La composante lexicale de la machine à dicter MAUD,

Kamel Smaïli, François Charpillet, Jean-Marie Pierrel, Jean-Paul Haton

LA COMPOSANTE LEXICALE DE LA MACHINE A DICTER MAUD

K. Smaïli, F. Charpillet, J.M. Pierrel, J.P. Haton
CRIN-CNRS & INRIA
B.P. 239, 54506 Vandœuvre lès Nancy Cedex
email : smaili, charp, jmp, jph@loria.crin.fr

RESUME

Dans cet article, nous présentons la composante lexicale de la machine à dicter MAUD. Nous montrons que cette composante est fondée sur un grand lexique qui ne doit pas être considéré comme une simple liste de mots, mais comme une base de données de laquelle, on peut extraire des informations nécessaires pour tous les niveaux de la reconnaissance. Dans la dernière partie de cet article, nous présenterons la stratégie lexicale de MAUD, et nous donnerons également quelques résultats lexicaux.

Mots clés : machine à dicter, lexique, classification contraintes lexicales.

ABSTRACT

In this paper, we discuss the lexical component of the MAUD dictation machine. This component is based on a large vocabulary. The lexicon is not considered as a simple list of words but as a data base from which we can extract information necessary at every level of the recognition process. We end this paper by giving some details concerning the lexical strategy, and by presenting some lexical results.

Key words : dictation machine, large vocabulary, classification, lexical constraints.

I INTRODUCTION

I.1 *Les applications de type machine à dicter*

La machine à écrire à entrée vocale ou machine à dicter (MAD) constituera sûrement à terme l'une des retombées les plus prometteuses du domaine de la reconnaissance de la parole. Les applications de type MAD suscitent actuellement un regain d'intérêt stimulé par les résultats des premières études de faisabilité. Ainsi la société Dragon commercialise déjà une première réalisation DRAGON DICTATE permettant la dictée de textes lus en mot à mot. Un éditeur évolué permet à l'utilisateur de guider le système de reconnaissance après l'énonciation de chaque mot si une erreur a été commise. Citons également le système TANGORA développé par IBM qui possède des caractéristiques similaires. Il existe pour ces deux systèmes différentes versions permettant une utilisation de la machine à dicter dans 6 langues européennes : anglais, allemand, espagnol, français, italien et néerlandais. Si ces premières approches constituent un réel progrès dans le domaine, il reste néanmoins un effort considérable de recherche à mener pour que ce type d'application devienne accessible au grand public. En effet, les contraintes actuelles de ces machines ne sont acceptables que pour des utilisateurs ne pouvant utiliser d'autres moyens que la parole pour la dictée (essentiellement les handicapés moteurs).
Parmi les améliorations envisageables, le mode d'élocution en parole continu et le recours à de très grands vocabulaires sont des éléments importants susceptibles d'améliorer l'ergonomie des machines à dicter. Le projet MAUD que nous développons actuellement va dans ce sens. Pour ce faire, les connaissances linguistiques mises en œuvre doivent aller bien au delà des modéles bi et tri-classes actuellement utilisés dans les MAD actuelles. Parmi ces connaissances, la composante lexicale tient une place primordiale. En effet, les interactions du lexique avec les autres composantes d'une MAD sont essentielles. C'est pourquoi le lexique doit renfermer aussi bien des informations infra-lexicales (phonétiques, phonologiques et morphologiques) que des informations supra-lexicales (syntaxiques, sémantiques ou pragmatiques).

I.2 *Notre approche dans le projet MAUD*

On retrouve dans MAUD les deux grandes composantes d'une machine à dicter, à savoir un éditeur évolué permettant la manipulation et la correction des textes saisis et un système de reconnaissance de parole continue. Le système que nous développons autorise la dictée de textes lus du français courant sans restrictions majeures. Les mots énoncés doivent cependant avoir été préalablement définis dans un lexique. Actuellement, nous disposons de deux lexiques de 7000 et 37 000 mots dont certaines informations ont été extraites de la base de données lexicale BDLEX. Le système de reconnaissance s'articule autour de trois composantes :

- APHODEX qui fournit un décodage acoustico-phonétique sous forme d'un treillis de phonèmes [Fohr 86],
- la composante lexicale que nous détaillerons dans la suite de cet article,
- la composante syntaxico-sémantique fondée sur la notion de bi et tri-classe et un certain nombre de contraintes locales [Smaïli 91b].

La composante lexicale joue un rôle central dans MAUD puisqu'elle s'articule avec la composante linguistique et la composante phonétique avec lesquelles elle interagit pour

identifier dans le continuum de parole les mots pouvant être mis en correspondance avec le signal vocal. Cette tâche d'identification ne peut s'effectuer sans une organisation efficace du lexique. Les fonctions d'accès doivent en effet faciliter la mise en place d'un ensemble de filtres dont la combinaison doit permettre d'extraire du lexique des sous-ensembles aussi restreints que possible qui pourront alors être mis efficacement en correspondance avec une portion du signal vocal. Pour celà, deux types de composantes, infra et supra-lexicale, doivent être définies pour faciliter la définition de filtres linguistiques et phonétiques [Pierrel 89]:

- la composante supra-lexicale permet l'interaction avec les niveaux syntaxico-sémantiques. Les hypothèses linguistiques constituées de classes syntaxiques, de traits grammaticaux (par exemple genre et nombre d'un nom), ou dans certains cas, de traits sémantiques permettent alors de restreindre la recherche lexicale à un sous ensemble du lexique. Par ailleurs, les informations supra-lexicales permettent une vérification de la cohérence syntaxique ou sémantique des suites de mots déjà reconnus ;
- la composante infra-lexicale permet l'interaction avec le niveau phonétique dont les résultats doivent permettre d'extraire du lexique un ensemble de mots dont la forme phonétique s'identifie le mieux avec le treillis phonétique. La composante lexicale peut exploiter un filtrage cherchant à mettre en correspondance une portion du treillis phonétique avec les patrons phonétiques, ou avec la forme phonétique complète des mots recherchés. Un filtrage sur la longueur des mots peut également être mis en œuvre lorqu'une pause ou un marqueur prosodique permettent de mesurer la durée du mot à reconnaître. Notons également que la composante lexicale peut interagir avec la composante accoustico-phonétique dans le but de vérifier certaines des hypothèses émises lors de la mise en correspondance entre un mot et le treillis.

La mise en place de procédures d'accès au lexique doit être aussi efficace que possible tant les accès sont nombreux pendant la reconnaissance. L'efficacité doit porter à la fois sur la rapidité et sur la capacité à extraire des sous-vocabulaires restreints. Cependant, l'utilisation d'un grand lexique, l'imprécision et l'incertitude du décodage acoustico-phonétique, la quasi absence de marqueurs de segmentation entre les mots en parole continue sont des facteurs qui tendent à augmenter les ambiguïtés et donc la complexité de ces filtres.

II LES CHOIX LEXICAUX DE MAUD

Pour répondre aux besoins des industries de la langue, des matériaux lexicaux doivent être fournis. Parmi ces outils, les lexiques informatisés de plus en plus complexes jouent un rôle primordial. Un lexique ne peut être défini comme une simple liste de mots mais plutôt comme une base de données dans laquelle des informations de nature diverses sont extraites grâce à des fonctions d'accès adaptées. Pour les applications de type MAD des informations de nature phonétiques, morphologiques, phonologiques, prosodiques, syntaxiques ou sémantiques sont nécessaires.

III LA COMPOSANTE INFRA-LEXICALE

Dans l'approche analytique que nous avons choisie, une description des entités lexicales en terme de sons élémentaires est nécessaire. On peut distinguer essentiellement trois niveaux infra-lexicaux [Pierrel 89] : la morphologie, la phonétique, et la phonologie.

Les informations infra-lexicales de MAUD sont extraites des dictionnaires de BDLEX, base de données lexicale du français écrit et parlé développé à l'IRIT dans le cadre du GDR-PRC Communication Homme-Machine [Pérennou 87]. L'organisation de BDLEX en tant que base de données est de type relationnel. Chaque entrée lexicale des dictionnaires de MAUD est composée de 8 champs dont 3 sont de type infra-lexicaux :

- la forme orthographique qui est une représentation graphique accentuée de l'entrée lexicale,
- la forme phonétique qui donne la représentation phonétique de l'entrée à l'exception de phonèmes terminaux sujets à variations,
- la finale phonologique qui permet de traiter les phénomènes de frontières de mots.

III.1 *La forme orthographique*

MAUD fonctionne actuellement avec les deux lexiques LEX 7000 et LEX 37000 qui contiennent respectivement 7000 et 37000 entrées. Le problème de la représentation du lexique se pose en ces termes : faut-il ou non coder en machine l'ensemble des entités lexicales ou au contraire se limiter au codage des morphèmes de base et utiliser ensuite des règles permettant d'engendrer les entités lexicales de surface. Cette deuxième possibilité paraît séduisante, mais suppose que l'on soit capable de filtrer le lexique sur les morphèmes. Ceci nous paraît difficile en raison de l'incertitude des résultats du décodage phonétique sur lesquels devrait reposer ces filtres.

III.2 *La forme phonétique*

C'est le champ le plus important du dictionnaire. En effet, la fonction permettant de lier les entités lexicales aux différentes portions du treillis phonétique utilise ce champ ainsi que celui de la finale phonologique pour mettre en correspondance le mot de référence et le treillis phonétique. La composante lexicale utilise également des données phonétiques qui ne sont pas mémorisées dans le dictionnaire, mais qui sont indispensables dans le calcul des distances inter-phonèmes. Ces données phonétiques sont réparties dans les 4 matrices suivantes :

• La matrice de confusion du système

Cette matrice de confusion modélise le comportement global du décodeur acoustico-phonétique. Les données de cette matrice permettent de prendre en compte les erreurs du décodeur acoustico-phonétique : APHODEX [Fohr 86].

• La matrice de confusion générale

Cette matrice de confusion est indépendante du système de décodage utilisé. Elle modélise les similitudes inter-phonèmes indépendamment de tout décodeur. L'intérêt d'utiliser deux matrices de confusion est de permettre de diversifier les sources de connaissances, et donc de ne pas laisser la prise de décision à un seul "décideur". La combinaison de ces deux matrices permet un renforcement mutuel des coefficients.

• La matrice d'élision

Il s'agit en fait d'un vecteur contenant les probabilités d'élision de chacun des phonèmes de la langue.

• La matrice d'insertion

Comme pour la matrice d'élision, cette matrice est un vecteur contenant les probabilités d'insertion des phonèmes de la langue.

III.3 *La forme phonologique*

Ce champ comporte des informations permettant de prendre en compte les phénomènes phonologiques apparaissant à la jonction de deux mots.

Il arrive que des mots voient certains de leurs phonèmes connaître une assimilation. Une assimilation correspond à un transfert d'une caractéristique, ou trait phonétique d'un son sur un son immédiatement voisin. C'est ainsi qu'on trouvera dans les dictionnaires de MAUD des alternatives de prononciation de ces phonèmes assimilés comme dans l'exemple suivant :

gros, gro, [s"/z"]

Dans cet exemple, il s'agit d'une assimilation du trait de voisement de la fricative qui selon le cas peut être sourde ou sonore.

Ce champ comporte également des informations concernant la présence ou non du schwa en fin de mot.

En revanche, dans nos dictionnaires, il manque les altérations dues aux accents ou prononciations particulières qui peuvent correspondre soit à des élisions, soit à des substitutions de phonèmes. Le plus souvent, ces altérations phonologiques se produisent à l'intérieur d'un mot.

IV LA COMPOSANTE SUPRA-LEXICALE

Comme nous l'avons précisé précédemment, un lexique doit véhiculer également des informations syntaxico-sémantiques qui sont utilisées par les niveaux supérieurs, soit pour l'émission d'hypothèses syntaxico-sémantiques, soit pour filtrer les solutions. Cela nécessite une répartition des entrées lexicales dans des classes syntaxico-sémantiques. Cette opération est connue chez les linguistes sous le terme de classification.

IV.1 *La classification du lexique*

La classification du lexique est l'opération qui permet de subdiviser le dictionnaire en plusieurs sous-vocabulaires. Les mots d'un même sous-vocabulaire doivent avoir des comportements syntaxiques identiques ou très proches. Chacun de ces sous-ensembles est désigné par une classe syntaxique et regroupe un certain nombre de mots sur la base de propriétés syntaxiques communes.

IV.2 *Le principe de classification du lexique de MAUD*

On dira qu'un mot x appartient à une classe C, si quelque soit le mot y de cette classe, x peut se substituer à y dans n'importe quelle phrase sans que celle-ci ne perde sa validité syntaxique. Chaque classe peut être considérée comme une classe d'équivalence. Le principe de classification du lexique de MAUD se résume comme suit : en partant des douze classes syntaxiques de BDLEX dans lesquelles sont répartis les mots du lexique, et après transformation de cette classification en une classification simple utilisant les classes grammaticales de base du français, on a procédé à une décomposition "dichotomique" des classes. En effet, chaque classe a été partitionnée en deux :

- la classe ouverte qui reçoit les mots qui peuvent être construits à partir de racines morphologiques. Cette classe reçoit également les mots pouvant poser des problèmes de classification,

- le deuxième sous-ensemble contient les mots pour lesquels on a estimé qu'ils véhiculent des informations syntaxico-sémantiques très importantes comme par exemple la classe HEX qui contient les prépositions excepté et hormis qui véhiculent l'idée de l'exception. D'autres mots sont séparés de la classe ouverte pour des raisons purement syntaxiques. C'est le cas par exemple des conjonction et / ou.

Les classes issues du deuxième sous-ensemble jouent un rôle important dans la prédiction et/ou la sélection syntaxique. Elles sont obtenues par affinement dichotomique en appliquant le principe de classification. Ce procédé nous a conduit à une classification large de 201 classes. La répartition de ces classes syntaxico-sémantiques est donnée dans le tableau de la figure 1.

Nous utilisons ici le terme syntaxico-sémantique pour qualifier nos classes, car beaucoup d'entre elles véhiculent des informations sémantiques. En effet, l'examen de certaines de ces classes montrent que les éléments qui les composent partagent un même concept. Par exemple, la classe ANL est composée des adjectifs indéfinis aucun et nul, qui suggèrent une idée de quantité zéro.

Classe majeure	Nombre de sous-classe
Adverbes	85
Adjectifs	13
Articles	3
Conjonctions	12
Noms	4
Prépositions	25
Pronoms	37
Verbes	8
Autres	14

Fig 1 : Répartition des classes syntaxico-sémantiques de MAUD dans les classes majeures

Il n'a pas été possible de reprendre une classification existante car ces dernières sont soit trop détaillées, comme celle proposée par GROSS pour les verbes [Gross 75], soit insuffisamment détaillés pour véhiculer un plus grand nombre d'informations.

IV.3 *Vers une triple classification du modèle linguistique*

Le modèle linguistique développé dans le cadre du projet MAUD est fondé sur la classification présentée ci-dessus. Cela, nous conduit à évoquer les problèmes de convergence liés à l'apprentissage du modèle [Smaïli 91b]. En effet, l'opération permettant de collecter les statistiques nécessaires à l'apprentissage a été effectuée sur un corpus comprenant environ 374 000 mots. Ce corpus a une taille relativement faible par rapport au nombre de classes que nous avons définies. Cela nous a conduit à remettre en cause cette classification, sans pour en autant abondonner les avantages. Pour ce faire, nous avons décidé d'adjoindre deux nouvelles classifications du lexique, ce qui se traduit au niveau du modèle linguistique par l'utilisation de modèles plus grossiers, mais plus convergents. Ces deux nouvelles classifications utilisent respectivement 138 et 38 classes syntaxico-sémantiques. Pour le modèle linguistique, cela entraîne l'utilisation de trois modèles

markoviens qui sont combinés de manière à favoriser au cours de l'interprétation le modèle le plus détaillé. Nous donnons dans les figures 2 et 3 les courbes d'évolutions des biclasses et triclasses en fonction du nombre de mots pour les trois classifications

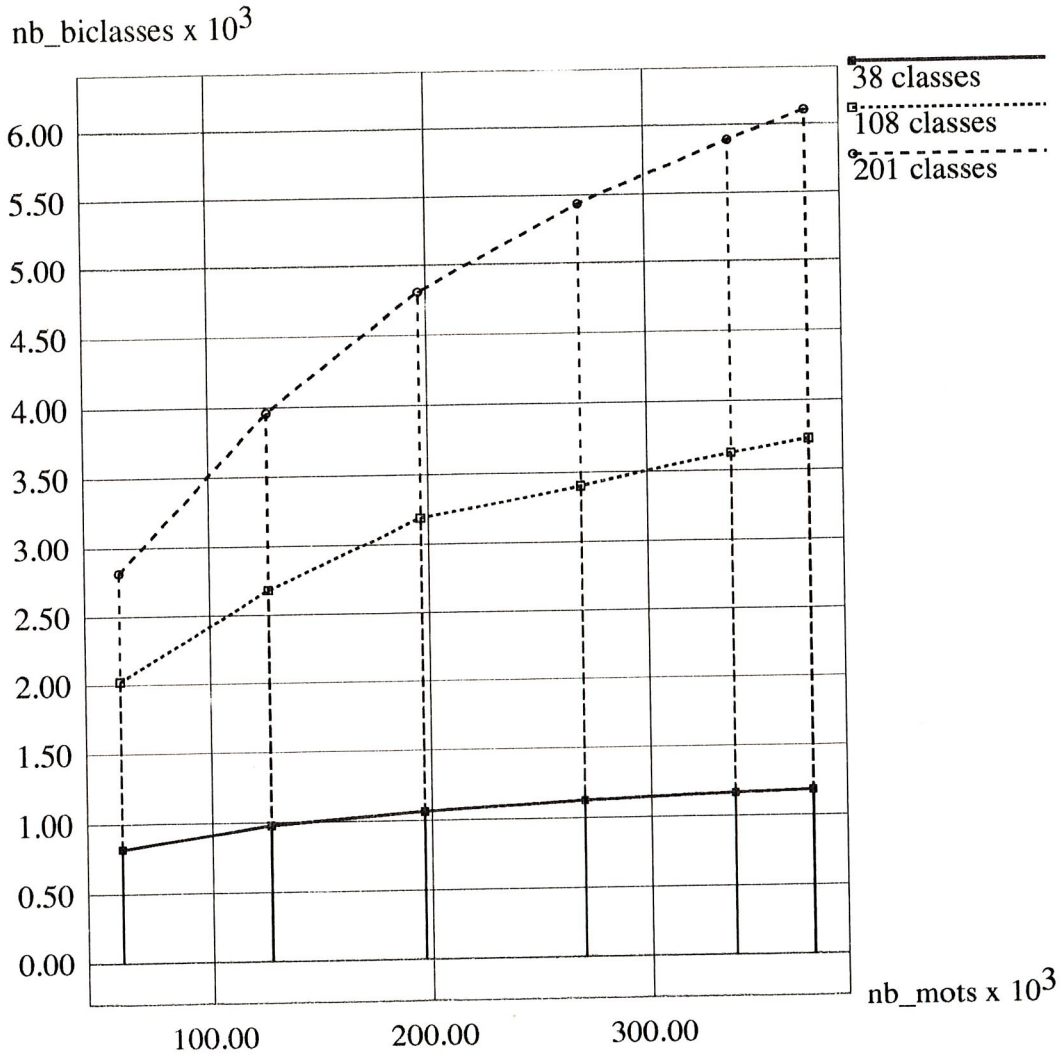

Fig 2

Variation du nombre de biclasses en fonction du nombre de mots pour les trois classifications

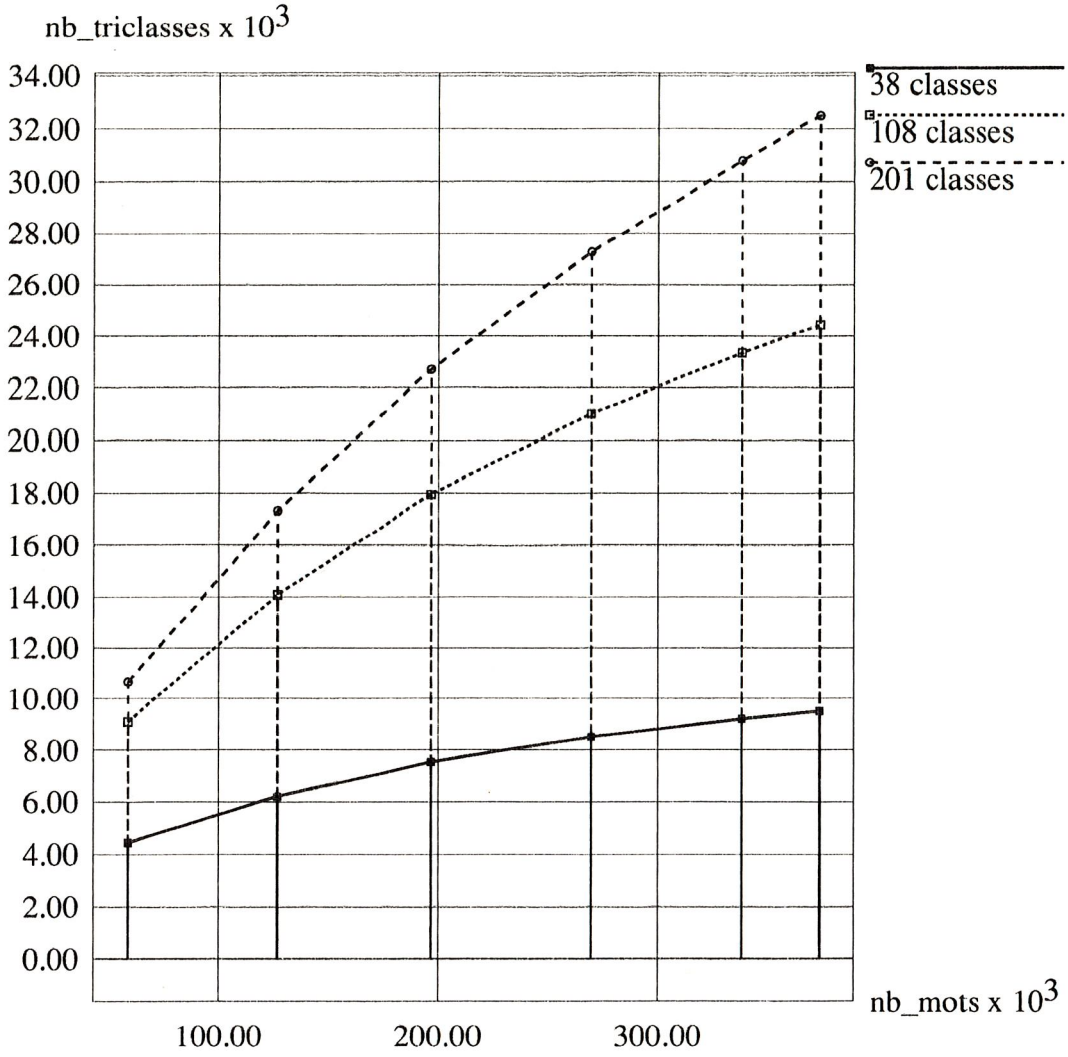

Fig 3

Variation du nombre de triclasses en fonction du nombre de mots pour les trois classifications

IV.4 *Les fréquences d'occurrence des mots*

Dans un souci d'amélioration des performances du modèle linguistique et d'enrichissement des dictionnaires, nous avons jugé nécessaire d'ajouter une nouvelle information concernant les occurrences de mots. Pour ce faire, deux types d'occurrences ont été introduites.

IV.4.1 La fréquence relative

A partir des textes qui ont servi à l'apprentissage du modèle linguistique [Smaïli 91], nous avons calculé la fréquence d'apparition de chacune des entrées du lexique. Ces fréquences sont dites relatives car elles dépendent des corpus d'apprentissage dont nous

disposons. Nous donnons dans le tableau de la figure 4, les mots les plus fréquents ainsi que leur fréquence.

IV.4.2 La fréquence absolue

Dans le but d'utiliser des fréquences plus représentatives, nous avons ajouté dans le dictionnaire un nouvel attribut permettant de donner la fréquence absolue de chaque entrée lexicale. Ces fréquences ont été extraites de la liste orthographique de base du français LB2 [Catach 84]. La figure 5 donne les mots les plus fréquents d'après cette liste.

Quelques remarques peuvent être faites à partir des figures 4 et 5. Tout d'abord l'ordre fréquentiel est à peu près équivalent dans les deux listes. Les 4 premières lignes des deux tableaux sont identiques. Les 5ème et 6ème lignes sont inversées. Les différences entre les deux tableaux peuvent s'expliquer par le fait que les corpus qui nous ont servi à calculer les fréquences relatives ne sont pas aussi représentatives que ceux utilisés pour les fréquences absolues. Avec un corpus plus important, on devrait arriver à une équivalence des deux listes modulo quelques altérations qui sont dues aux styles des auteurs des corpus.

l le la les	28 619
de	21 527
un des une	11 169
être est étais ...	7 192
à	6 815
et	5 621
avoir ai ...	4 906
en	3 628
il ils	3 153
se	2 937
ne	2 717
dans	2 677
pour	2 484
qui	2 314

Fig 4

Les mots les plus fréquents d'après leur fréquence relative

l le la les
de
un des
être est étais ...
et
à
il ils
ne
que
du des
avoir a ai...
ce ces cet cette
qui
se

Fig 5

Liste des mots les plus fréquents d'après la liste de CATACH

Ces deux fréquences sont combinées dans le modèle linguistique de manière à renforcer leurs valeurs. Le résultat de la combinaison est utilisé comme paramètre du modèle Markovien.

IV.5 *L'organisation des dictionnaires*

La reconnaissance d'un mot dans MAUD est dirigé par la syntaxe et ce choix a conditionné l'organisation des dictionnaires. En effet, pour faciliter l'accès à une classe syntaxique particulière, il est plus judicieux de regrouper les mots par classe syntaxique. Les mots d'une même classe syntaxique sont rangés par fréquences décroissantes, mais chaque entrée lexicale est suivie immédiatement par son entrée homophone dans le cas ou elle existe. Ainsi, le calcul de distance ne se fera pas deux fois pour une même forme phonétique. Enfin l'accès à une entrée lexicale se fait en deux étapes : on accède d'une manière directe à la classe syntaxique d'un mot et ensuite d'une manière séquentielle au mot en question.

V LA STRATEGIE LEXICALE DE MAUD

La composante lexicale de MAUD (COLEMAUD) doit assurer entre autres la fonction de choisir dans le dictionnaire un nombre réduit de mots candidats à la reconnaissance. Ces mots doivent être les plus proches possibles (au sens de la distance inter-mots) du mot prononcé. Pour ce faire, COLEMAUD fait appel à la composante syntaxique. Celle-ci initialise la table des hypothèses syntaxico-sémantiques, ce qui permet ensuite de désigner le sous-vocabulaire dans lequel s'effectue la recherche lexicale. Les mots de ce sous-vocabulaire vont servir à engendrer d'autres hypothèses. Ce processus cyclique s'arrête lorsque tout le treillis phonétique aura été parcouru (fig 6).

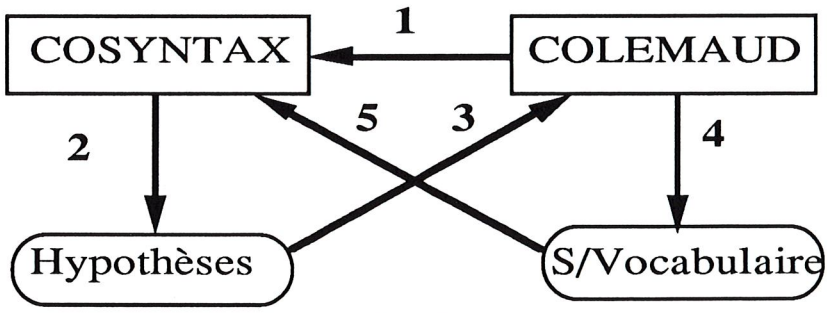

Fig 6
Schéma simplifié de la stratégie lexicale de MAUD

COLEMAUD utilise un certain nombre de procédures permettant d'une part de réduire au maximum le nombre de mots candidats à la reconnaissance, et d'autre part d'accélérer le processus de reconnaissance en minimisant le nombre d'appel aux fonctions de calculs de distance. Nous donnons ci-dessous un aperçu de ces procédures.

V.1 *La contrainte de tolérance aux erreurs*

Le décodeur acoustico-phonétique que nous avons utilisé ne commet pas n'importe quelle erreur. Par conséquent, nous avons imposé aux mots retenus de ne comporter qu'un nombre d'erreurs phonétiques raisonnable. Ce nombre est calculé par une fonction prenant en compte la longueur du mot.

V.2 *La contrainte sur la longueur des mots*

Cette contrainte est une conséquence directe de la précédente. Elle permet de comparer seulement les mots de longueur voisine. Cette contrainte joue un rôle important dans la réduction de l'espace de recherche. En effet, les tests ont montré qu' avec un lexique de 37 000 entrées, cette contrainte permet de ne retenir que 27% des mots proposés, pour les phases ultérieures de la reconnaissance.

V.3 *La mémoire cache*

Le français est l'une des langues d'Europe qui compte le plus d'homophones, ce qui est une source supplémentaire d'ambiguïté pour la reconnaissance. Pour limiter le calcul des distances (opération de calcul coûteuse), il nous a semblé intéressant de faire le calcul une seule fois pour une même liste d'homophones, et de récupérer les distances ainsi calculées lorsque cela devient nécessaire. Pour ce faire, COLEMAUD remplit au fur et à mesure une mémoire cache avec les n derniers mots pour lesquels les distances ont été calculées. Ainsi, à chaque requête de calcul de distance, la composante lexicale commence d'abord par consulter la mémoire cache, et c'est seulement si le mot n'y figure pas que le calcul s'effectue.

V.4 *Le treillis phonétique de rejet (TPR)*

Le TPR est construit par MAUD à partir du treillis phonétique délivré par APHODEX, en ne conservant dans celui-ci que les phonèmes ayant été reconnus avec un très mauvais score. Le TPR est utilisé ensuite par COLEMAUD pour pénaliser fortement les mots candidats à la reconnaissance qui seraient composés de phonèmes appartenant à ce treillis.

V.5 *La distance lexicale*

Lorsque plusieurs hypothèses de mots candidats sont possibles, on a souvent recours à un classement de ces propositions. Il est pour cela nécessaire de déterminer une mesure qui doit être un bon estimateur de la probabilité qu'une chaîne phonétique corresponde au mot effectivement prononcé. Une distance souvent utilisée dans les modules lexicaux est la distance de Levenshtein [Adda 87]. Cette distance permet de calculer le coût minimal de transformation d'une chaîne X en une chaîne Y, en pondérant le nombre de substitutions, d'élisions et d'insertions. Cette méthode présente l'inconvénient de ne pas prendre en compte les caractéristiques du décodeur acoustico-phonétique. Dans MAUD, la distance entre deux chaînes phonétiques est calculée en fonction du nombre d'élisions, d'insertions, de substitutions ainsi que des mesures inter-phonèmes, et des pénalités d'élisions et d'insertions (informations données par le décodeur acoustico-phonétique). Ensuite, ces mots sont retenus si leur score est supérieur à un seuil déterminé dynamiquement au cours de l'interprétation en fonction de la longueur du mot reconnu. Nos tests nous ont montré que seulement 1,1% des mots candidats ont été retenus pour les phases ultérieures de la reconnaissance. Ce taux correspond à une moyenne de 160 mots retenus par phrase pour un lexique de 7 000 entrées. Avec un lexique de 37 000 entrées, ce taux passe à 0,6%, ce qui correspond à 360 mots en moyenne.

VI CONCLUSION

Dans cet article, nous avons montré le rôle important joué par le lexique dans un système de reconnaissance de la parole et surtout dans un système type machine à dicter. Un problème important est celui de l'intégration de connaissances sémantiques dans les lexiques de grande taille. Quelles sont les connaissances sémantiques de base à rajouter lorsque le domaine d'application n'est pas restreint ? Ce sont des questions pour lesquelles il n'y a pas de consensus général. Pour notre part, nous avons travaillé dans un domaine non restreint ; la sémantique a été introduite d'une manière indirecte dans les classes syntaxico-sémantiques en mettant dans une même classe les mots partageant un ensemble de propriétés syntaxico-sémantiques.

La stratégie lexicale utilisée dans MAUD combine les informations infra et supra-lexicales, en utilisant une stratégie mixte. Cette approche nous paraît très intéressante lorsque le traitement exige l'utilisation d'un lexique de taille importante et que la chaîne d'entrée servant au traitement est entachée d'erreurs. Les filtres utilisés dans COLEMAUD ont permis un filtrage graduel du lexique. Ainsi, sur l'ensemble des mots du lexique proposés par le modèle syntaxique, 73% des mots ont été rejetés par la contrainte de la longueur. Et sur les 27% restant, le seuillage des distances a permis de garder seulement 0.6% pour les phases ultérieures de la reconnaissance.

Dans la nouvelle version de MAUD qui est en cours de développement, nous menons une réflexion sur l'utilisation des trois classifications que nous avons présentées dans cet article. En effet, ces trois classifications exigent une représentation particulière des lexiques de manière à garder le même schéma directeur d'accès aux classes syntaxico-sémantiques. IL faudrait également trouver une bonne fonction permettant de combiner d'une manière efficace les scores émanant des trois modèles linguistiques résultant.

REFERENCES

[Adda 87] : G.Adda " Reconnaissance de grands vocabulaires : une étude syntaxique et lexicale", Thèse de Docteur-ingénieur, Université de Paris XI, 1987.

[Catach 84] : N.Catach " Les listes orthographiques de base du français", Nathan recherche, 1984.

[Fohr 86] : D.Fohr " APHODEX : un système expert en décodage acoustico-phonétique de la parole continue", Thèse de Docteur-Ingénieur, Université de Nancy I, 1986.

[Gross 75] : M.Gross " Méthodes en syntaxe régime des constructions complétives", Herman, 1975.

[Pierrel 89] : JM.Pierrel " Lexique et compréhension automatique da la parole", Lexique n°8, p137-166, 1989.

[Pérennou 87] : G.Pérennou " BDLEX : a data and cognition base of spoken French", Research and development in language processing, JP.Haton et G.Pérennou eds, INRIA, 1987.

[Smaïli 91a] : K.Smaïli, F.Charpillet, JM.Pierrel, JP.Haton " Vers une première réalisation d'une machine à dicter destinée aux grands vocabulaires", Deuxièmes journées nationales du GRECO-PRC Communication Homme-Machine, Toulouse 1991.

[Smaïli 91b]: K.Smaïli " Conception et réalisation d'une machine à dicter à entrée vocale destinée aux grands vocabulaires : Le système MAUD", Thèse de Doctorat de l'université de Nancy I, 1991.