

REPRESENTATION D'UN LEXIQUE POUR LA R.A.P.C A L'AIDE DE CONNAISSANCES PHONOLOGIQUES

J. Gispert, H. Meloni

G.I.A. Faculté de Luminy
70, route Léon Lachamp
13288 Marseille Cedex 09 France

I. INTRODUCTION

Le système présenté fait partie d'un ensemble complet de Reconnaissance Automatique de la Parole Continue. Il fait suite à l'étude présentée par (Méloni 1985).

Son but est de permettre la reconnaissance de mots à partir d'un treillis de phonèmes produit par le décodage acoustico-phonétique.

Le système porte sur la composante phonétique du lexique ; la composante graphique nécessaire à la sortie des résultats sera ajoutée ultérieurement sous forme d'un nouveau module.

Les règles de Phonologie Générative, énoncées déclarativement, permettent la dérivation des formes phonétiques des mots à partir de leur forme sous-jacente construite par combinaison d'éléments lexicaux. Parallèlement, des règles morpho-syntaxiques permettent de déduire les informations diverses provenant du regroupement des morphèmes.

Les accès aux mots représentés par ces structures sont codés séparément. Ils sont constitués au moyen d'informations syntaxiques et phonétiques, et des données fournies par le décodage Acoustico-Phonétique.

Le codage du lexique est automatique ; toutefois, certaines ambiguïtés doivent être levées manuellement.

Un ensemble de prédicats permet d'extraire toutes les informations codées : structure phonologique, morphèmes, caractéristiques syntaxiques, forme phonétique.

II. CONNAISSANCES UTILISEES

On ne peut envisager de coder explicitement chaque mot sous sa forme phonétique pour conserver au lexique un volume raisonnable. Il convient donc de se tourner vers un système linguistique tirant partie des parentés entre mots différents, construits à l'aide d'un radical et de préfixes et suffixes éventuels.

Dans le cas des conjugaisons, (Bescherelle 1980) fournit un catalogue des particularités verbales, selon une approche descriptive. J. Pinchon (Pinchon 1981) opère quelques regroupements et aboutit à un ensemble de formes un peu plus restreint.

La Phonologie Générative (Chomsky 1973) offre un cadre analytique dans lequel les parentés entre mots trouvent des explications par le biais de variations phonologiques. Nous avons donc choisi cette approche, pour laquelle on dispose, pour le Français, des travaux de (Schane 1968), (Plénat 1981,1985) et (Dell 1973,1984). On traite également des connaissances lexicales :

- découpage des mots en morphèmes,
 - groupements de morphèmes possibles ou interdits,
- et des connaissances grammaticales :
- catégories et attributs syntactico-sémantiques.

III. CODAGE DES REGLES

Les règles utilisent les traits articulatoires suivants :
consonantique vocalique haut bas avant rond nasal tendu.

Chaque phonème dérivé ou sous-jacent est décrit par une clause associant son identificateur à son vecteur de traits, dans l'ordre ci-dessus.

Exemple :

<aa,vocalique.bas.tendu> → i ;

<Aa,vocalique.bas> → i ;

Aa représente le phonème /a/ sous-jacent lâche, et aa le phonème /a/ tendu, sous-jacent ou dérivé.

Des prédicats permettent :

- d'atteindre un trait dans la liste, par exemple *trait-cons(l,c)* donne la valeur positive ou négative du trait consonantique de l,
- de caractériser des classes de phonèmes (voyelle-orale, consonne ...),
- de modifier la liste de traits selon les besoins des règles (négation, affirmation, échange ...).

Les règles sont codées à l'aide de clauses Prolog sous une forme très proche de celle proposée par les linguistes. La tête de clause présente la séquence de phonèmes initiale et celle qui en dérive. La queue de la clause détermine les contraintes contextuelles d'application de la règle et les liens existant entre les deux formes.

Exemple :

nasalization(v.c.q.v'.q) →

voyelle-orale(v)

consonne-nasale(c)

nasaliser(v,v')

non-vocalique(q) ;

IV. INTERPRETATION DES REGLES

Les règles que nous utilisons sont (partiellement) ordonnées. Des méta-règles déclaratives décrivent cet ordre.

Chaque règle, à son tour, est appliquée de toutes les manières possibles sur la chaîne à transformer. La stratégie de synthèse force l'application d'une règle dès l'instant où son environnement est satisfait, puisque ceci correspond à la possibilité d'un phénomène phonologique.

Par contre, en analyse, il ne convient pas d'appliquer systématiquement toutes les règles possibles. En effet, certains phonèmes sont à la fois sous-jacents et dérivés : ils peuvent figurer tels quels dans la forme phonologique ou bien être obtenus par dérivation. Il existe donc diverses formes sous-jacentes conduisant à la même forme phonétique. Les connaissances phonologiques ne peuvent résoudre cette ambiguïté, qui est levée par consultation du lexique.

Le système doit être capable de reconnaître une phrase, même prononcée en violation de certains phénomènes comme par exemple les liaisons. Les règles qui décrivent ces phénomènes pourront être facultatives. On admettra ainsi des formes phonétiques fausses sur le plan théorique, mais d'emploi assez fréquent.

V. CODAGE DU LEXIQUE

V. 1- Représentation des Morphèmes

Les éléments lexicaux sont des morphèmes. Ils sont représentés par des clauses indiquant la nature du morphème (préfixe, radical, marques de genre et de nombre, ...) et des données dépendant de cette information.

- L'identificateur du morphème est construit avec les phonèmes qui le constituent,
- les suffixes portent une information indiquant la catégorie syntaxique qu'ils produisent,
- les radicaux décrivent l'ensemble des mots construits autour d'eux, grâce à un terme qui indique les préfixes et suffixes possibles. Ce terme est constitué de doublets <préfixe,suffixe>, des opérateurs *et* et *ou* et de la fonction *event* (éventuellement),
- tous les morphèmes peuvent comporter des données particulières sur l'emploi des règles.

Exemple :

ddOoll(radical, ou (<vide, ou (vide, et (Oorr, event(Oozz)))>, <Aann, et (Oorr, iirr)>)) → ;
Oorr(suffixe, nom) → ;
Oozz(suffixe, adjectif) → ;
Aann(prefixe, appris) → ;
iirr(suffixe, <verbe(3), appris>) → ;

V. 2- Codage des accès

Le radical permet d'accéder à tous les morphèmes constituant les mots qui lui correspondent. Par contre, les préfixes et suffixes n'offrent pas cette possibilité, le nombre de radicaux associés à chacun étant trop grand.

Pour la décomposition d'un mot, il est préférable d'avoir accès à chacun de ses morphèmes pour éviter de rechercher le radical n'importe où. L'analyse se fait de gauche à droite par identification de préfixes, puis du radical et enfin de suffixes. Cette recherche est bien entendu non déterministe.

Le phonème le plus à gauche dans un morphème donne un accès naturel pour une analyse de gauche à droite.

Exemple :

acces-morpheme(Oo, Oorr, Oo.rr.nil) → ;
acces-morpheme(dd, ddOoll, dd.Oo.ll.nil) → ;
acces-morpheme(Aa, Aann, Aa.nn.nil) → ;
acces-morpheme(ii, iirr, ii.rr.nil) → ;
acces-morpheme(Oo, Oozz, Oo.zz.nil) → ;

VI. TRANSFORMATION DES REGLES

Les résultats présentés par (Gispert 1986) montrent la nécessité de transformer les règles.

L'utilisation d'une règle choisie sur un contexte totalement indéterminé, produit les formes les plus générales transformées par cette règle. En appliquant à ces formes toutes les règles possibles de façon non déterministe, jusqu'à obtenir d'un côté une forme phonologique et de l'autre une forme phonétique, on définit tous les usages qu'il est possible de faire de cette règle. A chaque solution, on fait correspondre une macro-règle qui représente l'enchaînement des règles qui l'ont produite.

L'usage de ces macro-règles est possible grâce à des accès définis sur les phonèmes gauches des deux formes concernées. L'analyse d'un mot se fait maintenant ainsi :

- accès à une macro-règle par le premier phonème,
- unification de la forme phonétique donnée par la règle avec le mot à analyser,
- accès à une autre macro-règle par le premier des phonèmes restant à analyser, etc.

Avec ces méta-règles, les temps de calcul sur VAX 750 sont de l'ordre de la seconde.

Le remplacement des règles phonologiques par des macro-règles revient à déduire un catalogue des différents cas particuliers. Cependant, ce catalogue est obtenu automatiquement à partir des connaissances que les linguistes souhaitent manipuler. Il ressort donc que ce système peut convenir à la fois à la mise au point d'un jeu de règles et à son exploitation en situation de reconnaissance. Ceci justifie le détour par les règles de phonologie.

VII. CONSTRUCTION AUTOMATIQUE DU LEXIQUE

Un mot nouveau est proposé au système sous sa forme phonétique, avec ses attributs syntaxiques (catégorie, type d'emploi...). Le système en fait d'abord l'analyse phonologique, qui propose une forme sous-jacente dont il pourrait dériver.

L'analyse en morphèmes de cette forme ne peut être envisagée de toutes les manières possibles sans référence à des morphèmes connus. On imposera donc que les préfixes, suffixes et désinences soient tous répertoriés à priori dans le lexique. Seul le radical pourra donc être inconnu.

Ainsi ne peuvent subsister que certaines ambiguïtés d'analyse résultant de la confusion d'une partie du radical avec un préfixe ou un suffixe existant. Le choix de l'une ou l'autre forme est déterminé par d'autres mots de la même famille, qui possèdent le même radical. Ces ambiguïtés sont levées par l'utilisateur qui doit fournir au système un mot de même famille.

VIII. CONCLUSION

Cette étude met en évidence certains aspects de l'approche choisie par rapport au traitement automatique du problème :

- l'interprétation des connaissances proposées est délicate, certains aspects n'étant pas explicités,
- il est difficile de mélanger des règles provenant d'auteurs différents, celles-ci utilisant par exemple des jeux de traits différents,
- le système a permis de valider l'hypothèse de faisabilité sous Prolog (moyennant la compilation des règles),
- il constitue un outil de test pour de nouvelles théories phonologiques que l'on pourrait appliquer de la même manière.
- il est utilisable en reconnaissance

IX BIBLIOGRAPHIE

Bescherelle

Le Nouveau Bescherelle Hatier 1980

Chomsky N., Halle M.

The Sound Pattern of English Cambridge, Mass. MIT Press 1968

Dell F.

Les règles et les Sons, introduction à la Phonologie Générative Hermann Paris 1973

Dell F., Vergnaud J.-R.

Les développements récents en Phonologie : quelques idées centrales Forme Sonore du Langage Hermann 1984

Gispert J.

Représentation d'un Lezique pour la R.A.P.C. à l'aide de Connaissances Phonologiques

Séminaire Lexique GRECO GALF Toulouse Janvier 1986

Meloni H., Gispert J., Guizol J.

Un Système Expert pour l'Identification Analytique de Mots dans le Discours Continu. 5^{èmes} Journées Internationales Systèmes Experts Avignon 1985

Perennou G.

Base de Données Lezicale Rapport scientifique GRECO communication parlée Juin 1984 CRIN Nancy 1983

Pinchon J., Coute B.

le Système Verbal du Français Nathan 1981

Plenat M.

L'autre conjugaison, ou de la régularité des verbes irréguliers Cahiers de Grammaire n°3 Avril 81 Centre de Sociologie et de Dialectique Sociale Toulouse 1981

Plenat M.

Sur quelques aspects de la nasalisation en Français standard Cahiers de Grammaire n°9, Toulouse 1985

Schane S.A.

French Phonology and Morphology Cambridge, Mass. MIT Press 1968