PHONÉTISATION AUTOMATIQUE : ÉVALUATION ET ENRICHISSEMENT D'UNE GRAMMAIRE DE RÈGLES POUR LE FRANÇAIS

R. Belrhali, L. Libert, L.J. Boë

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 French. By constituting a lexicon of 60.000 words and systematically examining their transcriptions, we formulated a large body of new rules, which were added to a pre-existing base set, making a total of 900 rules. The resulting system gives a correct phoneticization of 99.75% of the base 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 du français peut être décrite essentiellement par règles de correspondance entre graphèmes et phonèmes. Cette correspondance se décrit par la contrainte de contexte sur la chaîne graphémique et va même jusqu'à la correspondance lexicale. Les phénomènes de phonétisation tiennent comptent de niveaux linguistiques supérieurs au mot que sont : la valeur catégorielle (souvent [suvɑ̃], chantent [ʃɑ̃t]), la fonction syntaxique (les portions [pɔʀsjɔ̃], nous portions [pɔʀtjɔ̃]), la structure syntaxique : les liaisons, (un savant [Ø] aveugle (nom - adjectif), un savant [t] aveugle (adjectif -nom)), la valeur sémantique (fils [fis], fils [fil]).

Ce travail concerne le niveau lexical de la phonétisation du français.

Avec la mise en place de bases de données faciles d'accès, de manipulation et de reconfiguration, il est maintenant possible d'élaborer, tester et améliorer les formalisations possibles. Notre effort s'est concentré sur les relations entre codes orthographiques et de vastes corpus de notations phonétiques.

2. PRÉSENTATION DU LOGICIEL TOPH

L'outil *TOPH* (Transcription Orthographique-Phonétique) est un phonétiseur multilangue qui propose une syntaxe pour décrire des grammaires de phonétisation. Ce transducteur fonctionne sur texte libre. Il permet de réécrire une chaîne d'entrée graphémique en une chaîne de sortie phonémique. Les avantages de TOPH par rapport à d'autres logiciels (cf. [3], [4], [6], [8]) sont certains. Nous pouvons mentionner sa facilité d'utilisation (traces d'application, statistiques) ainsi que la formalisation de ses règles transparente à l'utilisateur, permettant des modifications aisées. L'expert formalise son raisonnement sous la forme d'une grammaire déterministe de règles de réécriture contextuelles.

A chaque classe de règles il introduit un ordre local défini par l'ordre d'écriture des règles. La grammaire comprend :

1°- des ensembles prédéfinis de caractères orthographiques décrivant des phénomènes de nature très différente :

- *ensembles linguistiques* : "Consonnes non nasales" = (b, c, ç, d, f, g, h, j,k, l, p, q, r, s, t, v, w, x, z)
- *ensembles d'exceptions* : "Exception : fin en g " = (barlong, bastaing, basting, bourg, oing, seing, dugong, écang, étang, hareng, harfang, joug, kaoliang, long, pacfung, parfaing, rang, sampang, sanderling, sang, shampoing, shampooing, trépang, tripang)

2°- des *commentaires* pouvant être une chaîne quelconque bornée par '!' et insérée dans n'importe quelle portion de la grammaire.

3°- des *règles* partitionnées en classes ; la classe d'une règle étant déterminée par le premier caractère de la chaîne à transcrire.

3. MÉTHODOLOGIE

Afin d'enrichir la grammaire de phonétisation existante, un lexique de grande taille est nécessaire. Nous avons donc, dans un premier temps, constitué une *base de données* de 60 000 mots implantée sur *Macintosh*. L'environnement *Hypercard* et le langage *Hypertalk* ont rendu possible la mise au point de programmes de recherche de chaînes orthographiques de longueur quelconque. Elles ont été recherchées dans trois positions : *initiale, interne, finale*. A partir des listes obtenues nous avons systématiquement relevé la transcription phonétique de la chaîne étudiée en prenant comme référence de prononciation le *Petit Robert 1*. Nous avons ensuite vérifié l'existence de la (ou des) règle(s) correspondante(s) à la (ou aux) transcription(s) phonétique(s) de la chaîne de caractères étudiée. Dans le cas contraire, nous avons écrit de nouvelles règles.

Illustration de la méthode de travail par un exemple : la classe du 'b'.

Nous avons obtenu 2394 mots commencant par 'b-', 7210 mots contenant au moins un '-b-' en position interne et 32 mots se terminant par '-b'. Après le relevé de la prononciation du graphème 'b' dans tous les mots et dans toutes les positions nous avons établi les règles suivantes :

(les caractères syntaxiques sont notés en gras)

- (radou, lom) +b+ ("#", s) = [ ]

Cette règle concerne radoub et les mots se terminant par 'lomb' comme plomb, coulomb, surplomb, aplomb, dont la réalisation du '-b' en position finale est muette (ces mots peuvent être suivis d'un 's', marque du pluriel).

- +b+ (s, t) = [p]

Cette règle concerne tous les mots contenant la suite de caractères 'bs' ou 'bt' et dont le 'b' se réalise [p] ; il s'agit ici d'un cas d'assimilation régressive.

- (sub) +b+ (sidence, sidiaire, sist) = [b]
- (lam) +b+ (swool) = [b]

Ces deux dernières règles sont des exceptions à la précédente.

- +b+ (c, k) = [p]

Cette règle concerne tous les mots contenant la suite de caractères 'bc' ou 'bk' dont le 'b' se réalise [p]. Nous avons ici un autre cas d'assimilation régressive. La seule exception à cette règle est la suivante :

- (su) +b+ (carpatique) = [b]
- +bb+ = [b]

Toutes les géminées de la classe du 'b' obéissent à une règle unique.

- (“#”) +b+ (“#”) = [be]
Cette règle est uniquement applicable à la lettre de l'alphabet.
- +b+ = [b]
Il s'agit de la règle la plus générale.

Classement suivant l'ordre d'application des règles :
(radou, lom) +b+ (“#”, s) = [ ]
(lam) +b+ (swool) = [b]
(sub) +b+ (sidence, sidiaire, sist) = [b]
+b+ (s, t) = [p]
(su) +b+ (carpatique) = [b]
+b+ (c, k) = [p]
+bb+ = [b]
(“#”) +b+ (“#”) = [be]
+b+ = [b]

4. RÉSULTATS

La grammaire de base [1] contenait 200 règles et 12 ensembles d'exceptions. Actuellement *900 règles* et *16 ensembles d'exceptions* (décrivant 1 000 mots) permettent de phonétiser automatiquement les 60 000 mots de notre base de données avec un taux de réussite de *99,75%* (problème de polyphonie des mots du type 'plus' pouvant se prononcer [plys] ou [ply]). La langue, matériau vivant, est en constante évolution d'où la nécessité d'une *réactualisation systématique* de notre base de données et de la grammaire.

5. CONCLUSION

Au-delà des applications évidentes en synthèse et reconnaissance de la parole, le passage du niveau orthographique au niveau phonétique renvoit à des problèmes linguistiques fondamentaux et constitue un champ de validation privilégié des formalisations linguistiques et phonétiques. Le développement des Industries de la Langue constitue à la fois une stimulation et une possibilité directe d'application.

6. RÉFÉRENCES

[1] AUBERGE V. (1985)
Contribution à la phonétisation automatique des langues alphabétiques : le langage "TOPH". Rapport de DEA, CRISS, Département d'Informatique et Mathématiques appliquées aux Sciences Sociales, Université des Sciences Sociales de Grenoble.

[2] CATACH N. (1984)
La phonétisation automatique du français. Edition du CNRS., Paris.

[3] DIVAY M. & GUYOMARD M. (1979)
Le compilateur de règles de réécriture TOP et son utilisation à la transcription du français en vue de la synthèse. *10èmes J.E.P.-G.A.L.F.*, Grenoble, 202-211.

[4] FERVERS H., LE ROUX J.,& MICLET L. (1976)
Programme de transcription orthographique-phonémique en langue française. ENST. Paris.

[5] GACK V.G. (1976)
L'orthographe du français. Edition Selac.

[6] LETY M. (1980)
Transcription orthographique-phonétique : un système interpréteur. Thèse de 3ème Cycle. Université Scientifique et Médicale de Grenoble.

[7] PETIT ROBERT 1 (1990)

[8] PROUTS B. (1980)
Contribution à la synthèse à partir du texte ; transcription graphème-phonème en temps réel sur micrprocesseur. Thèse de 3ème Cycle. Université Paris -Sud-Centre d'Orsay.