taille du vocabulaire N	*20k*	*30k*	*40k*	*50k*	*60k*	*65k*
taux de MHV sur $\mathcal{T}$ (dévelop.)	6.38	4.30	3.15	2.36	1.95	1.79

Table 1: Evolution du taux de MHV en fonction de la taille N du vocabulaire. Les vocabulaires sont définis comme les N mots les plus fréquents dans le corpus d'apprentissage T_0 (40M mots *Monde87-88*.)

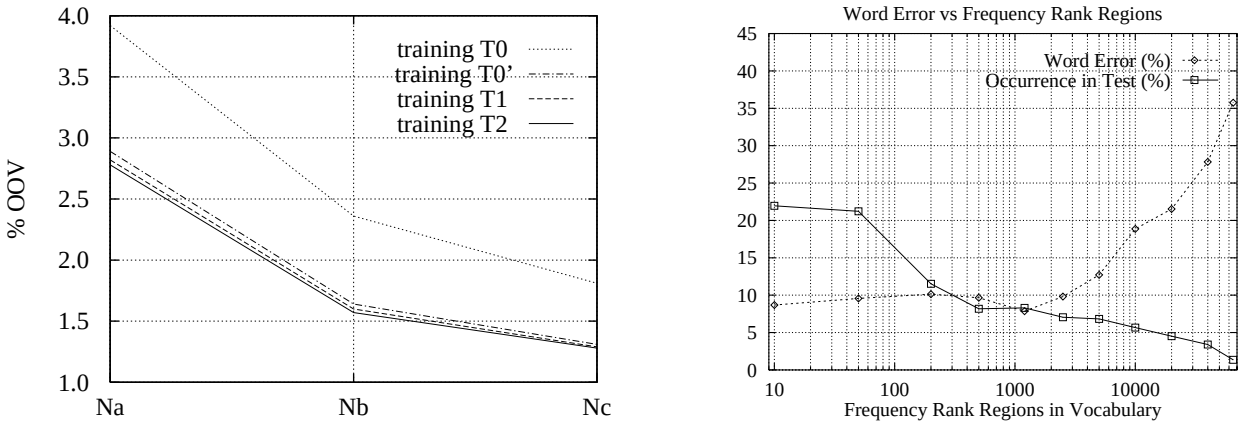

Figure 1: **(gauche)** Taux de MHV sur un texte de dévelopment pour 3 versions de normalisation N_a, N_b, N_c et des vocabulaires 65k mots à partir de données d'apprentissage T_0, T_0', T_1, T_2 du *Monde*, avec:
T_0 : années 1987-88 (40M mots, ressources communes dans le projet de Dictée Vocale del'AUPELF.)
T_0' : années 1994-95 (40M mots)
T_1 : années 1987-95 (185M mots)
T_2 : années 1991-95 (105M mots, T_2 moins de données que T_1, mais en moyenne plus récentes.)
(droite) Taux d'erreur et d'occurrences de mots en fonction de régions de rang fréquentiel (Frequency Rank Region: FFR) dans le vocabulaire 65k. Chaque point de l'axe des X définit la limite supérieure d'une région FFR. 11 FFR ont été définies et distribuées logarithmiquement de 1 à 65.000.

système	apprentissage ML	**MHV** ($\mathcal{T}$)	**Err.Mots** ($\mathcal{T}$)	**MHV** ($\mathcal{T}'$)	**Err.Mots** ($\mathcal{T}'$)
20k	T_0	6.38%	23.85%	3.60%	17.3%
$\widehat{20k}$	T_1 + *MD* + *AFP*	6.38%	21.76%	3.60%	14.6%
65k	T_1 + *MD* + *AFP*	1.34%	12.86%	0.45%	8.83%

Table 2: Résultats de reconnaissance avec des systèmes à 20k and 65k mots sur les données de développement. 20k et $\widehat{20k}$ diffèrent par la quantité d'apprentissage des ML: 40M et 255M de mots respectivement (*MD=LeMondeDiplomatique, AFP=AgenceFrancePresse*). $\mathcal{T}$ comprend les 600 phrases de développement, alors que $\mathcal{T}'$ correspond au sous-ensemble de 300 phrases à taux de MHV contrôlé.