MATHÉMATIQUES ET SCIENCES HUMAINES

CHRISTEL BEAUJARD
MICHÈLE JARDINO

Classifications de mots non étiquetés par des méthodes statistiques

Mathématiques et sciences humaines, tome 147 (1999), p. 7-23
<http://www.numdam.org/item?id=MSH_1999__147__7_0>

Math. Inf. Sci. hum., (37e année, n° 147, *1999*, pp. 7-23)

CLASSIFICATIONS DE MOTS NON ÉTIQUETÉS PAR DES MÉTHODES STATISTIQUES

Christel BEAUJARD, Michèle JARDINO[1]

RÉSUMÉ – *Notre thématique de recherche est le développement de modèles de langage robustes pour la reconnaissance de la parole. Ces modèles doivent prédire un mot connaissant les mots qui le précèdent. Malgré le nombre croissant de données textuelles électroniques, toutes les possibilités de la langue ne sont pas présentes dans ces données, un moyen de les obtenir est de généraliser la représentation textuelle en regroupant les mots dans des classes. Les modèles de langage fondés sur des classes présentent alors une plus large couverture de la langue avec un nombre réduit de paramètres permettant une reconnaissance plus rapide des mots par les systèmes de reconnaissance de la parole dans lesquels ils sont introduits. Nous décrivons deux types de classification automatique de mots, appris statistiquement sur des textes écrits de journaux et de transcriptions de parole. Ces classifications ne nécessitent pas d'étiquetage des mots, elles sont réalisées suivant les contextes locaux dans lesquels les mots sont observés. L'une est basée sur la distance de Kullback-Leibler et répartit tous les mots dans un nombre de classes fixé à l'avance. La seconde regroupe les mots considérés comme similaires dans un nombre de classes non prédéfini. Cette étude a été réalisée sur les données d'apprentissage en français de domaines, de taille et de vocabulaire différents.*

MOTS-CLÉS – Classification, partitionnement, distance, statistiques, optimisation, reconnaissance de la parole, modèles de langages probabilistes.

SUMMARY – Clustering unlabelled words with statistical methods

Our goal is to develop robust language models for speech recognition. These models have to predict a word knowing its history. Although the increasing size of electronic text data, all the possible word sequences of a language cannot be observed. A way to generate these non encountered word sequences is to map words in classes. The class-based language models have a better coverage of the language with a reduced number of parameters, a situation which is favourable to speed up the speech recognition systems. Two types of automatic word classification are described. They are trained on word statistics estimated on texts derived from newspapers and transcribed speech. These classifications do not require any tagging, words are classified according to the local context in which they occur. The first one is a mapping of the vocabulary words in a fixed number of classes according to a Kullback-Leibler measure. In the second one, similar words are clustered in classes whose number is not fixed in advance. This work has been performed with French training data coming from two domains, both different in size and vocabulary.

KEYWORDS – Classifications, mapping, distance, statistics, optimization, speech recognition, language modeling.

---

[1] LIMSI-CNRS, Groupe Traitement du Langage Parlé, Université Paris-Sud, bät. 508, BP 133, 91403 Orsay Cedex, e-mail : beaujard@limsi.fr et jardino@limsi.fr

1. INTRODUCTION

Nous décrivons deux classifications de mots non étiquetés dont l'apprentissage est réalisé à partir de statistiques obtenues sur des textes écrits. Notre domaine de recherche est la reconnaissance de la parole continue. Celle-ci nécessite la connaissance a priori de modèles de langage prédisant un mot connaissant les mots précédents.

Si on considére la chaîne de mots $m_1 m_2 m_3 ... m_i$, le modèle de langage le plus général fournit la probabilité conditionnelle d'obtenir le mot m_i sachant son histoire $h(m_i) = m_1 m_2 m_3 ... m_{i-1}$, soit :

$$p(m_i|h(m_i)) = p(m_i|m_1 m_2 m_3 ... m_{i-1}). \quad (1)$$

L'ensemble de ces probabilités est impossible à estimer, aussi réduit-on l'histoire de chaque mot en considérant qu'il y a suffisamment d'information dans le ou les deux mots précédents pour donner une estimation fiable de $p(m_i|h(m_i))$ [15]. Si on ne considère que le mot précédent en réduisant la chaine $m_1 m_2 m_3 ... m_{i-1} m_i$ à $m_{i-1} m_i$, on estime la probabilité d'obtenir le mot m_i par la probabilité conditionnelle, $p(m_i|m_{i-1})$.

Pour calculer cette probabilité, on projette l'ensemble des mots m_i d'un texte d'apprentissage sur un vocabulaire $V = \{v_k\}$ de taille L_V, en supposant que les mots du texte qui sont hors vocabulaire, sont représentés par une seule forme v_0. Si v_j et v_k sont les projections respectives de m_{i-1} et m_i sur V, la probabilité $p(m_i|m_{i-1})$ est alors donnée à partir des fréquences d'observation de v_j et v_k et des séquences $v_j v_k$. Le modèle de langage est entièrement défini par L_V^2 paramètres[2], qui correspondent aux séquences de deux mots consécutifs, que nous appelons "bigrammes de mots". Si on prend l'exemple de la dictée vocale avec un vocabulaire limité de 20 000 mots, il faudrait observer 400 millions de bigrammes de mots différents pour couvrir le modèle.

Plusieurs méthodes sont utilisées pour estimer les bigrammes de mots non observés, l'une d'entre elles est de regrouper les mots du vocabulaire dans des classes et d'estimer la probabilité $p(m_i|m_{i-1})$ à partir non plus des probabilités de succession des mots $v_j v_k$, mais à partir des probabilités de succession des classes qui les contiennent. Si on suppose que chaque mot du vocabulaire ne peut appartenir qu'à une seule classe et que $C(v_j)$ et $C(v_k)$ sont les classes des mots v_j et v_k, on a :

$$p(m_i|m_{i-1}) = p(v_k|v_j) = p(v_k|C(v_k)) * p(C(v_k)|C(v_j)). \quad (2)$$

A chaque séquence de deux mots consécutifs est associée une séquence de deux classes, appelée "bigramme de classes". Si L_C est le nombre de classes, le nombre de paramètres à estimer pour un modèle à base de classes, est alors $L_C * (L_C - 1) + L_V - L_C$ (voir note 2). Il correspond principalement au nombre de bigrammes de classes possibles. Par exemple en utilisant 1000 classes pour un vocabulaire de 20 000 mots, il y environ un million de paramètres à déterminer pour couvrir le modèle, à comparer aux 400 millions de paramètres nécessaires pour couvrir un modèle fondé sur des bigrammes de mots.

Ainsi, pour une même quantité de données, on obtient une meilleure couverture du modèle, en générant des transitions entre mots non observées dans le texte

---

[2]En fait, il n'y a que $L_V(L_V - 1)$ paramètres à estimer, L_V paramètres étant déterminés à partir des relations de normalisation des probabilités conditionnelles.

d'apprentissage. Cette meilleure couverture ou robustesse du modèle est obtenue au détriment de sa précision, puisque la probabilité de transition entre mots est moyennée par la probabilité de transition entre classes (équation 2), qui est la même pour tous les mots contenus dans les classes correspondantes.

Le problème est alors de savoir quelle classification utiliser. La classification morpho-syntaxique classique, attribuant à chaque mot ou à chaque séquence de mots une liste d'étiquettes incluant genre, nombre,... est loin d'être unique et il existe une grande variabilité de représentations [1]. C'est pourquoi nous avons développé des méthodes pour regrouper des mots à partir des informations statistiques de positions et de fréquences relatives, observées sur des données textuelles de taille suffisante, ceci sans aucune attribution d'étiquette aux mots.

Nous décrivons deux sortes de classification que nous appliquons à deux ensembles de données textuelles. Ces deux classifications réalisent des niveaux de généralisation différents, l'une est une partition du vocabulaire en L_C classes, l'autre regroupe les mots "similaires". Ces deux méthodes ont été développées pour obtenir des modèles de langage réalisant un bon compromis entre robustesse et précision compte tenu des données disponibles pour l'apprentissage.

2. TEXTES D'APPRENTISSAGE ET D'EVALUATION

Deux types de textes écrits ont été utilisés : des textes provenant de journaux et des transcriptions de parole spontanée. Pour chaque type de texte, nous avons utilisé deux ensembles disjoints, l'un pour la classification et l'autre pour son évaluation.

Ces deux types de textes correspondent à deux situations typiques dans le domaine de la reconnaissance de la parole, l'une est associée à la dictée vocale et doit couvrir le plus largement possible la langue utilisée, l'autre correspond à des dialogues homme-machine pour lesquels le domaine langagier est plus restreint et souvent de style plus relâché.

2.1. TEXTES DE JOURNAUX

Les textes de journaux sont constitués d'articles issus du Monde (1987-1997) et de dépêches de l'AFP. Ce corpus contient environ 300 millions de mots pour un vocabulaire $V = \{v_k\}$ correspondant aux 750 000 mots différents observés dans ce corpus. Le tableau 1 donne, selon une fréquence seuil F (colonne 1) d'observation des mots v_k dans le texte d'apprentissage TA :

- colonne 2 : $|E_F|$, le cardinal de l'ensemble E_F où E_F est défini par : $E_F = \{v_k \in V / N(v_k) = F\}$, où $N(v_k)$ est la fréquence d'observation de v_k.
- colonne 3 : $|E_{tot,F}|$, le cardinal de l'ensemble $E_{tot,F}$ où $E_{tot,F}$ est défini par : $E_{tot,F} = \{v_k \in V / N(v_k) \leq F\}$.
- colonne 4 : $|V_F|$, le cardinal de l'ensemble V_F où V_F est défini par : $V_F = \{v_k \in V / N(v_k) \geq F\}$.

On peut noter que :

$$|E_{tot,F}| = \sum_{f=1}^{F} f \times |E_f|, \text{ et } |V_F| = \sum_{f=F}^{F_{max}} |E_f| \text{ avec } F_{max} = 14\ 757\ 645$$

Tableau 1: Statistiques du texte d'apprentissage en fonction de la fréquence d'observation des mots différents

F	$\lvert E_F \rvert$	$\lvert E_{tot,F} \rvert$	$\lvert V_F \rvert = L_{V_F}$
1	329 948	329 948	768 243
2	99 107	528 162	438 295
3	51 218	681 816	339 188
4	33 103	814 228	287 970
5	23 032	929 388	254 867
—	—	—	—
649	22	14 975 395	20 001
—	—	—	—
6 172 292	1	253 544 974	5
6 265 106	1	259 810 080	4
8 163 622	1	267 973 702	3
11 449 354	1	279 423 056	2
14 757 645	1	294 180 701	1

Comme on peut le voir dans ce tableau, environ la moitié des mots de V n'apparaissent qu'une seule fois dans le texte. On peut également remarquer que si l'on réduit le vocabulaire aux 20 000 mots les plus fréquents du corpus, seulement 5% des mots du texte ne sont pas identifiés.

La figure 1 montre l'évolution de la couverture du texte d'apprentissage, soit :

$$\frac{|E_{tot,F_{max}}| - |E_{tot,F}|}{|E_{tot,F_{max}}|}, \tag{3}$$

en fonction de la taille du vocabulaire, $|V_F|$, lorsque celui est constitué des mots les plus fréquents.

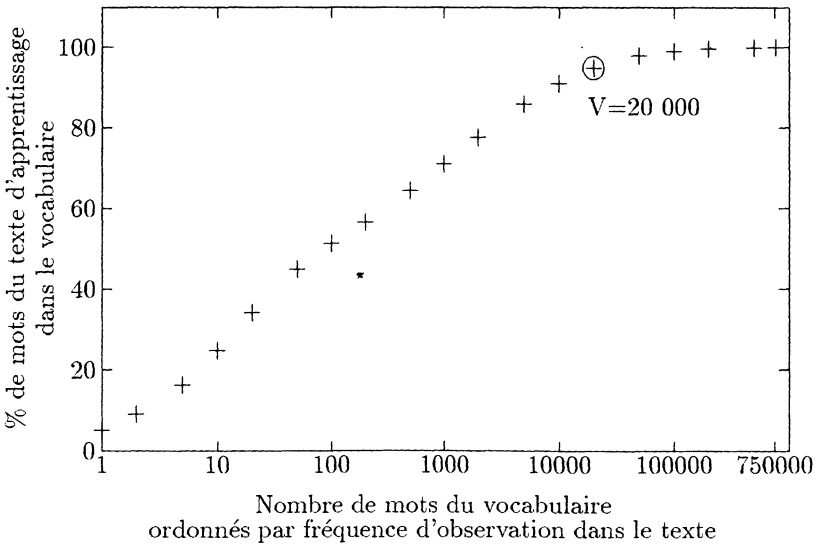

Figure 1: Couverture du texte d'apprentissage en fonction de la taille du vocabulaire

Le choix d'un vocabulaire de 20 000 mots est un compromis permettant de gérer moins de données avec une perte raisonnable d'informations. Il faut noter que les systèmes de reconnaissance de la parole les plus performants actuellement utilisent au maximum 65 534 mots ($2^{16} - 1$) de vocabulaire et ne fonctionnent pas en temps réel [11]. La réduction du vocabulaire permet de se rapprocher des conditions du temps réel. Aux 15 millions de mots inconnus résultant du choix d'un vocabulaire de 20 000 mots est associée une seule forme dans le vocabulaire. Le nombre de bigrammes de mots observés est de 19 millions pour un vocabulaire d'environ 750 000 mots, il se réduit à environ 8 millions pour un vocabulaire de 20 000 mots.

Sur les courbes de la figure 2a sont représentées les distributions des probabilités conditionnelles d'obtenir "mot" connaissant les mots *le* et *Jacques*. Ces probabilités sont calculées à partir des fréquences des bigrammes de mots correspondants observés dans les textes de journaux et rangés en ordre décroissant selon le rang de prédiction de "mot". Par exemple, à l'abscisse 1 on trouve *président* comme mot le plus probable après *le* et *Chirac* après *Jacques*, ce qui n'est pas étonnant pour le corpus utilisé.

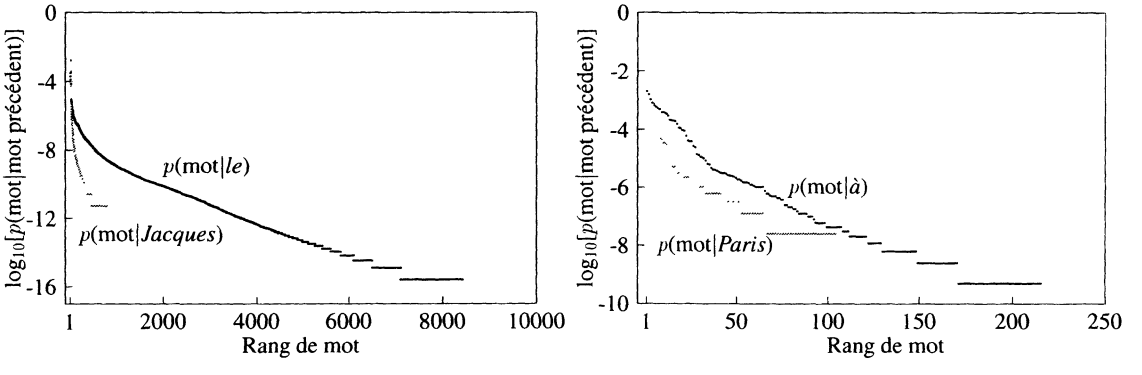

Figure 2a : Distributions relatives à *le* et *Jacques*

Figure 2b : Distributions relatives à *à* et *Paris*

Figure 2: Distributions ordonnées des probabilités conditionnelles relatives aux mots *le* et *Jacques* dans le journal Le Monde (1987-1997) (Figure 2a) et aux mots *à* et *Paris* dans les transcriptions de renseignements ferroviaires (Figure 2b)

Ces courbes montrent la diversité du comportement des mots dans les textes. On obtient deux distributions très différentes, l'une répartie sur presque la moitié du vocabulaire pour le mot *le* et l'autre sur un espace 10 fois plus petit pour le mot *Jacques*. *Jacques* est plus prédictif que *le*. Une distribution pointue permet de prédire plus facilement un mot. Dans le cas limite d'une distribution de Dirac un seul mot serait prédit.

Le texte d'évaluation provient du Monde Diplomatique (1990-1996) et contient environ 400 000 mots.

2.2. TRANSCRIPTIONS DE PAROLE SPONTANÉE

Les transcriptions de parole spontanée ont été réalisées à partir d'enregistrements de dialogues Homme/Machine [17]. Ces dialogues concernent les renseignements sur les prestations et services ferroviaires en France. Le texte d'apprentissage, ainsi constitué, contient environ 420 000 mots. Le vocabulaire, d'une taille de 1131 mots, est constitué des mots du texte d'apprentissage, orthographiés correctement et observés plus d'une fois sauf s'il s'agit d'un nom de ville. Etant donné le faible nombre de formes différentes, le

repérage des mots incorrects et celui des villes a été réalisé manuellement. Les 381 mots inconnus du texte d'apprentissage non répertoriés dans le vocabulaire, sont regroupés sous une même forme. Ce sont soit des segments de mots interrompus (vou-, résu-, par-, ...), soit des mots n'apparaissant qu'une seule fois (tabac, rouge, avion, raconter, empêche, ...). Nous avons choisi de ne pas tenir compte des évènements rares à l'apprentissage, peu significatifs statistiquement, à l'exception des noms de ville qui sont indispensables à la tâche de renseignements. Le pourcentage en fréquence des mots inconnus dans le texte d'apprentissage est alors inférieur à 0,5%. Le nombre de bigrammes de mots observés est d'environ 14 000.

Sur les courbes de la figure 2b sont représentés les distributions des probabilités conditionnelles d'obtenir un mot connaissant les mots *à* et *Paris*, observés dans le texte d'apprentissage provenant des transcriptions de parole.

Le texte d'évaluation, soit 12 000 mots, provient de transcriptions de données non utilisées à l'apprentissage.

3. PARTITION DU VOCABULAIRE, C_P

Le nombre de partitions de L_V mots en L_C classes est le nombre de Stirling du deuxième ordre [6]. Pour nos applications où les vocabulaires contiennent plusieurs milliers de mots , l'espace de recherche, devient très rapidement gigantesque quand le nombre de classes augmente. Une recherche exhaustive de la partition optimale est alors impossible en un temps polynomial, aussi utilise-t-on une exploration au hasard (Monte Carlo) pour rechercher une solution qui minimise notre critère de classification en un temps réalisable.

3.1. CRITÈRE DE SÉLECTION, DISTANCE DE KULLBACK-LEIBLER, D_P

Nous recherchons la classification qui assure une distance minimale entre la distribution de bigrammes des mots classés, et celle des mots non classés.

On suppose que la distribution, p, de mots non classés est complètement décrite par l'ensemble des probabilités conditionnelles, $p(v_k|v_j)$, d'obtenir un mot v_k connaissant le mot précédent v_j. Ces probabilités sont calculées à partir des fréquences observées dans le texte d'apprentissage, TA, des bigrammes $v_j\ v_k$ et des mots v_j et v_k. Quand on regroupe les mots dans des classes, ces probabilités sont moyennées et calculées à partir des fréquences des séquences de classes $C(v_j)\ C(v_k)$. A chaque classification sont générés de nouveaux bigrammes de mots résultant des interactions classe-classe, d'où une nouvelle distribution, $q(v_k|v_j)$, des probabilités conditionnelles.

Pour obtenir une classification optimale, on utilise la distance de Kullback-Leibler appliquée à des probabilités conditionnelles [7], entre la distribution p de référence et les distributions q explorées, $D_P(p||q)$:

$$D_P(p||q) = \sum_{v_j} p(v_j) \sum_{v_k} p(v_k|v_j) * log \frac{p(v_k|v_j)}{q(v_k|v_j)} \tag{4}$$

Les valeurs des probabilités sont estimées par les fréquences (notation $N(.)$) des évènements correspondants dans le texte d'apprentissage TA, soit :

$$p(v_k|v_j) = \frac{N(v_j\ v_k)}{N(v_j)} \tag{5}$$

$$q(v_k|v_j)] = \frac{N(v_k)}{N(C(v_k))} * \frac{N(C(v_j)\,C(v_k))}{N(C(v_j))} \tag{6}$$

$$p(v_j) = \frac{N(v_j)}{N} \tag{7}$$

où N est le nombre total de mots du texte TA.

3.2. ALGORITHME DE C_P

L'algorithme cherche aléatoirement, la répartition qui minimise la distance D_P entre la distribution des mots non classés et celle des mots classés [13, 14]. Le nombre de classes est défini *a priori*. Initialement tous les mots sont regroupés dans une même classe, ce qui correspond à une distribution initiale q qui ne dépend que des fréquences des mots. La recherche aléatoire est réalisée de la manière suivante : à chaque pas on choisit un mot au hasard et on lui affecte une nouvelle classe également au hasard. Si la distance D_P diminue on garde la nouvelle affectation, sinon on conserve l'ancienne. On arrête le processus quand il n'y a plus de changement ou quand les variations de distance deviennent infimes.

A chaque pas, on balaye l'ensemble des bigrammes de mots associés au mot choisi aléatoirement. Si NB_V est le nombre total de bigrammes de mots observés dans le texte d'apprentissage, il y a alors $2NB_V / L_V$ bigrammes à explorer car le mot peut être soit le premier terme soit le deuxième terme du bigramme. Ensuite on calcule les variations des fréquences des bigrammes de classes associés, soit une exploration de $4L_C$ bigrammes de classes. Expérimentalement, on atteint un quasi-équilibre après avoir effectué une centaine d'essais de changement de classe sur chaque mot. Si *N(essais)* est ce nombre, la classification finale est donc obtenue en *N(essais)*$(2NB_V + 4L_C L_V)$ opérations. A titre indicatif, une partition en 1000 classes du corpus de 300 millions de mots est réalisée en 5 heures sur une station de travail SGI R10 000.

Notre méthode se rattache à la méthode des nuées dynamiques [6], la recherche aléatoire permettant un accès rapide à une solution quasiment optimale. Nous avons vérifié que cette solution se rapprochait de l'optimum global, en la comparant à une solution obtenue par recuit simulé [13]. D'autres heuristiques ont été proposées pour réaliser une classification automatique de mots non étiquetés avec le même critère de sélection. Une des méthodes [5] correspond à une classification hiérarchique ascendante. Les mots sont initialement chacun dans une classe et on regroupe les mots puis les groupes de mots qui minimisent l'accroissement de D_P. Cet algorithme est très coûteux, et n'a pu être appliqué sur de grands nombres de mots. Une deuxième méthode [18] est également du type nuées dynamiques mais l'exploration de l'espace est réalisée sytématiquement et dépend crucialement des conditions initiales, cet algorithme est également plus coûteux en temps.

4. CLASSIFICATION DES MOTS SIMILAIRES, C_S

La méthode précédente utilise une approche descendante qui nécessite une connaissance *a priori* du nombre de classes dans lesquelles on autorise la répartition de tous les mots. A l'inverse, la méthode suivante utilise une approche ascendante, cherchant à regrouper une partie des mots dans un nombre de classes non défini à l'avance.

4.1. SIMILARITÉ

Dans cette classification, les mots sont regroupés selon un critère de similarité. On définit la similarité entre deux mots en considérant leurs voisins dans le texte d'apprentissage.

On rappelle que les voisins ou contextes d'un mot sont l'ensemble des mots qui peuvent se trouver immédiatement avant lui (voisins de gauche) ou immédiatement après lui (voisins de droite) dans le texte d'apprentissage. La classification d'un mot nécessite de connaître à la fois son contexte gauche et son contexte droit. En effet, dans le terme $P(C(v_k)|C(v_j))$ de l'équation (2), les classes interviennent aussi bien dans la prédiction (rôle de $C(v_k)$) que dans la condition (rôle de $C(v_j)$).

Si maintenant on considère les deux mots du vocabulaire, v_j et v_n, dont on veut mesurer le degré de similarité, on définit l'intersection des ensembles des voisins de gauche de ces deux mots par l'ensemble des contextes gauches communs à v_j et v_n, noté $CGC(v_j, v_n)$. On détermine de la même manière $CDC(v_j, v_n)$, l'ensemble des contextes droits communs à v_j et v_n.

On définit les similarités S_G et S_D, applications de $V \times V$ dans R^+, à partir des probabilités conditionnelles $p(v_i|v_j)$ et $p(v_i|v_n)$ d'observer le mot v_i connaissant les mots v_j et v_n, le mot v_i pouvant être soit à droite soit à gauche des mots v_j et v_n. On a :

$$\begin{aligned} S_G(v_j, v_n) &= \sum_{v_i \in CGC(v_j, v_n)} [\, p(v_i|v_j) + p(v_i|v_n)\,], \text{ avec } p(v_i|v_*) = \frac{N(v_i\, v_*)}{N(v_*)} \\ \text{et } S_D(v_j, v_n) &= \sum_{v_i \in CDC(v_j, v_n)} [\, p(v_i|v_j) + p(v_i|v_n)\,], \text{ avec } p(v_i|v_*) = \frac{N(v_*\, v_i)}{N(v_*)}. \end{aligned} \tag{8}$$

Les mots similaires sont ceux pour lesquels les probabilités conditionnelles associées sont distribuées sur quasiment le même sous-espace. S_G et S_D sont bien des similarités car elles vérifient les deux propriétés suivantes :

1. $\forall (v_j, v_n) \in V \times V \quad \begin{cases} S_G(v_j, v_n) = S_G(v_n, v_j) \\ \text{et } S_D(v_j, v_n) = S_D(v_n, v_j) \end{cases}$
2. $\forall (v_j, v_n) \in V \times V \quad \begin{cases} S_{G,max} = S_G(v_j, v_j) = S_G(v_n, v_n) = 2 \geq S_G(v_j, v_n) \\ \text{et } S_{D,max} = S_D(v_j, v_j) = S_D(v_n, v_n) = 2 \geq S_D(v_j, v_n) \end{cases}$

A partir de ces similarités, on définit la similarité, S, entre les deux mots v_j et v_n par :

$$S(v_j, v_n) = min(S_G(v_j, v_n), S_D(v_j, v_n)). \tag{9}$$

S est bien une similarité car :

1. $\forall (v_j, v_n) \in V \times V \quad \begin{aligned} S(v_j, v_n) &= min(S_G(v_n, v_j), S_D(v_j, v_n)) \\ S(v_j, v_n) &= min(S_G(v_j, v_n), S_D(v_n, v_j)) \\ S(v_j, v_n) &= S(v_n, v_j) \end{aligned}$
2. d'une part $\forall (v_j, v_n) \in V \times V \quad S(v_j, v_n) = min(S_G(v_n, v_j), S_D(v_j, v_n))$

   or $\forall (v_j, v_n) \in V \times V \quad \begin{cases} S_G(v_n, v_j) \leq S_{G,max} \\ S_D(v_n, v_j) \leq S_{D,max} \end{cases}$ et $S_{G,max} = S_{D,max}$

   donc $\forall (v_j, v_n) \in V \times V \quad S(v_j, v_n) \leq S_{max} = S_{G,max} = S_{D,max}$

   d'autre part $\forall (v_j) \in V \quad \begin{aligned} S(v_j, v_j) &= min(S_G(v_j, v_j), S_D(v_j, v_j)) \\ &= S_{G,max} = S_{D,max} = S_{max} \end{aligned}$

   donc $\forall (v_j, v_n) \in V \times V \quad S(v_j, v_n) \leq S_{max} = S(v_j, v_j) = S(v_n, v_n)$

La similarité entre deux mots est d'autant plus grande que l'intersection de leurs contextes est grande. Cette définition de la similarité S est une extension de celle utilisée en [3], elle a l'avantage de générer une seule matrice, carrée et symétrique.

4.2. ALGORITHME DE C_S

A l'état initial, les mots sont isolés les uns des autres. Ensuite ils sont comparés itérativement deux par deux dans l'ordre alphabétique. L'ordre n'a pas d'importance parce que les mots sont toujours comparés dans leur état initial. Les mots sont regroupés dans un même ensemble si leur similarité est supérieure à un seuil fixé initialement.

Trois cas se présentent :

- Les deux mots comparés n'appartiennent encore à aucun groupe de mots, un nouvel ensemble, contenant ces deux mots, est alors créé.
- Un des deux mots est déjà regroupé avec d'autres, alors le mot qui était seul est ajouté au groupe déjà existant.
- Les deux mots sont déjà regroupés dans des ensembles différents, alors les deux ensembles sont fusionnés.

Cette classification correspond aux composantes connexes d'un graphe seuil [9]. Tous les mots du vocabulaire ne trouvent pas un mot similaire, ce qui signifie qu'à la fin de la classification automatique, certains mots appartiennent à des ensembles d'au moins deux mots et d'autres pas.

On compare chaque mot du vocabulaire v_j à l'ensemble des mots du vocabulaire pour lesquels le regroupement avec v_j n'a pas été essayé, ce qui correspond à $(L_V - 1)/2$ opérations en moyenne. A chaque comparaison de v_j avec un v_n , on repère les contextes de v_j parmis ceux de v_n, soit $2NB_V/L_V$ opérations[3] en moyenne. Les contextes de v_j étant représentés par un vecteur de taille moyenne NB_V/L_V, la classification de l'ensemble des mots est achevée en $(2NB_V/L_V + 2NB_V/L_V * (L_V - 1)/2)L_V \approx NB_V L_V$ opérations.

A titre indicatif, la classification réalisée sur les textes du Monde (1987-1988), s'effectue en 2 heures environ sur une station de travail SGI R10 000.

La notion de similarité a déjà été utilisée pour regrouper des mots non étiquetés, mais avec des définitions différentes du critère de sélection (fréquence des mots , information mutuelle [7]) et de l'espace sur lequel il doit être appliqué [10], [8]. Notre méthode qui prend en compte le contexte immédiat de chaque mot est plus adaptée au point de vue markovien des systèmes de reconnaissance de la parole.

5. CLASSIFICATIONS OBTENUES

Les classifications ont été apprises, indépendamment l'une de l'autre, sur les deux types de textes décrits en section 2.2. Les résultats sont donnés quantitativement et qualitativement avec des exemples de classes.

5.1. CARACTÉRISTIQUES GÉNÉRALES

Les paramètres initiaux, le nombre de classes pour C_P et le seuil de similarité pour C_S sont déterminés expérimentalement sur des données de développement (différentes de l'apprentissage). Pour la classification C_P le nombre de classes est choisi de telle sorte qu'il corresponde à une prédiction optimale sur ces données [13, 14]. Pour la classification C_S, les critères d'optimalité décrits en [3, 4] fixent le seuil de similarité. Le tableau 2

[3] On rappelle que NB_V est le nombre de bigrammes de mots observés dans le texte d'apprentissage et L_V est la taille du vocabulaire

donne, à l'état final, le nombre de classes contenant plus d'un mot, le pourcentage de mots classés, le nombre de bigrammes de classes, le nombre de bigrammes de mots générés par la classification (cf. section 1) et le nombre de bigrammes de mots observés. Les valeurs en italique illustrent la classification des données de journaux par C_S, elles ne correspondent pas à des critères d'optimalité, elles ont été obtenues avec le même seuil que celui des données de parole, et sont données à titre de comparaison.

Tableau 2: Caractéristiques de l'état final des classifications C_P et C_S

Textes	**Journaux**		**Parole**	
Modèles	C_P	C_S	C_P	C_S
Nombre de classes	859	*27*	123	41
Pourcentage de mots classés	99%	*2%*	98%	17%
Bigrammes de classes	750 000	*8 millions*	6 340	10 605
Bigrammes de mots générés	350 millions	*13 millions*	453 030	47 743
Bigrammes de mots observés	8 millions		14 403	

Dans ce tableau on observe une différence significative des nombres de mots classés. Ceci est essentiellement dû au fait que la classification C_P contraint tous les mots à se répartir dans un nombre de classes fixe, alors que la classification C_S ne regroupe que les mots similaires. Rappelons que notre objectif est de trouver un équilibre entre robustesse, précision et complexité, selon les textes d'apprentissage utilisés et non pas de classer tous les mots. On peut également constater que la complexité est réduite puisque le nombre de bigrammes de classes est inférieur au nombre de bigrammes de mots, notablement pour la classification C_P sur les textes de journaux.

5.2. CONTENU DES CLASSES

L'étude *a posteriori* du contenu des classes est subjective. Rappelons que les mots que nous classons ne sont pas étiquetés et que nous n'utilisons pas de classes prédéfinies. Nous présentons quelques classes significatives des classifications apprises sur les textes de journaux et sur les transcriptions de dialogues.

On constate que les classes ont une certaine typologie que l'on peut qualifier *a posteriori*. Cependant, une partie d'entre elles ne peut pas recevoir de dénomination car ces classes ont un contenu sans cohérence apparente. On qualifie une classe lorsque la majorité des mots qu'elle contient a la même typologie.

Rappelons que ces classes de mots sont construites connaissant les mots ou les ensembles de mots voisins et reflètent donc une certaine syntaxe, voire une certaine sémantique de la langue, mais réduites au contexte le plus proche. C'est ainsi que l'on pourra observer des classes de mots au pluriel car ces mots sont précédés d'un même article au pluriel très fréquent, des classes de prénoms car ils sont suivis majoritairement par des noms propres, etc.

5.2.1. Apprentissage sur les textes de journaux

Nous donnons ici quelques exemples de la répartition des 20 000 mots du vocabulaire, apprise à partir des textes de journaux, les nombres entre parenthèses indiquent la fréquence des mots dans le texte.

Dans la classification effectuée par C_P et sur 100 classes, on reconnaît :

- des classes avec un seul mot très fréquent, dont certains sont considérés comme des mots-outils *à* (5 625 414), *de* (14 556 521), *les* (5 019 641), *s'* (1 044 080), *cent* (1 929 492)
- des classes grammaticales :
  - une classe de 544 noms féminins singuliers : *Action* (3 027), *Agence* (9 448), *abbaye* (1 243), *abolition* (1 986), *abondance* (1 798), *absence* (24 847), *absorption* (915), ...
  - une classe de 860 noms communs pluriels : *Amis* (679), *Cahiers* (1 396), *Côtes* (819), *abords* (3 285), *académies* (809), *accents* (1 756), *accusés* (10 869), *adjoints* (2 183), ...
  - une classe de 600 participes passés au masculin singulier : *abaissé* (942), *abandonné* (6 890), *abattu* (5 000), *abordé* (2 437), *abouti* (4 620), *absorbé* (913), ...
  - une classe de 665 verbes à l'infinitif : *baisser* (5 755), *balayer* (699), *barrer* (702), *basculer* (1 555), *battre* (8 330), *bloquer* (3 047), *bombarder* (1 357), *boucler* (1 590), ...
  - une classe de 320 adverbes de manière : *abondamment* (1 502), *abusivement* (605), *accessoirement* (841), *activement* (3 296), *actuellement* (50 161), *admirablement* (942), ...
  - ...
- des classes "sémantiques", par exemple :
  - une classe de 600 prénoms : *Abdel* (4 628), *Abel* (1 121), *Abou* (6 052), *Abraham* (1 696), *Achille* (1 408), *Adam* (1 855), *Adrien* (1 153), *Agnès* (1 531), *Ahmad* (1 548), ... et entités se comportant comme un prénom : *al* (11 115), *cheikh* (3 842), *del* (3 004), *der* (1 835), *di* (2 126), *el* (4 755), *ex* (5 244), *in* (8 671), *lord* (1 713), *m* (1 492), ...
  - une classe de mois : *janvier* (99 658), *février* (79 700), *mars* (104 005), *avril* (91 267), *mai* (99 653), *juin* (109 381), *juillet* (95 756), *août* (74 540), *septembre* (93 639), *octobre* (96 196), *novembre* (85 314), *décembre* (97 751) plus un substantif temporel *semestre* (8 153).
  - des classes de nombres : *cinquante* (175 374), *cinquante-deux* (25 088), *cinquante-quatre* (23 632), *dix* (219 063), *dix-huit* (98 119), *dix-sept* (92 084), *douze* (133 116), ...
  - ...

Dans ces classes, il y a une faible proportion de mots qui ne correspondent pas à l'étiquette donnée *a posteriori*. En revanche, d'autres classes peuvent difficilement recevoir une étiquette par leur inhomogénéité. Par exemple, la classe qui contient les 12 mots suivants : *Amnesty* (3 876), *Antenne* (1 990), *Manchester* (3 836), *New* (31 137), *divers* (45 534), *environ* (64 508), *inscr.* (4 965), *quelque* (90 031), *seules* (10 214), *seuls* (25 517), *tous* (212 041), *toutes* (107 133). Quelques mots, dans cette classe, ont une typologie comparable, mais il y a des mots sans lien apparent avec les autres. Ceci est vraisemblablement dû au fait que chaque mot ne peut être que dans une seule classe, ce qui introduit une distorsion d'autant plus grande que le nombre de classes est faible.

La classification provenant de C_S donne la même typologie que la classification C_P, qui traduit une syntaxe de proximité, mais avec un nombre réduit de mots classés :

- des classes grammaticales :
  - une classe de 15 noms au pluriel : *Internationaux* (2 575), *abords* (3 285), *alentours* (3 518), *allures* (2 751), *centaines* (23 694), *chances* (13 166), *chefs* (33 607), *côtés* (15 083), ...

- une classe de 131 noms au singulier : *Mission* (2 862), *Monde* (81 155), *Quai* (3 869), *Quotidien* (2 059), *baisse* (41 427), *bien-fondé* (1 800), *bord* (29 986), *bordure* (1 383), ...
- une classe de 25 participes passés masculins singuliers.
- plusieurs classes de verbes à l'infinitif :
  obtenir (35 239), *éviter* (33 001).
  assurer(29 303), *empêcher* (16 187).
  ⋮
- ...

• des classes "sémantiques" :
- une classe de 23 chiffres.
- la classe des 12 mois.
- la classe des 2 mots : *matin* (52 069), *soir* (77 536).
- la classe des 2 mots : *heures* (216 297), *minutes* (83 541).
- ...

Pour illustrer comment l'estimation des probabilités varie avec les classifications C_P et C_S, nous avons représenté dans la figure 3 les distributions des probabilités conditionnelles, $q(\text{mot}|Jacques)$ et $q(\text{mot}|le)$, générées par les deux classifications, les distributions initiales sont représentées sur la figure 2a.

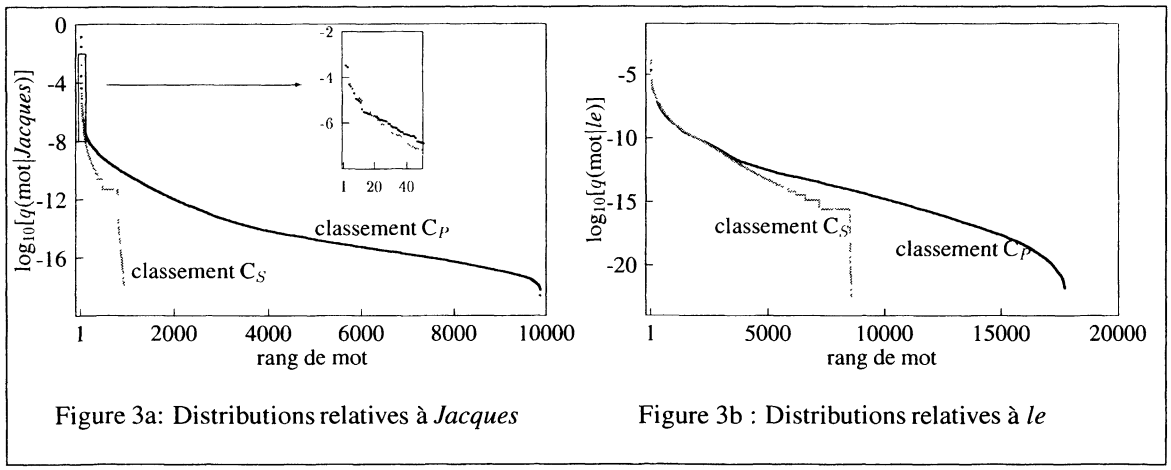

Figure 3a: Distributions relatives à *Jacques*

Figure 3b : Distributions relatives à *le*

Figure 3: Distributions des probabilités conditionnelles $q(\text{mot}|Jacques)$ (Figure 3a) et $q(\text{mot}|le)$ (Figure 3b) dans le journal Le Monde (1987-1997) après classification

Dans les figures 3a et 3b, on peut remarquer que la classification C_P étend davantage les distributions que la classification C_S. Celle-ci suit de très près la distribution p de la figure 2a. Dans le zoom de la figure 3a, les distributions se croisent vers le 20^e rang. Sur la figure 3b, les distributions s'écartent vers le 2000^e rang.

5.2.2. Apprentissage sur les données des dialogues

La taille moins élevée du vocabulaire pour les données provenant des dialogues, permet une analyse plus détaillée du contenu des classes pour les deux classifications.

La classification C_P crée des classes grammaticales, comme :

- une classe de 19 noms et 2 adjectifs, féminins singuliers : *Gare-de-l'Est* (1), *couchette* (18), *date* (57), *disposition* (2), *durée* (2), *fois* (3), *journée* (44), *ligne* (41), *liste* (406), *matinée* (247), ...
  et *meilleure* (12), *mieux* (3).
- une classe de 19 noms pluriels : *abonnements* (2), *allers* (1), *animaux* (20), *avantages* (8), *bars* (6), *billets* (105), *changements* (335), *chats* (9), *choix* (6), ...
- une classe de 8 verbes à l'infinitif : *acheter* (11), *choisir* (31), *essayer* (1), *mettre* (5), *payer* (144), *prendre* (797), *transporter* (3), *trouver* (2) et 1 substantif adjectif *retraité* (1).
- ...

Par contre la classification C_S a très peu de classes grammaticales, la quasi-totalité des classes sont de type "sémantique", par exemple les villes, les mois, les jours, les chiffres (ou mots formant des nombres) et les ordinaux...

On peut donc attribuer *a posteriori* une étiquette à certaines classes ce qui permet d'étudier de manière plus fine leur contenu à l'aide de la terminologie de l'extraction des connaissances [2], en utilisant les mesures de rappel (r), précision (pr) et manque (m) définies plus loin sur un exemple.

Les étiquettes "Villes", "Mois", "Chiffres" et "Ordinaux" ont été attribuées aux classes comportant plus de 50% de ces éléments, plusieurs classes peuvent bénéficier de l'attribution de la même étiquette. Sur l'exemple des villes, en considérant l'ensemble des "Villes", $\mathcal{S}_\vartheta$, l'ensemble des classes portant l'étiquette "Villes", $\mathcal{C}_\vartheta$, l'ensemble des classes ne portant pas l'étiquette "Villes" $\bar{\mathcal{C}_\vartheta}$, et v_k un mot du vocabulaire V, on a :

- $r(\text{"Villes"}) = 100\dfrac{|\{v_k \in \mathcal{C}_\vartheta \cap \mathcal{S}_\vartheta\}|}{|\mathcal{S}_\vartheta|}$,
- $pr(\text{"Villes"}) = 100\dfrac{|\{v_k \in \mathcal{C}_\vartheta \cap \mathcal{S}_\vartheta\}|}{|\mathcal{C}_\vartheta|}$,
- $m(\text{"Villes"}) = 100\dfrac{|\{v_k \in \bar{\mathcal{C}_\vartheta} \cap \mathcal{S}_\vartheta\}|}{|\mathcal{S}_\vartheta|}$.

Le tableau 3 donne, pour les deux classifications et par type "sémantique", le nombre $L_c(*)$ de classes correspondantes, le nombre moyen de mots par classe $N(\mathcal{C}_*)$, la précision, $pr(*)$, le rappel, $r(*)$, et le manque, $m(*)$, pour les "Villes", "Mois", "Chiffres" et "Ordinaux". Le nombre entre parenthèses à côté de chaque type représente le nombre de mots du type contenus dans le vocabulaire.

Tableau 3: Comparaison quantitative de classes sémantiques dans les classifications de C_P et de C_S

Types "sémantiques"	**Classification C_P**					**Classification C_S**				
	$L_c(*)$	$N(\mathcal{C}_*)$	$r(*)$	$pr(*)$	$m(*)$	$L_c(*)$	$N(\mathcal{C}_*)$	$r(*)$	$pr(*)$	$m(*)$
Villes (167)	8	21	90%	89%	10%	3	25	45%	95%	0%
Mois (12)	4	3-4	92%	73%	8%	1	11	92%	100%	8%
Chiffres (23)	4	2-3	44%	100%	17%	4	3-4	57%	100%	0%
Ordinaux (13)	1	12	85%	92%	0%	2	5	77%	100%	0%

On voit que pour les quatre types, la classification C_S est plus précise et a moins de manques que la classification C_P. En revanche, le rappel est, en général, plus élevé

pour C_P que pour C_S. On peut remarquer que pour les chiffres, les ordinaux et les villes, dans le cas de la classification C_S, la somme des rappels et des pertes n'est pas égale à un. En effet les mots non classés ou seuls dans une classe ne sont comptabilisés ni dans les manques, ni dans les rappels, la différence donne donc le pourcentage de mots non classés dans le type.

Sur la figure 4, sont tracées les distributions des probabilités conditionnelles $q(\text{mot}|Paris)$ et $q(\text{mot}|\grave{a})$ obtenues avec les classifications C_S et C_P et réalisées sur les transcriptions de dialogues.

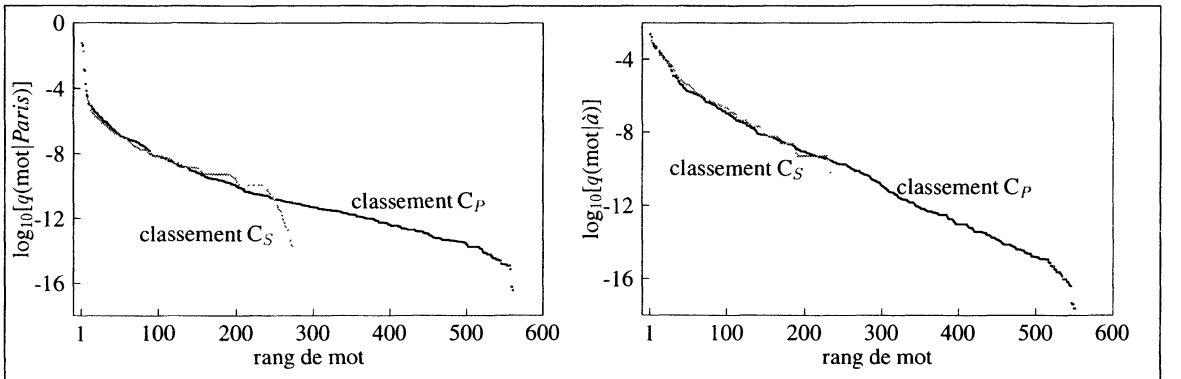

Figure 4a : Distributions relatives à *Paris*

Figure 4b : Distributions relatives à *à*

Figure 4: Distributions des probabilités conditionnelles $q(\text{mot}|Paris)$ (Figure 4a) et $q(\text{mot}|\grave{a})$ (Figure 4b) dans les transcriptions de parole après classification

Dans les figures 4a et 4b, les distributions sont proches jusqu'au 250^e rang environ, la distribution C_P s'étendant davantage.

6. EVALUATION

Les classifications apprises permettent de générer de nouveaux bigrammes de mots et de nouvelles probabilités conditionnelles q. Nous avons évalué l'effet de généralisation de nos classifications sur un texte test, décrit en section 2.2., non utilisé à l'apprentissage. L'ensemble des résultats de l'évaluation est décrit dans les tableaux 4 et 5, le modèle de référence correspond aux bigrammes de mots observés dans le texte d'apprentissage.

6.1. COUVERTURE DES BIGRAMMES DE MOTS DU TEST

Nous avons calculé la couverture des bigrammes de mots du test par les bigrammes de mots générés par les classifications. Les pourcentages indiqués confirme que cette couverture est effectivement plus élevée que celle donnée par les bigrammes de mots initiaux. Une partie de ces bigrammes correspond à des séquences de mots sensées. La classification C_P a une plus grande couverture que la classification C_S.

6.2. QUALITÉ PRÉDICTIVE DES MODÈLES DE LANGAGE

Chaque ensemble des probabilités p ou q constitue ce que nous appelons un modèle de langage en reconnaissance de la parole. Pour évaluer les qualités prédictives d'un modèle de langage, on calcule l'inverse de la moyenne géométrique des probabilités de prédiction

Tableau 4: Couverture du test par différents modèles de langage

Textes	**Journaux**			**Parole**		
Modèles	**Réf.**	C_P	C_S	**Réf.**	C_P	C_S
Nombre de classes	—	859	*27*	—	123	41
Couverture	93,8%	99,8%	*94,2*	85%	95%	88%%

des mots du test, données par le modèle considéré. Cette grandeur est appelée perplexité [16], elle correspond au nombre moyen de mots qui peuvent être prédits à la place de chacun des mots du test. La valeur maximale de la perplexité est V, la taille du vocabulaire, sa valeur minimale est 1, ce qui correspond à un modèle entièrement déterministe. Le modèle de langage est d'autant plus précis que la perplexité est faible. On ne peut comparer des perplexités que lorsqu'elles sont calculées à partir de modèles construits avec les mêmes vocabulaires. Le logarithme de la perplexité est analogue à l'entropie de Shannon [7], si on suppose que le test utilisé donne une bonne représentation statistique de l'ensemble des bigrammes possibles.

Comme la couverture des bigrammes n'est pas complète, toutes les probabilités q ne peuvent être estimées directement à partir du texte d'apprentissage. Pour pallier cet inconvénient, nous avons recours à une méthode d'interpolation couramment utilisée pour générer des modèles de langage statistiques complets. Pour chaque bigramme généré par la classification, on réduit sa fréquence d'une petite quantité, d. La somme des déductions est ensuite répartie sur l'ensemble de tous les bigrammes possibles avec une distribution qui respecte la normalisation des probabilités [18]. Si $\tilde{p}(C(v_k)|C(v_j))$ est l'estimation de la probabilité que la classe $C(v_k)$ soit prédite après la classe $C(v_j)$, on a :

$$\tilde{p}(C(v_k)|C(v_j)) = \frac{Max[N(C(v_j)C(v_k)) - d, 0]}{N(C(v_j))} + d \times \alpha(C(v_j)) \times p(C(v_k)) \qquad (10)$$

avec $\sum_{v_k} p(C(v_k)) = 1$ et $\alpha(C(v_j))$ défini par $\sum_{v_k} \tilde{p}(C(v_k)|C(v_j)) = 1$

Dans le tableau 5 on peut voir l'effet de la classification sur la prédiction des mots.

Tableau 5: Perplexités du test pour différents modèles de langage

Textes	**Journaux**			**Parole**		
Modèles	**Réf.**	C_P	C_S	**Réf.**	C_P	C_S
Nombre de classes	—	859	*27*	—	123	41
Perplexité	149	196	*159*	14,2	15,7	14,9

Les perplexités sont augmentées pour les modèles à base de classes, faiblement pour la classification provenant de C_S et plus fortement pour celui de C_P, on voit ici l'effet de la classification qui réduit la précision des modèles.

Les modèles ne sont pas utilisés tels quels dans le système de reconnaissance. Ils sont combinés avec les modèles de référence, auxquels ils apportent la robustesse qui leur manque, tout en conservant les qualités prédictives du modèle de référence.

7. CONCLUSION

Nous avons décrit deux types de classification automatique de mots appris sur des textes écrits de journaux et de transcriptions de parole. Ces classifications ne nécessitent pas d'étiquetage des mots, elles sont réalisées suivant les contextes locaux dans lesquels les mots ont été observés. L'une des méthodes , basée sur la distance de Kullback-Leibler, répartit tous les mots dans un nombre fixe de classes, en utilisant un algorithme de Monte-Carlo. L'autre méthode regroupe une partie des mots dans un nombre de classes déterminé par l'algorithme, selon un critère de similarité entre les mots pris deux à deux.

Les classes obtenues reflètent une partie de la syntaxe de la langue, voire une certaine sémantique, mais réduites au contexte le plus proche. Les classifications permettent de généraliser les phénomènes observés dans les textes d'apprentissage. Les modèles de langage construits à partir de ces classes ont une meilleure couverture de la langue.

Les méthodes utilisées sont performantes car elles permettent la classification de très grands vocabulaires en des temps relativement courts, de la minute à quelques heures selon la taille du vocabulaire et le type de classification.

BIBLIOGRAPHIE

[1] ADDA G., MARIANI J., LECOMTE J., PAROUBEK P. et RAJMAN M., "The GRACE French Part-of-Speech Tagging Evaluation Task", Actes de *Language Resources and Evaluation Conference*, (1998), 433-441.

[2] AGOSTI M., SMEATON A., *Information Retrieval and Hypertext*, Kluwer Academic Publishers, 1996.

[3] BEAUJARD C., JARDINO M., BONNEAU-MAYNARD H, "Evaluation of a Class-Based Language Model in a Speech Recognizer", Actes de *International Workshop on Speech and Computer*, (1997), 45-50.

[4] BEAUJARD C. et JARDINO M , "Un Modèle de Langage Mixte Basé sur la Similarité des Mots dans un Système de Reconnaissance de Parole", Actes des *Journées d'Étude sur la Parole*, (1998), 343-346.

[5] BROWN P.F. et al., "Class-based n-gram Models of Natural Language", *Computational Linguistics*, (1992), vol.18 n°4 .

[6] CELEUX G. et al., *Classification Automatiques des Données*, Paris, Dunod Informatique, 1989.

[7] COVER T., THOMAS J., *Elements of Information Theory*, Wiley & sons, 1991.

[8] DAGAN I., MARCUS S. et MARKOVITCH S., "Contextual Word Similarity and Estimation from Sparse Data", *Computer Speech and Language*, (1995), vol.9, 123-152.

[9] DUDA R.O., HART P.E., *Pattern Classification and Scene Analysis*, Wiley & sons, 1973.

[10] FARHAT A., ISABELLE J.F. et O'SHAUGHNESSY D., "Clustering Words for Statistical Language Models Based on Contextual Word Similarity", Actes de *IEEE International Conference on Acoustics Speech and Signal Processing*, (1996), vol.1, 180-183.

[11] GAUVAIN J.L., LAMEL L.F. et ADDA G., "The LIMSI 1997 Hub-4E Transcription System", Actes de *DARPA Broadcast News Transcriptions and Understanding workshop*, (1998), 75-79.

[12] HUCKLE C., "Grouping Words Using Statistical Context", Actes de *meeting of the Association for Computational Linguistics*, (1995).

[13] JARDINO M., "Multilingual Stochastic n-gram Class Language Models", Actes de *IEEE International Conference on Acoustics Speech and Signal Processing*, (1996), vol.1 , 161-164.

[14] JARDINO M., BEAUJARD C., "Rôle du Contexte dans les Modèles de Langage n-classes , Application et Evaluation sur MASK et RAILTEL", Actes des *Journées Scientifiques et Techniques*, (1997), 71-74.

[15] JELINEK F., *Statistical Methods for Speech Recognition*, MIT Press, (1998)

[16] JELINEK F., MERCER R.L. et BAHL L.R., "The Developpment of an Experimental Discrete Dictation Recognizer", *IEEE*, (1985), vol.73 n°11 , 1616-1624.

[17] LAMEL L.F. et al., "Developpment of Spoken Language Corpora for Travel Information", Actes de *European Conference on Speech Communication and Technology*, (1995), vol.3, 1961-1964.

[18] NEY H., ESSEN U. et KNESER R., "On Structuring Probabilistic Dependences in Stochastic Langage Modelling", *Computer Speech and Language*, (1994), vol.8.