	audio			*textes*	
Langue	*sources radio-télé*			journaux	trans.com.
Anglais	ABC, CNN, CSPAN, NPR, PRI, VOA	200h	1,9Mmots	790Mmots	240Mmots
Français	Arte, TF1, A2, France-Info, France-Inter	50h	0,8Mmots	300Mmots	20Mmots
Allemand	Arte	20h	0,2Mmots	260Mmots	-
Mandarin	VOA, CCTV, KAZN	20h	0,7Mcar.	200Mcar.	-

Table 1 : Données audio avec transcriptions fines et données textuelles utilisées lors de l'estimation des modèles acoustiques et des modèles de langage. Pour les données textuelles on distingue les textes de journaux (*journaux*) et les transcriptions commerciales (*trans.com.*), en millions de mots (ou de caractères pour le mandarin).

		Lexique		*ML*		*Test*	
Langue	*#phon.*	*taille*	*couverture*	*N-gram*	*ppx*		*%Merr*
Angl.	48	65k (mots)	99.5%	24M tg, 9M bg	123	3,0h	20
Franç.	37	65k (mots)	98.8%	14M tg, 6M bg	107	3,0h	23
Allem.	51	65k (mots)	96.5%	11M tg, 8M bg	225	2,0h	25(j)-35(d)
Mand.	39	40k (m) /5k (car.)	99.7%	10M tg, 3M bg	201	1,5h	20

Table 2 : Pour chaque langue sont indiqués: le nombre de phones utilisés pour les prononciations, la taille du lexique ainsi que la couverture obtenue sur les données de tests, la taille des modèles de langage ainsi que la perplexité sur le test, puis la durée des données de test (en heures) et les taux d'erreurs de mot. Pour l'allemand on distingue entre (j) et (d) pour données journalistiques et documentaires.

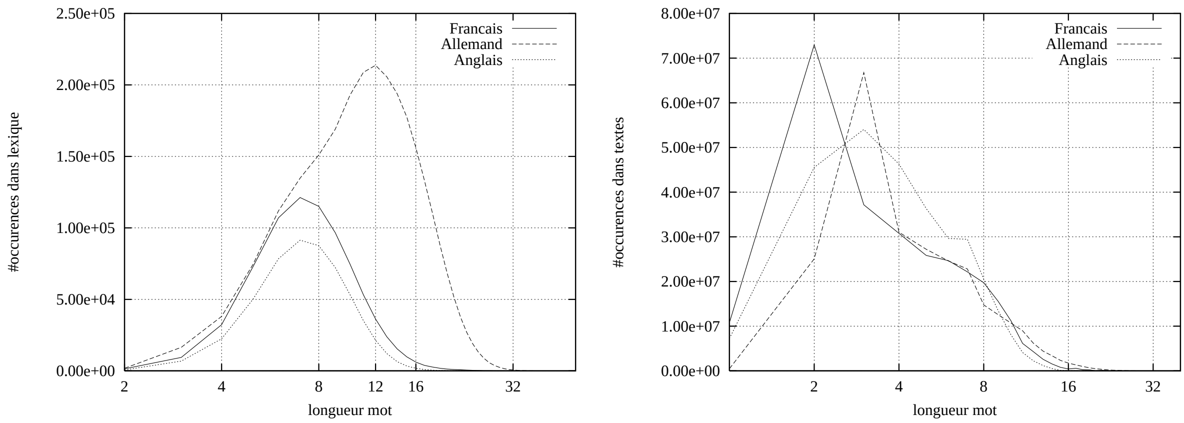

Fig 1 : Comparaison de la variété lexicale en fonction de la longueur des mots (en nombre de caractères) pour l'anglais, l'allemand et le français à partir de corpus de textes de journaux (environ 300M de mots par langue).
(Gauche) : Variété lexicale dans le lexique. L'anglais produit le plus faible nombre d'entrées lexicales. Le plus grand nombre de mots est observé autour de 7 caractères pour l'anglais et le français. L'allemand a une variété lexicale significativement plus élevée liée surtout à la composition de mots. La plus grand nombre d'entrées lexicales se situe pour l'allemand autour de 12 caractères.
(Droite) : Variété lexicale dans les textes. Pour toutes les langues les mots courts sont les plus fréquents, la majorité d'entre eux correspondant aux mots outils (articles, prépositions). Le français admet le plus grand nombre de mots courts avec un pic important pour une longueur de deux caractères. Pour l'allemand et l'anglais le pic correspond aux mots à 3 caractères.