*TALN 2005, Dourdan, 6–10 juin 2005*

# Évaluation des Modèles de Langage $n$-gramme et $n/m$-multigramme

P. Alain, O. Boëffard
IRISA – Université de Rennes 1 / ENSSAT
6, rue de Kerampont, 22305 Lannion
{pierre.alain,olivier.boeffard}@irisa.fr

**Mots-clefs :** Modèles de Langage statistiques, $n$-gramme, multigramme, évaluation

**Keywords:** Statistical Language Models, $n$-grams, phrase multigrams

**Résumé** Cet article présente une évaluation de modèles statistiques du langage menée sur la langue Française. Nous avons cherché à comparer la performance de modèles de langage exotiques par rapport aux modèles plus classiques de $n$-gramme à horizon fixe. Les expériences réalisées montrent que des modèles de $n$-gramme à horizon variable peuvent faire baisser de plus de 10% en moyenne la perplexité d'un modèle de $n$-gramme à horizon fixe. Les modèles de $n/m$-multigramme demandent une adaptation pour pouvoir être concurrentiels.

**Abstract** This paper presents an evaluation of statistical language models carried out on the French language. We compared the performance of some exotic models to the one of the more traditional n-gram model. The experiments show that the variable $n$-gram models can drop more than 10% of the average perplexity for a fixed $n$-gram model. $n/m$-multigram models require an adaptation to be able to compete.

## 1 Introduction

La modélisation du langage est un problème crucial et très largement abordé en traitement automatique de la langue écrite ou parlée[1]. À partir de l'observation de séquences de mots, il s'agit de construire un modèle dont l'objectif est de prédire avec succès de nouvelles séquences. On peut distinguer déjà deux problèmes, d'une part celui du choix du modèle et de sa méthodologie de construction et d'autre part celui de la méthodologie d'évaluation d'un modèle de langage. Concernant le premier point, on peut distinguer des approches déterministes qui tiennent compte de l'organisation profonde des mots liées notamment à la syntaxe, des approches probabilistes qui s'intéressent essentiellement à la forme de surface (Rosenfeld, 2000).

L'évaluation est un point relativement délicat dans la mesure où elle peut être dépendante du modèle choisi. La mesure la plus communément adoptée consiste à calculer l'entropie croisée entre un modèle de langage et la distribution réelle des données observées, mais inconnue. En supposant que les données suivent une distribution stationnaire et ergodique, le calcul de l'entropie-croisée peut être estimé à partir d'un corpus suffisamment grand[2]. La perplexité d'un modèle de langage n'est qu'une autre manière de représenter le degré d'incertitude d'un modèle et se calcule directement à partir de l'entropie-croisée du modèle sur un jeu de phrases de test. Pour un mot à prédire, la valeur de la perplexité représente le

[1] On peut citer le domaine de la reconnaissance automatique de la parole mais aussi celui de la reconnaissance de texte manuscrit ou encore celui de la traduction automatique.

[2] Théorème de Shannon-MacMillan-Brieman, (Shields, 1998). En respectant ces hypothèses de stationnarité et d'ergodicité, un corpus de parole de longueur finie peut refléter la distribution réelle des données.

nombre d'hypothèses moyennes de branchement[3]. Plus la perplexité est faible, plus le facteur moyen de branchements d'un mot vers un autre est bas et plus le modèle de langage est efficace. Pour les modèles $n$-gramme, un mot est prédit en tenant compte d'un historique relativement limité des mots qui le précèdent. Ces modèles connaissent finalement très peu des raisons profondes de l'organisation des mots dans une phrase. En revanche, l'utilisation de probabilités conditionnelles et un apprentissage réalisé à partir de quelques millions de phrases permettent d'obtenir de bonnes performances. Leur principal défaut réside dans la complexité spatiale sous-jacente. Théoriquement, plus la séquence de l'historique du modèle s'allonge ($n$ augmente), plus le modèle répartit efficacement la masse de probabilités sur des mots qui reviennent souvent après une valeur particulière de l'historique. Cependant, plus $n$ augmente, plus les observations se raréfient compte-tenu de la nature hyperbolique de la distribution de ces événements[4]. Pour des situations expérimentales réelles, les valeurs courantes de $n$ dépassent rarement 4 (Siu & Ostendorf, 2000). De nombreuses solutions ont été apportées au problème de l'explosion combinatoire et à celui de la raréfaction des événements (Rosenfeld, 2000). Des techniques de lissage permettent notamment de répondre à la difficulté de l'estimation d'une distribution de probabilité lorsque les événements sont rares. On peut citer le principe du lissage qui n'effectue l'estimation des points de la densité au sens du maximum de vraisemblance que pour des événements dont l'occurrence est supérieure à un seuil de *cut-off*. Une partie de la masse de probabilité est répartie sur des événements dont l'occurrence est inférieure au seuil, (Katz, 1987). (Chen & Goodman, 1999) propose une évaluation des principales techniques de lissage les plus utilisées.

Les modèles de $n$-gramme pour lesquels la longueur de l'historique est variable sont une alternative aux $n$-gramme classiques pour lesquels la longueur de l'historique reste fixe. Le principe consiste à ne pas retenir un historique de longueur $n$ si la contribution du $n$-gramme correspondant n'améliore pas la performance du modèle. Toute la difficulté réside dans la décision d'abandon d'un $n$-gramme pour un $(n-k)$-gramme avec $1 \leq k < n$, (Niesler & Woodland, 1994)(Siu & Ostendorf, 2000).

Les modèles multigramme sont des modèles de type $n$-gramme où la tête peut avoir une longueur supérieure à 1.(Bimbot *et al.* , 1995)(Deligne & Bimbot, 1995) présentent un cadre théorique pour des multigramme formés sur des modèles d'uni-gramme (longueur d'historique nulle). Les expériences rapportées concernent une application avec un vocabulaire limité de 900 mots pour un corpus d'apprentissage de 100 000 phrases et un corpus de test de 1 000 phrases (dont 52 occurrences de mots hors-vocabulaire). Les modèles de type multigramme obtiennent une perplexité meilleure que les $n$-gramme classiques lorsque $n > 3$. Compte-tenu de la taille relativement limitée des corpus, les conclusions sont difficilement transposables directement sur des corpus plus importants. (Deligne & Sagisaka, 2000) se place dans un contexte de multigramme de classes de mots sur des modèles de bi-gramme. Les expériences sont menées avec un vocabulaire d'environ 3 000 mots, 100 000 phrases pour le corpus d'apprentissage et environ 700 phrases pour le test. Deux types de mesure sont rapportés : d'une part la perplexité pour les modèles de type multigramme et d'autre part le taux d'erreur d'un système de reconnaissance de la parole. Les résultats entre multigramme et $n$-gramme classiques (bi- et tri-gramme) sont difficilement comparables. En effet, pour ces derniers, les valeurs de perplexité sont absentes et les modèles de $n$-gramme semblent avoir été non réduits[5]. (Zitouni, 2002) propose des modèles de multigramme où les probabilités de co-occurrence de mots sont conditionnées par rapport à des classes. Les expériences concernent deux années du journal "Le Monde" (années 1987 et 1988) pour un vocabulaire de 20 000 mots. L'utilisation de ces multigramme permet de réduire de 7% la perplexité des tri-gramme classiques. Encore une fois, il est difficile de retrouver sur cette expérience une comparaison entre modèles à nombre de paramètres constant.

Cet article propose une étude expérimentale sur les performances relatives des modèles de language

[3] Il s'agit d'une moyenne géométrique.

[4] Il s'agit de distributions à queue lourde où beaucoup d'événements sont extrêmement rares et peu sont très fréquents. La loi de Zipf est un cas particulier de lois puissances caractéristiques de ce phénomène.

[5] Le comportement d'un $n$-gramme est non-linéaire, il est possible de réduire de façon importante le nombre de paramètres sans trop dégrader ses performances.

de type $n$-gramme à horizon fixe, à horizon variable et multigramme. La section 2 présente un cadre théorique pour ces trois types de modèles de langage statistiques. La section 3 expose la problématique d'une telle expérimentation ainsi que nos hypothèses de travail. Une évaluation a été menée sur environ un million de phrases en français. La section 4 décrit la méthodologie suivie. La section 5 expose les expériences mises en œuvre. Enfin, la section 6 présente les résultats et une interprétation du comportement des modèles en fonction des données traitées.

## 2 Cadre théorique

Un modèle de langage statistique est un ensemble de distributions de probabilité sur des séquences observées de symboles. Comme, en pratique, il est impossible de caractériser de telles distributions, les modèles de langage se distingueront entre eux par les hypothèses choisies pour réduire la complexité combinatoire et améliorer leur capacité de généralisation. Après une présentation des notations utilisées, nous discutons du modèle de $n$-gramme à horizon fixe, du modèle de $n$-gramme à horizon variable et enfin du modèle $n/m$-multigramme.

Soit une séquence de mots $W = (w_1, w_2, ..., w_N)$ avec $w_i$ une variable représentant le mot de rang $i$ dans la séquence $W$. Les valeurs possibles pour $w_i$ appartiennent à un vocabulaire $\mathcal{V}$. Il peut s'agit souvent d'un vocabulaire fermé dans le cadre de systèmes de dialogue, nous considérons ici l'étude de la langue naturelle, nous choisissons un vocabulaire ouvert. Nous pouvons décrire cette séquence par une suite de variables aléatoires $w_i$. La probabilité conjointe des variables de la séquence $W$ peut se développer de la manière suivante en faisant apparaître des probabilités conditionnelles :

$$p(W) \quad = \quad p(w_1) \times \prod_{i=2}^{N} p(w_i | w_1, ...w_{i-1}) \tag{1}$$

L'objectif d'un modèle de langage consiste à calculer cette probabilité conjointe, c'est-à-dire à estimer des valeurs pour chacune des probabilités conditionnelles. L'estimation de ces probabilités conditionnelles est en pratique impossible car le nombre de paramètres croît de manière exponentielle avec la longueur de la suite de mots. Pour contrer cette difficulté, un modèle de langage pose une probabilité conditionnelle approchée $p^*(.)$ en simplifiant la loi conjointe, équation 1.

On note $\mathcal{G}$ l'ensemble des groupes de mots formés sur le vocabulaire $\mathcal{V}$. On note $\mathcal{S}$ l'ensemble des séquences formées sur les éléments de $\mathcal{G}$. On note $\mathcal{S}^* \subset \mathcal{S}$, l'ensemble des séquences de $\mathcal{S}$ qui correspondent à $W$. On note $S$ une séquence particulière de $\mathcal{S}^*$. Par exemple, pour $W = (w_1,\ w_2,\ w_3)$, on a :

$$\mathcal{S}^* \quad = \quad \left\{ \begin{array}{l} [w_1][w_2][w_3] \\ [w_1, w_2][w_3] \\ [w_1, w_2, w_3] \\ [w_1][w_2, w_3] \end{array} \right.$$

On note $|S|$ le nombre de groupes de mots dans la séquence $S$. Soit $k$ le groupe de mots de rang $k$ dans la séquence $S$, on note $i(S(k))$ l'indice dans la séquence $W$ du premier mot de $S(k)$. On note $l(S(k))$ le nombre de mots de $S(k)$.

On note $h_{i,j}(W)$ la chaîne des variables aléatoires représentant l'apparition conjointe de tous les mots $w_u$ de $W$ pour $u \in [i, i + (j - 1)]$.

$$h_{i,j}(W) \quad = \quad \left\{ \begin{array}{ll} w_i, w_{i+1}, ..., w_{i+(j-1)} & \text{si } i + (j-1) \leq N \\ w_i, w_{i+1}, ..., w_N & \text{sinon} \end{array} \right.$$

On définit également l'opérateur $t_{i,j}(W)$ qui représente les $j$ mots précédents le mot $w_i$. On a donc $t_{i,j}(W) = h_{i-j,j}(W)$. $t_{i,j}(W)$ correspond à un horizon ou historique (un groupe de mots qui précède

l'observation d'un mot). $h_{i,j}(W)$ correspond à la tête d'un paramètre du modèle de langage (pour les modèles $n$-gramme à horizon fixe ou variable, la variable aléatoire de tête est dégénérée, et ne contient qu'un seul mot). Les modèles de langage cherchent d'une part à réduire au maximum la longueur d'un horizon (minimisation du nombre de paramètres) et d'autre part, pour un horizon donné, à estimer la distribution de probabilité des historiques pour calculer la probabilité d'apparition du mot $w_i$. Soit $W$ associée à une séquence de découpage $S$, la loi conjointe estimée par le modèle de langage peut alors se réécrire sous la forme suivante avec $n$ l'ordre du $n$-gramme :

$$p_S^*(W) \quad = \quad p(h_{i(S(1)),l(S(1))}(W)) \times \prod_{k=2}^{|S|} p(h_{i(S(k)),l(S(k))}(W)|t_{i(S(k)),n-1}(W)) \tag{2}$$

## 2.1 Les modèles $n$-gramme à horizon fixe

Pour un $n$-gramme à horizon fixe, on fait une hypothèse d'indépendance conditionnelle du mot $w_i$ avec les mots présents dans la séquence à une distance de plus de $n-1$ mots (pour $n=2$, ce modèle est un modèle de bi-gramme ; la probabilité $P(W)$ correspond à celle d'une chaîne de Markov. Pour $n=3$, on parle de tri-gramme et pour $n=4$ de quadri-gramme). Comme nous l'avons déjà souligné, ce modèle est très simple, mais le nombre de paramètres croît de manière exponentielle avec $n$. Pour cette raison, les modèles de $n$-gramme les plus utilisés le sont pour des valeurs de $n$ de l'ordre de 3 ou 4. Pour corriger le problème des événements rares, il existe des techniques de lissage des probabilités conditionnelles, couplées à des techniques de *back-off* permettant de corriger celles d'événements manquants lors de l'apprentissage, (Katz, 1987). Cette correction s'effectue en pondérant la probabilité du $(n-1)$-gramme par un coefficient de *back-off* de telle manière que la distribution de probabilité des $n$-gramme somme toujours à 1. Le terme produit de l'équation 2 se simplifie alors :

$$p(h_{i(S(k)),l(S(k))}(W)|t_{i(S(k)),n-1}(W)) \quad \stackrel{\Delta}{=} \quad p(h_{i,1}(W)|t_{i,n-1}(W)) \tag{3}$$

Pour ce modèle, $S=W$, on obtient simplement :

$$p_{ML}(W) \quad = \quad p_S^*(W)$$

## 2.2 Les $n$-gramme à horizon variable

Forcer l'estimation du terme produit de l'équation 2 à un historique de longueur $n$ introduit un double biais. D'une part les occurrences sont plus faibles, on a donc tendance à faire du lissage et à être moins précis. D'autre part, on introduit des distributions conditionnelles sur $w_i$ qui ne servent pas à grand chose (augmentation injustifiée du nombre de paramètres). Autoriser une variation de la longueur de l'historique pour prédire $w_i$ permet de régler ce problème de sur-apprentissage. Les $n$-gramme à horizon variable définissent une probabilité en adaptant une longueur d'historique optimale en fonction de $w_i$. L'approche traditionnelle pour ce type de modèles consiste à déterminer au moment de l'apprentissage les longueurs optimales à retenir, (Bonafonte & Mariño, 1996)(Siu & Ostendorf, 2000). Dans cette situation un $n$-gramme est remplacé par un $(n-k)$-gramme avec $1 \leq k < n$. Les $n$-gramme à horizon variable peuvent apparaître intéressants pour un double enjeu : d'une part à nombre de paramètres fixé, il peuvent répondre à une amélioration de la performance des $n$-gramme à horizon fixe et d'autre part, à perplexité fixée, ils peuvent être utiles à la diminution du nombre de paramètres d'un modèle de langage.

Au moment du test, lors du calcul de la perplexité d'une phrase, pour ce modèle de $n$-gramme à horizon variable, le terme produit de l'équation 2 s'écrit :

$$p(h_{i(S(k)),l(S(k))}(W)|t_{i(S(k)),n-1}(W)) \quad \stackrel{\Delta}{=} \quad \max_{1 \leq v \leq n-1} \{p(h_{i,1}(W)|t_{i,v}(W))\} \tag{4}$$

Cette écriture signifie que pour chaque mot $w_i$ à prédire, on cherche à maximiser la probabilité en se basant sur des modèles allant du bi-gramme au $n$-gramme (équation 3). Pour ce modèle, $S = W$, on obtient simplement :

$$p_{ML}(W) \quad = \quad p^*_S(W)$$

### 2.3 Les $n/m$-multigram

Un $n/m$-multigramme correspond à une probabilité conditionnelle où la tête du $n$-gramme peut être plus longue qu'un mot unique. $m$ représente le nombre maximum de mots dans un groupe de mots en tête. Lors du test du modèle de langage, pour une découpe $S$ donnée, nous cherchons la meilleure probabilité suivant l'équation :

$$p(h_{i(S(k)),l(S(k))}(W)|t_{i(S(k)),n-1}(W)) \quad \triangleq \quad \max_{1\leq u\leq m, 1\leq v\leq m\times(n-1)} \{p(h_{i,u}(W)|t_{i,v}(W))\} \qquad (5)$$

Il suffit ensuite de prendre la meilleure solution sur toutes les séquences $S \in \mathcal{S}^*$ :

$$p_{ML}(W) \quad = \quad \arg\max_{S\in\mathcal{S}^*}\{p^*_S(W)\}$$

## 3 Problématique et hypothèses méthodologiques

Notre objectif est de vérifier l'intérêt des modèles $n$-gramme à horizon variable par rapport à des modèles à horizon fixe et à des modèles de type $n/m$-multigramme. La difficulté de mise en œuvre d'une telle évaluation réside dans le problème du contrôle explicite des paramètres lors de la construction des modèles. Différents facteurs sont responsables de la qualité d'un modèle de langage. Certains influent directement le processus d'apprentissage alors que d'autres déterminent la mesure de performance d'un modèle.

Tout d'abord l'estimation des probabilités conditionnelles provient directement de la détection de $n$-uplets. Avec peu de séquences, on défavorise notamment les modèles de $n$-gramme d'ordre supérieur. Le nombre de paramètres d'un modèle de langage de type $n$-gramme est proportionnel à $|\mathcal{V}|^n$. Le *cut-off* est une technique simple et relativement efficace pour limiter le nombre de paramètres (Chen & Goodman, 1999). Il s'agit de ne pas retenir les $n$-uplets qui apparaissent sous un seuil d'occurrence. Ainsi, un *cut-off* à 1 signifie qu'un mot doit apparaître au moins 2 fois pour être intégré au modèle de langage. Cependant, compte-tenu de la forme des distributions de probabilité (fonction puissance), la réduction conséquente du nombre de paramètres n'est pas linéaire en fonction de la valeur de *cut-off*. En introduction, nous avons souligné le rôle de la perplexité comme outil de mesure de la qualité d'un modèles de langage.

Un autre facteur clé que l'on doit maintenir entre les différents modèles de langage pour pouvoir comparer les valeurs de perplexité est le nombre de mots hors-vocabulaire. Plus la taille du vocabulaire est faible (et donc plus le nombre de paramètres est faible), plus le taux des mots hors-vocabulaire augmente avec des valeurs de perplexité qui s'améliorent. Il s'agit d'un facteur calculé a posteriori, une fois le modèle construit. Il est donc difficile d'intervenir explicitement sur cette valeur.

Le calcul de la perplexité peut varier notamment par la prise en compte ou non des mots hors-vocabulaire sur l'ensemble de test. On peut décider de ne pas prédire un mot hors vocabulaire ; dans ce cas l'accumulation de la perplexité est plus faible mais le nombre de mot prédit n'augmente pas. Le calcul de la perplexité fait intervenir une hypothèse de stationnarité et d'ergodicité qu'il faudrait vérifier en pratique. La performance d'un modèle de langage dépend donc étroitement du couple ensemble d'apprentissage/ensemble de test. Il faut que ces ensembles contiennent un nombre suffisant de

séquences pour pouvoir conclure à des résultats stables. Au cours de nos expériences, nous avons essayé de minimiser l'influence de chacun de ces facteurs de manière à favoriser la comparaison entre structures de modèles ($n$-gramme à horizon fixe, $n$-gramme à horizon variable et $n/m$-multigramme).

Nous avons considéré les hypothèses méthodologiques suivantes. Une année du journal "Le Monde" a été choisie comme univers linguistique (année 1997). Après extraction des phrases et tirage aléatoire, ce corpus est reparti en deux sous-corpus : $70\%$ pour le corpus d'apprentissage et $30\%$ pour le corpus de test. Le choix d'un corpus fixe est suffisant pour valider une comparaison entre modèles, mais ne permettra pas de conclure sur la robustesse des résultats. Des analyses complémentaires seront donc nécessaires. Nous avons considéré trois ensembles de mots : un premier vocabulaire à 3 000 mots, un deuxième à 30 000 mots et un dernier à 60 000 mots (il s'agit à chaque fois des plus fréquents sur l'ensemble d'apprentissage). Les valeurs de perplexité et les taux de mots hors vocabulaire dépendent directement de ces trois ensembles. Nous avons cherché à contrôler explicitement le nombre de paramètres de nos modèles. Deux approches complémentaires ont été mises en œuvre : d'une part par application de seuils de coupure sur les différents types de $n$-gramme et d'autre part par la conservation des co-occurrences de $m$-uplets de mots les plus fréquentes pour les $n/m$-multigramme. Dans le premier cas, nous balayons un spectre de valeurs de *cut-off* et nous observons a posteriori le nombre de paramètres. Ce nombre nous sert ensuite à ajuster le nombre de multigramme autorisés à entrer dans le vocabulaire et se placer ainsi à nombre de paramètres constant (avec une tolérance de 1%). La perplexité calculée ne tient pas compte des mots hors vocabulaire qu'ils soient présents dans la tête ou dans l'historique d'un $n$-gramme.

Notre système de référence est celui des $n$-gramme classiques (que nous avons nommé $n$-gramme à horizon fixe). Nous avons choisi des valeurs communément admises pour $n$ et introduit des modèles de bi-, tri- et quadri-gramme. L'estimation de ces modèles utilise le lissage des probabilité de Good-Turing, selon les recommandations classiques de lissage, *discounting*, et *back-off* (Chen & Goodman, 1999). Nous cherchons tout d'abord à comparer les $n$-gramme à horizon fixe avec des $n$-gramme à horizon variable. Les $n$-gramme à horizon variable sont mis en œuvre lors du test, en appliquant l'équation 4. Notre objectif n'est pas de valider une technique de réduction de paramètres au moment de la construction du modèle, (Siu & Ostendorf, 2000)(Niesler & Woodland, 1994), mais plutôt de débrider un modèle de $n$-gramme à horizon fixe pour en faire un modèle de $n$-gramme à horizon variable. Notre manière de procéder introduit un coût de calcul supplémentaire, mais il reste acceptable car les longueurs des historiques sont faibles devant le nombre de mots à traiter.

Nous cherchons enfin à situer les modèles $n/m$-multigramme par rapport aux deux approches précédentes. L'intérêt du multigramme réside dans sa capacité à prédire une séquence de mots avec un seul paramètre. En moyenne on baisse le nombre de termes impliqués dans le calcul de la perplexité ; il s'agit alors d'une situation favorable. Cependant, le risque est de répartir une masse de probabilités sur plus de termes[6]. Pour que la compétition entre modèles reste équitable, nous avons choisi de travailler avec des modèles $n/m$-multigramme dont la taille maximale (en nombre de mots) est soumise à une contrainte.

## 4 Estimation des paramètres des modèles

Les expériences sont réalisées à partir de la suite de programme *HTK* (Woodland & Young, 1993). Cet ensemble de librairies et d'outils correspond à une chaîne complète permettant de construire et de tester un modèle de langage. La gestion des $n$-gramme à horizon variable n'est pas écrite dans la distribution standard de *HTK*. La modification du programme de test du modèle de langage a été nécessaire pour introduire le traitement proposé équation 4. La gestion des $n/m$-multigramme n'est pas non plus écrite. Les modifications à faire sont d'une part dans le programme d'apprentissage, afin de

[6]Un $n$-gramme classique estime, pour chaque historique, une densité de probabilité dont la complexité spatiale est celle du vocabulaire. Les multigramme avec des têtes de longueur au plus $m$ ont une complexité spatiale bornée par $|\mathcal{V}|^m$, les probabilités tendent vers 0.

parcourir systématiquement toutes les unités de multigramme possibles Il est également nécessaire de modifier, comme pour les $n$-gramme à horizon variable, le programme de test, pour pouvoir parcourir toutes les têtes et tous leurs historiques possibles.

Pour les $n$-gramme à horizon fixe, la perplexité du modèle de langage est déterminée directement grâce à l'équation 6. Si le mot de tête du $n$-gramme n'est pas dans le vocabulaire sélectionné, il est alors compté comme mot hors vocabulaire.

$$\begin{aligned} PP &= 2^{H^*} \quad \text{avec} \\ H^* &= -\frac{1}{m}\log_2\left(P(w_1, w_2, \ldots, w_m)\right) \end{aligned} \tag{6}$$

Pour les $n$-gramme à horizon variable, la situation est différente : pour chaque mot plusieurs choix sont possibles (le choix se fait entre un 2-gramme, un 3-gramme, ..., ou un $n$-gramme). Il suffit de choisir *le meilleur* $k$-gramme parmi les $n-1$ possibles (choix parmi toutes les longueurs d'historique autorisées). Cet algorithme est appliqué phrase par phrase (hypothèse d'indépendance des phrases entre elles). En fin de traitement d'une phrase on connaît la perplexité évaluée sur cette phrase, le nombre de mots prédits ainsi que le nombre de mots hors vocabulaire.

Pour les $n/m$-multigramme, la situation est encore différente. Maintenant plusieurs têtes sont disponibles, et pour chacune d'elles, plusieurs choix sont possibles. Nous avons volontairement limité la taille maximale du $n/m$-multigramme à un nombre fixe de mots : avec des multigramme de taille au plus 2, nous pourrions former 4 bi-gramme : $P(w_i|w_{i-1})$, $P(w_i|[w_{i-2}\ w_{i-1}])$, $P([w_i\ w_{i+1}]|w_{i-1})$, $P([w_i\ w_{i+1}]|[w_{i-2}\ w_{i-1}])$. En limitant le nombre maximum de mots dans le $n/m$-multigramme, nous pouvons choisir le modèle de langage avec lequel nous entrons en concurrence. Par exemple, avec des multigramme de taille au plus 2, et une somme à 3, nous n'avons plus que 3 choix possibles : $P(w_i|w_{i-1})$, $P(w_i|[w_{i-2}\ w_{i-1}])$, et $P([w_i\ w_{i+1}]|w_{i-1})$. Dans le programme de test, afin de sélectionner la meilleure découpe de la phrase selon le $\max$ de l'équation 5, nous avons mis en place une recherche du meilleur chemin dans un graphe[7] orienté et valué selon l'algorithme de Dijkstra.

## 5 Méthodologie expérimentale

Les expériences sont réalisées sur un corpus de texte du français : tous les articles parus pendant l'année 1997 dans le journal "Le Monde" (ressource ELRA). Ce corpus est découpé en phrases par un logiciel d'analyse syntaxique (logiciel Cordial de Synapse). Les phrases sont uniformisées par une réécriture systématique en majuscules et la suppression de toute ponctuation. Le corpus ainsi obtenu contient 1 131 135 phrases pour un vocabulaire de 219 034 mots. Il s'agit de la taille exacte du vocabulaire (mots variants en genre et en nombre, ainsi que les verbes rencontrés sous une forme conjuguée), le nombre d'occurrence des mots est de 23 999 626. L'apprentissage se fait sur 70% du corpus, le test est réalisé sur les 30% restant. La répartition des phrases a été réalisée de manière aléatoire à partir du corpus d'origine.

Pour faire baisser le nombre de paramètres d'un modèle de taille $n$, on fait évoluer la valeur de *cut-off* sur des $n$-gramme à horizon fixe. On conserve une valeur de *cut-off* à 1 sur les paramètres d'ordre inférieur (horizon de longueur inférieure à $n$). Ainsi, pour faire baisser le nombre de paramètres d'un modèle de tri-gramme, on va augmenter la valeur du *cut-off* sur les probabilités conditionnelles des tri-gramme, et laisser constante la valeur de *cut-off* pour les probabilités de bi-gramme et d'uni-gramme. Pour les $2/2$-multigramme, on peut faire baisser le nombre de paramètres en limitant le nombre de multigramme autorisés dans le modèle de langage[8]. On peut ainsi fixer le nombre de paramètres du modèle $n/m$-

[7] L'algorithme de Viterbi permet de rechercher la meilleure solution a priori, nous lui préférons l'algorithme de Dijkstra qui permet d'obtenir la meilleure solution a posteriori.

[8] avec un nombre de multigramme à 0, on obtient un modèle de bi-gramme ; cela est visible sur la figure 1 en prolongeant la courbe de perplexité des $n/m$-multigramme.

multigramme de façon précise grâce à un algorithme de dichotomie qui sélectionne le bon nombre de multigramme à prendre en compte dans la suite.

Si au moins un des mots de l'historique n'est pas présent dans le vocabulaire alors le modèle de langage déclare ne pas pouvoir prédire le $n$-gramme. Le mot de tête du $n$-gramme est alors déclaré non prédit, et la perplexité n'évolue pas. Dans la situation où tous les mots sont présents dans le vocabulaire, mais où la probabilité du $n$-gramme n'a pas été apprise par le modèle de langage, le système de *back-off* déjà présenté s'applique. Dans le cas des $n$-gramme à horizon variable, la probabilité est évaluée de manière identique, le mot en tête du $n$-gramme est déclaré non prédit si au moins un mot de son horizon est hors vocabulaire. Si tous les mots de l'horizon sont dans le vocabulaire, le choix de la meilleure probabilité est réalisé selon l'équation 4. Pour les multigramme, l'algorithme de Dijkstra permet de déterminer la meilleure solution au sens de l'équation 5, parmi toutes les solutions possibles.

## 6 Résultats et commentaires

La figure 1 présente l'évolution de la perplexité en fonction du nombre de paramètres des différents modèles pour différentes tailles de vocabulaire. Le modèle de bi-gramme à horizon variable est exactement le modèle de bi-gramme à horizon fixe, les courbes de perplexité sont donc confondues. On peut observer que le modèle de $n/m$-multigramme tend à avoir un comportement de bi-gramme de mots quand le nombre de multigramme autorisés diminue.

La perplexité d'un modèle de langage augmente quand le nombre de paramètres utilisé baisse. Cela est parfaitement normal, car le pouvoir de prédiction d'un mot de la langue est moins important avec un nombre de paramètres inférieur. Un modèle de $n$-gramme semble avoir une perplexité plus importante qu'un modèle de $n+1$-gramme. Cependant (Bonafonte & Mariño, 1996) rapporte que la perplexité des $n$-gramme augmente à partir de $n = 5$. Un modèle de tri-gramme avec un seuil de *cut-off* à 2 a une perplexité et un nombre de paramètres plus faible qu'un modèle de bi-gramme avec un seuil à 0 ; le modèle de tri-gramme est donc préférable dans ce cas. Selon (Rosenfeld, 2000), l'intérêt comparé d'un modèle de langage apparaît lorsque la mesure de perplexité baisse de plus de 10%. Le modèle de tri-gramme est donc notablement plus intéressant que le modèle de bi-gramme. Tout comme le modèle de quadri-gramme est plus intéressant que le modèle de tri-gramme.

Un modèle de $n$-gramme à horizon variable, comparativement au $n$-gramme concurrent, à horizon fixe, obtient une perplexité[9] plus faible. Cette baisse de la perplexité est due pour partie au calcul de la probabilité maximum ; en effet, par construction, on obtient une probabilité au moins supérieure à celle déterminée par le modèle à horizon fixe. Le gain obtenu par des $n$-gramme à horizon variable provient également de l'utilisation du coefficient de *back-off* par le modèle de $n$-gramme. En effet, le modèle de $n$-gramme utilise un coefficient de *back-off* pour obtenir une probabilité de $n$-gramme à horizon fixe à partir de la probabilité du $n-1$-gramme qui lui correspond si le $n$-gramme n'est pas trouvé. Le modèle de $n$-gramme à horizon variable permet de mélanger les probabilités des différents $(n-k)$-gramme avec $m \in [1, n-2]$, et ce sans pénaliser des $(n-k)$-gramme d'ordre inférieur.

Les $n/m$-multigramme se montrent moins performants que le modèle de $n$-gramme de même ordre (c'est à dire a nombre de mot considérés constants). En effet, l'équation 5 semble indiquer que le choix de la meilleure probabilité se fait entre un bi-gramme de mots, un tri-gramme de mots, et un bi-gramme ayant 2 mots en tête (dans le cas où la taille maximum d'un multigramme est de 2 mots, et la somme des mots du bi-gramme est d'au plus 3). Le choix ne peut donc par construction qu'être au moins aussi bon qu'un tri-gramme de mots. Cependant, nous pouvons constater que pour obtenir un nombre de paramètres équivalent afin de comparer les différents modèles, il faut interdire un nombre conséquent de multigramme parmi ceux disponibles. Nous devons alors chercher à améliorer ce modèle $n/m$-multigramme.

[9] Bien sûr, il ne s'agit pas exactement d'une mesure de perplexité qui devrait être calculée à partir d'une distribution de probabilité.

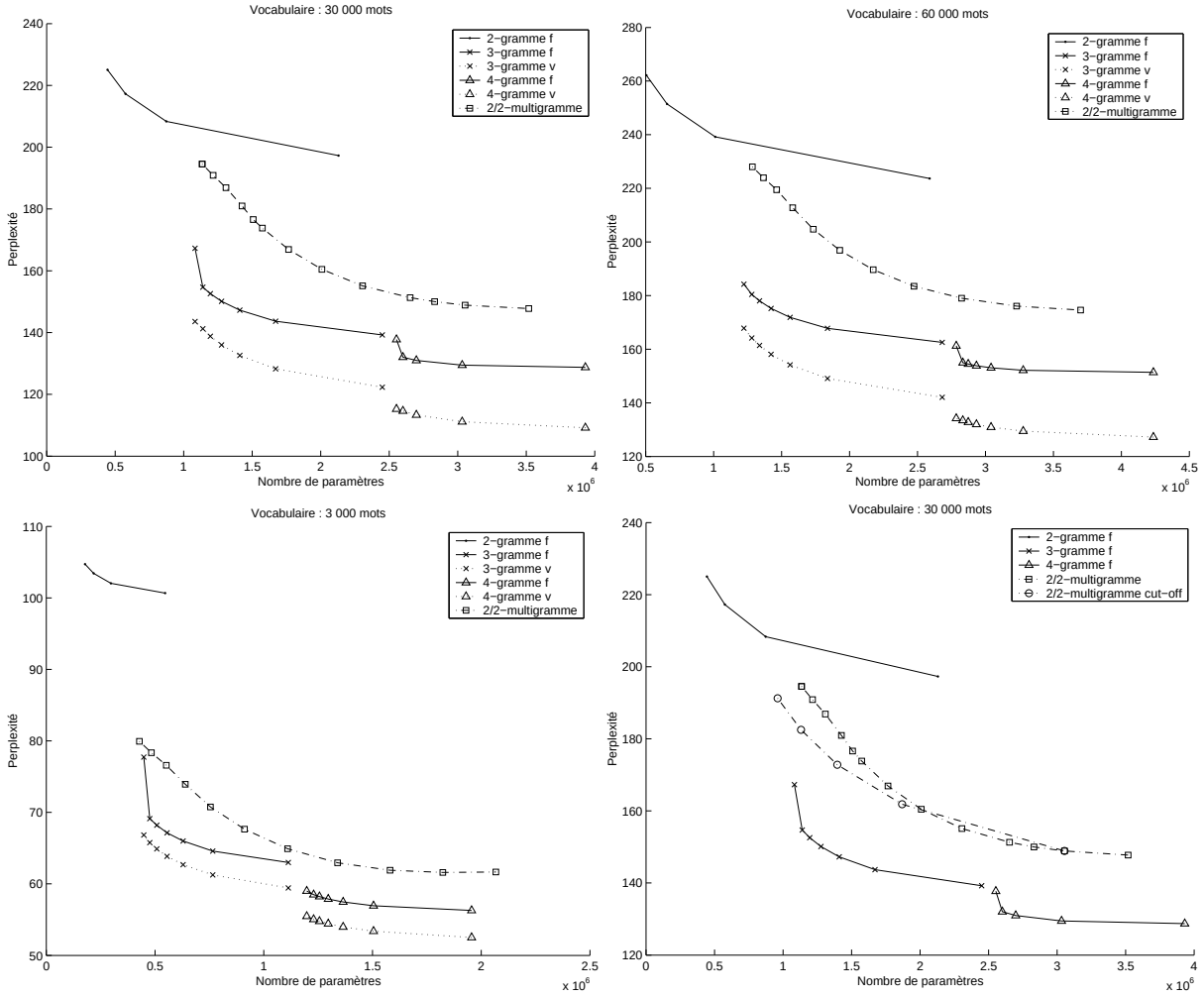


Figure 1: Comparaison de l'influence du nombre de paramètres sur la perplexité des modèles de $n$-gramme à horizon fixe ($n$-gramme-f) ou variable($n$-gramme-v) pour $n \in [2, 4]$, et du $2/2$-multigramme pour différentes tailles de vocabulaire, et influence sur la perplexité de la méthode de *cut-off* pour réduire les paramètres du modèle de $2/2$-multigramme avec un vocabulaire de 30 000 mots.

Pour améliorer la situation, on peut tout d'abord chercher à n'inclure dans les multigramme autorisés que ceux qui apportent un gain vis à vis de l'équation5. Nous avons constaté par des expériences que ces multigramme n'améliorent pas significativement la perplexité (nous n'avons pas la place pour rapporter ces expériences). Cela semble indiquer que les multigramme qui apportent le plus gros gain en terme de perplexité sont déjà inclus dans la liste des plus fréquents. Une expérience similaire consiste à définir la liste des multigramme en changeant le seuil de *cut-off*. En effet, on peut observer une baisse significative du nombre de paramètres, qui s'accompagne d'une augmentation de la perplexité (environ 10 points) quand on passe d'un bi-gramme de mots avec un *cut-off* à 0 (respectivement 1) à un bi-gramme de mots avec un *cut-off* à 1 (respectivement 2). La figure 1 montre l'évolution de la perplexité en conservant les multigramme les plus fréquents (100 000 multigramme pour un vocabulaire de 30 000 mots). On peut constater une baisse du nombre de paramètres sans hausse de la perplexité ; cette solution semble donc convenir. Enfin, étant donnée la baisse significative de la perplexité observée avec peu de multigramme entre un bi-gramme de mots avec un *cut-off* à 1, et un $2/2$-multigramme avec le même *cut-off*. On peut souhaiter généraliser l'usage des multigramme aux $n$-gramme. Cependant la complexité risque d'augmenter de manière exponentielle avec $n$.

## 7 Conclusion

Cet article a présenté des résultats concernant des modèles de langage statistiques de type $n$-gramme à horizon fixe ou variable et des $n/m$-multigramme. À taux de mots hors-vocabulaire fixe, le comportement des $n$-gramme classiques fait baisser la perplexité pour des valeurs de $n$ de 3 à 4, mais au prix d'une baisse du nombre de mots prédits (environ 7 millions pour un modèle de bi-gramme, 6.8 millions pour un tri-gramme, et un peu plus de 6.6 millions pour un quadri-gramme). Plus on reconnaît des mots, plus la probabilité conjointe va être faible, on peut donc trouver discutable de comparer entre eux des modèles de $n$-gramme qui ne se trouvent pas tout à fait sur le même pied d'égalité. Ce problème ne se pose pas pour les $n$-gramme à horizon variable, ou les $n/m$-multigramme, car le nombre de mots prédits est à chaque fois celui du modèle de bi-gramme. Les résultats de perplexité obtenus avec des vocabulaires de taille plus importante nous montrent à la fois une augmentation de la perplexité, et une augmentation du nombre de paramètres. Cette augmentation est due encore une fois à une augmentation du nombre de mots prédits (pour un modèle de bi-gramme, nous avons près de 4.9 millions de mots prédits pour un vocabulaire de 3 000 mots, 7 millions pour 30 000 mots, et 7.3 millions pour 60 000 mots). Le taux de mots hors vocabulaire sur le corpus de test baisse de 19.38% pour 3 000 mots à 1.65% pour 60 000 mots. Nous avons montré que le modèle de multigramme le plus simple, un 2/2-multigramme (c'est-à-dire un bi-gramme de séquences comprenant au plus deux mots) se comporte comme un modèle situé entre un bi-gramme et un tri-gramme classique. Notre objectif consiste à pousser un peu plus loin ces modèles en augmentant notamment l'ordre et en réglant le nombre de paramètres par des techniques de cut-off.

## Références


BIMBOT, F., PIERACCINI, R., LEVIN, E., & ATAL, B. 1995. Variable-Length Sequence Modeling: Multigrams. *IEEE Signal Processing Letters*, **2**(6), 111–113.

BONAFONTE, A., & MARIÑO, J. 1996. Language Modeling Using X-grams. *Pages 394–397 of: Proceedings of the International Conference on Spoken Language Processing*.

CHEN, S.F., & GOODMAN, J. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, **13**(4), 359–394.

DELIGNE, S., & BIMBOT, F. 1995. Language modeling by variable length sequences: theoretical formulation and evaluation of multigrams. *In: IEEE International Conference on Acoustics and Speech Signal Processing*.

DELIGNE, S., & SAGISAKA, Y. 2000. Statistical language modeling with a class-based n-multigram model. *Computer Speech and Language*, **14**, 261–279.

KATZ, S.M. 1987. Estimation of Probailities from Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE transactions on Acoustics, Speech and Signal Processing*, **35**, 400–401.

NIESLER, T.R., & WOODLAND, P.C. 1994. Variabl-length category n-gram language models. *Computer Speech and Language*, **13**, 99–124.

ROSENFELD, R. 2000. Two decades of statistical language modeling: where do we go from here? *Proceedings of the IEEE*, **88**(8), 1270–1278.

SHIELDS, P.C. 1998. The Interactions Between Ergodic Theory and Information Theory. *IEEE Transactions on Information Theory*, **44**, 2079–2093.

SIU, M., & OSTENDORF, M. 2000. Variable n-grams and extensions for conversational speech language modeling. *IEEE transactions on Speech and Audio Processing*, **8**(1), 63–75.

WOODLAND, P.C., & YOUNG, S.J. 1993. The HTK Continuous Speech Recogniser. *Pages 2207–2219 of: Proceedings of the Eurospeech conference*.

ZITOUNI, I. 2002. A Hierarchical Language Model Based on Variable-Length Class Sequences: The $MC_n^\nu$ Approach. *IEEE Transactions on Speech and Audio Processing*, **10**(3), 193–198.