Elements pour la mise au point de système de reconnaissance grand vocabulaire en français

Martine Adda-Decker, Gilles Adda, Jean-Luc Gauvain, Lori Lamel

Groupe Traitement du langage parlé
LIMSI-CNRS, BP 133, 91403 Orsay cedex, FRANCE
{madda,gadda,gauvain,lamel}@limsi.fr
http://www.limsi.fr/TLP

ABSTRACT

In previous studies concerning speech recognition in French, lexical coverage has been identified as a major contributor to recognition error rates. High lexical variety also implies poorer language model (LM) training for a fixed text corpus, even though the LM component is of prime importance in French to discriminate homophones. The use of combined word and class N-gram language models is investigated. To address the coverage problem the system vocabulary has been extended from 20k to 65k words and different text normalisation and selection methods are addressed. Comparing 20k and 65k systems, a relative error reduction of 40% is observed. Word error rates analysed as a function of word frequencies highlight the importance of the LM in the recognition process. Part of this work has been carried out within the ARC *Linguistics, Computer Sciences & Spoken Corpora* supported by the AUPELF-UREF. During the official AUPELF'97 evaluation of French recognizers our 65k system obtained a 11.2% word error rate.

1. INTRODUCTION

Le français est une langue inflectionnelle; cette spécificité, qu'elle partage avec d'autres langues (mais pas avec l'anglais), conduit à une grande variété lexicale, ce qui n'est pas sans causer quelques soucis aux systèmes de reconnaissance de la parole : un grand nombre de mots hors vocabulaires (MHV) et des modèles de langage (ML) avec des perplexités élevées. Et il se superpose à cette variété lexicale un fort taux de mots homophones, qui ne peuvent être discriminés que par le modèle de langage. Une étude comparative portant sur des textes de journaux en français et en anglais a montré que si 20% des mots du texte anglais avaient une transcription phonémique (supposée correcte) ambiguë, ce nombre montait à 75% pour le texte français[Gau94]; et pour illustrer la faiblesse de la couverture lexicale en français, la même étude a montré qu'il fallait doubler la taille du lexique en français pour obtenir un couverture comparable à celle de l'anglais. On peut ainsi mieux mesurer les problèmes que rencontre la reconnaissance en français.

Dans cet article, nous présentons les améliorations que nous avons apportées au système de reconnaissance du LIMSI, afin de lui permettre de traiter efficacement la langue française, en insistant sur l'importance de la modélisation lexicale et syntaxique. En premier lieu, nous abordons les moyens de diminuer le nombre de MHV; nous montrons ensuite l'impact sur l'estimation du ML et sur les résultats de reconnaissance, de l'utilisation de textes de 40 à 255 M de mots. Nous donnerons les résultats que nous avons obtenus avec les systèmes 20k et 65k, sur des corpus de tests avec et sans contrôle du taux MHV, en analysant l'impact que ces MHV produisent sur l'efficacité globale du système. Nous présentons une étude des taux d'erreur en fonction de la fréquence des mots erronés, pour mettre en lumière de quelle manière le ML intervient dans le processus de reconnaissance. Cette étude confirme la nécessité de disposer de modèles de langage plus efficaces, les modèles de type n-gramme étant trop limités dans de nombreuses situations, pour la langue française.

2. LE PROJET "DICTÉE VOCALE" DE L'ARC B1 DE L'AUPELF

Une part importante des travaux décrits dans cet article a été produite lors du développement et de l'évaluation à laquelle nous avons participé dans l'ARC B1 de l'AUPELF-UREF[1], portant sur l'évaluation des systèmes de reconnaissance en français. Différentes catégories d'évaluation ont été définies, se différenciant par la taille du lexique (20k, 65k) et par le fait que le taux de MHV était contrôlé ou non. Une évaluation officielle pour la langue française, menée dans le projet LRE-SQALE [Lam95] avait eu lieu, quelques années auparavant, où le lexique était limité à 20k mots et où le taux MHV était contraint à moins de 2% (ce taux, sans contraintes, devant être de 5 à 6%). Pour l'évaluation AUPELF [Dol97], pour le développement et pour l'évaluation officielle, 2 ensembles de test $\mathcal{T}$ de 600 phrases ont été sélectionnés sans contraite sur le taux MHV. On a ensuite extrait de chaque ensemble $\mathcal{T}$, un sous-ensemble $\mathcal{T}'$ de paragraphes totalisant 300 phrases, les paragraphes étant choisis pour obtenir le taux de MHV minimal. Nous présentons ici les résultats obtenus lors du développement sur les 2 ensembles $\mathcal{T}$ et $\mathcal{T}'$.

3. NOUVELLES AVANCÉES POUR LA LANGUE FRANÇAISE

Lors du projet SQALE précédemment évoqué, un des résultats importants qui avaient été mis en évidence, est que le taux de MHV est très différent pour une même taille de vocabulaire d'une langue à l'autre. Etant donné l'impact de ce taux sur l'efficacité de la reconnaissance, il est donc fondamental d'améliorer la couverture lexicale pour les langues inflectionnelles comme l'allemand ou le français.

[1] AUPELF: Association des Universités Partiellement ou Entièrement de Langue Française - UREF: Université des Réseaux d'Expression Française. ARC:Action de Recherche Concertée.

3.1. *Couverture lexicale*

Nous avons étudié l'impact de 3 phénomènes sur la couverture lexicale : la taille du vocabulaire, la définition d'un mot du lexique, et la sélection des mots du lexique. Pour ce faire, nous avons utilisé différents extraits du corpus de texte du journal *Le Monde* (années 1987 à 1995) :
T_0 : années 1987-88 (40M mots)[2], T'_0 : années 1994-95 (40M mots), T_1 : années 1987-95 (185M mots), T_2 : years 1991-95 (105M words).

Taille du vocabulaire La table 1, donnant le taux de MHV pour différentes tailles de vocabulaire, montre que la couverture s'accroît toujours de manière sensible en augmentant la taille du vocabulaire. Ainsi, le taux de MHV passe de plus de 6% pour 20k mots à moins de 2% pour 65k mots, sur le corpus $\mathcal{T}$.

Table 1: Taux de MHV sur le corpus de développement $\mathcal{T}$ pour des listes de mots de taille variant de 10k à 65k mots. La liste est choisie en sélectionnant les N mots les plus fréquents du texte T_0.

lexique	*20k*	*30k*	*40k*	*50k*	*60k*	*65k*
% MHV	6,38	4,30	3,15	2,36	1,95	1,79

Définition du mot Pour un vocabulaire de taille fixe, la couverture peut grandement varier suivant la définition que l'on adopte de ce qu'est un "mot". Une étude détaillée de l'impact des différentes définitions peut être trouvée dans [Add97a]. Dans la figure 1, le taux de MHV est réduit de 50% lorsque l'on utilise une normalisation significative (N_b, N_c) du texte, en partant d'un texte (N_a) sans traitement particulier. La normalisation N_b est obtenue à partir du texte N_a en séparant les ponctuations ambiguës (- , ', .), en décapitalisant à bon escient les majuscules de début de phrase, et en normalisant les nombres et les sigles; N_c diffère de N_b en ce qu'aucune disctinction minuscule-majuscule n'a été conservée, que les signes diacritiques ont été retirés, et que ponctuations ambiguës ont été systématiquement décomposées. La forme N_b sera la normalisation utilisée dans les expériences suivantes.

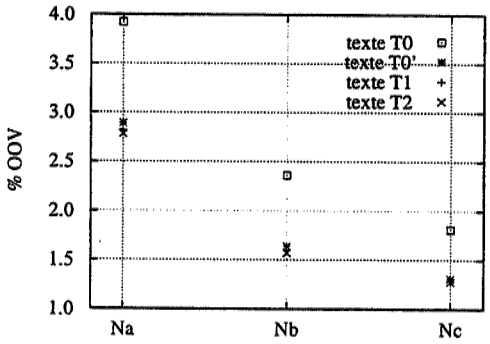

Figure 1: Taux de MHV sur le corpus de développement $\mathcal{T}$ pour différentes normalisations N_a, N_b, N_c et différents textes d'apprentissage T_0, T'_0, T_1, T_2 avec un vocabulaire de 65k mots.

[2]Ce texte constitue la ressource textuelle de base fournie à tous les partenaires de l'ARC B1

Table 2: Nombre de 2-grammes et de 3-grammes dans le ML, et perplexité pour différents textes d'apprentissage *LeM*, *LeM + MD*, *LeM + MD + AFP*.

	LeM	*LeM + MD*	*LeM + MD + AFP*
nb mots	185 M	191 M	255 M
nb bg	11.9 M	12.1 M	13.5 M
nb tg	13.9 M	14.3 M	18.1 M
ppx.	137.7	137.3	135.2

Sélection des mots de vocabulaire Une méthode classique de sélection d'un vocabulaire de taille N, consiste à prendre les N mots les plus fréquents d'un texte représentatif de la tâche de reconnaissance. La sélection du texte le plus représentatif conduit alors à une liste de mots ayant une meilleure couverture, comme l'illustre la figure 1 : les résultats montrent que la couverture obtenue à partir de textes de taille différentes 40M (T'_0), 105M (T_2) et 185M (T_1) est comparable, mais que l'utilisation de textes plus anciens (T_0) conduit à une dégradation nette de la couverture lexicale. Ce résultat conduit à la constatation que la date du texte est plus importante que la taille du texte. Seuls des gains mineurs peuvent encore être obtenus si l'on choisit un texte récent de grande taille (T_2) à un texte plus grand encore (T_1), mais contenant d'autres textes plus anciens.

3.2. *Modèles de langage*

La grande variété lexicale du français implique non seulement des problèmes de couverture lexicale, mais également une estimation plus difficile des modèles de langage. L'utilisation de textes de très grande taille est alors primordiale. Par rapport au système évalué dans SQALE où le texte d'apprentissage avait une taille de 40M de mots, nous avons accru significativement la quantité de texte; nous avons utilisé dans nos expériences 255 M de mots :
LeM: 185 M de mots du journal *Le Monde* années 87-96,
MD: 6 M de mots du journal *Le Monde Diplomatique*, années 89-96,
AFP: 64 M de mots de l'*Agence France Presse*, années 94-96.
Dans la table 2, nous avons indiqué la taille des modèles 3-gramme en fonction des textes utilisés. Tous les modèles de langage présentés ont été obtenus avec une technique dite de repli ("*Back-off*") de Good-Turing. Dans ces modèles, les 3-grammes singletons ont été exclus.

Afin de limiter plus encore l'effet de cette mauvaise estimation, nous avons interpolé le modèle 3-gramme de mots avec un modèle 2-gramme de classes [Jar96], qui, par une plus grande généralisation, gère mieux ce phénomène. La formule d'interpolation utilisée est la suivante :

$$P_{int}(w) = \lambda P_{tg}(w) + (1 - \lambda) P_{bc}(w)$$

où $P_{int}(w), P_{tg}(w)$ et $P_{bc}(w)$ sont respectivement la probabilité interpolée de la séquence de mots, la probabilité obtenue par le modèle 3-gramme de mots et la probabilité obtenue par le modèle 2-gramme de classes. Une valeur optimale de $\lambda = 0.9$ a été obtenu expérimentalement. Cette interpolation conduit à une légère baisse de la perplexité (135 à 131) mais également à une baisse relative du taux d'erreur de 1,5%, en réduisant les fautes d'accord à court terme.

4. RÉSULTATS EXPÉRIMENTAUX

Les résultats présentés ici sont ceux obtenus par le système évalué dans le cadre de l'AUPELF.

4.1. *Le système* AUPELF

Le système de reconnaissance est décrit dans les détails dans [Add97b]. Nous en donnons ici uniquement les grandes lignes.

modèles acoustiques 39 paramètres cepstraux (y compris les dérivées du 1er et du 2ème ordre) constituent les paramètres acoustiques. Ils ont été obtenus à partir du spectre Mel estimé sur une bande de fréquence de 8 kHz. Chaque modèle acoustique est un CDHMM à 3 états, représentant un phone en contexte. Des modèles séparés pour les hommes et les femmes ont été construits. Les modèles ont été estimés sur un ensemble de 66 500 phrases prononcées par 120 locuteurs (BREF[Lam91]).

modèle de langage Nous utilisons des modèles 2 et 3-gramme utilisant une liste de 65k mots, estimés sur les 255 M de mots décrits précédemment, en excluant les 3-grammes singletons et en lissant à l'aide d'une technique de repli de Good-Turing.

lexique Les lexiques d'apprentissage et de reconnaissance ont été développés au LIMSI à partir d'une phonétisation automatique [Bou96]. Chaque entrée lexicale a été transcrite phonétiquement, à l'aide d'un ensemble de 34 phones (incluant le silence). Un graphe de prononciation est associé à chaque mot, afin de prendre en compte les variantes de prononciation.

décodeur Le décodage s'effectue en 3 passes. Lors de la première passe, nous générons un graphe de mots en utilisant un ML 2-gramme (2,2 M de 2-grammes) et 3000 modèles de triphones dépendants de la position dans le mot, et 8000 états partagés. La seconde passe filtre le graphe à l'aide d'un ML 3-gramme (14 M de 3-grammes, 22 M de 2-grammes), et des modèles de triphones indépendants de la position dans le mot (9000 états partagés distribués sur 5000 modèles). La troisième passe est précédée par une adaptation non supervisée, fondée sur la méthode MLLR [Leg95], en utilisant les hypothèses produites lors de la deuxième passe. Le ML interpolé évoqué précédemment n'est utilisé que dans cette passe.

résultats AUPELF Les résultats obtenus sur les corpus $\mathcal{T}$ de développement et de test, à l'aide du système 65k, sont représentés dans la table 3. Le corpus de développement était rendu plus difficile par la présence d'un locuteur dont la langue maternelle n'était pas le français. Le système du LIMSI a obtenu le meilleur résultat officiel dans cette catégorie, avec un taux d'erreur de 11,2%.

Table 3: Résultats officiels obtenus sur les corpus $\mathcal{T}$, en utilisant un lexique 65k et un ML interpolé.

test	*dev.* $\mathcal{T}$	*eval.* $\mathcal{T}$
65k	12,7	11,2

4.2. *Résultats 20k et 65k*

Une comparaison des résultats obtenus à l'aide des systèmes 20k et 65k est résumée dans la table 4.
2 systèmes 20k ont été testés: le $\widehat{20k}$ utilise un ML estimé sur 255 M de mots, alors que le ML du 20k est estimé sur le corpus T_0 (40 M de mots). Le gain observé en augmentant ainsi la taille du texte est faible : en réduction relative du taux d'erreur, le gain est de 9% sur $\mathcal{T}$ et 16% sur $\mathcal{T}'$. Au contraire, on peut observer que le gain obtenu en passant du système 20k au système 65k est important : 40% de réduction relative du taux d'erreur sur $\mathcal{T}$ et $\mathcal{T}'$. Ce gain s'explique par une synergie des gains aux niveaux lexical, syntaxique et acoustique. Il illustre ainsi l'importance que l'on doit apporter à disposer d'une couverture lexicale suffisante, pour espérer améliorer les résultats de reconnaissance par l'amélioration du ML.

Table 4: Résultats de reconnaissance obtenus par les systèmes 20k et 65k sur les corpus $\mathcal{T}$ et $\mathcal{T}'$

$\mathcal{T}$	*%MHV*	*%err.*
20k	6,38%	23,85%
$\widehat{20k}$	6,4%	21,8%
65k	1,3%	12,9%

$\mathcal{T}'$	*%MHV*	*%err.*
20k	3,6%	17,3%
$\widehat{20k}$	3,6%	14,6%
65k	0,5%	8,9%

4.3. *Contrôle du taux MHV*

Nous pouvons mesurer l'impact sur le taux d'erreur du contrôle artificiel du taux de MHV, en nous limitant au système 65k. Nous pouvons observer dans la table 4 une réduction du taux d'erreur de plus de 4 fois la réduction du taux MHV (-4% de taux d'erreur alors que la réduction du taux MHV est de 0,9%) lorsque l'on passe du corpus $\mathcal{T}$ au corpus $\mathcal{T}'$. Si l'on complète cette observation par la mesure de perplexité sur ces 2 corpus (135 sur $\mathcal{T}$ et 106 sur $\mathcal{T}'$, soit une réduction de 20%), nous pouvons en déduire que le contrôle du taux MHV conduit à un filtrage des phrases difficiles. Le corpus $\mathcal{T}'$ est donc non seulement mieux couvert par le vocabulaire 65k mais également mieux modélisé par le ML.

4.4. *Analyse des erreurs*

Nous pouvons attribuer une grande part des erreurs à la faiblesse du modèle de langage. Cette assertion repose sur notre analyse manuelle de la typologie des erreurs observées, mais est également étayée par l'analyse des erreurs en fonction de la fréquence des mots concernés.

L'analyse manuelle des erreurs de reconnaissance nous apprend qu'une grande proportion de celles-ci provient d'erreurs d'accord à court terme en genre, nombre ou personne; environ 40% des erreurs de confusion se sont portées sur les problèmes d'homophonie, erreurs pour lesquelles seul le ML est responsable. 15% des erreurs de substitution sont causés par la présence de noms propres qui sont difficiles à modéliser, à la fois au niveau syntaxique, lexicale et acoustique : les noms propres sont souvent des mots rares, dont la fréquence évolue rapidement dans le temps, et qui (en particulier pour les noms propres étrangers) conduisent à de nombreuses prononciations plus ou moins prévisibles.

Pour mieux comprendre la manière dont le taux d'erreur peut être lié aux capacités du modèle de langage, il nous a semblé intéressant de remplacer le moyennage du taux d'erreur par phrase, par un moyennage par classe de fréquence de mot. A cette fin, le vocabulaire a été partitionné en I classes de fréquence $]K_{i-1}, K_i]$, où les rangs $K_i, i = 1, \ldots, I$ sont distribués logarithmiquement le long de l'axe des rangs de fréquence de mots. Chaque mot w_n du corpus de test, dont la fréquence observée sur le

corpus d'apprentissage conduit au rang de fréquence k_n, est associé avec la région K_i, telle que $k_n \in]K_{i-1}, K_i]$. L'ensemble des mots hors vocabulaires présents dans le corpus de test (0,45%) ont été associés à une classe de rang $> 65k$ (fréquence 0). Le résultat de cette analyse du taux d'erreur par régions de fréquence est visualisé dans la figure 2.

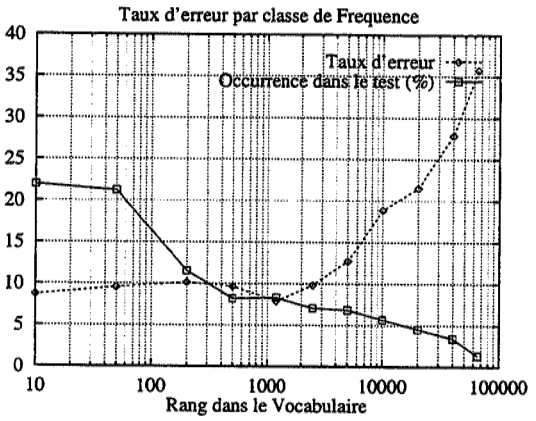

Figure 2: Taux d'erreur de mots et fréquence d'occurrence des mots du test, en fonction des régions de fréquence définies sur le vocabulaire 65k. Chaque point correspond à la limite supérieure de la région. 12 régions ont été définies et distribuées logarithmiquement entre 0 et 65000; la classe de rang $> 65k$, correspondant à 100% d'erreur n'a pas été représentée.

Nous pouvons observer sur la figure 2 que le taux d'erreur pour les rangs $k_n > 5000$ augmente de manière très importante, mais que cela ne concerne que 15% du corpus, c'est-à-dire la partie du texte non couverte par les 5000 mots les plus fréquents (les 7 premières régions). La première région contient les 10 mots les plus fréquents *de, la, l', le, à, et, les, des, d', un*, qui sont des mots très courts, voire monophones homophones, qui sont donc acoustiquement difficiles à discriminer. Les meilleurs résultats sont obtenus pour les mots de la 5^e région (rang entre 500 et 1200) : ces mots sont bien représentés dans le texte d'apprentissage, en général polysyllabiques, donc favorisés à la fois par la syntaxe et l'acoustique.

Seules des erreurs marginales peuvent être imputées à une phonétisation, comme par exemple des liaisons ou des e muets manquants dans le dictionnaire de prononciation. Cependant, on ne peut pas sous-estimer les progrès potentiels que nous pouvons réaliser grâce à une meilleure modélisation acoustique, ne serait-ce que pour expérimenter différentes pondérations entre les modèles acoustiques et de langage dans le décodeur.

5. Conclusion

Nous résumons ici les principales observations que nous avons pu accumuler lors des expériences décrites.
Des gains modestes en couverture lexicale peuvent être acquis en optimisant la normalisation et la sélection des textes. Pour cette dernière la date du texte semble être plus importante que sa taille. L'acroissement de la taille du texte d'apprentissage de 40 à 255 M de mots a permis une réduction significative de la perplexité et du taux d'erreur. L'interpolation du modèle 3-gramme de mots avec un 2-gramme de classes a conduit à un faible gain. L'utilisation d'un vocabulaire de 65k en lieu et place d'un vocabulaire 20k permet une réduction relative du taux d'erreur de 40%, ce qui peut être expliqué par la synergie d'améliorations simultanées aux niveaux lexical, syntaxique et marginalement acoustique. Nous avons montré expérimentalement que le contrôle artificiel du taux de MHV conduit à une contrainte sur la perplexité, ces 2 contrôles induisant une forte amélioration des résultats de reconnaissance. La plupart des erreurs de reconnaissance sont dues à des homophones, parmi lesquels réside une grande proportion de variantes en genre et en nombre. Enfin, nous avons mis en évidence que le taux d'erreur sur les mots les moins fréquents augmentait de manière très importante.
Une conclusion générique à l'ensemble de ces observations peut être que l'amélioration des techniques de modélisation linguistique est la voie de nous nous devons d'explorer dans les années à venir, en reconnaisance de la parole pour la langue française.

6. Remerciements

Une partie de ce travail a été effectuée dans le cadre des Actions de Recherche Concertée "Linguistique, Informatique et Corpus Oraux", financés par l'AUPELF-UREF.

Bibliographie

[Add97a] Adda G. et al., *Text Normalization and Speech Recognition in French* EuroSpeech'97, Rhodos, septembre 1997.

[Add97b] Adda G. et al., *Le système de dictée du LIMSI pour l'évaluation* AUPELF'97 1ères JST FRANCIL, Avignon, avril 1997.

[Bou96] Boula de Mareüil P. *Pour une approche par règles en transcription graphème-phonème* Séminaire GDR-PRC CHM Lexique et Communication Parlée, Toulouse'96 France.

[Dol97] Dolmazon J.M. et al., *ARC B1 - Organisation de la 1e campagne* AUPELF *pour l'évaluation des systèmes de dictée vocale* 1ères JST FRANCIL, Avignon, avril 1997.

[Gau94] Gauvain J.-L. et al., *Speaker-independent continuous speech dictation* Speech Communication **15**, pp. 21-37, septembre 1994.

[Jar96] M. Jardino M. *Multilingual stochastic n-gram class language models* IEEE ICASSP-96, Atlanta, 1996.

[Lam95] Lamel L. et al., *Issues in Large Vocabulary, Multilingual Speech Recognition* Eurospeech'95, Madrid, septembre 1995.

[Lam91] Lamel L. et al., *BREF, a Large Vocabulary Spoken Corpus for French* EuroSpeech'91, Genoa, septembre 1991.

[Leg95] C.J. Legetter C.J., Woodland P.C. *Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models* Computer Speech & Language, **9**, pp. 171-185, 1995.

[You97] Young S.J. et al. *Multilingual large vocabulary speech recognition: the European* SQALE *project* Computer Speech & Language, **11**(1), pp. 73-89, janvier 1997.