégorielles et configurationnelles pour déclencher la liaison. En effet plusieurs conditions sont nécessaires : un contexte de liaison valide (c-à-d une consonne latente de liaison suivie d'une voyelle ou d'un h muet), une paire de catégories valides (comme déterminant-nom *les enfants*, pronom-verbe *ils attendent, il les attend*) et la c-commande. Cette dernière condition est une configuration syntaxique bien précise, nécessaire à la liaison selon Selkirk (1978).

Une étude interne portant sur tous les facteurs influençant la réalisation de la liaison facultative dans un corpus de parole d'une durée totale de 5 heures (6 femmes et 4 hommes), a conclu que le débit avait aussi une

[3] Ainsi que les fausses liaisons et les pataquès, issus d'erreurs de production que nous ne traiterons pas.

influence sur la liaison. Nous affinons actuellement le traitement de la liaison afin de diviser les liaisons facultatives en plusieurs niveaux de réalisation. Ce taux de réalisation pourra être modifié directement ou indirectement selon le paramétrage du style de parole et du débit.

4. Prosodie

La génération de la prosodie est une étape importante dans le processus de synthèse de la parole car, au-delà de l'intelligibilité pour laquelle la phonétisation et la prosodie sont indispensables, la prosodie apporte également le naturel à la voix de synthèse en lui conférant une intonation, un rythme et une énergie ressemblant à une voix humaine. La prosodie, qui est composée de l'ensemble de ces paramètres, peut être déterminée par de nombreuses façons. La méthode choisie dépend de plusieurs critères tels :

- le modèle prosodique théorique sous-jacent : on distingue principalement les modèles phonologiques, les modèles linéaires et les modèles superpositionnels. Nous verrons plus loin quelle stratégie nous est dictée par l'utilisation de l'arborescence syntaxique fournie par l'analyseur ;
- les informations fournies par les modules en amont. Par exemple, le traitement linguistique détermine pour chaque mot si c'est un mot plein ou un mot outil, mais aussi l'arborescence des syntagmes que nous pensons intéressante à exploiter pour éviter des regroupements prosodiques malencontreux. D'autre part, la phonétisation donne la chaîne phonétique, et par ce biais le nombre de syllabes indispensable pour l'équilibrage des groupes prosodiques. De plus, l'élision du schwa, habituellement déterminée lors de la phonétisation, peut être différée au moment de la génération de la prosodie pour prendre part à l'équilibrage des groupes prosodiques, car nous pensons que l'eurythmie est un des facteurs de l'élision ;
- le codeur (l'ultime module qui transforme la chaîne phonétique et les paramètres prosodiques en signal sonore) qui peut être conçu pour restituer d'authentiques extraits de courbes mélodiques naturelles stockées ou, au contraire, pour produire une mélodie fidèlement aux paramètres prosodiques déterminés. Dans les synthétiseurs concaténant des unités de longueur variable, ces dernières sont stockées et restituées avec leur prosodie initiale, ce qui signifie que les paramètres prosodiques calculés peuvent être minimaux car les informations stockées assureront la finesse et le naturel de la voix synthétique. En contre partie, ces courbes sont figées et il est impossible de produire une intonation originale absente de la base de données. Les systèmes

concaténant des diphones « plats » sont plus souples mais nécessitent plus de précision dans la prédiction des cibles mélodiques car elles seules assurent la courbe mélodique globale.

Nous avons choisi un système puisant beaucoup d'information dans le traitement linguistique selon l'hypothèse d'un lien fort entre prosodie et syntaxe. Par manque de place, nous renvoyons à Goldman (1997a) et Mertens (2001a) pour la justification de ce choix et une description détaillée. Signalons quand même que ce système est basé sur le modèle prosodique de Mertens, et adopte une approche récursive exploitant la structure syntaxique issue de l'analyseur. Les étapes successives sont le calcul des groupes accentuels (GA), le calcul des groupes intonatifs (GI), la syllabation de la chaîne phonétique, l'accuentation et finalement la génération de la prosodie superposant la déclinaison globale et la modulation de l'intonation dans le registre du locuteur. Le calcul des durées se fait en établissant d'abord la durée de la syllabe en fonction de l'accentuation de celle-ci et de sa position dans le groupe intonatif, puis est calculée la durée des phonèmes de cette syllabe (en répartissant la durée syllabique vers chacun des phonèmes).

En plus de l'intonation d'une phrase neutre, le système est en mesure de détecter automatiquement des structures syntaxiques spéciales telles que les clivées, les incises et les extraposées, et pour lesquelles un patron intonatif original est appliqué (Goldman 1997b). Par exemple la phrase *c'est le lapin que j'ai adopté* sera prononcée différemment selon qu'elle est la réponse à *As-tu adopté le lapin ou le hamster ?* ou *Qu'as-tu dans les bras ?* Dans le premier cas, la réponse est une clivée et le groupe *le lapin* doit être mis en focus, alors que dans le second exemple, c'est une simple relative et une intonation plus neutre est attendue.

Pour ce qui concerne des tournures plus spéciales propre à l'analyse du discours, les limitations actuelles des outils de linguistique informatique ne nous permettent pas un traitement entièrement automatisé. Néanmoins, nous avons mis en place un système de balisage de type XML, autorisant une annotation préalable d'un texte. Chaque balise, indiquant par exemple le discours direct, ou bien un style emphatique ou ironique, indiquera au système qu'il doit adopter une accentuation spéciale pour la portion de texte concernée (Mertens 2001a).

5. Réalisation et conclusion

L'implémentation du système a été pensée de manière à ce que d'une part plusieurs applications puissent être développées autour du moteur d'analyse linguistique Fips et, d'autre part, que plusieurs langues soient

disponibles. En effet, des outils de traitement automatique de la langue comme des traducteurs automatiques, des dictionnaires bilingues parlants et des outils d'enseignement des langues sont développés autour de Fips, et nous tentons de les rendre utilisables dans plusieurs langues comme le français, l'anglais, l'allemand, l'italien, l'espagnol, le grec, le vietnamien,… Pour ce faire, une organisation modulaire est nécessaire et, pour éviter un temps de développement considérable pour chaque nouvelle langue, des modules génériques sont mis en œuvre, exploitant au mieux les bases linguistiques communes à toutes les langues tant au niveau de l'analyse syntaxique que des traitements phonétiques et prosodiques. Les modules et ressources propres à chaque langue sont les lexiques, les spécificités syntaxiques, les jeux de règles de phonétisation, les règles de génération de la prosodie, les durées par défaut des phonèmes, les bases de diphones.

L'évaluation d'un tel système est une tâche délicate car le nombre important de traitements successifs de la phrase à la parole rend difficile le diagnostic d'une erreur perçue dans la parole synthétique résultante. Des scores d'opinion sont généralement utilisés mais l'objectivité des auditeurs-test n'est pas facile à garantir tout au long de l'expérience. Seule la phonétisation peut être évaluée quantitativement car les informations résultantes peuvent être comparées à une solution de référence.

Le but initial du projet FipsVox qui est de montrer l'importance des données linguistiques pour le traitement de la parole a été atteint car nous avons justifié l'intérêt de l'apport d'informations syntaxiques dans la plupart des sous-tâches de la synthèse de la parole. Nous avons de plus réalisé un système de synthèse de la parole complet, multilingue et utilisable pour des applications diverses comme la lecture orale de textes (livres, articles on-line, mail) ou de messages plus courts (applications informatiques, serveurs téléphoniques, aide à la navigation sur internet…) par des non-voyants ou dans des situations où l'accès visuel à un écran n'est pas possible.

Les travaux futurs s'inscrivent dans le cadre multilingue des outils linguistiques développés au LATL puisque les lexiques phonétiques et les jeux de règles pour l'anglais, l'allemand, l'italien et l'espagnol sont en cours de réalisation. Nos efforts se portent aussi sur l'amélioration de la prosodie pour une voix plus naturelle tant au niveau phrastique que discursif.

Bibliographie

BOULA DE MAREUIL Ph. (1997), *Etude linguistique appliquée à la synthèse de la parole à partir du texte*, Thèse de doctorat, Université Paris XI.

BELHOULA K. (1993), « Rule-based grapheme-to-phoneme conversion of names », *Eurospeech* 881-884.

CARLSON R., GRANSTRÖM B., LINDSTRÖM A. (1989), « Predicting name pronunciation for reverse directory service », *Eurospeech* 113-116.

DARD A., GUELAT L., JAEGER C. (2001), *Variation dans la réalisation des liaisons et des élisions de schwas en français : effet du style de parole, du débit, du sexe et de la nature des liaisons et des élisions,*Mémoire de licence, Université de Genève.

DUTOIT T. (2001), *MBROLA : http://tcts.fpms.ac.be/synthesis/mbrola.html*

ENCREVÉ P. (1988), *La liaison avec et sans enchaînement*, Paris, Seuil.

GAUDINAT A. & WEHRLI E. (1997), « Analyse syntaxique et synthèse de la parole : le projet Fipsvox », *TAL* 38. pp.121-154

GOLDMAN J.-PH. (1997), *Génération automatique de l'intonation en français : cas des syntagmes extraposés*. Mémoire de DES, Université de Genève.

GOLDMAN J.-PH. & WEHRLI E. (1997), « Deriving prosodic patterns from syntactic structures : the case of extraposition, clefts and extraposition », *ESCA workshop on Intonation : Theory, Models and Applications, in Botinis A eds., Athens, Greece, pp.153-156.*

GOLDMAN J.-PH., LAENZLINGER C.(1998), *La micro-grammaire des nombres*, notes techniques internes.

GOLDMAN J.-PH., LAENZLINGER C., WEHRLI E.(1999), « La phonétisation de *plus*, *tous* et de certains nombres : une analyse phono-syntaxique », *TALN* Cargèse.

GRÉVISSE (1990), *Le bon usage. Précis de grammaire française*, Paris, Duchot.

MALÉCOT (1955), « The elision of the French mute-e within complex consonantal clusters », *Lingua* V, 1, 45-60.

MERTENS P., AUCHLIN A., GOLDMAN J.-PH., GROBET A. & GAUDINAT A. (2001), « Intonation du discours et synthèse de la parole : premiers résultats d'une approche par balises », ce volume.

MERTENS P., GOLDMAN J.-PH., WEHRLI E. et GAUDINAT A. (2001), « La synthèse de l'intonation à partir de structures syntaxiques riches », *TAL* 42/1.

MUSILLO G. (2001), *Rapport technique interne sur le système FipsPhonetizer*, 2001

NDIAYE M. (2001), *Identification de l'origine linguistique des noms propres*, mémoire de DEA Linguistique Informatique.

RACINE, I & GROSJEAN F. (2000), « La fenêtre ou la f'nêtre: le E caduc facultatif l'est-il réellement? », submitted.

SELKIRK E. (1974), « French liaison and the X-bar convention », *Linguistic Inquiry* 5, 573-590.

VITALE (1991), « An algorithm for high accuracy name pronunciation by parametric speech synthesizer », *Computational Linguistics* 17(3), 237-276.

WEHRLI E. (1997), *L'analyse syntaxique des langues naturelles: problèmes et méthodes*, Paris, Masson.