LA PHONÉTISATION DU CASTILLAN

CABRERA C., CONTINI M. et BOË L-J.

Institut de la Communication Parlée, URA CNRS n° 368
Grenoble, France

ABSTRACT

Our project was the establishment of a grammar for the automatic phoneticization of Spanish. By examining a lexicon of 65.000 words and systematically examining their transcriptions, we formulated a large body of rules. In a next step, we will use this knowledge for a text-to-speech synthesis application. We constituted a data base of 2500 words. The resulting system gives a correct phoneticization of 98% of the original lexicon. We here present the analysis method used on this large lexicon, as well as a selection of the rules derived.

1. INTRODUCTION

La phonétisation automatique consiste à passer d'une chaîne orthographique quelconque à une chaîne phonétique. Cette transcription en A.P.I. ou dans un autre code relève du domaine de la description linguistique et peut être utilisée dans une application telle que la synthèse de la parole. Ses intérêts sont multiples et apparaissent de plus en plus comme l'étape nécessaire pour l'établissement du dialogue homme-machine. Sur le plan hispànique, ce champ d'étude a déjà fait l'objet de récents travaux [1].

Résultat d'une collaboration étroite entre linguistes et informaticiens, l'outil de phonétisation multilangue qu'est TOPH [2], défini dans le cadre de la synthèse à partir du texte, présente l'avantage pour le linguiste de formaliser facilement sa connaissance. Conçu comme un module adaptable à chaque langue orthographique visée en l'occurrence le français, l'allemand et l'italien, cet outil a donné lieu au développement de grammaires de transcription pour chacune de ces langues. Cette étude se veut une description linguistique des phénomènes de phonétisation mais une étape ultérieure consistera à l'intégration de ces connaissances dans un système de synthèse (SYNTALIT).

Notre contribution consiste en l'établissement d'une grammaire de règles de transcription orthographique-phonétique utilisant le formalisme TOPH pour le castillan normatif (prononciation de l'espagnol madrilène cultivé).

2. PRÉSENTATION DE TOPH

Formalisation de grammaires de transcription, TOPH a été réalisé afin de proposer une description concise des phénomènes de phonétisation. Le logiciel élaboré s'articule autour des éléments

syntaxiques suivants :

- L'unité linguistique sélectionnée est la chaîne graphémique
- Déclaration d'ensembles de natures diférentes à savoir les ensembles linguistiques et les lexiques d'exceptions.
- Le linguiste formalise son raisonnement sous la forme d'une grammaire déterministe (à une quelconque sous-chaîne d'un mot correspond une seule transcription) de règles de réecriture contextuelles.
- Ordonnées du particulier au général, les règles sont regroupées par classes, avec un ordre local pour chaque règle, défini par son ordre d'écriture.
- Possibilité d'insertion de commentaires dans la grammaire bornés par !

L'intérêt de TOPH réside dans l'accès à des traces de réalisation des règles sollicitées de même qu'à des résultats statistiques sur ces dernières.

3. GRAMMAIRE DU CASTILLAN

La grammaire a été élaborée sur la base de 65000 entrées lexicales issues du dictionnaire SGEL [3] dont la particularité, outre les transcriptions attachées à chaque entrée, réside dans l'introduction de nombreux emprunts (anglicismes en majorité) plus ou moins assimilés au phonétisme du castillan. A l'aide d'un ensemble de règles la correspondance phonétique de chaque graphème est définie en tenant compte de toutes ses distributions possibles. L'apport constant de termes nouveaux auxquels une langue naturelle est soumise nécessitera une mise à jour régulière de notre grammaire. Ceci pose évidemment le problème de la pertinence des lexiques liée à leur actualisation.

Pour la prononciation standard du castillan nous nous référons à des ouvrages spécialisés [4], [5], [6] . Nous nous appuyons en outre sur le dictionnaire SGEL (mentionné précédemment) à partir duquel nous avons dressé des listes d'exceptions pour chacune des 29 lettres de la langue. Ces listes contiennent toutes les réalisations phonétiques déviantes ou supplémentaires par rapport aux règles mentionnées dans les travaux déjà signalés cela dans le but de répertorier toutes les occurrences allophoniques pour une chaîne graphémique donnée afin de construire une grammaire de phonétisation la plus complète possible. Après ce premier travail d'identification et de synthèse, nous nous sommes attachés à l'édification et à la codification de la grammaire proprement dite pour laquelle nous avons déclaré les éléments décrits ci-après :

- 12 ensembles répartis en ensembles linguistiques par exemple :

a) "semi-consonnes" = (y, w)

b) séparateur de mots
"#" = (-,., ,:, ;,",!)

c) "except : i" = (articulad, angular, unívoc, áxic, auricular, atómic, ocinética, odegradable, odegradación, odinámica, ofísica, ograf, ográfic, ógrafo, ología, ológico, ólogo, oluminiscencia, omasa, omecánica, ometría, opsia, oquímico, osfera, osíntesis, oterapia, otico, otita, otropismo, óxido, al, ofita,os, ozoo, alin, ato, ante, ogloso, oide,able, abilidad, enio, enal, edro, ásico, ar, ángulo)

Cet ensemble d'exceptions nous permet d'écrire la règle :
("#"+b,br,h,tr,v) +i+ ("except : i") = [i]
sachant que la règle générale (majoritaire)

est :
("consonne") +i+ ("voyelle sauf i") = [j]

4. RÉSULTATS

A la lumière des résultats, plusieurs remarques s'imposent. Il apparaît que si l'on ne considère que les règles de prononciation circonscrites aux phénomènes réguliers, autrement dit sans tenir compte des exceptions ou des emprunts, la phonétisation du castillan se résume à une soixantaine de règles élémentaires. A titre illustratif, nous nous limiterons au cas du graphème "g". Alors que ce graphème est communément défini comme se réalisant selon 3 allophones , il s'enrichit de nombreuses réalisations lorsque nous étendons la grammaire à l'étude des emprunts et autres exceptions (entigrecerse).
Ainsi si l'on considère le trait régulier, un graphème comme "g" sera traité par 3 règles:

+g+ (e,i) = [x]
("#", n) +g+ = [g]
+g+ = [ɣ]

En revanche, il en faudra 19 si l'on tient compte, par ailleurs, des emprunts:

("#"+gro,buldo) +g+ ("#") = [g]
("#"+zigza,iceber,basi) +g+ ("#") = [x]
("#"+ban,campin,smokin,bumeran,boomeran,rin,puddin,pin,pon,parkin,marketin,gon,dumpin,dopin) +g+ ("#") = [ ]
("#"+tun) +g+ (steno) = [ ]
("#"+rémin) +g+ (ton) = [ ]
("#"+gan) +g+ (ster+ismo,#) = [ ]
("#"+copyri,bri) +g+ (ht) = [ ]
("#"+neglig) +g+ (é) = [j]
("#"+sufra) +g+ (is+t,m) = [ɣ]
("#"+he) +g+ (elia+n,nismo) = [ɣ]
("#"+per) +g+ (ola) = [g]
("#"+lori) +g+ (a) = [g]
("#"+ideolo) +g+ (o) = [g]
("#"+enti) +g+ (recerse) = [g]
("#"+cat) +g+ (ut) = [g]
+g+ (e, i) = [x]
("#",n) +g+ =[g]
("#") +g+ ("#") = [xe]
+g+ = [ɣ]

Les règles ont été testées sur une base de données conséquente et notamment un dictionnaire de 2500 entrées, implémenté sur HYPERCARD (Macintosh) contenant formes orthographiques et phonétiques de référence. Actuellement 484 règles et 3 ensembles d'exceptions permettent de phonétiser automatiquement ce corpus. Nous obtenons un taux de succès de 98%.

5. CONCLUSION

Dépourvu d'homographes hétérophones, le castillan s'avère être une langue relativement régulière quant à un processus de phonétisation. Néanmoins si l'on considère la manière dont elle intègre les emprunts, nous constatons que ces apports lexicaux désorganisent quelque peu le phonétisme de cette dernière ou du moins n'obéissent pas aux règles de prononciation standard. Cependant quelquefois ils semblent avoir été pratiquement totalement assimilés par la langue (pour les plus anciens) et nous obtenons alors deux prononciations possibles pour une même unité lexicale, une se fondant sur le phonétisme castillan et l'autre conservant les traits de la langue d'origine (bridge, chauvinismo, chauvinista). Materiau vivant, la langue nécessitera pour son étude la constante réactualisation de nos règles ainsi que le renouvellement des lexiques établis.

RÉFÉRENCES

[6] ALCINA FRANCH & MANUEL BLECUA (1988), *Gramática española.* Editorial Ariel, Barcelona, 277-401.
[2] AUBERGÉ & al. (1987), TOPH : un outil de phonétisation multilangue, *Bulletin de l'Institut de Phonétique de Grenoble*, Vol 16, 155-176.
[3] *Gran Diccionario de la lengua española.* (1989), S.G.E.L., Madrid.
[4] NAVARRO TOMÁS (1970), *Pronunciación española.*, Decimoquinta edición, Publicaciones de la revista de filología española, Madrid.
[5] QUILIS & FERNÁNDEZ (1969), *Curso de fonética y fonología españolas*, Cuarta edición, C.S.I.C., Madrid.
[1] SANTOS & al. (1984), Real time text to speech conversion system for spanish, *IEEE-ICASSP*, San Diego, 1593-1596.